IBM InfoSphere BigInsights
Version 2.1

Installation Guide

GC19-4100-00

IBM InfoSphere BigInsights
Version 2.1

Installation Guide

GC19-4100-00

Note

Before using this information and the product that it supports, read the information in "Notices and trademarks" on page 89.

Contents

Chapter 1. Introduction to InfoSphere BigInsights

InfoSphere® BigInsights™ is a software platform for discovering, analyzing, and visualizing data from disparate sources. You use this software to help process and analyze the volume, variety, and velocity of data that continually enters your organization every day.

InfoSphere BigInsights helps your organization to understand and analyze massive volumes of unstructured information as easily as smaller volumes of information. The flexible platform is built on an Apache Hadoop open source framework that runs in parallel on commonly available, low-cost hardware. You can easily scale the platform to analyze hundreds of terabytes, petabytes, or more of raw data that is derived from various sources. As information grows, you add more hardware to support the influx of data.

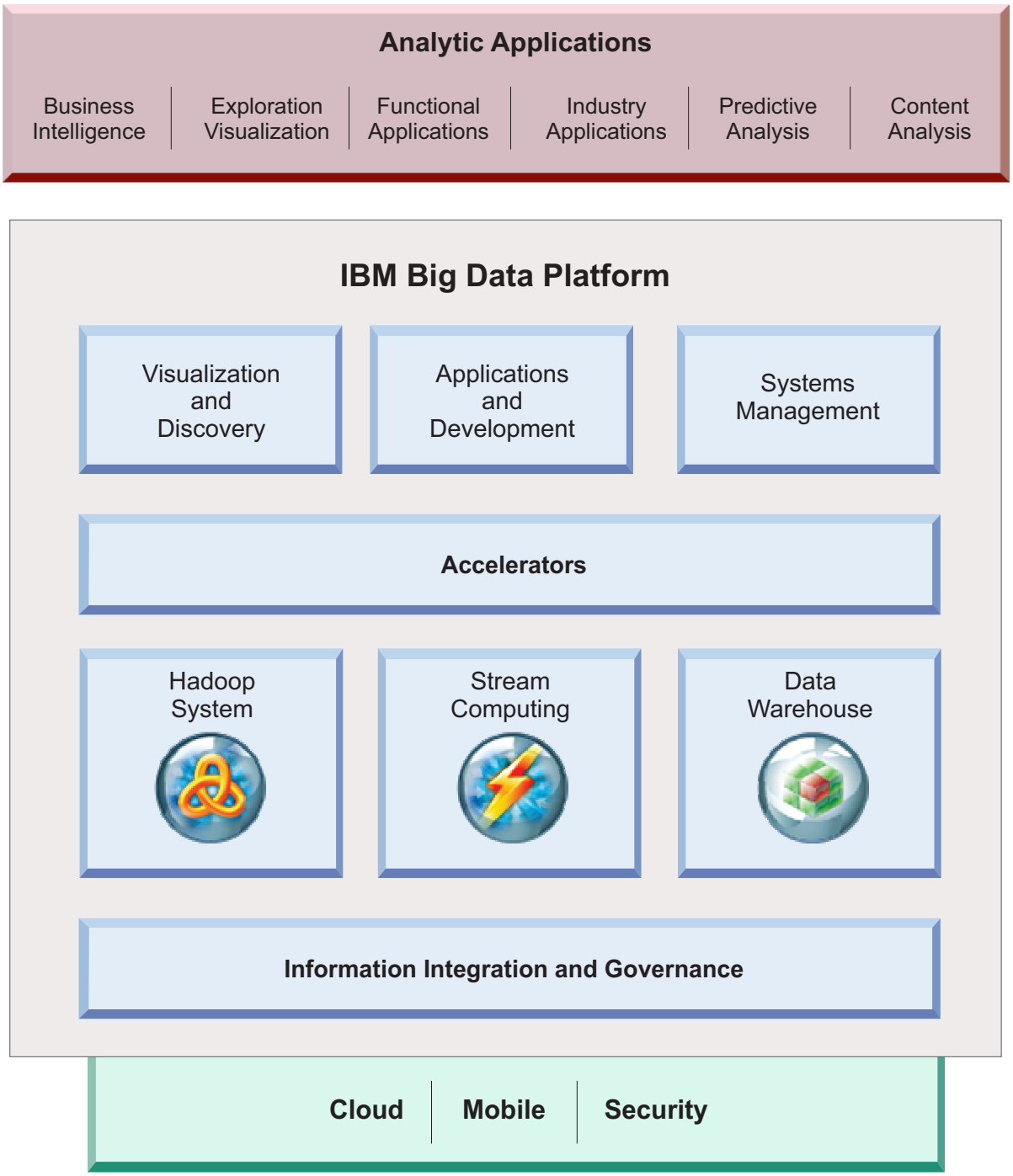
InfoSphere BigInsights helps application developers, data scientists, and administrators in your organization quickly build and deploy custom analytics to capture insight from data. This data is often integrated into existing databases, data warehouses, and business intelligence infrastructure. By using InfoSphere BigInsights, users can extract new insights from this data to enhance knowledge of your business.

InfoSphere BigInsights incorporates tooling for numerous users, speeding time to value and simplifying development and maintenance:

- Software developers can use the Eclipse-based plug-in to develop custom text analytic functions to analyze loosely structured or largely unstructured text data.
- Administrators can use the web-based management console to inspect the status of the software environment, review log records, assess the overall health of the system, and more.
- Data scientists and business analysts can use the data analysis tool to explore and work with unstructured data in a familiar spreadsheet-like environment.

InfoSphere BigInsights features and architecture

InfoSphere BigInsights provides distinct capabilities for discovering and analyzing business insights that are hidden in large volumes of data. These technologies and features combine to help your organization manage data from the moment that it enters your enterprise.



By combining these technologies, InfoSphere BigInsights extends the Hadoop open source framework with enterprise-grade security, governance, availability, integration into existing data stores, tools that simplify developer productivity, and more.

Hadoop is a computing environment built on top of a distributed, clustered file system that is designed specifically for large-scale data operations. Hadoop is designed to scan through large data sets to produce its results through a highly scalable, distributed batch processing system. Hadoop comprises two main components: a file system, known as the Hadoop Distributed File System (HDFS),

and a programming paradigm, known as Hadoop MapReduce. To develop applications for Hadoop and interact with HDFS, you use additional technologies and programming languages such as Pig, Hive, Jaql, Flume, and many others.

Apache Hadoop helps enterprises harness data that was previously difficult to manage and analyze. InfoSphere BigInsights features Hadoop and its related technologies as a core component.

Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) allows applications to run across multiple servers. HDFS is highly fault tolerant, runs on low-cost hardware, and provides high-throughput access to data.

Data in a Hadoop cluster is broken into smaller pieces called *blocks*, and then distributed throughout the cluster. Blocks, and copies of blocks, are stored on other servers in the Hadoop cluster. That is, an individual file is stored as smaller blocks that are replicated across multiple servers in the cluster.

Each HDFS cluster has a number of *DataNodes*, with one DataNode for each node in the cluster. DataNodes manage the storage that is attached to the nodes on which they run. When a file is split into blocks, the blocks are stored in a set of DataNodes that are spread throughout the cluster. DataNodes are responsible for serving read and write requests from the clients on the file system, and also handle block creation, deletion, and replication.

Also on each HDFS cluster is a single *NameNode*, which is a primary server that regulates access to files by clients, and tracks all data files in HDFS. The NameNode determines the mapping of blocks to DataNodes, and handles operations such as opening, closing, and renaming files and directories. All of the information for the NameNode is stored in memory, which allows for quick response times when adding storage or reading requests. The NameNode is the repository for all HDFS metadata, and user data never flows through the NameNode.

The NameNode and DataNodes run on computers that run the RedHat Linux, Linux on pSeries, or SUSE Linux operating systems. A typical HDFS deployment has a dedicated computer that runs only the NameNode, because the NameNode stores metadata in memory. If the computer that runs the NameNode fails, then metadata for the entire cluster is lost, so this computer is typically more robust than others in the cluster.

IBM General Parallel File System

IBM® General Parallel File System (GPFS™) is an enterprise file system that InfoSphere BigInsights supports as an alternative to HDFS.

GPFS supports local disks on cluster nodes and storage area networks (SANs). Logical isolation and physical isolation are supported so that file sets can be separate file systems inside of a file system (logical isolation), or can be part of separate storage pools (physical isolation). InfoSphere BigInsights uses a customized version of GPFS that supports all existing GPFS commands, and provides additional interfaces and commands.

GPFS supports thousands of nodes and petabytes of storage so that you can modify the scale to meet your most demanding needs. Data is replicated on multiple nodes so that no single point of failure exists, whereas the NameNode is a

single point of failure in HDFS. You can push updates asynchronously or synchronously allowing you to choose how you want to manage changes from a primary system to a secondary system.

If a node fails, changes are replicated to other nodes. When the failed node is operational, GPFS quickly determines which blocks must be recovered. Changes that occured while the node was down are copied to the previously failed node so that the node is syncrhonized with other nodes in the cluster.

Applications define their own logical block size by chunking data into file blocks. Each file block is determined based on the effective block size or chunk size. Applications can also determine replication layout by using either wide striping over the network, write affinity on a local disk, or a combination of both layouts. Allowing applications to dictate block size and replication layout provides greater performance and efficiency over HDFS.

GPFS includes several enterprise features that provide distinct advantages, such as the capability to take a logical, read-only copy of the file system at any point in time. You can use this snapshot for backing up your system and recovering from errors. The snapshot file does not occupy disk space until it is modified or delete, providing an efficient backup and recovery solution.

GPFS has Active File Management (AFM/Panache) support to enable asynchronous access and control of local and remote files.

The full set of GPFS documentation provides more detail and examples for existing GPFS commands and is available in the Cluster Products Information Center.

Note: GPFS is not included as part of the InfoSphere BigInsights Quick Start Edition.

In addition, if you install InfoSphere BigInsights and select GPFS as the distributed file system, no additional high availability feature is required, since GPFS has a built-in `NameNode`-equivalent high availability feature.

The following table provides a comparison of GPFS features to HDFS features:

Table 1. Comparison of GPFS features and HDFS features. The table includes key features and how they apply to GPFS and HDFS. GPFS includes several features that distinguish it from HDFS and provide extended capabilities.

Features	GPFS	HDFS
Hierarchical storage management	Allows sufficient usage of disk drives with different performance characteristics	
High performance support for MapReduce applications	Stripes data across disks by using metablocks, which allows a MapReduce split to be spread over local disks	Places a MapReduce split on one local disk

Table 1. Comparison of GPFS features and HDFS features (continued). The table includes key features and how they apply to GPFS and HDFS. GPFS includes several features that distinguish it from HDFS and provide extended capabilities.

Features	GPFS	HDFS
High performance support for traditional applications	• Manages metadata by using the local node when possible rather than reading metadata into memory unnecessarily • Caches data on the client side to increase throughput of random reads • Supports concurrent reads and writes by multiple programs • Provides sequential access that enables fast sorts, improving performance for query languages such as Pig and Jaql	
High availability	Has no single point of failure because the architecture supports the following attributes: • Distributed metadata • Replication of both metadata and data • Node quorums • Automatic distributed node failure recovery and reassignment	Has a single point of failure on the NameNode, which requires it to run in a separate high availability environment
POSIX compliance	Is fully POSIX compliant, which provides the following benefits: • Support for a wide range of traditional applications • Support for UNIX utilities, that enable file copying by using FTP or SCP • Updating and deleting data • No limitations or performance issues when using a Lucene text index	Is not POSIX compliant, which creates the following limitations: • Limited support of traditional applications • No support of UNIX utilities, which requires using the **`hadoop dfs get`** command or the **`put`** command to copy data • After the initial load, data is read-only • Lucene text indexes must be built on the local file system or NFS because updating, inserting, or deleting entries is not supported
Ease of deployment	Supports a single cluster for analytics and databases	Requires separate clusters for analytics and databases

Table 1. Comparison of GPFS features and HDFS features (continued). The table includes key features and how they apply to GPFS and HDFS. GPFS includes several features that distinguish it from HDFS and provide extended capabilities.

Features	GPFS	HDFS
Other enterprise level file system features	• Durable sync • Workload isolation • Security	
Data replication	Provides cluster-to-cluster replication over a wide area network	

Adaptive MapReduce

The adaptive MapReduce component is capable of running distributed application services on a scalable, shared, heterogeneous grid. This low-latency scheduling solution supports sophisticated workload management capabilities beyond those of standard Hadoop MapReduce.

Note: Adaptive MapReduce is not included as part of the InfoSphere BigInsights Quick Start Edition.

The adaptive MapReduce component can orchestrate distributed services on a shared grid in response to dynamically changing workloads. This component combines a service-oriented application middleware (SOAM) framework, a low-latency task scheduler, and a scalable grid management infrastructure. This design ensures application reliability while also ensuring low-latency and high-throughput communication between clients and compute services.

Hadoop has limited prioritization features, whereas the adaptive MapReduce component has thousands of priority levels and multiple options that you can configure to manage resource sharing. This sophisticated resource sharing allows you to prioritize for interactive workloads that are not possible in a traditional MapReduce environment. For example, with the adaptive MapReduce component, you can start multiple Hadoop jobs and associate those jobs with the same consumer. Within that consumer, jobs can share resources based on individual priorities.

For example, consider a 100 slot cluster where you start job “A” with a priority of 100. Job “A” starts, and consumes all slots if enough map tasks exist. You then start job “B” while job “A” is running, and give job “B” a priority of 900, which is nine times greater than the priority of job “A”. The adaptive MapReduce component automatically rebalances the cluster to give 90 slots to job “B” and 10 slots to job “A”, so that resources are distributed in a prioritized manner that is transparent to the jobs.

The adaptive MapReduce component is not a Hadoop distribution. It relies on a MapReduce implementation that includes Hadoop components like Pig, Hive, HBase, and a distributed file system. The scheduling framework is optimized for MapReduce workloads that are compatible with Hadoop. Because InfoSphere BigInsights is built on Hadoop, you can use the adaptive MapReduce component as a workload scheduler for InfoSphere BigInsights instead of the standard MapReduce scheduler. When coupled with InfoSphere BigInsights, the adaptive MapReduce component transparently provides improved performance at a lower cost of a variety of big data workload managers.

Hadoop MapReduce

MapReduce applications can process large data sets in parallel by using a large number of computers, known as clusters.

In this programming paradigm, applications are divided into self-contained units of work. Each of these units of work can be run on any node in the cluster. In a Hadoop cluster, a MapReduce program is known as a *job*. A job is run by being broken down into pieces, known as *tasks*. These tasks are scheduled to run on the nodes in the cluster where the data exists.

Applications submit jobs to a specific node in a Hadoop cluster, which is running a program known as the *JobTracker*. The JobTracker program communicates with the NameNode to determine where all of the data required for the job exists across the cluster. The job is then broken into `map` tasks and `reduce` tasks for each node in the cluster to work on. The JobTracker program attempts to schedule tasks on the cluster where the data is stored, rather than sending data across the network to complete a task. The MapReduce framework and the Hadoop Distributed File System (HDFS) typically exist on the same set of nodes, which enables the JobTracker program to schedule tasks on nodes where the data is stored.

As the name MapReduce implies, the `reduce` task is always completed after the `map` task. A MapReduce job splits the input data set into independent chunks that are processed by `map` tasks, which run in parallel. These bits, known as *tuples*, are key/value pairs. The `reduce` task takes the output from the `map` task as input, and combines the tuples into a smaller set of tuples.

A set of programs that run continuously, known as *TaskTracker* agents, monitor the status of each task. If a task fails to complete, the status of that failure is reported to the JobTracker program, which reschedules the task on another node in the cluster.

This distribution of work enables `map` tasks and `reduce` tasks to run on smaller subsets of larger data sets, which ultimately provides maximum scalability. The MapReduce framework also maximizes parallelism by manipulating data stored across multiple clusters. MapReduce applications do not have to be written in Java™, though most MapReduce programs that run natively under Hadoop are written in Java.

Additional Hadoop technologies

Hadoop includes many open source technologies that continue to increase as Hadoop is used in more applications. You use these technologies to interact with Hadoop and the Hadoop Distributed File System (HDFS).

HDFS is not compliant with the Portable Operating System Interface for UNIX (POSIX) file system, which means that typical interactions such as copying, creating, moving, and deleting files is not supported. Therefore, you must use HDFS interfaces, technologies, and APIs to interact with files in HDFS. You must write your own applications to interact with files, or learn the different HDFS commands to manage and manipulate files in the file system.

The following open source technologies are components of Hadoop that are included with InfoSphere BigInsights.

Avro A data serialization system that includes a schema within each file. A schema defines the data types that are contained within a file, and is

validated as the data is written to the file using the Avro APIs. Users can include *primary data types* and *complex type definitions* within a schema. When Avro data is written to a file, the schema that defines the data is written with it, so that files can be processed by any application. Data can be versioned by changes in the schema, because the latest version is included in addition to the older versions.

Version 1.7.2 is included in this release.

Chukwa
A data collection system for monitoring large distributed file systems. Chukwa is built on the HDFS and MapReduce framework, and includes a toolkit for displaying results from monitoring and analysis.

Version 0.5.0 is included in this release.

Flume A distributed, reliable, and highly available service for efficiently moving large amounts of data in a Hadoop cluster. Flume helps users to aggregate data from many sources, manipulate the data, and then add the data into your Hadoop environment. The entities that you work with in Flume are *sources*, *decorators*, and *sinks*. A source can be any data source, a decorator is a transformation operation that is completed on the data stream, and a sink is the target of a specific operation.

Version 1.3.0 is included in this release.

HBase A column-oriented database management system that runs on top of HDFS and is often used for sparse data sets. Unlike relational database systems, HBase does not support a structured query language like SQL. HBase applications are written in Java, much like a typical MapReduce application. HBase allows many attributes to be grouped into column families so that the elements of a column family are all stored together. This approach is different from a row-oriented relational database, where all columns of a row are stored together.

Version 0.94.3 is included in this release.

HCatalog
A table and storage management service for Hadoop data that presents a table abstraction so that you do not need to know where or how your data is stored. You can change how you write data, while still supporting existing data in older formats. HCatalog wraps additional layers around the Hive metadata store to provide an enhanced metadata service that includes functions for both MapReduce and Pig. Because HCatalog uses the Hive data model, you can use these functions to interact directly with MapReduce and Pig without translating the data type.

Version 0.4.0 is included in this release.

Hive A data warehouse infrastructure that facilitates data extract-transform-load (ETL) operations, in addition to analyzing large data sets that are stored in the Hadoop Distributed File System (HDFS). SQL developers write statements, which are broken down by the Hive service into MapReduce jobs, and then run across a Hadoop cluster. InfoSphere BigInsights includes a JDBC driver that is used for programming with Hive and for connecting with Cognos Business Intelligence software.

Version 0.9.0 is included in this release.

Jaql Query language that is designed for JavaScript Object Notation (JSON) and that is primarily used to analyze large-scale, semistructured data. Jaql helps users to select, join, group, and filter data that is stored in HDFS,

making it seem like a combination of Pig and Hive. Jaql can support additional semistructured data sources such as SQL, XML, CSV, and flat files.

Jaqlserver is deprecated as of InfoSphere BigInsights Version 2.0. To run the ad-hoc Jaql application, use the InfoSphere BigInsights REST APIs.

Lucene
A high-performance text search engine library that is written entirely in Java. When you search within a collection of text, Lucene breaks the documents into text fields and builds an index from them. The index is the key component of Lucene that forms the basis of rapid text search capabilities. You use the searching methods within the Lucene libraries to find text components. With InfoSphere BigInsights, Lucene is integrated into Jaql, providing the ability to build, scan, and query Lucene indexes within Jaql.

Version 3.3.0 is included in this release.

Oozie A management application that simplifies workflow and coordination between MapReduce jobs. Oozie provides users with the ability to define actions and dependencies between actions. Oozie then schedules actions to run when the required dependencies are met. Workflows can be scheduled to start based on a given time or based on the arrival of specific data in the file system.

Version 3.2.0 is included in this release.

Pig A programming language that is designed to handle any type of data. Pig helps users to focus more on analyzing large data sets and less time writing map programs and reduce programs. Pig comprises two components: PigLatin, the programming language, and a runtime environment where PigLatin programs run.

Version 0.10.0 is included in this release.

Sqoop A tool designed to easily import information from structured databases (such as SQL) and related Hadoop systems (such as Hive and HBase) into your Hadoop cluster. You can also use Sqoop to extract data from Hadoop and export it to relational databases and enterprise data warehouses.

Version 1.4.2 is included in this release.

ZooKeeper
A centralized infrastructure and set of services that enable synchronization across a cluster. ZooKeeper maintains common objects that are needed in large cluster environments, such as configuration information, distributed synchronization, and group services. Many other open source projects that use Hadoop clusters require these cross-cluster services. Having these services available in ZooKeeper ensures that each project can embed ZooKeeper without having to build new synchronization services into each project.

Version 3.4.5 is included in this release.

Text Analytics

To help your organization gain insight into vast repositories of text data, InfoSphere BigInsights includes Text Analytics, which extract information from unstructured and semistructured data. By using Text Analytics, your organization can analyze large volumes of text and produce annotated documents that provide valuable insights into unconventional data.

The goal of Text Analysis is to read unstructured text and distill insights. This process can involve searching for customer web browsing patterns in clickstream log files, finding fraud indicators through email analytics, or assessing customer sentiment from social media messages. To help facilitate text analysis, InfoSphere BigInsights includes developer tools, a programming language, a processing engine, and prebuilt text extractors. Built-in multilingual support for tokenization and semantic analysis enables InfoSphere BigInsights to understand multiple languages and different parts of speech.

A text analysis application can read a paragraph of text and derive structured information based on various rules. These rules are defined in *extractors*, which are programs that extract information from within a text field. The product of extractors is a set of annotated text that includes specific information that is important to your business. Extractors are written in the annotation query language (AQL), which is the core of Text Analytics in InfoSphere BigInsights.

AQL is a fully declarative Text Analytics language, which means that any module can be customized. This implementation results in a flexible Text Analytics language that is both highly expressive and fast. InfoSphere BigInsights includes an Eclipse-based development environment for creating and maintaining extractors written in AQL. By supporting modular AQL, InfoSphere BigInsights improves the usability and development productivity of Text Analytics applications by enabling and reusing common extractors, dictionaries, rules, and tables.

An optimizer is also included that generates an efficient runtime plan for the extractor specification. The runtime component receives a collection of documents and runs the extractor on each document to extract text entities. After extraction, you can test extractors against a subset of data to refine the precision and recall of the extractors. To begin building your own extractors, you can use the rich library of precompiled extractors that are included with InfoSphere BigInsights.

After you develop text analysis applications, you can deploy, run, and monitor these applications from the InfoSphere BigInsights console. You can link text analysis applications to other applications in the console, including integration with BigSheets. By linking applications, you can create new applications that include multiple analytic components.

IBM Big SQL

IBM Big SQL is a data warehouse system for Hadoop that you use to summarize, query, and analyze data. Big SQL provides SQL access to data that is stored in InfoSphere BigInsights by using JDBC or ODBC connections.

Big SQL provides broad SQL support that is typical of commercial databases. You can issue JDBC queries or ODBC queries to access data that is stored in InfoSphere BigInsights, in the same way that you access databases from your enterprise applications.

You can use Big SQL to issue queries of any size, including concurrent queries. Big SQL provides support for large ad-hoc queries by using MapReduce parallelism and point queries, which are low-latency queries that return information quickly to reduce response time and provide improved access to data.

The Big SQL server is multi-threaded, so scalability is limited only by the performance and number of cores in the computer that runs the server. If you

want to issue larger queries, you can increase the hardware performance of the server computer that Big SQL runs on, or chain multiple Big SQL servers together to increase throughput.

InfoSphere BigInsights Console

InfoSphere BigInsights includes an integrated console that you can use to view the health of your cluster, deploy applications, manage cluster instances, work with BigSheets, manage files, and schedule workflows, jobs, and tasks from a single location.

The console includes quick access to tasks for system administration, application deployment, data analysis, and file management. You can navigate between the tasks by using the navigation tabs and access common tasks from the Welcome panel to begin working with InfoSphere BigInsights.

You can create dashboards to monitor the status of multiple applications services, data services, and attributes of your cluster on a single page. In addition, you can monitor the overall status of your cluster, including the status of individual Hadoop components.

You can deploy and run applications from the console to make data available for analysis in BigSheets. BigSheets is a browser-based visualization tool that you can use to extend the scope of your business intelligence data. By using BigSheets, you can easily integrate massive amounts of unstructured data into consumable business contexts for specific situations.

InfoSphere BigInsights Tools for Eclipse

The InfoSphere BigInsights Tools for Eclipse add functionality to your Eclipse environment so that you can develop programs to run with InfoSphere BigInsights.

The Welcome page of the InfoSphere BigInsights console includes information about how to enable your Eclipse environment for developing applications by using InfoSphere BigInsights.

You can develop and test InfoSphere BigInsights programs from your Eclipse environment and publish applications that contain workflows, text analytics modules, BigSheets readers and functions, and Jaql modules to your cluster. After you deploy applications to your cluster, you can run them from the InfoSphere BigInsights console.

You can use the InfoSphere BigInsights Tools for Eclipse to develop programs that interact with Hadoop and the Hadoop Distributed File System (HDFS) and the General Parallel File System (GPFS):

- Create text analytics modules that contain text extractors by using an extraction task wizard and editor. You can then test the extractor by running it locally against sample data. Visualize the results of the text extraction and improve the quality of your extractor by analyzing how results were obtained.
- Create Jaql scripts or modules by using a wizard, and edit scripts with an editor that provides content assistance and syntax highlighting. Run Jaql `explain` statements in your scripts, and run the scripts locally or against the InfoSphere BigInsights server. You can open the Jaql shell from within Eclipse to run Jaql statements against the cluster.
- Create Pig scripts by using a wizard and edit the scripts with an editor that provides content assistance and syntax highlighting. Run Pig `explain` statements

and `illustrate` statements for aliases in your scripts, and then run the Pig scripts locally or against the InfoSphere BigInsights server. You can open the Pig shell from within Eclipse to run Pig statements against the cluster.

- Connect to the Hive server by using the Hive JDBC driver and run Hive SQL scripts and explore the results. Browse the navigation tree to explore the structure and content of the tables in the Hive server.
- Use the Java editor to write programs that use MapReduce, and then run these programs locally or against the InfoSphere BigInsights server. Open the InfoSphere BigInsights console to monitor jobs that are created by MapReduce.
- Create templates for BigSheets readers or functions and then use the Java editor to implement the classes.
- Write Java programs that use the HBase APIs and run them against the InfoSphere BigInsights server. Open the HBase shell from your Eclipse environment to run HBase statements against the cluster.

Integration with other IBM products

InfoSphere BigInsights compliments and extends existing business capabilities by integrating with other IBM products. These integration points extend existing technologies to encompass more comprehensive information types, enabling a complete view of your business.

IBM Cognos® Business Intelligence
By using the InfoSphere BigInsights Big SQL JDBC driver or the Hive JDBC driver, Cognos Business Intelligence can incorporate data from InfoSphere BigInsights into business intelligence analysis, reports, and statistics. You can use a business intelligence modeler to create Cognos reports, dashboards, and workspaces while using the InfoSphere BigInsights MapReduce capabilities.

IBM DB2®
You use the IBM DB2 for Linux, UNIX, and Windows user-defined functions to query InfoSphere BigInsights data and run jobs from DB2 for Linux, UNIX, and Windows. InfoSphere BigInsights and IBM Netezza have optimized JDBC connections that you can access by using Jaql database modules. From InfoSphere BigInsights, you use a Jaql server or Jaql module on the cluster to run SQL with DB2 for Linux, UNIX, and Windows in parallel. You can also use a generic Jaql module to connect with third-party databases such as Derby, Informix, Oracle, and Teradata. You can use the Big SQL LOAD FROM statement to import data from a DB2 for Linux, UNIX, and Windows table by using a JDBC driver into a Big SQL table.

IBM InfoSphere Data Explorer
InfoSphere Data Explorer enables navigation of data from existing enterprise information systems and the advanced analytics that are provided by InfoSphere BigInsights components. InfoSphere Data Explorer Engine servers can receive data in real time from a cluster of InfoSphere BigInsights servers or InfoSphere Streams servers. InfoSphere Data Explorer can also push relevant data to users of information applications that are created by using the InfoSphere Data Explorer Application Builder. InfoSphere Data Explorer also enables federated access to other IBM products, such as IBM Netezza, and provides a flexible mechanism for making data available for exploration and analysis.

IBM InfoSphere DataStage®
The connector to InfoSphere DataStage (shipped with that product) enables

InfoSphere BigInsights to quickly exchange data with any software product that can connect with InfoSphere DataStage. InfoSphere DataStage includes a stage that uses InfoSphere BigInsights data within a job, enabling powerful and flexible extract-transform-load (ETL) scenarios.

IBM InfoSphere Guardium®
Integration with InfoSphere Guardium provides InfoSphere BigInsights with enterprise-level security auditing capabilities that support many current compliance mandates. This integration provides insights into which users are running specific InfoSphere BigInsights requests, which MapReduce jobs each user is running, and whether file permission exceptions are being caused by restricted access to sensitive data.

InfoSphere Streams
The adapter for InfoSphere Streams enables streaming data to be stored directly in your InfoSphere BigInsights cluster, and for InfoSphere Streams applications to read data from the cluster. You can use this integration to create an infrastructure that responds to events in real time and includes existing data to inform the response. You can also use InfoSphere Streams as a large-scale data engine to filter a stream of data that you want to store in the cluster.

IBM Netezza
InfoSphere BigInsights includes a connector that enables bidirectional data exchange between an InfoSphere BigInsights cluster and an IBM Netezza appliance. The adapter is implemented as a Jaql module, which helps you to use the simplicity and flexibility of Jaql in your database interactions. You use this adapter to read from or write to DB2 for Linux, UNIX, and Windows tables and IBM Netezza tables. You can use the Big SQL LOAD FROM statement to import data from an IBM Netezza table by using a JDBC driver into a Big SQL table.

Rational and Data Studio
InfoSphere BigInsights integrates with IBM Rational Application Developer Version 8.5, Rational Team Concert, and Data Studio by installing the InfoSphere BigInsights Tools for Eclipse into your existing Rational Data Studio, Rational Team Concert, or Data Studio environment. This integration provides a more collaborative application development for big data.

Chapter 2. Planning to install InfoSphere BigInsights

Before you install InfoSphere BigInsights, review the system requirements, read the release notes, design your installation topology, and review the installation checklists.

Reviewing system requirements and release notes

Ensure that all computers meet the hardware and software requirements for the product components that you want to install.

About this task

To obtain the most current information about the installation requirements for InfoSphere BigInsights, see the following resources:

- The system requirements on ibm.com.
- The release notes in the InfoSphere BigInsights Information Center.

Directories created when installing InfoSphere BigInsights

When you install InfoSphere BigInsights, directories are created for each of the product components. These directories contain files that are required for the components to function properly.

Two environment variables point to the location of critical files for installed components.

$BIGINSIGHTS_HOME
The $BIGINSIGHTS_HOME environment variable points to the location where required files are created for each component. This environment variable is created by the installation program when you install InfoSphere BigInsights. The default directory is /opt/ibm/biginsights.

$BIGINSIGHTS_VAR
The $BIGINSIGHTS_VAR environment variable points to the location where log files are stored for each component. This environment variable is created by the installation program when you install InfoSphere BigInsights. The default directory is /var/ibm/biginsights.

The default log directory for Jaql, Hive, and Pig is /var/ibm/biginsights/*component_name*/logs/*user_name*/.

component_name is the name of the product component, such as Jaql.

user_name is the name of the user who installed InfoSphere BigInsights.

Component directories

The directories in the following table are created under the $BIGINSIGHTS_HOME/*directory_name* directory, where /*directory_name* is the name of the directory for the specified component.

Table 2. Directories where components and their associated files are created

Component	Directory	Directory contents	Deployed location
IBM Big SQL	`/bigsql`	Big SQL files and directories, including scripts, server configuration files, server jar files, JDBC and ODBC drivers, the Big SQL command line client, error messages, and the keystore file for SSL encryption between client and server (if SSL is enabled)	All cluster nodes
IBM Hadoop Cluster	`/jdk`	IBM Java Development Kit	All cluster nodes
	`/IHC`	IBM Hadoop Cluster	All cluster nodes
	`/hadoop-conf`	Hadoop configuration file that is used by IBM Hadoop Cluster	All cluster nodes
BigIndex	`/bigindex`	BigIndex, a runtime library for the web search engine	All cluster nodes
BigSheets	`/sheets`	BigSheets, a browser-based visualization tool	Management node
Text analytics	`/text-analytics`	Text analytics, a system for extracting information from text data	All cluster nodes
InfoSphere BigInsights Console	`/console`	InfoSphere BigInsights Console, a web interface that is the central management of your cluster	Management node
Derby	`/derby`	Derby, a Java relational database management system	All cluster nodes
Flume	`/flume`	Flume, a service for moving large amounts of data	All cluster nodes
HBase	`/hbase`	HBase, a non-relational distributed database	All cluster nodes
HCatalog	`/hcatalog`	HCatalog, a table and storage management service for Hadoop data	All cluster nodes

Table 2. Directories where components and their associated files are created (continued)

Component	Directory	Directory contents	Deployed location
Hive	`/hive`	Hive, a data warehouse system for Hadoop	All cluster nodes
Jaql	`/jaql`	Jaql, a query language that is designed for JavaScript Object Notation (JSON)	All cluster nodes
Jaqlserver	`/jaqlserver`	Jaqlserver, a Jaql user-defined function (UDF) access gateway to Jaql Jaqlserver is deprecated as of InfoSphere BigInsights Version 2.0. To run the ad-hoc Jaql application, use the InfoSphere BigInsights REST APIs.	Management node
Lucene	`/lucene`	Lucene, a high-performance text search engine library	All cluster nodes
Monitoring	`/monitoring`	Monitoring, an agent-based component for collecting and processing monitoring data against Hadoop components	All cluster nodes
Oozie	`/oozie`	Oozie, a workflow coordination service that manages data processing jobs	All cluster nodes
Pig	`/pig`	Pig, a platform for analyzing large data sets, comprised of a programming language for expressing data analysis programs (PigLatin), and a runtime environment	All cluster nodes
Sqoop	`/sqoop`	Sqoop, a tool designed to easily import information from SQL databases into your Hadoop cluster	All cluster nodes

Table 2. Directories where components and their associated files are created (continued)

Component	Directory	Directory contents	Deployed location
ZooKeeper	`/zookeeper`	ZooKeeper, a centralized service that enables synchronization across a cluster	All cluster nodes

InfoSphere BigInsights installed components

The InfoSphere BigInsights installation program installs several components during the installation process. Some components are installed on all nodes in the cluster, but are configured to run on one node only.

The following components are installed with InfoSphere BigInsights. The table indicates where each of the components are deployed and configured in your cluster.

Use the **`Short host name`** column to record the name of the node where each component is deployed and configured. For example, if your distributed file system is deployed on `servername0001.localdomain`, record `servername0001` in the **`Short host name`** column. You enter this name into the installation program when indicating the location of each component.

Table 3. Installation summary for components that are installed with InfoSphere BigInsights

Component	Deployed location	Configured location	Short host name	Required components
Avro	All nodes	Does not need to be configured		None
Big SQL	Dedicated Big SQL server	Single node		Hadoop HBase ZooKeeper
DataNode	Any node	All nodes except the nodes for the NameNode, JobTracker, and web console		Distributed file system
Derby	All nodes	Single node		None
Distributed file system	Installation node	Single node		None
Flume	All nodes	All nodes except the nodes for the NameNode, JobTracker, and web console **Note:** Flume is not configured by the installation program, and must be configured manually		None

Table 3. Installation summary for components that are installed with InfoSphere BigInsights (continued)

Component	Deployed location	Configured location	Short host name	Required components
Hadoop	All nodes	Single node		Distributed file system MapReduce
InfoSphere BigInsights Console	Installation node	Single node		Catalog Distributed file system MapReduce Oozie
InfoSphere Guardium collector host	Dedicated InfoSphere Guardium proxy server	Single node		Hadoop
InfoSphere Guardium host	External node, outside of InfoSphere BigInsights	NameNode		None
Jaql	All nodes	All nodes		MapReduce
Jaql UDF server	All nodes	Single node		None
JobTracker	Any node **Note:** Install on own node, and do not configure a DataNode or TaskTracker on this node	Single node		Distributed file system
Lucene	All nodes	All nodes		None
Monitoring	All nodes	All nodes that are configured with Hadoop components, HBase master node, HBase region server, Flume nodes, Oozie node, and Zookeeper node		HBase
NameNode	Any node **Note:** Install on own node, and do not configure a DataNode or TaskTracker on this node	Single node		Distributed file system
Oozie	All nodes	Single node		Distributed file system MapReduce

Table 3. Installation summary for components that are installed with InfoSphere BigInsights (continued)

Component	Deployed location	Configured location	Short host name	Required components
Orchestrator node	InfoSphere BigInsights Console node	Single node		MapReduce
Pig	All nodes	All nodes		None
Scheduler	All nodes	Any TaskTracker node		None
Secondary NameNode	Any node	Single node		Distributed file system
Text analytics	Installation node	Single node		None
Sqoop	All nodes	All nodes		Distributed file system
ZooKeeper	All nodes	Single node, but can be configured on additional nodes		None

InfoSphere BigInsights installation worksheet

The following worksheet includes installation directories and parameters that you encounter when installing InfoSphere BigInsights. Use this checklist to plan for any conflicts that might exist in your system, and then record the values for each directory and parameter before installing InfoSphere BigInsights.

When you reach the Summary panel of the installation program, you can print the installation summary, which includes the values that you selected for each option.

Cluster node list

Record information for your management node and all data nodes in your cluster where you are installing InfoSphere BigInsights. The following table includes examples that illustrate the format of each of the fields. You might need a larger area to include all of your data nodes.

For external facing components, such as the InfoSphere BigInsights Console, HttpFS, and the Big SQL server, you must configure the host name by using an external facing host name. This host name must be accessible from outside of the node where InfoSphere BigInsights is installed. Using an external facing host name ensures that the JDBC URL connection is binding properly to the external host names for external access.

Table 4. Nodes that are used in your cluster

Rack ID	Data LAN IP address	Qualified host name	Short host name
DR01	*192.0.2.0*	*server1.localdomain*	*server1*
DR02	*192.0.2.1*	*server2.localdomain*	*server2*

Table 4. Nodes that are used in your cluster (continued)

Rack ID	Data LAN IP address	Qualified host name	Short host name

Installation directories

The following table lists installation directories and provides you with the default value for each directory. Use the last column to record your value if you specify one that is different from the default value.

Table 5. Installation directories and default values

Directory	Description	Default value	Record your value
Data directory	Directory where you store application data for InfoSphere BigInsights	`/var/ibm/biginsights`	
DataNode data directory	Local directory where each DataNode stores data for your distributed file system	`/hadoop/hdfs/data`	
Hadoop system directory	System directory where Hadoop stores its configuration data	`/hadoop/mapred/system`	
HBase root directory	Directory within your distributed file system where HBase stores data	`/hbase`	
Installation directory	Local directory where you are installing InfoSphere BigInsights	`/opt/ibm/biginsights`	
MapReduce cache directory	Directory where MapReduce stores its cache data	`/hadoop/mapred/local`	
MapReduce log directory	Directory where MapReduce logs are written	`/var/ibm/biginsights/hadoop/logs`	
MapReduce system directory	Directory where MapReduce stores shared system files	`/hadoop/mapred/system`	
Secondary NameNode data directory	Local directory where the Secondary NameNode stores data	`/hadoop/hdfs/namesecondary`	

Installation ports

Table 6 lists installation parameters and provides you with the default value for each parameter. Use the last column to record your value if you specify one that is different from the default value.

Some parameters, indicated by the following symbols, are not valid with specific components.

Symbol	Description
†	Component is not valid for use with adaptive MapReduce.
‡	Component is not valid for use with GPFS.

Table 6. Installation parameters and default values

Installation parameter	Description	Default value	Record your value
Big SQL	Communication port for the Big SQL server	7052	
DataNode	Installation port for the DataNode server	50010	
	IPC port for the DataNode server	50020	
	HTTP port for the DataNode server	50075	
	Java Management Extensions (JMX) port for the DataNode server	8007	
Derby	Installation port for the Derby client	1528	
GPFS	Installation port for GPFS	1191	
	Privileged port for GPFS	1001	
HBase	Installation port for the HBase master	60000	
	Installation port for the HBase master user interface	60010	
	JMX port for the HBase master server	10101	
	HBase region server port	60020	
	User interface port for the HBase region server	60030	
	JMX port for the HBase region server	10102	
Hive	Installation port for the Hive client	10000	

Table 6. Installation parameters and default values (continued)

Installation parameter	Description	Default value	Record your value
	Installation port for the Hive web interface	9999	
HttpFS	Communication port for the HttpFS server, which is used to access and transfer data in your distributed file system	14000	
Jaql UDF server	Installation port for the Jaql user-defined server	8200	
JobTracker†	Installation port for the JobTracker daemon	9001	
	HTTP port for the JobTracker	50030	
	JMX port for the JobTracker	8006	
Monitoring	Control port for monitoring	9093	
	REST port for monitoring	9099	
NameNode‡	Installation port for the NameNode server	9000	
	HTTP port for the NameNode	50070	
	JMX port for the NameNode	8004	
Orchestrator node	Installation port for the orchestrator node	8888	
Oozie	Installation port for the Oozie client	8280	
Secondary NameNode	Installation port for the Secondary NameNode server	50090	
TaskTracker†	Installation port for the TaskTracker agent	50060	
Web console	Installation port for the web console if you choose HTTP	8080	
	Installation port for the web console if you choose HTTPS	8443	
ZooKeeper	Installation port for the ZooKeeper client	2181	
	JMX port for the ZooKeeper client	2182	

Planning for high availability

When planning for using the high availability solution with InfoSphere BigInsights, you must consider additional requirements.

The following requirements are necessary for implementing InfoSphere BigInsights with high availability:

- An external, shared, highly available Network File System (NFS)
- Additional unused IP addresses and host names with forward and reverse resolution
- External hardware fencing mechanisms (optional)

The NFS is used to maintain shared state information such as high availability system configuration or the NameNode transaction log between high availability nodes. The NFS must be highly available because a failure in the shared file system results in errors in the InfoSphere BigInsights cluster.

The reserved IP addresses and host names are used as the main access point for the NameNode process. The high availability system automatically creates a network interface alias in the selected NameNode host, and uses gratuitous Address Resolution Protocol (ARP) requests to inform all nodes in the cluster when a migration is in process.

Fencing is supported to ensure consistency after failures occur, and when configured, fencing is called by the high availability solution to isolate a failing node from the remainder of the cluster.

High availability configurations

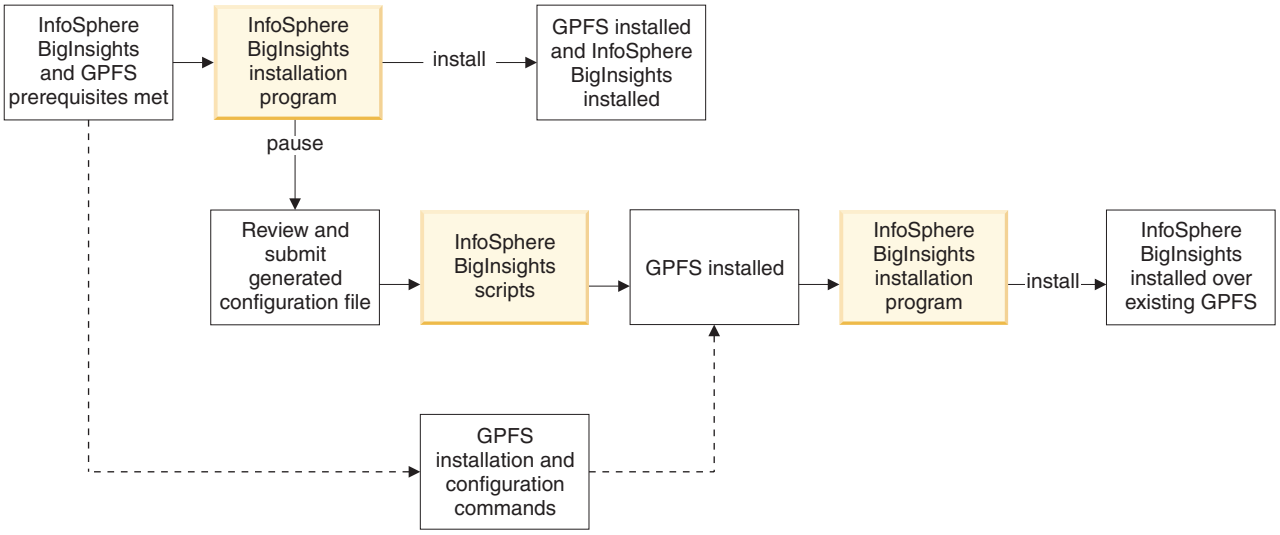
InfoSphere BigInsights supports two high availability configurations that differ by the number of dedicated high availability nodes.

The first configuration requires three dedicated hosts for the high availability management component, NameNode, and JobTracker. The NameNode and JobTracker components run in different hosts, while a third node runs in standby mode. After the NameNode or JobTracker components are migrated, the migration process is started on the node in standby mode. A second consecutive migration of the same component results in all processes being collocated in the remaining host. This configuration is best suited for deployments that require maximum stability.

The second configuration requires only two high availability nodes. The NameNode and JobTracker components run in different hosts like in the first configuration. In the event of a failure, the migrated process is collocated with the other service to ensure that the cluster supports one component failure, but might require administrative intervention to restore normal behavior in high load clusters. This configuration is best suited for deployments that require faster deployment time, but do not require as much stability.

GPFS installation paths

You can choose one of several paths for installing InfoSphere BigInsights and GPFS. Except for the file system component, the component directories and summary for an InfoSphere BigInsights over GPFS installation are the same as for over an HDFS installation.



Note: GPFS is not included as part of the InfoSphere BigInsights Quick Start Edition.

You can install GPFS in one of the following ways:

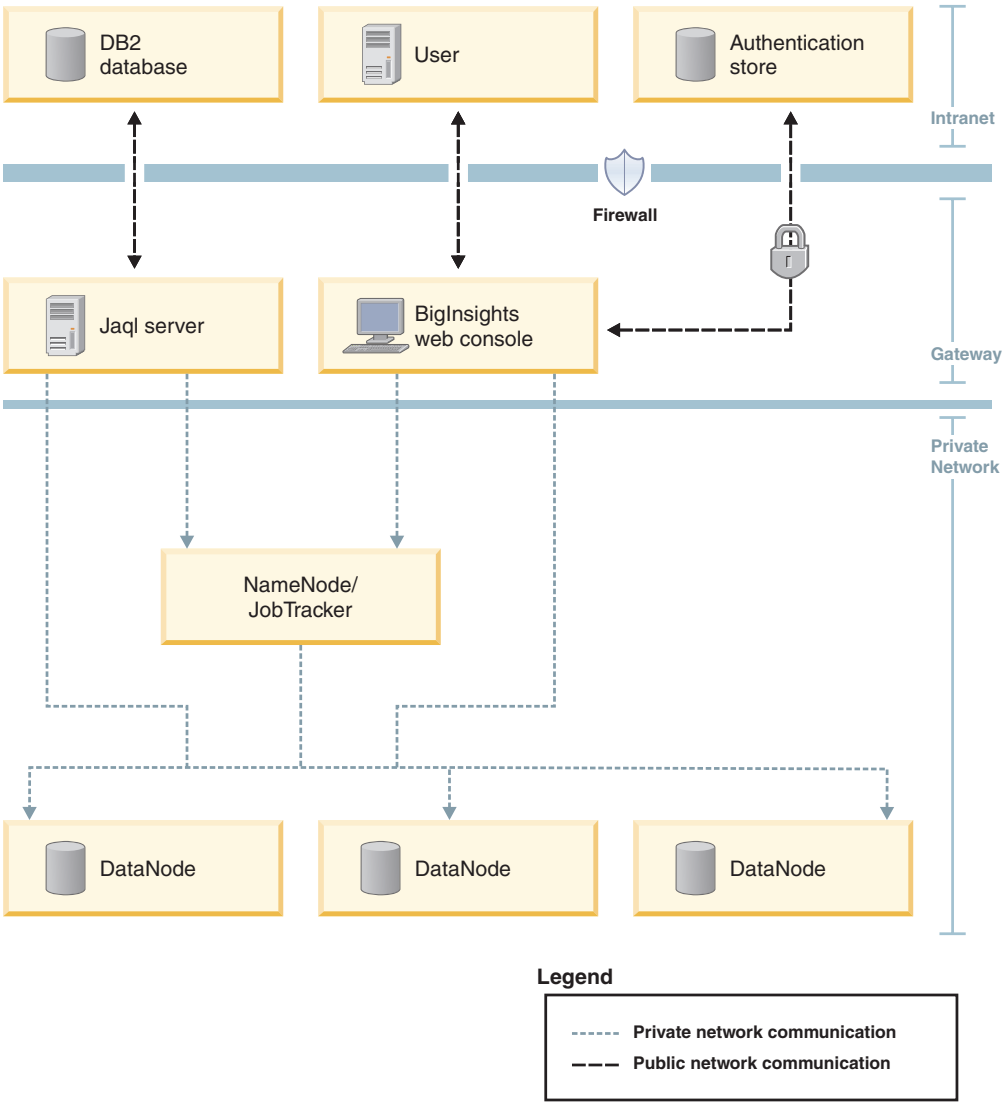
1. Install GPFS and InfoSphere BigInsights simultaneously by using the installation program.
2. Begin installing InfoSphere BigInsights by using the installation program, install GPFS by using scripts, and then complete the installation by using the installation program.
3. Install GPFS by using commands, and then use the installation program to install InfoSphere BigInsights.

Alternatively, you can complete a silent installation instead of using the InfoSphere BigInsights installation program.

Security architecture

InfoSphere BigInsights security architecture uses both private and public networks to ensure that sensitive data, such as authentication information and source data, is secured behind a firewall. Users roles are assigned distinct privileges so that critical Hadoop components cannot be altered by users without access.

Connect all machines in your InfoSphere BigInsights production cluster over a private network. Because cluster machines have unrestricted communication with one another, configure the InfoSphere BigInsights console to serve as a gateway into the InfoSphere BigInsights cluster. The console port (8080 by default) on the machine that hosts the console needs to be the only open HTTP port. Ensure that other HTTP content is served by using a reverse proxy.



Choosing user security and authentication

During installation, you must configure an administrative user. You can choose different authorization options for defining the administrative user, depending on the type of user security you want to use.

Authorization schemas and role authorization are closely related. A minimal role authorization schema is created for you by the installation program that is based on the authentication schema you select during installation. The following are the authentication schemas that are supported by InfoSphere BigInsights.

If you install InfoSphere BigInsights with one authentication schema and then decide to use a different authentication method, you can run the InfoSphere BigInsights installation program, select the **Upgrade** option, and then choose a different authentication schema.

No user authentication

If you plan to use no user authentication, no user or group restrictions are configured. Users can access the InfoSphere BigInsights web console by using their web browser, and can complete all available tasks.

Flat file authentication

If you plan to use flat file authentication, the installation program configures InfoSphere BigInsights so that the administrative user ID that you create can complete all tasks in the web console across all nodes in your cluster.

Lightweight Directory Access Protocol (LDAP)

If you plan to use LDAP authentication, add the administrative user to all groups defined in LDAP that will be mapped to the following roles during installation:

- BigInsightsSystemAdmin
- BigInsightsDataAdmin
- BigInsightsAppAdmin
- BigInsightsUser

Pluggable Authentication Modules (PAM)

If you plan to use PAM authentication with a Linux shadow password file, complete the following steps:

1. Modify the `/etc/shadow` file on all nodes in the cluster to include the InfoSphere BigInsights administrative user and the encrypted password.
2. Modify the `/etc/groups` file on all nodes in the cluster to map the InfoSphere BigInsights administrative user to all operating system groups that will be mapped to the following roles during installation:
   - BigInsightsSystemAdmin
   - BigInsightsDataAdmin
   - BigInsightsAppAdmin
   - BigInsightsUser

   Note that uid values for all users, and gid values for all groups used by InfoSphere BigInsights, must be consistent across all nodes in the cluster. For example, if you have a group named usergroup with gid 446 on the install node, then the gid value for usergroup on all the data nodes must also be 446. To ensure that the cluster is configured in a consistent manner for each user, you may issue the id command (for example, id user1) on all nodes in the cluster to compare results. If the output matches exactly, then that user is setup correctly.

If you plan to use PAM authentication by communicating with an LDAP server, add the InfoSphere BigInsights administrative user to all groups defined in LDAP that will be mapped to the following roles during install:

- BigInsightsSystemAdmin
- BigInsightsDataAdmin
- BigInsightsAppAdmin
- BigInsightsUser

Mapping users and groups to roles

InfoSphere BigInsights supports four predefined roles. During installation, you can map users and groups in your enterprise to the four InfoSphere BigInsights roles.

The mapping of enterprise users and groups to InfoSphere BigInsights roles is stored in the `install.xml` in the `$BIGINSIGHTS_HOME/conf` directory.

Roles

A *role* defines a set of user privileges and determines the actions that a user can complete, including the data that a user can see.

A user can be associated with a role directly or indirectly by being a member of a group that is assigned to a role. The InfoSphere BigInsights console uses the following default roles:

BigInsightsSystemAdministrator
: Completes all system administration tasks, such as monitoring cluster health and adding, removing, starting, and stopping nodes.

BigInsightsDataAdministrator
: Completes all data administration tasks, such as creating directories, running Hadoop file system commands, and uploading, deleting, downloading, and viewing files.

BigInsightsApplicationAdministrator
: Completes all application administration tasks, such as publishing and unpublishing (deleting) applications, deploying and undeploying applications to a cluster, configuring icons, applying application descriptions, changing runtime libraries and categories of applications, and assigning application permissions to a group.

BigInsightsUser
: Runs applications that the user has permission to run and views the results, data, and cluster health.

  This role is typically the most common role granted to cluster users who complete non-administrative tasks.

When a user who is authorized to run an application is logged into a secure InfoSphere BigInsights console, the application workflows are submitted under the user ID and primary group ID of that user. The InfoSphere BigInsights console verifies the roles of the current user to verify that the user is authorized to complete certain actions.

If you are using the HDFS file system, and the user is authorized to browse that file system, all existing files and directories are listed. However, file-level access control lists (ACLs) determine whether the user can read or write file contents. A data administrator can use the **`hadoop fs -chmod ...`** command to change HDFS ACLs.

Groups

A *group* associates a set of users who have similar business responsibilities. Groups separate users at the business level rather than the technical level.

You might define one or more groups that associate users into logical sets based on business needs and constraints. If a group of users must be given access to the InfoSphere BigInsights console, then the group must be associated with at least one InfoSphere BigInsights role.

Users

A *user* is an entity that can be authenticated and typically represents an individual user. Each user authenticates to the system by using credentials, including a user

name and password, and might belong to one of more groups. A user inherits the roles associated with all groups of which that user is a member.

If you want to grant a user access to the InfoSphere BigInsights console, the user must be associated with or must be a member of a group that is associated with at least one InfoSphere BigInsights role.

Chapter 3. Preparing to install InfoSphere BigInsights

Before you install InfoSphere BigInsights, ensure that you have the correct installation media and that your computers are ready for installation.

Choosing a user to install the product with

You can install InfoSphere BigInsights as the root user or as a non-root user. The user that you run the installation program with determines the privileges and level of security that you install the product with.

You can run the installation program with one of the following privilege levels:

Root privileges
The root user is used directly by logging in to each node in the cluster. During installation, you input the password for the root user. If additional nodes require a different password, you can set passwords for each node.

Non-root with `sudo` privileges
You can select from two different options for running the installation program as a non-root user with **`sudo`** privileges:

- The current user that you are running the installation program with uses **`sudo`** privileges to gain privileged access, and then uses passwordless SSH from the root user to gain privileged access to other nodes in the cluster. If you select this option, you must input the password for the current user that you are running the installation program as.
- The current user that you are running the installation program with uses **`sudo`** privileges to gain access to all nodes in the cluster. If you select this option, you can optionally provide the password for the current user. If you specify the password, it must be identical across all nodes.

Configuring your browser

To run the installation program successfully, configure your browser.

All browsers

- Verify that you have a supported browser. For details, see the system requirements.
- Make sure that JavaScript is enabled in your browser.

Microsoft Internet Explorer

Make sure that the security zone for the computer allows the installation program to run:

1. From Microsoft Internet Explorer, in the **Tools** menu, select **Internet Options**.
2. Click the **Security** tab.
3. Click the **Trusted Sites** icon.
4. Click **Sites**.
5. In the **Add this Website to the zone** field, type `http://`*`hostname`* where *hostname* is the host name of the computer where you plan to run the installation program.

6. Click **Add**.
7. Click **OK**.
8. Click **OK**.

Mozilla Firefox

Make sure that images load automatically and that JavaScript is enabled:

1. In the **Tools** menu, click **Options**. The Options window displays.
2. Click **Content**.
3. Enable **Load images automatically**.
4. Enable **Enable JavaScript**.
5. Click **OK**.

Obtaining InfoSphere BigInsights software

Obtain the InfoSphere BigInsights software and any applicable fix packs before you begin the installation process.

Before you begin

To ensure uninterrupted connectivity to the installation media, copy the contents of the installation media to a local file system or optical drive. Run the installation program from this location.

About this task

The node from where you start the installation or upgrade must be part of your cluster.

Procedure

1. Acquire the installation media.

Option	Description
If you have the installation media in physical form	Check that you have all of the installation disks.
If you do not have the installation media in physical form	Download the installation image files from Passport Advantage. Use the Knowledge Collection support document to determine the image files that are appropriate for your operating systems and configurations.

2. Download any applicable fix packs from Fix Central.
3. Extract the InfoSphere BigInsights installation files onto a cluster node with a supported operating system.

Preparing to run the installation program

In addition to product prerequisites, there are tasks common to all InfoSphere BigInsights installation and upgrade paths. You must complete these common tasks before you start an installation or upgrade.

Procedure

1. Return a list of available disks in your cluster. You use the disk partition names when specifying the cache and data directories for your distributed file system.

   ```
   df -h
   ```

2. Ensure that adequate disk space exists for the following required directories.

Directory	Available disk space
/	10 GB
/tmp	40 GB
/$BIGINSIGHTS_HOME The default directory for this variable is /opt/ibm.	15 GB
/$BIGINSIGHTS_VAR The default directory for this variable is /var/ibm.	5 GB
/home/$USER_HOME The default directory for this variable is /home/biadmin.	5 GB

3. Create directories for the data files and cache files for your distributed file system. These directories must be owned by biadmin:biadmin.

   ```
   mkdir /disk_name/directory
   ```

   For example, the following directories exist on a server with mount points disk1 through disk10.

   ```
   /disk1/hadoop/hdfs && /disk1/hadoop/mapred.../disk10/hadoop/hdfs &&
   /disk10/hadoop/mapred
   ```

4. Create the biadmin user and group.
   a. On every node in your cluster, as the root user, create the biadmin group and then add the biadmin user to it.
      1) Add the biadmin group.

         ```
         groupadd -g 168 biadmin
         ```

      2) Add the biadmin user to the biadmin group.

         ```
         useradd -g biadmin -u 168 biadmin
         ```

      3) Set the password for the biadmin user.

         ```
         passwd biadmin
         ```

   b. On the master node, add the biadmin user to the sudoers group.
      1) Edit the sudoers file.

         ```
         sudo visudo -f /etc/sudoers
         ```

      2) Comment out the line that reads Defaults requiretty.
      3) Locate the following line...

         ```
         # %wheel  ALL=(ALL)  NOPASSWD: ALL
         ```

         ...and replace that line with the following line.

         ```
         %biadmin ALL=(ALL)  NOPASSWD: ALL
         ```

5. Configure your network.
   a. In the /etc directory, edit the hosts file to include the IP address, fully qualified domain name, and short name of each host in your cluster, separated by spaces. You must edit this file on each server in your cluster.

The format is `IP_address domain_name short_name`. For example,

```
127.0.0.1 localhost localhost.localdomain
123.123.123.123 server_name server_name.server_domain.com
```

b. Configure passwordless SSH between every node and the master node, between the master node and itself, and for both the biadmin user and root user.

1) On every node in your cluster, run the following command as both the biadmin user and root user. Select the default file storage location and leave the password blank.

```
ssh-keygen -t rsa
```

2) On the master node, run the following command as both the biadmin user and the root user to each node, and then from each node back to the master.

```
ssh-copy-id -i ~/.ssh/id_rsa.pub user@server_name
```

c. Run the following commands in succession to disable the firewall on all nodes in your cluster.

Important: Ensure that you reenable the firewall on all nodes in your cluster after installing InfoSphere BigInsights.

```
service iptables save

service iptables stop

chkconfig iptables off
```

d. Disable IPv6 on all servers in your cluster.

1) From the `/etc` directory, open the `modprobe.conf` file and add the following line to it.

```
ipv6 /bin/true
```

2) From the `/etc/sysconfig/network` directory, add the following parameters.

```
NETWORKING_IPV6=no
IPV6INIT=no
```

6. Ensure that the `ulimit` properties for your operating system are configured.

a. From the `/etc/security` directory, open the `limits.conf` file.

b. Ensure that the `nofile` and `nproc` property contain the following values or greater. The `nofile` parameter sets the maximum number of files that can be open, and the `nproc` property sets the maximum number of processes that can run. The following values are the minimum values that are required.

```
nofile - 16384
nproc  - 10240
```

7. Synchronize the clocks of all servers in the cluster by using an internal or external Network Time Protocol (NTP) source. The InfoSphere BigInsights installation program synchronizes the other server clocks with the master server during installation. You must enable the NTP service on the management node and allow the clients to synchronize with the master node.

a. From the `/etc` directory, open the `ntp.conf` script.

```
vi /etc/ntp.conf
```

b. In the `ntp.conf` script, search for the line that begins with `# Please consider joining the pool (http://www.pool.ntp.org/join.html.` After this line, insert one or more of the following time servers.

```
server 0.rhel.pool.ntp.org
server 1.rhel.pool.ntp.org
server 2 rhel.pool.ntp.org
```

c. Update the NTP service with the time servers that you specified.

```
chkconfig --add ntpd
```

d. Start the NTP service.

```
service ntpd start
```

e. Verify the time service by displaying the offsets for each of the connected servers.

```
ntpq -p
```

8. Ensure that the shell interpreter for the administrator user ID is `bash`.
   a. Navigate to the `/etc` directory.
   b. Run the following command to show the default shell interpreter for the administrator user ID.

```
$grep "^biadmin:" passwd
biadmin:x:500:500::/home/biadmin:/bin/bash
```

   c. If the value for the administrator user ID (by default, `biadmin`) is not `/bin/bash`, open the `passwd` file in the `/etc` directory and change the value.
9. Check port availability and resolve host names.
   a. Ensure that all required ports are available. The following command lists the state of all ports on your system, their current state, and the process ID that is running on each port.

      Port number 8300 must be available for the installation program to run. For a list of required components and their ports, see installation worksheet.

```
netstat -ap | more
```

   b. Ensure that the host names for all cluster nodes are resolved. The host names must be configured to the same IP addresses as the actual servers, because InfoSphere BigInsights does not support dynamic IP addresses. You can resolve host names by using DNS servers, or by ensuring that the host names are mapped correctly in the `/etc/hosts` file across all nodes in the cluster.
   c. In the `/etc/hosts` file, ensure that `localhost` is mapped to the loopback address `127.0.0.1`, as shown in the following example.

```
# Do not remove the following line, or various programs
# that require network functionality will fail.
127.0.0.1    localhost.localdomain localhost
::1          localhost6.localdomain6 localhost6
192.0.2.*    server_name.com server_name
```

10. Update the firmware for your disk controllers. If disks are not used for an extended period of time, they might enter sleep mode. This behavior might be perceived as a delay when the InfoSphere BigInsights installation program or related processes try to access the disks. Disks might require a longer response time because the disks start only when accessed.
11. On the node where you plan to run the installation program, verify or install the Linux Expect package.
    a. Verify if the Linux Expect package is installed.

```
rpm -qa | grep expect
```

    b. If the package is not installed, then run the following command to install it.

```
yum install expect
```

Configuring LDAP authentication

If you want to use LDAP or PAM authentication for your users and groups, you must configure your LDAP server before running the InfoSphere BigInsights installation program. You must complete this procedure on every node in your cluster.

Before you begin

You must have an LDAP server configured and working. You need the following information to complete this procedure. You can find this information in the `nslcd.conf` file in the `/etc` directory.

- LDAP server URI, such as `ldap://10.0.0.1`.
- LDAP server search base, such as `dc=example,dc=com`.

Procedure

1. Install the Name Service Switch (NSS) package. Applications use the NSS to authenticate by using LDAP.
   a. Check to see whether the NSS package is installed.
      ```
      rpm -qa | grep nssldapd
      ```
      If the NSS package is installed, then output is returned that shows the package name and version.
   b. If the NSS package is not installed, then run the following command to install it.

Table 7. Commands to install the NSS package based on operating system

Operating system	Command
Red Hat Linux	`yum install nss-ldapd`
SUSE Linux	`yum install nss_ldapd`

2. Modify the NSS configuration file to add the LDAP option to related services.
   a. Navigate to the `/etc` directory and open the `nssswitch.conf` file.
   b. Add `ldap` to each of the services listed. With this configuration, the system searches for services in system files. If no value is returned, then the LDAP server is queried to obtain a value for each service.
      ```
      passwd:  files ldap
      shadow:  files ldap
      group:   files ldap
      ```
   c. Save and close the `nssswitch.conf` file.
3. Modify the LDAP client configuration file to include the name of your LDAP server and the name of the search base.
   a. Navigate to the `/etc` directory and open the `libnss-ldap.conf` file.
   b. Add the host of your LDAP server and the distinguished name of the search base.
      ```
      # Your LDAP server.
      uri ldap://10.0.0.1

      # The distinguished name of the search base.
      base dc=domain_name
      ```

domain_name is the distinguished name that is used to bind to the directory server for lookups. For example, if you specify `base dc=example,dc=com`, then LDAP searches only for users that exist in the `example,dc=com` search base.

c. Save and close the `libnss-ldap.conf` file.

What to do next

1. Install InfoSphere BigInsights.
2. Add users and groups by using PAM.

Selecting prerequisite checker options

The prerequisite checker is an embedded utility in the InfoSphere BigInsights installation program that checks your cluster environment to determine that software requirements are met before starting the installation process. This utility checks the operating system level, key libraries, and essential configurations to ensure that your cluster environment is ready to install the product.

About this task

By default, the prerequisite checker utility runs checks on all prerequisite software. You can select the options that you want the utility to check when it runs, or disable the utility completely. If the prerequisite checker detects errors during installation, then you must resolve the errors before resuming the installation process.

Procedure

1. Navigate to the *extract_dir*`/installer/bin` directory, where *extract_dir* is the directory where you extracted the InfoSphere BigInsights installation package.
2. Run the **enableOrDisablePrechecker.sh** script to modify how the prerequisite checker utility runs. For information about the usage of this command, specify the `help` argument:

   ```
   enableOrDisablePrechecker.sh --help
   ```

Option	Description
To run the utility with options	Run the **enableOrDisablePrechecker.sh** script with the options that you want to include: `enableOrDisablePrechecker.sh disable item1,item2,item3` *item* is the option that you want the utility to run.
To disable the utility	Run the **enableOrDisablePrechecker.sh** script with the `all` argument: `enableOrDisablePrechecker.sh disable all`

What to do next

Install the product by using the installation wizard or by using a response file.

Preparing to install GPFS

If you are installing IBM General Parallel File System (GPFS) as your file system, ensure that you meet all disk prerequisites and install the required RPM packages.

Before you begin

On SUSE Linux, the GPL configure script uses the `/lib/modules/$kernel_string/build` path to determine the directories to build. The `/lib/modules/$kernel_string/build` directory must correspond to where the `include/linux/version.h` file is located.

Note: GPFS is not included as part of the InfoSphere BigInsights Quick Start Edition.

Procedure

1. When designing your disk infrastructure, ensure that your disks are unmounted, and that they meet the following criteria.
   - Enough disk space exists on the computers where you are installing GPFS and InfoSphere BigInsights. See the detailed system requirements on ibm.com for hardware requirements.
   - Raw disks or raw partitions that you use are not already in use by (or formatted for) a file system. You can use tools like **parted** or **fdisk** to create or remove a partition table (disklabel). The tool that you use depends on the type of partition tables that are involved. You can use a tool like **partprobe** to request the operating system to reread the changed partition table. Rebooting a system helps to detect these changes after you make them.
   - Disks do not contain empty partition tables. You can erase an empty partition table by using the Linux **dd** command, but doing so is considered a system administration task, not a user task.
   - A separate storage pool is used for your data that is independent from the system storage pool. For optimal performance, use separate disks for metadata and data where possible. In addition, using an entire disk for GPFS is preferable to using a partitioned disk, regardless of whether all partitions of a disk are used for GPFS.
2. Install the RPM packages that are required to build the open source portability layer.
   a. Run the **uname –r** command to determine the kernel version of your system.
   b. Use the **yast** command or **zypper** command on SUSE Linux, or the **yum** command on RedHat Linux to install the following RPM packages on all cluster nodes.

      Note: The package version must match the kernel version reported through the **uname –r** command. If the versions do not match, the InfoSphere BigInsights installation program might fail.
      - kernel-headers
      - kernel-devel (RedHat Linux) or kernel-source (SUSE Linux)
      - imake (RedHat Linux) or automake (SUSE Linux)
      - gcc-c++
      - libstdc++
      - compat-libstdc++-33
      - redhat-lsb (RedHat Linux) or lsb (SUSE LInux)

c. Optional: If you do not obtain the versions that you need by using one of the package installation commands, increment the kernel version. For example, for kernel-source, if **uname –r** returns `2.6.32.12-0.7-default`, try installing `kernel-source-2.6.32.12-0.7.1`. Otherwise, consider updating your kernel to a newer version.

   If an error message displays that reads, `Cannot find a valid kernel include dir`, check that the directory `/lib/modules/$kernel_string/build/include` exists, where `$kernel_string` is the value returned by the **uname –r** command.

   If the updated kernel-level packages are different from the running kernel, the build symbolic link might be incorrect. Use the following command to update the symbolic link:

   ```
   ln -sf ../../../usr/src/kernels/2.6.18-238.12.1.el5-x86_64 build
   ```

3. Optional: Enable monitoring for GPFS. If you decide not to enable monitoring, you can enable monitoring for GPFS after you complete the installation process.
   a. Install the Net-SNMP package by using the `yum` utility.

      ```
      yum install net-snmp
      ```

   b. Create symbolic links for each of the following libraries if they do not already exist. The GPFS subagent expects to find these libraries without version information appended to the name, so you must create symbolic links as shown in the following example.

      ```
      cd /usr/lib64
      ln -s libnetsnmpagent.so.5.1.2 libnetsnmpagent.so
      ```

      Repeat this process for each of the following libraries. The GPFS subagent expects these libraries to exist in the following directories from the specified components.

      Important: The `libcrypto.so` library depends on the `openssl_devel.x86_64` package for GPFS monitoring to function properly. Install the `openssl_devel.x86_64` package by using the following command before configuring symbolic links.

      ```
      yum install openssl_devel.x86_64
      ```

Library	Where GPFS expects the library	Directory
`libnetsnmpagent.so`	Net-SNMP	`/usr/lib64`
`libnetsnmphelpers.so`	Net-SNMP	`/usr/lib64`
`libnetsnmpmibs.so`	Net-SNMP	`/usr/lib64`
`libnetsnmp.so`	Net-SNMP	`/usr/lib64`
`libwrap.so`	TCP wrappers	`/lib64`
`libcrypto.so`	OpenSSL	`/usr/lib64`
`libperl.so`	Net-SNMP	`/usr/lib64`

   c. Start the Net-SNMP Agent Daemon service (`snmpd`) to ensure that the service is running.

Operating system	Command
Red Hat Enterprise Linux	`/sbin/service snmpd start`
SUSE Linux Enterprise Server	`/etc/rc.d/snmpd start`

Discovering devices with uncommon names

If you have a configuration where the storage multipath driver does not use a common device name, use the `gpfs-nsddevices` script so that GPFS can detect your devices.

About this task

By default, GPFS supports common device names. For example, GPFS automatically detects devices that begin with `/dev/sd` on Linux systems. To detect devices with uncommon names, edit the `gpfs-nsddevices` script.

If you install GPFS by itself, you can use the `nsddevices` user exit to detect uncommon device names. Because the file that the `nsddevices` user exist requires is available after you install GPFS, the InfoSphere BigInsights installation program packages the required file so that you can customize it before installing GPFS.

Procedure

1. Navigate to *$extract_directory*`/install/hdm/bin` and open the `gpfs-nsddevices` script.

   $extract_directory is the directory where you extracted the `tar.gz` file for the installation program.
2. Modify the script to include the device names that you want to discover. For example, the following file detects device names that begin with `dm-`.

   ```
   #/bin/ksh
   #
   # this script ensures that we are not using the raw /dev/sd\* devices for GPFS
   # but use the multipath /dev/dm-\* devices instead
   for dev in $( cat /proc/partitions | grep dm- | awk '{print $4}' )
   do
       echo $dev generic
   done

   # skip the GPFS device discovery
   return 0
   ```

   Specifying `return 0` indicates that you want only the device names that match your discovery criteria that is returned.

   Specifying `return 1` indicates that you want the device names that match the GPFS default discovery criteria that is returned. If you specify `return 1`, the device names that match your discovery criteria are returned and the device names that match the GPFS default discovery criteria are returned.
3. Run the `gpfs-nsddevices` script to discover your devices.
4. Review the discovered device list to ensure that the discovered disks are the disks that you want GPFS to use. Not doing so can lead to irreversible changes that can result in data loss.

Enabling adaptive MapReduce

If you want to install InfoSphere BigInsights with adaptive MapReduce instead of Apache MapReduce, then you must enable this option in a configuration file before you run the InfoSphere BigInsights installation program.

About this task

Your cluster must contain at least three nodes to install InfoSphere BigInsights with high availability. If you enable the option to install adaptive MapReduce with high

availability, the option to configure high availability is enabled when you run the InfoSphere BigInsights installation program.

Procedure

1. Navigate to the directory where you extracted the `biginsights-enterprise-linux64_release_number.tar.gz` file, where *release_number* is the release number that you are installing.
2. Open the `install.properties` file and modify the appropriate property depending on whether you want to install adaptive MapReduce with high availability. If both of the following properties are set to `false`, then Apache MapReduce is installed.

Option	Description
Install adaptive MapReduce instead of Apache MapReduce	`AdaptiveMR.Enable=true`
Install adaptive MapReduce with high availability instead of Apache MapReduce	`AdaptiveMR.HA.Enable=true`

   The following table shows the combinations of these parameters and which capabilities are installed as a result.

Table 8. Combination of settings in the `install.properties` file for adaptive MapReduce and high availability

Settings	Result
`AdaptiveMR.HA.Enable=true` `AdaptiveMR.Enable=false`	Adaptive MapReduce is installed and high availability is enabled
`AdaptiveMR.HA.Enable=true` `AdaptiveMR.Enable=true`	Adaptive MapReduce is installed and high availability is enabled
`AdaptiveMR.HA.Enable=false` `AdaptiveMR.Enable=true`	Adaptive MapReduce is installed
`AdaptiveMR.HA.Enable=false` `AdaptiveMR.Enable=false`	Apache MapReduce is installed

3. Save and close the `install.properties` file, and then run the InfoSphere BigInsights installation program.

Creating a private SSL certificate for a secure InfoSphere BigInsights Console

If you want to configure your InfoSphere BigInsights Console and HttpFS for HTTPS, you can create a private certificate authority, use it to sign an SSL certificate, and import the public SSL certificate to your browser.

Before you begin

OpenSSL for Linux must be installed before you create your SSL certificate.

About this task

When you install InfoSphere BigInsights, you can upload your SSL certificate directly in the installation program and enter a password. During the installation process, the SSL certificate is used to configure the InfoSphere BigInsights Console and HttpFS to run with the HTTPS protocol.

If you do not want to create your own private SSL certificate, then you must purchase one from a vendor. Ensure that the certificate contains the wildcard value for your SSO domain name. For example, if the host name for your server is `myserver.ibm.com`, then enter `*.myserver.ibm.com`.

Note: If you want to install InfoSphere BigInsights without HTTPS, and decide later that you want to configure your InfoSphere BigInsights Console for HTTPS, then you can upgrade InfoSphere BigInsights and select the option to use HTTPS.

Procedure

1. Create the `/certs`, `/private`, and `/crl` directories.
2. Create an empty text file and save it as `index.txt`.
3. Run the following command to create a new file named `serial`, with the initial contents of `01`.

   ```
   echo "01" > serial
   ```

4. Configure the properties for your certificate authority.
   a. Copy the sample code from "Certificate authority sample" on page 43 into a new text file and save it as `ca.txt`.
   b. Open the `ca.txt` file in a text editor and locate the sections that begin with `# MODIFY`. Replace each section with the appropriate information. The comments in the file describe the changes that you must make.
   c. Save the `ca.txt` file as `ca.cnf`. OpenSSL requires this file extension.
5. From the `root` directory, run the following command to create a private certificate authority.

   ```
   OPENSSL=ca.cnf openssl req -x509 -nodes -days 3650 \
               -newkey rsa:2048 -out certs/ca.pem \
               -outform PEM -keyout private/ca.key
   ```

6. Create a certificate request. The command prompts you for information that will be saved in the certificate request.

   For the common name, which is the host name for your server, enter the wildcard value for your SSO domain name. For example, if the host name for your server is `myserver.ibm.com`, then enter `*.myserver.ibm.com`.

   ```
   openssl req -newkey rsa:1024 -nodes -sha1 \
             -keyout cert.key -keyform PEM -out cert.req -outform PEM
   ```

7. Sign the certificate request that you generated previously.

   After running this command, the `root` directory contains the `cert.key` file and the `cert.pem` file.

   ```
   OPENSSL_CONF=ca.cnf openssl ca -batch -notext -in cert.req -out cert.pem
   ```

8. Package the `cert.key` file and the `cert.pem` file into a PSCK12 file. When prompted, enter a password that is at least six characters in length.

   ```
   openssl pkcs12 -export -in cert.pem -inkey cert.key -out server.p12
                -name biginsights  -CAfile certs/ca.pem -caname root
   ```

9. Download the `server.p12` file to a local directory on your computer.
10. From the `root/certs` directory, download the `ca.pem` to a local directory on your computer. You use this file to configure your browser to automatically accept the SSL certificate that you created.

What to do next

1. "Installing InfoSphere BigInsights by using the wizard" on page 54. When you reach the Components 1 panel, enter the SSO domain name for your InfoSphere BigInsights Console, upload your SSL certificate (`server.p12`), and enter the password for your certificate.

2. Import your SSL certificate into your browser.

Certificate authority sample

The following sample includes code that you use to configure the properties for your certificate authority. After you configure your certificate authority, you can use it to sign an SSL certificate, and then import the certificate into your browser.

```
[ ca ]
# MODIFY: Assuming [root_dir] is where the private CA is configured and create.
# Change this line to be default_ca = name of the [root_dir]. For example:
# If the full path of [root_dir] is /home/ssl/testCA2, then default_ca = testCA2
default_ca = testCA2

# MODIFY: change this line to be [ name of the [root_dir] ]. For example:[ testCA2 ]
[ testCA2 ]

# IMPORTANT: If you change the root directory, change the default_keyfile
# in the [req] section below.

# Location of all system files.
# MODIFY: replace [root_dir] with the absolute path of the [root_dir]
dir = [root_dir]/testCA2

# Location of the issued certificates.
certs = $dir/certs

# Location of the issued certificate revocation list.
# You use this list to blacklist SSL certificates.
crl_dir = $dir/crl

# Database index file.
database = $dir/index.txt

# Default location of new certificates.
new_certs_dir = $dir/certs

# Location of the certified authority certificate.
certificate = $dir/certs/ca.pem

# The current serial number.
serial = $dir/serial

# The current certificate revocation list.
crl = $dir/crl/crl.pem

# IMPORTANT: If you change the current certificate revocation list,
# change the default_keyfile in the [req] section below.

# The private key
private_key = $dir/private/ca.key

# Location of the private random number file.
RANDFILE = $dir/private/.rand

# The extentions to add to the certificate.
x509_extensions = usr_cert

# Determines how long to certify the certificate.
default_days = 365

# Specifies how long before the next certificate revocation list.
default_crl_days= 30

# Indicates what function to use to generate the fingerprint of an SSL certificate.
# The default is SHA-1, as shown below.
default_md = sha1
```

```
# Indicates whether to preserve the order of the Distinguished Name
# fields to match the order passed in.
preserve = no

# Directives that are used when requesting or signing certificates.
policy = mypolicy
x509_extensions = certificate_extensions

[ mypolicy ]
# Use the supplied information for the policy.
commonName = supplied
stateOrProvinceName = supplied
countryName = supplied
emailAddress = supplied
organizationName = supplied
organizationalUnitName = optional

[ certificate_extensions ]
# If set to false, the signed certificate cannot be used as
# the certificate authority.
basicConstraints = CA:false

[ req ]
# Same as private_key.
# MODIFY: change [root_dir] to the absolute path of the [root_dir].
default_keyfile = [root_dir]/private/ca.key

# Indicates which hash to use.
default_md = sha1

# Indicates no prompts will be used.
prompt = no

# The following properties are for the certificate authority.
subjectKeyIdentifier=hash
authorityKeyIdentifier=keyid:always,issuer
string_mask = utf8only
basicConstraints = CA:true
distinguished_name = root_ca_distinguished_name
x509_extensions = root_ca_extensions

[ root_ca_distinguished_name ]
# MODIFY: Edit the following fields for stateOrProvinceName, countryName,
# emailAddress and organizationName.
commonName = IBM InfoSphere BigInsights
stateOrProvinceName = ON
countryName = CA
emailAddress = none@ibm.com
organizationName = IBM

[ root_ca_extensions ]
basicConstraints = CA:true
```

Chapter 4. Installing InfoSphere BigInsights software

The InfoSphere BigInsights installation program supports installations, overlays, and upgrades by using the installation console or by completing a silent installation.

Installing GPFS by using InfoSphere BigInsights scripts

You can use the InfoSphere BigInsights installation program to help you install GPFS, and then to install InfoSphere BigInsights over GPFS.

About this task

Note: GPFS is not included as part of the InfoSphere BigInsights Quick Start Edition.

Procedure

1. Start the InfoSphere BigInsights installation program.
2. On the Installation Type panel, select **Cluster installation**.
3. On the File System panel, select **Install General Parallel File System GPFS**.
   a. Enter the GPFS mount point.
   b. Optionally, select a privileged port for GPFS.

   Important: HBase requires that your distributed file system supports the **sync** call. This call pushes data through the write pipeline and blocks it until the data receives acknowledgement from all three nodes in the pipeline. If you select GPFS as your file system when installing InfoSphere BigInsights, the `hbase.fsutil.gpfs.impl` property in the `hbase-site.xml` file is set to `org.apache.hadoop.hbase.util.FSGPFSUtils`.
4. Select the user that you want to install InfoSphere BigInsights with, and enter the required information for that user.
5. On the Nodes panel, click **Discover Disks** to discover all disks that are available.

   Review the GPFS disks that are discovered by the InfoSphere BigInsights installation program. If the discovered disks are correct for your cluster, complete the remaining panels in the installation program to install InfoSphere BigInsights over GPFS. Otherwise, click **Cancel** to save your current configuration and exit the installation program.

   Note: The InfoSphere BigInsights installation program generates two additional files that are used for diagnostics:

   - The `policyfile` file contains policies that determine data replication and placement for the different file sets.
   - The `clusterfile` file contains information about the cluster.
6. Navigate to the directory where you extracted the InfoSphere BigInsights `.tar.gz` file, and edit the following configuration files to match your environment. These files facilitate GPFS installation and configuration.
   - `nodefile`
   - `diskfile`
   - `bi_gpfs.cfg`

7. After editing the configuration files, use the `gpfs-create_clusterfs.sh` script to install GPFS. The installation script is located in the directory where you extracted the InfoSphere BigInsights .tar file.

   Restriction: Because the scripts use supporting scripts, you must use the script from the installed location. Copying the script to a different location, running the supporting scripts directly, and modifying the scripts directory structure is not supported.

   For example:

   ```
   installer/hdm/bin/gpfs-create_clusterfs.sh -f bi_gpfs.cfg
      -n nodefile -d diskfile
   ```

   The `gpfs-create_clusterfs.sh` script creates a GPFS cluster.

   ```
   Syntax: installer/hdm/bin/gpfs-create_clusterfs.sh
               -f param-file
               -n nodefile
               -d diskfile
               [ -i | -z ]
               [ -h ]
   ```

Table 9. Options for the gpfs-create_clusterfs.sh script. The table includes the options for the gpfs-create_cluster.sh script, including which options are required, and a detailed description of each option.

Option	Required	Description
`-f param-file`	Yes	Specify the path to the parameter file that contains the parameters to use in creating the GPFS cluster. The default name of this file is `bi_gpfs.cfg`.
`-n nodefile`	No	Specify the name of the file that indicates how nodes are to be treated in GPFS.
`-d diskfile`	No	Specify the name of the file that indicates disks to be used.
`-i` `-z`	No	If **`-i`** is specified, install only the GPFS `.rpm` files. If **`-z`** is specified, force reuse of disks marked as in use by GPFS.
`-h`	No	Print usage message and exit.

8. After you use the InfoSphere BigInsights scripts to install GPFS, restart the InfoSphere BigInsights installation program.
   a. On the File System panel, select **Use an existing General Parallel File System**.
   b. Complete the remaining panels in the installation program to install InfoSphere BigInsights over the GPFS that you created.

Node descriptor file

The node descriptor file is defined by GPFS and is used to provide information about each node in the cluster. The InfoSphere BigInsights installation program generates a node descriptor file that you can edit before you use the InfoSphere BigInsights script for installing GPFS.

When using the `gpfs-create_clusterfs.sh` script to install GPFS, specify the node descriptor file with the `-n` option. The following example shows the format of the node information that the script generates:

```
hdtest1.xxx.yyy.com:quorum:
hdtest2.xxx.yyy.com:quorum:
hdtest3.xxx.yyy.com:quorum:
```

Stanza file

The stanza file is defined by GPFS and is used to provide information about each disk that is included in the cluster. In addition, the stanza file defines the storage pools that each Network Storage Device is assigned to.

The InfoSphere BigInsights installation program generates a stanza file that you can edit before you use the InfoSphere BigInsights script for installing GPFS.

When using the `gpfs-create_clusterfs.sh` script to install GPFS, specify the stanza file with the `-d` option of the script.

The following example shows the format of the information that the script generates:

```
%pool:
pool=system
blockSize=1024K
layoutMap=cluster
allowWriteAffinity=no

%pool:
pool=datapool
blockSize=1024K
layoutMap=cluster
allowWriteAffinity=yes
writeAffinityDepth=1
blockGroupFactor=128

%nsd: device=/dev/sdb1 servers=hdtest1.xxx.yyy.com
  usage=dataAndMetadata failureGroup=1 pool=system
%nsd: device=/dev/sdb2 servers=hdtest1.xxx.yyy.com
  usage=dataOnly failureGroup=1,0,0 pool=datapool
%nsd: device=/dev/sdb3 servers=hdtest1.xxx.yyy.com
  usage=dataOnly failureGroup=1,0,0 pool=datapool
%nsd: device=/dev/sdb1 servers=hdtest2.xxx.yyy.com
  usage=dataAndMetadata failureGroup=2 pool=system
%nsd: device=/dev/sdb2 servers=hdtest2.xxx.yyy.com
  usage=dataOnly failureGroup=2,0,0 pool=datapool
%nsd: device=/dev/sdb3 servers=hdtest2.xxx.yyy.com
  usage=dataOnly failureGroup=2,0,0 pool=datapool
%nsd: device=/dev/sdb1 servers=hdtest3.xxx.yyy.com
  usage=dataAndMetadata failureGroup=3 pool=system
%nsd: device=/dev/sdb2 servers=hdtest3.xxx.yyy.com
  usage=dataOnly failureGroup=3,0,0 pool=datapool
%nsd: device=/dev/sdb3 servers=hdtest3.xxx.yyy.com
  usage=dataOnly failureGroup=3,0,0 pool=datapool
```

bi_gpfs.cfg configuration file

The `bi_gpfs.cfg` configuration file contains the parameters that are used to create the GPFS cluster.

You specify the `bi_gpfs.cfg` configuration file with the `-f` option of the `gpfs-create_clusterfs_opt.sh` script. The following example shows what the contents of the `bi_gpfs.cfg` file might look like:

```
##### BEGIN - PARAMETERS ######

PRIMARY="hdtest3.xxx.yyy.com"
SECONDARY="hdtest1.xxx.yyy.com"
```

```
TSCTCPPORT="1001"
DEFAULT_METADATA_REPLICATION=3
MAX_METADATA_REPLICATION=3
DEFAULT_DATA_REPLICATION=3
MAX_DATA_REPLICATION=3
BLOCK_ALLOCATION="cluster"
BLOCK_SIZE=1M
BLOCK_GROUP_FACTOR=128
WRITE_AFFINITY_DEPTH=1
ESTIMATED_CLUSTER_SIZE=32
TMP_FILESET="tmp"
LOG_FILESET="log"
MRL_FILESET="/hadoop/mapred/local"
MC_LIST="hdtest3.xxx.yyy.com hdtest1.xxx.yyy.com hdtest2.xxx.yyy.com"
NUM_QUORUM_NODES=3
PAGEPOOL_MEMPCT=25
GPFS_DISKS="/dev/sdb1 /dev/sdb2 /dev/sdb3"
ROOT_DIR="/root/gpfs"
GPFS_NAME="bigpfs"
GPFS_MOUNT="/mnt/bigpfs"
MOUNT_OPTION="yes"
USER=biadmin
GROUP=biadmin

##### END - PARAMETERS ######
```

Table 10. Parameters in the `bi_gpfs.cfg` file. The table lists the parameters in the bi_gpfs.cfg configuration file, including a detailed description of each parameter.

Parameter	Description
PRIMARY	Specifies the primary GPFS cluster server configuration node. For more information, see **mmcrcluster** *command* in the Cluster Products information center.
SECONDARY	Specifies the secondary GPFS cluster server configuration node. For more information, see **mmcrcluster** *command* in the Cluster Products information center.
TSCTCPPORT	Specifies the port for GPFS communication.
DEFAULT_METADATA_REPLICATION	Specifies default number of metadata replicas. Increasing this value enables large clusters to better tolerate failures. The specified value cannot be greater than the specified number of failure groups.
MAX_METADATA_REPLICATION	Specifies maximum number of metadata replicas. Increasing this value enables large clusters to better tolerate failures. The specified value cannot be greater than the specified number of failure groups.
DEFAULT_DATA_REPLICATION	Specifies default number of data replicas. Increasing this value enables large clusters to better tolerate failures. The specified value is cannot be greater than the specified number of failure groups.
MAX_DATA_REPLICATION	Specifies maximum number of data replicas. Increasing this value enables large clusters to better tolerate failures. The specified value cannot be greater than the specified number of failure groups.

Table 10. Parameters in the `bi_gpfs.cfg` file (continued). The table lists the parameters in the bi_gpfs.cfg configuration file, including a detailed description of each parameter.

Parameter	Description
BLOCK_ALLOCATION	Specifies the GPFS block allocation map type. For more information, see **mmcrfs** *command* in the Cluster Products information center.
BLOCK_SIZE	Specifies the size of the data blocks. The default is set to 1M. For more information, see **mmcrfs** *command* at the Cluster Products information center.
METADATA_BLOCK_SIZE	This parameter is currently not used.
BLOCK_GROUP_FACTOR	Specifies a logical grouping of blocks that enables the group to behave as one large block. The block group factor is used to determine the metablock size.
WRITE_AFFINITY_DEPTH	Specifies the number of copies of a file that are localized, as opposed to being striped across the disks in a cyclical fashion.
ESTIMATED_CLUSTER_SIZE	Specifies an estimated number of file system nodes. For more information, see the GPFS **mmcrfs** *command* at the Cluster Products information center.
TMP_FILESET	Specifies a fileset for temporary files.
LOG_FILESET	Specifies a fileset for log files.
MRL_FILESET	Specifies a fileset for MapReduce Local files.
MC_LIST	Specifies the list of all cluster nodes in space-delimited format.
NUM_QUORUM_NODES	Specifies the number of quorum nodes. This value is ignored if you run the **gpfs-create-clusterfs.sh** script and specify the `nodefile` parameter. This value must be an odd number, and the maximum is 7.
PAGEPOOL_MEMPCT	Specifies the GPFS value for the `pagepoolMaxPhysMemPct` variable. The specified value indicates the percentage of physical memory that can be assigned to the page pool. The InfoSphere BigInsights installation program initially sets the percentage to 25 unless you override the value. For more information, see **mmchconfig** *command* at the Cluster Products information center.
GPFS_DISKS	Specifies the raw disk device names in space-delimited format. **Important:** Do not use or modify this parameter. To specify your disk devices, use stanza files.
ROOT_DIR	Specifies the directory on the web console node where the files generated by the InfoSphere BigInsights installation program are located. These files are used by the GPFS installation scripts.
GPFS_NAME	Specifies the name for the GPFS file system.
GPFS_MOUNT	Specifies the mount point for the GPFS file system.

Table 10. Parameters in the `bi_gpfs.cfg` file (continued). The table lists the parameters in the bi_gpfs.cfg configuration file, including a detailed description of each parameter.

Parameter	Description
`MOUNT_OPTION`	Specifies whether to mount the GPFS file system on GPFS startup. The default value is `yes`.
`USER`	Specifies the InfoSphere BigInsights administrator user name.
`GROUP`	Specifies the InfoSphere BigInsights administrator group name.

Installing GPFS by using administration commands

You can install GPFS by using GPFS administration commands before you install InfoSphere BigInsights.

About this task

By using the GPFS administration commands, you can create a GPFS file system on nodes where no GPFS file system exists.

After you install GPFS by using these commands, you must install InfoSphere BigInsights over GPFS by using the installation program.

All of the InfoSphere BigInsights nodes that you plan to include in your cluster must also be included in your GPFS cluster.

Note: GPFS is not included as part of the InfoSphere BigInsights Quick Start Edition.

You must complete the following procedure as the root user.

Procedure

1. From the **`$EXTRACTION_DIRECTORY`**`/artifacts/gpfs` directory on the product media, locate and copy the self-extracting archive file, `gpfs_install-3.5`**`version`**, to a local directory. The `.rpm` files that GPFS requires are included in the archive file.

   `version` is the version of GPFS for your hardware platform and Linux distribution.

   `$EXTRACTION_DIRECTORY` is the directory where you extracted the installation files from the product media.

   For more details, refer to Installing GPFS on Linux nodes in the Cluster Products Information Center.
2. Install the GPFS `.rpm` files. In the directory where you copied the packages, run the **`rpm`** command to install the GPFS software on every node.
   a. Install the version 3.5.0.0 RPM packages.

      ```
      rpm -ivh gpfs*.rpm
      ```

   b. Install additional RPM packages.

      ```
      rpm -U gpfs*.rpm
      ```

The packages that you install differ depending on the version of InfoSphere BigInsights that you are installing. The following table shows the versions of InfoSphere BigInsights and the corresponding RPM package versions.

Table 11. Versions of InfoSphere BigInsights and the corresponding RPM package versions

Software version	RPM package version
2.1	3.5.0.9
2.1.0.1	3.5.0.11

c. Verify that the software is installed by running the following command on every node.

```
rpm –qa | grep gpfs
```

The output indicates whether the GPFS `rpm` files are installed on the selected nodes.

3. Run the following commands in succession on every node to build the GPFS portability layer. The GPFS portability layer is a loadable kernel module that is built for a specific Linux kernel version. The portability layer enables the GPFS daemon to interact with the Linux operating system.

```
cd /usr/lpp/mmfs/src
make Autoconfig
make World
make InstallImages
```

Important: You must update the GPFS kernel module any time that you update or patch the Linux kernel. Updating the GPFS kernel module after you update a Linux kernel requires rebuilding and installing a new version of the module.

4. Create the cluster by using the **mmcrcluster** command. The following command uses options that are shown in the following table to specify cluster parameters.

```
mmcrcluster -N node01:quorum,node02 -p node01 -s node02 -r /usr/bin/ssh
  -R /usr/bin/scp
```

*Table 12. Options for the **mmcrcluster** command, including a brief description of each option.*

Option	Description
-N	Comma-separated list of nodes (in the format *NodeName*:*NodeDesignation*:*AdminNodeName*)
-p	Primary configuration server
-s	Secondary configuration server
-r	ssh as the remote shell
-R	scp as the remote file copy command

5. Set the license designation for each node by using the **mmchlicense** command. The following example associates a server license with each of the cluster nodes.

```
mmchlicense server --accept -N node01,node02
```

6. Start GPFS on a node in the cluster.

```
mmstartup -a
```

7. From the **$EXTRACTION_DIRECTORY**/installer/hdm/bin/ directory, customize the gpfs-nsddevices script. The GPFS daemon uses this script during disk discovery processing.
   a. Open the gpfs-nsddevices script and modify it to match your system configuration. Refer to the script for a detailed description of how to modify it.
   b. Copy the modified script into the /var/mmfs/etc directory and assign permissions.
      ```
      chmod 0744 /var/mmfs/etc/nsddevices
      ```
8. Create the NSDs by using the **mmcrnsd** command as shown in the following example.
   ```
   mmcrnsd -F stanzaFile
   ```
   stanzaFile is the stanza file that defines the storage pools and disks that you want to use for the file system.
9. Use the **mmcrfs** command to configure the GPFS file system. The following example creates a file system named gpfs1 and mounts it at /gpfs1 automatically when GPFS starts.
   ```
   mmcrfs gpfs1 -F diskdef.txt -T /gpfs1 -A yes -m 1
     -M 3 -r 1 -R 3 -Q no -B 1024K -j cluster -S yes -E no
   ```
10. Mount the file system, which is shown as /gpfs1 in the following example.
    ```
    mmmount /gpfs1 -a
    ```
11. Run the **mmchconfig** command to set the readReplicaPolicy to local. The local option indicates that a read is done from a replica that is available from a local NSD server.
    ```
    mmchconfig readReplicaPolicy=local
    ```
12. Create the /var/mmfs/bi directory and ensure that the directory is owned by the InfoSphere BigInsights administrator.
    a. Create the bi directory in a temporary location on your file system.
       ```
       mkdir /tmp/bi
       ```
    b. Assign the owner and group of the directory to the InfoSphere BigInsights administrator. In the following example, biadmin is the name of both the user and group for the InfoSphere BigInsights administrator.
       ```
       chown biadmin:biadmin /tmp/bi
       ```
    c. Move the bi directory to the /var/mmfs directory.
       ```
       mv /tmp/bi /var/mmfs
       ```
13. Create a default fileset for the MapReduce **mapred.local.dir** property.

    The following commands create a fileset on the GPFS file system that you can configure to use a data replication factor of 1, which helps to avoid replicating intermediate MapReduce data. The following example assumes that you specify the cache directory in the InfoSphere BigInsights installation program as /hadoop/mapred/local.

    Important: InfoSphere BigInsights currently does not support hosting the MapReduce directory outside of the GPFS file system. If you want this directory to exist on your local file system, change the value for the **mapred.local.dir** property after you successfully install InfoSphere BigInsights.
    a. Create a fileset for MapReduce.
       ```
       mmcrfileset device mrl_set
       ```
       device is the device name for GPFS, which is typically /dev/**gpfs_name**, where **gpfs_name** is the name of your GPFS installation.

b. Create the MapReduce directory on the GPFS file system.

```
mkdir -p mount_point/hadoop/mapred
```

mount_point is the mount point for GPFS.

c. Create a junction that references the root directory of the GPFS fileset.

```
mmlinkfileset device mrl_set -J mount_point/hadoop/mapred/local
```

For an example of applying a policy rule for specifying a data replication of 1, see this step.

14. Create user directories for GPFS and set permissions for them. The following commands assumes that /gpfs1 is the mount point for GPFS, and that biadmin is the InfoSphere BigInsights administrator user and group.

```
mkdir -p /gpfs1/user

chmod 1777 /gpfs1/user

chown -R biadmin:biadmin /gpfs1

chmod -R +rx /gpfs1
```

15. Copy the following files into the /var/mmfs/etc directory on every node in your cluster and assign permissions.

```
$EXTRACTION_DIRECTORY/artifacts/gpfs/platform/gpfs-connector-daemon
$EXTRACTION_DIRECTORY/installer/hdm/components/gpfs/binary/gpfs-callback*sh

##Assign permissions to the files that you copied##
chmod 0744 /var/mmfs/etc/gpfs-callback*sh
chmod 0100 /var/mmfs/etc/gpfs-connector-daemon
```

platform is your system type, which can be either Linux-amd64-64 or Linux-ppc64-64.

16. Install callbacks.

```
$EXTRACTION_DIRECTORY/installer/hdm/bin/gpfs-callbacks.sh --add
```

17. Run the following commands to change configuration parameters for GPFS.

```
/usr/lpp/mmfs/bin/mmchconfig readReplicaPolicy=local
/usr/lpp/mmfs/bin/mmchconfig restripeOnDiskFailure=yes -i
/usr/lpp/mmfs/bin/mmchpolicy device policyFile -I yes
```

policyFile typically refers to a storage pool with file placement optimizer (FPO) properties. The following example includes rules to set the data replication factor for a fileset to 1, and to use that data pool as the storage pool.

```
rule 'R1' SET POOL 'datapool' REPLICATE (1,3) FOR FILESET (mrl_set)
rule default SET POOL 'datapool'
```

18. Enable monitoring for GPFS.

a. Ensure that monitoring runs successfully.

```
$EXTRACTION_DIRECTORY/installer/hdm/bin/precheck-snmp.sh
```

b. If the **precheck-snmp.sh** script completes successfully, then run the following command to enable monitoring for GPFS.

```
$EXTRACTION_DIRECTORY/installer/hdm/bin/gpfs-setup-snmp.sh
```

19. Restart GPFS.

```
/usr/lpp/mmfs/bin/mmumount /dev/bigpfs -a
/usr/lpp/mmfs/bin/mmshutdown -a
/usr/lpp/mmfs/bin/mmstartup -a
```

20. Start the InfoSphere BigInsights installation program.

a. On the Installation Type panel, select **Cluster installation**.

b. On the File System panel, select **Use existing General Parallel File System GPFS**, and then enter the GPFS mount point.

Important: HBase requires that your distributed file system supports the **sync** call. This call pushes data through the write pipeline and blocks it until the data receives acknowledgement from all three nodes in the pipeline. If you select GPFS as your file system when installing InfoSphere BigInsights, the `hbase.fsutil.gpfs.impl` property in the `hbase-site.xml` file is set to `org.apache.hadoop.hbase.util.FSGPFSUtils`.

c. Select the user that you want to install InfoSphere BigInsights with, and enter the required information for that user.

d. Complete the remaining panels in the installation program to complete the installation.

Installing InfoSphere BigInsights by using the wizard

You can use a web browser to run the InfoSphere BigInsights installation program. Run the installation program on the cluster node where you want to install the software.

Before you begin

Complete each of the following tasks before you start the installation program.

- Review the system requirements and release notes
- Choose a user to install the product with
- Configure your browser
- Obtain the product software
- Complete all prerequisite tasks
- Optional: Select the options that you want the prerequisite checker utility to run with

Installing with high availability: If you are installing with high availability or want to deploy your cluster with adaptive MapReduce, complete the following tasks before starting the installation program:

- Enable adaptive MapReduce with high availability.
- Ensure that the shared Network File System (NFS) is mounted on all high availability nodes, and that it is readable and writeable by the cluster administrator.
- Ensure that the reserved IP address for the NameNode is not used by any host.
- If you run the installation program as a non-root user, ensure that you can run the **nfsstat** command on the high availability nodes through SSH. For example, log in as the cluster administrator and run the following command from the installation host, where *server_name.example.com* is the name of your high availability server.

  ```
  ssh server_name.example.com nfsstat
  ```

  If the output reads `bash:nfsstat: command not found`, ensure that the `nfs-utils` RPM package is installed on each of the high availability nodes. In addition, add the location of the **nfsstat** command to the PATH variable in the `.bashrc` file for the system administrator on each of the high availability nodes in your cluster. The default location of the **nfsstat** command is `/usr/sbin`.

Installing with GPFS: If you are installing GPFS as your distributed file system, ensure that you read and understand the following considerations.

- When installing over an existing version of GPFS, ensure that the InfoSphere BigInsights administrator user has read, write, and execute permissions on the GPFS mount point.
- The installation program supports only GPFS FPO configurations when you install InfoSphere BigInsights and GPFS at the same time. The installation program does not support installing GPFS only. Alternatively, you can install GPFS by using scripts or commands.
- HBase requires that your distributed file system supports the **sync** call. This call pushes data through the write pipeline and blocks it until the data receives acknowledgement from all three nodes in the pipeline. If you select GPFS as your file system when installing InfoSphere BigInsights, the `hbase.fsutil.gpfs.impl` property in the `hbase-site.xml` file is set to `org.apache.hadoop.hbase.util.FSGPFSUtils`.
- GPFS is not included as part of the InfoSphere BigInsights Quick Start Edition
- GPFS does not currently support the high availability feature.

Procedure

1. Navigate to the directory where you extracted the `biginsights-enterprise-linux64_`*`release_number`*`.tar.gz` file, where *release_number* is the release number that you are installing.
2. Run the `start.sh` script.

   ```
   ./start.sh
   ```

   The script starts WebSphere® Application Server Community Edition on port 8300. The script provides you with a URL to the installation wizard, which is available at:

   ```
   http://server_name:8300/Install/
   ```

   server_name is the server where you extracted the `.tar.gz` file. Multiple URLs might be provided if the server has multiple IP addresses; pick one that is accessible from your browser.
3. Complete the remaining panels in the installation wizard.
   a. Review the Welcome panel, and then click **Next**.
   b. Review the License Agreement panel, accept the terms in the license agreement, and then click **Next**.
   c. On the Installation Type panel, select **Cluster installation**, and then click **Next**.
   d. On the File System panel, make selections for your distributed file system, and then click **Next**.
      - Select the distributed file system that you want to install, either HDFS or GPFS.

        Installing with high availability: If you are installing with high availability, select **Install Hadoop Distributed File System (HDFS)**.
      - Specify the installation directories for InfoSphere BigInsights.
      - Expand **MapReduce general settings** and specify the MapReduce settings that you want to use.

Table 13. MapReduce general settings

Directory	Description
Cache directory	Directory where the MapReduce intermediate data (map output data) is stored
Log directory	Directory where MapReduce logs are written to
MapReduce system directory	System directory where Hadoop stores its configuration data

e. On the Secure Shell panel, specify the user that you want to install the product with, and then click **Next**. For more information, see "Choosing a user to install the product with" on page 31.

f. On the Nodes panel, click **Add Node** to add single nodes, or **Add Multiple Nodes** to add several nodes simultaneously. For each node, use the `Short host name` and `Rack ID` that you recorded in "InfoSphere BigInsights installation worksheet" on page 20.

Installing with GPFS: If you are installing GPFS as your distributed file system, on the Add Nodes panel or the Add Multiple Nodes panel, enter the disks that you want to use for GPFS in the **Disks to use for GPFS** field. You can also click **Discover Disks** on the Nodes panel to discover all disks that are available.

The installation program overwrites all disks that you specify for each node. Ensure that you specify the correct disks to avoid losing data.

Installing with high availability: If you are installing with high availability, add all nodes in the cluster, including the nodes intended as high availability nodes.

After you finish adding nodes, click **Next**.

g. On the Components 1, Components 2, and Components 3 panels, specify the host names and port numbers for each of the components that you are installing. For more information, see "InfoSphere BigInsights installed components" on page 18.

Important: Use a dedicated node for ZooKeeper to ensure that the HBase region server connection to ZooKeper does not time out. You can specify a dedicated node for ZooKeeper on the Components 3 panel in the InfoSphere BigInsights installation program.

Installing with high availability:

- On the Components 1 panel, ensure that no service is assigned to any of the high availability nodes.
- On the Components 2 panel, select the **Configure High Availability** option.
  1) Next to the **High Availability nodes** field, click **Assign**. Add the high availability nodes by selecting them in the left pane and clicking the right arrow. You must select at least two nodes, but cannot select more than three nodes.
  2) In the **Virtual NameNode FQDN** field, enter the fully qualified domain name for the NameNode, which should resolve to the unassigned virtual IP address for the NameNode.

3) In the **Virtual NameNode IP address** field, enter the unassigned virtual IP address for the NameNode.
4) In the **NFS server information** field, enter the server and NFS directory in the following format: *server:shared_directory*. For example, `nfs-server.com:/remote/path`.
5) In the **NFS local mount point** field, enter the path to the mount point of the NFS shared directory.
6) Next to the **NameNode** field, click **Assign** to specify which high availability nodes you want to run the NameNode and JobTracker processes. The Secondary NameNode cannot be one of the high availability nodes.

h. On the Security panel, specify the type of authentication that you want to use, and then click **Next**. For more information, see "Choosing user security and authentication" on page 26.

4. When you reach the Summary panel, review the information for the settings, nodes, and components.
5. Click **Install** to start the installation. The installation progress shows so that you know how much time is remaining in the installation process.
6. When the installation completes, click **Finish** to stop the web server. Alternatively, you can run the `start.sh shutdown` script after the installation completes.
7. Optional: To clear disk space, you can remove the extracted installation files from your system.

Installing InfoSphere BigInsights by using a response file

You can use silent mode to run an unattended installation of InfoSphere BigInsights. In silent mode, the installation program does not display a user interface. Instead, it reads settings from a response file that you create, completes a prerequisites check, and installs the software if the check succeeds.

About this task

This procedure describes how to run an unattended installation by modifying the sample response file. Alternatively, you can run the InfoSphere BigInsights installation program and choose to create a response file without installing the product. You can then use this response file to install InfoSphere BigInsights in silent mode.

Procedure

1. Choose a path for the installation program such as `$BigInsightsInstaller`, to point to the installation directory you created when you extracted the InfoSphere BigInsights `tar.gz` file.
2. Navigate to the `$BIGINSIGHTSInstaller/silent-install` directory. This directory contains several sample response files named `sample-install-`*`security_type`*`.xml`, where *`security_type`* is the type of security that you want to run the installation program with.
3. Edit the response file that contains the security configuration that you want to use to install the software. These files are located in the `/silent-install` directory. Each file has a standard format that must be preserved during editing.

sample-install-noSecurity.xml
Install the InfoSphere BigInsights web console with no user authentication.

sample-install-enterprise-ldap.xml
Install the InfoSphere BigInsights web console with Lightweight Directory Access Protocol (LDAP) authentication.

sample-install-enterprise-defaultFlat.xml
Install the InfoSphere BigInsights with default security configuration and flat file authentication.

sample-install-enterprise-customizedFlat.xml
Install the InfoSphere BigInsights web console with customized security configuration and flat file authentication.

sample-install-enterprise-pam.xml
Install the InfoSphere BigInsights web console with Pluggable Authentication Module (PAM) authentication.

4. If you want to install InfoSphere BigInsights with adaptive MapReduce, you must modify one of the following settings in the response file that you edited, depending on whether you want to install adaptive MapReduce with high availability.

```
<hadoop>
  <general>
    ...
    #Set this value to true if you want to install adaptive MapReduce
     instead of Apache MapReduce.
    <apache-mapred>true</apache-mapred>
  </general>
  ...
  <high-availability>
    #Set this value to true if you want to install adaptive MapReduce
     with high availability.
    <configure>true</configure>
  </high-availability)
```

5. Save and close the sample response file that you edited.
6. In the/`silent-install` directory, run the **silent-install.sh** script, including the file name of the sample response file.

```
./silent-install.sh sample-install-noSecurity.xml
```

Results

The silent installation process returns a log file named `silent-install.log_`*time_stamp* in the `$BigInsightsInstaller/silent-install` directory. For example, the log file name might be something like `silent-install_2013-05-03T03.50.36.593_PDT.log`. This log file contains information about the installation process, and can be used to troubleshoot problems with your installation.

Installing the InfoSphere BigInsights Tools for Eclipse

Installing InfoSphere BigInsights Tools for Eclipse adds capabilities to your Eclipse development environment so that you can develop applications to run with InfoSphere BigInsights.

Before you begin

Install and configure all required software for your operating system before you run the installation program. Both Java and Eclipse must have the same bit level, either 32-bit or 64-bit.

Installing with GPFS

If you want to run Jaql, Pig, and HBase programs from the InfoSphere BigInsights Tools for Eclipse against an InfoSphere BigInsights cluster that uses GPFS as the distributed file system, then you must install your Eclipse client on a computer where GPFS is mounted.

Operating system	Procedure
All operating systems	• Install and configure a supported web browser. • Download and install the Eclipse IDE 3.6.2 for Java EE developers from the Eclipse website. You can install the InfoSphere BigInsights Tools for Eclipse into the IBM Data Studio or IBM Rational product environments that are based on Eclipse 3.6.2. • Remove the Eclipse Data Tools Platform (DTP) if you previously installed Version 1.9 or higher.
Linux	• Download and install the Runtimes for Java Technology, Version 6.0.11.0 from IBM Fix Central. • Specify the JVM for Eclipse to run on. For more information, see "Specifying the JVM" on the Eclipse website.
Microsoft Windows	• Install and configure Cygwin on the computer where Eclipse is installed. Download and install Cygwin from the Cygwin website. • Add the *install_dir*\cygwin\bin directory to your PATH environment variable. *install_dir* is the directory where you installed Cygwin. • If you need a version of IBM Java, download the IBM Development Package for Eclipse from developerWorks®. • Specify the JVM for Eclipse to run on. For more information, see "Specifying the JVM" on the Eclipse website. • Ensure that Eclipse is installed in a local directory, and that your Eclipse workspace is saved in a local directory. Using a Universal Naming Convention (UNC) path is not supported.

Procedure

1. Log in to the InfoSphere BigInsights Console.

Option	Description
In a non-SSL installation	Enter the following URL in your browser: `http://host_name:8080` *host_name* is the name of the host where the InfoSphere BigInsights Console is running.
In an SSL installation	Enter the following URL in your browser: `https://host_name:8443` *host_name* is the name of the host where the InfoSphere BigInsights Console is running.

2. From the Welcome panel, click **Enable your Eclipse development environment for BigInsights application development**.

Option	Description
To install from the web server	Copy the `http://server_name:port/updatesite/` URL. *server_name* is the name of your InfoSphere BigInsights server. *port* is the port number where InfoSphere BigInsights is running. The default port is `8080` for HTTP and `8443` for HTTPS.
To download an archive that includes the plugins	Download the `BigInsightsEclipseTools.zip` file.

3. Start Eclipse. From the Eclipse taskbar, click **Help** > **Install New Software**.

   Important: If you installed IBM Data Studio or IBM Rational products into a virtualized directory like the `Program Files` directory, you must run IBM Data Studio or IBM Rational products as an administrator user to install the InfoSphere BigInsights Tools for Eclipse.
4. In the Install window, click **Add**.

Option	Description
To install from the web server	Enter the URL for your InfoSphere BigInsights server that you copied previously.
To install by using the file that you downloaded	Click **Archive** to browse to the location of the `BigInsightsEclipseTools.zip` file.

5. Click **OK**. In the lower section of the Install window, select the **Group items by category** check box to compress the list of options to install.
6. Select the check box for **IBM InfoSphere BigInsights**, and all entries under that category.
7. Follow the steps in the installation wizard to install the InfoSphere BigInsights client for Eclipse.

   Attention: If you use the internal browser in Eclipse on Red Had Enterprise Linux or Suse Linux Enterprise Server, Eclipse can potentially display error messages continuously or hang. The problem is known to occur with Eclipse

3.6.2, but the problem can occur with other versions of Eclipse. For more information, see Issue with the Eclipse browser on the IBM Support website.

On Microsoft Windows, configure Eclipse to use the external web browser and select a supported web browser version.

Configuring access to the default task controller

During installation, the InfoSphere BigInsights installation program installs and configures a task controller for you. If you ran the installation program as a user that is not root, and modified the path of the task controller configuration file, you must provide root access to the directory that you specified.

About this task

The default task controller configuration path and file name is `/var/bi-task-controller-conf/taskcontroller.cfg`.

Procedure

1. Navigate to the directory that you specified for the default task controller when you ran the installation program.
2. Configure root access to the task controller configuration file and all of its parent directories.

Installing and configuring a Linux Task Controller

If you install or upgrade InfoSphere BigInsights with a user ID that does not have root privileges and you need to enable security, you must manually install and configure a Linux Task Controller.

Procedure

1. Run the following command to create the task controller binary:

   ```
   ant task-controller -Dhadoop.conf.dir=/setuid_conf
   ```

   /setuid_conf is the directory where you want to deploy the Linux Task Controller configuration file.
2. Copy the task controller binary to the `$BIGINSIGHTS_HOME/hdm/hadoop-conf-staging/` directory. The directory must have 6050 or `--Sr-s---` permissions and be owned by the root user and the InfoSphere BigInsights administrator group.
3. In the `$BIGINSIGHTS_HOME/hdm/hadoop-conf-staging/` directory, open the `mapred-site.xml` file and add the following properties.

   ```
   <property>
        <name>mapred.task.tracker.task-controller</name>
        <value>org.apache.hadoop.mapred.LinuxTaskController</value>
   </property>

   <property>
        <name>mapreduce.tasktracker.group</name>
       <value>biadmin</value>
   </property>
   ```
4. From the `$BIGINSIGHTS_HOME/bin` directory, run the **syncconf.sh** command to synchronize the nodes.

   ```
   ./syncconf.sh
   ```

5. Move the taskcontroller.cfg task controller configuration file to the /setuid_conf directory. The /setuid_conf directory is owned by the root user, and has 755 permissions. The taskcontroller.cfg file has the following permission:

   ```
   -r-------- 1 root biadmin
   ```

   The taskcontroller.cfg file contents are:

   ```
   mapred.local.dir=/hadoop/mapred/local
   hadoop.log.dir=/var/ibm/biginsights/hadoop/logs
   mapreduce.tasktracker.group=biadmin
   min.user.id=100
   ```

   a. Set a min.user.id that is appropriate for your deployment.
   b. Ensure that the mapred.local.dir directory and the hadoop.log.dir directory match the corresponding settings of the $HADOOP_CONF_DIR directory in the cluster.

What to do next

If you configured flat file authentication, run the **createosusers.sh** script from the $BIGINSIGHTS_HOME/bin directory to create a group for your users. The first group from the list of groups is selected by default.

The variation of the command that you run depends on the type of authentication that you selected when you ran the installation program. You must also run this command after you add a new node to your cluster.

- No authentication

  ./createosusers.sh *$BIGINSIGHTS_HOME*/console/conf/security/biginsights_group.xml

- Root authentication

  ./createosusers.sh *$BIGINSIGHTS_HOME*/console/conf/security/biginsights_group.xml root *password*

- Non-root authentication

  ./createosusers.sh *$BIGINSIGHTS_HOME*/console/conf/security/biginsights_group.xml nonrootuser *password*

Chapter 5. Upgrading InfoSphere BigInsights software

You can upgrade an existing installation of InfoSphere BigInsights to a more recent version, or to repair an existing installation. Use the following procedures to upgrade InfoSphere BigInsights and the InfoSphere BigInsights client for Eclipse.

Upgrading from Version 1.3.0.1 or later to Version 2.1

You can upgrade directly from InfoSphere BigInsights Version 1.3.0 fix pack 1 or later to InfoSphere BigInsights Version 2.1.

Important: If the cluster that you are upgrading was not installed with the high availability option, then you can choose to install either adaptive MapReduce or Apache MapReduce as your MapReduce option. You can upgrade from Apache MapReduce to adaptive MapReduce, or from adaptive MapReduce to Apache MapReduce.

Upgrading from Version 1.3 to Version 2.1

While you cannot upgrade directly from InfoSphere BigInsights Version 1.3 to InfoSphere BigInsights Version 2.1, you can upgrade from Version 1.3 to Version 1.3.0 fix pack 1, and then to Version 2.1.

If you follow this upgrade path, clear your web browser cache after you finish the upgrade from Version 1.3 to Version 1.3.0 fix pack 1, and again before you start the upgrade from Version 1.3.0 fix pack 1 to Version 2.1. If you do not clear the cache, an incorrect version number might display in the web browser.

Preparing to upgrade software

Before you begin upgrading software, ensure that you complete all prerequisite tasks.

About this task

Your distributed file system must be in a healthy state, which means that no missing or corrupted blocks exist in your distributed file system.

Procedure

1. Check the health of your distributed file system. The following command generates a report that indicates the health of your distributed file system, including the number of corrupt and missing files and blocks.

   ```
   hadoop fsck /> fsck.output
   ```

   The following example shows what a corrupted distributed file system looks like. If your distributed file system is corrupted in any way, fix all reported files and blocks before proceeding with an upgrade.

   ```
   Total size:    15411396714043 B
    Total dirs:    180
    Total files:   1458
    Total blocks (validated):      82230 (avg. block size 187418177 B)
     ********************************
     CORRUPT FILES:        12
     MISSING BLOCKS:       22
   ```

```
 MISSING SIZE:         10902081596 B
 CORRUPT BLOCKS:       22
 ********************************
Minimally replicated blocks:   82208 (99.97324 %)
Over-replicated blocks:        7163 (8.710933 %)
Under-replicated blocks:       0 (0.0 %)
Mis-replicated blocks:         74584 (90.70169 %)
Default replication factor:    3
Average block replication:     2.94215
Corrupt blocks:                22
Missing replicas:              0 (0.0 %)
Number of data-nodes:          18
Number of racks:               2
FSCK ended at Fri Mar 08 12:57:14 EST in 1268 milliseconds


The filesystem under path '/' is CORRUPT
```

2. Confirm that the InfoSphere BigInsights Administrator User that starts the InfoSphere BigInsights installation program has read and write permissions on the following directories.

The table lists the directories that the InfoSphere BigInsights Administrator User must have read and write permissions for.

Directory	Description
BIGINSIGHTS_HOME	The InfoSphere BigInsights installation directory. If *biginsights-install-directory* is a relative path, the value of *directory-prefix* in the `install.xml` file is the parent directory. The default value is `/opt/ibm/biginsights`.
BIGINSIGHTS_VAR	The InfoSphere BigInsights components log directory. The directory value is the same as the *biginsights-data-log-directory* value in the `BIGINSIGHTS_HOME/conf/install.xml` file on the Console node. If *biginsights-data-log-directory* is a relative path, the value of *directory-prefix* in the `install.xml` file is the parent directory. The default directory value is `/var/ibm/biginsights`.
Hadoop data directories	The Hadoop DataNode data storage directories. The directory values are the same as the *dfs.dta.dir* values in the `BIGINSIGHTS_HOME/hdm/hadoop-conf/hdfs-site.xml` file on the individual Hadoop Data Nodes.
Hadoop name directory	The Hadoop NameNode data storage directory. The directory value is the same as the *dfs.name.dir* value in the `BIGINSIGHTS_HOME/hdm/hadoop-conf/hdfs-site.xml` file.
Hadoop check point directory	The Hadoop check point data storage directory. The directory value is the same as the *fs.checkpoint.dir* value in the `BIGINSIGHTS_HOME/hdm/hadoop-conf/core-site.xml` file on the Hadoop Secondary NameNode.

The table lists the directories that the InfoSphere BigInsights Administrator User must have read and write permissions for.

Directory	Description
Hadoop MapReduce log directory	The Hadoop MapReduce intermediate data file storage directory. The directory value is the same as the *mapred.local.dir* value in the `BIGINSIGHTS_HOME/hdm/hadoop-conf/mapred-site.xml` file on each Hadoop TaskTracker node. On the Console node, the directory value is found in the `$HADOOP_SECURITY_CONF_DIR/taskcontroller.cfg` file.
Hadoop log directory	The Hadoop log storage directory. The directory value is the same as the *log-directory* value in the Hadoop section in the `BIGINSIGHTS_HOME/conf/install.xml` file on the Console node. The default directory value is `$BIGINSIGHTS_VAR/hadoop/logs`.
Hadoop temp directory	The base directory for Hadoop to create temporary directories. The directory value is the same as the *hadoop.tmp.dir* value in the `BIGINSIGHTS_HOME/hadoop-conf/core-site.xml` file on the Console node.

Upgrading InfoSphere BigInsights

You can upgrade InfoSphere BigInsights version 1.3.0 fix pack 1 or later, version 1.4.0, or version 2.0.

Before you begin

Ensure that you have completed all prerequisite tasks in "Preparing to upgrade software" on page 63.

About this task

During an upgrade, the installation program reads and uses your current XML file settings, including any nodes that were added after your previous installation. The installation program then runs a script that renames the `$BIGINSIGHTS_HOME` directory and the `$BIGINSIGHTS_VAR` directory, and backs up the `HDFS` name directory.

The following lists describe the information that is preserved and deleted when upgrading InfoSphere BigInsights.

Preserved information

- Application permissions
- Application state
- BigSheets workbooks
- Custom applications

If you upgrade from Version 1.3 fix pack 1 or Version 1.4 to Version 2.0, workbook data is not preserved, and you must manually run your BigSheets workbooks as needed.

If you upgrade from an installation that does not use security to an installation that uses security, BigSheets workbooks that were created before upgrading are not visible. If you roll back to a previous installation, the BigSheets workbooks are visible.

Deleted information

- Application security settings
- Application history
- Application status
- Authentication schema
- Changes to application descriptions
- Custom configuration changes to any of the InfoSphere BigInsights components
- Customized icons
- Shared libraries
- Tags

Procedure

1. Reassign any of your current users who are associated with the `biadmin` group to a new group, such as `biusers`.
2. If any of your current nodes are not functioning properly, remove them.
3. On your current HDFS environment, stop all InfoSphere BigInsights processes, put Hadoop into safe mode, create a directory to store the command results, and then start all HDFS processes.
   a. Stop your current services by running the **stop-all.sh** command from the `$BIGINSIGHTS_HOME/bin` directory.
      ```
      ./stop-all.sh
      ```
   b. Start your current HDFS by running the **start-dfs.sh** command from the `$BIGINSIGHTS_HOME/IHC/bin` directory.
      ```
      ./start-dfs.sh
      ```
   c. Create a directory named `mkdir $BIGINSIGHTS_VAR/__upgrade` to store the results of running the **fsck** command.

      Important: The directory must start with two underscores or the directory is moved during the upgrade.
   d. Put Hadoop into safe mode.
      ```
      hadoop dfsadmin -safemode enter
      ```
   e. Run the **hadoop fsck** command, including the directory that you created to store the command results.
      ```
      $BIGINSIGHTS_HOME/IHC/bin/hadoop fsck /
        -files -blocks -locations > $BIGINSIGHTS_VAR/__upgrade/dfs-fsck-old.log
      ```
   f. Run the **tail** command to read the last 20 lines of the `dfs-fsck-old.log` file to determine whether the HDFS file system is healthy. The command should return a line like, `The filesystem under path '/' is HEALTHY`.
      ```
      tail -20 $BIGINSIGHTS_VAR/__upgrade/dfs-fsck-old.log
      ```
   g. Generate a list of all Hadoop directories and files starting from the / (root) directory.

```
$BIGINSIGHTS_HOME/IHC/bin/hadoop fs -lsr / > $BIGINSIGHTS_VAR/__upgrade/dfs-lsr-old.log
```

h. Generate a snapshot of all Datanodes and their current status as recognized by the NameNode node.

```
$BIGINSIGHTS_HOME/IHC/bin/hadoop dfsadmin -report > $BIGINSIGHTS_VAR/__upgrade/dfs-report-old.log
```

i. Stop all InfoSphere BigInsights processes by running the **stop-all.sh** command from the $BIGINSIGHTS_HOME/bin directory.

```
./stop-all.sh
```

4. Verify your current permissions.
   - Your InfoSphere BigInsights administrator must have permissions to create the following directories under the /hadoop/hdfs and /hadoop/mapred directories. The following directory names are the default names. You might have specified different directories when you installed the latest version of InfoSphere BigInsights.
     - /hadoop/hdfs/name, the name directory
     - /hadoop/hdfs/namesecondary, the secondary NameNode data directory
     - /hadoop/hdfs/data, the data directory
     - /hadoop/mapred/local, the cache directory
   - If the current log directory is not in the $BIGINSIGHTS_VAR directory, your InfoSphere BigInsights administrator must also have permissions to create the /var/ibm/biginsights/hadoop/logs log directory.
5. Start the InfoSphere BigInsights installation program by running the start.sh installation script, and complete the installation program as the existing InfoSphere BigInsights administrator.

```
./start.sh
```

6. After the installation program completes, stop and then start all InfoSphere BigInsights processes.
   a. Stop all services by running the **stop-all.sh** command.
   b. End any InfoSphere BigInsights Java processes that are still running.
   c. Start HDFS by running the **start-dfs.sh** command.
   d. Run the **hadoop fsck** command, including the directory that you created to store the command results.

```
$BIGINSIGHTS_HOME/IHC/bin/hadoop fsck /
  -files -blocks -locations > $BIGINSIGHTS_VAR/__upgrade/dfs-fsck-new.log
```

   e. Generate a list of all Hadoop directories and files starting from the / (root) directory.

```
$BIGINSIGHTS_HOME/IHC/bin/hadoop fs -lsr / > $BIGINSIGHTS_VAR/__upgrade/dfs-lsr-new.log
```

   f. Generate a snapshot of all Datanodes and their current status as recognized by the NameNode node.

```
$BIGINSIGHTS_HOME/IHC/bin/hadoop dfsadmin -report > $BIGINSIGHTS_VAR/__upgrade/dfs-report-new.log
```

7. Run the **diff** command to compare log files before and after you run the installation program. If the **diff** command identifies differences in the new and old log files, examine the logs carefully for corruption or errors.
   a. Compare the log files from running the **hadoop fsck** command:

```
diff -q $BIGINSIGHTS_VAR/__upgrade/dfs-fsck-new.log
  $BIGINSIGHTS_VAR/__upgrade/dfs-fsck-old.log
```

   b. Compare the log files from running the **lsr** command:

```
diff -q $BIGINSIGHTS_VAR/__upgrade/dfs-lsr-new.log
  $BIGINSIGHTS_VAR/__upgrade/dfs-lsr-old.log
```

   c. Compare the log files from running the **report** command:

```
diff -q $BIGINSIGHTS_VAR/__upgrade/dfs-report-new.log
  $BIGINSIGHTS_VAR/__upgrade/dfs-report-old.log
```

8. If you modified any configuration files from your previous installation, these changes must also be applied in the upgraded installation. The installation program stores a copy of all configuration files from your previous installation in the $BIGINSIGHTS_HOME/__biattic__biginsights/ directory.
9. If you discover you need to roll back to your previous installation, run the following commands to return to your previous installation.
   a. Stop all services by running the **stop-all.sh** command.
   b. Verify that no InfoSphere BigInsights processes are running.
      ```
      ps -ef | grep java | grep biadmin
      ```
   c. Stop any processes that are still running.
      ```
      kill -9 process_ids
      ```
      process_ids is the list of processes that were indicated as running.
   d. Run the **rollbackUpgrade.sh** script from the BIGINSIGHTS_HOME/bin directory to start the rollback process.
      ```
      ./rollbackUpgrade.sh
      ```

   Important: You cannot run the **rollbackUpgrade.sh** script after you run the **finalizeUpgrade.sh** command in the next step.
10. If you successfully completed the upgrade, run the **finalizeUpgrade.sh** command from the $BIGINSIGHTS_HOME/bin directory to:
    - Delete the old $BIGINSIGHTS_HOME and $BIGINSIGHTS_VAR directories.
    - Finalize the HDFS upgrade.

Upgrading the InfoSphere BigInsights Tools for Eclipse

You can upgrade the installed version of the InfoSphere BigInsights Tools for Eclipse.

About this task

If you are upgrading the InfoSphere BigInsights Tools for Eclipse in a typical deployment, complete the following procedure.

Procedure

Upgrade the InfoSphere BigInsights Tools for Eclipse.

1. If you registered the update site URL for a previous version of the InfoSphere BigInsights Tools for Eclipse, remove the update site URL from the list of available update sites.
2. Restart Eclipse.
3. Follow the instructions for "Installing the InfoSphere BigInsights Tools for Eclipse" on page 58.

Migrating from HDFS to GPFS

You can migrate from the Hadoop Distributed File System (HDFS) to the IBM General Parallel File System (GPFS). The following procedures outline the steps that are required to migrate your distributed file system.

About this task

For assistance with migrating your distributed file system from HDFS to GPFS, contact IBM Support.

Procedure

1. Install the latest version of InfoSphere BigInsights to upgrade your existing HDFS.
2. Choose the disks to use for GPFS.
3. Generate the required files to create the GPFS cluster.
4. Complete the migration process and sync configuration changes.
5. Update InfoSphere BigInsights catalog tables that reference HDFS if necessary.

Chapter 6. Removing InfoSphere BigInsights software

To remove InfoSphere BigInsights software and features, use the provided scripts, or manually run individual commands to complete the software removal process.

Removing InfoSphere BigInsights by using scripts

To remove InfoSphere BigInsights software from the nodes in your cluster, run the script that is provided with the software.

About this task

This script ends all InfoSphere BigInsights processes on all cluster nodes, and deletes all files in the `$BIGINSIGHTS_HOME` directory, `$BIGINSIGHTS_VAR` directory, and all other Hadoop data directories.

Procedure

1. Log on to the management node as the InfoSphere BigInsights administrator user. The management node is the node where the InfoSphere BigInsights web console is installed.
2. Optional: If you are not the InfoSphere BigInsights administrator user, run the following command to log on as `BIGINSIGHTS_USER`. This user role gives you permissions to run the `uninstall.sh` script.
3. If you installed GPFS as your file system, stop GPFSand then remove it.
   a. Stop any processes that use GPFS, including Linux shells that have a current working directory that points to GPFS.
   b. If you installed GPFS as your file system, from the `$BIGINSIGHTS_HOME/hdm/bin` directory, run the `gpfs-delete_clusterfs.sh` script, which calls GPFS remove commands.

      Attention: Removing the file system results in loss of data and cannot be reversed. If you want to retain data, ensure that you back up your file system before running the cluster removal script.

      The following example shows the syntax of the `gpfs-delete_clusterfs.sh` command.

      ```
      installer/hdm/bin/gpfs-delete_clusterfs.sh
                  [ -f param-file ]
                  [ -u {1 | 2} ]
                  [ -h ]
      ```

      param_file is the path name of the file that contains the parameters that were used to create your GPFS cluster. For example, `bi_gpfs.cfg`.

      - To delete the underlying cluster and file system, and remove GPFS, specify `-u 2`.
      - To remove GPFS only, specify `-u 1`.

      If you do not specify a value for the `-u` option, then the underlying cluster and file system are deleted.
4. From the `$BIGINSIGHTS_HOME/bin` directory, run the **`uninstall.sh`** command to begin the software removal process.

Option	Description
Attended removal	1. Run the `./uninstall.sh` command. 2. When prompted by the shell console, enter `yes` to start removing software from your system. 3. When asked whether you want to continue, enter `yes`.
Unattended removal	Run the `./uninstall.sh` command with the `--silent` parameter. `./uninstall.sh --silent` This command removes InfoSphere BigInsights and deletes all files from the `$BIGINSIGHTS_HOME` directory, `$BIGINSIGHTS_VAR` directory, and all other Hadoop data directories.

Removing InfoSphere BigInsights manually

If a previous installation or removal failed, you might be required to manually remove InfoSphere BigInsights from the nodes in your cluster. You remove the software manually only after you try to use the removal scripts that are provided with the software.

Procedure

1. Stop InfoSphere BigInsights processes.
   a. Run the following command as the InfoSphere BigInsights administrator user from the `$BIGINSIGHTS_HOME/bin` directory on the management node.
      `./stop-all.sh`
   b. Run the following command from the `$BIGINSIGHTS_HOME/bin` directory to stop any component processes that are still running.
      `./stop` *`component-name`*

      `component-name` is the name of the component that you want to stop.
   c. Run the following command on each node in the cluster to locate any InfoSphere BigInsights that might still be running. Manually stop any processes that the command locates.
      `ps aux | grep $BIGINSIGHTS_HOME`
2. Remove the InfoSphere BigInsights related environment values from the `~/.bashrc` file on each of the cluster nodes.
3. Remove the line `source $BIGINSIGHTS_HOME/conf/biginsights-env.sh` from the `~/.bashrc` file.
4. Run the following command to remove remaining files from the `$BIGINSIGHTS_HOME` and `$BIGINSIGHTS_VAR` directories.
   `rm -rf $BIGINSIGHTS_HOME $BIGINSIGHTS_VAR`
5. Optional: Remove the remaining directories that exist on each of the nodes in the cluster.

   Tip: These directories might have been customized during the installation process, so you might want to temporarily save the `$BIGINSIGHTS_HOME/conf/install.xml` file (on the management/console node) until you complete the manual uninstall in case you need to go back to the customized directories.

/hadoop/hdfs/name
Contains Hadoop Distributed File System (HDFS) data that is maintained by the NameNode.

Important: You might want to keep this directory intact because the NameNode server contains information about all data files in HDFS. Removing this directory results in permanent data loss in your cluster.

/hadoop/hdfs/namesecondary
Contains HDFS data that is maintained by the Secondary NameNode.

Important: You might want to keep this directory intact because the Secondary NameNode server contains checkpoint data for HDFS. Removing this directory results in permanent data loss in your cluster.

/hadoop/hdfs/data
Contains HDFS data that is maintained by each DataNode in your cluster.

Important: You might want to keep this directory intact because the DataNode servers contain information about the storage that is attached to each nodes in the cluster. Removing this directory results in permanent data loss in your cluster.

/hadoop/mapred/local
Contains MapReduce cache data. The cache directory is the local file system path on which each node writes temporary MapReduce data.

/hadoop/mapred/system
Contains MapReduce system data. The MapReduce system directory is the HDFS path on which the MapReduce framework stores system files.

/tmp/biginsights
Contains temporary data for InfoSphere BigInsights.

Removing the InfoSphere BigInsights Tools for Eclipse

The procedure to uninstall the InfoSphere BigInsights Tools for Eclipse is the same, regardless of how InfoSphere BigInsights is configured.

Procedure

1. Remove the InfoSphere BigInsights software from your Eclipse environment.

   Important: If you installed IBM Data Studio or IBM Rational products into a virtualized directory like the `Program Files` directory, you must run IBM Data Studio or IBM Rational products as an administrator user to remove the InfoSphere BigInsights Tools for Eclipse.

   a. Click **Help** > **About Eclipse SDK**.
   b. In the About window, click **Installation Details** to show the software that is installed in your Eclipse environment.
   c. In the Installation Details window, select **IBM InfoSphere BigInsights** from the list of installed software and click **Uninstall**.
   d. Follow the steps in the software removal wizard to complete the software removal process.
   e. Restart your Eclipse environment when prompted.

2. After you restart your Eclipse environment, if you see the InfoSphere BigInsights perspective or the InfoSphere BigInsights Text Analytics Workflow perspective, remove the perspectives manually from your Eclipse environment.
   a. Click **Window** > **Preferences**.
   b. In the Preferences window, expand the **General** section and click **Perspectives**.
   c. In the list of available perspectives, select and delete any entries that include `<BigInsights>` or `<BigInsights Text Analytics Workflow>`.

      Do not remove any entries that are not contained by angle brackets (< >). If these perspectives exist, then the product is still installed. Complete the procedure in step 1 on page 73 to remove the InfoSphere BigInsights software from your Eclipse environment, and then remove the perspectives.
   d. Click **Apply** and then **OK** to apply your changes and close the Preferences window.

Chapter 7. Installation problems and workarounds

Common problems and questions that are related to the installation are identified.

Installation program hangs and progress does not update

The InfoSphere BigInsights installation program hangs and the progress of the installation does not update.

Symptoms

The installation program runs for more than 15 minutes without updating the progress of the installation, or the installation progress does not proceed further than 0% after more than one minute.

Causes

WebSphere Application Server Community Edition is not functioning properly.

Diagnosing the problem

Run the **ps aus | grep install.sh** command, and then look for a response similar to the following example.

```
root  17983  0.0  0.0 106192  1416 ?
S    07:04   0:00 bash /opt/ibm/builds/biginsights-enterprise-linux64_b20130408_0527/
                  installer/bin/install.sh fullinstall.xml
```

Resolving the problem

Use one of the following methods to install the product.

Installation method	Instructions
Silent	Run a silent installation by using the `.xml` file that exists in the directory where you extracted the `.tar` file for the installation program. • If you are running a new installation, use the `fullinstall.xml` file. • If you are upgrading to a new version, use the `upgrade.xml` file. For instructions on running a silent installation, see Installing by using a response file.

Installation method	Instructions
Installation program	1. Run the **/start.sh shutdown** command to stop the installation program. 2. Extract the `.tar` file for the installation program to a new directory. 3. Copy the response file from the directory where you first ran the installation program to the directory that you created. • If you are running a new installation, use the `fullinstall.xml` file. • If you are upgrading to a new version, use the `upgrade.xml` file. 4. Run the **/start.sh** command from the directory that you created to start the installation program.

Cannot install the Linux Expect package

The InfoSphere BigInsights installation program cannot install the Linux Expect package.

Symptoms

An error message indicates that the Linux Expect package cannot be installed. The following example shows what this message looks like:

```
[ERROR] Cannot find /user/bin/expect, please install Expect and run the installer again.
```

Causes

The InfoSphere BigInsights installation program installs the Linux Expect package by default. However, the installation program might not be able to install the Linux Expect package because the operating system of the computer is not compatible.

Resolving the problem

You must manually install the Linux Expect package. You must have `root` access to install this package.

1. Install the Linux Expect package.

Option	Procedure
If the **yum** facility is supported on your system	Run the **yum** command to install the package: `yum install expect`

Option	Procedure
If the **yum** facility is not supported on your system	1. Download the package for your operating system from the Linux Expect website. 2. Save the package that you downloaded in the *install_dir*/artifacts directory. *install_dir* is the directory where you extracted the contents of the installation package. 3. Run the **rpm** command to install the package: # rpm -i */path_name/file_name*.rpm path_name is the full path name of the directory to the Linux Expect package that you want to install. file_name is the file name of the package that you want to install.

2. Follow the instructions for installing InfoSphere BigInsights.
   - Installing by using the installation wizard
   - Installing by using a response file

Installing optional components

You cannot install more components after the initial installation of the product.

Symptoms

You want to install more optional components to an existing InfoSphere BigInsights cluster.

Causes

InfoSphere BigInsights does not support incremental installation to add optional components to an existing cluster environment.

Resolving the problem

Perform a new installation to add more components, such as the Pig components, by using the installer.

CAUTION: This process overwrites your existing installation, so ensure that you install all necessary components and back up your data and directories before you begin the new installation.

Incorrect hostname information for monitoring adaptor

The monitoring adaptor for Hadoop DataNode is added without the correct hostname information.

Symptoms

When you install a single node cluster, by using all default installation options, the monitoring adaptor for Hadoop DataNode gets added without the correct hostname information.

Resolving the problem

1. Using Monitoring REST API, start monitoring from the Enterprise Console.

2. From a web browser, open the url `http://host:monitoring_agent_port/rest/v1/adaptor`.
3. Examine the XML file that is returned from the URL for the adaptor ID of the incorrect adaptor. The `monitoring_agent_port` is what the user defines on the Installer UI. It is 9090 by default.
4. By using a REST client, issue the command **HTTP DELETE** against `http://host:monitoring_agent_port/rest/v1/adaptor/adaptor_id`, by using the `adaptor_id` from step c.
5. By using a REST client, issue the command **HTTP POST** against `http://host:monitoring_agent_port/rest/v1/adaptor`. Set the request header to have Content-Type as `application/json;charset=UTF-8`. Set the request body as

```
{Offset=0,
  AdaptorClass=
   org.apache.hadoop.chukwa.datacollection.adaptor.JMXAdaptor,
  DataType=
   DatanodeProcessor, AdaptorParams=<host_name> 8007 60 Hadoop:*}
```

<host_name> is the actual hostname of the single node cluster.

Installation failure due to insufficient prerequisites

Installing InfoSphere BigInsightsV1.3.0.1 without the correct prerequisites results in an installation failure.

Symptoms

When you install InfoSphere BigInsights V1.3.0.1, you might receive a message similar to the following message:

```
IUI0005E:
  The installation failed with fatal error:
  System pre-check fails, some prerequisite is not fulfilled.
  Check log for detail.
```

Causes

Check the log for messages similar to the following message:

```
[WARN] ###
  /hadoop/hdfs/namesecondary ### 3 files/dirs exist
[ERROR]
  7 invalid directories are found.
```

This line indicates that the V1.3.0.1 installer detected files in the specified installation directories.

Environment

Optional. Describe any environmental details that are not already in the title or short description.

Diagnosing the problem

Optional. Clearly state the steps necessary to diagnose the problem. Optionally, include appropriate response role elements. Alternatively, imbed a task topic or conref that provides the steps to diagnose the problem.

User response: Optional. When you have particular actions that are performed by particular users, use one or more of the ts*Response elements.

Resolving the problem

You can do one of the following actions:

User response:

- Select the check box next to **Write over existing files and directories** on the File System page of the Installation UI.
- Move the existing files to another location.
- Change the installation directory.

Hadoop data nodes are in uncertain status

Hadoop data nodes are in an uncertain status, and multiple components fail the health check.

Symptoms

Hadoop data nodes are in uncertain status, and multiple components fail the health check.

Causes

The hostname that is specified for the Hadoop name node is listening on the wrong IP address.

Resolving the problem

Check `/usr/sbin/lsof -i :9000`, where `9000` is the name node RPC port:

```
java 23100 biadmin 138u IPv4 41853749 0t0
  TCP 127.0.0.1:cslistener (LISTEN)
```

When 127.0.0.1 cannot be accessed by remote nodes, then none of the Hadoop data nodes can connect to the name node. Typically, this issue is caused by mapping the hostname to the local loop address 127.0.0.1 or ::1 in `/ect/hosts`, like

```
127.0.0.1 svltest101.svl.ibm.com svltest101
::1 svltest101.svl.ibm.com svltest101
```

Commenting out such lines can resolve this issue. You do not need to reinstall InfoSphere BigInsights.

Local names do not match the managed nodes

Local names and managed nodes do not match.

Symptoms

You receive a message that none of the local names match the managed nodes.

Resolving the problem

Reconfigure `/etc/hosts`, ensuring that the hostname and IP address specified for the InfoSphere BigInsights Console node can be resolved by either:

```
hostname -s
hostname -a
hostname -f
valid inet addr defined in /sbin/ifconfig
```

NameNode in safe mode causes errors

A fatal error, IUI0005E, occurs when NameNodes are in safe mode.

Symptoms

You see message IUI0005E, that the installation failed with a fatal error:

```
Failed to install BigInsights component(s).
Installation failed on component(s): Flume
Verification failed on component(s): Jaql, Hive, Oozie, BigInsights orchestrator, Jaql UDF server

org.apache.hadoop.ipc.RemoteException: org.apache.hadoop.hdfs.server.namenode.SafeModeException:
Cannot create file/user/biadmin/jaqltest_bdvm039.svl.ibm.com.dat. NameNode is in safe mode.
Resources are low on NN. Safe mode must be turned off manually.
```

Causes

Before you install InfoSphere BigInsights, Hadoop daemons must be started, and the Hadoop NameNode must not be in safe mode.

Resolving the problem

Check the NameNode log to determine which of these issues caused the failure, and resolve that issue or issues.

Incorrect HBase Sudo policy

The defined policy for the sudo privileges of a user is not correct.

Symptoms

You see the following error message:

```
[ERROR]
DeployManager - hbase failed root@svltest367.svl.ibm.com s password:
sudo: sorry, you must have a tty to run sudo
>java.io.IOException: exit code: 1 -- /opt/ibm/biginsights/hdm/bin/_ssh-remote.exp
{xor}Dz4sLChvLTs= ssh -o NumberOfPasswordPrompts=1 root@svltest367.svl.ibm.com sudo -u hbase
/usr/lib/hbase/get_endpoints.sh /usr/java/default /usr/lib/hbase /usr/lib/hadoop
root@svltest367.svl.ibm.com s password:
sudo: sorry, you must have a tty to run sudo
at com.ibm.xap.mgmt.util.ExecUtil.exec(ExecUtil.java:81)
at com.ibm.xap.mgmt.util.ExecUtil.exec(ExecUtil.java:28)
```

Causes

The `#Defaults requiretty` line must be uncommented.

Resolving the problem

Edit `/etc/sudoers`, comment out the `#Defaults requiretty` line by removing the `#` symbol.

Administrative user is not listed in `AllowUsers` property

You might receive a failure in the check of **`passwordless`** SSH.

Symptoms

When you install InfoSphere BigInsights V1.4, you might receive a message that indicates that the check of passwordless SSH setup failed.

Causes

Check the log for message similar to following message:

```
>java.io.IOException:
  exit code: 1 --
  sudo -S -p SUDOPWD: scp -r -o StrictHostKeyChecking=no -o
BatchMode=yes
  /home/installuser1/.ssh/id_rsa.pub
  /opt/ibm/biginsights/hdm/bin/_root-setup-biadmin-remote.sh
    bdvm070.svl.ibm.com:/tmp
Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).
lost connection
```

Resolving the problem

Ensure that the InfoSphere BigInsights administrative user is listed in **AllowUsers** property in `/etc/ssh/sshd_config` file. If you modify this file, restart the sshd service with the **service sshd** restart command.

Disk discovery fails due to node passwordless SSH errors

When installing InfoSphere BigInsights with GPFS, the InfoSphere BigInsights installation program fails on disk discovery.

Symptoms

The installation program fails with the following error message.

```
The disk list is unavailable.  Please make sure that the current user root has
already been configured with sudo privileges and root has already been configured
for passwordless SSH access.
```

Causes

This error might occur during an InfoSphere BigInsights installation for the following reasons:

1. Passwordless SSH is not configured.
2. Nodes are specified without editing the nodes to provide passwords.

Resolving the problem

In the GPFS, the InfoSphere BigInsights installation program, return to the Nodes panel and edit the main node to add the root password.

HBase status shows as “Unavailable” during installation

During an InfoSphere BigInsights installation, the InfoSphere BigInsights installation program shows HBase as unavailable when HBase is running correctly.

Symptoms

HBase shows as unavailable when installing InfoSphere BigInsights, though HBase is already running.

Causes

This error can occur when HBase (either the client or server) and Zookeeper are not both installed on the same node as the web console node.

Resolving the problem

Ensure that HBase (either client or server) and Zookeeper are both available on the console node. To correct the problem after installation, click **Add Node** from the Cluster Status panel in the InfoSphere BigInsights Console to ensure that HBase (either client or server) and Zookeeper are both available on the same node as the console node.

A previous GPFS installation failed

If a previous attempt to install GPFS failed, reinstalling GPFS can result in error messages.

Symptoms

A message displays that GPFS is already partially installed.

```
Detected either an existing or partial installation of GPFS on:
node1 node2 node3 node4 node5
Aborting the attempt to install GPFS.
[FATAL] Failed to setup a GPFS file system
[FATAL] Failed to install GPFS
```

Causes

The previous installation of GPFS was not removed completely.

Resolving the problem

1. Uninstall the package by using the `gpfs-delete_clusterfs.sh` script. The following example shows how to run this script.

   ```
   sudo installer/hdm/bin/gpfs-delete_clusterfs.sh -f bi_gpfs.cfg -u 2
   ```

2. Delete the `/var/mmfs` directory and the `/tmp/mmfs` directory if they still exist.
3. Reattempt the installation.

Linux Standard Base package is not installed

If the Linux Standard Base (LSB) package is not installed, the GPFS autoload feature can sometimes fail.

Symptoms

The following message displays when installing GPFS.

```
/var/tmp/rpm-tmp.GFyDsp: line 164: /usr/lib/lsb/install_initd: No such file or directory
```

Causes

The LSB package for Linux x86 64-bit is not installed.

Resolving the problem

Before installing GPFS, ensure that the `lsb.x86_64` package is installed. The exact version depends on the kernel and package levels on your specific Linux instance. You must determine the x86-64 version of the package for your Linux system. For example, the package for Linux RedHat 6.2 is `redhat-lsb-4.0-3.el6.x86_64.rpm`.

Linux system does not have prerequisite kernel or C++ packages

If your Linux system does not have all of the prerequisite kernel or C++ packages, errors can occur when installing InfoSphere BigInsights with GPFS.

Symptoms

Errors can occur that vary depending on the package that is missing. The following examples show some example errors.

If the `kernel-headers` or `kernel-devel` are missing, the following error might display.

```
 + ssh root@<hostname> 'cd /usr/lpp/mmfs/src; make CleanConfig; make Autoconfig;
   make World; make InstallImages'
rm -f config/def.mk config/env.mcr
cd /usr/lpp/mmfs/src/config; ./configure --genenvonly; if [ $? -eq 0 ];
   then /usr/bin/cpp -P def.mk.proto > ./def.mk; exit $? || exit 1; else exit $?; fi
Cannot find a valid kernel include dir
make: *** [Autoconfig] Error 1
```

If the `gcc-c++` package or other dependent packages are missing, the following error might display.

```
+ ssh root@<hostname> 'cd /usr/lpp/mmfs/src; make CleanConfig; make Autoconfig; make World;
   make InstallImages'
rm -f config/def.mk config/env.mcr
cd /usr/lpp/mmfs/src/config; ./configure --genenvonly; if [ $? -eq 0 ]; then /usr/bin/cpp
   -P def.mk.proto > ./def.mk; exit $? || exit 1; else exit $?; fi
Verifying that tools to build the portability layer exist....
cpp present
gcc missing! Verify that it is installed.
make: *** [VerifyBuildTools] Error 1
(cd gpl-linux; /usr/bin/make InstallImages; \
exit $?) || exit 1
make[1]: Entering directory `/usr/lpp/mmfs/src/gpl-linux'
Pre-kbuild step 1...
make[2]: gcc: Command not found
make[2]: Entering directory `/usr/src/kernels/2.6.18-128.el5-x86_64'
DEPMOD 2.6.18-128.el5
make[2]: Leaving directory `/usr/src/kernels/2.6.18-128.el5-x86_64'
cat: //usr/lpp/mmfs/src/gpl-linux/gpl_kernel.tmp.ver: No such file or directory
cat: //usr/lpp/mmfs/src/gpl-linux/gpl_kernel.tmp.ver: No such file or directory
cat: //usr/lpp/mmfs/src/gpl-linux/gpl_kernel.tmp.ver: No such file or directory
/usr/bin/install: cannot stat `lxtrace': No such file or directory
```

Resolving the problem

Ensure that the following packages are installed or available. The versions depend on the exact kernel and package levels on your specific Linux instance. You can install the packages by using the `rpm -ivh` command.

```
kernel-headers-2.6.18-128.el5.x86_64.rpm
kernel-devel-2.6.18-128.el5.x86_64.rpm
libgomp-4.3.2-7.el5.x86_64.rpm
glibc-headers-2.5-34.x86_64.rpm
glibc-devel-2.5-34.x86_64.rpm
gcc-4.1.2-44.el5.x86_64.rpm
libstdc++-devel-4.1.2-44.el5.x86_64.rpm
gcc-c++-4.1.2-44.el5.x86_64.rpm
```

Stale mounts cause installation errors

A stale Network File System (NFS) file handle can result in the old mount being detected as an existing, but not usable, file system.

Symptoms

You might not be able to install GPFS because stale mounts are detected.

Causes

A stale mount might occur if an attempt was made to remove GPFS while processes are still accessing the GPFS file system.

Diagnosing the problem

To check for this condition, run the following command. The stale mount is indicated on `/mnt/bigpfs`.

```
$ df -T
Filesystem Type 1K-blocks Used Available Use% Mounted on
/dev/sda3 ext3 93347880 84713772 3815816 96% /
/dev/sda1 ext3 101086 11846 84021 13% /boot tmpfs
tmpfs 8188960 0 8188960 0% /dev/shm
df: `/mnt/bigpfs': Stale NFS file handle
```

Resolving the problem

Run the following command to unmount the stale mount.

```
$ sudo /bin/umount /mnt/bigpfs
```

Unable to load one or more GPFS kernel extensions

If a previous version of GPFS is not completely removed, installing GPFS with InfoSphere BigInsights can cause errors.

Symptoms

The following message displays when attempting to install InfoSphere BigInsights.

```
1692 ? S<l 0:00 /usr/lpp/mmfs/bin/lxtrace-2.6.18-194.el5 on /log/mmfs
mmfsenv: The /lib/modules/2.6.18-194.el5/extra/mmfslinux.ko kernel extension does not exist.
mmfsenv: Unable to verify kernel/module configuration.
Unable to unload one or more GPFS kernel extensions.
You may need to reboot the node.
```

Causes

InfoSphere BigInsights does not support a GPFS installation over another GPFS installation.

Diagnosing the problem

If you see this message, run the **ps** command to find information about process 1692, and then stop it. For example:

```
$ ps -ef | grep 1692
root 1692 1 0 Jul13 ? 00:00:00 /usr/lpp/mmfs/bin/lxtrace-2.6.18-194.el5 on
   /log/mmfs/lxtrace.trc.hdtest030
   -s 400000000 -r -b 1310720
$ sudo kill -9 1692

$ rpm -qa | grep gpfs
gpfs.base-3.4.0-0
gpfs.gpl-3.4.0-7
gpfs.msg.en_US-3.4.0-7
gpfs.docs-3.4.0-7

$ sudo rpm -e 'rpm -qa | grep gpfs'
rpm: no packages given for erase
```

Resolving the problem

Stop this process on every node where the error occurs. If the error persists, reboot the node.

Installing GPFS by using the `mmcrfs` command fails

When installing GPFS by using the **mmcrfs** command, the generated configuration attempts to set the default metadata replication and data replication to 3. This setting can result in errors when attempting to install GPFS on clusters with an insufficient number of GPFS failure groups.

Symptoms

The **mmcrfs** command fails with an error that is similar to the following example.

```
The following disks of bigpfs will be formatted on node node1.xxx.yyy.com:
    gpfs1nsd: size 976562483 KB
    gpfs2nsd: size 976562500 KB
    gpfs3nsd: size 976562500 KB
Formatting file system ...
Disks up to size 8.1 TB can be added to storage pool system.
Disks up to size 8.0 TB can be added to storage pool datapool.
Incompatible parameters: Unable to create file system.
Change one or more of the following as suggested and try again:
    increase the number of failure groups
```

```
    decrease the value for -r
mmcrfs: tscrfs failed.  Cannot create bigpfs
mmcrfs: Command failed.  Examine previous error messages to determine cause.
```

Causes

This issue might be an indirect result of the number of nodes and disks that you specified for the cluster.

Resolving the problem

Edit the `fullinstall.xml` file or the `bi_gpfs.cfg` configuration file to decrease the values that correspond to the variables that are used for the `-r` and `-m` parameters.

Edit the `fullinstall.xml` file

1. In the directory where you extracted the `biginsights-enterprise-linux64_`*`release_number`*`.tar.gz` file, open the `fullinstall.xml` file.

   release_number is the release number that you are installing.
2. Decrease the value of the `default-data-replication` and `default-metadata-replication` parameters. For example, if the value of these parameters is 3, change the value to 1.

   ```
   <default-data-replication>1</default-data-replication>
   <default-metadata-replication>1</default-metadata-replication>
   ```
3. Run the InfoSphere BigInsights installation program in silent mode to install GPFS.

Edit the `bi_gpfs.cfg` configuration file

1. In the directory where you extracted the `biginsights-enterprise-linux64_`*`release_number`*`.tar.gz` file, open the `bi_gpfs.cfg` configuration file.

   release_number is the release number that you are installing.
2. Decrease the value of the `DEFAULT_DATA_REPLICATION` and `DEFAULT_METADATA_REPLICATION` parameters. For example, if the value of these parameters is 3, change the value to 1.

   ```
   DEFAULT_DATA_REPLICATION=1
   DEFAULT_METADATA_REPLICATION=1
   ```
3. Run the **`gpfs-create_clusterfs.sh`** script and specify the `bi_gpfs.cfg` configuration file as a parameter to install GPFS.

   ```
   gpfs-create_clusterfs.sh -f bi_gpfs.cfg -n nodefile -d diskfile
   ```

   `nodefile` is the name of the file that indicates how nodes are treated in GPFS.

   `diskfile` is the name of the file that indicates the disks to use.

Cluster status displays as “Running”, even when the file system is down

After you successfully install GPFS, the status for **General Parallel File System** should display as Running in the Cluster Status panel of the InfoSphere BigInsights Console.

Symptoms

The status for GPFS displays as Running, even though the file system is down.

Causes

GPFS might have been shut down, and did not restart correctly.

Diagnosing the problem

Use the **mmgetstate** command to check the status of GPFS outside of the InfoSphere BigInsights Console. For more information, see the Cluster Products Information Center.

Resolving the problem

This is a known problem with no current workaround.

Applications hang when running as a non-administrator user

Applications can hang when running from the InfoSphere BigInsights Console as a non-administrative user.

Symptoms

You run an application from the InfoSphere BigInsights Console, but the application hangs.

Causes

This problem typically occurs if you installed InfoSphere BigInsights with LDAP security and GPFS.

Resolving the problem

Add lines to the `core-site.xml` file to impersonate an administrator user.

1. Change permissions on the following directory.

   ```
   chmod 1777 GPFS_mount_point/hadoop/mapred/local/* GPFS_mount_point/user
   ```

   GPFS_mount_point is the GPFS mount point that you specified during installation.

2. From the `/opt/ibm/biginsights/bin` directory, stop Hadoop.

   ```
   ./stop.sh hadoop
   ```

3. Add the following lines to the `core-site.xml` file in the `/opt/ibm/biginsights/hdm/hadoop-conf-staging` directory.

   ```
   <property>
     <name>hadoop.proxyuser.oozie.hosts</name>
     <value>*</value>
   </property>

   <property>
     <name>hadoop.proxyuser.oozie.groups</name>
     <value>*</value>
   </property>
   ```

4. From the `/opt/ibm/biginsights/bin` directory, sync your changes with Hadoop.

   ```
   ./syncconf.sh hadoop
   ```

5. Start Hadoop.

   ```
   ./start.sh hadoop
   ```

6. From the InfoSphere BigInsights Console, run your application.

Users cannot log in to the InfoSphere BigInsights Console when using LDAP authentication

If you do not provide the LDAP user names when installing the product, users might not be able to log in to the InfoSphere BigInsights Console.

Symptoms

Users cannot log into the InfoSphere BigInsights Console.

Causes

Users were not added to your LDAP configuration when installing InfoSphere BigInsights.

Resolving the problem

You must add the users to your LDAP configuration. For more information, see *Adding users* in the InfoSphere BigInsights Information Center.

Notices and trademarks

This information was developed for products and services offered in the U.S.A.

Notices

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte character set (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web

sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
J46A/G4
555 Bailey Avenue
San Jose, CA 95141-1003 U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to

IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. _enter the year or years_. All rights reserved.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at www.ibm.com/legal/copytrade.shtml.

The following terms are trademarks or registered trademarks of other companies:

Adobe is a registered trademark of Adobe Systems Incorporated in the United States, and/or other countries.

Intel and Itanium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows and Windows NT are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

The United States Postal Service owns the following trademarks: CASS, CASS Certified, DPV, LACS^{Link}, ZIP, ZIP + 4, ZIP Code, Post Office, Postal Service, USPS and United States Postal Service. IBM Corporation is a non-exclusive DPV and LACS^{Link} licensee of the United States Postal Service.

Other company, product or service names may be trademarks or service marks of others.

Providing comments on the documentation

You can provide comments to IBM about this information or other documentation.

About this task

Your feedback helps IBM to provide quality information. You can use any of the following methods to provide comments:

Procedure

- Send your comments by using the online readers' comment form at www.ibm.com/software/awdtools/rcf/.
- Send your comments by e-mail to comments@us.ibm.com. Include the name of the product, the version number of the product, and the name and part number of the information (if applicable). If you are commenting on specific text, include the location of the text (for example, a title, a table number, or a page number).

Printed in USA

GC19-4100-00