
IBM's InfoSphere BigInsights: Smart Analytics for Big Data



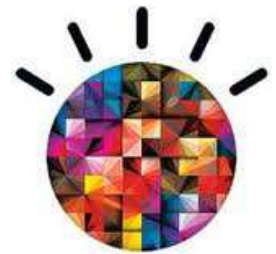
Burak İLTER

IBM Disclaimer

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Agenda

- **The “Big Data” challenge: smarter analytics for a smarter planet**
- **IBM’s approach**
 - The big picture
 - Details on BigInsights
 - How BigInsights fits in your software stack
- **How IBM can help you get off to a quick start**

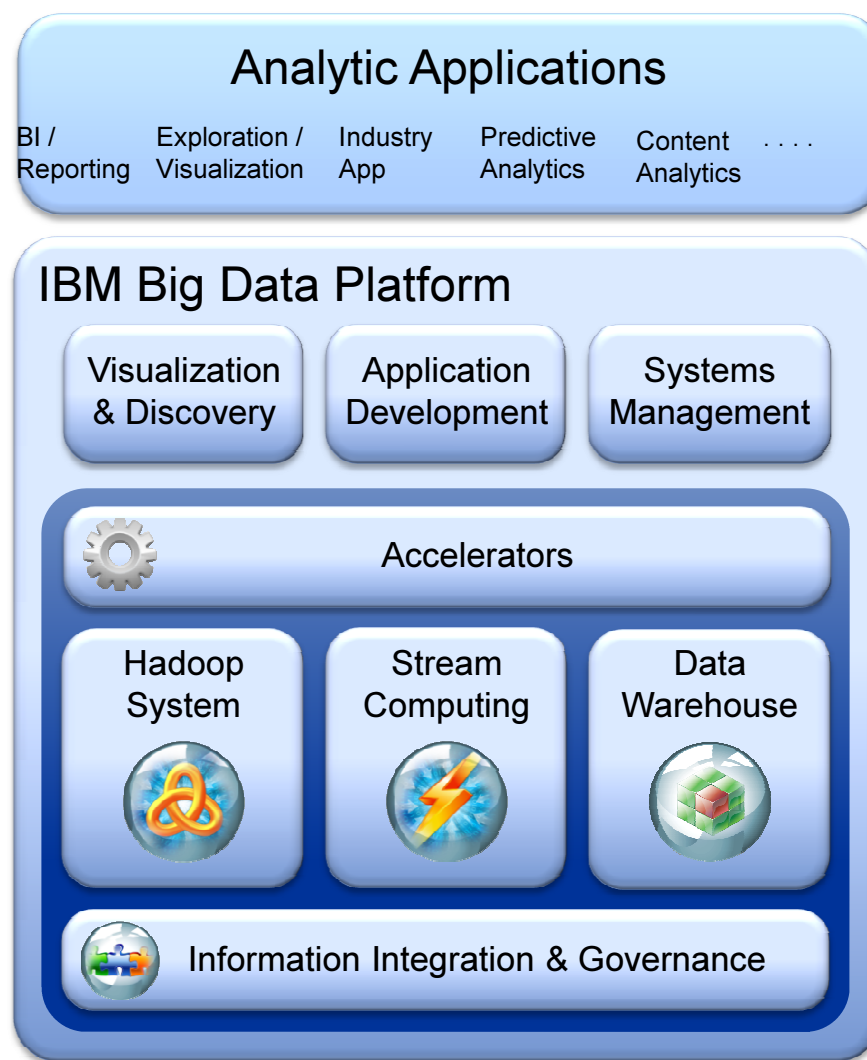


IBM's approach



IBM Big Data Platform Strategy

- Integrate and manage the full variety, velocity and volume of Big Data
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis
- Development environment for building new analytic applications
- Support workload optimization and scheduling
- Provide for security and governance
- Integrate with enterprise software



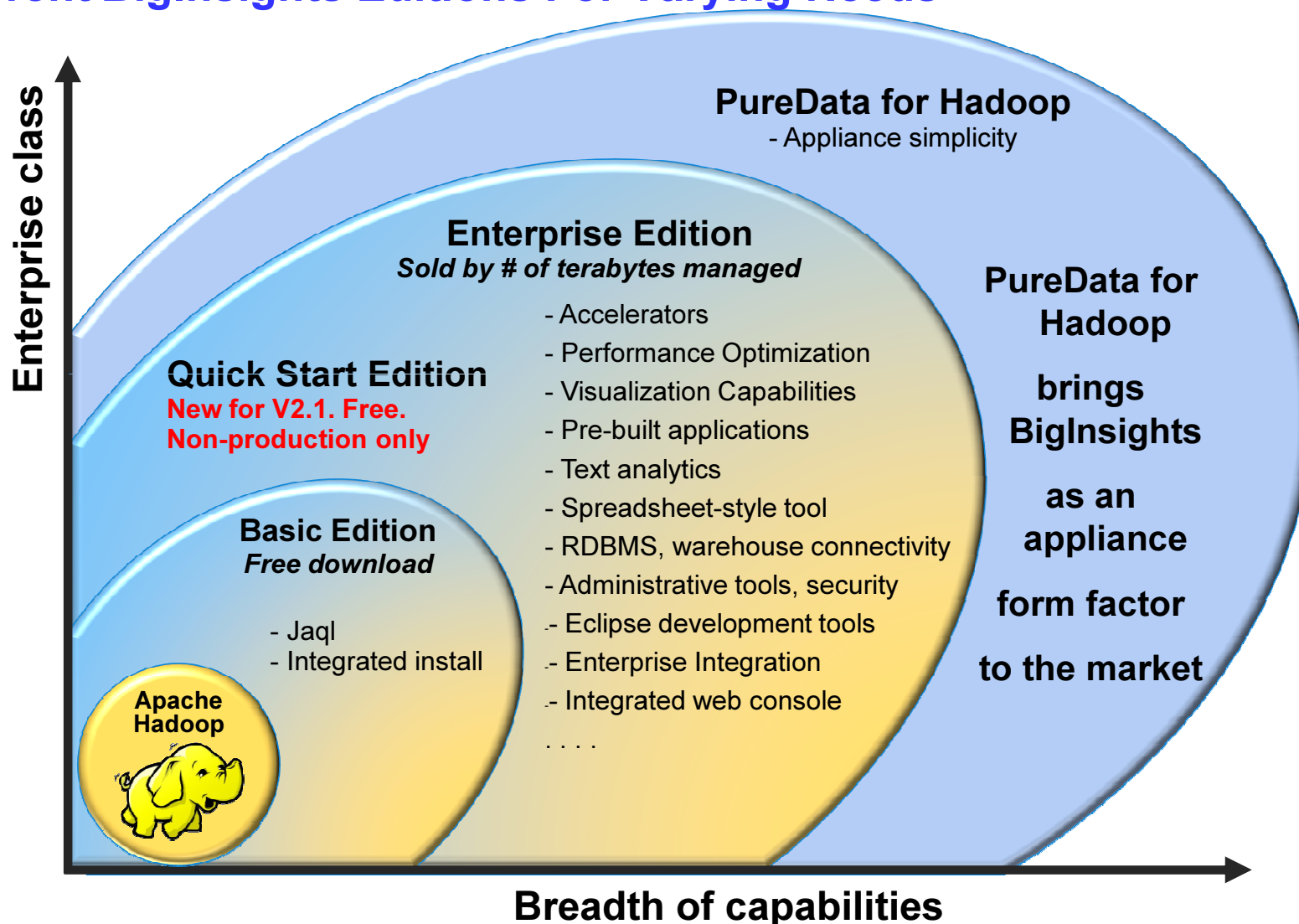
BigInsights Brings Hadoop to the Enterprise



- **BigInsights = analytical platform for persistent Big Data**
 - Based on open source & IBM technologies
 - Managed like a start-up Emphasis on deep customer engagements, product plan flexibility
- **Distinguishing characteristics**
 - Built-in analytics *Enhances business knowledge*
 - Enterprise software integration *Complements and extends existing capabilities*
 - Production-ready platform with tooling for analysts, developers, and administrators. . . . *Speeds time-to-value; simplifies development and maintenance*
- **IBM advantage**
 - Combination of software, hardware, services and advanced research



From Getting Starting to Enterprise Deployment: Different BigInsights Editions For Varying Needs



BigInsights Content

Function	Version	Basic Edition	Enterprise Edition
Integrated Install		Inc	Inc
Hadoop (including common utilities, HDFS, MapReduce framework)	1.1.1	Inc	Inc
Pig (programming / query language)	0.10.0	Inc	Inc
Flume (data collection/aggregation)	1.3.0	Inc	Inc
Hive (data summarization/querying)	0.9.0	Inc	Inc
Lucene (text search)	3.3.0	Inc	Inc
Zookeeper (process coordination)	3.4.5	Inc	Inc
Avro (data serialization)	1.7.2	Inc	Inc
HBase (real time read/write)	0.94.3	Inc	Inc
HCatalog (table and storage management service)	0.4.0	Inc	Inc
Sqoop (RDBMS bulk data transfer)	1.4.1	Inc	Inc
Oozie (workflow/ job orchestration)	3.2.0	Inc	Inc
Jaql (query and scripting language)		Inc	Inc
Online documentation		Inc	Inc
Integration with JDBC sources through general-purpose Jaql module		Inc	Inc
Integration with DB2 (sample functions to submit jobs, read data)		Inc	Inc

BigInsights Content (cont'd)

Function	Basic Edition	Enterprise Edition
Big SQL (standard SQL query support, JDBC/ODBC drivers, LOAD from DB2, Netezza, Teradata)	n/a	Inc
Integration with Netezza, DB2 LUW with DPF from Jaql. Integration with R (Jaql module to invoke R statistical capabilities from BigInsights)	n/a	Inc
LDAP authentication, Guardium support, etc.	n/a	Inc
Integrated Web Console	n/a	Inc
Business process accelerators (social data, machine data analytics)	n/a	Inc
Platform enhancements (GPFS-FPO, Adaptive MapReduce, efficient processing of compressed text files, flexible job scheduler, high availability, etc.)	n/a	Inc
Text analytics	n/a	Inc
Eclipse tools for text analytic development, Jaql, Hive, Java	n/a	Inc
Applications for data import/export, Web crawl, machine learning, etc.	n/a	Inc
Web-based application catalog	n/a	Inc
Spreadsheet-like analytical tool	n/a	Inc
IBM support	Opt	Inc
Streams, Data Explorer, Cognos BI (limited use licenses)	n/a	Inc
Unlimited storage	n/a	Inc

BigInsights: Value Beyond Open Source



Enterprise Capabilities

Visualization & Exploration

Development Tools

Advanced Engines

Connectors

Workload Optimization

Administration & Security

Open source
components

IBM-certified
Apache Hadoop
and related projects

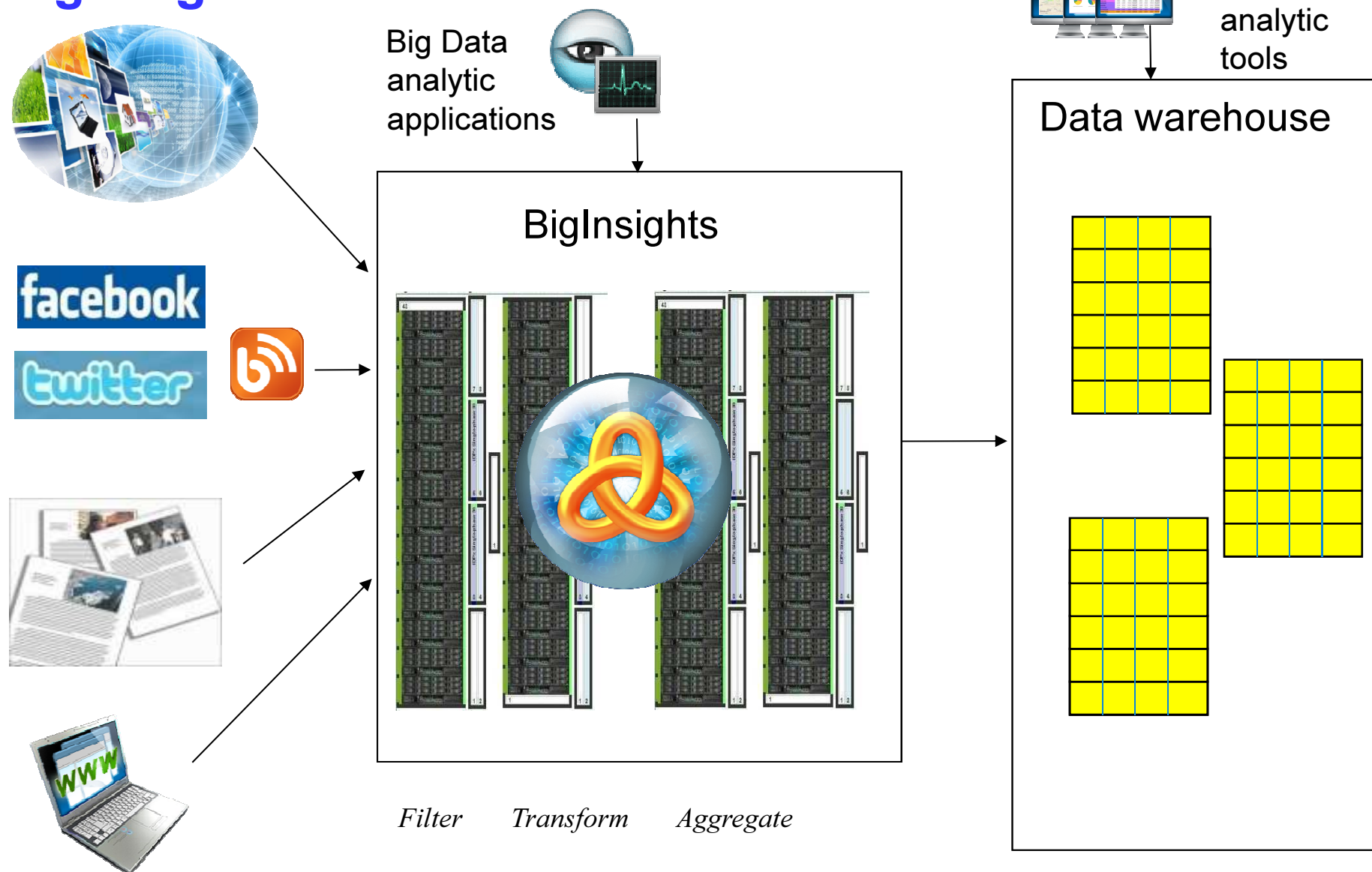
Key differentiators

- Built-in text analytics
- Enterprise software integration
- SQL support
- Spreadsheet-style analysis
- Integrated installation of supported open source and other components
- Web Console for admin and application access
- Platform enrichment: additional security, performance features, GPFS (alternative file system), . . .
- World-class support
- Full open source compatibility

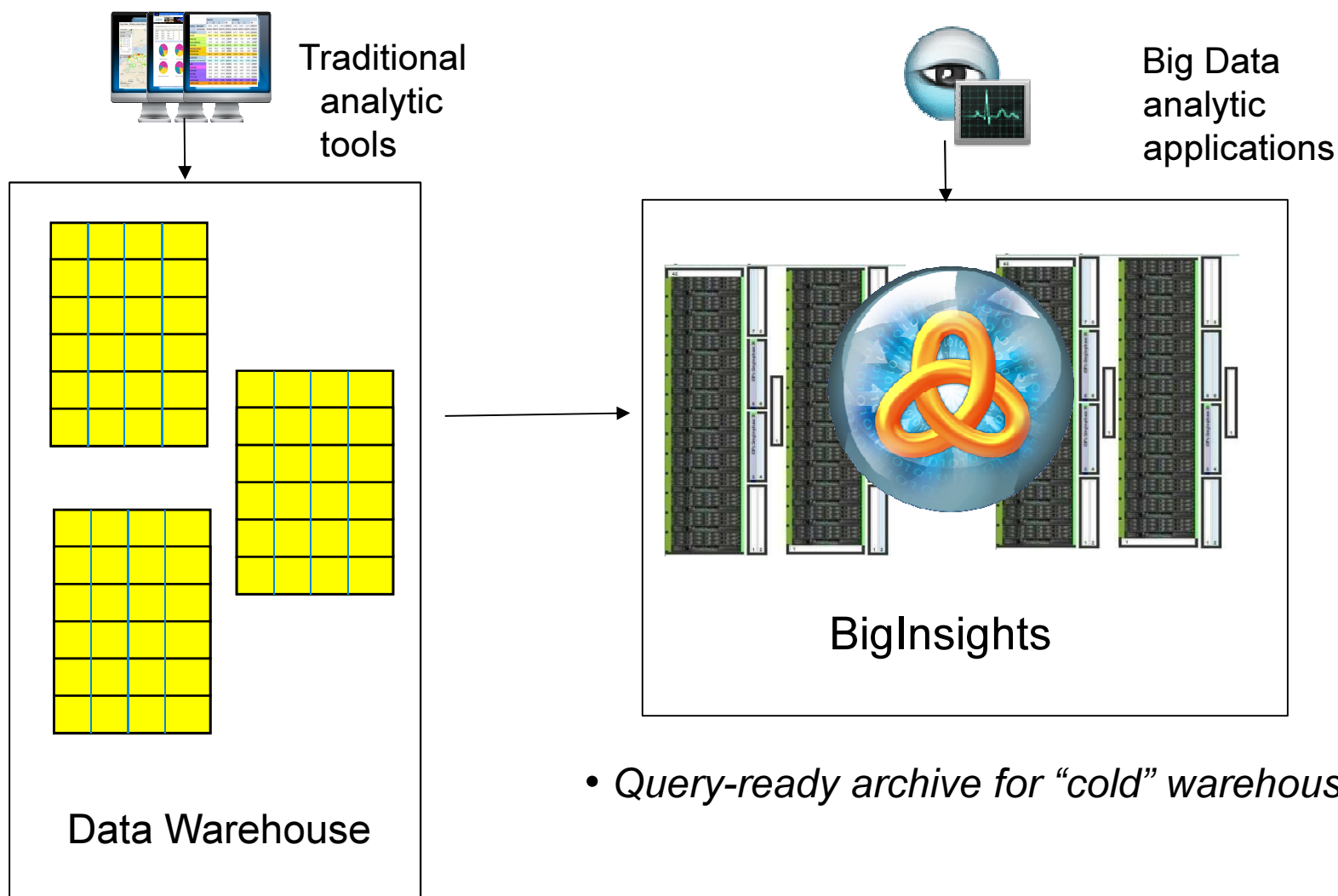
Business benefits

- Quicker time-to-value due to IBM technology and support
- Reduced operational risk
- Enhanced business knowledge with flexible analytical platform
- Leverages and complements existing software

BigInsights and the data warehouse



BigInsights and the data warehouse



Growing Ecosystem of Solutions

IBM Solutions



Partner Solutions



... with more to come

Installing Hadoop

3 Different Approaches

- **Roll Your Own (RYO) Hadoop**
- **Non-IBM Hadoop Distribution**
- **IBM Hadoop = BigInsights**

Roll Your Own Hadoop

Steps Required

- **Choose Hadoop components and their versions**
- **Create and setup a Hadoop user account**
- **Download each component and install them on cluster**
- **Configure SSH for user account and copy keys to computers**
- **Configure Hadoop**
- **Configure HDFS**
- **Define global variables**

Non-IBM Hadoop Distribution

- **Graphical installer**
- **Only for Hadoop, does not include other components**

BigInsights

Function	Version	Basic Edition	Enterprise Edition
Integrated Install		Inc	Inc
Hadoop (including common utilities, HDFS, MapReduce framework)	1.1.1	Inc	Inc
Pig (programming / query language)	0.10.0	Inc	Inc
Flume (data collection/aggregation)	1.3.0	Inc	Inc
Hive (data summarization/querying)	0.9.0	Inc	Inc
Lucene (text search)	3.3.0	Inc	Inc
Zookeeper (process coordination)	3.4.5	Inc	Inc
Avro (data serialization)	1.7.2	Inc	Inc
HBase (real time read/write)	0.94.3	Inc	Inc
HCatalog (table and storage management service)	0.4.0	Inc	Inc
Sqoop (RDBMS bulk data transfer)	1.4.1	Inc	Inc
Oozie (workflow/ job orchestration)	3.2.0	Inc	Inc
Jaql (query and scripting language)		Inc	Inc
Online documentation		Inc	Inc
Integration with JDBC sources through general-purpose Jaql module		Inc	Inc
Integration with DB2 (sample functions to submit jobs, read data)		Inc	Inc

BigInsights

- All components tested to work together
- Plus IBM components running on Hadoop

Introducing the BigInsights Web Console



Insert your name & email here

What is the Web Console?

- **Graphical interface for**
 - Administering and monitoring your cluster
 - Administering and monitoring BigInsights applications
 - Launching applications and analyzing data
 - Creating custom dashboards
 - Getting started with BigInsights!
- **Installation**
 - Completed as part of BigInsights Enterprise Edition installation
 - Installed on the node where the `start.sh` script is run.

Overview of Web Console Capabilities

- **Manage BigInsights**
 - Inspect / monitor system health
 - Add / drop nodes
 - Start / stop services
 - Launch / monitor jobs
 - Explore / modify file system
 - Create custom dashboards
 - . . .

The screenshot displays the IBM InfoSphere BigInsights Web Console interface. The top navigation bar includes links for Welcome, Dashboard, Cluster Status, Files, Applications, Application Status, and BigSheets. The main content area is divided into several sections:

- Understand IBM big data tools: Explore before doing**: Contains links to 'Learn about BigInsights' and 'Accelerate machine log, social, and telecommunications analytics'.
- Tasks**: A list of tasks including 'Create a dashboard', 'Chain (or link) applications', 'Explore and update data using sheets', 'Run an application', 'Deploy or remove an application', 'Add or remove a node', and 'View, start, or stop a service'.
- Quick Links**: A list of links for accessing secure cluster servers, running Big SQL queries, enabling the Eclipse development environment, downloading applications, Big SQL client drivers, Hive JDBC package, and Netezza UDFs.
- Learn More**: Links to accelerator demos, documentation, and information centers.

The bottom section shows the 'Application Status' tab, which displays a table of running jobs. The table has columns for Status, Name, ID, Progress, Created, Last Modified, Start Time, and End Time. The jobs listed are:

Status	Name	ID	Progress	Created	Last Modified	Start Time	End Time
✓	java-main-wf	0000000-121015103516380-oozie-biad-W	100%	Oct 15, 2012 11:27:26 AM	Oct 15, 2012 11:29:48 AM	Oct 15, 2012 11:27:27 AM	Oct 15, 2012 11:29:48 AM
✓	_SDA_Global_Analy	0000006-121014205133736-oozie-biad-W	100%	Oct 14, 2012 11:20:43 PM	Oct 14, 2012 11:55:50 PM	Oct 14, 2012 11:20:44 PM	Oct 14, 2012 11:55:50 PM
✓	_SDA_Local_Analy	0000005-121014205133736-oozie-biad-W	100%	Oct 14, 2012 10:51:45 PM	Oct 14, 2012 11:20:38 PM	Oct 14, 2012 10:51:45 PM	Oct 14, 2012 11:20:38 PM
✓	_SDA_Config_BMF	0000004-121014205133736-oozie-biad-W	100%	Oct 14, 2012 10:45:45 PM	Oct 14, 2012 10:51:40 PM	Oct 14, 2012 10:45:45 PM	Oct 14, 2012 10:51:40 PM
✓	App-Chaining-wf	0000003-121014205133736-oozie-biad-W	100%	Oct 14, 2012 10:45:37 PM	Oct 14, 2012 11:55:53 PM	Oct 14, 2012 10:45:38 PM	Oct 14, 2012 11:55:52 PM

■ Launch applications

Welcome Tab: Your Starting Point

Tasks: Where and how to begin performing common administrative or analytical tasks

Quick links to common functions

Understand IBM big data tools: Explore before doing

Learn about BigInsights
Understand the tools for analyzing data at rest and gaining business insights.

Tasks

- Accelerate machine log, social, and telecommunications analytics**
If you have installed one of the IBM accelerators, you can run applications to jump-start your big data analytics.
- Create a dashboard**
Create a dashboard to monitor your application...
- Chain (or link) applications**
Chain together several applications to run in a predefined sequence.
- Explore and update data using sheets**
Explore your data set to discover, analyze, and visualize your data.
- Run an application**
Run an application once, immediately.
- Deploy or remove an application**
Deploy an application on a cluster, or remove an application from a cluster.
- Add or remove a node**
Add a node and its service or services to a cluster, or remove a node from a cluster.
- View, start, or stop a service**

Quick Links

- Access secure cluster servers
- Run Big SQL Queries
- Enable your Eclipse development environment for BigInsights application development
- Download applications (Eclipse projects)
- Download the Big SQL Client drivers
- Download Hive JDBC package
- Download Netezza UDFs

Learn More

- Accelerator demos and documentation
- Infosphere BigInsights Information Center
- Infosphere Streams Information Center
- Communities and forums
- Support
- IBM big data on the Web

Learn more through external Web resources

Dashboard Tab

- Monitor overall system, data, and application services
- Create your own dashboard with supplied or custom widgets



Cluster Status Tab

- Inspect status of specific services, start/stop services
- Add / remove node(s)
- Activate monitoring to drive dashboard displays

Welcome | Dashboard | **Cluster Status** | Files | Applications | Application Status | BigSheets

Refresh Interval: 15 seconds ↕

Nodes ✔ 4

Map/Reduce ✔ Running

HDFS ⚠ Warning

Big SQL ✔ Running

Catalog ✔ Running

HBase ✔ Running

Hive ✔ Running

HttpFS ✔ Running

JAQL Server ✔ Running

Monitoring ✖ Unavailable

Oozie ✔ Running

Zookeeper ✔ Running

HDFS Summary

▶ Start ⏹ Stop | Balance Cluster

Namenode: hdfs://hdtest097.svl.ibm.com:9000

Status: ✔ Running

Process ID: 29380

Capacity: 6.78 TB (100%)

Used: 491.05 MB (0.01%)

Remaining: 6.78 TB (99.99%)

Datanodes

Start(1) Stop Add Data Directories...

Host & Port	Status	Capacity	Used	Remaining	Process ID	Last Contact	Data Directories
No filter applied							
hdtest100.svl.ibm.com:50020	✖ No heartbeat	0 B	0 B	0 B	9900	0s	/bi_hdfs /hadoop /hdfs/data
hdtest098.svl.ibm.com:50020	✔ Running	3.39 TB	245.53 MB	3.39 TB	7268	0s	/bi_hdfs /hadoop /hdfs/data
hdtest099.svl.ibm.com:50020	✔ Running	3.39 TB	245.53 MB	3.39 TB	12383	0s	/bi_hdfs /hadoop /hdfs/data

Files Tab

- Navigate the distributed file system to see what's stored
- Create / remove / rename directory
- Upload / download / move / rename files
- Execute Hadoop file system shell commands

The screenshot displays the Hadoop Files Tab interface. The top navigation bar includes links for Welcome, Dashboard, Cluster Status, Files, Applications, Application Status, and BigSheets. The left sidebar shows the HDFS file tree with the path `hdfs://streams.ibm.com:9000/` expanded, showing various directories like `IOD_Demo`, `accelerators`, `biginsights`, `hadoop`, `hbase`, `hdm-tera-input`, `hdm-tera-output`, `hdm-tera-report`, `tmp`, `user`, `applications`, `biadmin`, `staging`, `credstore`, `oozie-biad`, `sampleData`, `IBM_Watson_Jan_June_2012`, `WordCount_Output`, and `documents`. The `Lic_en.txt` file is selected in the `documents` directory.

The main content area shows the details for the selected file `Lic_en.txt` at the path `/user/biadmin/sampleData/documents/Lic_en.txt`. A table lists the file's metadata:

Name	Size	Block Size	Time
Lic_en.txt	59.1 KB	128.0 MB	Oct 15, 2012 2:21:54 PM

Below the table, the file's content is displayed, showing the "International License Agreement for Early Release of Programs".

Two modal windows are overlaid on the interface:

- Upload**: A window for uploading files. It includes a message: "Uploading large files may result in your web browser timing out." and a text input field for the upload destination, currently set to `/temp/mySample`. There is a "Browse..." button next to the input field. Below the input field, there is a section for "Files to Upload:" which shows a file named `labex.txt` with a red "X" icon, indicating an error.
- Hadoop File System Shell Command**: A window for executing Hadoop file system shell commands. It has a text input field for the command, currently containing `hadoop fs -ls /user/idcuser`. Below the input field, there is a section for "Hadoop Shell Command Output" which displays the results of the command:


```
Found 2 items
drwxr-xr-x - idcuser supergroup 0 2011-11-10 19:02 /user/idcuser/oozie-idcu
drwxr-xr-x - idcuser supergroup 0 2011-11-22 20:23 /user/idcuser/testBoardReader
```

Applications Tab

- Manage, execute, and link applications
 - Browse available applications
 - Deploy / undeploy applications
 - Launch (or schedule for launch) a deployed application
 - Monitor job (application) execution status

The screenshot displays the IBM Applications Tab interface. On the left, a sidebar titled 'Applications' shows a grid of application icons: Ad hoc Hive, Ad hoc Jaql, Ad hoc Pig query, and Boardreader. The main panel shows details for the 'Ad hoc Jaql query' application. It includes a description, an execution section with a 'Run' button, and a parameters section with a text area containing '\$test = [1,2,3];' and '\$test;'. At the bottom, an 'Application History' table shows the execution status of the 'JaqlTest' job.

Status	Execution Name	Progress	Start Time	Elapsed Time	Details
No filter applied					
	JaqlTest	100%	Nov 22, 2011 8:59:07 PM	21(sec)	

- Application Status

Show: **Workflows** | Scheduled Workflows | Jobs Auto Refresh: 5 seconds

Status	Name	ID	Progress	Run	Created	Last Modified	Start Time	End Time	
	No filter applied								
	perl-adhoc	0000000-11110902700037 000w-efcu-W	<div><div>100%</div></div>	0	Nov 22, 2011 8:59:07 PM	Nov 22, 2011 8:59:28 PM	Nov 22, 2011 8:59:07 PM	Nov 22, 2011 8:59:28 PM	
	java-main-wf	0000001-11110001500037 000w-efcu-W	<div><div>0%</div></div>	0	Nov 30, 2011 7:02:51 PM	Nov 22, 2011 9:04:01 PM	Nov 30, 2011 7:02:51 PM	N/A	
	map-reduce-wf	0000000-11110902700037 000w-efcu-W	<div><div>85%</div></div>	0	Nov 9, 2011 10:50:35 PM	Nov 9, 2011 10:50:53 PM	Nov 9, 2011 10:50:35 PM	Nov 9, 2011 10:50:53 PM	

The screenshot shows the JIRA Workflows page for the 'JIRA' project. The page title is 'Workflows (Scheduled Workflows) - JIRA'. Below the title, there is a table of workflows. The table has columns for 'Name', 'ID', 'Status', 'Start Time', 'End Time', 'Last Update', and 'Owner'. The 'JIRA' workflow is highlighted in blue. The 'JIRA' workflow has the ID 'jira_201111200147_0000' and is in the 'Active' state. The 'Start Time' is 'Nov 20, 2011, 9:58:00 PM' and the 'End Time' is 'Nov 20, 2011, 9:58:00 PM'. The 'Last Update' is 'Nov 20, 2011, 9:58:00 PM' and the 'Owner' is 'jira-admin'. Below the table, there is a section for 'All Tasks' with columns for 'Type', 'Start Time', 'End Time', 'Status', 'Priority', 'Assignee', 'Reporter', and 'Created Time'. The 'All Tasks' section shows a list of tasks with their respective details.

Name	ID	Status	Start Time	End Time	Last Update	Owner
JIRA	jira_201111200147_0000	Active	Nov 20, 2011, 9:58:00 PM	Nov 20, 2011, 9:58:00 PM	Nov 20, 2011, 9:58:00 PM	jira-admin

BigSheets Tab

- Spreadsheet-like analytical tool for non-programmers

The screenshot displays the BigSheets application interface. At the top, there is a navigation bar with tabs: Welcome, Dashboard, Cluster Status, Files, Applications, Application Status, and BigSheets. Below this, a 'Workbooks' section shows a list of workbooks with buttons for 'New Workbook', 'Purge', and 'Settings'. A 'New Workbook' dialog is open, showing a file tree on the left and a 'Line Reader' configuration on the right. The 'Line Reader' configuration includes a 'Select a reader' dropdown menu with options like 'Basic Crawler Data', 'Character Delimited Data', 'Comma Separated Value (CSV) Data', 'Hive Reader', 'JSON Array', 'Line Reader', 'Sheets Data', and 'Tab Separated Value (TSV) Data'. The 'Line Reader' option is selected. Below the configuration, a table of data is displayed with columns for 'Crawled', 'FeedInfo', 'Language', 'PostSize', 'PostTitle', and 'Publ'. To the right of the table, a pie chart titled 'Language Coverage for "IBM Watson" in blogs' is shown, with a legend indicating the percentage of each language. At the bottom, a 'Select a type of sheet:' section offers various sheet types: Filter, Macro, Load, Pivot, Join, Union, Intersection, Complement, Limit, Distinct, Copy, and Formula.

Supplemental slides

Starting or Stopping a Service (Cluster Status)

Welcome
Dashboard
Cluster Status
Files
Applications
Application Status
BigSheets

Refresh Interval: 15 seconds

Nodes
1
Map/Reduce
Running
HDFS
Running
Big SQL
Running
Catalog
Running
Flume
Running
HBase
Running
Hive
Running
JAQL Server
Running
Monitoring
Running
Oozie
Running
Zookeeper
Running

Monitoring Summary

Start
Stop

Version: 0.5.1
Running Agents: 1
Stopped Agents: 0

Monitoring Agents

Start
Stop

Host	Status	Adaptor Count	Process ID
streams.ibm.com:9093	Running	9	19086

Welcome
Dashboard
Cluster Status
Files
Applications
Application Status
BigSheets

Refresh Interval: 15 seconds

Nodes
1
Map/Reduce
Running
HDFS
Running
Big SQL
Running
Catalog
Running
Flume
Running
HBase
Running
Hive
Running
JAQL Server
Running
Monitoring
Unavailable
Oozie
Running
Zookeeper
Running

Monitoring Summary

Start
Stop

Version: 0.5.1
Running Agents: 0
Stopped Agents: 1

Monitoring Agents

Start
Stop

Host	Status	Adaptor Count	Process ID
streams.ibm.com:9093	Stopped		

30

© 2013 IBM Corporation

Adding or Removing a Node (Cluster Status)

The screenshot shows the IBM InfoSphere BigInsights web interface. The 'Cluster Status' tab is active, displaying a list of services on the left and a table of nodes on the right. The 'Add nodes' button in the 'Nodes' section is circled in green. An 'Add Nodes' dialog box is open in the foreground, showing a dropdown menu for 'Service' with 'DataNode/TaskTracker' selected. The dialog includes fields for 'Start IP/Host', 'Number of nodes', 'End IP', 'Rack', 'Root password', and 'Confirm_Root password'. The 'Nodes' field at the bottom is empty.

IBM InfoSphere BigInsights Cluster Status

Services and Status:

- Map/Reduce: Running
- Hive: Running
- JAQL Server: Running
- Flume: Running
- Zookeeper: Running
- Distributed File System: Running
- HBase: Running
- Oozie: Running
- Cataglog: Running

Nodes Table:

Host	Status	Roles
localhost.localdomain	Running	hive-server, secondarynamenode, zookeeper-client-port, hive-web-interface, bigsheets-web-interface, flume-node, flume-master, hbase-regionserver, datanode, namenode, tasktracker, jobserver, hbase-master, jobtracker

Add Nodes Dialog:

- Service: DataNode/TaskTracker
- Start IP/Host:
- Number of nodes:
- End IP:
- Rack:
- Root password:
- Confirm_Root password:
- Nodes:

Accessing Cluster Services

IBM InfoSphere BigInsights

Quick Links

- Access secure cluster servers
- Browse files
- View cluster status**
- View application, workflow, or job status
- Enable your Eclipse development environment for BigInsights application development

Learn More

URL Alias

http://vhost0018.dc1.co.us.compute.ihost.com:80030	libase-regionserver
http://vhost0018.dc1.co.us.compute.ihost.com:80010	libase-master
http://vhost0018.dc1.co.us.compute.ihost.com:50090	secondarynamenode
http://vhost0018.dc1.co.us.compute.ihost.com:50075	datanode
http://vhost0018.dc1.co.us.compute.ihost.com:50070	namenode
http://vhost0018.dc1.co.us.compute.ihost.com:50060	tasktracker
http://vhost0018.dc1.co.us.compute.ihost.com:50030	jobtracker

Done

IBM InfoSphere BigInsights

vhost0018 Hadoop Map/Reduce Administration

State: RUNNING
 Started: Wed Nov 09 22:43:08 GMT 2011
 Version: 0.20.2.1
 Compiled: Sat Oct 22 19:41:47 PDT 2011 by pshen
 Identifier: 201111092243

Cluster Summary (Heap Size is 12.56 MB/1000 MB)

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes
1	0	9	1	0	4	12.00	0













Scheduling Information

Queue Name	Scheduling Information
default	N/A

Filter (JobId, Priority, User, Name)

Example: "user.smith 3200" will filter by "smith" only in the user field and "3200" in all fields.

Managing Applications

Applications					
 Deploy  Delete					
Icon	Application Name	Categories	Status	Created By ▾	Configuration
	Ad hoc Hive query	Query,SQL	NOT DEPLOYED	samples	
	Ad hoc Jaql query	Query,SQL	NOT DEPLOYED	samples	
	Ad hoc Pig query	Query	NOT DEPLOYED	samples	
	Bivariate Statistics	Descriptive Statistics	NOT DEPLOYED	samples	
	Bivariate Statistics Decode	Descriptive Statistics	NOT DEPLOYED	samples	
	Boardreader	Web,Import	DEPLOYED	samples	
	Cluster aggregation	Monitoring	DEPLOYED	samples	
	Data Sample	Test	NOT DEPLOYED	samples	

Executing Applications

Welcome
Dashboard
Cluster Status
Files
Applications
Application Status
BigSheets

Manage
Execute
Link

Applications

Lead Generation Finance Local Analysis

Lead Generation Finance Real-time

Lead Generation Retail Configuration

Lead Generation Retail Configure - Local - Global

Lead Generation Retail Global Analysis

Lead Generation Retail Local Analysis

Lead Generation Retail Real-time Analysis

Word Count

Name: Word Count

Description:
The Word Count application reads text files and determines the frequency with which certain words occur.

Execution
Execution Name:

Parameters
* **Input path:**
* **Output path:**

[Schedule and Advanced Settings](#)

Application History

Status	Execution Name	Progress	Start Time	Elapsed Time (sec)	Output	Details
No filter applied						
	WC-Test	<div>100%</div>	Oct 15, 2012 2:22:47 PM	33		

1 - 1 of 1 items
10 | 25 | 50 | 100 | All
1

Linking or Chaining Applications



BigInsights Monitoring



<<Speaker Name Here>>

<<Speaker Title Here>>

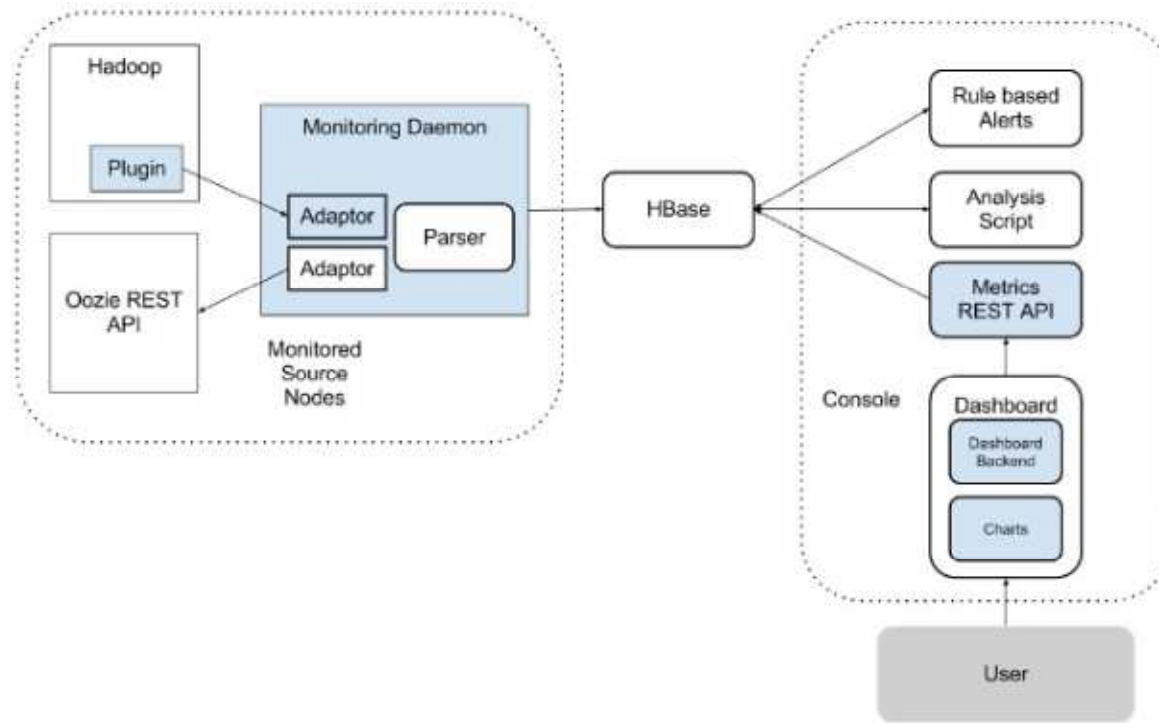
<<For questions about this presentation contact Speaker Name speaker@us.ibm.com>

Monitoring Overview

- **BigInsights Monitoring System is custom designed to run within Hadoop, Oozie and HBase Frameworks**
- **Monitoring data collected through JMX, REST and Sigar interfaces**
- **Timeseries data stored in HBase**
- **Visualization on console dashboard**
- **Monitoring for Hadoop, Oozie, HBase, Zookeeper, and physical system health**

Monitoring System Architecture

- Monitoring Agent supports both push and pull of data and transforms data into writable HBase structure
- Stateless REST API for query data from HBase
- MapReduce based analysis scripts to aggregate the data
- Personalized dashboard to visualize monitoring data



Operating Monitoring System

- **Monitoring is disabled by default; need to start monitoring agents post install**
- **Four monitoring applications are provided which need to be deployed and scheduled**

Start Monitoring Agents

- BigInsights Console -> Cluster Status tab, select Monitoring, and click Start

IBM InfoSphere BigInsights

About | Information Center

Welcome | Dashboard | **Cluster Status** | Files | Applications | Application Status | BigSheets

Refresh Interval: 15 seconds

Nodes 3

- Map/Reduce Running
- HDFS Running
- Catalog Running
- Flume Running
- HBase Running
- Hive Running
- JAQL Server Running
- Monitoring Unavailable**
- Oozie Running
- Zookeeper Running

Monitoring Summary

Start Stop

Version: 0.5.1

Running Agents: 0

Stopped Agents: 3

Monitoring Agents

Start Stop

Host	Status	Adaptor Count	Process ID
No filter applied			
bdvm157.svl.ibm.com:9093	Stopped		
bdvm156.svl.ibm.com:9093	Stopped		
bdvm158.svl.ibm.com:9093	Stopped		

1 - 3 of 3 items 10 | 25 | 50 | 100 | All

Deploy Monitoring Applications

- Click on Applications Tab
- Click on Manage link
- Select Monitoring from Categories
- Deploy Cluster Aggregation, Data Retention, DownSampling, and LogCollection applications

The screenshot displays the 'Applications' tab in the IBM Big Data Platform interface. The top navigation bar includes 'Welcome', 'Dashboard', 'Cluster Status', 'Files', 'Applications' (selected), 'Application Status', and 'BigSheets'. Below the navigation bar, there are links for 'Run', 'Manage' (selected), and 'Link'. On the left, a 'Categories' sidebar shows a tree view with 'Monitoring' selected, containing sub-items: 'Cluster Aggregation', 'Data Retention', 'DownSampling', and 'LogCollection'. The main area, titled 'Applications', features an 'Undeploy' button and a table of deployed applications.

Icon	Application Name	Categories	Status	Created By	Created At
	Cluster Aggregation	Monitoring	DEPLOYED	samples	
	Data Retention	Monitoring	DEPLOYED	samples	
	DownSampling	Monitoring	DEPLOYED	samples	
	LogCollection	Monitoring	DEPLOYED	samples	

Dashboards

- **Dashboards provide a centralized place within the InfoSphere BigInsights Console to:**
 - Visually monitor system usage and statistics
 - Display BigSheets workbooks and charts
- **There are two widget categories that can be added to a dashboard**
 - Workbook widgets (Top)
 - Visually display existing BigSheets workbooks or charts. The list of widgets available is based on the existing workbooks and charts that are listed on the BigSheets tab
 - Monitoring widgets (Bottom)
 - Visually display some type of application service, data service, or system metric



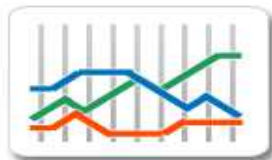
BM_ME_GA_Buzz
[Add Widget](#)



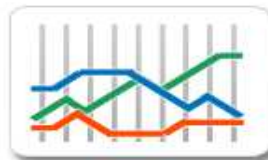
LG_Retail_LA_Buzz
[Add Widget](#)



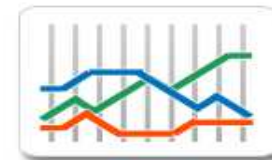
LG_Finance_LA_Buzz
[Add Widget](#)



Averaged Combined System CPU Usage Percentage
Averaged Combined System CPU Usage Percentage
[Add Widget](#)



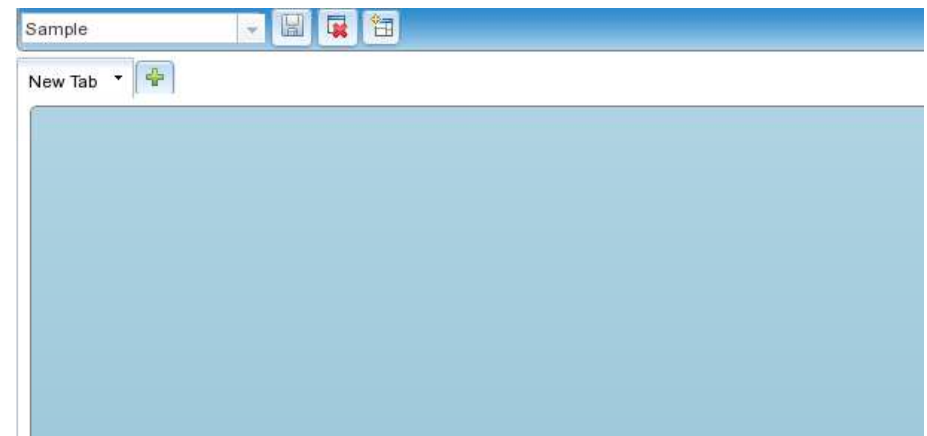
Averaged Number of Physical Disk Bytes Written
Averaged Number of Physical Disk Bytes Written
[Add Widget](#)



Averaged Percentage of Used System Memory
Averaged Percentage of Used System Memory
[Add Widget](#)

Creating a Custom Dashboard

- **You can create a custom Dashboard by**
 - Starting with a blank dashboard
 - Click New Dashboard
 - In the New Dashboard dialog, enter a unique name for the dashboard
 - Click OK
- **Once the dashboard is created you can then add tabs and widgets to display the information you want**



Monitoring Dashboard

- The BigInsights console dashboard contains widgets for components grouped under three categories
 - Application Services
 - Data Service
 - System



Monitoring Dashboard - Application Services

- **Widgets for visualizing Oozie and Map-Reduce related metrics**

- **MapReduce Activities**

- Metrics related to map-reduce jobs, mappers and reducers
 - Submitted, running and failed for each of these entities

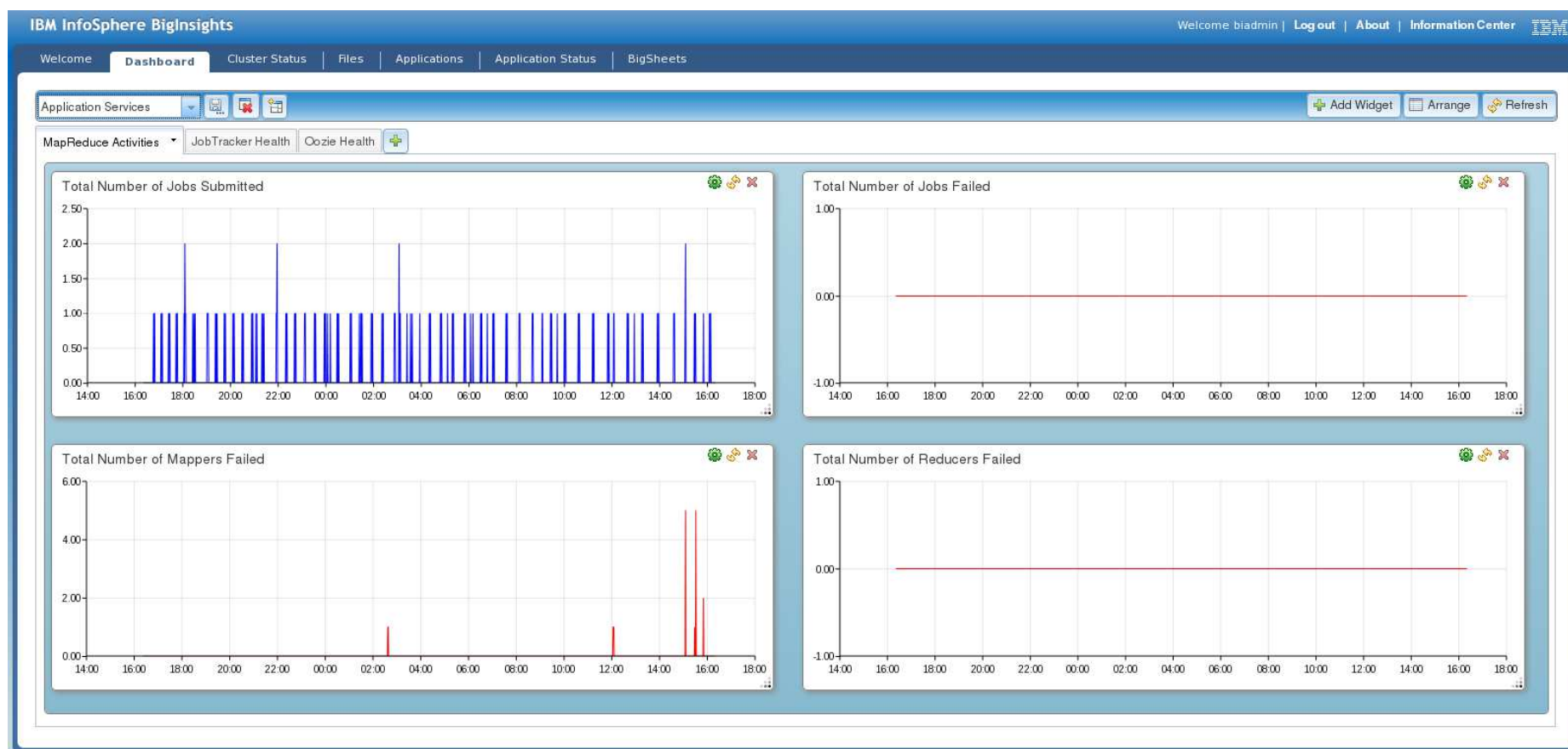
- **Jobtracker Health**

- Metrics related to job tracker JVM memory, garbage collection and heartbeats from task trackers

- **Oozie Health**

- Oozie JVM memory, garbage collection and heartbeats from task trackers
 - Monitor locks used to synchronize workflows

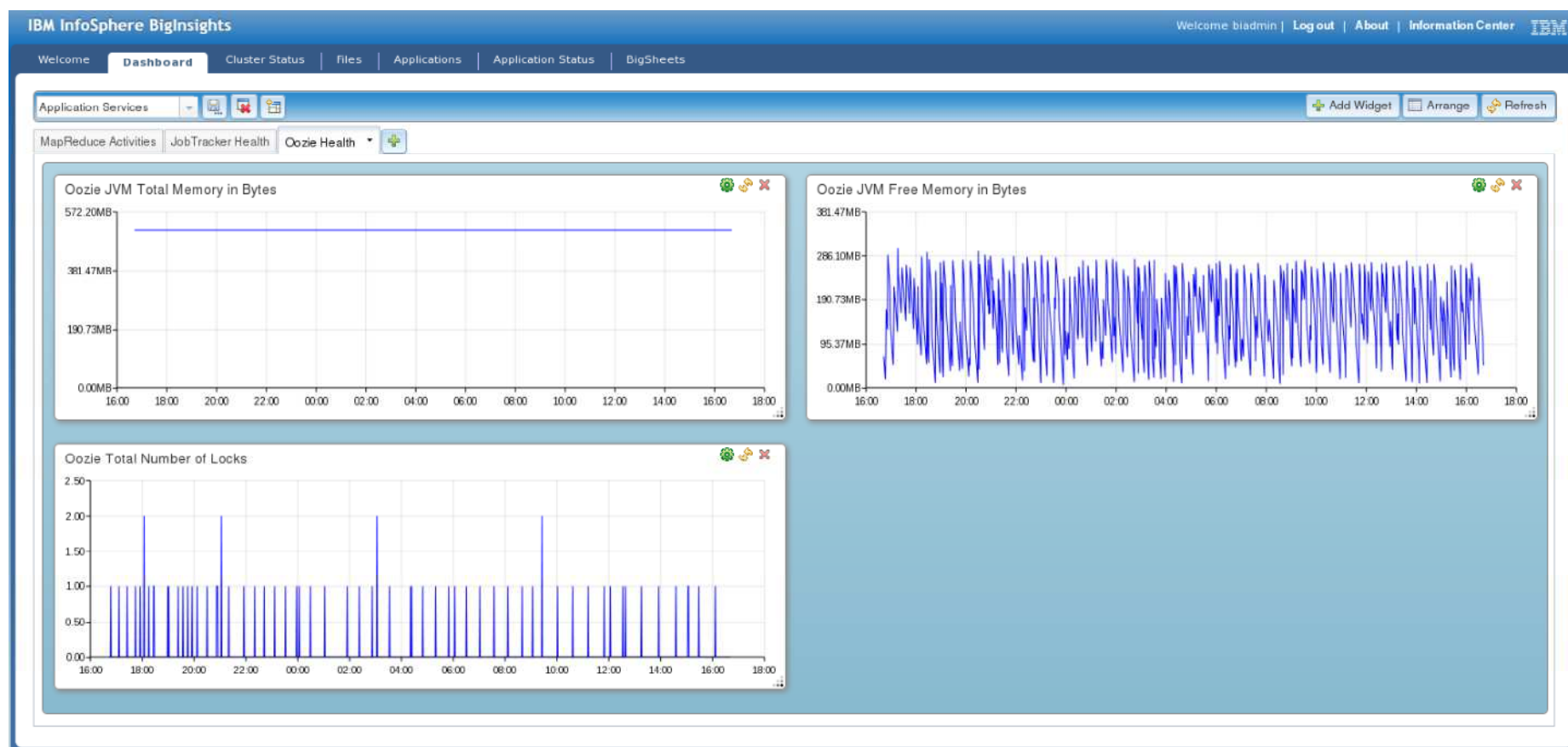
MapReduce Activities



JobTracker Health



Oozie Health



Monitoring Dashboard – Data services

▪ Widgets for visualizing HDFS, HBase, and Zookeeper related metrics

–HDFS

- Collected from the namenode and datanode processes
- Monitoring filesystem in terms of files and blocks
- Monitoring namenode health using JVM memory, garbage collection metrics
- Monitoring datanodes health using heartbeats and heartbeats per second
- Monitoring load and namenode through “Number of bytes received” widget

–HBase

- Metrics collected from the HBase masters and region servers
- Monitoring region splits
- Compaction queue
- Stores, store files
- Load and HBase region server using client requests
- Block caching – hit percentage, cache size and cache bytes available
- DFS performance as seen from HBase – number and latencies of read/write/sync operations

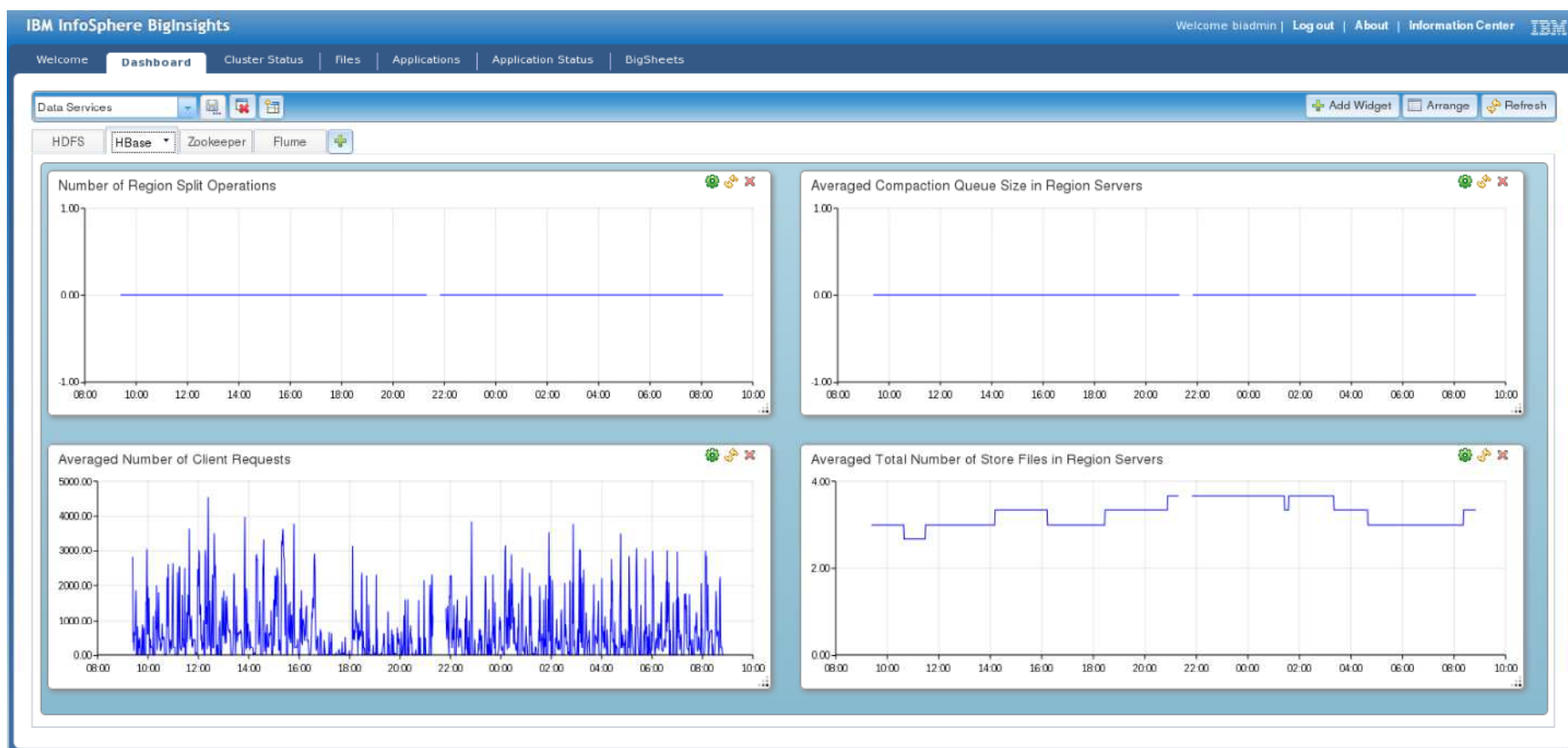
–Zookeeper

- Performance in terms of outstanding requests and request latencies (max/min/avg)
- Load in terms of packets received, packets sent and znodes
- ~~Watches count~~

HDFS Health Monitoring



HBase Health Monitoring



Zookeeper Health Monitoring



Monitoring Dashboard – System

▪ Widgets for visualizing cluster and node metrics

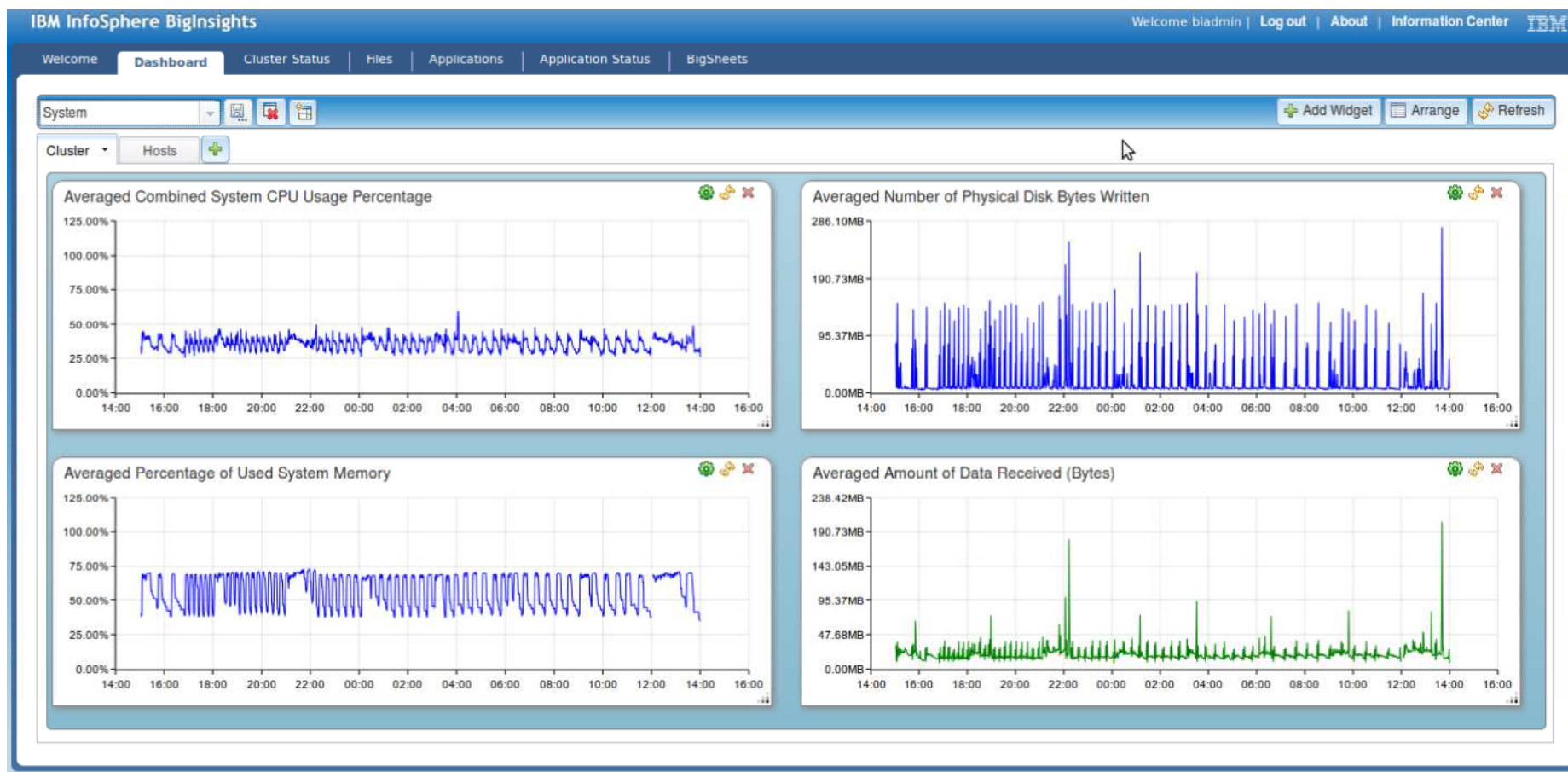
–Cluster

- Contains metrics for system metrics, averaged over the entire cluster
- CPU – as percentage usage, kernel time, idle time, IO wait time and interrupt services time across all nodes in the cluster
- Disk – bytes written, bytes read, number of reads, number of writes across all nodes in the cluster
- Memory – percentage of free and used system memory
- Network – for sent and received, we have amount of data (bytes), number of packets and number of packet errors.

–Hosts

- Contains a grid view of all nodes in a cluster
- Metrics related to CPU, memory, disk and network

Cluster Health



Hosts Health

IBM InfoSphere BigInsights

Welcome bladmin |
Log out |
About |
Information Center

Welcome

Dashboard

Cluster Status

Files

Applications

Application Status

BigSheets

System

Add Widget
Arrange
Refresh

Cluster

Hosts

Total Number of Available Hosts in the Cluster

Host	CPU	Memory Free	Disk Read	Disk Write	Network Received	Network Sent
bdvm027.svl.ibm.com	24%	63%	0	5984256	4709406.75	10138972.25
bdvm028.svl.ibm.com	44%	52%	78848	7570432	13848654.75	34156073
bdvm029.svl.ibm.com	23%	73%	2743296	6304768	5407494.5	12056562.25
bdvm030.svl.ibm.com	46%	24%	551936	6933504	22168577	21983785.25
bdvm023.svl.ibm.com	66%	8.0%	1222656	10760192	73241870	27110443.25
bdvm024.svl.ibm.com	0.88%	76%	0	5442560	157757	166092
bdvm025.svl.ibm.com	26%	80%	0	6159360	5324133.5	14753955.5
bdvm026.svl.ibm.com	25%	80%	0	5344256	4686273.5	9362744.25

Notes (IBM Internal Use Only)

- **Monitoring not recommended for use in VMware images due to resource consumption**
- **Monitoring shines in a multi-node cluster not in a single-node cluster**