



Search & Analytics Group IBM Research - Almaden

February 8, 2014

© 2010 IBM Corporation



Introduction to Text Analytics and BigInsights capabilities

Introduction

- What is text analytics?
- Example text analytics applications and data sources
- Critical success factors
- Previous approaches

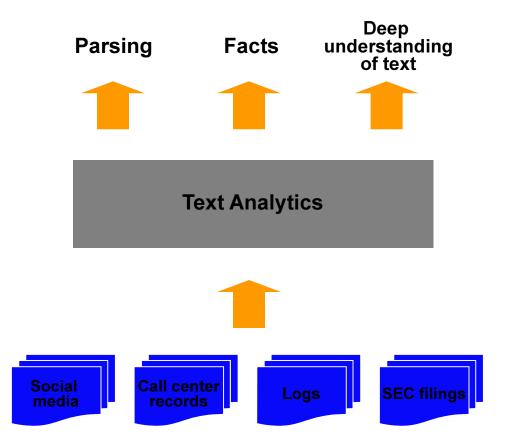
BigInsights Text Analytics Overview

- Architecture
- Key advantages

Tutorial Overview



What is Text Analytics ?





Verticals and Horizontals using Text Analytics

Verticals

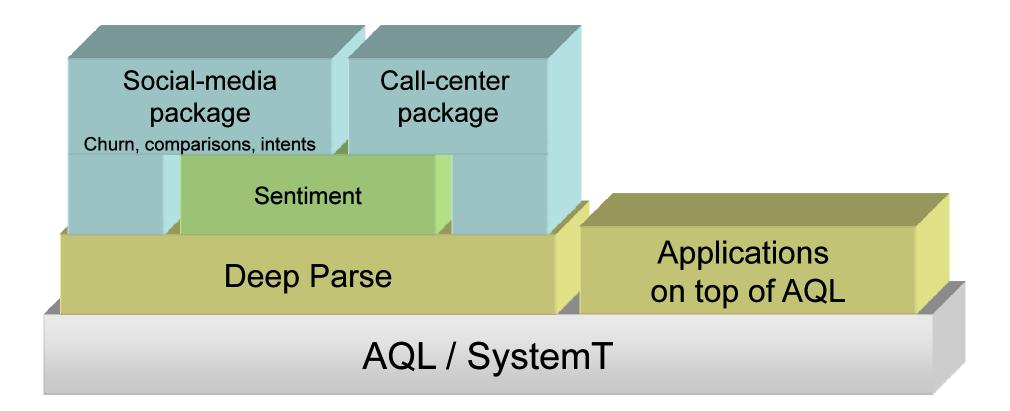
- Financial: financial events, company earnings, key players, etc.
- Healthcare: drugs / diseases, patient history, chemical compounds
- General: named entities (person, organization, location, phone, URL, email)

Horizontals

- CRM Analytics
 - Voice of customer
 - Product / Services gap analysis and in combination with Social Media predicting churn etc.
- Social Media Analytics
 - Retail applications such as intent identification and customer churn
 - Reputational Risk applications such timeliness of response
- Digital Piracy
 - Illegal broadcast of streaming and video content
 - Illegal dissemination of copyrighted digital material
- Log Analytics
 - Log parsing into fields, IP addresses, exception stack trace
- Data Redaction
 - Identify sensitive information (people names, DOB, SSN)
- Regulatory Compliance



Building Applications on top of BigInsights Text Analytics: A Conceptual View





Data Sources for Text Analytics are Highly Heterogeneous

- Variations in content quality: from formal to informal (noisy)
 - Formal: news reports, financial reports, patent applications
 - Informal: email, blogs, Twitter/Facebook posts

Variations in structure: from unstructured to semi-structured

- Unstructured: news reports
- Semi-structured: system/application logs, web pages, financial reports (SEC)

Variations in size: from very small to very large

- Small (bytes): Twitter posts
- Medium (Kilobytes): email, blogs, news reports
- Large (Megabytes): financial reports, patents



Critical Success Factors

Quality: Drives effectiveness of entire application

Need high accuracy and coverage

Performance: Dominant cost is CPU

 Process large documents and large number of documents with high throughput

Explainability

Determine the cause of errors and fix it without affecting the remaining correct results

Reusability: easily adaptable for a different domain

 The development platform must enable layers of abstractions to be built and easily reused in a different domain



Previous Approaches to Text Analytics

Statistical (Machine Learning)

- Labeled data required to train a model
- Model must be retrained for each domain
- The model is opaque

Rule-based

- No unified formal language
- Performance and expressivity limitations



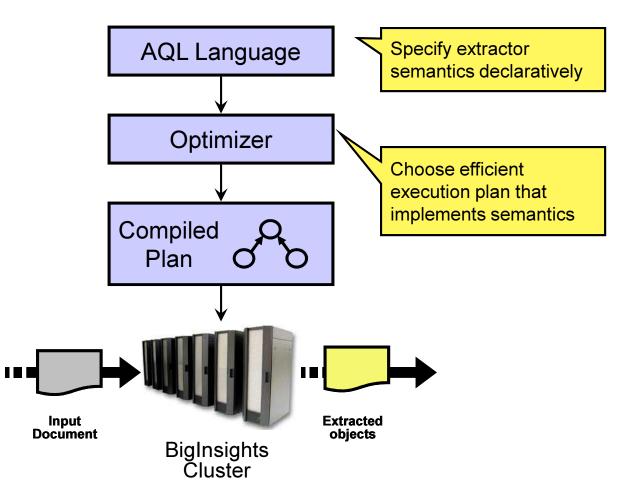
Outline

Introduction

- What is text analytics?
- Example text analytics applications and data sources
- Critical success factors
- Previous approaches
- BigInsights Text Analytics Overview
 - Architecture
 - Key advantages
- Tutorial Overview

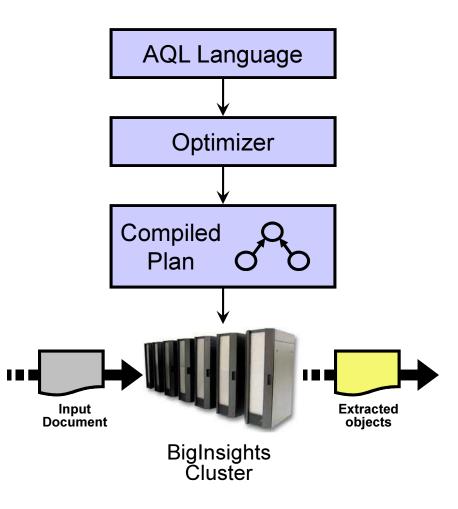


BigInsights Text Analytics Architecture





BigInsights Text Analytics Components



- Eclipse Tools
 - Develop and maintain extractors in AQL

Pre-compiled extractor library

- Western languages: Named Entities (person, organization, location, phone, URL, email, date/time) and financial events (merger, acquisition, company earnings)
- Chinese/Japanese: Named Entities (Person, Organization, Location)

Jaql Text Analytics module

 Execute extractors on the cluster from Jaql

Text Analytics Java API

 Invoke Text Analytics directly from your application



AQL: A Declarative Language to Specify Extraction Patterns



create view FirstCaps as
select CombineSpans(F.name, C.name) as name
from First F, Caps C
where FollowsTok(F.name, C.name, 0, 0);

Choice of SQL-like syntax for AQL motivated by wider adoption of SQL



The Expressivity of AQL

Feature Extraction primitives

- Regular Expressions
- Dictionary

Text-specific primitives

- Span-based predicates
- Multi-lingual support
 - Tokenization: Arabic, Chinese, Czech, Danish, Dutch, English, German, Greek, Spanish, French, Italian, Japanese, Korean, Finnish, Norwegian (Nynorsk and Bokmal), Polish, Portuguese, Russian, Swedish
 - Parts-of-speech: Chinese, English, German, Japanese, Spanish, French

Set-level primitives

- Join
- Block
- Consolidation
- Group By

AQL Reference in Info Center

http://publib.boulder.ibm.com/infocenter/bigins/v1r3/topic/com.ibm.swg.im.in fosphere.biginsights.doc/doc/biginsights_aqlref_con_aql-overview.html



Why "Declarative" Language ?

Semantics are separate from execution

- The developer expresses "what" to extract
- The system determines "how"

Advantages

- Explainability
 - Easy to understand and debug
- Global optimization
 - The system can determine an efficient execution plan
 - The developer does not worry about performance

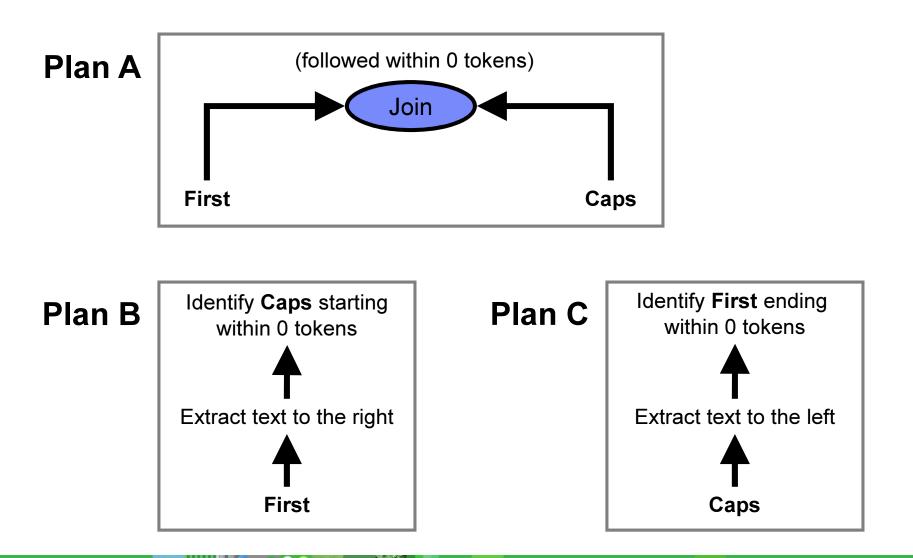


Scalability of BigInsights Text Analytics

- **1.** Better throughput via query optimization
- **2.** Massive scale-out on BigInsights

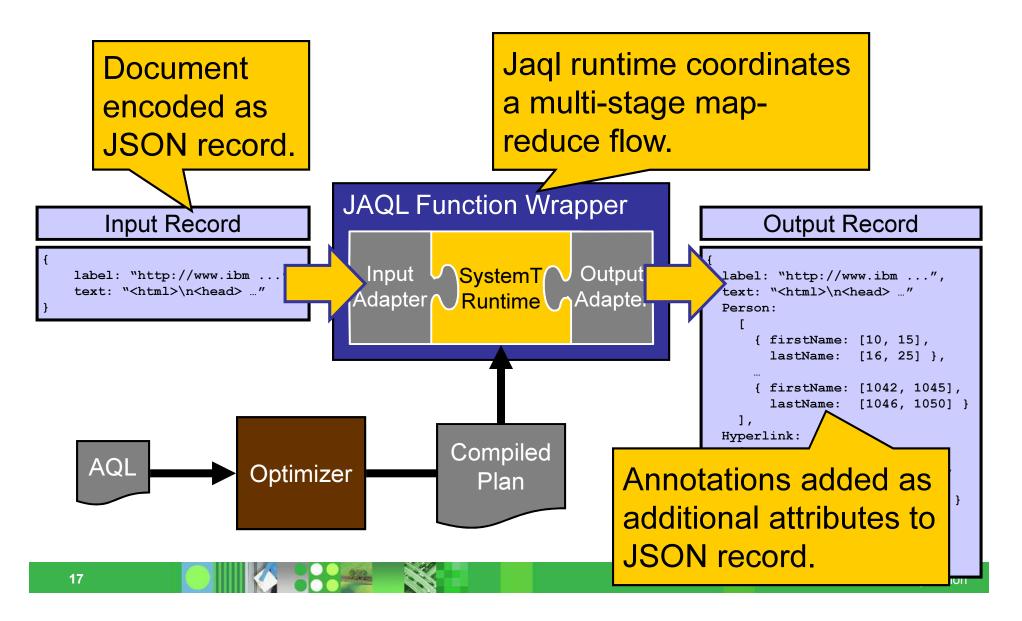


Scalability: Flavor of Optimization





Scaling Up on BigInsights Clusters



Information Management



Eclipse Tools Overview

Ease of Programming

Automatic Discovery AQL Editor: syntax highlighting, auto-complete, hyperlink navigation

Result Viewer: visualize/compare/evaluate

Explain: show how each result was generated

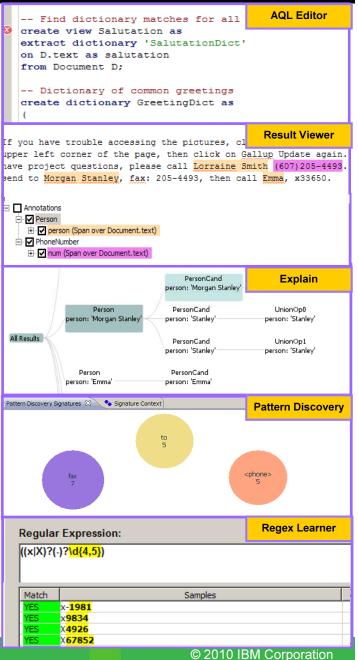
Workflow UI: enable novice users to become experts in a short time

Pattern Discovery: identify patterns in the data

Regex Generator: generate regular expressions from examples

Performance Tuning

Profiler: identify performance bottlenecks to be hand tuned





Key Advantages of BigInsights Text Analytics

- Quality: Drives effectiveness of entire application
 - Highly expressive AQL language → Easy to express complex concepts and to improve quality

Performance: Dominant cost in text analytics is CPU

- Optimizer \rightarrow Developers don't need to worry about performance
- Ease of Development, Explainability and Reusability
 - Looks like SQL \rightarrow Low learning curve
 - Declarative language \rightarrow Output can be explained by automatic tools
 - Eclipse Development Tools \rightarrow Rule editing/discovery

Best of rule-based and statistical approaches!

- \Box Rule-based at Runtime \rightarrow quality and performance
- Statistical approaches for building rules \rightarrow ease of development



Outline

Introduction

- What is text analytics?
- Example text analytics applications and data sources
- Critical success factors
- Previous approaches
- BigInsights Text Analytics Overview
 - Architecture
 - Key advantages
- Tutorial Overview



Developing an Extractor with AQL

- Combinations of syntactic patterns using regular expressions, dictionaries, span operations, relational operators and consolidate
 - Virtually everything you would need is exposed in the language
 - Material covered in Days 1 and 2
- Different domain \rightarrow the extractor must be customized
 - Domain adaptation guidelines in place
 - Material covered in Day 3
- Different language \rightarrow the extractor must be customized
 - Language adaptation process worked out with business partner



Existing AQL Extractor Library

- We developed many extractors by writing AQL rules
 - Combinations of syntactic patterns using regular expressions, dictionaries, span operations, relational operators and consolidate
 - Examples:
 - Simple entities: numbers, IP addresses, error messages, ...
 - Complex entities: person, organization, location
 - Names of drugs, diseases, chemical compounds,...
 - Financial facts: merger/acquisition, earnings, key players,...

Adapted some of the extractors for:

- Multiple industry verticals: financial, healthcare
- Multiple data sources: news, email, blogs, log records
- Multiple languages:
 - Named Entities (person, location, organization): Western languages (DE, EN, ES, FR, IT), Chinese, Japanese

Developed tooling for writing and customizing AQL

- Generate regular expressions and discover dictionaries
- Explain rule output and help fix mistakes
- Organize complex rule sets
- Material covered in Days 1-2, and partly Day 3



Fact Extraction versus Deeper Understanding

- A large percentage of use cases can be addressed by using syntactic patterns, building dictionaries, span operations, relational operators and consolidate
 - Extraction of facts

But there is also a need for deeper, semantic, extraction

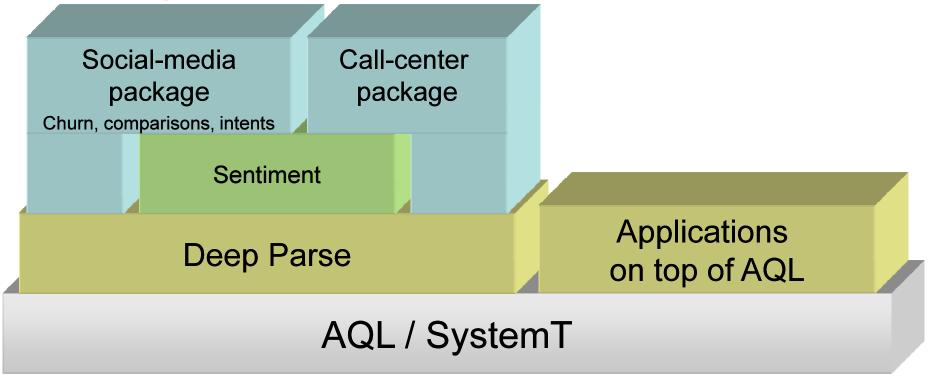
- Involves deeper understanding of linguistics and meaning of text
- BigInsights Text Analytics is sufficiently powerful for such analysis, but the rules and domain adaptation will be complex, BUT can be learned

Need to hide the complexity

- We are building higher levels of abstraction for semantic extraction
 - Deep parsing layer built using AQL
 - Sentiment layer built on top of deep parsing layer
- Working on guidelines for domain adaptation of the two layers
 - Introduction to this material will be given on Day 3



Building Applications on top of BigInsights Text Analytics: A Conceptual View





References (1/2) – Peer-reviewed Publications

- Overview
 - Rajasekar Krishnamurthy, Sriram Raghavan, and Huaiyu Zhu: "Evolution of Rule-Based Information Extraction: From Grammars to Algebra", Tutorial given at CIKM 2008.
 - Laura Chiticariu, Yunyao Li, Sriram Raghavan, and Frederick Reiss: "Enterprise Information Extraction: Recent Developments and Open Challenges". SIGMOD 2010 (tutorial)

Runtime Engine and Extractor Library

- Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, Shivakumar Vaithyanathan: "An Algebraic Approach to Rule-Based Information Extraction". ICDE 2008: 933-942
- Rajašekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, Shivakumar Vaithyanathan, Huaiyu Zhu: "SystemT: a system for declarative information extraction". SIGMOD Record 37(4):7-13 (2008)
- Eirinaios Michelakis, Rajásekar Krishnamurthy, Peter J. Haas, Shivakumar Vaithyanathan: "Uncertainty management in rule-based information extraction systems". SIGMOD Conference 2009: 101-114
- Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, Shivakumar Vaithyanathan: "SystemT: An Algebraic Approach to Declarative Information Extraction". ACL 2010.
- Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, Shivakumar Vaithyanathan: "Domain Adaptation of Rule-based Annotators for Named-Entity Recognition Tasks". EMNLP 2010.
- Daisy Zhe Wang, Long Wei, Yunyao Li, Frederick Reiss and Shivakumar Vaithyanathan. Selectivity Estimation for Extraction Operators over Text Data. ICDE 2011

Tooling

- Yunyao Li, Rajasekar Krishnamurthy, Sriram Raghavan, Shivakumar Vaithyanathan, H. V. Jagadish: "Regular Expression Learning for Information Extraction". EMNLP 2008: 21-30
- Bin Liu, Laura Chiticariu, Vivian Chu, H.V. Jagadish, Frederick Reiss: "Automatic Rule Refinement for Information Extraction". PVLDB 2010.
- Yunyao Li, Vivian Chu, Sebastian Blohm, Huaiyu Zhu, Howard Ho. "Facilitating Pattern Discovery for Relation Extraction with Semantic-Signature-based Clustering". CIKM 2011



References (2/2) – US Patents

Runtime Engine

- US Patent Publication 20090198646: "Systems, Methods and Computer Program Products for an Algebraic approach to Rulebased Information Extraction"
- US Patent Publication 20100174718: "Indexing for Regular Expressions in Text-Centric Applications"
- US Patent Application 12/788,142: "An Extensible System for Information Extraction in a Data Processing System"

Tooling

- US Patent Application 12/269,216: "User-Guided Regular Expression Learning"
- US Patent Application 12/788,407: "Method for Automatic Refinement of Information Extraction Rules",
- US Patent Application 13/117,570: "A Semantic-Signature-based Method for Contextual Clue Pattern Discovery of Information Extraction"





Introducing Big SQL for BigInsights

IBM's SQL Query Interface for Hadoop

<<Speaker Name Here>> <<Speaker Title Here>> <<For questions about this presentation contact Speaker Name speaker Ous.ibm.com>

February 8, 2014

© 2010 IBM Corporation



SQL

Executive Summary

- What is Big SQL?
 - Industry-standard SQL query interface for BigInsights data
- Why Big SQL?
 - Easy on-ramp to Hadoop for SQL professionals
 - Support familiar SQL tools / applications (via JDBC and ODBC drivers)
- What SQL operations are supported?
 - Create tables (and, optionally, HBase indexes)
 - Load data into tables (from local files, distributed files, RDBMSs)
 - Query data (project, restrict, join, union, sub-queries)
- What Hadoop-based storage mechanisms are supported?
 - Hive
 - HBase
 - Distributed file system



Agenda

Big SQL: motivation and architecture

Using Big SQL

- Invocation options
- Creating tables
- Populating tables with data
- Querying data
- Developing applications and working with tools
- . . . And a peek at some additional topics
- What RDBMS professionals should know about





Agenda



Using Big SQL

- Invocation options
- Creating tables
- Populating tables with data
- Querying data
- Developing applications and working with tools
- . . . And a peek at some additional topics

What RDBMS professionals should know about

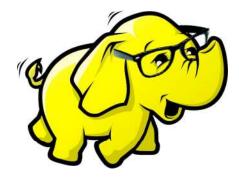


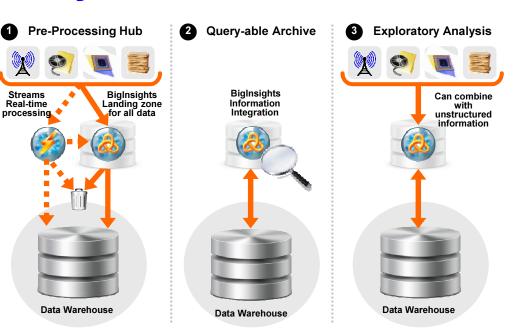
Information Management



SQL Access for Hadoop: Why?

 Data warehouse augmentation is a leading Hadoop use case





Hadoop often perceived as difficult

- MapReduce Java API requires programming expertise
- Unfamiliar languages (such as Pig) also require special skills

SQL support opens the data to a much wider audience

- Familiar, widely known syntax
- Common catalog for identifying data and structure





Big SQL Architecture and Feature Overview

Standard SQL syntax and data types

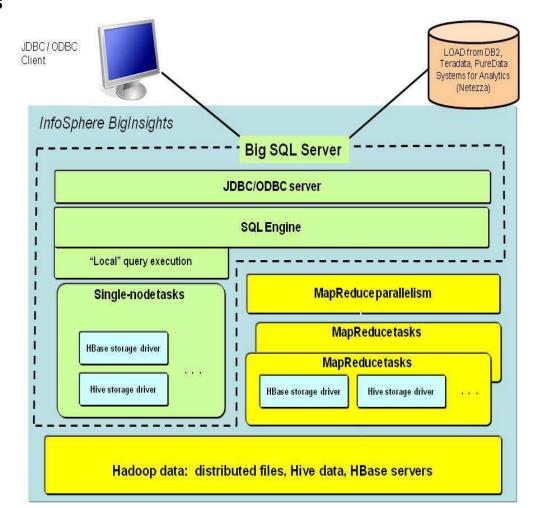
- Joins, unions, aggregates . . .
- VARCHAR, decimal, TIMESTAMP, ...

JDBC/ODBC drivers

- Prepared statements
- Cancel support
- Database metadata API support
- Secure socket connections (SSL)

Optimization

- MapReduce parallelism or...
- "Local" access for low-latency queries
- Varied storage mechanisms appropriate for Hadoop ecosystem
- Integration
 - Eclipse tools
 - DB2, Netezza, Teradata (via LOAD)
 - Cognos Business Intelligence
 - **—** , , ,





Agenda

Big SQL: motivation and architecture

Using Big SQL^{*}

- Invocation options
- Creating tables
- Populating tables with data
- Querying data
- Developing applications and working with tools
- . . . And a peek at some additional topics
- What RDBMS professionals should know about





Invocation options provided with BigInsights

- **Command-line interface (JSqsh shell)**
- Web-based interface (BigInsights web console)
- Eclipse (BigInsights plug-in) - 6 🖓 🗖 🖨 Task Launcher for Big Data 🛛 📓 TestSQL.sql JDBCapp.java 🖙 Data Project E 🛛 E 2 SampleProject (Big SQL) Overview Accelerate ! Design | Develop Publish and run Preferences X Properties for Big SQL - hdtest097.svl.ibm.com symphony Tasks **Big SQL JDBC Connection Properties** type filter text Create a text extractor **Big SQL JDBC Con** Follow a step-by-step wizard to creat Drivers: IBM Big SQL JDBC Driver 2.0 Default ▼ 😤 🛆 Common extractor (including generating regula Data Privacy Mod Properties , visualize the results of running the e Default Schema Fi evaluate extractor quality. Discover p General Optional view differences in the results of two **Default Stored Prc** Catalog: default runs, and interpret the lineage of extr Default Table Filte Host: hdtest097.svl.ibm.com Version Create a BigInsights program Port number: 7052 Ξ Create a Jagl script or module, a BigS or reader, a BigInsights Java program User name: biadmin MapReduce program, a SQL script, c Password: (III Create a configuration and run a B program Create a configuration to run a Jagl, ? OK Cancel MapReduce, or Java program. < Ⅲ Properties SQL Results X 🙀 Data Source E 🛛 - 6 Type query expression here Status Result1 8 5 6 8 4 4 Operation Connection Pro A id . Status Date name ➢ Configuration Repo: ▲ ✓ Succeeded load hive data inpath '... 4/26/1... Big SQL - hdtes 1 111 "The Business Journals" 🗁 Database Connectio 🗏 ✓ Succeeded create external table ... 4/26/1... Big SQL - hdtes = 2 222 "CNN" Big SQL - hdtest0 ✓ Succeeded select id, name from ... 4/26/1... Big SQL - hdtes 3 333 "CIO Today"



Creating a Big SQL Table

 BigSQL supports CREATE TABLE and many data types including varchar, decimals, etc. Non-ISO standard clauses leverage Hadoop ecosystem

CREATE TABLE TPCH.CUSTOMER (C_CUSTKEY INTEGER, C_NAME VARCHAR(25), C_ADDRESS VARCHAR(40), C_NATIONKEY INTEGER, C_PHONE CHAR(15), C_ACCTBAL FLOAT, C_MKTSEGMENT CHAR(10), C_COMMENT VARCHAR(117)) row format delimited fields terminated by '|' stored as textfile;

• Hive does not support data types like varchar and decimal*



CREATE TABLE TPCH.CUSTOMER (C CUSTKEY INTEGER, C NAME VARCHAR(25), C ADDRESS VARCHAR(40), C NATIONKEY INTEGER, C PHONE CHAR(15), C ACCTBAL FLOAT, C MKTSEGMENT CHAR(10), C COMMENT VARCHAR(117)) row format delimited fields terminated by '|' stored as textfile;

> *Hive 0.11 added DECIMAL

> > © 202013BBM Corporation



Results from CREATE TABLE ...

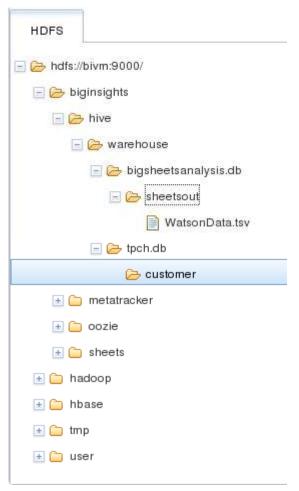
- Table
 - Subdirectory created in warehouse directory /biginsights/hive/warehouse/tablename/
 - External tables may have their data stored anywhere in the DFS
 - Populated tables contain 1 or more data files

Schema (or database)

- Tables may be organized by schemas
- Schema is just a collection of tables
- Creating a schema creates a subdirectory in the warehouse to hold the tables

/biginsights/hive/warehouse/schema.db/
 tablename/

Catalog data (more later)





Big SQL Extensions to CREATE TABLE

 Additional data types: BINARY(N), VARCHAR(N), DECIMAL(P,S)

NULL/NOT NULL indicators

- These are advisory only not enforced
- Big SQL query re-write software takes advantage of this info

Table hints

- Certain optimizer himestante ante attached to tables
- Hint will automatically apply when the table is used in a query

```
with hints (tablesize=`small')
```

³⁷Explicit syntax for HBase tables (column mappings. columnoration

Populating Tables

- Data can be LOADed from
 - Local file system
 - Distributed file system
 - Remote Netezza, DB2, or Teradata RDBMS

Example

CREATE TABLE EMPLOYEE (EMPNO INT, NAME STRING, AGE INT) ...;

// Overwrite any existing data with new data from a local file LOAD HIVE DATA LOCAL INPATH '/home/user1/employee.data' OVERWRITE INTO TABLE EMPLOYEE;

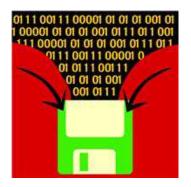
// Append new data from a file in HDFS to the table LOAD HIVE DATA INPATH '/user/biadmin/employee.data' INTO TABLE EMPLOYEE;

What LOAD does:

- Copies or moves the data, but doesn't manipulate it
- Format of the input file must match the format of the table

HBase notes:

- Similar LOAD syntax (LOAD HBASE). Composite keys, indexes, column encoding handled.
- A single row INSERT may be used against HBase table 38





© 2010 IBM Corporation



Querying data: Overview of SQL Support

Projection

SELECT col1, col2 FROM t1

Restriction

SELECT * FROM t1 WHERE col1 > 5

Union

SELECT EMPNO FROM EMPLOYEE WHERE WORKDEPT LIKE 'E%' UNION SELECT EMPNO FROM ACTIVITIES WHERE PROJNO IN('MA2100', 'MA2110', 'MA2112')

Difference (EXCEPT)

(SELECT * FROM T1) EXCEPT ALL (SELECT * FROM T2)

Intersection

Subqueries

(SELECT * FROM T1) INTERSECT (SELECT * FROM T2)

Joins

39

© 2010 IBM Corporation



SQL Support - Joins

Big SQL supports both common and ANSI join syntax



select ...
from tpch.orders, tpch.lineitem
where o orderkey = 1 orderkey



select ... from tpch.orders join

tpch.lineitem
 on o_orderkey =
l_orderkey

Hive supports joins via ANSI join syntax only select



from tpch.orders, tpch.lineitem where o orderkey = 1 orderkey from tpch.orders join

tpch.lineitem
 on o_orderkey =
l_orderkey



SQL Support – Subqueries

Big SQL supports subqueries



select c1,
(select count(*) from t2)
 from t1
 ...

select c1
 from t1
where c2 > (select ...)

Hive does not support subqueries



select c1, (select count(*) from t2) from t1

. . .

select c1
from t1
where c2 > (select ...)



SQL Support – Aggregates

Big SQL supports windowed aggregates

year	total_sales	ranked_sales
2006	1495891100.90	1
2005	1159195590.16	2
2007	1117336274.07	3
2004	914352803.72	4

• Hive does not support windowed aggregates



SQL Support – Functions (partial list)

Wide variety of built-in functions

- Numeric

abs	ceil	floor	ln	log10
mod	power	sqrt	sign	width_bucket

- Trigonometric

cos	sin	tan	acos	asin
atan	cosh	sinh	tanh	

- Date

_add_days	_add_months	_add_years	localtimestamp	_age
_day_of_week	_day_of_year	_week_of_year	_days_between	_months_between
_years_between	_ymdint_between	_first_of_month	_last_of_month	extract

- String

char_length	bit_length	octet_length	upper	lower
substring	position	index	translate	trim
json_get_object				

- Aggregates, etc.



Catalog Tables (HCatalog)

[localhost][foo] 1> select * from syscat.tables where tablename='users';

| schemaname | tablename |

| default | users

1 row in results (first row: 0.14s; total: 0.15s)

[localhost][foo] 1> select * from syscat.columns where tablename='users';

schemaname	tablename	name	type	precision	scale
default	users	id	INT	10	0
default	users	office_id	INT	10	0
default	users	name	STRING	0	0
default	users	children	ARRAY	0	0

4 rows in results (first row: 0.19s; total: 0.21s)

Other BigInsights catalog tables track index and schema information

Information Management



Using Existing Standard SQL Tools: Eclipse

BigInsights Servers 🛛 🕈 🕈	□ □			🗇 🕼 Help 📢 Data Source Explorer 🛛
BigInsights Servers	<pre>1 Supplicient 1 1 SELECT 2 L_RETURNFLAG, 3 L_LINESTATUS, 4 SUM (L_QUANTITY) AS SUM_QTY, 5 SUM (L_EXTENDEDPRICE) AS SUM_BASE 6 SUM (L_EXTENDEDPRICE* (1-L_DISCOUNT) 7 SUM (L_EXTENDEDPRICE* (1-L_DISCOUNT) 8 AVG (L_QUANTITY) AS AVG_QTY, 9 AVG (L_EXTENDEDPRICE) AS AVG_PRICE 10 AVG (L_DISCOUNT) AS AVG_DISC, 11 COUNT (*) AS COUNT_ORDER 12 FROM LINEITEM 13 WHERE L_SHIPDATE <= '1998-09-02' 14 GROUP BY L_RETURNFLAG, L_LINESTATUS 15 ORDER BY L_RETURNFLAG, L_LINESTATUS 16 16 17 Where Console * Call Hierarchy X Search 16 Search 17 To display the call hierarchy, select one or more methods, content 10 AVG (L_MARCH, Select one or more methods, content 11 Count (*) Search 12 FROM LINEITEM 13 WHERE L_SHIPDATE <= '1998-09-02' 14 GROUP BY L_RETURNFLAG, L_LINESTATUS 15 ORDER BY L_RETURNFLAG, L_LINESTATUS 16 Search 16 Search 16 Search 17 Odisplay the call hierarchy, select one or more methods, content 17 Odisplay the call hierarchy select one or more methods, content 18 OF Search 19 OF Search 10 OF Search 10</pre>	 AS SUM_DISC_ * (1+L_TAX)) AS Properties for BigSQL - 1 Big SQL JDBC Connec Common Default Schema Filter Default Stored Procec Default Table Filter Version 	SUM_CHARGE, SUM_CHARGE, Big SQLJDBC Connecti Drivers: IBM Big SQLJDBC D Properties General Optional Catalog: default Host: 170.224 Port number: 7052 User name: biadmin Password: ••••••	Big SQL - bdvm072.svl.ibm.co BigSQL - 170.224.193.37 (IBM Constraints) Big default Big gosales Big gosa
eclipse				Test Connec

Information Management



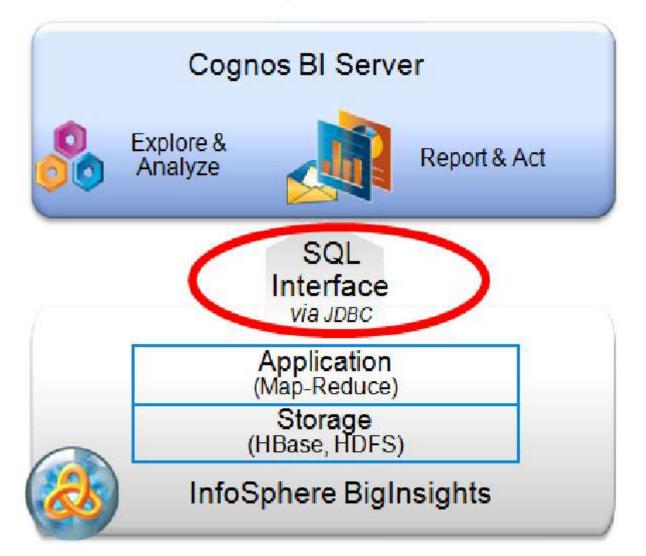
Using Existing Standard SQL Tools: SQuirreL SQL

SQuirreL SQL Client Versi	ion 3,4.0											
Elle Drivers Aliases Plugins	Session Windows Help											
Connect to: BigSQL vm.	📽 🖉 Active Session: 1 - BigSQL vm072 () as bladmin 💌 👧 🛞 🖉 📑 14											
Allanes -	× (] 1-BigSQL vm072 () as bladmin											
* * * / 13 *		Q @ @ @ Ø	6º 6º									
RigSQL vm072	Objects SQL											
	SELECT L_RETURNFLAG, L_LINESTATUS, SUM(L_QUANTITY) AS SUM_GTY, SUM(L_EXTENDED PRO	CE) AS SUM_BASE_PRICE	SUMIL_EXTENDE V 1 E ELimit rows 100									
Drivers	SUM(L_EXTENDED PRICE) AS SUM_BASE_PRICE, SUM(L_EXTENDED PRICE*(1-L_DISCOUNT)) AS SUM_DISC_PRICE, SUM(L_EXTENDED PRICE*(1-L_DISCOUNT)*(1+L_TAX)) AS SUM_CHARGE,											
	AVG(L_OUAMTITY) AS AVG_OTY, AVG(L_EXTERDEDYRICE) AS AVG_PRICE, AVG(L_DISCOUNT) AS AVG_DISC, COUNT(*) AS COUNT_ORDER	Add Alias										
	FROM tpch.LINEITER WHERE L SHIPDATE - '1998-09-02'	Add Alias										
	GROUP BY L PETURNFLAG, L LINESTATUS	Name:	BigSQL on bdvm072.svl.ibm.com									
	COPPER DI L'ALIDRIARO, L'ALIBEDIRINO,	Driver:	✓ IBM Big SQL JDBC Driver 2.0 Default ▼ New									
	select count(*) SELECT L RETURN	URL:	identities of //indum 070 and item come:7050/default									
	Rows 4: SELECT L RETURNETAG, L LINESTATUS, SUN(L QUANTITY) AS SUM CTY, SUM(L E	UKL.	jdbc:bigsql://bdvm072.svl.ibm.com:7052/default									
	Results MetaData Info Overview	User Name:	biadmin									
	L_returnflag L_linestatus SUM_QTY SUM_BASE_PRICE S	Password:	•••••									
	A F 380,456 532,348,211,54083 505,82 N F 8,971 12,384,801,3941 11,798 N O 742,802 1,041,502,841,74463 989,73 R F 381,449 534,594,465,51434 507,99	🔲 Auto logon	Connect at Startup									
	2R P 0361,449 0394,943,51434 0507,33		Properties									
		Warning - Pass	words are saved in clear text									
SQuirreL SQ.			OK Close Test									





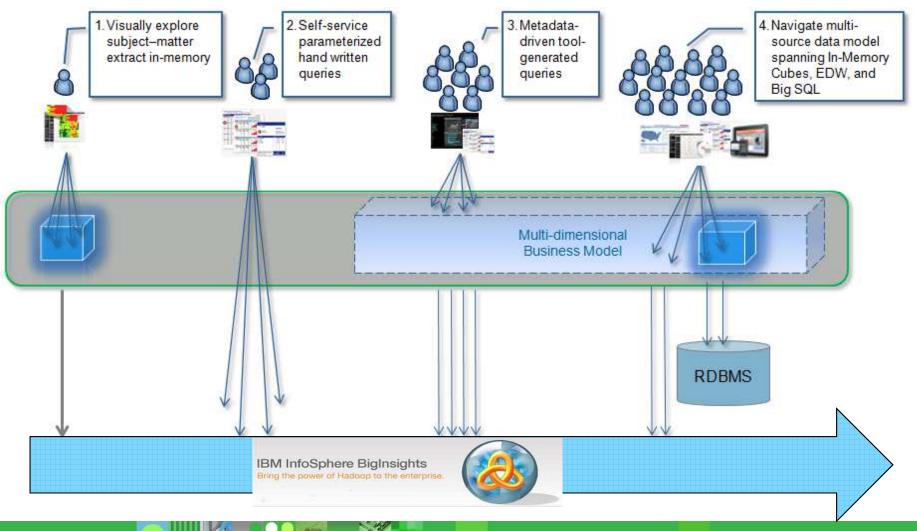
Cognos Business Intelligence





MicroStrategy use of Big SQL

MicroStrategy



© 2010 IBM Corporation

Information Management



MS Excel: Big SQL integration via ODBC

1 Big SQL ODBC Client Setup		icrosoft E		-	rad F		Tech	Det	Minday, Hala								
	1.1.1.0000000000	Eile Ed							a <u>W</u> indow <u>H</u> elp Sort			10					
ig SQL Server configuration						1 相對以	B v 💚	Z*		Arial		- 10	- B 1	<u>n</u> e a		\$ %	%
Please enter the myBigSQL DSN's informati		A1	-		fx O		-		Eilter	•		1				1 2	-
	1	A	t	3	С		D		F <u>o</u> rm		l J	K	L	M	N		0
Database:	2						-		Su <u>b</u> totals					_			
default	3								Validation							-	-
lost:	4								Table								
na cine	5								Text to Columns								
sdsvm691013.svl.ibm.com	6												-		-	_	_
Port:	7 8						_		Co <u>n</u> solidate				0				_
7052	9								<u>G</u> roup and Outline	•						-	_
	10							13	PivotTable and PivotChart Report					-		-	_
lser ID:	11			1			1		Import External Data	• 🚡	Import Data.						-
biadinin	12						1		List	10000000			2				
assword:	13						_			•	New Web Qu		-				
	14		- 1			- 6	Data Co	onnec	ion Wizard		a	e Query			-	-	_
•••••	15								tabase and Table		×		-				_
hield	17						Sele	ct the	Database and Table/Cube which contains the data yo	u want.		roperties	-			-	_
· · · · · · · · · · · · · · · · · · ·	10						-										-
Data Connection Wizard							Select		abase that contains the data you want:								
									a specific table:								
Connect to ODBC Data So							Name			Modified Crea	ted Type *						
Choose the ODBC data source y	ou wani	t to conn	ect to.						ame_lookup gosalesdw		TABLE						
								_regio	n_dim gosalesdw action_dim gosalesdw		TABLE TABLE						
ODBC data sources:							💷 ga	_time_	dim gosalesdw		TABLE						
BI_DSN_1				*				edia_te	quarter_lookup gosalesdw st gosalesdw		TABLE TABLE						
dBASE Files								edia_te	st_a gosalesdw		TABLE						
Excel Files MS Access Database								edia_ts	AND		TABLE TABLE						
myBigSQL		_	8					rk_acu rk_adv			TABLE						
MyDSN55			6				💷 m	rk_bun	fle_group_lookup gosalesdw		TABLE					_	100
									paign_lookup gosalesdw _survey_targ_fact gosalesdw		TABLE			1	-		-
									uct_survey_dim gosalesdw		TABLE E						
							🛄 m	rk_pro	uct_survey_fact gosalesdw		TABLE					1.00	
									notion_dim gosalesdw		TABLE TABLE						1
									notion_fact gosalesdw notion_plan_fact gosalesdw		TABLE						
							0 m	rk_rtl_:	urvey_dim gosalesdw		TABLE						
									urvey_fact gosalesdw		TABLE						-
1				*				rk_rtl_: vprod	urvey_targ_fact gosalesdw prand gosalesdw		TABLE TABLE						-
									orand1 gosalesdw		TABLE *					_	-
	Car	ncel	< E	Back	Ne	ext >					•						



A word about . . . SerDes

Custom serializers / deserializers (SerDes)

- Read / write complex or "unusual" data formats (e.g., JSON)
- Commonly used with Hive, HBase
- Developed by user or available from open source community

Using SerDes with Big SQL

- Add the SerDe .jar file to \$BIGSQL_HOME/userlib and \$HIVE_HOME/lib
- Stop / restart Big SQL service
- Specify SerDe class name (not .jar file name) when creating table

Example

/* Create a table for JSON data. Use open source hive-json-serde-0.2.jar SerDe */
create table socialmedia-json (Country String, FeedInfo String, . . .)
row format serde 'org.apache.hadoop.hive.contrib.serde2.JsonSerde'
stored as textfile;

load hive data inpath '</hdfs_path>/WatsonBlogsData.json' overwrite into table
 socialmedia-json;

sélect * from socialmedia-json;



Sample JSON input for previous example

[biadmin@bdvm327 twitter]\$ cat WatsonNewsBlogsData.json|more

[{"PostSize":5775, "ThreadId":"4f129a8be", "Crawled":"2012-01-15 09:21:15", "FeedInfo":"{\"Tit le\":\"www.ibm.comnews\", \"Id\":\"44032787\", \"ExtKey\":\"879cd3257c296614160914c3d96f9b85\ ",\"Url\":\"http://www-03.ibm.com\"}", "Published":"2012-01-15 09:21:15", "Url":"http://www.i bm.com/innovation/us/watson/?lnk=ftpl", "Country":"US", "SubjectHtml":"<Keyword>IEM<\/Keyword > - <Keyword>Watson<\/Keyword>", "Inserted":"2012-05-29 00:52:57", "Language":"English", "Text Html":"<![CDATA[<Keyword>IBM<\/Keyword> - <Keyword>Watson<\/Keyword>\n\n Call to find out how Watson's capabilities could benefit your business.\n 1-800-426-7630\n \nRel ated content\nDesigning the Computer for a Smarter Planet\nThere\u2019s an enormous amount of science included when <Keyword>Watson<\/Keyword> answers a single Jeopardy! question, ho w does it all work together?\nExplore <Keyword>Watson<\/Keyword>\NBeyond Jeopardy! The Busi ness Implications of <Keyword>Watson<\/Keyword>\n<Keyword>IBM<\/Keyword>\Netword>\Keyword>IBM<\/Keyword>\Netword>\Netword> Keyword>IBM<\/Keyword>Watson<\/Keyword>IBM<\/Keyword>\Netword>\Netword> he Busi ness Implications of <Keyword>Watson<\/Keyword>\n<Keyword>IBM<\/Keyword> Keyword> passed its first test on Jeopardy! in February 2011, but the real test will be i

applying the underlying systems, data management and analytics technology in business and a cross different industries. Watch the webcast now and learn about the present and future bu siness implications of Deep QA and the other technologies behind <Keyword>Watson<\/Keyword> from David Ferrucci and other <Keyword>IBM<\/Keyword> executives.\nRegister now\nBetter Bu

JSON-based social media data to load into Big SQL Table socialmedia-json defined with SerDe



Sample Big SQL query output for JSON data

authorinfo	country	crawled	1	feedinfo	+-	id	inserted	++ published
{"Nick":"" ,"Id":""," Name":""," Url":""}	υs	2012-04-11 03:31:47		www.ibm.co mnews","Id ":"4403278 7","ExtKey ":"879cd32 57c2966141 60914c3d96 f9b85","Ur l":"http:/		31859312 32	2012-05-23 20:18:08	2012-04-11 03:31:47
{"Nick":"" ,"Id":""," Name":""," Url":""}	υs	2012-02-22 23:50:14		/www-03.ib m.com") {"Title":" www.ibm.co mnews","Id ":"4403278 7","ExtKey ":"879cd32 57c2966141 60914c3d96 f9b85","Ur 1":"http:/ /www-03.ib m.com")		32535948 48	2012-05-26	2012-02-22 23:50:14

Sample output: Select * from socialmedia-json



A word about . . . performance

Tuning options

- Table design (e.g., storage formats for Hive, key & column family definitions for HBase)
- Hints in queries
- Hints in table definition
- Secondary indexes (HBase tables only)
- MapReduce job properties
- Query hints provided in comments: /*+ name=value [, ...] +*/

```
select * from foo /*+ accessmode=`local' +*/ where c1 < 1000;</pre>
```

Access mode hint

- Causes query to be executed in the Big SQL server
- HBase indexed queries can return extremely rapidly
- Local access can be forced on for your entire session





Agenda

Big SQL: motivation and architecture

Using Big SQL

- Invocation options
- Creating tables
- Populating tables with data
- Querying data
- Developing applications and working with tools
- And a peek at some additional topics

What RDBMS professionals should know about





Big SQL – what RDBMS experts should know

Big SQL provides industry-standard query support for Hadoop-based storage managers

- Exploits Hadoop environment
- Includes Hadoop-specific extensions
- Introduces Hadoop-specific concepts
- Copes with "unconventional" data structures and formats (e.g., JSON) via SerDes, other features

RDBMS = more than query & storage management

- Transaction management
- Views
- Stored procedures
- INSERT / UPDATE / DELETE
- GRANT / REVOKE
- 3GL language support (e.g., COBOL)
- Rich catalog statistics and decades of cost-based optimization development
- Bottom line: Big SQL provides SQL experts with on-ramp to Hadoop, but doesn't turn Hadoop into one big relational database



Want to learn more?

- Big SQL tutorial (product Information Center)
- Videos, articles, downloads, etc.
 - Technical portal at http://tinyurl.com/biginsights

InfoSphere BigInsights Tutorials

Manage

Within minutes, dive into the world of big data with robust, browser-based control.



Easily develop your first big data application by using the InfoSphere BigInsights Eclipse plugin.



Collect and import data for exploration and analysis that helps you make sense of seemingly unrelated data.

Query

Quickly master the intricacies of SQL queries for Hadoop with IBM Big SQL.



Delve into BigSheets, an intuitive spreadsheet-like tool, to create analytic queries without any previous programming experience.

Extract

Discover the power of Text Analytics by creating extractors to derive valuable insights from text documents.

nsights [•] My profil [•] My courses • Courses • My profil • My p

Navig

My ho Site p

Big Da Learn from the inc		search course You are logged in as Cynthia Sa	racco (Logout)
COURSES Site pages	DOWNLOADS RESOURCES	ELke 2.9	Tweet 1,339
ation III	Courses		
ome pages gs roofile ourses rses rses	Courses in belia (preview) PSQL Access for Hadoop PSpL Access for Hadoop PJava Fundamentals PJava Fundamentals PHadoo Reporting and Analysis PHadoo Reporting and Analysis PSU Boata Amylics - Demos PHadoop Fundamentals I PHadoop and the Amazon Cloud PHadoop and the Amazon Cloud PHadoop and the Amazon Cloud PHadoop and the MSmart/Cloud Enterprise Ptot Analytics Essentials I		0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
ech	 Text Analytics Essentials - Case Study Text Analytics Essentials - Case Study Text Big Date Meetip - Real-time analytics University-sponsored courses DB2 Maccilianeous Collarse allExcand all 	nent Wiki	0 0 0 0 0

BigInsights Technical Enablement Wiki



Get up to speed on InfoSphere BigInsights, IBM's software platform designed to help firms store, manage, and analyze "big data".



BigInsights forum on IBM developerWorks



Big SQL JDBC driver definition (Eclipse)

- A JDBC driver for Big SQL 2.1 is created automatically when BigInsights server is added
- New driver can be added and customized from Preference.
- Window → Preferences → Data Management → Connectivity → Driver Definition → "Add..."

type filter text	Driver Definitions									
 ▷ General ▷ Ant ▷ BigInsights 	Filter: All									
	Name	System Vendor	System Version							
 Data Management Connectivity Database Connectic Driver Definitions 	Hive JDBC Driver 0.9.0 IBM Big SQL JDBC Driver 2.1	Hive JDBC Big SQL JDBC	0.9.0 2.1							
Edit Driver Definition										
- Provide Driver Details	7	uired jars, and								
Provide Driver Details Modify details in the fields below available and applicable property	values.	uired jars, and								



JSqsh – Big SQL's CLI

JSqsh ("jay-skwish" – Java SQL Shell)

- Open source command line JDBC client
 - (http://jsqsh.wiki.sourceforge.net)
- Works with any JDBC driver, not just Big SQL

It can be started with

\$ \$BIGSQL HOME/bin/jsqsh --driver=bigsql --user=biadmin --password=biadmin JSqsh Release 1.5-ibm, Copyright (C) 2007-2013, Scott C. Gray Type \help for available help topics. Using JLine. [localhost][biadmin] 1> select * from syscat.tables; tablename schemaname columns syscat syscat tables syscat schemas syscat indexcolumns dual system integers system



BigInsights Web Console

In Quick Links, select to run Big SQL queries from the console



Type in query, or cut and paste from SQL script. Hit Run.

BM InfoSphere BigInsights Enterprise Edition — Big SQL
<pre>SELECT SUM(L_EXTENDEDPRICE*L_DISCOUNT) AS REVENUE FROM tpch_hive.LINEITEM WHERE L_SHIPDA SELECT SUM(L_EXTENDEDPRICE*L_DISCOUNT) AS REVENUE FROM tpch_hive.LINEITEM WHERE L_SHIPDATE >= '1994-01-01' AND L_SHIPDATE < '1995-01-01' AND L_DISCOUNT BETWEEN .06 - 0.01 AND .06 + 0.01 AND L_QUANTITY < 24;</pre>
Run Status Result
Number of results returned: 1 REVENUE 1.2314107801567236E8

Information Management



Tableau: Big SQL integration via ODBC

Connect Using																	
DSN: bigsql_poc	*																
	😥 Tableau - Book2	-															
Driver: IBM BIG SQL ODBC DRIVER	<u>File Data W</u> orksheet Dash <u>b</u> oa	rd <u>A</u> nalysi	is <u>M</u> ap I	F <u>o</u> rmat	Server	Window	<u>H</u> elp										
Connect	♦ ♦ ■ •0 ₫• ₫	ď• 5	• 3 •	P	FL :	@ Abc	,	- N	ormal	-	12.						
Connection Attributes:	Data +	Pages	12 inte		Colum	ns											
Key Value	- 🔂 mk_machine				-	-	-		-								
	Dimensions III 🖉 👻				Rows	C	cur_d	/c_typ_ds	5C								
port 7052 server	🖻 🖩 mk_machine 🔺	Filters						1									
username rapentak	Abc bmc				12112 201	enner -	Drop										
dbname tableau	Abc category_cd				_cur_dv MTG1		Ab	-									
password *******	Abc category_dsc																
	Abc cur_dvc_num	Marks		1	MTG2		Ab										
Schema:	Abc cur_dvc_typ_dsc		0.000 4.020	-	MTG4	he la	Ab										
tableau	Abc cur_lcnse_lvl_nm	Abc Au	utomatic		MTG5	8	Ab	e l									
Table:	Abc cur_sbscr_lvl_dsc	Text∙	(MTG6		Ab										
Single Table I Multiple Tables Custom SQL	Abc dmc	Color-			MTG7		Ab	5									
	Abc iud_flag	Color	_	=	MTG8		Ab										
Table Alias Foreign Key	Abc iud_ts Abc jdlink_is_actv																
mk_machine	Abc jdlink_rgstn_ts																
mk_msrmt_agrgt [mk_machine].[mach_id]	Abc mach_partn_k category_ds																
	mfg_dt Rows																
Add Table Edit Remove																	
	Abs makt made key																
Give the connection a <u>n</u> ame for use in Tableau:	Abc native pin																
mk_machine	Abc pin																
	Measures																
OK Car	□ □ ■ mk_machine									ped_cbp-	tm / category						
	# cur_dvc_typ_i					1/3/2008	1:30:00 AM					1/3/2	009 1:30:00 AM			1/3	3/2010 1:
	# mach id			Machine /	Machine M	lachine Ma	achine Ma	hine Mach	nine Machi	ne Machine	Machine	Machine	Machine Mar	hine Mach	ine Machine	Machine	Mach
		m_loc_long	m_loc_lat			eries 4 se		es 6 serie						es 6 serie			series
	# mk_msrmt_ag																
	=# Number of Recora 111	-78.948237	35.929673			(
	# Measure Values									-			$\mathbf{}$			-	
	# Preusore values									_						-	
								12.11					174				
	112	-77.032000	38.889510														
										-						-	
k a na n								12		C							-
++ableau	9 117	-72.830290	33.379450											C			
t t abieau [,]					1000	6-10-10-10-10-10-10-10-10-10-10-10-10-10-						1.00			1.4		
S O F T W A R E				e													

Information Management

Lotus Symphony: Big SQL integration via JDB

		<u>W</u> indow <u>H</u> elp						
V	🥹 🕒 🖪	New Spreadsheet 3	×					
	é 🗖							
7		\checkmark f $\otimes \Sigma =$	ETHIOPIA				₹.	🚔 My Widgets 🔫
	A	В	С			D	L	🏟 Spreadsheet and Da 📲
	n_nationkey		n_regionkey					
	0	ALGERIA	0	final acco	unts wake quickly.	special reques		
3	1	ARGENTINA	1			tealthily. regular instructions wake carefully blith		
	2	BRAZIL	1		nding pinto beans s			DB Drivers
5	3	CANADA	1		ong the bold request		6	
6	4	EGYPT	4	pending a	ccounts haggle furio	ously. furiously bold accounts detect. platelets a	<u> </u>	
7	5	ETHIOPIA	0	fluffily ruth	nless requests integ	rate fluffily. pending ideas wake blithely acco		
8	6	FRANCE	3	even requ	ests detect near the	e nendin =	\oslash	
9	7	GERMANY	3	blithely ir	0	Modify Aliase:bigsql 🛛 🔿 🖯		
0	8	INDIA	2	ironic pac	Modify Aliase:big		Æ	
1	9	INDONESIA	2	unusual e	Moully Allase.blg	əyi		Data Sources
2	10	IRAN	4	blithely e	Name:	hissel		
3	11	IRAQ	4	express,		bigsql		
4	12	JAPAN	2	blithely fi	Loaded Driver:	Big SQL JDBC Driver		·
.5	13	JORDAN	4	blithe, ex		Lig cyclobo birdi		🗣 🥒 🗈 🙁
.6		KENYA	0	ironic req	URL:	jdbc:bigsql://hdtest102.svl.ibm.com:7052/defai		, TEE
.7	15	MOROCCO	0	ideas acc	Lie er Nemer			bigsql
	16	MOZAMBIQUE	0	ironic cou	User Name:	biadmin		bigadi
.9		PERU	1	final, fina	Password:	***		
0	18	CHINA	2	bold acco	Password.			
21		ROMANIA	3	deposits				
	20	SAUDI ARABIA	4	fluffily fin				
3		VIETNAM	2	doggedly				
24		RUSSIA	3	slowly pe		OK Close Test		
5		UNITED KINGDOM	-	fluffily rec				
26		UNITED STATES	1		gular deposits serve	e furiously blithely regular warthogs! slyly fi		
7				ionarioty re		- tenency with a regime mentiodo, alyry if		
8								
9								
0								< III >
1								
32								
23								Import



© 202013BBM Corporation



More about SerDes

SerDe (Serializer/Deserializer)

- A Hive concept
- A java class responsible for converting a **record** produced by an InputFormat into a Hive **row**, based upon the table definition
- The Hive LazySimpleSerDe
 - Expects records from a TextInputFormat a record is just a single line of text
 - The SerDe parses the line of text using delimiters
 - The table inition indicates with the maximum are too be) converted to which data type

SerDe

- Storage Handler
 - A Hive concept
 - A java class that interacts with an external data source

InputFormat

- Contains an InputFormat and SerDe to communicate data
- Is presented with query projection and predicates to optimize data access
 - HBase is currently the only Storage Handler



Hive at a Glance

- Open source data warehouse framework for Hadoop
- Data stored in DFS files, but programmers create / query tables
- Provides SQL-like interface (Hive Query Language, HQL)
 - Language constructs cover a subset of commercial SQL
 - Queries run as MapReduce jobs under the covers
 - Programmers can create custom Mappers, Reducers

From Hive wiki:

Hive is not designed for OLTP workloads and does not offer real-time queries or row-level updates. It is best used for batch jobs over large sets of append-only data (like web logs)





Hive storage

Warehouse directory in DFS

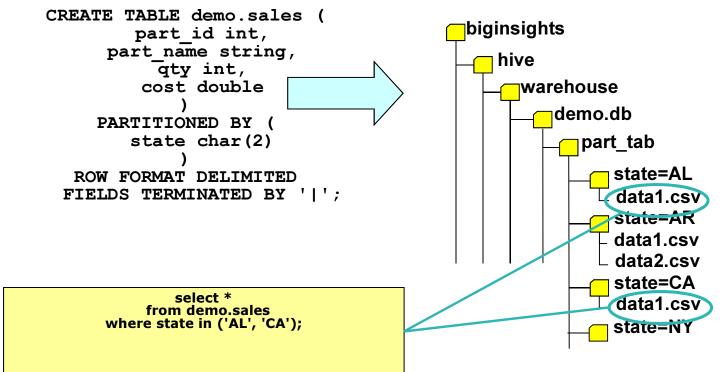
- Specified by "*hive.metastore.warehouse.dir*" in hive-site.xml
- /biginsights/hive/warehouse the default location for BigInsights
- One can think tables, partitions and buckets as directories, subdirectories and files respectively

Hive Entity	Sample	Sample location in DSF					
database	test	\$WH/test.db					
table	Т	\$WH[/test.db]/T					
partition	date='01012013'	\$WH/T/date=01012013					
bucketing column	userid	\$WH/T/date=01012013/000000_0 \$WH/T/date=01012013/000032_0					



Partitioned Tables

- All tables except HBase can be partitioned
- Partitioning is on one or more columns
- Each unique value becomes a partition
- Query predicates can be used to eliminated scanned partitions







Why Another Data Warehousing System? A perspective from Facebook

Problem: Data, data and more data

- 200GB per day in March 2008
- 2+TB(compressed) raw data per day today

The Hadoop Experiment

- Superior to availability/scalability/manageability of commercial DBs
- Efficiency not that great, but throw more hardware
- Partial Availability/resilience/scale more important than ACID

Problem: Programmability and Metadata

- Map-reduce hard to program (users know sql/bash/python)
- Need to publish data in well known schemas
- Solution: HIVE



Excerpt from 2008 presentation by Facebook



HBase at a Glance

Open source key-value data store for Hadoop

- Based on Google's Bigtable paper [2006]
- Implemented as a sparse, consistent, distributed, multi-dimensional, persistent, sorted map
- Key and value are byte arrays

Strengths

- Efficient read/write access using row key, small range scan
- Very good for "sparse data" (no fixed columns)
- Highly scalable
- Rich set of Java APIs and extensible frameworks

HBASE

Different from relational databases

- No types: all data is stored as bytes
- No schema: Rows can have different set of columns
- No native SQL support
- No multi-row transactions
- ⁶⁷ Not optimized for N-way joins scanning large portions of data²Set^{M Corporation}



HBase Data Model

• Table	HBTABLE					
 Contains column-families 	Row key	Value				
 Column family – – – – – – – – – – – – – – – – – – –	11111	cf_data: {'cq_name': 'name1', 'cq_val': 1111} cf_info:				
 Column – – – – – – – – – – – – – – – – – – –		{ 'cq_desc' : 'desc11111' }				
 Can have multiple versions Each row can have different set of columns Each column identified by it's key 	22222	cf_data: {'cq_name': 'name2', 'cq_val': 2013 @ ts = 2013, 'cq_val': 2012 @ ts = 2012				
Row key		5				

- Implicit primary key
- Used for storing ordered rows
- Efficient queries using row key

HFile 11111 cf_data cq_name name1 @ ts1 11111 cf_data cq_val 1111 @ ts1 22222 cf_data cq_name name2 @ ts1 22222 cf_data cq_val 2013 @ ts1 22222 cf_data cq_val 2012 @ ts 2

HFile

11111 cf_info cq_desc desc11111 @ ts1



HBase Support

- Robust HBase support is a major Big SQL focus
- HBase is different than most other data sources in Hadoop
 - Client/server database
 - Fetching rows/columns requires a network hop
 - Efficiently querying HBase requires pushing as much to the server(s) as possible
 - Pushing down query predicates as filters to region servers
 - Fetching only columns needed by the query
 - All HBase tables are ordered and accessed by primary key
 - Big SQL leverages this

Hive



Creating HBase Tables

- Hive syntax for HBase tables is cumbersome
 - Difficult to read
 - Cannot express composite keys and columns
- Big SQL provides explicit syntax for defining tables in

create hbase table sales (prod id int not null, sales date int not null, quantity int not null, double not null price column mapping key mapped by (prod id, sales date) encoding binary, cf1:sales data mapped by (quantity, price) separator '|' encoding string column family options cfl compression gz, bloom filter none, in memory hbase table name 'PROD SALES' default encoding binary default column family options compression none



HBase Hints

rowcachesize (default=2000)

- Used as scan cache setting
- Also used to determine number of get requests to batch in index lookups

colbatchsize (default=100)

useindex ('false' to avoid index usage)

```
select o_orderkey from orders /*+ rowcachesize=10000 +*/ where o_custkey>5000

o Log discard
1450136 rows in results(first row: 22.67s; total: 27.46s)
```

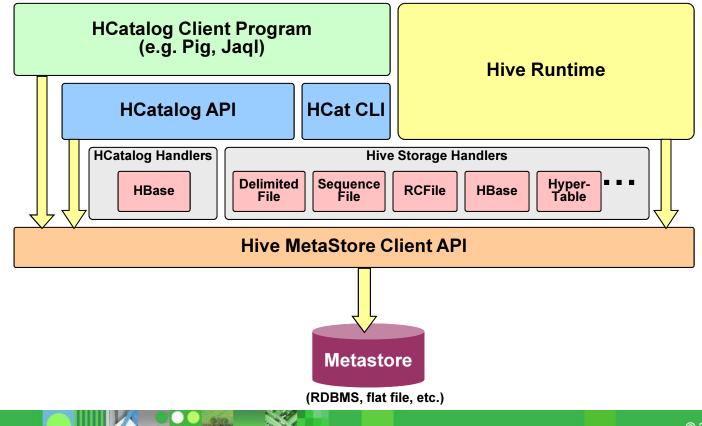
HBase scan details:{..., caching=10000, ...}

- rowcachesize can also be set using the set command:
 - set hbase.client.scanner.caching=10000;



HCatalog

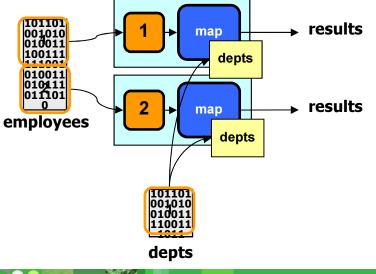
- API's to read/write data directly from Hive "tables"
- Tools for manipulating metadata (create/drop/alter tables)
- Provide enhanced HBase support for use cases beyond Hive





Dynamic Query Optimization

- During query execution Big SQL dynamically re-evaluates its options SELECT c1, c2 FROM T1
 - Queries that cannot be assisted by MapReduce run in the server
 - If a given step (e.g. GROUP or SORT) involves only "small" data, the step is executed in the server
 - If all tables are small, the whole query will be run in-server
 - If one table is large and one or more are small, a memory (hash) join is performed





Performance and Tuning

- Big SQL will dynamically adjust query strategies
 - Execute steps in the server
 - Automatically choose memory joins vs. redistribution join
- Server config settings can adjust how decisions are made
 - And can be adjust at the session level using ${\tt SET}$ command
- Query join order (in the FROM clause) can impact performance
 - Order you provide is honored
 - Most selective data sets should be first
- Query hints and be used to fine-tune performance
 - Table hints can adjust join strategy for specific tables
 - Table access hints and fine-tune access to specific data sources (e.g. HBase row fetch sizes)
 - Specific joins can be forced to execute locally (in-server) vs.
 MapReduce



Developing, publishing and deploying your first Big SQL application with InfoSphere BigInsights

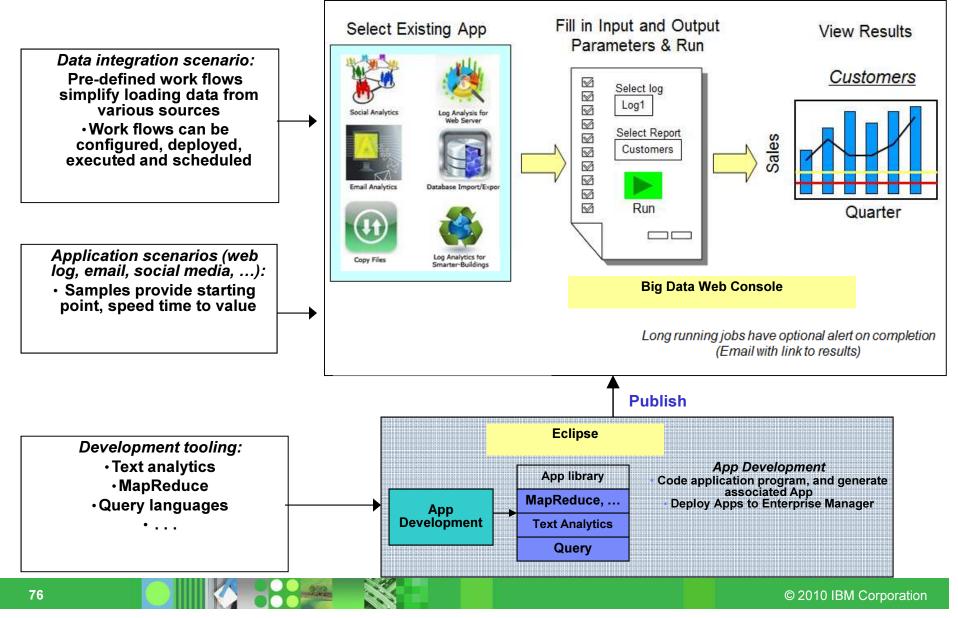
< insert your name here>

February 8, 2014

© 2010 IBM Corporation



Big Data Application Ecosystem





BigInsights Applications Catalog (Web Console)

- Browse available applications
- Manage and deploy applications (administrators only)
- Execute (or schedule execution of) a deployed application
- Monitor job (application) status
- Link or chain applications for sequential execution

and I Execute	Link							
nage Execute	LIDK	1						
plications			Name: Boardreader					
arch	E ∂		Description					
æ			The Boardreader application querie	as the web and retrieves information	on, based on user-specified parameters.			
and the second								
	a	- Execution						
		and the second s	IBM Watson Jan-Jun 2012					
oardReader	BoardReader	- Parameters	IBM Watson Jan-Jun 2012	l.				
gs Ingestion	Boards Ingestion							
	(2)	* Results path:	/user/bladmin	n/sampleData/IBM_Watson_Jan_J	June_2012	Browse		
TEP .	(A)(A)	* Maximum ma	tches: 10000					
		Application History						
		Status	Execution Name	Progress	Start Time	- Elapsed Time (sec)	Output	Details
and the second sec		No filter applied	0.					
loardreader	Brand	[a mill ran miner obliging a					525	
loardreader	Management		IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	142	D	1000
		and the second se	IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	142	1.2	164
	Management	and the second se	IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	142	L#	100
	Management	and the second se	IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	142		
	Management	and the second se	IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	I note	14	
î Î	Management Finance	and the second se	IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	142	19	
Boardreader	Management	and the second se	IBM Watson Jan-Jun 2012	100%	Oct 15, 2012 11:27:26 AM	I note		L > H



Overview of Application Development Lifecycle

- Configure your Eclipse environment (one-time set up)
- Develop your application using BigInsights tools
- Test your application
- Package and publish your application
- Deploy your application on the cluster

First Steps Image: Complexity of the explore a graphic to learn how InfoSphere BigInsights enables you to accomplish goals. Drill down on roles and tasks to discover how to use the tools most effectively for your needs. Image: Complexity of the explore a graphic to learn how InfoSphere BigInsights enables you to accomplish goals. Drill down on roles and tasks to discover how to use the tools most effectively for your needs. Image: Complexity of the explore a graphic to learn how InfoSphere BigInsights enables you to accomplish goals. Drill down on roles and tasks to discover how to use the tools most effectively for your needs. Image: Complexity of the explore a graphic to a BigInsights server before you can upload data to a cluster, use the BigInsights console, or publish and run an application in a cluster. Image: Complexity of the explore a term of the explore to the ex	Design Develop Develop Develop Develop code to query and analyze text, big data at rest, or streams of big data in motion. Image: Publish and Run Publish your applications from the BigInsights console.
Learn More IBM big data on the Web IBM big data community InfoSphere BigInsights support InfoSphere BigInsights Information Center	Quick Links Open Project Explorer Open the BigInsights console Switch to the BigInsights perspective



Configure your Eclipse environment

- One-time set up
- Download and install BigInsights tools (Eclipse plug-ins)
 - Welcome tab of BigInsights Web console includes pre-req info, download & installation instructions

Quick Li	nks	
-	Access secure cluster servers	^
sol	Run Big SQL Queries	E
	Enable your Eclipse development environment for BigInsights application development	
	Download applications (Eclipse projects)	
	Download the Big SQL Client drivers	-



Develop your application – Big SQL example

Open the BigInsights perspective in Eclipse
Create a BigInsights project

Image: Section of the cools most effectively for your needs. Image: Section of the bigInsights server connection of the bigInsights server before you can cluster, use the BigInsights Administration Console, or papplication in a cluster. Image: Section of the bigInsights project is the bigInsights pr	<u></u>	Learn about BigInsights Interactively explore a graphic to learn how InfoSphere B enables you to accomplish goals. Drill down on roles and	tasks to discover
Create a new BigInsights project Before creating code, create a BigInsights project to con Use default location	ľ	You must connect to a BigInsights server before you can cluster, use the BigInsights Administration Console, or p	Create a new BigInsights project.
		Create a new BigInsights project	

tion - Big SQL example



(cont'd)

Create a BigInsights program with a SQL script (file)

tion - Big SQL example



×

(cont'd)

Create (or select) a Big SQL connection

New Connection Profile	×	New Connection Profile	•
Connection Profile Create a BigSQL JDBC connection profile.		iver and Connection Details r from the drop-down and provide login de n.	atails for
Connection Profile Types: Lype filter text Big SQL JDBC	Drivers: IBM E	Big SQL JDBC Driver Default	• • •
 DB2 for Linux, UNIX, and Windows DB2 for i DB2 for z/OS Derby Flat File Data Source Generic JDBC HSQLDB Hive JDBC Informix Ingres 	Schema: Host: Port number User name: Password: ✓ Save par Connection	biadmin	
Na <u>m</u> e: New Big SQL JDBC <u>D</u> escription (optional):		nen the wizard completes rery time the workbench is started	Test Connection



Develop your application – SQL example (cont'd)

Populate your SQL file with the desired code

create table if not exists media_csv (id integer not null, name varchar(50), url varchar(50), contactdate string) row format delimited fields terminated by ',' stored as textfile;

load hive data local inpath //home/biadmin/sampleData/DBMS/RDBMS_data.csv

> -- overwrite into table media_csv;



Test your application

Run your application from Eclipse

		🖹 Problems 📮	Console 🔲 SQL	Results 🛿		= × 🖗 🛅 🍰 🏹 🗖 🗖
		Type query exp	pression here			Status
		Status	Operation	Date	Connection Profile	LoadLocal.sql
🖄 Task Launcher for Big Data 🛛 🧊 GetFbMsgs.jaql	I Cu <u>t</u>	🕨 🗸 Succeede	ed LoadLocal.sq	1/21/14 1	1: Big SQL - bivm	create table if not exists media_csv (id integer not null, name varchar(50),
Connection: Big SQL - bivm	<u>C</u> opy <u>P</u> aste					url varchar(50), contactdate string) row format delimited fields terminated by ','
create table if not exists media_csv (id integer not null, name varchar(50), url varchar(50), contactdate string) row format delimited	<u>R</u> un As Debug As <u>P</u> rofile As <u>V</u> alidate					stored as textfile load hive data local inpath '/home/biadmin/sampleData/DBMS/RDBM into table media_csv Total execution time => 14 s: 159 ms
fields terminated by ',' stored as textfile; load hive data local inpath '/home/biadmin overwrite	T <u>e</u> am Comp <u>a</u> re W Rep <u>l</u> ace Wit				•	
into table media_csv;	Preferences	k				
	 Content Ass Format SQL Toggle Com Validate Stat 	ment		Ctrl+Spa Ctrl+Shift Ctr	VINCKITY	
[+]	Use Databa	- Connection				
Problems 🛛 📮 Console	🗐 Run S <u>Q</u> L		maxatt (P-0)	1994	F5	
0 items	📃 Set Statema	nt Terminator				
Description Resource	🗌 Validate Dat	abase Object	References			



Publish your application to the BigInsights catalog

Package and publish your application from Eclipse
Specify application name, workflow requirements, etc.

e	BigInsights Application Publish	e	Bi	ginsights Application Publish	×	
Specify Applie The informatio	cation n specified here will be saved in the application.xml file.	Application Type Select the type of application	n to be published.			
Location 🖒 🖡	Application 🔿 Type 🔿 Workflow 🖘 Text Analytics 🕤	Location \Leftrightarrow Application \Leftrightarrow Select the type of application		Text Analytics 🗳 BigSheets 💠 Paramete kaged into the application will depend on th		
Application	Create New Application Provide Contraction	Application Type Workflow BigSheets	Specify Workflow		jhts Application Publish w.xml file.	×
	O Replace Existing Application	 Jaql Module Text Analytics 	_	cation 🗢 Type 🗢 Workflow 🐟 Text A	Analytics 👶 BigSheets 🗢 Parameters 🗢 Publish	
Name:	* BigSQLTestApp		/home/biadmin/	vorkspace/BigSQLTestApp/BIApp/workflov	v/workflow.xml	Browse
Description:	Load data from local file into Big SQL table		③ Create a new s Workflow:	ingle action workflow.xml file.		
lcon:	/home/biadmin/sampleData/DBMS/load-icon.jpeg		Action Type: * Big		Darterander	•
Preview Icon:	🦥		Name script	Value LoadLocal.sql	Variable false	<u>N</u> ew
	Comma separated list of category names. load		credentials_r	rop /user/biadmin/credstore/pri	Vate/mykeys_Bigl.txt faise Biginsignts Application Fublish Zp and Publish Application Specify the application zp file content: Rabish the application to Biginsignts.	
?	< Back Next >	Cancel	Finish		Location © Application © Type © Woodflow © Tool Analysis © BigBheets © Parame Zp Preview: © Billepp 320 © Bil	eters ∲ Putson <u>Add.</u> Berrove Coarge form.
U					Bource:	Drado jar. Drado jar.



Deploy your application on the cluster

- Access the Applications tab of the Web console
- "Manage" the published applications
- Locate your new application and deploy it
- Create credentials store file in DFS (if needed) -- see next chart
- Optionally, execute the application after it's been deployed

M InfoSphere	BigInsights Quick	Start Editi	ion (for Non-Production Environr	nent)	About Information C	enter 📘
/elcome Das	hboard Cluster Statu	s Files	Applications Application 5	tatus BigSheets		
n <u>Manage</u> Lin Categories	nk	Aparlineti	ons			
Search	€ 2	Q Depl	oy 🔀 Delete			
Export		lcon	Application Name	Categories	Status	Created
ቿ 🛅 Files ቿ 🛅 Import		3	BigSQLTestApp	load	NOT DEPLOYED	biadmi
- 🕞 load		-				
BigSQLT	estApp	=				
E 🗀 Query						
🗄 🧰 R		0				
🗄 🛅 SQL		-	'n	ĥ.		Þ
🕂 🧰 Test		- 1 - 1 of 1 i	tems 10.1	25 50 100 All	H 4 1	5 H T

Information Management



Simple credential store properties file

6	biadmin			🍃 bigsql.pro	p (~/sampleData/DBI	MS) 🗆 🗙
<u>F</u> ile <u>E</u> dit <u>V</u> iew	<u>P</u> laces <u>H</u> elp			<u>File Edit Vie</u>	w <u>S</u> earch <u>T</u> ools <u>D</u> oo	cuments <u>H</u> elp
	sampleData dit <u>V</u> iew <u>P</u> laces <u>H</u> elp	~		New Open		🥱 🥱 🖌 Indo Redo
sa	DBM File Edit View Places Help	5	×	server=lo	nin [encode]biadmin	
	No templates installed	Create <u>F</u> older Create <u>D</u> ocumer		port=7052		
		Arrange Items Clean <u>U</u> p by Nar	⊕ ⊕	🏄 💥 🔤 🛃 2	Path: //user/biadmin/credstore/private	e/bigsql.prop
() sam		Paste	metatracker for oozie for sheets		bigsql.prop	137 B 128.0 MB
<mark>i iadmin ▼</mark> "sa	DBMS 0 items, Free space: 15.5	Soom <u>O</u> ut Normal Si <u>z</u> e	 		HigInsights Credential Store #BigInsights Credential Store #Tue Jan 21 14:16:16 EST 2014 port=7052 user=biadmin password=[base64]PTY+0zI2MQ\:	e file 4
			applications for the second	Output	server=localhost	



Run your Big SQL application on the cluster

Applications		BigSQLTestApp				-		
earch 12 33 2	• Execution	Description Load data from local file into Big				=		
BigSQLTestApp Database Import	Execution Name:	Default 🛛 🕨 Ru	n					
	Application History Status	Execution Name	Progress	↓ Start Time	Elapsed Time (sec)	Details		
	🗱 🔝 No filter applied							
Distributed File Copy		Test	100%	2014 1 21 12:18	58			
	1 - 1 of 1 items	Welcome Dashb	oard Cluster Status	Files	Applications	Application Status	s BigSheets	
		🗈 🕂 🖾 🍋	@ 🥴 🗙 😼 💹	2	Path: /biginsights/hive	e/warehouse/media_	_csv/RDBMS_data,/	CSV
		HDFS			Name		Size	Block Size
		🖃 🧽 hdfs://bivm:9000)/		RDBMS_da	ata.csv	606 B	128.0 MB
		🖃 🗁 hive			Edit Viewing Size	: 10KB 👻 🍥 Tex	t 💿 Sheet	



Upgrade your application (optional)

- Satisfy evolving business requirements, improve flexibility
 - Example: add input parm(s)
- Modify code and re-package, re-publish, re-deploy

create table if not exists \$TABLE (id integer not null, name varchar(50), url varchar(50), contactdate string) row format delimited fields terminated by ',' stored as textfile;

load hive data local inpath '/home/biadmin/sampleData/DBMS/RDBMS_data.csv' -- overwrite into table \$TABLE;



Upgrade your application (con'td)

Re-publish your application from Eclipse Adjust workflow specs for input parm

	G	BigInsights Application Publish X	
The information specified here will be saved in the application xml file.	Application Type Select the type of application to be published.		
Application Create New Application Replace Existing Application Application to replace Name: BigSQLTestApp Description: Big SQL sample application to load data into a table. Uses 1 input p		다 Text Analytics 다 BigSheets 다 Parameters 다 Publish packaged into the application will depend on the type selected.	
Description: Big SQL sample application to load data into a table. Uses 1 input p	 BigSheets Jaql Module Text Analytics 	Eiginsights Application Publish Zip and Publish Application Specify the application zip file content. Publish the application to BigInsights.	×
Icon: //home/biadmin/workspace/BigSQLTestApp/defaultApp_L.png Preview Icon:		Location 🗘 Application 🗢 Type 🗘 Workflow 🗢 Text Analytics 🗢 BigSheets 🗢 Parameters 🍄 Publish Zip Preview:	
The information specified here will be saved in the workflow.xml file.	eets ↔ Parameters ↔ Publish Browse Variable faise t faise faise faise faise	▼ BIApp.zip Add ▼ ≥ application Remove Image: Second Sec	ule
90		Source:	sh



Summary

- Eclipse tools simplify big data application development for BigInsights
 - Wizards
 - Context-sensitive help
 - Oozie workflow generation
 - Built-in test environment
 - Etc.
- Application catalog provides easy way to locate and launch apps of interest
 - Developers use Eclipse tools to package/publish their applications to this catalog
 - Application upgrades easily managed