InfoSphere BigInsights on IBM SmartCloud with Cognos and Hive

Gerhard Wenzel
IT Architect Big Data

Wilfried Hoge
IT Architect Big Data

IBM Information Management Forum, 15. bis 16. April 2013, Berlin
The IT infrastructure

IBM Smart Cloud Enterprise

Master Node

Data Nodes

Big Insights (Hadoop) Cluster

Hive

JDBC

Cognos BI Server

Web Browser

Cognos Framework Manager

Laptop
Deploy BigInsights on IBM SmartCloud Enterprise

• Your own Hadoop cluster on the IBM cloud in less than 45 minutes
• No need for hardware, install, patch, maintain
• Locate your Hadoop cluster in one of IBM Cloud data centers worldwide
• Low hourly charges starting at $0.30/cluster/hour.
• Explore using BigInsights Basic seamlessly transition to BigInsights Enterprise when ready
Cognos BI 10.2 and BigInsights 2.0

- Cognos BI server
  - Explore & Analyze
  - Report & Act

- Cognos Insight

- Hive via JDBC
  - Hive
    - Application (Map-Reduce)
    - Storage (HBase, HDFS)

- BigSheets
  - Text Analytics
  - REST API

- InfoSphere BigInsights
Using JDBC to access BigInsights from Cognos

• Copy JDBC driver to Cognos BI server
  - hadoop-core-1.0.3.jar
  - hive-exec-0.9.0.jar
  - hive-jdbc-0.9.0.jar
  - hive-metastore-0.9.0.jar
  - hive-service-0.9.0.jar
  - libfb303.jar
  - libthrift.jar

• Configure JDBC connection to BigInsights master node

Verbindungszeichenfolge bearbeiten - JDBC

Bearbeiten Sie die Parameter zur Erstellung einer Verbindungszeichenfolge für: JDBC.

<table>
<thead>
<tr>
<th>Typ:</th>
<th>IBM InfoSphere BigInsights (Hive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JDBC-URL:</td>
<td>//192.168.102.81:10000/default</td>
</tr>
</tbody>
</table>

Verbindungseigenschaften:

| Treiberklasse: | org.apache.hadoop.hive.jdbc.HiveDriver |

Verbindungszeichenfolge:

;LOCAL;JDBC;URL=jdbc:hive://192.168.102.81:10000/default;DRIVER_NAME=org.apache.hadoop.hive.jdbc.HiveDriver;
Using Hive to access CSV files with SQL

• Copy CSV files to HDFS
  - hadoop fs -mkdir /user/biadmin/BTS/D_AIRPORTS
  - hadoop fs -copyFromLocal $i /user/biadmin/BTS/D_Airports.csv

• Define Table in Hive
  - CREATE EXTERNAL TABLE D_Airports(
    Airport_Code string,
    Airport_Name string,
    Airport_Country string,
    Airport_State_Region string,
    Airport_Market string
  ) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\';
  LOCATION ' /user/biadmin/BTS/D_Airports';
What is Big SQL?

• SQL access to all data stored in BigInsights
• Via JDBC/ODBC
• Using rich standard SQL
• Leveraging Map/Reduce parallelism OR achieving low-latency
Big SQL Architecture

- Big SQL shares catalogs with Hive via the Hive metastore
  - Each can query the others tables

- SQL engine analyzes incoming queries
  - Separates portion(s) to execute at the server vs. portion(s) to execute on the cluster
  - Re-writes query if necessary for improved performance
  - Determines appropriate storage handler for data
  - Produces execution plan
  - Executes and coordinates query
Cognos BI 10.2 + and BigInsights 2.1

- Cognos BI server can push down many computations to BigInsights
  - Big SQL directs this processing to happen on BigInsights instead of the Cognos BI Server

- Faster response times
  - Increased opportunity for query processing to occur closer to the data

- Free from the limitations of Hive (latency, SQL language support)
Where to start with BigInsights?

• Learn it at BigDataUniversity.com

• Try it on Smart Cloud Enterprise: ibm.biz/Bdx8FF

• Read about it in “Harness the Power of Big Data” at ibm.biz/Bdx8RP

• Learn about Big Data at www.ibmbigdatahub.com

• Register for “Big Data at the speed of business” event on April 30th at ibm.co/bigdataevent

• Try BigSQL: bigsql.imdemocloud.com

• YouTube Videos - Big Data Channel: youtube.com/user/ibmbigdata
IBM SmartCloud Enterprise

• Get information about IBM SmartCloud Enterprise

• Read all about the IBM SmartCloud

• YouTube Videos – IBM Cloud Computing
Please Note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion.

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.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.