BI 4.0 on Apache Hadoop Hive

Marc Daniau  marc.daniau@sap.com
Introducing Apache Hadoop and Hive

- **Hadoop**
  a framework for storing and processing petabytes of data

- **Hive**
  a data warehouse based on Hadoop

- **Hive QL**
  a simple language based on SQL
A solution leveraging the BI 4.0 architecture

BI Front-end tools
- Web Intelligence
- Crystal Reports
- Dashboards
- Explorer

Common user experience
- Query Panel

Best access method for each specific data source
- Direct Access
- Universe Access

Data sources
- SAP BW
- SAP HANA
- SAP ERP
- OLAP
- Application Database
- Customer Database
- Excel
- Text
- XML
- Web Service
Here are the client tools that support the Hadoop Universe:

- Web Intelligence
- Crystal Reports Enterprise
- Dashboards (Xcelsius)
- Explorer
Explorer on Hadoop Hive
Demo landscape

Hadoop On-premise

Hadoop in-the-cloud

Information Design Tool

Web Intelligence

Crystal Reports

Dashboards

Explorer
Connecting to Hadoop Hive

- We use a JDBC driver to connect to Hadoop Hive

The driver for Hadoop Hive in-the-cloud using Amazon EMR is planned for a future release.
Prerequisites before connecting to Hive

- You must copy the Hive JAR files under the connection server directory in order to connect to Hive

Setting up a Universe against Hadoop

• A data foundation against a Hive schema

The support of multi-source universe on Hadoop Hive is available in the SP4 release.
Querying Hive data

- The business user can get data out of Hadoop in a non-technical manner using the query panel.
- When the user runs the query, SAP generates a HiveQL statement under the cover and sends it to Hadoop Hive.

```
[Query] script
SELECT
  Table_2.country_name,
  sum(Table_6.quantity_sold),
  sum(Table_6.revenue),
  (sum(Table_6.revenue)) / (sum(Table_6.cost_of_sales)) * 100
FROM
  sdm.calendar Table_1 JOIN sdm.sales Table_6 ON
  (Table_1.time_key=Table_6.time_key)
  JOIN sdm.country Table_2 ON (Table_2.country_id=Table_6.country_id)
WHERE
  Table_1.the_year = '2010'
GROUP BY
  Table_2.country_name
HAVING
  (sum(Table_6.revenue)) / (sum(Table_6.cost_of_sales)) * 100 > 250
```
Querying Hive data

- Hive translates the HiveQL statement into MapReduce tasks.

Hive transmits the HiveQL statement to the Hadoop cluster, which then translates it into MapReduce tasks. Each MapReduce task operates on a subset of the data and processes it independently. The results from all tasks are then combined to produce the final output.
Combining data from Hadoop Hive and SAP HANA

- We loaded actual sales in Hadoop Hive.

```sql
hive> show tables;
OK
calendar
country
product
product_family
region
sales
Time taken: 0.053 seconds
hive> describe sales;
OK
country_id  int
time_key  int
revenue  float
cost_of_sales  float
quantity_sold  float
Time taken: 0.064 seconds
```

```sql
hive> select * from sales limit 27;
OK
2 63 20090101 138.0 56.0 4.0
2 63 20090201 43.0 17.0 4.0
2 63 20090301 56.0 19.0 3.0
2 63 20090401 36.0 12.0 1.0
2 64 20090501 34.0 12.0 2.0
2 65 20090201 67.0 23.0 2.0
2 65 20090301 33.0 11.0 1.0
2 65 20090401 35.0 16.0 2.0
2 65 20090501 67.0 27.0 2.0
2 65 20090601 19.0 9.0 2.0
2 66 20090101 127.0 59.0 5.0
2 66 20090201 43.0 14.0 2.0
2 67 20090101 33.0 15.0 1.0
2 67 20090201 67.0 26.0 2.0
2 67 20090301 77.0 36.0 3.0
2 67 20090401 33.0 16.0 1.0
2 70 20090301 66.0 29.0 3.0
2 70 20090501 24.0 10.0 1.0
2 71 20090401 83.0 27.0 3.0
2 72 20090301 95.0 46.0 3.0
2 72 20090401 46.0 19.0 2.0
2 72 20090501 118.0 53.0 4.0
2 73 20090101 54.0 22.0 4.0
2 73 20090201 80.0 40.0 3.0
2 73 20090601 76.0 30.0 4.0
2 74 20090201 108.0 42.0 3.0
2 74 20090501 80.0 30.0 4.0
Time taken: 0.08 seconds
```
Combining data from Hadoop Hive and SAP HANA

• We loaded planning data in SAP HANA.
• A plan can have multiple versions.
Combining data from Hadoop Hive and SAP HANA

- We compare the actual sales coming from Hadoop Hive against the plan in SAP HANA using Web Intelligence.
Combining data from Hadoop Hive and SAP HANA

• One can refresh the SAP HANA query (#2) with no latency in order to try different planning versions.
Analyzing Text data

- We loaded 3 famous speeches in natural language in Hive.
Analyzing Text data

- We find the most frequent words.
- The extraction and count of words are done by Hadoop Hive.
Analyzing Text data

- We find the most frequent word combinations.
- We must tell Hive how many words we want to combine.

**Speech: John_F_Kennedy(1961-01-20)**

<table>
<thead>
<tr>
<th>Group of words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>[can, do, for]</td>
<td>4</td>
</tr>
<tr>
<td>[Let, both, sides]</td>
<td>4</td>
</tr>
<tr>
<td>[a, call, to]</td>
<td>3</td>
</tr>
<tr>
<td>[of, the, world]</td>
<td>3</td>
</tr>
<tr>
<td>[all, forms, of]</td>
<td>2</td>
</tr>
<tr>
<td>[and, the, weak]</td>
<td>2</td>
</tr>
<tr>
<td>[is, little, we]</td>
<td>2</td>
</tr>
<tr>
<td>[that, we, shall]</td>
<td>2</td>
</tr>
<tr>
<td>[the, instruments, of]</td>
<td>2</td>
</tr>
<tr>
<td>[the, world, ask]</td>
<td>2</td>
</tr>
</tbody>
</table>

Group size is 3

**Speech: Martin_Luther_King(1963-08-28)**

<table>
<thead>
<tr>
<th>Group of words</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>[I, have, a, dream]</td>
<td>8</td>
</tr>
<tr>
<td>[will, be, able, to]</td>
<td>8</td>
</tr>
<tr>
<td>[have, a, dream, that]</td>
<td>6</td>
</tr>
<tr>
<td>[Let, freedom, ring, from]</td>
<td>6</td>
</tr>
<tr>
<td>[a, dream, that, one]</td>
<td>5</td>
</tr>
<tr>
<td>[be, satisfied, as, long]</td>
<td>5</td>
</tr>
<tr>
<td>[dream, that, one, day]</td>
<td>5</td>
</tr>
<tr>
<td>[freedom, ring, from, the]</td>
<td>5</td>
</tr>
<tr>
<td>[satisfied, as, long, as]</td>
<td>5</td>
</tr>
<tr>
<td>[Now, is, the, time]</td>
<td>4</td>
</tr>
</tbody>
</table>

Group size is 4
Statistical Analysis

- We loaded numerical data (Salary, Age, ...) in Hadoop Hive.

<table>
<thead>
<tr>
<th>value</th>
<th>sequence</th>
<th>dataset</th>
<th>individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>420000</td>
<td>425</td>
<td>National League</td>
<td>Guzman, Cristian</td>
</tr>
<tr>
<td>800000</td>
<td>426</td>
<td>Baseball Salaries</td>
<td>Hernandez, Livon</td>
</tr>
<tr>
<td>319000</td>
<td>427</td>
<td>(2005)</td>
<td>Horgen, Joe</td>
</tr>
<tr>
<td>145000</td>
<td>428</td>
<td>National League</td>
<td>Johnson, Nick</td>
</tr>
<tr>
<td>290000</td>
<td>429</td>
<td>Baseball Salaries</td>
<td>Lozito, Esteben</td>
</tr>
<tr>
<td>318000</td>
<td>430</td>
<td>(2005)</td>
<td>Metko, Henry</td>
</tr>
<tr>
<td>275000</td>
<td>431</td>
<td>National League</td>
<td>Ohka, Tomo</td>
</tr>
<tr>
<td>800000</td>
<td>432</td>
<td>Baseball Salaries</td>
<td>Ocuna, Antonio</td>
</tr>
<tr>
<td>322500</td>
<td>433</td>
<td>(2005)</td>
<td>Patterson, John</td>
</tr>
<tr>
<td>200000</td>
<td>434</td>
<td>National League</td>
<td>Schneider, Brian</td>
</tr>
<tr>
<td>345000</td>
<td>435</td>
<td>Baseball Salaries</td>
<td>Sledge, Terrel</td>
</tr>
<tr>
<td>657000</td>
<td>436</td>
<td>(2005)</td>
<td>Tucker, TJ</td>
</tr>
<tr>
<td>325000</td>
<td>437</td>
<td>National League</td>
<td>Vergas, Claudio</td>
</tr>
<tr>
<td>700000</td>
<td>438</td>
<td>Baseball Salaries</td>
<td>Vidro, Jose</td>
</tr>
<tr>
<td>305000</td>
<td>439</td>
<td>(2005)</td>
<td>Wilkerson, Brad</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Washington</td>
</tr>
<tr>
<td>61</td>
<td>2</td>
<td>Age of U.S. presidents at inauguration</td>
<td>J.Adams</td>
</tr>
<tr>
<td>57</td>
<td>3</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Jefferson</td>
</tr>
<tr>
<td>57</td>
<td>4</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Madison</td>
</tr>
<tr>
<td>58</td>
<td>5</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Monroe</td>
</tr>
<tr>
<td>57</td>
<td>6</td>
<td>Age of U.S. presidents at inauguration</td>
<td>J.Q.Adams</td>
</tr>
<tr>
<td>61</td>
<td>7</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Jackson</td>
</tr>
<tr>
<td>54</td>
<td>8</td>
<td>Age of U.S. presidents at inauguration</td>
<td>VanBuren</td>
</tr>
<tr>
<td>68</td>
<td>9</td>
<td>Age of U.S. presidents at inauguration</td>
<td>W.Harrison</td>
</tr>
<tr>
<td>51</td>
<td>10</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Tyler</td>
</tr>
<tr>
<td>49</td>
<td>11</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Polk</td>
</tr>
<tr>
<td>64</td>
<td>12</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Taylor</td>
</tr>
<tr>
<td>50</td>
<td>13</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Fillmore</td>
</tr>
<tr>
<td>48</td>
<td>14</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Pierce</td>
</tr>
<tr>
<td>65</td>
<td>15</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Buchanan</td>
</tr>
<tr>
<td>52</td>
<td>16</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Lincoln</td>
</tr>
<tr>
<td>56</td>
<td>17</td>
<td>Age of U.S. presidents at inauguration</td>
<td>A.Johnson</td>
</tr>
<tr>
<td>46</td>
<td>18</td>
<td>Age of U.S. presidents at inauguration</td>
<td>Grant</td>
</tr>
</tbody>
</table>
Statistical Analysis

- We discover the data distribution.
- The bins definition and frequency estimation are done by Hive.
Statistical Analysis

- We summarize the data using descriptive statistics.

```
<table>
<thead>
<tr>
<th>Summary</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set size</td>
<td>Minimum: 300,000</td>
</tr>
<tr>
<td>Null values</td>
<td>5th percentile: 316,000</td>
</tr>
<tr>
<td>Central tendency</td>
<td>Quartile 1: 338,750</td>
</tr>
<tr>
<td>Mean: 2,585,804.41</td>
<td>Quartile 3: 3,250,000</td>
</tr>
<tr>
<td>Median: 800,000</td>
<td>95th percentile: 10,108,333</td>
</tr>
<tr>
<td>Spread</td>
<td>Maximum: 22,000,000</td>
</tr>
<tr>
<td>Variance: 12,254,957,460,152.1</td>
<td>Shape</td>
</tr>
<tr>
<td>Range: 21,700,000</td>
<td>Skew: 2.1395</td>
</tr>
<tr>
<td>Standard deviation: 3,500,708.14</td>
<td>Kurtosis: 5.0402</td>
</tr>
<tr>
<td>Interquartile range: 2,911,250</td>
<td></td>
</tr>
</tbody>
</table>
```
We aggregate the data over-time in an ad-hoc manner.
Key Learnings

• We saw how a Designer can define a connection and prepare a business layer against Hadoop Hive using the information design tool version 4.0 Support Pack 4

• We saw how a Business User can define a query and run it against Hadoop Hive via a BusinessObjects Universe

• We saw how a WebI User can combine data coming from Hadoop Hive with data coming from SAP HANA

• We saw examples of text analysis and statistical analysis performed on Hadoop Hive using Web Intelligence
Thank you for participating.

Please provide feedback on this session by completing a short survey via the event mobile application.

SESSION CODE: 1210

Learn more year-round at www.asug.com
No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

Microsoft, Windows, Excel, Outlook, and PowerPoint are registered trademarks of Microsoft Corporation.

IBM, DB2, DB2 Universal Database, System i, System i5, System p, System p5, System x, System z, System z10, System z9, z10, z9, iSeries, pSeries, xSeries, zSeries, eServer, z/VM, z/OS, i5/OS, S/390, OS/390, AS/400, S390 Parallel Enterprise Server, PowerVM, Power Architecture, POWER6+, POWER6, POWER5+, POWER5, POWER, OpenPower, PowerPC, BatchPipes, BladeCenter, System Storage, GPFS, HACMP, RETAIN, DB2 Connect, RACF, Redbooks, OS/2, Parallel Sysplex, MVS/ESA, AIX, Intelligent Miner, WebSphere, Netfinity, Tivoli and Informix are trademarks or registered trademarks of IBM Corporation.

Linux is the registered trademark of Linus Torvalds in the U.S. and other countries.

Adobe, the Adobe logo, Acrobat, PostScript, and Reader are either trademarks or registered trademarks of Adobe Systems Incorporated in the United States and/or other countries.

Oracle and Java are registered trademarks of Oracle and/or its affiliates.

UNIX, X/Open, OSF/1, and Motif are registered trademarks of the Open Group.

Citrix, ICA, Program Neighborhood, MetaFrame, WinFrame, VideoFrame, and MultiWin are trademarks or registered trademarks of Citrix Systems, Inc.

HTML, XML, XHTML and W3C are trademarks or registered trademarks of W3C®, World Wide Web Consortium, Massachusetts Institute of Technology.

SAP, R/3, SAP NetWeaver, Duet, PartnerEdge, ByDesign, SAP BusinessObjects Explorer, StreamWork, and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries.

Business Objects and the Business Objects logo, BusinessObjects, Crystal Reports, Crystal Decisions, Web Intelligence, Xcelsius, and other Business Objects products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Business Objects Software Ltd. Business Objects is an SAP company.

Sybase and Adaptive Server, iAnywhere, Sybase 365, SQL Anywhere, and other Sybase products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of Sybase, Inc. Sybase is an SAP company.

All other product and service names mentioned are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

The information in this document is proprietary to SAP. No part of this document may be reproduced, copied, or transmitted in any form or for any purpose without the express prior written permission of SAP AG.