Tapping Into Hadoop and NoSQL Data Sources with MicroStrategy

Presented by: Jeffrey Zhang and Trishla Maru
Agenda

- Big Data Overview

- All About Hadoop
  - What is Hadoop?
  - How does MicroStrategy connects to Hadoop?
  - Customer Case Studies
  - Demo

- MicroStrategy Now Supports MongoDB

- MicroStrategy Consumes Unstructured Data with Text Analytics
  - What is Text Analytics?
  - Use Cases of Text Analytics
  - Introducing New Text Analytics Module
  - Demo

- Q&A
What is Big Data, Really?

The Three Vs of Big Data According to Gartner

**Volume**
- Orders of magnitude bigger than conventional data (Terabytes, Petabytes, Exabytes)
- Cost-prohibitive or practically impossible to store, manage or analyze in typical database software

**Variety**
- Structured, semi-structured, unstructured formats
- Diverse sources - complex event processing, application logs, machine sensors, social media data

**Velocity**
- Speed of ingesting incoming data streams
- Processing and real-time analysis of streaming and complex event data
# Use Case Categories for Big Data

Four broad categories of Big Data sources and their value

<table>
<thead>
<tr>
<th>Traditional sources becoming bigger</th>
<th>Digital exhaust from interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SOURCE</strong> Company, Government, Financial sector, Business and consumer studies, Surveys, Polls</td>
<td><strong>SOURCE</strong> Online click-stream, Application logs, Call/service records, ID scans, Security cameras</td>
</tr>
<tr>
<td><strong>VALUE</strong> All business performance drivers – Operational efficiency, Revenue management, Strategic planning</td>
<td><strong>VALUE</strong> New revenue sources, Consumer promotions, Risk management, Fraud detection</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Web 2.0 phenomenon</th>
<th>Internet of things</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SOURCE</strong> Content generated from social media posts, tweets, blogs, pictures, videos, ratings</td>
<td><strong>SOURCE</strong> Machine generated sensor data and “connected device” communication</td>
</tr>
<tr>
<td><strong>VALUE</strong> Customer engagement, Customer service, Brand management, Viral marketing</td>
<td><strong>VALUE</strong> Operational efficiency, Cost control, Risk avoidance</td>
</tr>
</tbody>
</table>
No Data Left Behind

Optimized access to your entire Big Data ecosystem as if it were a single database

- MapReduce & NoSQL Databases
  - cloudera
  - Hortonworks
  - MAPR
  - mongoDB
  - Amazon Web Services
  - Elastic Map Reduce
  - IBM
  - BigInsights
  - Teradata
  - Aster
  - Intel
  - Distribution

- Columnar Databases
  - Amazon Redshift
  - Sybase IQ
  - Vertica
  - PARACCEL
  - Qumulo
  - Infobright
  - Exasol
  - Google BigQuery

- Data Warehouse Appliances
  - Teradata
  - Netezza
  - Oracle Exadata
  - SAP HANA
  - Microsoft SQL Server 2008 R2 Parallel Data Warehouse
  - Greenplum

- Relational Databases
  - IBM DB2
  - MySQL
  - IBM Informix
  - Oracle
  - Microsoft SQL Server
  - Sybase
  - PostgreSQL
  - MariaDB

- Multidimensional Databases
  - SAP
  - Hyperion
  - Microsoft SQL Server Analysis Services
  - Cognos
  - TM1

- SaaS-Based App Data
  - salesforce
  - connection cloud
  - zendesk
  - Intacct
  - Google
  - elucida

- User / Departmental Data
  - Excel
  - CSV
  - Access
  - TXT
Why Hadoop and No SQL Sources Are Such A Major Force in the Big Data Scene?

**Extreme Scalability and Reliability**
These sources provide scalable and reliable data storage that is designed to span large clusters of commodity servers.

**Affordable Data Storage**
Prior to Hadoop data storage was expensive. And the need to store increasingly large amounts of data and be able to easily get to it for a wide variety of purposes, makes Hadoop special.

**Highly Flexible**
Hadoop and No SQL sources bypass the need to specify a schema/structure the data. Allows to dump the data and ask questions later.
Agenda

• Big Data Overview

• All About Hadoop
  • What is Hadoop?
  • How does MicroStrategy connects to Hadoop?
  • Customer Case Studies
  • Demo

• MicroStrategy Now Supports MongoDB

• MicroStrategy Consumes Unstructured Data with Text Analytics
  • What is Text Analytics?
  • Use Cases of Text Analytics
  • Introducing New Text Analytics Module
  • Demo

• Q&A
What is Hadoop?

A scalable fault-tolerant distributed system for data storage and processing, open sourced under the Apache License.

**Hadoop Distributed File System:** A highly reliable, high-bandwidth clustered storage used for buffering, copying and transferring data.

**Map Reduce:** A software framework for distributed parallel processing of data.
Where does Hadoop fit in the Enterprise?

Hadoop is a complement to a relational data warehouse
Enterprises are generally not replacing their relational EDW with Hadoop

Hadoop’s strengths
• Inexpensive
• High reliability
• Extreme scalability
• Flexibility: Data can be added without defining a schema

Hadoop’s weaknesses
• Hadoop is not an interactive query environment
• Processing data in Hadoop requires writing code
Query Execution Times in an Environment with Hadoop
How does MicroStrategy Integrate with Hadoop?

- MicroStrategy certifies Cloudera Impala, Google Big Query and Pivotal HAWQ as a data source.
- MicroStrategy optimizes and certifies Hadoop/Hive as a data source.
- MicroStrategy also provides a connector to execute Freeform Pig-Latin reports.
Easier Connection to Hadoop with Each MicroStrategy Version

- **MicroStrategy 9.3**
  - Q2 2012
    - Apache ODBC
    - Cloudera ODBC 1.0
    - MicroStrategy Thrift Connector

- **MicroStrategy 9.3.1**
  - Q1 2013
    - MicroStrategy Universal ODBC driver
    - Cloudera ODBC 2.0

- **MicroStrategy Analytics Platform**
  - Q4 2013
    - Hortonworks Simba ODBC driver
    - MSTR Greenplum driver for HAWQ

New
Usage Patterns for MicroStrategy with Hadoop as a Data Source

1. Visually explore subject-matter extract in-memory through a one-time query to Hadoop
2. Self-service parameterized queries directly to Hadoop
3. Model-driven access to Hadoop.
4. Query multi-source schema model and drill down among Intelligent Cubes, EDW, Hive

Maturity of Data Access
Key BI Characteristics:

INDUSTRY: Entertainment
BI COMPONENTS: 1 Application; Traditional Reports
USERS: ~200
DATABASE: Hadoop, Teradata
HADOOP DISTRIBUTION: Amazon EMR
VOLUME OF DATA: Petabytes
TYPE OF DATA: Log and Events data
APPLICATIONS: Sales Analysis

Business Use and Benefits

• Sales Analysis generally with a new launch in new region, quick report analysis to understand the new accounts, number of hours of viewing etc.

• Directly querying and reporting from MicroStrategy on logs via Hive

• Able to make better Sales decisions
Key BI Characteristics:

INDUSTRY: E-commerce
BI COMPONENTS: 1 Application; Reports, Dashboards, VI
USERS ~200
DATABASE: Hadoop, Oracle
HADOOP DISTRIBUTION: Apache
VOLUME OF DATA Petabytes
TYPE OF DATA Web Logs, Online behavior
APPLICATIONS: Sales Analysis

Business Use and Benefits

• Analyzing web logs/online behavior stored in Hadoop. Dashboards and VI analysis run against our in-memory cubes. And ad-hoc reports run live against Hive.

• End users do not need to code with MapReduce

• Developers are more productive delivering self service BI through a tool instead of coding custom user interface.
Key BI Characteristics:

INDUSTRY: Electronics and Media
BI COMPONENTS: 1 Application; Reports, VI, Dashboards
DATABASE: Hadoop, Hive
HADOOP DISTRIBUTION: Cloudera Impala
VOLUME OF DATA: Over 1 Billion traffic attribute combinations
APPLICATIONS: Traffic Attribute Multiplier

Business Use and Benefits

• The Traffic Attribute Multiplier application is helping Adconion to get precisely target their digital ads, shorten the time to prepare and tune models and better ad delivery ROI for their customers

• Leveraging MicroStrategy’s integration to Impala and the rich visualizations library, making it easy to be consumed by business users.

• Achieved 2.4% improvement in ad budgets spending efficiency
Agenda

• Big Data Overview

• All About Hadoop
  • What is Hadoop?
  • How does MicroStrategy connects to Hadoop?
  • Customer Case Studies
  • Demo

• MicroStrategy Now Supports MongoDB

• MicroStrategy Consumes Unstructured Data with Text Analytics
  • What is Text Analytics?
  • Use Cases of Text Analytics
  • Introducing New Text Analytics Module
  • Demo

• Q&A
Demo – Log Analytics on Alert Loyalty Platform

MicroStrategy’s capability to query and analyze web logs stored in HBase.

This application is used internally by MicroStrategy Alert Loyalty Platform

- Flume is configured to automatically stream logs
- Data is further stored in HBase table
- MicroStrategy queries HBase via Impala, using the certified Impala driver
Demo – Log Analytics
MicroStrategy’s capability to query and analyze web logs stored in HBase.

```
| Hbase (main):003:0> scan 'AlertEXAccessLog', LIMIT = 23 |
|-----------------------------|-----------------------------|
| column: Fapi, timestamp: 1389639739005, value: /admin/store/users |
| column: FclientIP, timestamp: 1389639739005, value: 10.26.9.21 |
| column: Fcontentlength, timestamp: 1389639739005, value: 1 |
| column: Freqsize, timestamp: 1389639739005, value: 50 |
| column: Fencoding, timestamp: 1389639739005, value: |
| column: Fmethod, timestamp: 1389639739005, value: GET |
| column: Fparameter, timestamp: 1389639739005, value: store_id=3068324082&segment_id=0 |
| column: Fprotocol, timestamp: 1389639739005, value: |
| column: FserverIP, timestamp: 1389639739005, value: |
| column: Fstatus, timestamp: 1389639739005, value: 200 |
| column: Fthread, timestamp: 1389639739005, value: http-apr-0102-exec-10 |
| column: Ftimestamp, timestamp: 1389639739005, value: 2014-01-12 00:00:01 |
| column: Ftoken, timestamp: 1389639739005, value: |
| column: FUseragent, timestamp: 1389639739005, value: |
| column: Fapi, timestamp: 1389639739005, value: /admin/store/users |
| column: FclientIP, timestamp: 1389639739005, value: 10.26.9.21 |
| column: Fcontentlength, timestamp: 1389639739005, value: 1 |
| column: Freqsize, timestamp: 1389639739005, value: 50 |
| column: Fencoding, timestamp: 1389639739005, value: |
| column: Fmethod, timestamp: 1389639739005, value: GET |
| column: Fparameter, timestamp: 1389639739005, value: store_id=3068324082&segment_id=0 |
| column: Fprotocol, timestamp: 1389639739005, value: |
| column: FserverIP, timestamp: 1389639739005, value: |
| column: Fstatus, timestamp: 1389639739005, value: 200 |
| column: Fthread, timestamp: 1389639739005, value: http-apr-0102-exec-10 |
| column: Ftimestamp, timestamp: 1389639739005, value: 2014-01-12 00:00:01 |
| column: Ftoken, timestamp: 1389639739005, value: |
| column: FUseragent, timestamp: 1389639739005, value: |
```

Data stored in HBase table
Demo – Log Analytics

Performance Dashboard - Overview

Overview

<table>
<thead>
<tr>
<th>API</th>
<th>Count of Executions</th>
<th>Average Process Time (s)</th>
<th>Max Process Time (s)</th>
<th>Min Process Time (s)</th>
<th>Average Data Size (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/account/password/update_token</td>
<td>1</td>
<td>22.06</td>
<td>22.06</td>
<td>22.06</td>
<td>0.11</td>
</tr>
<tr>
<td>/account/create</td>
<td>75,376</td>
<td>7.12</td>
<td>1,794.63</td>
<td>0.00</td>
<td>0.84</td>
</tr>
<tr>
<td>/account/link</td>
<td>3,123</td>
<td>4.49</td>
<td>176.61</td>
<td>0.01</td>
<td>0.20</td>
</tr>
<tr>
<td>/store/store_id/admins</td>
<td>66</td>
<td>3.90</td>
<td>38.73</td>
<td>0.01</td>
<td>2.46</td>
</tr>
<tr>
<td>/user/me/init</td>
<td>480,578</td>
<td>3.49</td>
<td>426.14</td>
<td>0.00</td>
<td>1.23</td>
</tr>
<tr>
<td>/message/message_id</td>
<td>172,532</td>
<td>3.44</td>
<td>1,994.46</td>
<td>0.01</td>
<td>1.54</td>
</tr>
<tr>
<td>/v1/devices_token/registrations/pass.com.str</td>
<td>5</td>
<td>3.19</td>
<td>9.36</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>/category/toc</td>
<td>147</td>
<td>2.64</td>
<td>85.56</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>/account/password/reset</td>
<td>14,022</td>
<td>2.38</td>
<td>88.63</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>/account/profile/picture</td>
<td>52,048</td>
<td>2.37</td>
<td>349.15</td>
<td>0.04</td>
<td>0.18</td>
</tr>
<tr>
<td>/news/news_id/share</td>
<td>1,810</td>
<td>2.26</td>
<td>52.13</td>
<td>0.01</td>
<td>0.21</td>
</tr>
<tr>
<td>/v1/passes/pass.com.strategy.alert/pass_id</td>
<td>5</td>
<td>2.00</td>
<td>7.62</td>
<td>0.19</td>
<td>39.31</td>
</tr>
<tr>
<td>/account/confirm_token</td>
<td>62</td>
<td>1.81</td>
<td>37.03</td>
<td>0.01</td>
<td>0.10</td>
</tr>
<tr>
<td>/user/me/categories</td>
<td>434,563</td>
<td>1.75</td>
<td>98.42</td>
<td>0.00</td>
<td>285.13</td>
</tr>
<tr>
<td>/account/create</td>
<td>10</td>
<td>1.64</td>
<td>5.67</td>
<td>0.01</td>
<td>0.26</td>
</tr>
</tbody>
</table>

API Overview Performance

(choose one from the grid)

Concurrency

Overall System Load and Average Backend Process Time

MicroStrategy World 2014
Demo – Log Analytics
Observed Benefits of using HBase/Impala Solution vs. MySQL

- **Scalability/Reliability**
  - Storing more than 30 days of logs data in MySQL degraded the performance
  - HBase also proved to be more reliable

- **Performance**
  - For Dashboard cube publishing with similar data as in MySQL, Impala was 10x faster!
  - (22 million records, 200s for Impala on 8GB/1 CPU vs. 1900s for MySQL on 64 GB/2 CPU)

- **Affordability**
  - We achieved the same performance with way lesser hardware on HBase/Impala
Full and Flexible Capabilities for Hadoop Analytics

The dominant BI solution for interactive access to Hadoop data

All Leading Distributions

Multiple Connection Options

In-Memory Cubes for Interactive Hadoop Analytics

Certified integration with the most popular Hadoop and MapReduce solutions.

Point-and-click query builder. No need to train analysts in HiveQL or MapReduce.

Transform PBs of slow big data into GBs of “agile-ready” data
Agenda

• Big Data Overview

• All About Hadoop
  • What is Hadoop?
  • How does MicroStrategy connects to Hadoop?
  • Customer Case Studies
  • Demo

• MicroStrategy Now Supports MongoDB

• MicroStrategy Consumes Unstructured Data with Text Analytics
  • What is Text Analytics?
  • Use Cases of Text Analytics
  • Introducing New Text Analytics Module
  • Demo

• Q&A
MicroStrategy Taps into MongoDB

- MongoDB 2.4 is supported as a warehouse with MicroStrategy Analytics Enterprise 9.4.1
- We certify the 32-bit Simba MongoDB ODBC driver on Windows and Linux.
- MicroStrategy can also retrieve data from MongoDB’s REST interface via web services
MicroStrategy Taps into MongoDB
How does Simba ODBC Driver works?

Sends SQL

MicroStrategy Analytics Enterprise

Simba Mongo DB ODBC Driver

Converts to MongoDB query

Converts to tabular data

Mongo Proprietary

Limitation:
• The Simba ODBC driver does not support Temp Table creation at this point.
**MicroStrategy Taps into MongoDB**

**Connectivity Details**

Where can you find the Simba driver?

![ODBC Data Source Administrator](image)

- **MongoDSN**
- **Operational Data Mart**
- **Retail Demo Hive_WH**
- **Retail Demo MD**
- **Salesforce**
- **SAM_WH_AC**

**Conversion from Unstructured data to tables/columns**

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Source Name</th>
<th>Source Type</th>
<th>Hide Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>_id</td>
<td>SQL_DOUBLE</td>
<td>_id</td>
<td>NumberDouble</td>
<td>Show</td>
</tr>
<tr>
<td>awards</td>
<td>SQL_VARCHAR</td>
<td>awards</td>
<td>Array</td>
<td>Hide</td>
</tr>
<tr>
<td>awards_0</td>
<td>SQL_VARCHAR</td>
<td>awards.0</td>
<td>Object</td>
<td>Hide</td>
</tr>
<tr>
<td>awards_0_award</td>
<td>SQL_VARCHAR</td>
<td>awards.0.award</td>
<td>String</td>
<td>Show</td>
</tr>
<tr>
<td>awards_0_by</td>
<td>SQL_VARCHAR</td>
<td>awards.0.by</td>
<td>String</td>
<td>Show</td>
</tr>
<tr>
<td>awards_0_year</td>
<td>SQL_DOUBLE</td>
<td>awards.0.year</td>
<td>NumberDouble</td>
<td>Show</td>
</tr>
<tr>
<td>awards_1</td>
<td>SQL_VARCHAR</td>
<td>awards.1</td>
<td>Object</td>
<td>Hide</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><code>awards_1_year</code></th>
<th><code>awards_2_award</code></th>
<th><code>awards_2_by</code></th>
<th><code>awards_2_year</code></th>
<th><code>awards_3_award</code></th>
<th><code>awards_3_by</code></th>
</tr>
</thead>
</table>
Agenda

• Big Data Overview

• All About Hadoop
  • What is Hadoop?
  • How does MicroStrategy connects to Hadoop?
  • Customer Case Studies
  • Demo

• MicroStrategy Now Supports MongoDB

• MicroStrategy Consumes Unstructured Data with Text Analytics
  • What is Text Analytics?
  • Use Cases of Text Analytics
  • Introducing New Text Analytics Module
  • Demo

• Q&A
What is Text Analytics?

Text Analytics

Text Mining

Unstructured Data

Structured Data

- Process of deriving high quality information from text, using linguistic, statistical, and machine learning algorithms

- Often heard together with “Big Data”, since 70~80% of big data information originates in unstructured text.
Use Cases of Text Analytics

Security Applications

Sentiment Analysis

Media Monitoring

Customer Satisfaction

Market Intelligence

Product Reviews

Brand Management

Medical Applications
Open the Gates for New Insights from Text Analytics
Bring unstructured text into MicroStrategy and get insights!

MicroStrategy Text Analytics Module

- Introducing MicroStrategy’s New Text Analytics Module that translates unstructured data into consumable csv/tabular data.
- It is a standalone tool and you will hear more about it in our Futures session!
MicroStrategy Text Analytics Module Workflow

Raw Text Data File(s)
Posts Comments Reviews Survey Responses etc.

Text Processing
MicroStrategy Text Analytics Module
Tunable with domain-specific input

Reporting Data
Topics Sports, Weather...
Sentiment Positive/Negative
Entities Company, Product, Place...
Phrases and Words

Reports & Dashboards
How Text Analytics Module Works?

INPUT
Unstructured Text Data

Prepare Text /Parsing

Tokenization/Term
POS Tagger
Chunker
Sentence Breaker

OUTPUT
Structured Data

Topic: classification; what subject matter of the text
Entities: who/what are the people, places, proper nouns
Phrases/Themes: meaningful word phrases
Sentiment: Grade text on a positive/negative numeric scale

Parsed Text
Agenda

• Big Data Overview

• All About Hadoop
  • What is Hadoop?
  • How does MicroStrategy connects to Hadoop?
  • Customer Case Studies
  • Demo

• MicroStrategy Now Supports MongoDB

  • MicroStrategy Consumes Unstructured Data with Text Analytics
    • What is Text Analytics?
    • Use Cases of Text Analytics
    • Introducing New Text Analytics Module
    • Demo

• Q&A
Demo – Text Analytics
Showcases MicroStrategy’s Capability to Get Insights from Unstructured Text

**Use Case:** Analyzes unstructured text collected from a car dealership customer survey forms.
## Demo – Text Analytics

### Car Dealer Survey Data

Real Data from a car dealer is mocked up as “Tesla”
Demo – Text Analytics

Text Analytics Dashboard after text processing

Customer Sentiment Analysis

Dealer Overview

Category Detail

Sentiment Distribution

- Positive
- Neutral
- Negative
- Count

Trend Chart

Top 3 High Sentiment Scores in: CA

<table>
<thead>
<tr>
<th>Dealer</th>
<th>Sentiment Score</th>
<th>Survey Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALAMANDRA, CA</td>
<td>55</td>
<td>115</td>
</tr>
<tr>
<td>ANAHEIM, CA</td>
<td>52</td>
<td>99</td>
</tr>
<tr>
<td>AVENAL, CA</td>
<td>52</td>
<td>69</td>
</tr>
</tbody>
</table>

Bottom 3 Low Sentiment Scores in: CA

<table>
<thead>
<tr>
<th>Dealer</th>
<th>Sentiment Score</th>
<th>Survey Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALAMEDA, CA</td>
<td>38</td>
<td>194</td>
</tr>
<tr>
<td>AGUAMILLA HILLS, CA</td>
<td>34</td>
<td>96</td>
</tr>
<tr>
<td>APTOS, CA</td>
<td>31</td>
<td>220</td>
</tr>
</tbody>
</table>

No data returned for this view. This might be because the applied filter excludes all data.

MicroStrategy World 2014
MicroStrategy’s big data solution is backed up with its organic, tightly integrated technology and customer success!
- Facebook
- Netflix
- Yahoo! and many more!

MicroStrategy provides a certified and reliable solution for integrating with Hadoop/Hive and other SQL on Hadoop solutions.

Latest support:
- Google Big Query
- Pivotal HAWQ

MicroStrategy taps into No SQL sources like Mongo DB. Support for more such sources in future!

MicroStrategy consumes unstructured data with Text Analytics Module. Contact your account representatives or beta programs to try this exciting product!
Q&A

Questions?