Tapping into Hadoop and NoSQL Data Sources in MicroStrategy

Presented by: 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

#mstrworld
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**
Distribute the same analytics across the organization
Reuse, improve and extend easily
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></td>
<td><strong>SOURCE</strong></td>
</tr>
<tr>
<td>Company, Government, Financial sector, Business and consumer studies, Surveys, Polls</td>
<td>Online click-stream, Application logs, Call/service records, ID scans, Security cameras</td>
</tr>
<tr>
<td><strong>VALUE</strong></td>
<td><strong>VALUE</strong></td>
</tr>
<tr>
<td>All business performance drivers – Operational efficiency, Revenue management, Strategic planning</td>
<td>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></td>
<td><strong>SOURCE</strong></td>
</tr>
<tr>
<td>Content generated from social media posts, tweets, blogs, pictures, videos, ratings</td>
<td>Machine generated sensor data and “connected device” communication</td>
</tr>
<tr>
<td><strong>VALUE</strong></td>
<td><strong>VALUE</strong></td>
</tr>
<tr>
<td>Customer engagement, Customer service, Brand management, Viral marketing</td>
<td>Operational efficiency, Cost control, Risk avoidance</td>
</tr>
</tbody>
</table>

#mstrworld
# No Data Left Behind

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

<table>
<thead>
<tr>
<th>MapReduce &amp; NOSQL Databases</th>
<th>cloudera</th>
<th>MAPR</th>
<th>amazon Elastic Map Reduce</th>
<th>IBM</th>
<th>PIVOTAL</th>
<th>TERADATA</th>
<th>ASTER</th>
<th>MarkLogic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hortonworks</td>
<td>mongoDB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columnar Databases</td>
<td>amazon Redshift</td>
<td>Sybase</td>
<td>VERTICA</td>
<td>PARAccel</td>
<td>Qubole</td>
<td>EXASOL</td>
<td>kognitio</td>
<td>Google bigquery</td>
</tr>
<tr>
<td>Data Warehouse Appliances</td>
<td>TERADATA</td>
<td>netezza</td>
<td>ORACLE</td>
<td>SAP</td>
<td>HANA</td>
<td>SQL Server 2008 R2</td>
<td>parallel</td>
<td>Greenplum</td>
</tr>
<tr>
<td>Relational Databases</td>
<td>IBM DB2</td>
<td>MySQL</td>
<td>IBM Informix</td>
<td>ORACLE</td>
<td>SQL Server</td>
<td>SYBASE</td>
<td>PostgreSQL</td>
<td>MariaDB</td>
</tr>
<tr>
<td>Multidimensional Databases</td>
<td>SAP</td>
<td>Hyperion</td>
<td>SQL Server Analysis Services</td>
<td>COGNOS</td>
<td>TM 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SaaS-Based App Data</td>
<td>Salesforce</td>
<td>connectioncloud</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User / Departmental Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#mstrworld
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 DWH 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
Interfaces Available to Analyze Data Stored in Hadoop

**SQL on Hadoop**
- Interactive SQL interface for Hadoop typically used for BI, analytics, and ad-hoc queries.

**Apache Hive**
- Data warehouse infrastructure built on top of Hadoop. Allows SQL-like queries to be submitted to a Hadoop cluster.

**Apache Shark/Spark**
- In-memory data analytics cluster computing framework on the HDFS. Promises better performance when compared to map reduce.

**Apache Pig**
- High level procedural language, typically useful for batch data flow workloads.

**Streaming/Java MapReduce**
- The “native” way of accessing data in Hadoop but requires writing code in Java.
How does MicroStrategy integrates 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 certifies Spark/Shark on HDFS.
- 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 9.4.1
- Q4 2013
  - Hortonworks Simba ODBC driver
  - MSTR Greenplum driver for HAWQ
  - Splunk ODBC driver

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
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

---

**Making Better Sales Decisions by Getting Insights from Web Logs in Hadoop**

**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.

Precise Targeting of Digital Ads
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

#mstrworld
Demo – Log Analytics on Alert Loyalty Platform
MicroStrategy’s capability to query and analyze web logs stored in HBase.

- 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.

Data stored in HBase table
# Demo – Log Analytics

**Performance Dashboard - 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/create2</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/admin</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>85.55</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>58.63</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>434,851</td>
<td>1.75</td>
<td>98.42</td>
<td>0.00</td>
<td>289.13</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>

**Concurrency**

API Overview Performance

(choose one from the grid)

**Overall System Load and Average Backend Process Time**

Series: AVG Process Time (s)
Category: 05/14/2013 Value: 3.09

# mstrworld
Demo – Log Analytics
Performance Dashboard - Concurrency

Overall Concurrency Usage (i/day)

Daily Concurrency Usage (i/min)

Count of Executions
Average Process Time
Demo – Log Analytics
Performance Dashboard – API Overview

Overview

<table>
<thead>
<tr>
<th>(All)</th>
<th>Admin</th>
<th>Client</th>
<th>Internal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>19/12/2013</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

API Overview

Concurrence

Executions < 500

API Quadrant Plot

Executions > 500

<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>/category/news</td>
<td>53</td>
<td>1.04</td>
<td>1.61</td>
<td>0.19</td>
<td>180.28</td>
</tr>
</tbody>
</table>

Rectangle Size

Rectangle Color: Count...
Demo – Log Analytics
Performance Dashboard – API Overview

Overview

<table>
<thead>
<tr>
<th>Date</th>
<th>Admin</th>
<th>Client</th>
<th>Internal</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROD NODE 01 (m19)</td>
<td>PROD NODE 04 (m14)</td>
<td>PROD NODE 08 (m11)</td>
<td></td>
</tr>
<tr>
<td>PROD NODE 03 (m19)</td>
<td>PROD NODE 12 (m19)</td>
<td>PROD NODE 10 (m12)</td>
<td></td>
</tr>
<tr>
<td>PROD NODE 06 (m20)</td>
<td>PROD NODE 09 (m12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROD NODE 02 (m19)</td>
<td>PROD NODE 13 (m27)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROD NODE 07 (m19)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

API

Concurrence

API Overview

Performance Trend

Global Usage

Executions < 500

API Quadrant Plot

Series: Count of Executions
Category: /admin/jctors_id/campaigns
Count of Executions: 25
Average Process Time: 5.13.78
Average Data Size: 5.777464357

Executions > 500

API

<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 (tb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/category/news</td>
<td>33</td>
<td>1.04</td>
<td>1.61</td>
<td>0.19</td>
<td>180.28</td>
</tr>
</tbody>
</table>

#mstrworld
Demo – Log Analytics
Drill down from Performance Overview: System Usage
Demo – Log Analytics
Benefit of using HBase/Impala Solution vs. MySQL

- **Scalability/Reliability**
  - With MySQL we can only store 30 days of data with existing customers
  - Reliability is not an issue with HBase

- **Performance**
  - For Dashboard cube publishing with similar data, Impala was 10x faster using less hardware for MicroStrategy Server
  - (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
Full and Flexible Capabilities for Hadoop Analytics

The dominant BI solution for interactive access to Hadoop data

All Leading Distributions

Certified integration with the most popular Hadoop and MapReduce solutions.

Multiple Connection Options

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

In-Memory Cubes for Interactive Hadoop Analytics

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

1. Extract

2. Accelerate > 100x

3. Interact and Discover

#mstrworld
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

#mstrworld
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
More About Simba ODBC Driver

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

• 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.

Text Mining

Unstructured Data

Structured Data
Use Cases of Text Analytics

- Security Applications
- Sentiment Analysis
- Media Monitoring
- Customer Loyalty
- Market Intelligence
- Product Reviews
- Brand Management
- Medical Applications

#mstrworld
Open the Gates for New Insights from Text Analytics

Bring unstructured text into MicroStrategy and get insights!

Introducing MicroStrategy’s New Text Analytics Module that translates unstructured data into consumable csv/tabular data.

Currently in Beta, please contact our beta programs for trying it on your data.

MicroStrategy Analytics Platform

Traditional BI Platform based on structured data with well defined hierarchies.

Consumes Structured Data
MicroStrategy Text Analytics Module Workflow

Raw Text Data File(s)

Posts Comments Reviews Survey Responses etc.

Text Processing

MicroStrategy Text Analytics Module

Tune able 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

**Parsed Text**

**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

#mstrworld
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

#mstrworld
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
DeLorean Car Dealer Survey Data

Excel spreadsheet showing customer feedback on Tesla vehicles, including positive and negative comments about the dealerships and customer service.

#mstrworld
MicroStrategy’s big data solution is backed up with its organic and tightly integrated technology. And therefore these leading industries use MicroStrategy:

- Facebook
- Netflix
- Yahoo! and many more!

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

MicroStrategy provides a certified and reliable solution for connecting and analyzing data from Hadoop/Hive and other map reduce solutions.

Latest support:

- Google Big Query
- Pivotal HAWQ
- Spark/Shark

MicroStrategy consumes unstructured data with Text Analytics Module. Contact our beta programs to try this exciting product!
Questions?