Analytics with Hadoop Distributions
Leveraging Cloudera, Hortonworks, and MapR for Data Discovery
Andrew Kern, MicroStrategy World 2018
Safe Harbor Notice

This presentation describes features that are under development by MicroStrategy. The objective of this presentation is to provide insight into MicroStrategy’s technology direction. The functionalities described herein may or may not be released as shown.

This presentation contains statements that may constitute “forward-looking statements” for purposes of the safe harbor provisions under the Private Securities Litigation Reform Act of 1995, including estimates of future technology releases. Forward-looking statements inherently involve risks and uncertainties that could cause actual results of MicroStrategy Incorporated and its subsidiaries (collectively, the “Company”) to differ materially from the forward-looking statements.

Factors that could contribute to such differences include: the Company’s ability to develop, market and deliver on a timely and cost-effective basis new or enhanced offerings that respond to technological change or new customer requirements; delays in the Company’s ability to develop or ship new products; the extent and timing of market acceptance of MicroStrategy’s new offerings; continued acceptance of the Company’s other products in the marketplace; competitive factors; general economic conditions; and other risks detailed in the Company’s registration statements and periodic reports filed with the Securities and Exchange Commission. By making these forward-looking statements, the Company undertakes no obligation to update these statements for revisions or changes after the date of this presentation.
Data is the new science. Big Data holds the answers.

-Pat Gelsinger, CEO VMware, Inc.
Agenda

- Hadoop
- Hortonworks
- Cloudera
- MapR
- Connectivity Options
- Q&A
Connecting to Data
Traditional Enterprise DW

RDBMS Platforms
Traditional SQL based access for reporting and dashboarding
ANSI SQL standard allows relatively consistent experience across relational database platforms
Individual gateways can be optimized by default using MicroStrategy VLDB Properties

Schema Model
Created through MicroStrategy Architect
Best practice around star/snowflake schemas, focus on normalized data structures
Allows for optimal join paths, performant prompting
Additional implementations in MicroStrategy
  • Dataset caching
  • Prompting/filtering capabilities
  • Security Filters
Multi-Source & Bring-your-own-Data
Leveraging disparate data sources

Multiple SQL Platforms
In-house data sources for different departments
Leverage combination of multiple sources for additional powerful insights
Leverages MicroStrategy’s ability to:
• Store data within the MicroStrategy Analytical Engine
• Write data back into Relational data sources
• Create a projects schema, allowing mapping of an attribute across multiple database platforms

Data Import
Support quick analysis like flat files
Additionally support non-relational web based data sources (e.g. Twitter, Google Analytics, etc.)
Concerns around data integrity, data recency, data security
MicroStrategy makes it possible to:
• Restrict access to import data to specific users/groups
• Import data from a shared location to specify schedules for cube publication
• Ability to map imported data to defined schema attributes with security
Evolution of Big Data

Industry Drivers

• Increasing volumes of incoming data via web applications
• Need to provide quick performance on read actions for web users
• Need for consistency and high availability
• Need to access large volumes of data with computations in reasonable times

Evolution

• Hadoop/HDFS
• Columnar Databases
• MPP platforms
• Key Value stores
• Document data stores (JSON, XML)
Big Data on Hadoop

Hadoop and HDFS
- HDFS – Hadoop Distributed Files systems
- Open Source
- MapReduce – Batch processing program
- Distribution platforms have emerged to provide underlying data storage, access, management, and security solutions
- These large, unstructured sources of data are often referred to as data lakes

MicroStrategy’s overall Big Data strategy
- Enable all users to tap into their data lakes, and provide an end-to-end user experience in MicroStrategy right from connecting, wrangling, modeling, storing, querying, calculating, to building Dossiers and dashboards.
Big Data Use Cases
Analyze Large Data Volumes with a Platform that is Scalable and Performant

Selected use Cases

• Real time transactions reporting
• Interactive analysis on high-volume clickstream data
• Merchant analytics for a credit card issuer
• Prevent fraud and waste
• Equipment and medicine tracking

Leverage scheduling and distribution capabilities to perform **Offline – Processing**

Create and publish Intelligent Cubes with **High Data Scalability**

Push down analytical calculations to Hadoop using the **Powerful SQL Engine**
Enterprise Data Warehouse Optimization
Leveraging MicroStrategy’s Integrated Architecture for Data Blending
Hortonworks
Based on Apache Hadoop Distribution

Hortonworks Data Platform (HDP) is secure, enterprise-ready open source Apache™ Hadoop® distribution based on a centralized architecture (YARN).
Connectivity
SQL on Hadoop

Diagram

Supported Options

1. Hive and Map-Reduce
2. Hive on Tez
3. Spark
4. HBase via Phoenix
5. Hadoop Gateway
6. Druid
1. Connectivity via Hive on Map-Reduce

Best for Batch Queries

- The most popular way of querying Hadoop. Hive uses the map reduce framework to query data from HDFS. It is fault tolerant, but has high latency.
- Best suited for complex, long-running tasks and analyses on large sets of data.
- Use Case: Ideal for offline execution
  - Cube publication workflows.
  - Scheduled report delivery.
- ODBC Driver for Hive is shipped out of the box
- JDBC driver is shipped with Mac
2. Connectivity via Hive on Tez

Best for Interactive Queries

- Tez delivers fast batch and interactive SQL, overcoming long-latency of Map-Reduce
- Ideal for data discovery workflows using Visual insight
- For Hive to use Tez execution engine instead of the Map-reduce engine:
  - Enable for all queries: set hive.execution.engine=tez
  - Enable on a report-by-report basis from the report’s VLDB settings; or for all the reports via the DB Instance->VLDB Settings->Pre/Post statements and adding the above line to enable tez engine
- Utilizes the same ODBC/JDBC driver as when connecting to Hive via MapReduce
3. Connectivity via Spark
To enable interactive queries

- Apache Spark is one of the fastest evolving open source communities, popular for it’s MPP capabilities on top of Hadoop and providing advanced analytical workflows

- Benefits:
  - Spark has wide community support.
  - Can pin tables in memory
  - Enables advanced use cases such as Streaming
  - MicroStrategy ships SparkSQL ODBC Driver
4. Connectivity to HBase via Phoenix
Connect to structured and unstructured data

- Apache HBase™ is a distributed, scalable, No-SQL Big Data store.
- It’s primarily a key-value store for transactional processing.
- Connectivity through Phoenix, a relational layer over HBase.
- Use cases: To enable retrieving transaction-level data from HBase.
- Example: Provide current status of a particular account.
- MicroStrategy supports Phoenix 4.x, connectivity through JDBC
5. Connectivity via Hadoop Gateway

To natively access HDFS

- Native YARN based application that enables direct access to HDFS
- New in 10.0; substantially enhanced through 10.10
- Supports both in-memory and live connect
- Use cases
  - In-memory cube publishing at scale
  - Data preparation/wrangling at scale
6. Connectivity via Druid

- Designed for Business Intelligence queries based on event data
- Has been implemented as a proof of concept using MicroStrategy’s Generic DBMS object
- Developing out of the box connector
- Use cases
  - Event based data (Single table structure, no joins)
  - Real time analytics using native integration with Kafka
Best Practices – Design and Query
Optimize Query Performance

- Optimize your warehouse design
- Write good SQL
- Use views – Hive views are natively supported in MicroStrategy
- Partition smartly
- Indexing
- Bucketing
  - Populating Partitioned and Bucketed table
  - Enable pre/post SQL statements from the VLDB settings
Best Practices
System Parameters

- Use ORC file format
- Use Compressed Sequence Files
- Use Tez execution engine instead of Map-Reduce to speed up queries
  - Set hive.execution.engine=tez;
- Vectorization
  - Set hive.vectorized.execution.enabled=true;
  - Set hive.vectorized.execution.reduce.enable=true;
- Enable cost-based query optimization
  - Set hive.cbo.enable=true;
  - Set hive.compute.query.using.stats=true;
  - Set hive.stats.fetch.column.stats=true;
  - Set hive.stats.partition.stats=true
- VLDB setting: DB Instance->VLDB Settings->Pre/Post statements
Cloudera Enterprise

Adaptive and High-Performance Analytic Database

- Provides a management system for connecting into disparate data stores such as HDFS
- Differentiator is Impala, used for fast, interactive SQL queries against HDFS
Cloudera Impala - Fast Interactive SQL

High-performance, low-latency SQL queries on data stored in popular Apache Hadoop file formats.

- The fast response for queries enables interactive exploration and fine-tuning of analytic queries.
- Familiar SQL interface that data scientists and analysts already know.
- Query high volumes of data in Apache Hadoop.
- Single system for big data processing and analytics.
MicroStrategy’s Integration with Cloudera

• MicroStrategy certifies Cloudera CDH 5.x through connectivity to Impala and Hive (Including up to Impala 2.x)

• MicroStrategy ships ODBC drivers for both Hive and Impala

• Security integration with Kerberos, Delegation and Sentry

• MicroStrategy’s Hadoop Gateway supports the Cloudera Platform
Best Practices with Impala in MicroStrategy

• Store tables in Parquet File format

• This should include the creation of temporary tables
  • VLDB Property: Intermediate Table Type -> Permanent Table
  • VLDB Property: Create Post String: STORED AS PARQUET;
  • VLDB Property: Table Creation Type: Explicit

• HDFS caching
  • hdfs cacheadmin -addPool four_gig_pool -owner impala -limit 4000000000

• Optimize queries on Parquet Tables by issuing COMPUTE STATS after substantial
data is loaded into/appended into a table
MapR
Converged Data Platform

OPEN SOURCE ENGINES AND TOOLS

PROCESSING
- Hadoop
- Spark
- Apache Drill
- Search and Others

COMMERCIAL ENGINES AND APPLICATIONS

- Cloud and Managed Services
- Vertica
- SAP
- MySQL
- Custom Apps

WEB-SCALE STORAGE
MAPR-FS
- High Availability

DATABASE
MAPR-DB
- Real Time
- Unified Security
- Multi-tenancy
- Disaster Recovery
- Global Namespace

EVENT STREAMING
MAPR STREAMS

DATA

UNIFIED MANAGEMENT AND MONITORING

HDFS API
POSIX API
HBase API
JSON API
Kafka API

ENTERPRISE-GRADE PLATFORM SERVICES

MicroStrategy
Apache Drill

Overview

• A SQL Query Layer against NoSQL Data Sources
• JSON Data Model
• Data-driven Query – Compile Query on-the-fly without knowing the schema ahead
• Columnar Execution – Shredded, In-Memory Columnar Execution
MicroStrategy Integration

Drill

- ODBC/JDBC Connectivity available
- ODBC Driver shipped with the MicroStrategy platform
- MicroStrategy is certified on MapR 5.x (Drill 1.10) in 10.10

<table>
<thead>
<tr>
<th>Supported Data Sources</th>
<th>Apache Drill</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HBase</td>
</tr>
<tr>
<td></td>
<td>MongoDB</td>
</tr>
<tr>
<td></td>
<td>MapR-DB</td>
</tr>
<tr>
<td></td>
<td>HDFS</td>
</tr>
<tr>
<td></td>
<td>MapR-FS</td>
</tr>
<tr>
<td></td>
<td>Amazon S3</td>
</tr>
<tr>
<td></td>
<td>Azure Blob Storage</td>
</tr>
<tr>
<td></td>
<td>Google Cloud Storage</td>
</tr>
<tr>
<td></td>
<td>Swift</td>
</tr>
<tr>
<td></td>
<td>NAS and local files</td>
</tr>
</tbody>
</table>
Best Practices with Drill in MicroStrategy

- Exclude schemas
  - ExcludedSchemas=hive.default,sys;

- Parquet file format

- Metadata caching
  - Hive/Parquet

- Access data through Drill Views
Connectivity Options

Supported on a wide range of platforms and connection modes

Products
- Desktop Windows
- Desktop Mac
- Enterprise Windows
- Enterprise Linux

Capabilities
- Live-Connect
- Multi-source
- In-memory

Connections
- DSN
- DSN-less