Transitioning from Relational to Hadoop

Jochen Demuth, Senior Director – Partner Engineering
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.
Why?
Live Access to Data Sources with Unparalleled Performance
Quickly generate multi-pass SQL and leverage push down functionality for optimized performance

Dynamic Sourcing
MicroStrategy has the ability to allow queries to seamlessly and dynamically drill across multiple sources, and the server is able to auto recognize situations and intelligently direct queries against in-memory cubes when possible.

Generate Multi-pass SQL
MicroStrategy is able to easily and quickly generate multi-pass SQL queries to provide greater analytical power and minimize the amount of data that is pulled back to the mid-tier. The sophisticated SQL engine can deliver high performance for the most complex SQL computations.

Pushdown architecture
MicroStrategy leverages the database to its fullest extent by pushing data joins and analytic calculations to the database when possible. Every connector is optimized for high performance by pushing down functions to leverage the power of the database in conjunction with the Server.
Progression of Data Technologies
Innovations have made data more accessible and faster to access and analyze than ever before

Products providing this technology:
- Microsoft SQL Server
- Oracle Database
- IBM DB2
- Teradata
- HP Vertica
- MongoDB
- Cassandra
- Amazon Redshift
- Snowflake
- Splunk
- ElasticSearch
- ParStream
- PipelineDB

- Faster to query
- Larger data volumes
- Columnar
- Faster to build apps
- Easier to modify and extend
- Faster to deploy
- Easier to grow and extend
- Easier data access
- Faster data access
- Real-time analysis

• Enterprise data warehouse with 360 degree view
• Single version of the truth
• Faster data access
• Real-time analysis
Hadoop Offers A Mature Ecosystem of Vendors and Technologies

- Cloudera
- MapR
- Hive
- Solr
- Hortonworks
- AtScale
- Drill
- Spark
- Jethro
- Impala
- HBase
- Kafka
Transitioning from Relational to Hadoop
Top things to consider while moving from relational sources into Hadoop platforms

In the best case, the transition is as easy as changing the DB type in the DBinstance and pointing to a new DSN

**Things to consider when:**
- Migrating Data
- Switching Application to Hadoop
- Optimizing application for Hadoop

**Today, a growing number of organizations are looking to supplement or replace their existing legacy systems with Hadoop-based platforms. Attend this session to learn more about:**
- The ongoing transformation and evolution of modern data architectures and how it impacts your organization
- Examples of technical and business scenarios where it makes sense to move data from relational sources
- How MicroStrategy delivers an adaptive model to dynamically harness the value of data from disparate sources
- The need for a metadata-based infrastructure to tie disparate sources together to deliver a trusted system of record repository
- Options for accessing and blending data from Hadoop and other data sources
Considerations

• Plenty of new turf and it’s continually changing

• SQL is still key as NoSQL or NewSQL
  • Add in semi-structured and data mining

• SQL on Hadoop products are evolving quickly with rapid increases in functionality and performance
  • Performance can double, triple or more in some cases in 6 month periods

• Ultimate decisions depend on testing with your data

• Schema on read is key for iteration and flexibility in the back end

• Schemas are still important to operationalize analytics to large number of users

• Use of aggregation, agg-aware sql generation, caching, cubes, and dynamic sourcing help with performance and scalability

• Memory Caching is a key infrastructure component
What?
MicroStrategy Delivers a Single Version of the Truth on Top of Data

Leverage the unified metadata repository to design reusable objects for rapid development, governance and scale.
MicroStrategy Interacts with Hadoop in Same Way as Relational Data

Enables incrementally leveraging Hadoop capabilities

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
5. Agg aware SQL engine, dynamic sourcing with cubes, caching and SQL access

Maturity of Data Access
How?
Big Data Apps Leverage Dynamic Sourcing and SQL Engine

Query multi-source schema model and drill down from Intelligent Cubes to Hadoop

Agg-aware SQL engine, dynamic sourcing with cubes, caching and SQL access
Create Independent Analysis of New Data Sets
Example: Introducing the 1.3 Billion Row Table

- **Analyzing 1.1 Billion NYC and Uber Trips with a Vengeance**
  Now there’s over 1.3 Billion trips
- There’s a large body of work out there based on this data
Analyzing the Taxi Data
Sizing a Hadoop Cluster Balances Cost versus Performance

Key metrics that impact performance and cost of a Hadoop deployment

Adding more nodes
- Nodes can be added/removed dynamically
- Nodes add compute and storage capacity both in memory and on disk
  - Each capacity can be configured dynamically
  - Querying detail vs. querying an intermediate aggregate
  - Pushing down analytics - using complex SQL

Adding more memory
- Does the whole table fit in memory?
- How much does a (peta-)byte of memory cost (per month)?
  - Caching tables
  - Keeping the most important things in memory
    - Constantly reviewing what is "most important"
    - Changing what's in memory

Other related Systems
- Disk and Network
Tuning Hadoop Cluster for MicroStrategy Workloads

**Hadoop Tuning**
- HDFS – YARN – Queues – Memory settings - JVM

**SQL on Hadoop Tuning**
- Hive – Impala – Presto – Spark SQL – Drill – Jethro – AtScale
- Collecting Statistics and Physical Data Modeling/De-normalization
- Storage Type: AVRO, Parquet, ORC

**MicroStrategy Tuning**
- Governing, Memory, Queues
- Statistics
- Aggregate tables and cubes
Spark Table Caching
Moving from one Hadoop SQL Engine to Another

A relatively simple process ensures quality in the transition

Create functional and performance baseline
- Integrity Manager can validate functional, data correctness, single user performance
- Enterprise Manager collects actual multi-user use and performance

Repoint database instance
- Evaluate configuration changes, especially non-default VLDB settings
- Use Integrity Manager for baseline comparison and resolve issues

Compare performance against performance baseline
- Tune and resolve issues
Lift and Shift - Migrating Existing Applications to Hadoop
Watch Out for the Following Deployment-specific Things
Most production deployments don’t exclusively use product defaults

Data deployment
• Modified DB object names. E.g. User names, name spaces, table names, column names, etc.
• Unsupported data types

Data source specific syntax encapsulated in your MicroStrategy metadata
• Non-default VLDB settings (e.g. pre-/post-SQL)
• Pass-through functions (e.g. applysimple)
• Transaction Services – SQL write-back
• Logical tables (essentially a view encapsulating a SELECT statement)
• FFSQL/QB reports
Leverage the Hadoop Ecosystem
Consider products which pre-aggregate data for suitable workloads

• Mixing structured and unstructured data in the same project
• User-defined functions
• Hadoop distributed cache options to speed up performance
Thank You!
Hadoop offers a mature ecosystem of modern technology
The tools available in Hadoop enable capabilities that required separate independent product deployments or were not available in a pure relational environment
Transitioning MicroStrategy to adopt Hadoop can happen in multiple stages
Hadoop technologies are as complex as traditional sources in some cases not quite as polished.
Migrate existing projects only after the Hadoop technologies are mastered.