Governed Analytics on Hadoop

How to deploy governed analytics on top of Hadoop

Benjamin Reyes, Product Management, Data
Governed Analytics on Hadoop
How to deploy governed analytics on top of Hadoop

Agenda

• Introduction – The need for governance: Self-Service Big Data Analytics

• The evolution of Hadoop-based analytics

• MicroStrategy Metadata: a consistent business definition of data

• Enforcing security policies with the MicroStrategy Security Model

• Leveraging existing Hadoop-level user security via Kerberos Authentication

• Governed Data Science workflows and operationalizing predictive analytics on Hadoop

• Customer anecdotes from the field

• Q&A
Introduction
The need for governance: Self-Service Big Data Analytics
Introduction
The evolution of Hadoop-based analytics

• Hadoop-based technologies have evolved from batch analytics scenarios to low latency, real-time decision support and interactive data analysis.

• Open source SQL-on-Hadoop engines have become more performant in their most recent releases, with new variants of Hive running on more faster processing engines such as Spark or Tez.

• Commercial products are also taking advantage of native implementations, proprietary file formats and indexing.
Self-Service BI Applications on Hadoop
Business agility and enterprise governance

• Applications that traditionally queried data in relational sources and multidimensional databases are increasingly sourced by the data lake.

• Applications include operational dashboards, interactive, exploratory, ad-hoc reporting for self-service and even embedded analytics uses cases.

• Data lakes have become great platforms for data discovery and rapid application development but they need business context and governance.
Self-Service BI Applications on Hadoop
Business agility and enterprise governance

• Organizations want to enable self-service applications but within governed models and definitions to avoid silo applications.

• Modeled schemas provide governing and confidence in the data by promoting reusable business definitions of data i.e attributes, metrics, hierarchies.

• A fully-fledged MicroStrategy project schema can be built on top of Hadoop sources by leveraging SQL-enabled gateways.
MicroStrategy Unified Metadata Model

A, governed, common business definition of your data

Reusable Objects, Built on a Unified Model, Deliver Rapid Development, Governance, and Scale

The MicroStrategy metadata repository is at the heart of every analytics project. It stores and catalogs logical business definitions of enterprise data and forms the **reusable building blocks** for reports, dossiers, and applications across an analytics deployment.

- A unified model helps ensure a **single version of the truth**
- **Reusable** objects enable efficient, high scale development
- Object change proliferation reduces the IT burden, and
- Centralized metadata enables user **personalization** and **security**
MicroStrategy Metadata Model
A governed, common business definition of your data
MicroStrategy Security Model
A unified model to secure, restrict and control access to objects, capabilities and data.
Users and User Groups
A unified model to secure, restrict and control access to objects, capabilities and data

User Groups

• User Groups are a collection of users that share a common set of privileges and permissions. User groups provide a convenient way to manage a large number of users.

• Instead of assigning privileges to hundreds of users individually, organizations can assign privileges to a group. Groups may also be assigned permissions to specific objects.

• Object Access Control, Privileges and Security Filters can be set at the User Group level.
User Administration
Control data access through object-level access control lists (ACLs)

Object Access Control

- MicroStrategy gives administrators the tools that they need to control data and application access via privileges, access control lists (ACLs), and security filters.

- ACLs govern the security permissions of individual metadata objects. Each object has its own ACL, which grants users a specific set of privileges for the object (including: browse, read, write, full control, use, execute)
User Administration
Control data access through object-level access control lists (ACLs)

MicroStrategy Developer
User Administration
Control what actions or features are available to users or groups

Role-based Privileges

- MicroStrategy employs over 160 unique privileges that assign specific application functionality to user groups, user roles, and individual users.

- Roles are a convenient way to assign an extensive range of privileges to a user or group of users based on their role within the organization e.g. Analysts, Consumers, Administrators, etc.

- Roles are assigned at the project level for complete flexibility.

- Users can also be given specific privileges on an individual basis, over-riding their role definition.
User Administration
Control what actions or features are available to users or groups

MicroStrategy Developer
Security Filters
Secure data according to users security policies and group membership

Enforce data security policies

• Security filters allow administrators to control row and column level data access across reports and dossiers, so that each user or user group only has access to data that is appropriate for their role within the organization.

• Security filters hold expressions that qualify all queries executed against the data source, regardless of the objects on the report or dossier.

• Filters can be assigned at the group or individual user level and may have complex expressions across multiple dimensions.
Security Filters
Secure data according to users security policies and group membership

Define complex filter expressions

• Filters can be assigned at the group or individual user level and may have complex expressions across multiple dimensions.

• A manager who oversees the Engineering department may only see information for his own department.

• All report data will be filtered so every dimension reflects only his specific department.

• The filtering condition is appended to every query executed by the user.
Object Certification
Stamp trusted datasets and dossiers

Give select users the authority to “Certify” objects

- Certifiers may stamp datasets and dossiers as “certified” objects.
- Certified objects will show a visual indicator of a “seal” on all tools and client applications.
- Self-service users will confidently create reports and Dossiers from “certified” datasets.
Leveraging Hadoop Security Policies
Enforce user-level security via Kerberos Authentication

Kerberos Authentication and User Delegation

• Leverage Hadoop user-level security policies by executing all queries under a specific user’s identity

• Use Kerberos Authentication and user delegation to submit queries for each individual user

• Re-use data security policy defined by the Hadoop administrator

Learn more

TRACK 9: BIG DATA
Kerberos Authentication for Hadoop: How to deliver secure analytics on big data with Kerberos
Wednesday 17 | 4:05 PM - 4:30 PM
Governed Data Science on Hadoop
Enable Data Science workflows on governed datasets

Machine learning on MicroStrategy
data in any platform or language

• Train cutting-edge models on trusted data, and put predictions in decision makers’ hands with application, dashboard, and report integration

• Open-source libraries for Python, R, and other new entrants

• Train models in the cloud with connectors to ML-as-a-service platforms like Azure ML, AWS Machine Learning, and Alteryx

• All the benefits of control and scalability of MicroStrategy with the openness and ease of deployment data scientists need

Learn more

TRACK 5: DATA SCIENCE, AI, AND MACHINE LEARNING
Machine Learning with Python: Train models on trusted data to equip users with predictive analytics
Wednesday 17 | 3:30 PM - 4:30 PM
Customer stories from the field

Moacyr Passador, Senior Sales Engineer
A Couple of Customer Anecdotes
Moving from a relational warehouse to “Big Data”

What would it take to migrate to a Big Data source?

Moving the data, updating pipelines, data validation…

Months? years?

For MicroStrategy it will be a few clicks!
A Couple of Customer Anecdotes
Moving from a relational warehouse to “Big Data”

At a Social Media Company

RDBMS → presto

1 month, including data validation

At a Technology Company

ORACLE DATABASE → druid

1 week
Q&A
Thank you!