Don’t build a data swamp

Hadoop governance case studies for financial services

Mark Donsky, Product Management, Cloudera
Chang She, Software Engineering, Cloudera
Compliance + Productivity = Hadoop Adoption

Compliance
• Am I prepared for an audit?
• Who’s accessing what data?
• What are they doing with the data?
• Is sensitive data governed and protected?

End-user productivity
• How can I explore data on my own?
• Can I trust what I find?
• How do I use what I find?
• How do I find and use related data sets?
### Governance: the foundation of data management

#### Compliance
- Track, understand and protect access to data
  - Am I prepared for an audit?
  - Who’s accessing what data?
  - What are they doing with the data?
  - Is sensitive data governed and protected?

#### Stewardship
- Manage and organize data assets at Hadoop scale
  - How do efficiently manage data lifecycle, from ingest to purge?
  - How do I classify data efficiently?
  - How do I make data available to my end users efficiently?

#### Data Science
- Effortlessly find and trust the data that matters most
  - How can I explore data on my own?
  - Can I trust what I find?
  - How do I use what I find?
  - How do I find and use related data sets?

#### Administration
- Boost user productivity and cluster performance
  - How is data being used today?
  - How can I optimize for future workloads?
  - How can I quickly take advantage of Hadoop risk-free?

---

**Hadoop Governance Foundation**
- Unified metadata
- Unified lineage
- Unified auditing
- Universal policies
Compliance Use Cases

Compliance
Track, understand and protect access to data

- Am I prepared for an audit?
- Who's accessing what data?
- What are they doing with the data?
- Is sensitive data governed and protected?

HADOOP DATA GOVERNANCE & MANAGEMENT

- Unified metadata
- Unified lineage
- Unified auditing

ENTERPRISE METADATA REPOSITORY

- informatica
- Data Advantage Group
- IBM
- adaptive

ENTERPRISE AUDITING & SECURITY

- splunk
- IMPERVA
- IBM
- RSA

Common use cases:
- Security breach detection
- Data access tracking for PCI compliance
- Audit defense
Compliance Case Study

- A top US-based bank holding company that grew through many acquisitions
- Many disparate systems – some Hadoop, but many others, including EDW, relational databases, etc.
- All governance artifacts were centralized in a custom-built lineage and metadata enterprise framework

Problem
- Basel reporting, AML reporting, Dodd Frank, etc. required a single source of truth
- Automatic controls weren’t in place for Hadoop
- Difficulty answering “how do you know these numbers are correct?”

Solution
- Auditing
  - Captured a full audit trail of user behaviour and integrated with enterprise tools
- Lineage
  - Captured 100% of attribute-level transformations and integrated with enterprise systems

Results
- Costs of passing an annual PCI audit were cut in half

© Cloudera, Inc. All rights reserved.
Stewardship Use Cases

**Stewardship**
Manage and organize data assets at Hadoop scale

- How do efficiently manage data lifecycle, from ingest to purge?
- How do I classify data efficiently?
- How do I make data available to my end users efficiently?

**Define Business Metrics & Glossary**

**Ingest & Prepare: Landing Area**

**Analyze, Discover, Search Data**

**Clean, Transform, Refine Data**

**Deliver Visualizations, Analytics, Reporting Across Systems**

HADOOP DATA GOVERNANCE & MANAGEMENT
Stewardship Use Cases

Stewardship
Manage and organize data assets at Hadoop scale

- How do efficiently manage data lifecycle, from ingest to purge?
- How do I classify data efficiently?
- How do I make data available to my end users efficiently?

Define Business Metrics & Glossary
- informatica
- Collibra
- Data Advantage Group
- adaptive
- IBM

Ingest & Prepare: Landing Area
- informatica
- syncsort
- pentaho
- IBM
- TRIFACTA
- talend
- Paxata

Analyze, Discover, Search Data
- informatica
- pentaho
- WATERLINE DATA SCIENCE
- TRIFACTA
- Datameer
- Paxata

Clean, Transform, Refine Data
- informatica
- pentaho
- TRIFACTA
- Datameer
- platfora
- IBM

Deliver Visualizations, Analytics, Reporting Across Systems
- tableau
- SAP BusinessObjects
- zoomdata
- platfora
- SAS

HADOOP DATA GOVERNANCE & MANAGEMENT
Stewardship Case Study

- Fortune 500 financial services company with a strong data-driven culture was building an extremely rich data repository that included website traffic, audio calls, credit card transactions, ACH payments, credit scores, and more
- End users didn’t know which data assets were available, how to use them, or who to contact with questions
- Administrators were concerned about data quality, compliance and potential misuse of data
- This is really a case study in using stewardship to enable compliance & productivity

Problem

- Data scientists needed to find, trust, collaborate on, and use datasets that were growing by 20,000 files per day
- All access needed to comply with regulation

Solution

- Built a stewardship portal and discovery platform
- Classified data upon ingest: PII flagging, risk level assignment, business definitions
- Automated the stewardship workflow

Results

- Created a single, trusted catalog for hundreds of data scientists and dozens of stewards
- Delivered automatic enforcement of metadata and governance requirements
Administration Use Cases: Performance

**Administration**
- Boost user productivity and cluster performance

  - How is data being used today?
  - How can I optimize for future workloads?
  - How can I quickly take advantage of Hadoop risk-free?

**Visibility**
- Distribution of data objects
- Workloads by engine

**Patterns**
- Data churn over time
- Table clusters
- Frequent users

**Optimization**
- Sub-optimal query patterns
- “Rogue” users
- Capacity planning
Administration Use Cases: Recovery

1. All Hive tables are suddenly missing

2. Use audit information to identify who ran `rm -rf /user/hive/warehouse/`

3. Recreate databases and tables using metadata and lineage
Integration Use Case: Lineage Augmentation

• Custom applications and ISVs build applications on top of Hadoop
  • e.g., Trifacta, Syncsort, Talend, Cask
• These applications need governance, too

Problem
• Holistic and consistent governance is required across all high-level and low-level tools

Solution
• Programmatically link application operations with the underlying Hadoop operations

Benefits
• Hadoop developers don't need to build their own governance solution
• Metadata is unified across the stack
Integration Use Case: Lineage Export

• Enterprises have more data systems than just Hadoop
  • Need enterprise-wide metadata management system
  • Export Hadoop lineage via integration layer
  • Flexible APIs that support both bulk and incremental exports

• Navigator SDK
  • Cloudera has introduced the Navigator SDK, an Apache-licensed set of APIs, available at http://github.com/cloudera/navigator-sdk
Lessons Learned

• Compliance and productivity can be complementary

• Governance is the foundation
  • Improve performance
  • Drive adoption
  • Protect against catastrophes

• Hadoop governance is a part of a broader ecosystem
  • Look for interoperability with many different tools
  • Metadata must flow freely between enterprise systems
Cloudera Navigator
Overview & Demo
Please stop by the Cloudera booth

• See a demo of Cloudera Enterprise, including our governance solution that’s used by hundreds of production customers!

• Find out what makes Cloudera Enterprise the only PCI-certified Hadoop distribution

• Learn about our vast partner ecosystem
Questions
Thank You!

md@cloudera.com
chang@cloudera.com