How to avoid building a data swamp

Case studies in Hadoop data management and governance

Mark Donsky, Product Management, Cloudera
Naren Korenu, Engineering, Cloudera
Abstract

How can you make sure your data doesn’t get lost in Apache Hadoop? How can users across your company find the data they care about, and know that it’s trustworthy? How can you protect yourself against a data breach? How can you ingest more and more data while meeting these needs?

To take advantage of the rising volume, variety, and speed of data, companies must rethink their data management architectures so they can meet current and future business needs.

Find out how some of the world’s most sophisticated Hadoop deployments are addressing these data challenges head-on, while preserving Hadoop’s flexibility, through an integrated data management and governance approach for Hadoop. We’ll discuss how users can discover, trust, protect, and govern the data that matters most, for its entire lifecycle. This presentation will feature a live demo and cover audit, lineage, unified metadata, policy management, as well as leading partner integrations for end-to-end visibility.
The benefits of Hadoop...

One place for unlimited data

- All types
- More sources
- Faster, larger ingestion

Unified, multi-framework data access

- More users
- More tools
- Faster changes
...can create a data swamp: leading to trust, visibility, and governance challenges

<table>
<thead>
<tr>
<th>Compliance Officer</th>
<th>Data Stewards &amp; Curators</th>
<th>Data Scientists &amp; BI Users</th>
<th>Hadoop Admins &amp; DBAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Track, understand and protect access to data</td>
<td>Manage and organize data assets at Hadoop scale</td>
<td>Effortlessly find and trust the data that matters most</td>
<td>Boost user productivity and cluster performance</td>
</tr>
<tr>
<td>Am I prepared for an audit?</td>
<td>How do efficiently manage data lifecycle, from ingest to purge?</td>
<td>How can I explore data on my own?</td>
<td>How is data being used today?</td>
</tr>
<tr>
<td>What are they doing with the data?</td>
<td>How do I make data available to my end users efficiently?</td>
<td>How do I use what I find?</td>
<td>How can I quickly take advantage of Hadoop risk-free?</td>
</tr>
<tr>
<td>Is sensitive data governed and protected?</td>
<td>How do I find and use related data sets?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hadoop needs a governance foundation

Compliance Officer
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?

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

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

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

Hadoop Governance Foundation
- Unified metadata
- Unified lineage
- Unified auditing
- Universal policies
Enterprise metadata is the foundation of governance

Metadata enables you to put context and meaning to data

- **Business**
  - Business Glossary
  - Enterprise Taxonomy
  - Ontology

- **Technical**
  - Database Schema
  - File Definition
  - ETL Job Design
  - BI Report Definition
  - Data Model

- **Operational**
  - Job Run-Time Stats
  - Report Run Information
  - Hardware Usage
  - Scheduler Stats

---

Audit Logs  
Lineage  
Policies  
Technical Metadata  
Business Metadata
Lineage provides context

• Lineage describes the relationships between data sets:
  • Provenance: what data sets were used to create this data set?
  • Impact: what data sets are derived from this data set?

• ETL-derived lineage doesn’t cover end-user transformations, yet you typically need to capture 100% of lineage for compliance

• Lineage is embedded in Hadoop technical metadata, but it’s hard to access
Data policies enable scalability and predictability

• Information is of limited use unless it is actionable

• There is a treasure trove of actionable information in the metadata that the various Hadoop services emit
  • Archival of unused data (date created > 7 years ago)
  • Encryption of sensitive data (tags:sensitive)
  • Remediation of incorrect permissions (permissions:rwxrwxrwx and tags:sensitive)

• Triggers should be configurable based on user-defined criteria

• Hadoop does not offer a sufficient policy engine or action framework
Hadoop Governance Case Studies
Compliance Use Cases

Bringing together Hadoop data governance & management, enterprise metadata repository, and enterprise auditing & security.

**Compliance Officer**
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**

**Enterprise Auditing & Security**

**Common use cases:**
- Security breach detection
- Data access/usage for PCI compliance
Stewardship & Curation Use Cases

Data Stewards & Curators
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 & Curation Use Cases

Data Stewards & Curators
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
- IBM

Ingest & Prepare: Landing Area

- Informatica
- Synsor
- Pentaho
- IBM
- Trifacta
- Talend
- Paxata

Analyze, Discover, Search Data

- Informatica
- Talend
- Pentaho
- IBM
- Trifacta
- Datameer
- Paxata

Clean, Transform, Refine Data

- Informatica
- Talend
- Pentaho
- IBM
- Trifacta
- Datameer
- Paxata

Deliver Visualizations, Analytics, Reporting Across Systems

- Tableau
- SAP BusinessObjects
- QOMVIA
- platfora
- SAS

HADOOP DATA GOVERNANCE & MANAGEMENT
A few more data management use cases

Self-service discovery portal built on top of curated artifacts
- Curators set up business metadata
- IT creates a portal using the business metadata
- This use case is common in healthcare and pharmaceuticals

Recreating deleted artifacts
- Someone runs `rm -rf` on their directory
- Use lineage and audit queries to reconstruct data sets
Metadata interoperability is most important

- Governing only Hadoop is never enough
  - All metadata must flow freely between enterprise systems, throughout the entire organization
    - From Hadoop to enterprise lineage system
    - From Hadoop to SIEM tools for threat detection

- This requires open interoperability with all the leading tools

- Cloudera has introduced the Navigator SDK, an Apache-licensed set of APIs for interoperability, available at http://github.com/cloudera/navigator-sdk
Cloudera Navigator
Overview & Demo
Learn more

Please stop by the Cloudera booth!

• See a demo of Cloudera Enterprise, including our governance solution that’s used by over 200 production customers for over two years!
• Find out what makes Cloudera Enterprise the only PCI-certified Hadoop distribution
• Learn about our 2000+ partner ecosystem
Questions
Thank You!

md@cloudera.com
naren@cloudera.com