Big Data Governance for the Hybrid Cloud

Best Practices for Data Governance

Mark Donsky, md@cloudera.com
Compliance + Productivity = Hadoop Adoption

End-User Productivity
- How can I find and explore data sets 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 & Compliance
- Am I prepared for an audit?
- Who’s accessing sensitive data?
- What are they doing with the data?
- Is sensitive data governed and protected?
Governance is the Foundation of Data Management

<table>
<thead>
<tr>
<th>Compliance</th>
<th>Stewardship</th>
<th>End User Productivity</th>
<th>Administration</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 can I efficiently manage data lifecycle, from ingest to purge?</td>
<td>How can I find explore data sets on my own?</td>
<td>Is my data optimized to support current access patterns?</td>
</tr>
<tr>
<td>Who’s accessing sensitive data?</td>
<td>How can I efficiently organize and classify all my data?</td>
<td>Can I trust what I find?</td>
<td>How can I optimize for future workloads?</td>
</tr>
<tr>
<td>What are they doing with the data?</td>
<td>How can I efficiently make data available to my end users?</td>
<td>How do I use what I find?</td>
<td>How can I migrate workloads to Hadoop risk-free?</td>
</tr>
<tr>
<td>Is sensitive data governed and protected?</td>
<td></td>
<td>How do I find and use related data sets?</td>
<td></td>
</tr>
</tbody>
</table>

**Hadoop Governance Foundation**

- Centralized audits
- Unified data catalog
- Comprehensive lineage
- Data policies
What makes governance so difficult?

### Hadoop governance challenges
- Variety, Volume, Velocity
- Multiple compute types: Spark, Hive, Pig, MR, MR2, Sqoop, etc.
- Multiple third-party tools

### Cloud governance challenges
- Multiple storage types: HDFS, S3, ADLS, etc.
- Transient clusters
- Long-running clusters
- Shared Hive Metastores

Yet the business still needs one set of trusted governance artifacts
Requirements for Successful Big Data Governance

- Both compliance and end-user productivity needs must be addressed
- Observation is better than disclosure
- Interoperability and extensibility are critical: one size doesn’t fit all
- All data must be governed, whether it’s on-prem, in the cloud or mixed
Common Governance
Use Cases
Data Stewardship and Governance Activities

Project management
Policy management
RACI
Stewardship workflows
ETL
Centralized curation
Centralized glossaries

Privitar
Datagise

Security profiling
Compliance: BCBS239, GDPR

Informatica
Talend

Data quality
Uniqueness
Data valuation
Data profiling
Content enrichment

Unified technical metadata catalog
Extensible business metadata and glossary
Metadata rules engine
Comprehensive lineage
Unified audit/access logs
Dashboards and analytics
APIs for augmentation and consumption

End user collaboration
Crowdsourced metadata

Enterprise aggregation: metadata, lineage, SIEM, auditing

© Cloudera, Inc. All rights reserved.
Use Cases: Compliance

Compliance
Track, understand and protect access to data

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

ENTERPRISE METADATA REPOSITORY

ENTERPRISE AUDITING & SECURITY

HADOOP DATA GOVERNANCE & MANAGEMENT

Common use cases:
- Security breach detection
- Data access tracking for PCI compliance
- Audit defense
Use Cases: Stewardship

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

- How can I efficiently manage data lifecycle, from ingest to purge?
- How can I efficiently organize and classify all my data?
- How can I efficiently make data available to my end users?

---

Define Business Metrics & Glossary

Deliver Visualizations, Analytics, Reporting Across Systems

Ingest & Prepare: Landing Area

Analyze, Discover, Search Data

Clean, Transform, Refine Data

HADOOP DATA GOVERNANCE & MANAGEMENT

© Cloudera, Inc. All rights reserved. 9
Use Cases: Stewardship

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

- How can I efficiently manage data lifecycle, from ingest to purge?
- How can I efficiently organize and classify all my data?
- How can I efficiently make data available to my end users?

**Define Business Metrics & Glossary**

- informatica
- Collibra
- adaptive
- IBM

**Ingest & Prepare: Landing Area**

- informatica
- syncsort
- IBM
- TRIFACTA
talend
Paxata

**Analyze, Discover, Search Data**

- informatica
- trillium software
- Datameer
- Pentalo
- IBM

**Clean, Transform, Refine Data**

- informatica
- TRIFACTA
- Datameer
- Paxata
- IBM

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

- Tableau
- SAP BusinessObjects
- Qumulo
- platforma
- SAS

HADOOP DATA GOVERNANCE & MANAGEMENT
Use Cases: Administration

**Administration**
Boost user productivity and cluster performance

- Is my data optimized to support current access patterns?
- How can I optimize for future workloads?
- How can I migrate workloads to Hadoop risk-free?

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

**Optimization**
- Sub-optimal query patterns
- “Rogue” users
- Capacity planning

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

**Unexpected Behaviour**
- Hive tables suddenly missing
- `rm -rf /usr/hive/warehouse`
Governance Maturity Progression

1. Initial
   - Chaos: “We don’t know what’s in our data hub”

2. Compliance
   - Basic compliance: Raw governance artifact capture

3. Discovery & Collaboration
   - Self-service data catalog: Data curation and business glossary
   - This is where most big data deployments need to focus

4. Data Stewardship
   - Information lifecycle automation: Data stewardship and lifecycle automation

5. Optimization & Refactoring
   - Continuous improvement: ongoing optimization
Product Demo
Thank you!

md@cloudera.com

@markdonsky