Big Data Governance for the Hybrid Cloud

Best Practices for Data Governance

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

**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?

**End-User Productivity**
- How can I find 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 is the Foundation of Data Management

**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?

**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?

**End User Productivity**
Effortlessly find and trust the data that matters most
- How can I find 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?

**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?

---

**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
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
- Informatica
- Data Advantage Group
- IBM
- Adaptive

Enterprise Auditing & Security
- Splunk
- Imperva
- IBM Security

Hadoop Data Governance & Management
- Unified metadata
- Unified lineage
- Unified auditing

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

Ingest & Prepare: Landing Area

Analyze, Discover, Search Data

Clean, Transform, Refine Data

Deliver Visualizations, Analytics, Reporting Across Systems

Hadoop Data Governance & Management
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
- IBM
- Data Advantage Group
- adaptive

Deliver Visualizations, Analytics, Reporting Across Systems

- Tableau
- SAP BusinessObjects
- Tibco Spotfire
- plattora
- SAS

Ingest & Prepare: Landing Area

- Informatica
- Syncsort
- Pentaho
- IBM
- Talend
- Trifacta
- Paxata

Analyze, Discover, Search Data

- Informatica
- Talend
- Pentaho
- IBM
- Waterline Data Science
- Trifacta
- Datameer
- Paxata

Clean, Transform, Refine Data

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

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

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

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

**Unexpected Behaviour**
- Hive tables suddenly missing
- `rm -rf /usr/hive/warehouse`
Big Data Governance
Best Practices
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
   - Business metadata for self-service: Data curation automation

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

5. Optimization & Refactoring
   - Continuous improvement: ongoing optimization
## Big Data Governance Best Practices

<table>
<thead>
<tr>
<th>People</th>
<th>Processes</th>
<th>Product</th>
<th>Initial</th>
<th>Basic Compliance</th>
<th>Discovery &amp; Collaboration</th>
<th>Data Stewardship</th>
<th>Optimization &amp; Refactoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited awareness of the purpose or value of data governance</td>
<td>Everything is ad-hoc</td>
<td>No product support</td>
<td>Limited awareness of the purpose or value of data governance</td>
<td>Data governance group Captures raw governance artifacts to ensure compliance/full paper trails</td>
<td>Data curation group Implements an organization-wide business metadata classification scheme</td>
<td>Data stewardship group Automates all ingest-to-purge lifecycle activities</td>
<td>Ongoing involvement from business executives and subject matter experts</td>
</tr>
<tr>
<td>No formal roles defined for data governance</td>
<td>Basic governance artifact capture for compliance</td>
<td>- Unified auditing</td>
<td>- Technical metadata catalog generation</td>
<td>- Technical metadata catalog generation</td>
<td>- Business metadata</td>
<td>- Lifecycle policy actions, such as retention and purge</td>
<td>- Purpose-built add-on tools</td>
</tr>
<tr>
<td>- Lineage collection</td>
<td>- Unified audit collection</td>
<td>- Technical metadata catalog</td>
<td>- Comprehensive lineage</td>
<td>- Integration with enterprise frameworks</td>
<td>- End-user custom metadata</td>
<td>- Purpose-built add-on tools</td>
<td>- Purpose-built add-on tools</td>
</tr>
<tr>
<td>- Data governance group</td>
<td>- Discovery</td>
<td>- Collaborative tagging</td>
<td>- Metadata assignment on ingest</td>
<td>- Metadata assignment on ingest</td>
<td>- Automated authorization configuration</td>
<td>- Intelligent administration</td>
<td>- Continuous improvement</td>
</tr>
<tr>
<td>- Data curation automation</td>
<td>- Self-service</td>
<td>- Data quality and profiling</td>
<td>- Ingest-to-purge lifecycle automation</td>
<td>- Ingest-to-purge lifecycle automation</td>
<td>- Dynamic policy enforcement</td>
<td>- Ongoing refactoring of Levels 3+</td>
<td>- Dynamic threat detection</td>
</tr>
</tbody>
</table>
Thank you!
md@cloudera.com
@markdonsky