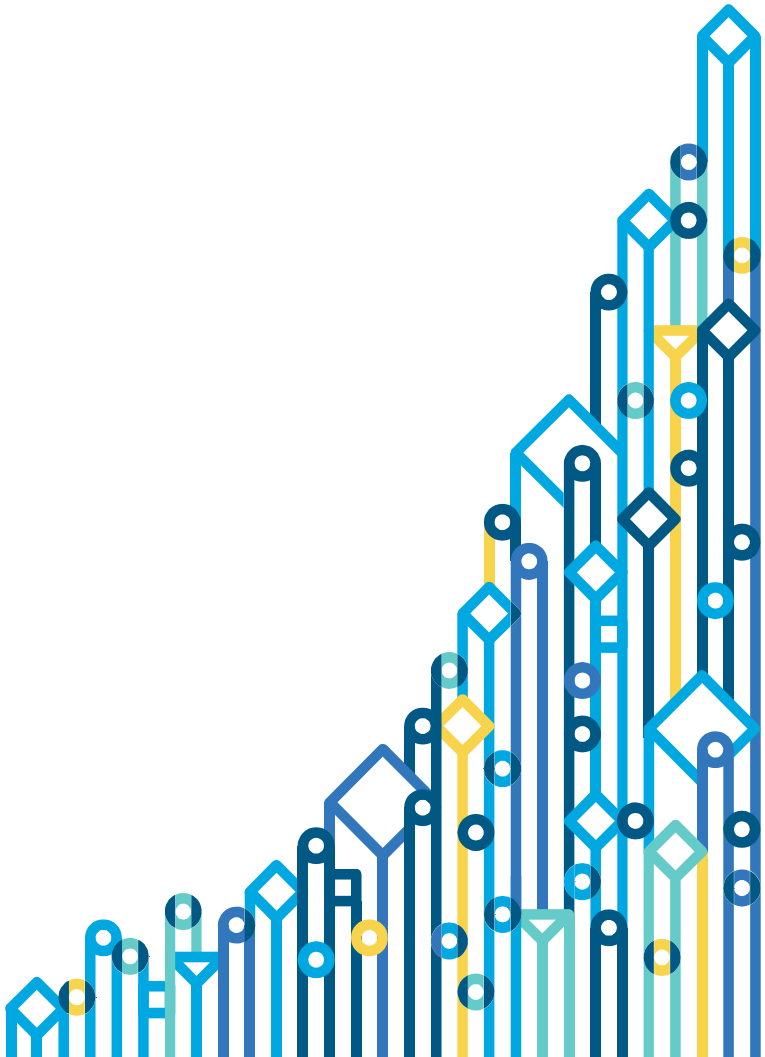
cloudera

Big Data Governance for the Hybrid Cloud

Best Practices for Data Governance

Mark Donsky, md@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?

cloudera

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

cloudera

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

cloudera

Common Governance Use Cases

Data Stewardship and Governance Activities

Application

collibra podium data informatica talend

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

Data quality
Uniqueness
Data valuation
Data profiling
Content enrichment

Alation Paxata Waterline Data TRIFACTA ARCADIA DATA

tableau SOFTWARE Qlik

Data wrangling
Data visualization
Query recommendations

Platform

PRIVITAR DATAGUISE

Security profiling
Compliance: BCBS239, GDPR

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

cloudera NAVIGATOR

Enterprise aggregation: metadata, lineage, SIEM, auditing

informatica Ab Initio IBM

Centralized Stewardship ⟷ End User Discovery

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 · RSA 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

cloudera

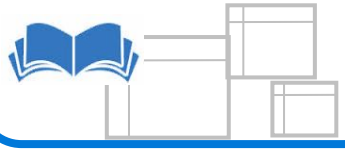
Use Cases: Stewardship

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

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 · Data Advantage Group · Collibra · adaptive · IBM

Deliver Visualizations, Analytics, Reporting Across Systems

tableau · SAP BusinessObjects · ZOOMDATA · platfora · sas

Ingest & Prepare: Landing Area

informatica · pentaho · syncsort · IBM · TRIFACTA · talend · Paxata

Analyze, Discover, Search Data

informatica · talend · WATERLINE DATA SCIENCE · pentaho · TRIFACTA · IBM · Datameer · Paxata

Clean, Transform, Refine Data

informatica · talend · pentaho · TRILLIUM SOFTWARE · Datameer · TRIFACTA · platfora · IBM

HADOOP DATA GOVERNANCE & MANAGEMENT

cloudera

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

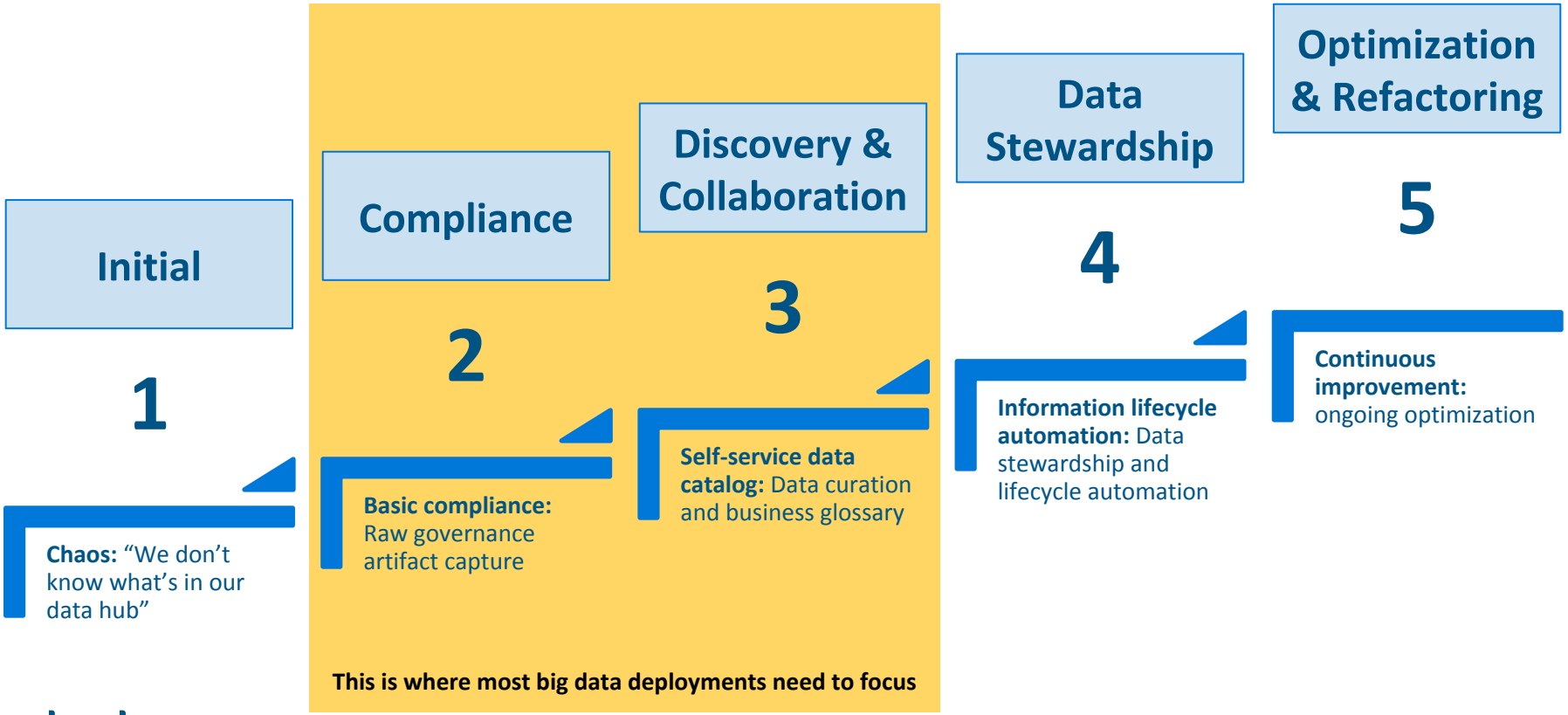
- Hive tables suddenly missing
- rm –rf /usr/hive/warehouse

cloudera

cloudera

Big Data Governance Best Practices

Governance Maturity Progression



cloudera

Product Demo

cloudera

Thank you!

md@cloudera.com

@markdonsky