Enhance your data lake with comprehensive data governance to improve adoption and meet compliance needs

Sanjeev Mohan
Gartner
Sept 12, 2018
But, What Is Metadata?

- Fast Food
- Savory
- 750 calories
- $5

- McDonald's
- Uber Eats
- Carrot sticks
- Chocolate Milkshake
Agenda

- Why is metadata important?
- Conceptual analytics workflow
- Reference architecture for metadata management
# Types of Metadata

<table>
<thead>
<tr>
<th>Technical (Definitional)</th>
<th>Operational (Descriptive)</th>
<th>Business (Descriptive)</th>
<th>Social (Descriptive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schemas</td>
<td>Output from processes</td>
<td>Ontology — classify and tag data</td>
<td>Metadata about party data relationships</td>
</tr>
<tr>
<td>Data types</td>
<td>ETL or actions on data</td>
<td>Metadata mapped to business relationships</td>
<td>User-generated content</td>
</tr>
<tr>
<td>Data models</td>
<td>Data lineage</td>
<td>Data provenance (reproducibility)</td>
<td>Tribal knowledge</td>
</tr>
<tr>
<td>Configurations</td>
<td>Data provenance</td>
<td>Performance</td>
<td></td>
</tr>
<tr>
<td>Functions</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
It's a Regulated World … We Just Live in It!

Your customers may:
- Demand to "clearly" see information collected about them (in 30 days)
- Change their "consent"
- Ask to be "forgotten"
- Invoke data **portability** option

Your architecture must:
- Be **designed** with security
- **Match** and **merge** identities
- Display end-to-end **lineage** and establish provenance
- **Reproduce** pipeline processes
Good news: Business and IT are now completely aligned
Data Catalog — Index and Search Metadata

Capabilities of a Data Catalog Solution

Curate inventory of information assets

Collaborate for accountability and governance

Communicate shared semantic meaning

Facilitate, Broker, Enable, Share, Orchestrate

Source: "Data Catalogs Are the New Black in Data Management and Analytics," (G00338777)
Data Lake Reference Architecture

Acquire

- RDBMS/EDW
- Logs/Emails
- Social Media
- IoT Sensor
- File

Data Lake Reference Architecture

Organize

- Raw/Landing/Secure Zone
  - CDC/Batch
  - Batch
  - Streaming
  - Spark
  - Custom
  - Kafka
  - Custom
  - MQTT
  - Custom
  - SFTP

- Enriched/Discovery Zone (Data Transformation)
  - HDFS/S3/DBMS
  - Hive/S3/DBMS

- Consumption Zone
  - HBase/S3/DBMS

Analyze

- Downstream Applications
- Operational Analytics
- Self-Service Dashboards
- BI Analysts
- Advanced Analytics
- Compliance Analytics
- BI Analysts
- Data Scientists
- Data Stewards

Source: "Applying Effective Data Governance to Secure Your Data Lake." (G00346975)
Unified Data Governance Reference Architecture

- **Raw/Landing/Secure Zone**: HDFS/S3/DBMS
- **Enriched/Discovery Zone (Data Transformation)**: Hive/S3/DBMS
- **Consumption Zone**: HBase/S3/DBMS
- **Self-Service Data Prep.**: MDM
- **Data Wrangling**: Profile, Classify, Tokenize, Masking
- **Data Quality**: Tools for data profiling, classification, tokenization, and masking
- **Data Access**: Encryption, Indexing, Searching, Cataloging, Metadata, Lineage, Auditing
- **Downstream Applications**: Operational Analytics, Self-Service Dashboards, Advanced Analytics, Compliance Analytics
- **Operational Analytics**: Tools for developer, data analysts, BI analysts, data scientists, data stewards

**Key Technologies**:
- S3 = Amazon Simple Storage Service
- Hive = Apache Hive
- HBase = Apache HBase

**Security and Governance**:
- API Governance: AD/LDAP/Kerberos
- RBAC, ABAC

**Data at Rest**
- RDBMS/EDW
- Logs/Emails
- Social Media
- IoT Sensor
- File

**Data in Motion**
- CSV
- Encrypt
- Encrypt
- Index
- Search
- Catalog

**Data Wrangling**
- Profile
- Classify
- Tokenize
- Masking

**Data Quality**
- Tools for data profiling, classification, tokenization, and masking

**Data Access**
- Encryption, Indexing, Searching, Cataloging, Metadata, Lineage, Auditing

**Operational Analytics**
- Tools for developer, data analysts, BI analysts, data scientists, data stewards

**API Governance**
- AD/LDAP/Kerberos
- RBAC, ABAC

**Key Definitions**
- S3 = Amazon Simple Storage Service
- Hive = Apache Hive
- HBase = Apache HBase
Invest early in metadata management, preferably under the leadership of the CDO or similar

Begin data governance journey at the POC stage. Don't make it an afterthought

Maintain focus on improving data quality

Start with the use case driving greatest business value and demand

Add other use cases over time and across initiatives