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

Sanjeev Mohan

Gartner

Sept 12, 2018

Strata
DATA CONFERENCE

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

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
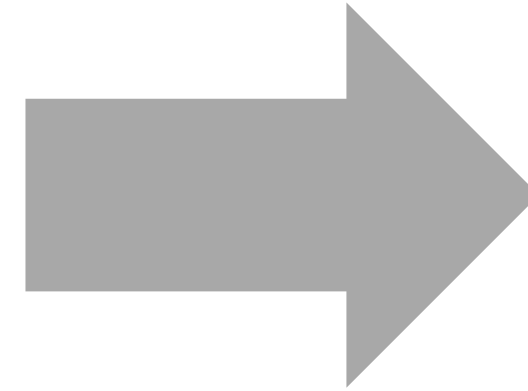
Your architecture must:

- Be **designed** with security
- **Match** and **merge** identities
- Display end-to-end **lineage** and establish provenance
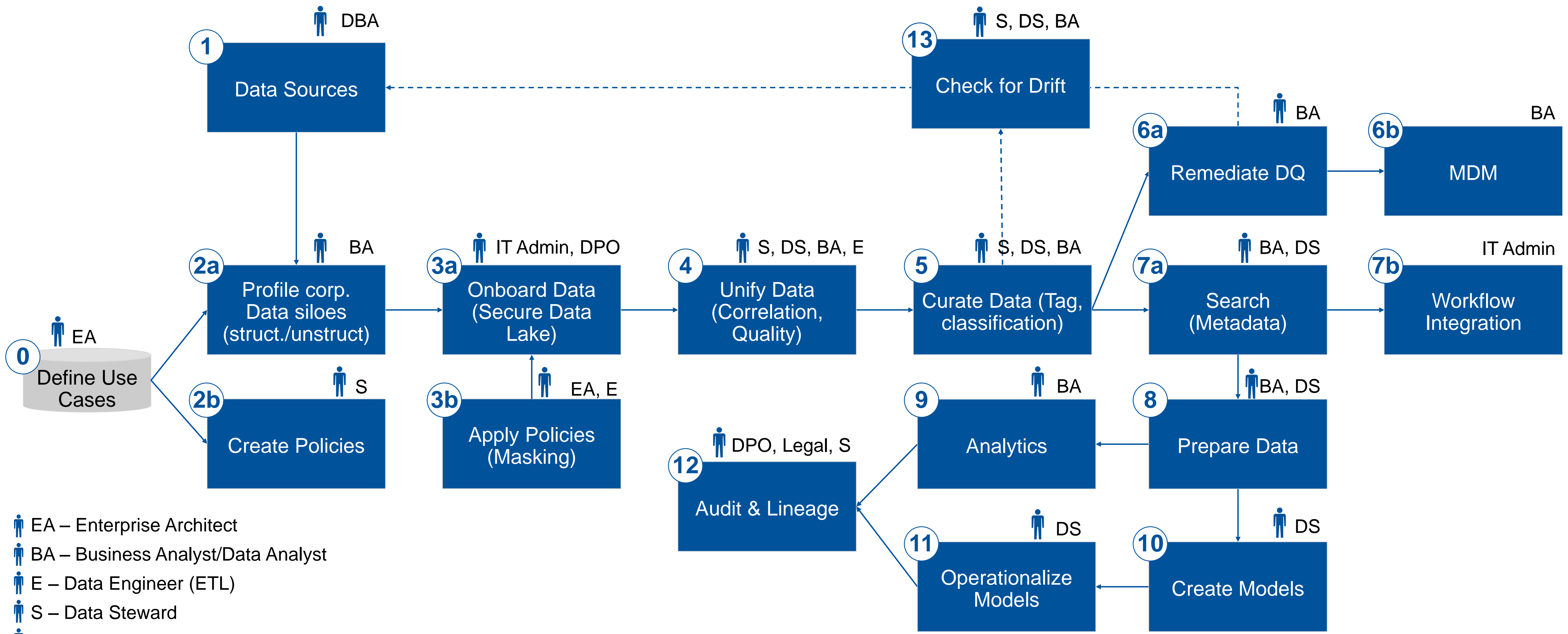- **Reproduce** pipeline processes

Strata
DATA CONFERENCE

Metadata Glues Together Analytics Workflow

1 — DBA — Data Sources

13 — S, DS, BA — Check for Drift

6a — BA — Remediate DQ

6b — BA — MDM

0 — EA — Define Use Cases

2a — BA — Profile corp. Data siloes (struct./unstruct)

2b — S — Create Policies

3a — IT Admin, DPO — Onboard Data (Secure Data Lake)

3b — EA, E — Apply Policies (Masking)

4 — S, DS, BA, E — Unify Data (Correlation, Quality)

5 — S, DS, BA — Curate Data (Tag, classification)

7a — BA, DS — Search (Metadata)

7b — IT Admin — Workflow Integration

8 — BA, DS — Prepare Data

9 — BA — Analytics

12 — DPO, Legal, S — Audit & Lineage

11 — DS — Operationalize Models

10 — DS — Create Models

EA – Enterprise Architect
BA – Business Analyst/Data Analyst
E – Data Engineer (ETL)
S – Data Steward
DS – Data Scientist
DBA – Database Administrator
DPO – Data Protection Officer

Gartner.

Data Catalog — Index and Search Metadata

Capabilities of a Data Catalog Solution

Curate inventory of information assets
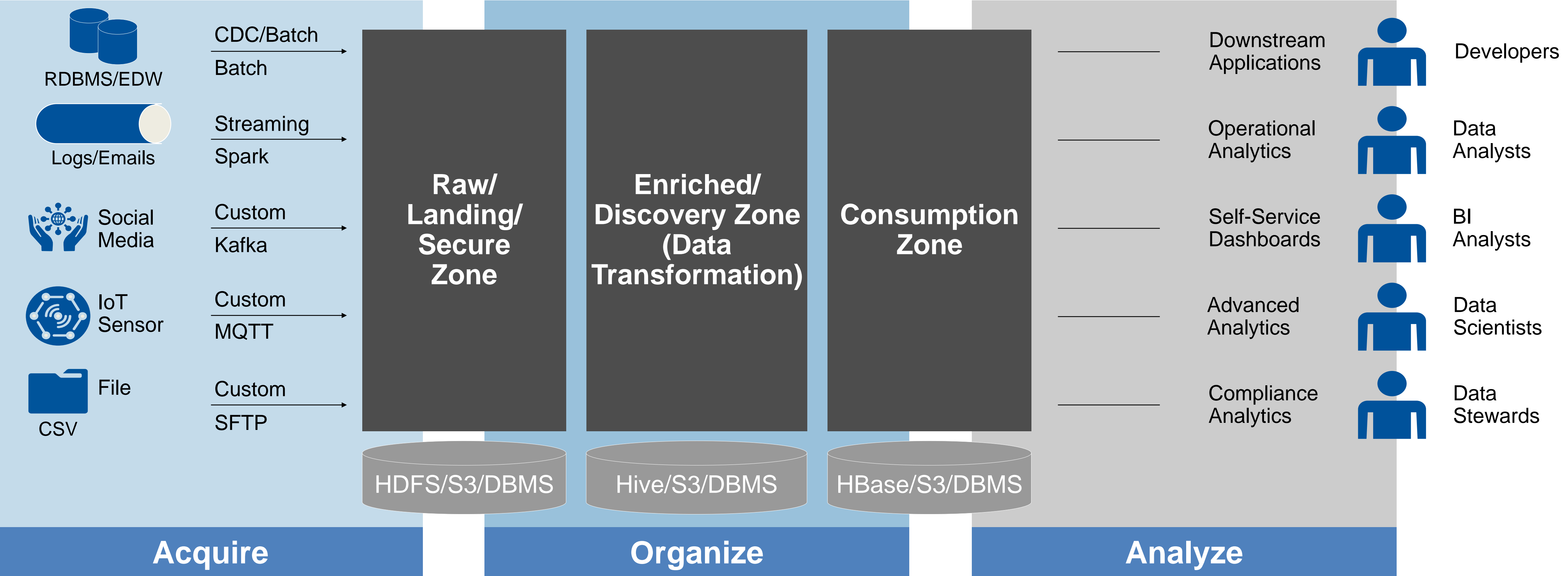
Collaborate for accountability and governance

Communicate shared semantic meaning

Facilitate, Broker, Enable, Share, Orchestrate

Gartner®

Data Lake Reference Architecture

Data Lake Reference Architecture

| Acquire | Organize | Analyze |

Acquire sources:
- RDBMS/EDW
- Logs/Emails
- Social Media
- IoT Sensor
- File / CSV

Transfer methods:
- CDC/Batch / Batch
- Streaming / Spark
- Custom / Kafka
- Custom / MQTT
- Custom / SFTP

Zones:
- Raw/Landing/Secure Zone — HDFS/S3/DBMS
- Enriched/Discovery Zone (Data Transformation) — Hive/S3/DBMS
- Consumption Zone — HBase/S3/DBMS

Analyze:
- Downstream Applications — Developers
- Operational Analytics — Data Analysts
- Self-Service Dashboards — BI Analysts
- Advanced Analytics — Data Scientists
- Compliance Analytics — Data Stewards

Source: "Applying Effective Data Governance to Secure Your Data Lake," (G00346975)

Unified Data Governance Reference Architecture
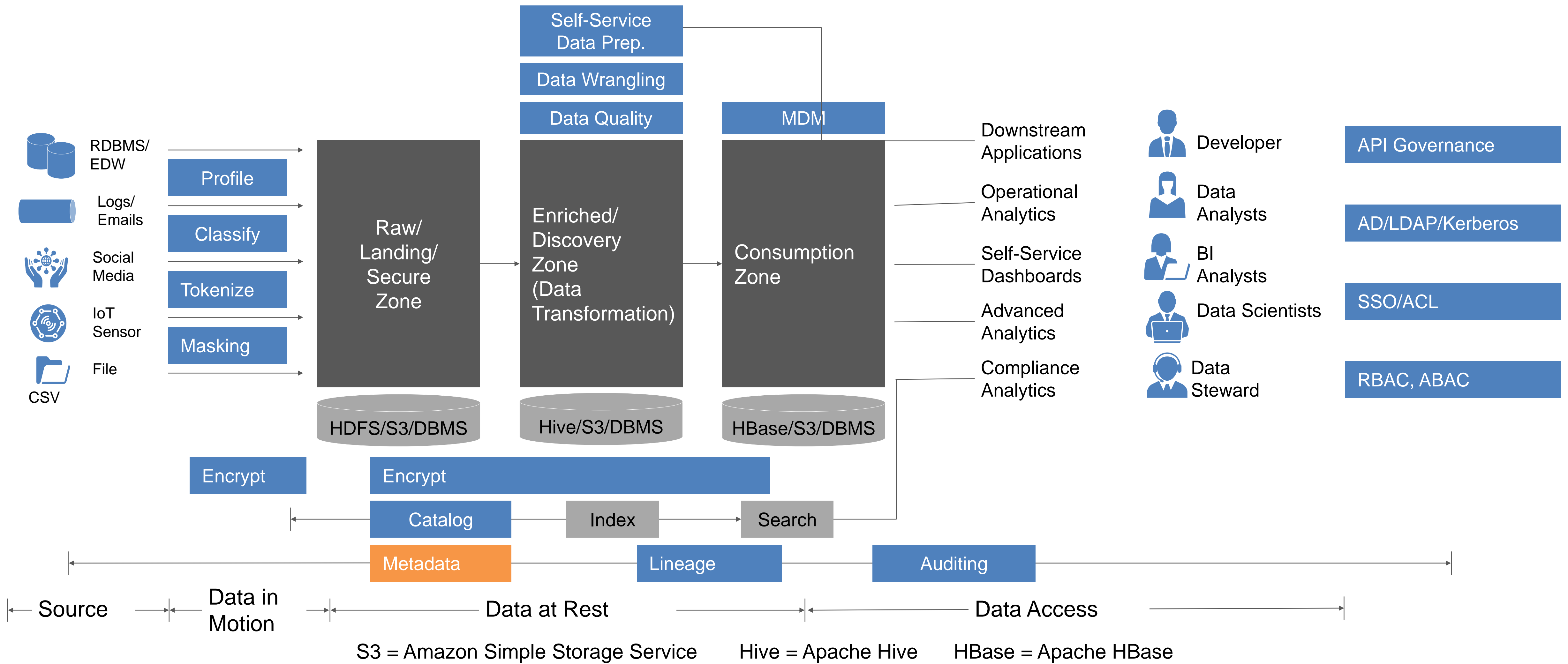
Recommendations

✓ 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