Enterprise Data and MDM
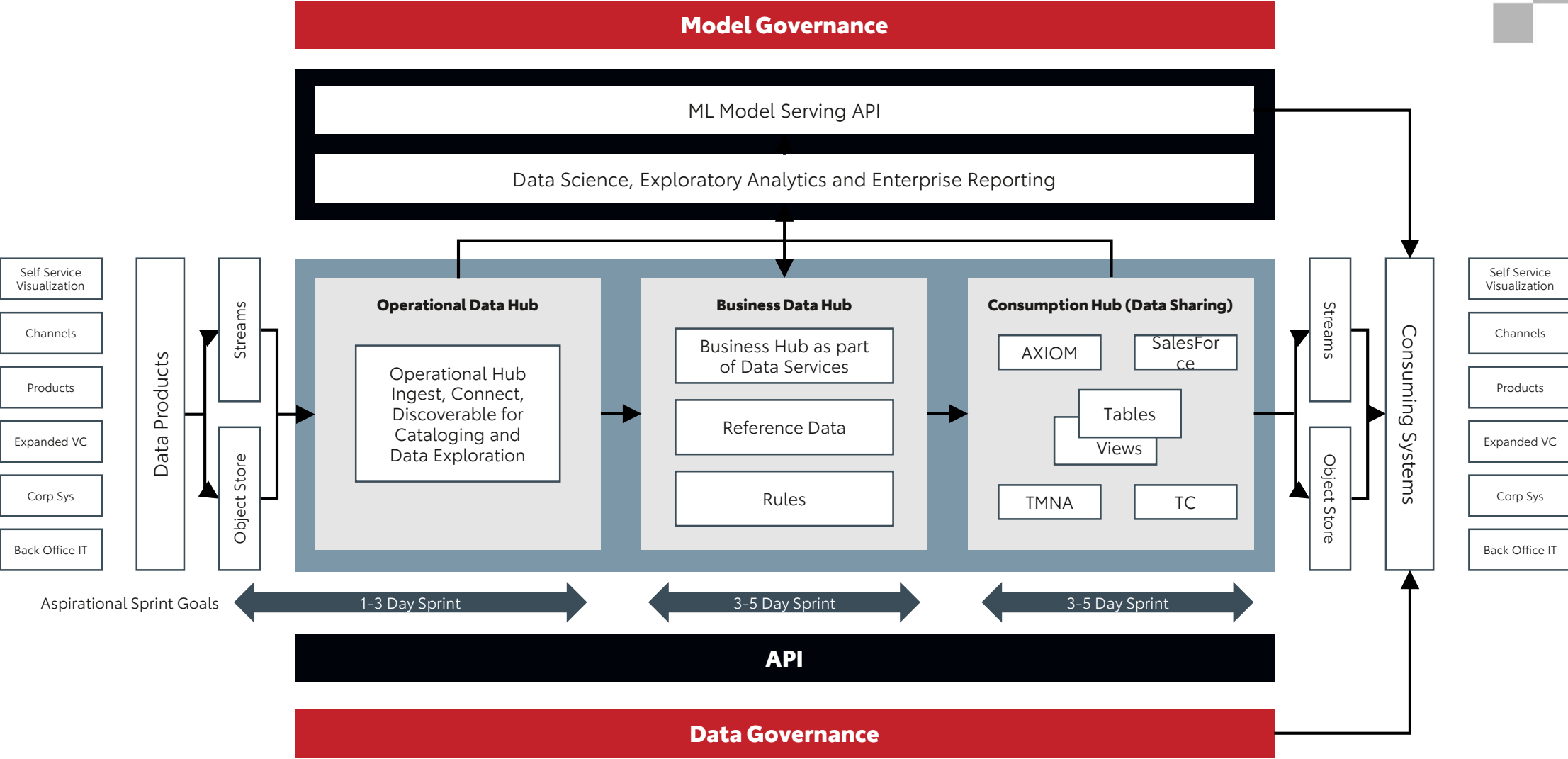
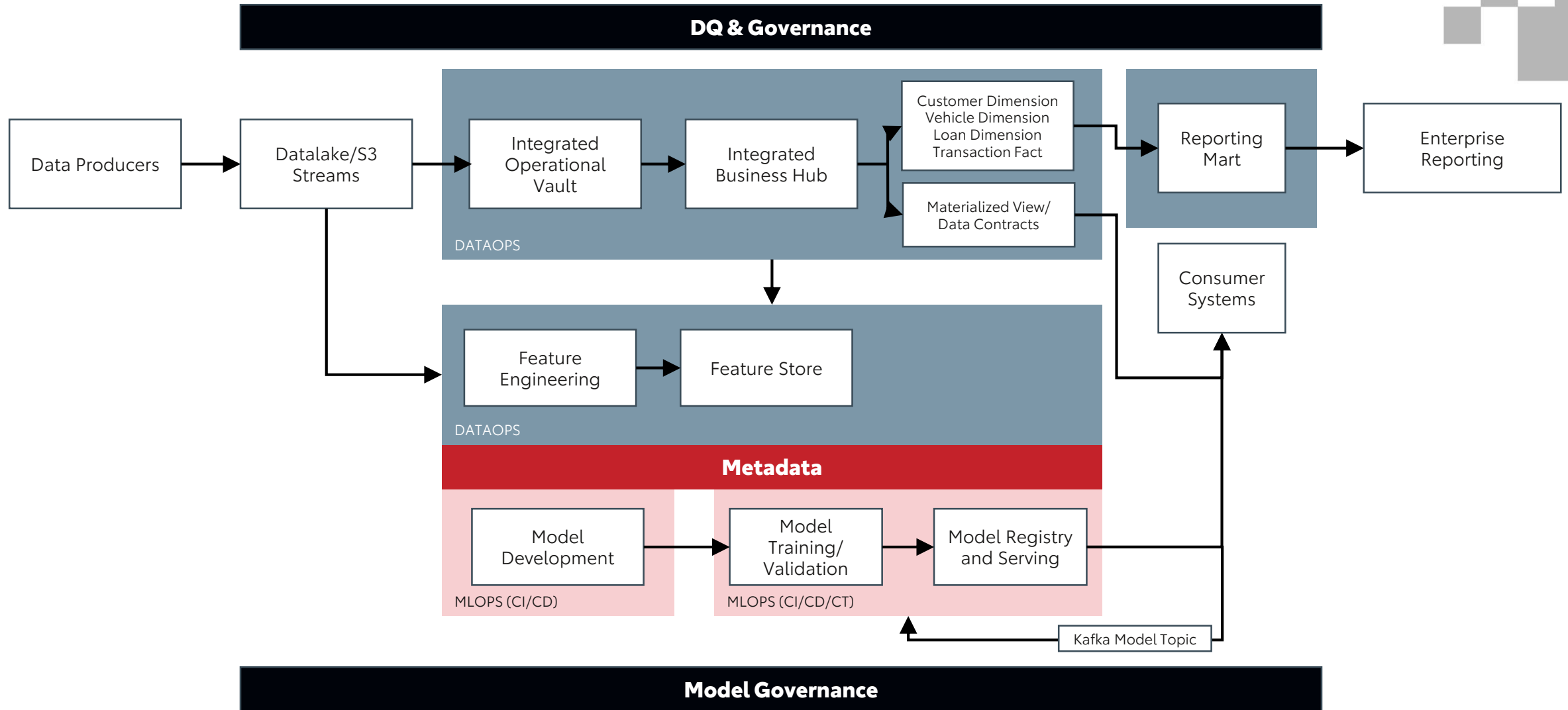# Clarifying Terms

Data Harmonization in an EDP Platform

▪ Data is in different formats with different field names and schemas in the various silos, but when it is ultimately served up, the records must look at least similar to be processed as a group. Harmonization falls into three categories:

1. **Naming Differences:** The simplest to handle, naming differences, occur when different silos have an identical value with identical meaning, but use a different name. E.g. "PERSON.firstName" vs. "IDENTITY.GIVEN_NAME" in two different source tables—but holding the same sort of data.

2. **Structural Differences:** Somewhat more complex than Naming Differences, structural differences occur when the number and combination of fields differ across silos. E.g. "boxes_available" in one system might be total boxes in the warehouse plus a count_per_box field to derive total items, but in another system, may directly represent total_items (regardless of boxing). Further, one may represent available inventory not already allocated to unfilled orders, where the other may represent all physical inventory, regardless of open orders.

3. **Semantic Differences:** These are the worst. One system may have three patient statuses, and another may have five. These statuses will often overlap and be hard to map to one another ({Scheduled, Needs_Followup, Inactive} vs. {Intake, Scheduled, Telemedicine-only, Discharged}).

Goal EDP 2.0 : Conceptual View

Model Governance

ML Model Serving API

Data Science, Exploratory Analytics and Enterprise Reporting

Self Service Visualization		Data Products	Streams	**Operational Data Hub**	**Business Data Hub**	**Consumption Hub (Data Sharing)**	Streams	Consuming Systems	Self Service Visualization

Operational Data Hub

Operational Hub Ingest, Connect, Discoverable for Cataloging and Data Exploration

Business Data Hub

Business Hub as part of Data Services

Reference Data

Rules

Consumption Hub (Data Sharing)

AXIOM | SalesForce

Tables

Views

TMNA | TC

Self Service Visualization

Channels

Products

Expanded VC

Corp Sys

Back Office IT

Data Products

Streams

Object Store

Consuming Systems

Self Service Visualization

Channels

Products

Expanded VC

Corp Sys

Back Office IT

Aspirational Sprint Goals

1-3 Day Sprint

3-5 Day Sprint

3-5 Day Sprint

API

Data Governance

Micro Batching and Streaming Framework

DQ & Governance

Data Producers → Datalake/S3 Streams

DATAOPS

Integrated Operational Vault → Integrated Business Hub → Customer Dimension / Vehicle Dimension / Loan Dimension / Transaction Fact

Materialized View/ Data Contracts

Reporting Mart → Enterprise Reporting

Consumer Systems

DATAOPS

Feature Engineering → Feature Store

Metadata

MLOPS (CI/CD)

Model Development → Model Training/ Validation → Model Registry and Serving

MLOPS (CI/CD/CT)

Kafka Model Topic

Model Governance

Reference Data Pipeline

Event Pattern Pipeline

Producer Systems

API

Data Share

Consumer Systems

APACHE nifi

Data Processing Pipeline (Automated Extract and Model Load)

Informatica
APACHE nifi

Data Processing: End System Transform

APACHE Spark™

Batch Pattern Pipeline

Platform Enablement Pipeline
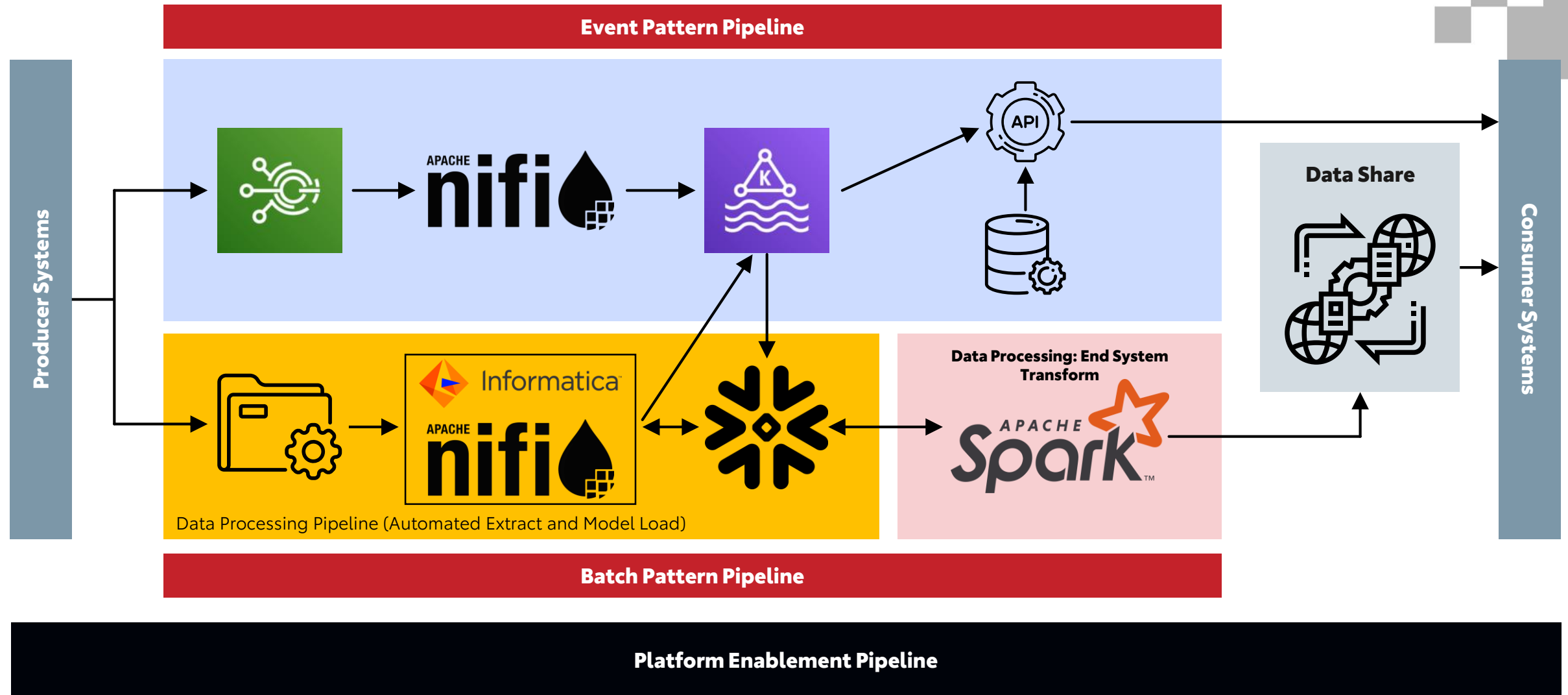
Data Vault Organization of Objects:

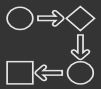HUB – Consists of unique list of Business Keys and Multi Tenancy

SATELLITE – Consists of descriptive data of parent Hub or Link that can change over time

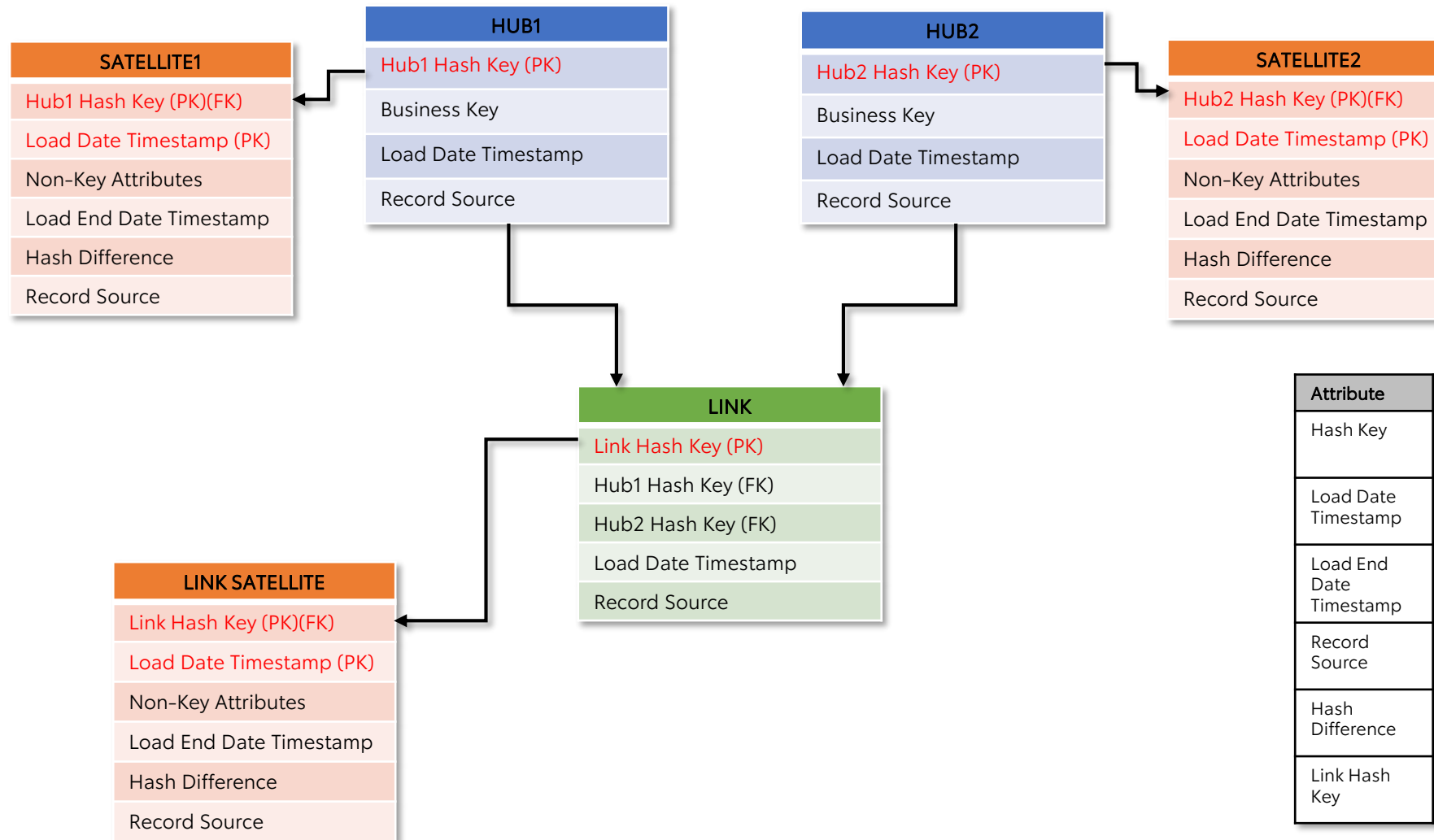LINK – Represents relationships/associations, hierarchies, transactions and events

SATELLITE
Hub Hash Key (PK)(FK)
Load Date Timestamp (PK)
Non-Key Attribute1
Non-Key Attribute2
Load End Date Timestamp
Hash Difference
Record Source

HUB
Hub Hash Key (PK)
Business Key
Load Date Timestamp
Record Source

LINK
Link Hash Key (PK)
Hub Hash Key (FK)
Load Date Timestamp
Record Source

Data Vault – Overview of Data Model

HUB1
- Hub1 Hash Key (PK)
- Business Key
- Load Date Timestamp
- Record Source

SATELLITE1
- Hub1 Hash Key (PK)(FK)
- Load Date Timestamp (PK)
- Non-Key Attributes
- Load End Date Timestamp
- Hash Difference
- Record Source

HUB2
- Hub2 Hash Key (PK)
- Business Key
- Load Date Timestamp
- Record Source

SATELLITE2
- Hub2 Hash Key (PK)(FK)
- Load Date Timestamp (PK)
- Non-Key Attributes
- Load End Date Timestamp
- Hash Difference
- Record Source

LINK
- Link Hash Key (PK)
- Hub1 Hash Key (FK)
- Hub2 Hash Key (FK)
- Load Date Timestamp
- Record Source

LINK SATELLITE
- Link Hash Key (PK)(FK)
- Load Date Timestamp (PK)
- Non-Key Attributes
- Load End Date Timestamp
- Hash Difference
- Record Source

Attribute	Description
Hash Key	Hash representation of Business Keys
Load Date Timestamp	Date and Time of data insert
Load End Date Timestamp	Date and Time of expiry data
Record Source	File name or Table name of the source system
Hash Difference	Hash representation of all non-key attributes and meta data
Link Hash Key	Hash representation of associative Hub Business Keys

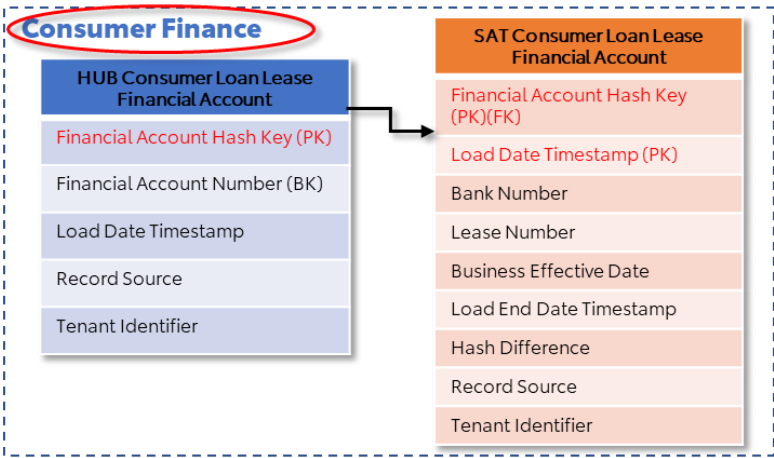# Multi Tenancy:

- Data segregation at Hub level with Tenant Identifier

- Different Vaults

- Different S3 buckets

Product Line of Business:

- Data segregation by Domain
  Example: Insurance, Commercial Finance, Consumer Finance, etc.

- Introducing Hubs, Satellites and Links by Product Line of Business

Harmonize & Materialize Data for Consumption

All data pulled into the Enterprise Data Platform will be harmonized and materialized for consumption by applications/factories.
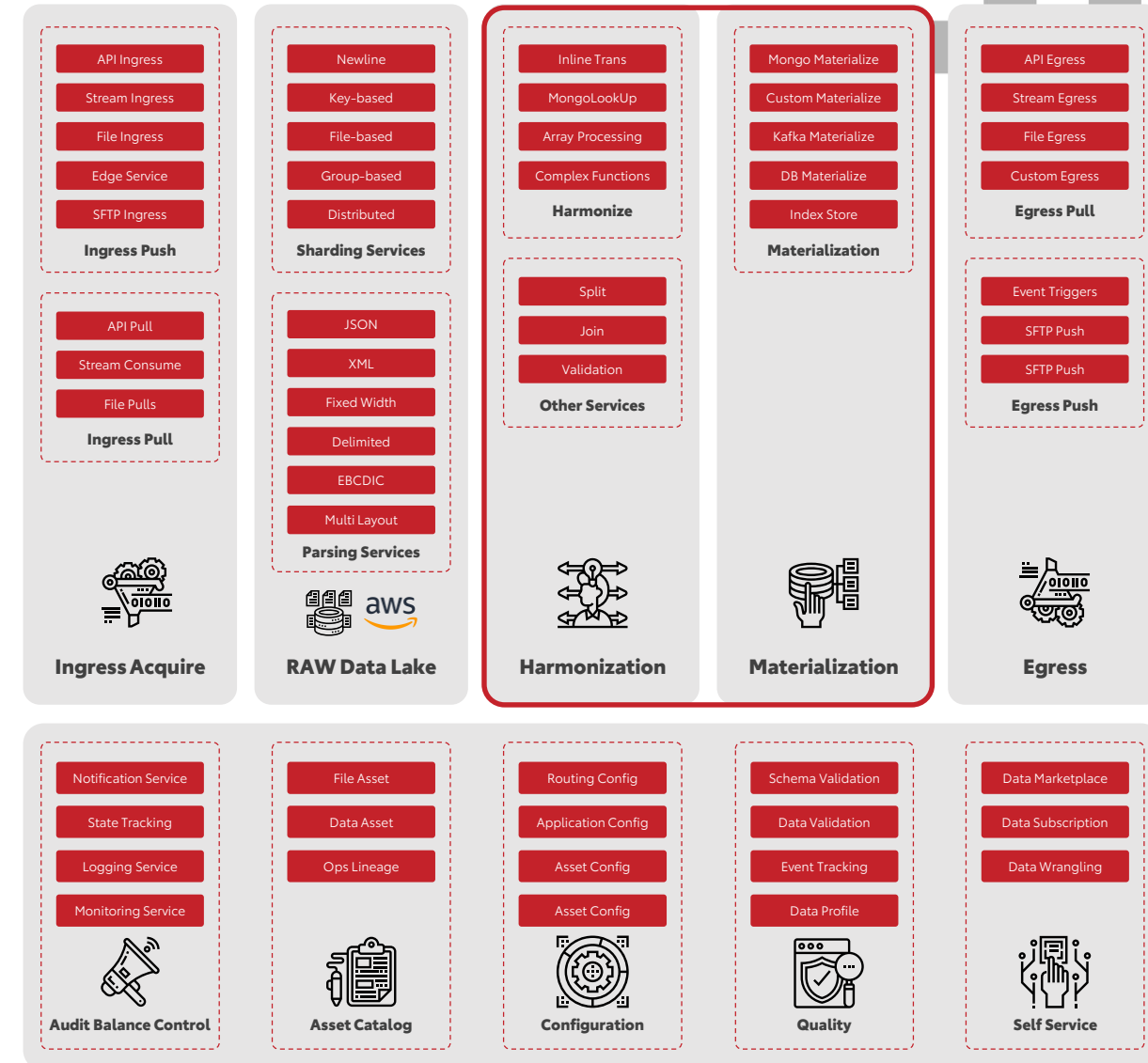
Factories need to provide the event and time-based specifics (transactional and operational) to achieve the expected data materialization for their consumption and use.

Acceptance Criteria

- ✓ Harmonized output stores fine-grained well-linked dataset in a standardized format with high cohesion for enterprise use.
- ✓ Structural data transformation meets and fulfills business needs.
- ✓ Ensure content level derivations, calculations, aggregation/summarization and splits are in alignment with business requirements.
- ✓ As functions are incrementally build and the systems may evolve, it utilizes a flexible data model to process new or changed data entities.
- ✓ Factory must demonstrate and adhere to the METER operational quality dimensions.

How

- ✓ Implement structural level (i.e.., structural data transformation) & content level (i.e., derivation, calculation, aggregation, summarization logic) semantic data integration & data quality functions from the shared enterprise perspective.
- ✓ Maintain versioning by utilizing flexible data model as data entities may arrive at different times in different sequences from varying data sources.
- ✓ Data enrichment and transformation; process the formatted source data with harmonization logic using a set of microservices and stores the outputs into the harmonized & materialized data stores.
- ✓ Ensure data quality is maintained and meets the end user expectations.
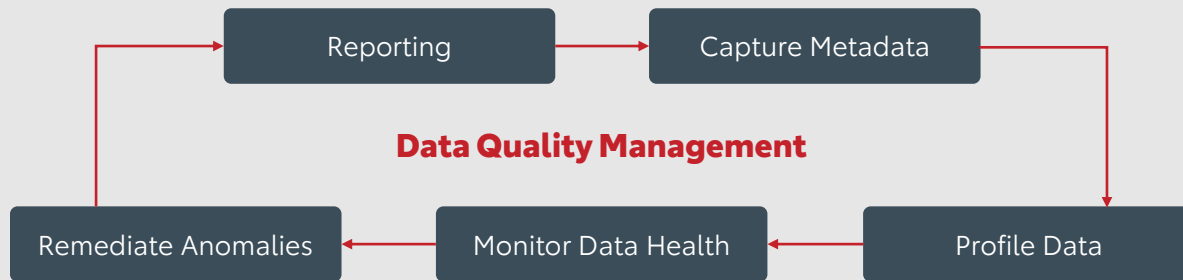

Ingress Acquire
- Ingress Push: API Ingress, Stream Ingress, File Ingress, Edge Service, SFTP Ingress
- Ingress Pull: API Pull, Stream Consume, File Pulls

RAW Data Lake
- Sharding Services: Newline, Key-based, File-based, Group-based, Distributed
- Parsing Services: JSON, XML, Fixed Width, Delimited, EBCDIC, Multi Layout

Harmonization
- Harmonize: Inline Trans, MongoLookUp, Array Processing, Complex Functions
- Other Services: Split, Join, Validation

Materialization
- Materialization: Mongo Materialize, Custom Materialize, Kafka Materialize, DB Materialize, Index Store

Egress
- Egress Pull: API Egress, Stream Egress, File Egress, Custom Egress
- Egress Push: Event Triggers, SFTP Push, SFTP Push

Audit Balance Control: Notification Service, State Tracking, Logging Service, Monitoring Service

Asset Catalog: File Asset, Data Asset, Ops Lineage

Configuration: Routing Config, Application Config, Asset Config, Asset Config

Quality: Schema Validation, Data Validation, Event Tracking, Data Profile

Self Service: Data Marketplace, Data Subscription, Data Wrangling

Data Quality Management

Data Quality Mgmt. Process Overview

Inputs:	Process/ Activities:	Outputs/ Deliverables:

Inputs:
- EDC Intake Form
- Define Data Quality Expectations
- Business Requirements
- Business Rules
- Data Requirements
- Business Metadata
- Technical Metadata
- Data Sources & Data Stores
- Data Lineage

Process/Activities:
- User Stories Created, Refined
- Perform Initial Data Quality Assessment
- Identify & Prioritize Improvements
- Develop & Deploy Data Quality Operations for Repeatability
- Review with TPO/BPO for Definition of Done
- Correct, Measure, Monitor & Report on Data Quality Levels & Findings

Outputs/Deliverables:
- Data Quality Framework
- Analyses from Data Profiling Based on Epic Definition
- Recommendations based on Root Cause Analysis of Issues
- DQM Procedures for Future Use
- Data Quality Reports
- Data Quality SLA
- DQ Policies & Guidelines

✓ Metadata Mgmt.   ✓ Data Quality Mgmt.   ✓ Data LifeCycle Mgmt.

Data Quality Improvement Lifecycle

Reporting → Capture Metadata → Profile Data → Monitor Data Health → Remediate Anomalies → Reporting

Data Quality Management

Applications provide real-time feedback which needs to be included in a day-to-day operations, in order to be effective. They also require access to quality data from any source, whether on-premise, in the cloud, or in any other new data platform.

Lack of front-office controls (e.g., poor quality of data entry at system of origin with no/limited validation) will impede in achieving the enterprise strategic goal of quality management.

Acceptance Criteria

- Ensure the data/information quality is maintained & matches to the user consumption
- All Factories shall plan, implement, and control activities that apply data quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers
- Ensure no entity instance will be recorded/identified more than once based upon how that instance is identified for uniqueness
- Ensure and validate the data is valid and it confirms to the syntax (format, type, range) of its definition
- Ensure the data is accurate, consistent and consumable

How

- **Factories must plan to identify the critical data, existing rules & patterns associated with their domains**
- **Factories, in collaboration with Data Quality Factory, will :**
  - conduct initial data quality assessment to identify, prioritize and perform root cause analysis of issues (baseline),
  - identify & prioritize actions based on business impact, develop preventive & corrective measures to improve data quality,
  - develop & deploy data quality operations to correct, measure, monitor & report data quality levels and findings

TOYOTA FINANCIAL SERVICES

Data Archival & Restoration

Data Archiving
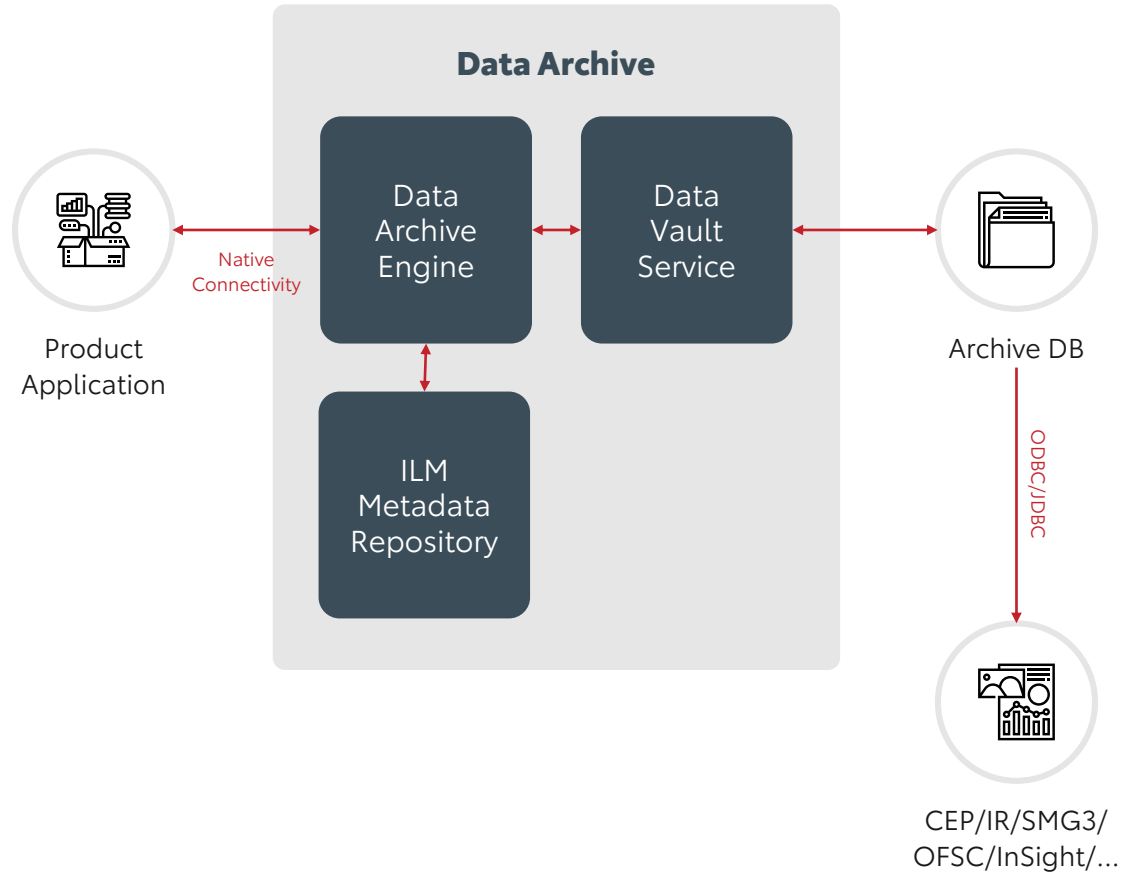
Data archiving is integral part of Data Supply Chain. Due to the expected growth in data over time we must have the archival and retrieval capability. The archived data must be performed in two key time spans (13 month and 13 years).

The archived data must be separated by the functional usage and will be stored according to the enterprise data retention guidelines. Special provision must be in place for email data archival and retrieval.

Acceptance Criteria

- All factories must ensure that they comply with the data archival/retention and retrieval policies
- All factories must ensure their data archiving/purging in a rolling fashion as per data retention policy
- Ensure for archiving and restore activities fulfills the business requirements
- Enterprise data archival/storage and restoration requires a well-defined process that is practical, reliable and trustworthy at all phases of data supply chain
- Where, how, and length of time must be defined and maintained
- Data ownership, access, storage, backup and backup schedule must be defined
- Store inactive data in a separate solution for long term retention
- Overall storage costs reduced, application performance is improved, and resources are not constrained

How

- **All factories must:**
  - Participate in the EDC intake process.
  - Engage with EDQ factory if they are out of compliance as per data storage and date retention policies to archive and purge data
  - Engage EDQ factory to build a process flow for automated archival, retrieval/restoration activities, if needed by the business

"PROTECTED" 関係者外秘