# MITS6005

# Big Data

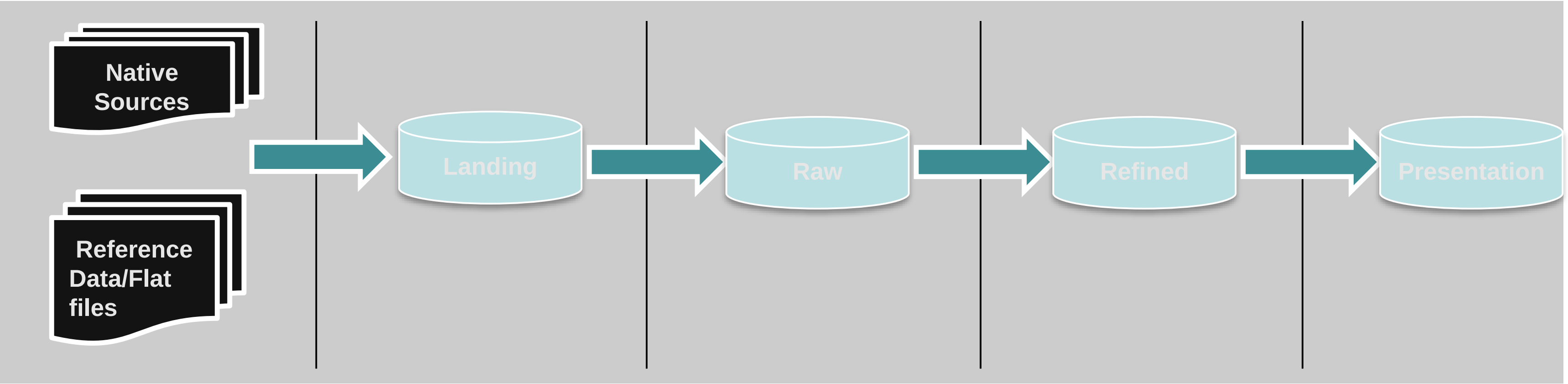
# Session 5b

# Data Ingestion

# Data Lake – What is it

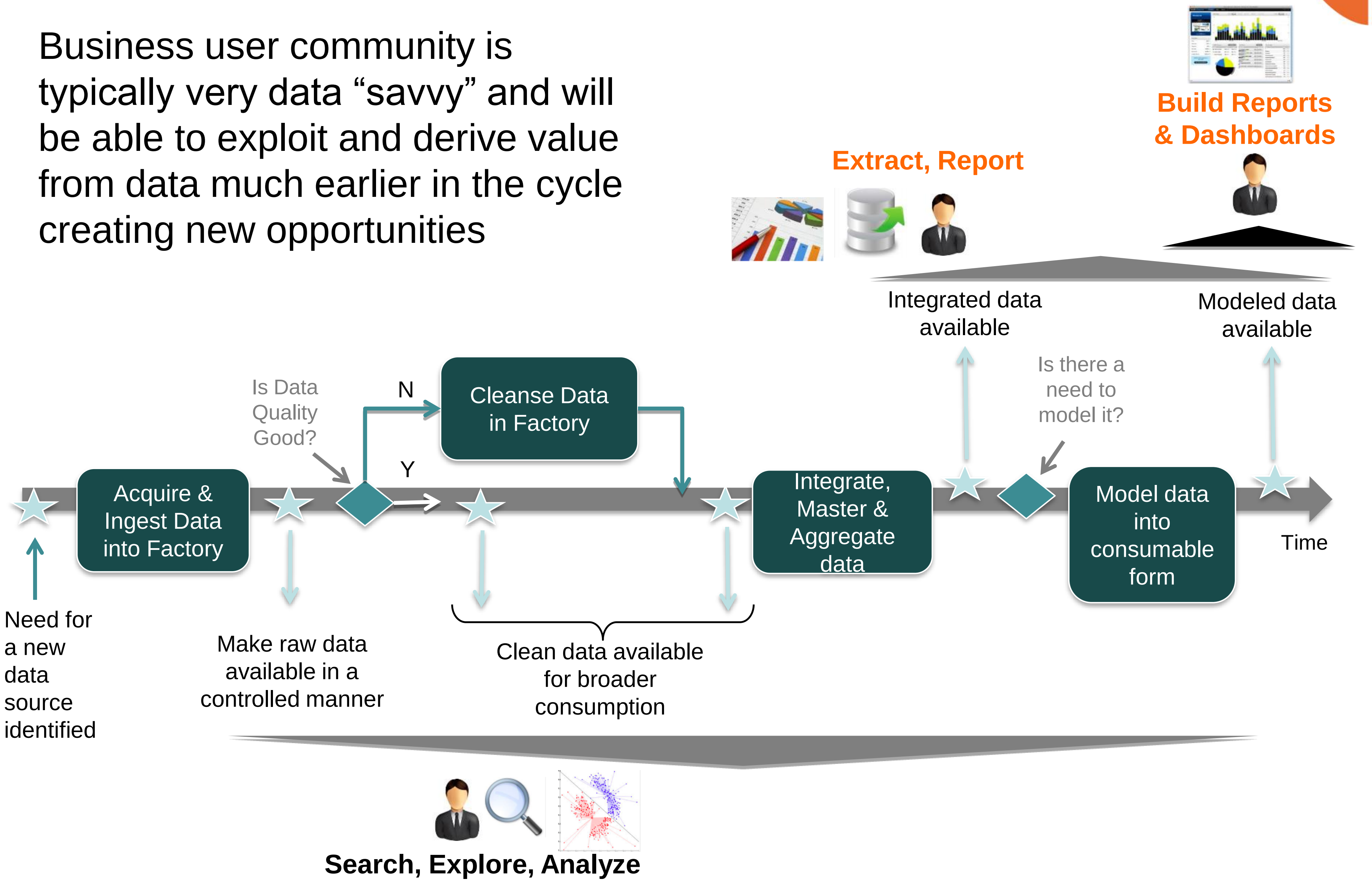
## Introduction

- Concept of Data lake took off with the advent of Big Data technologies and remains a fluid evolving concept at this time.
- Data Lake is an enterprise level repository of data on commodity hardware, running Big data applications like HADOOP.
- Data originates from multiple applications in the enterprise and is kept available “As Is” in pre-categorized and pre-manipulated state.
- Raw Data is then refined and made available based on the needs of an organization.
- Data lake Implementation projects originate because of the desire to integrate and store massive data sets at a single centralized location to enable cross functional analytics or to lay the basis for building the functional marts, Departmental sandboxes or enterprise warehouse.



# Value - Unlocking the Value of Data Earlier

Business user community is typically very data “savvy” and will be able to exploit and derive value from data much earlier in the cycle creating new opportunities


Build Reports
& Dashboards
Extract, Report
Integrated data available
Modeled data available
Is there a need to model it?
Is Data Quality Good?
N
Y
Cleanse Data in Factory
Acquire & Ingest Data into Factory
Integrate, Master & Aggregate data
Model data into consumable form
Time
Need for a new data source identified
Make raw data available in a controlled manner
Clean data available for broader consumption
Search, Explore, Analyze

# Data Lake Guiding Principles

- Keep original drivers and objectives “top of mind” and communicate them regularly:
  - “Enable easy integration of new data sources…”
  - “Minimize dependency on costly hardware…”
- Business engagement is essential to understand how data is created and used
- Adapt Incremental implementation approach to Succeed Fast or Fail Fast
- Constantly evaluate tendency to fall back into “old habits”:  Avoid “But that’s how we have always done it…”
- Just-enough data governance necessary to prevent data lakes turning into data swamps
- Select right tool for the job to provide better business value as fast as possible
- Collect metadata for the not only Ingestion automation, but for Data Catalog as a perquisite of Ingesting Data into the Data Lake
- Industry data models must be informed by “the art of the possible” as dictated by source system data structures and business use
- Establish and automate patterns for the ingestion of data into the Data Lake and out of the Data Lake
- Measure data quality upon Ingestion through implementing processes that immediately generate data profiles after ingestion. Create a data quality dashboard in a tool like Tableau to provide visibility into actual data quality

# The Data Lake Paradigm

|  | Data Lake | Data Warehouse |
|---|---|---|
| Data Breadth and Depth | • Store Everything As-is, With Complete History<br>• Structured, Semi-Structured and Unstructured | • The Data Warehouse With Aggregated Subsets<br>• Content Management Systems With Limited Metadata |
| Consumption Model | • Let Business Decide What They Need - On-Demand Views<br>• Support Rapid Change | • Pre-Defined Views, Curated By Experts.<br>• Long Change Cycles. |
| Business Driven | • Search Using Business Terminology<br>• Provide Data Lineage and History Tracking and Visualization | • Structured - Tables, Views, Reports. Limited Context<br>• Unstructured - Key-Word Search |
| Data Quality | • Data Quality Is Known And Tracked.<br>• Data Is Available In Various States from Raw to Fully Conformed and Standardized | • Data Available Only After Fully Conformed and Standardized<br>• Quality Metrics Often Not Available |
| Tools | • BYO Data Analysis Tools | • Fixed Set Of Business Intelligence Tools |

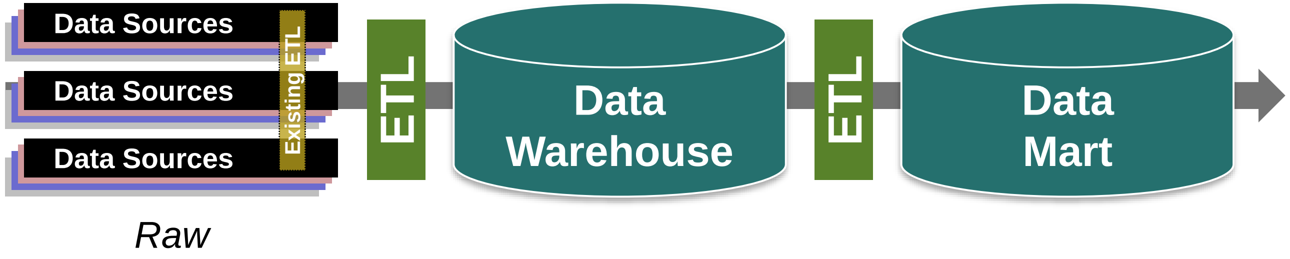
# Characteristics of Traditional DA vs. Big Data

| Characteristics | Traditional BI | Big Data |
|---|---|---|
| **Data volume** | Typically Terabytes | Tens to Hundreds of Terabytes, to Petabytes |
| **Velocity of change in scope** | Slower | Faster. Can adapt to frequent change of analytics needs |
| **Total Cost of Ownership** | TOC tends to be expensive | TOC tends to be lower due to lower cost storage and Open source tools |
| **Source data diversity, variety** | Lower | Higher |
| **Analysis driven** | Typically supports known analytics and required reporting | Inherently supports the data analysis and data discovery process by certain users |
| **Requirements driven** | Most of the time | Rarely |
| **Exploration & discovery** | Some of the time | Most of the time |
| **Structure of queries** | Robust | Un-structured |
| **Accuracy of Results** | Deterministic | Approximated |
| **Availability of results** | Slower (longer batch cycles) | Faster |
| **Stored data** | Schema is required to write data | No pre-defined schema is required |

# How things are different

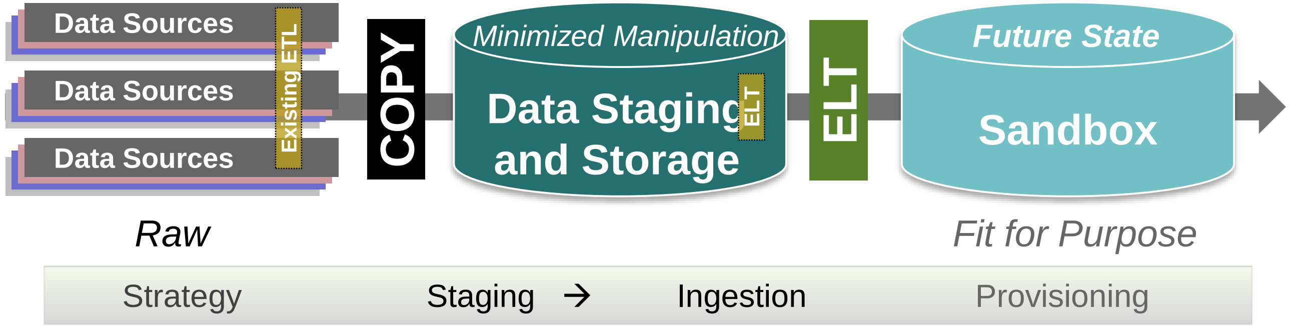
## Data Acquisition Methodologies

### Traditional Data Management



*Data is managed as part of a project, analytic, or other workflow*

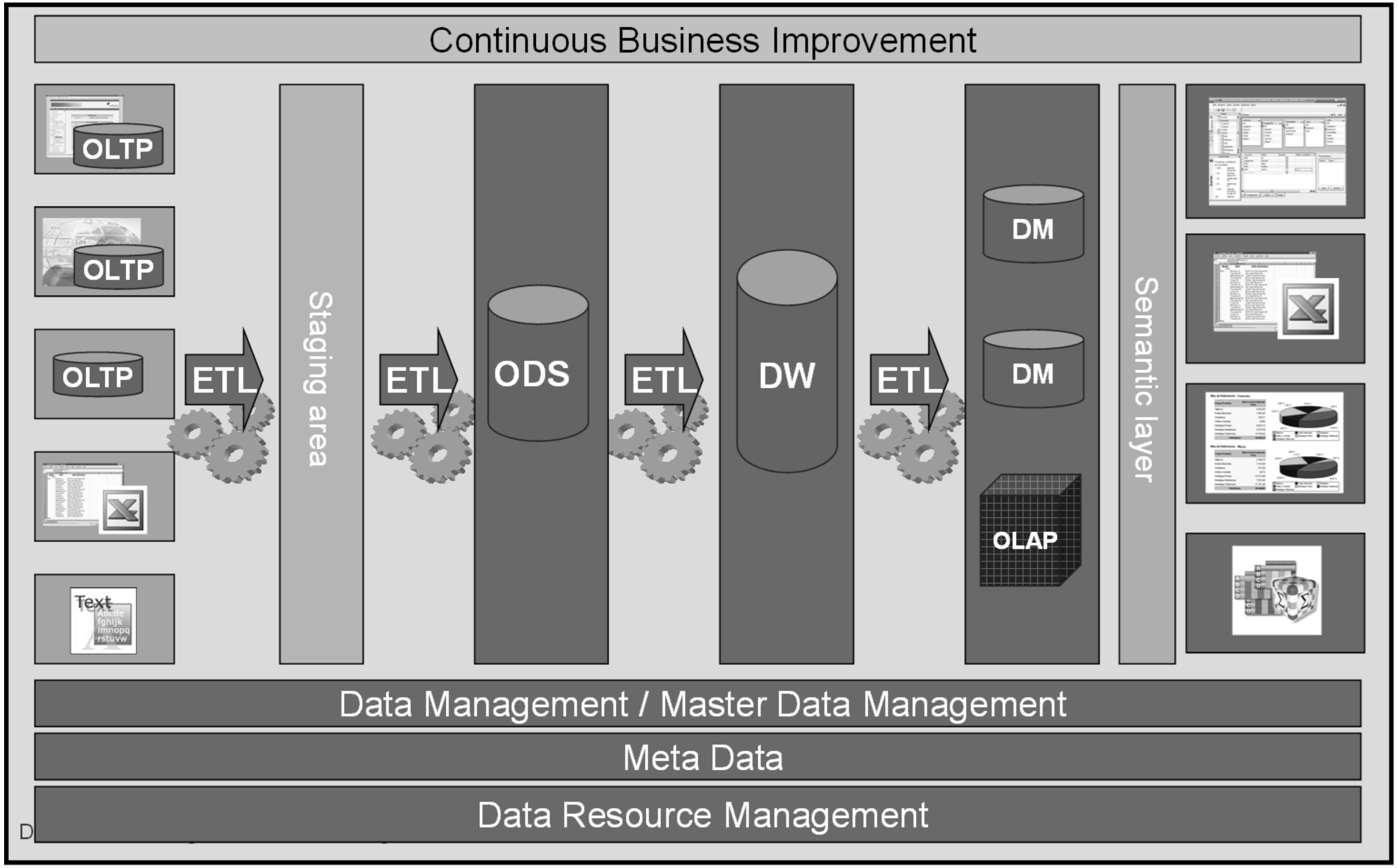
### "Big Data" Data Management



- Data sources are ingested raw (potentially enriched for identifier  resolution, searchabilty, quality, etc.)
- Data Services and Access Control move data between storage data stores and consolidate data for analytical data stores

***A principle shared by firms with successful Big Data capabilities is providing as much raw data  as possible in an easy to consume, trusted manner***

# Traditional Data Integration



# Traditional Data Integration

**Schema-on-Write (RDBMS):**

- Prescriptive Data Modeling:

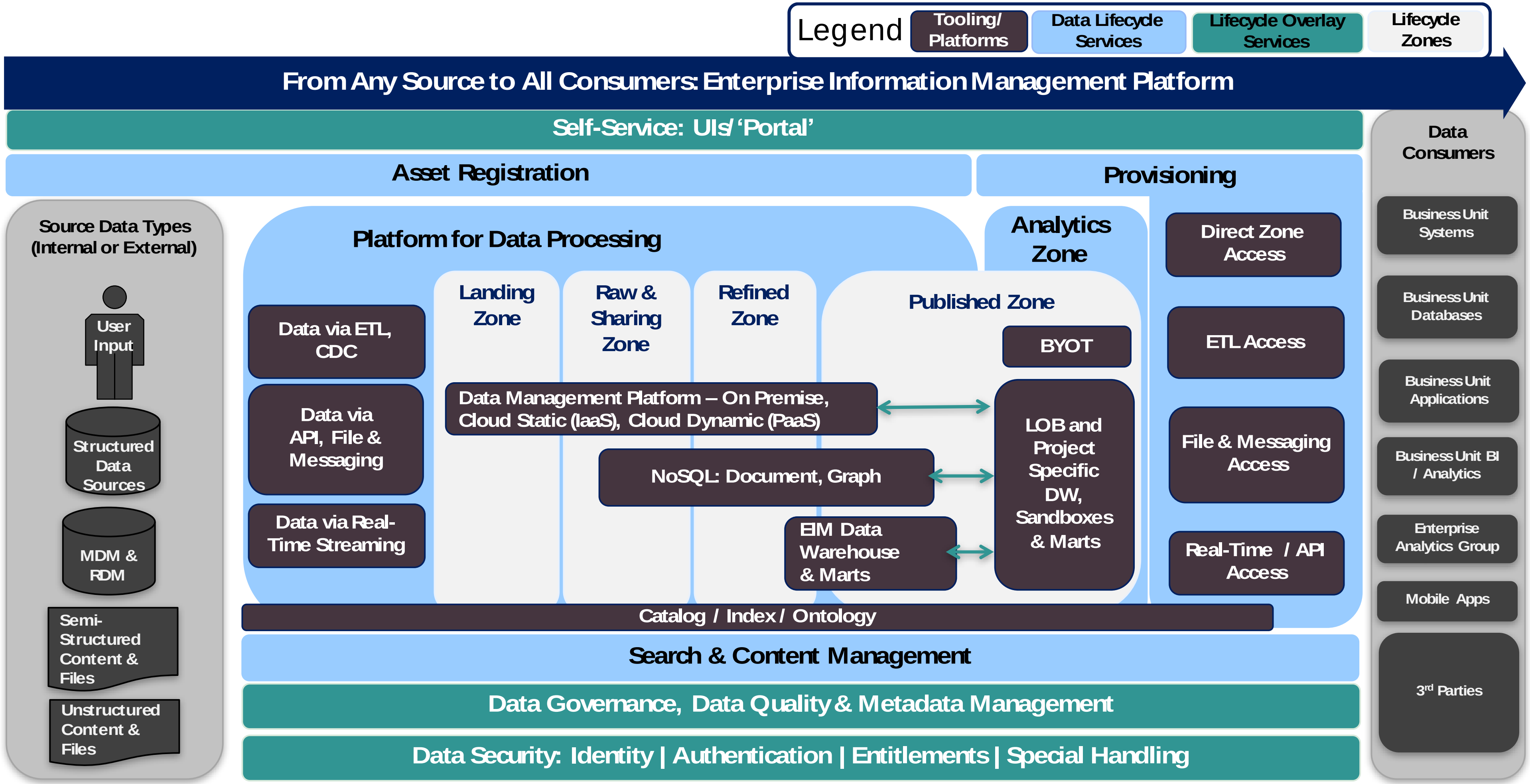
  - Create static DB schema

  - Transform data into RDBMS

  - Query data in RDBMS format

- New columns must be added explicitly before new data can propagate into the system.
- Tend to be quite expensive and slow to change
- Limited in terms of the scalability and processing the data as rapidly as the business wants
- **Good for Known Unknowns (Repetition)**

# Modern Day Data Lake Architecture



Legend: Tooling/ Platforms | Data Lifecycle Services | Lifecycle Overlay Services | Lifecycle Zones

From Any Source to All Consumers: Enterprise Information Management Platform

Self-Service: UIs/ 'Portal'

Asset Registration

Provisioning

Source Data Types (Internal or External)

User Input

Structured Data Sources

MDM & RDM

Semi-Structured Content & Files

Unstructured Content & Files

Platform for Data Processing

Data via ETL, CDC

Data via API, File & Messaging

Data via Real-Time Streaming

Landing Zone

Raw & Sharing Zone

Refined Zone

Analytics Zone

Published Zone

BYOT

Data Management Platform – On Premise, Cloud Static (IaaS), Cloud Dynamic (PaaS)

NoSQL: Document, Graph

EIM Data Warehouse & Marts

LOB and Project Specific DW, Sandboxes & Marts

Direct Zone Access

ETL Access

File & Messaging Access

Real-Time / API Access

Catalog / Index / Ontology

Search & Content Management

Data Governance, Data Quality & Metadata Management

Data Security: Identity | Authentication | Entitlements | Special Handling

Data Consumers

Business Unit Systems

Business Unit Databases

Business Unit Applications

Business Unit BI / Analytics

Enterprise Analytics Group

Mobile Apps

3rd Parties

# Modern Day Data Lake Architecture

**Schema-on-Read (Hadoop):**

- Descriptive Data Modeling:

  - Copy data in its native format

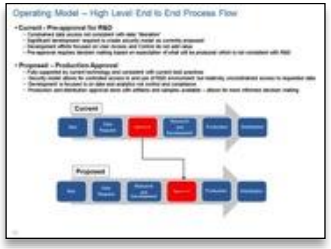
  - Create schema + parser

  - Query Data in its native format (does ETL on the fly)

- New data can start flowing any time and will appear retroactively once the schema/parser properly describes it.
- Flexibility and Scalability
- Rapid Data Ingestion
- **Good for Unknown Unknowns (Exploration)**

# Data Lake Integration - Strategy & Planning

Different views or perspectives on the Data Management Architecture will facilitate understanding of recommendations and implications

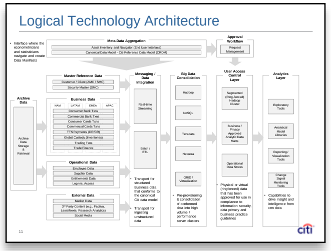


Operating Model – High Level End to End Process Flow

- Current - Preapproval for R&D
- Proposed – Production Approval

## Business Architecture

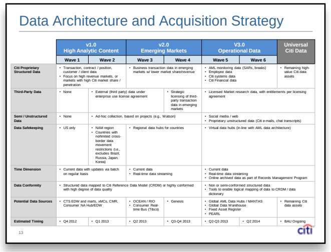
- Business Capability View: Business and management processes, their strategic objectives and required data analytics capabilities
- Operating Model / Functional View: Description of key policies, procedures and governance models (committees, review and approval points) required to achieve business objectives
- Organizational View: Description of teams, staffing requirements and reporting relationships in order to support the operational model



Logical Technology Architecture

## Technology Architecture

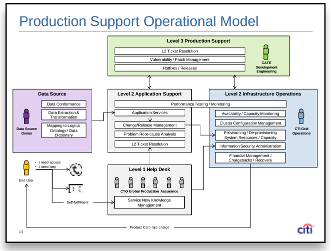
- Logical and physical blueprints for enabling technology capabilities (i.e., Hadoop and/or EDW repositories, Information Asset Inventory & Navigator)
- Vendor strategy and technology product selections



Data Architecture and Acquisition Strategy

## Data Architecture and Acquisition Strategy

- Identification of structured and unstructured data to expose for analytics, and their data sources
- Strategy / approach for conforming to data dictionary / ontology
- Prioritization and schedule for pre-provisioning data into production environment



Production Support Operational Model

Level 3 Production Support
Data Source
Level 2 Application Support
Level 2 Infrastructure Operations
Level 1 Help Desk
citi

## Physical Infrastructure Support  Architecture

- Organizational and process model for how to support end users in a production environment
- Includes model for help desk and ticket resolution, environment monitoring, provisioning and access control management, and financial chargeback/recovery
- Organizational model and Production Service Definition

# Key Capabilities

**EDH Platform:** Scalable and Agile Standards-based data management and analytic processing capability

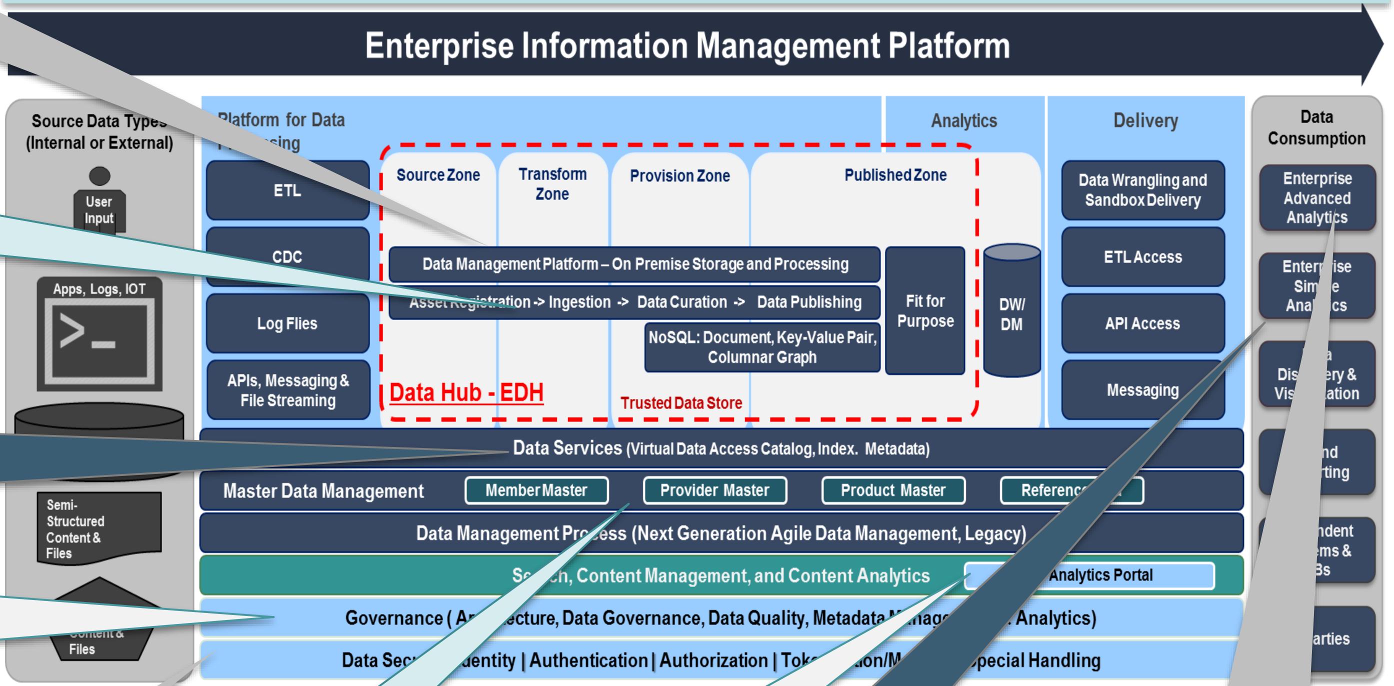
**Rapid Data Provisioning:** Automated, Metadata configure, data ingestion process sources new data in days

**Data Curation and Quality:** Data quality is measured, tracked, and improved.

**Data Catalog:** Business-focused data dictionary with Google-like search capabilities

**Scalable Data Governance:** Data governance enabled by the architecture.

**Integrated Data Security:** Leverage Tokenization and Encryption to secure critical data



Enterprise Information Management Platform
Source Data Types (Internal or External)
User Input
Apps, Logs, IOT
Semi-Structured Content & Files
Content & Files
Platform for Data
ssing
ETL
CDC
Log Flies
APIs, Messaging & File Streaming
Analytics
Delivery
Source Zone
Transform Zone
Provision Zone
Published Zone
Data Management Platform – On Premise Storage and Processing
Asset Registration -> Ingestion -> Data Curation -> Data Publishing
NoSQL: Document, Key-Value Pair, Columnar Graph
Fit for Purpose
DW/ DM
Data Hub - EDH
Trusted Data Store
Data Wrangling and Sandbox Delivery
ETL Access
API Access
Messaging
Data Consumption
Enterprise Advanced Analytics
Enterp ise Sim e Ana ics
Dis ery & Vis ation
nd rting
ndent ms & Bs
arties
Data Services (Virtual Data Access Catalog, Index. Metadata)
Master Data Management
Member Master
Provider Master
Product Master
Referenc
Data Management Pr ess (Next Generation Agile Data Management, Legacy)
Sc ch, Content Management, and Content Analytics
Analytics Portal
Governance ( Ar ecture, Data Governance, Data Quality, Metadata M nag Analytics)
Data Sec dentity | Authentication | Authorization | Tok ion/M pecial Handling

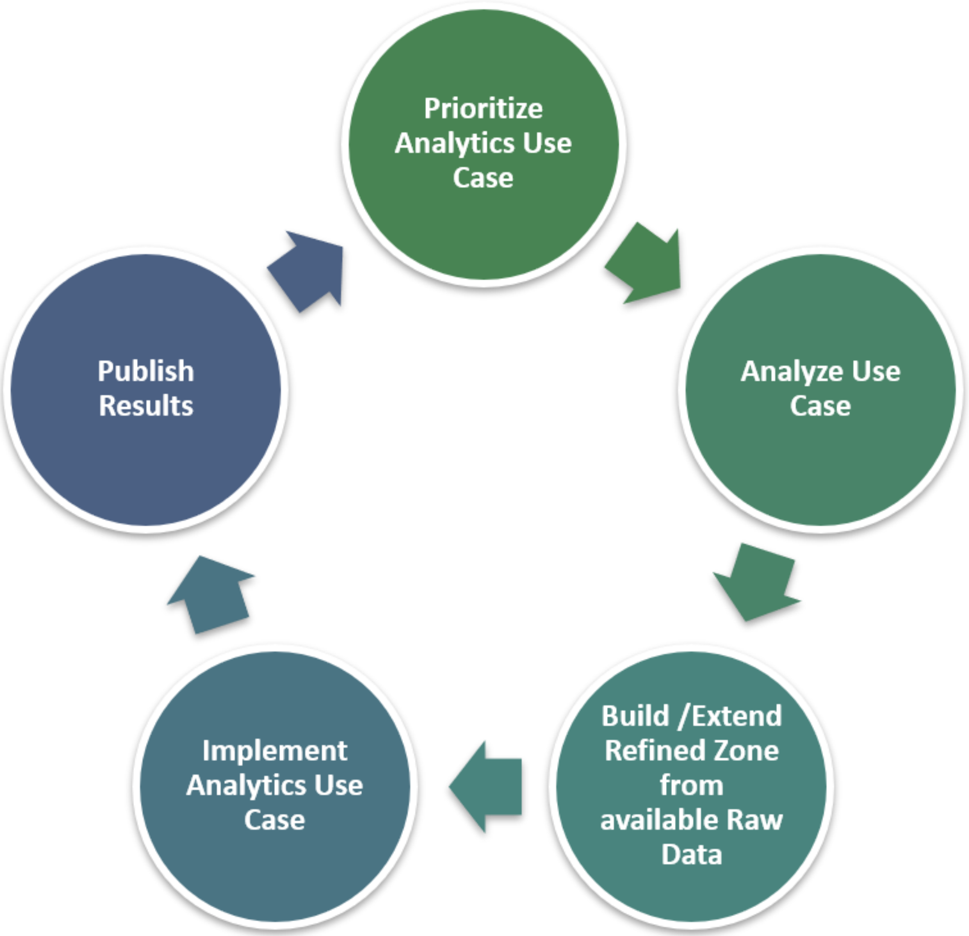
**Holistic View of Member, Provider:** Single view of high-value data.

**Collaborative Knowledge Sharing:** BI/Analytics Portal provides shared access to critical analytical content and best practice information

**Business Authored Reporting and Dashboards:** Data Discovery and Visualization tools enable better understanding of data, thus enabling better corporate performance

**Business Enabled Analytics:** Data Scientists are able to prepare data sets, perform advanced analytics, and publish their results

# Integration - Analytics Driven Approach


Prioritize Analytics Use Case
Analyze Use Case
Build /Extend Refined Zone from available Raw Data
Implement Analytics Use Case
Publish Results

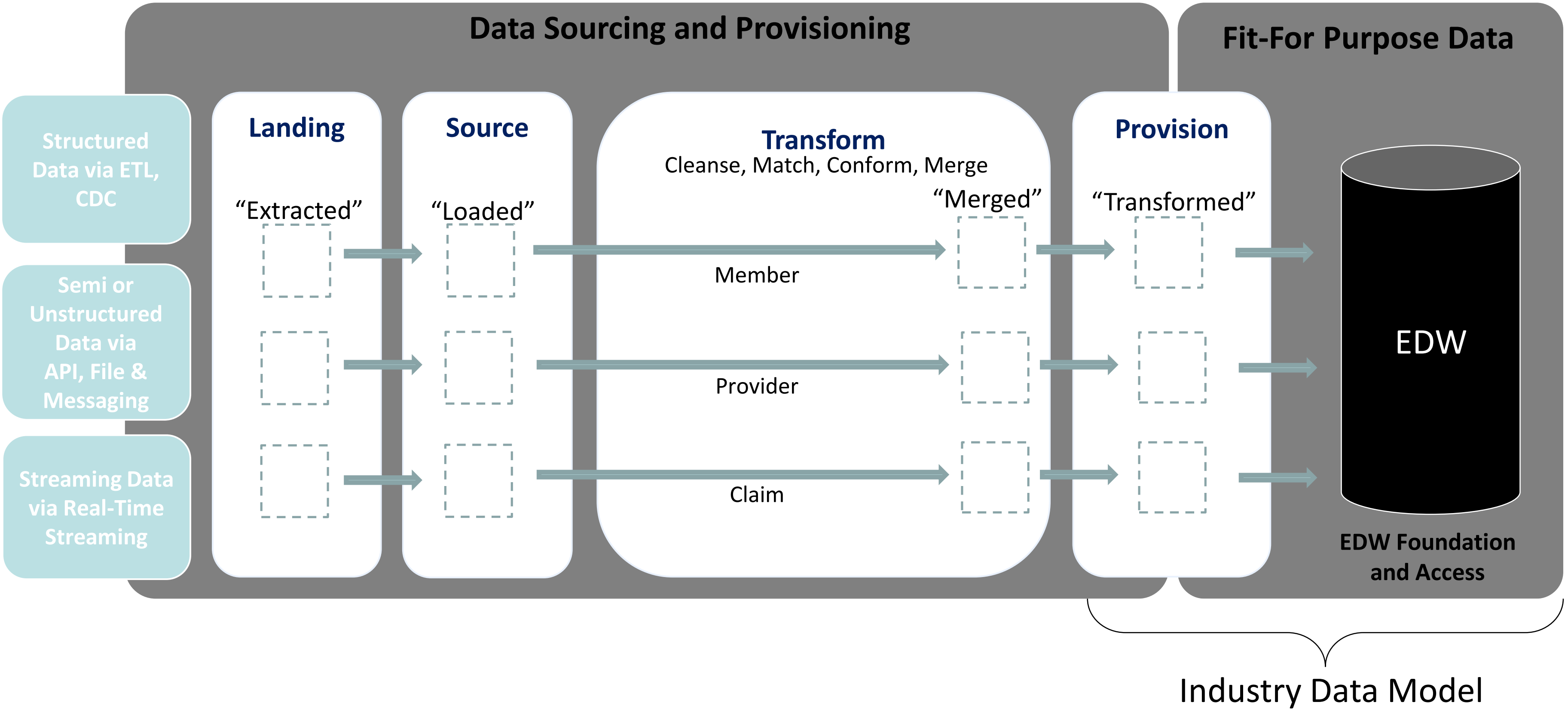
## Key Facts

- Business initiated programs
- Business outcome and value drives priorities
- Analytics-Ready dataset, the priority
- Earlier challenges include cross-domain and cross systems integration
  - Claims matching with Membership
  - Merging Membership from multiple source systems

# Integration - Domain Driven Approach

### Key Facts

- IT initiated programs
- IT value drives priorities
- Building an Enterprise level data model, a first priority
- Earlier challenges include cross systems integration
  - Merging Membership from multiple source systems



Data Sourcing and Provisioning
Fit-For Purpose Data
Structured Data via ETL, CDC
Semi or Unstructured Data via API, File & Messaging
Streaming Data via Real-Time Streaming
Landing
"Extracted"
Source
"Loaded"
Transform
Cleanse, Match, Conform, Merge
"Merged"
Member
Provider
Claim
Provision
"Transformed"
EDW
EDW Foundation and Access
Industry Data Model

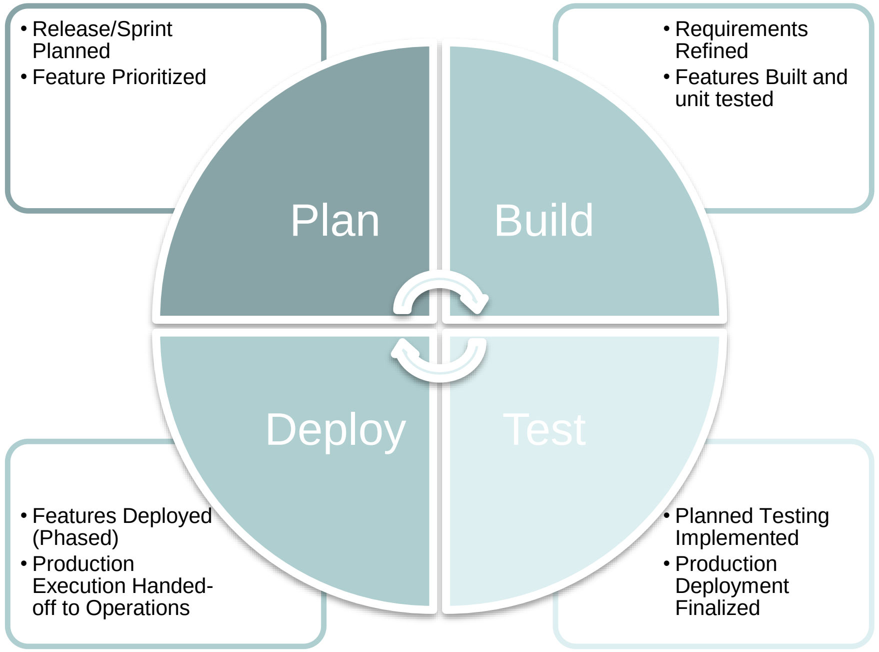
# Agile – Continuous Delivery Approach

**Initiate**

- Requirements Analysis
- Epics/Features Defined Groomed Backlog
- Architecture Defined
- Platform Stood Up
- Testing Approach Defined

**Release Definition**

- 8 Weeks Release
- 5 Two-Week Sprints
- 4 Build Sprints
- 1 HIP Sprint
- HIP Sprint Reserved for
  - Technical Debt
  - Utility Building
  - Sprint Planning



- Release/Sprint Planned
- Feature Prioritized

- Requirements Refined
- Features Built and unit tested

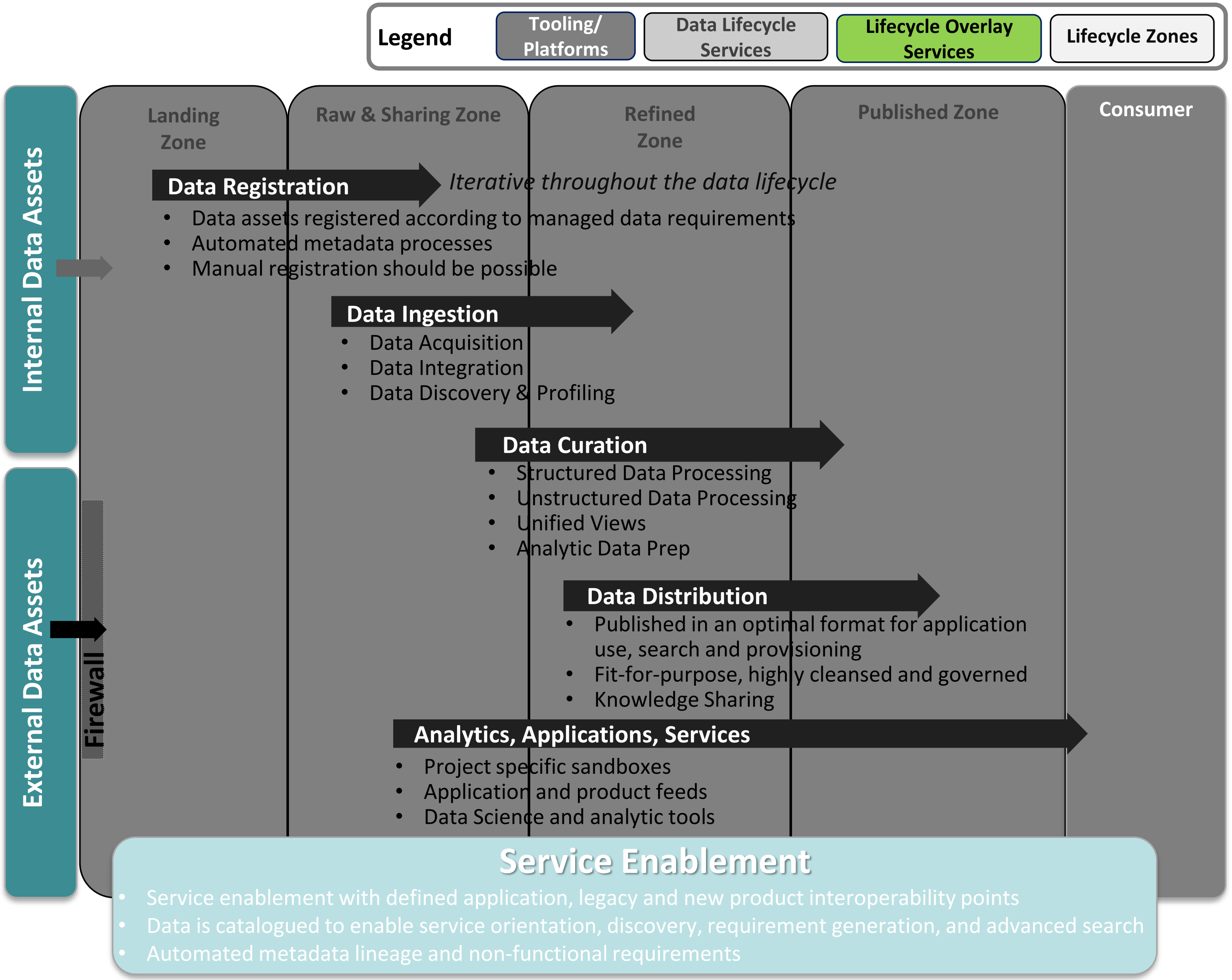
- Features Deployed (Phased)
- Production Execution Handed-off to Operations

- Planned Testing Implemented
- Production Deployment Finalized

**Typical Release**

Succeed Fast or Fail Fast

# Data Movement & Zones



**Data Registration**

*Iterative throughout the data lifecycle*

- Data assets registered according to managed data requirements
- Automated metadata processes
- Manual registration should be possible

**Data Ingestion**

- Data Acquisition
- Data Integration
- Data Discovery & Profiling

**Data Curation**

- Structured Data Processing
- Unstructured Data Processing
- Unified Views
- Analytic Data Prep

**Data Distribution**

- Published in an optimal format for application use, search and provisioning
- Fit-for-purpose, highly cleansed and governed
- Knowledge Sharing

**Analytics, Applications, Services**

- Project specific sandboxes
- Application and product feeds
- Data Science and analytic tools

**Service Enablement**

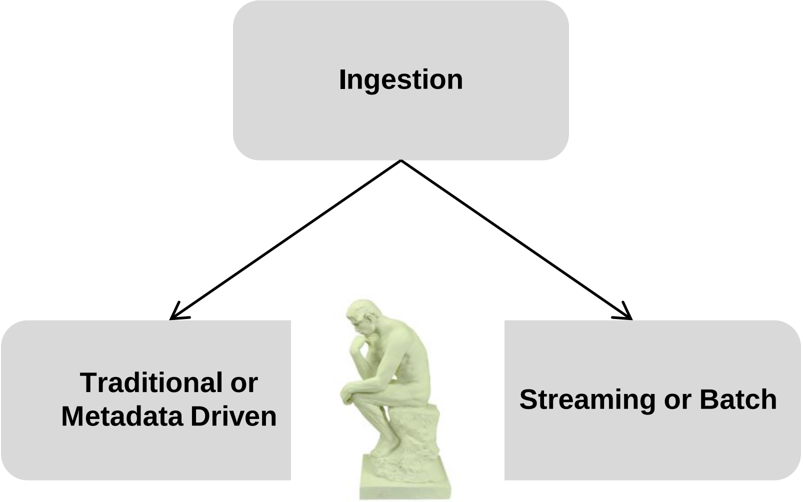
- Service enablement with defined application, legacy and new product interoperability points
- Data is catalogued to enable service orientation, discovery, requirement generation, and advanced search
- Automated metadata lineage and non-functional requirements

# Data Ingestion

- Process of Moving Data to the Data Lake

- Once Ingested, Data is Available
  - For Processing and Distribution
  - Discovery
  - Analytics

- Key Decisions
  - Streaming vs. Batch
  - Traditional vs. Metadata Driven



Ingestion

Traditional or Metadata Driven

Streaming or Batch

# Batch vs. Streaming

## *Batch Processing*

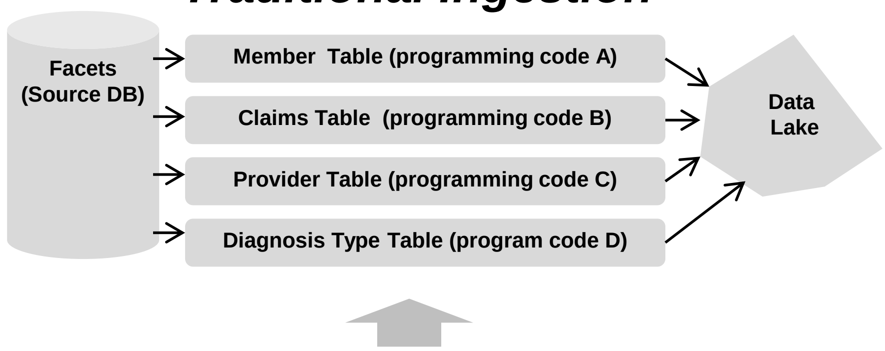
- Similar to Traditional data integration processing
- Good for groups or snapshots of data
- Time lag in data availability

## *Stream Processing*

- Fast-data
- Real-time needs
  - Fraud detection
  - Order processing
- Minimal lag in data availability
- Can be more scalable

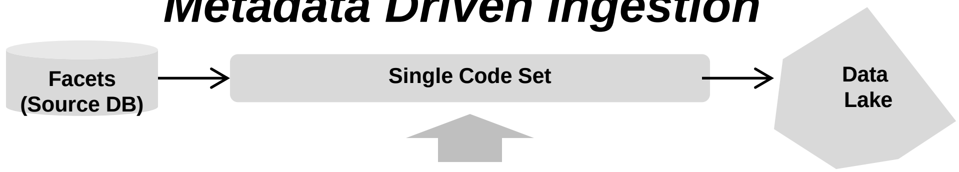
# Traditional vs. Metadata Driven Ingestion

## *Traditional Ingestion*


Facets
(Source DB)
Member  Table (programming code A)
Claims Table  (programming code B)
Provider Table (programming code C)
Diagnosis Type Table (program code D)
Data
Lake

- One set of code for each table ingested
- Good for custom coding or small number of ETL Jobs

## *Metadata Driven Ingestion*



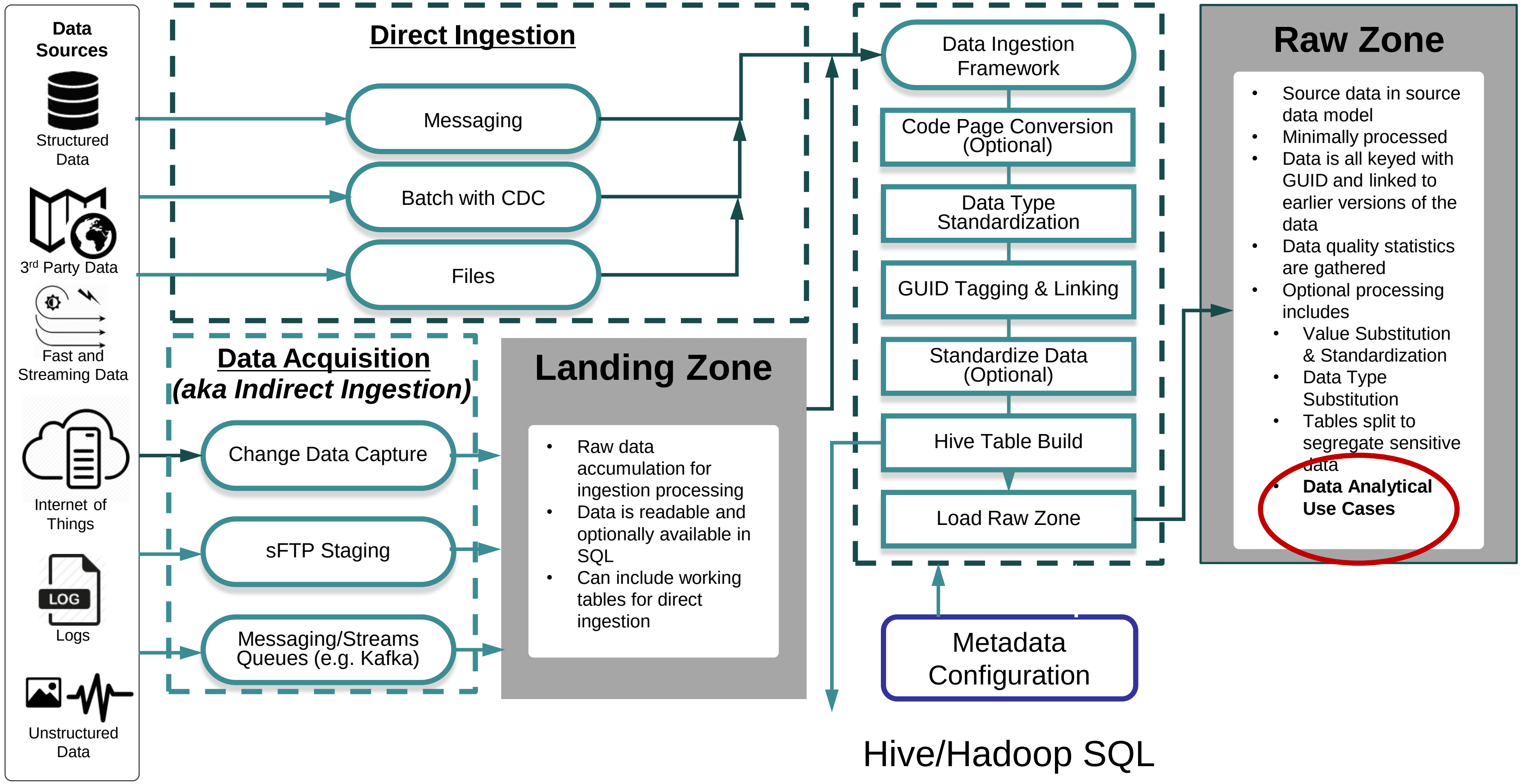
Metadata

| Sequence | Source | Table | Schedule |
|---|---|---|---|
| 1 | Facets | Member | Daily |
| 2 | Facets | Claims | Daily |
| 3 | Facets | Provider | Daily |
| 4 | Facets | Diagnosis Type | Monthly |

### *Benefits of Metadata Driven Ingestion*

- Extremely scalable (in development time) over 100s and 1000s of tables
- Increased consistency and supportability
- Increased quality of data
- Quick time to value: 5-10x faster  than custom coding or point-to-point usage of ETL tools

# Data Acquisition and Ingestion Detail



**Data Sources**

- Structured Data
- 3rd Party Data
- Fast and Streaming Data
- Internet of Things
- Logs
- Unstructured Data

**Direct Ingestion**

- Messaging
- Batch with CDC
- Files

**Data Acquisition** *(aka Indirect Ingestion)*

- Change Data Capture
- sFTP Staging
- Messaging/Streams Queues (e.g. Kafka)

**Landing Zone**

- Raw data accumulation for ingestion processing
- Data is readable and optionally available in SQL
- Can include working tables for direct ingestion

Data Ingestion Framework

- Code Page Conversion (Optional)
- Data Type Standardization
- GUID Tagging & Linking
- Standardize Data (Optional)
- Hive Table Build
- Load Raw Zone

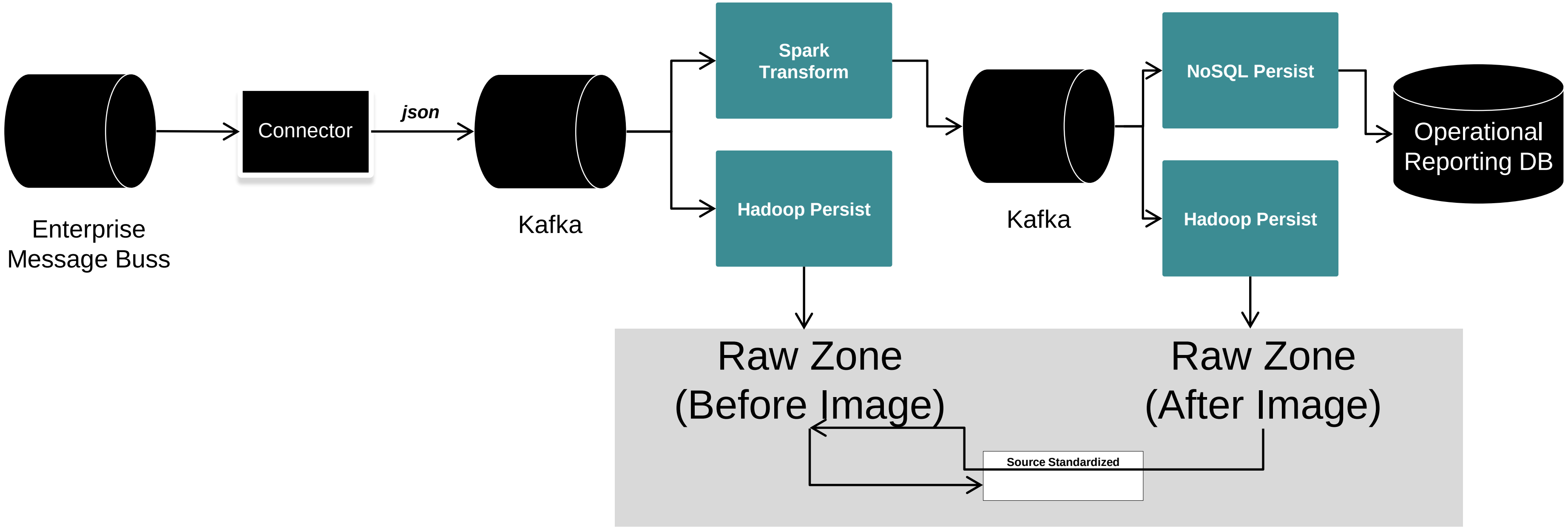
Metadata Configuration

Hive/Hadoop SQL

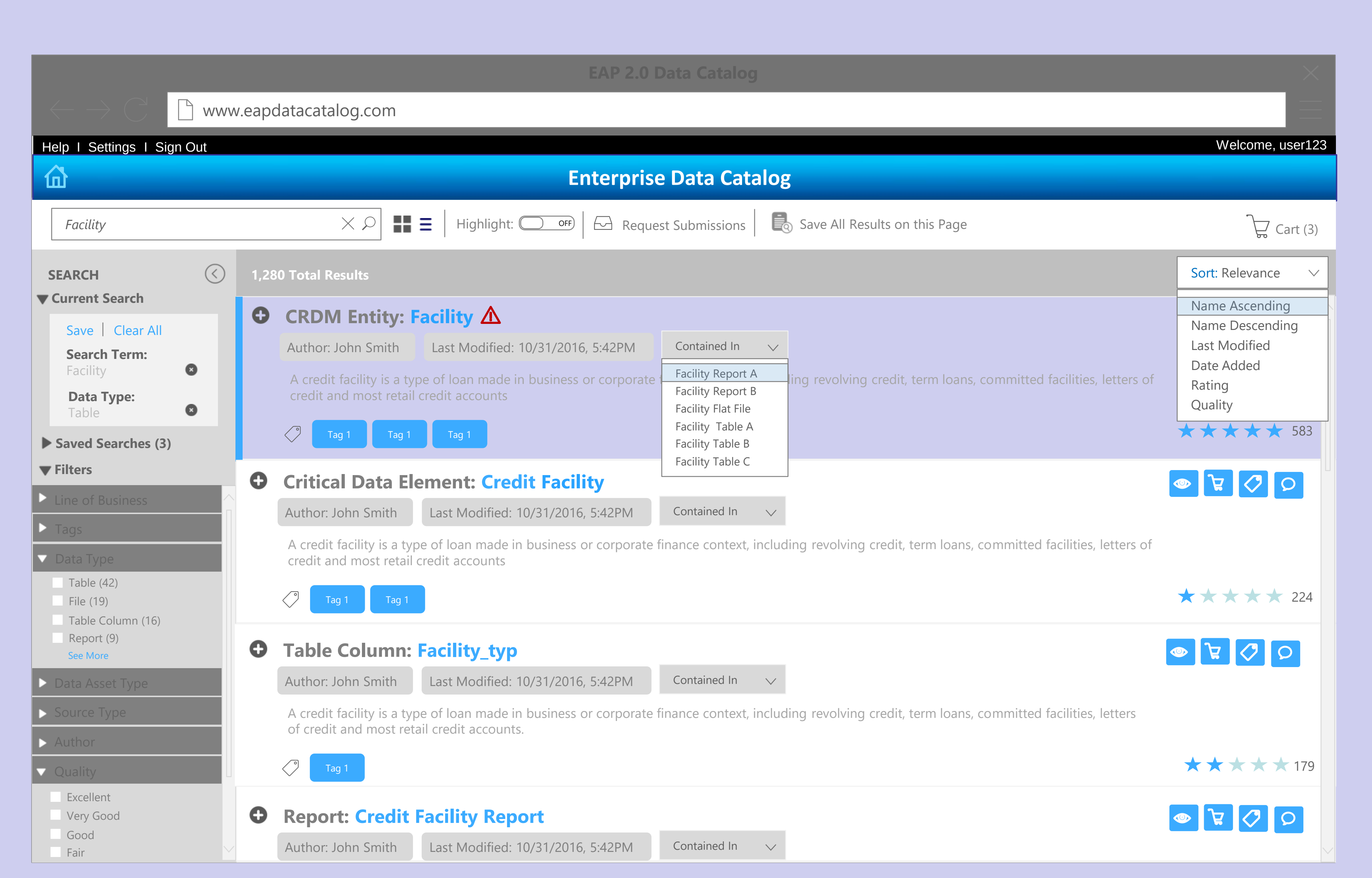
**Raw Zone**

- Source data in source data model
- Minimally processed
- Data is all keyed with GUID and linked to earlier versions of the data
- Data quality statistics are gathered
- Optional processing includes
  - Value Substitution & Standardization
  - Data Type Substitution
  - Tables split to segregate sensitive data
  - **Data Analytical Use Cases**

# High-Performance Streaming Ingestion



# Catalog - Search & Explore Data



EAP 2.0 Data Catalog

www.eapdatacatalog.com

Help I Settings I Sign Out

Welcome, user123

Enterprise Data Catalog

Facility

Highlight: OFF | Request Submissions | Save All Results on this Page | Cart (3)

SEARCH

▼ Current Search

Save | Clear All

Search Term: Facility

Data Type: Table

▶ Saved Searches (3)

▼ Filters

- ▶ Line of Business
- ▶ Tags
- ▼ Data Type
  - □ Table (42)
  - □ File (19)
  - □ Table Column (16)
  - □ Report (9)
  - See More
- ▶ Data Asset Type
- ▶ Source Type
- ▶ Author
- ▼ Quality
  - □ Excellent
  - □ Very Good
  - □ Good
  - □ Fair

1,280 Total Results

Sort: Relevance

- Name Ascending
- Name Descending
- Last Modified
- Date Added
- Rating
- Quality

**CRDM Entity: Facility** ⚠

Author: John Smith | Last Modified: 10/31/2016, 5:42PM | Contained In

- Facility Report A
- Facility Report B
- Facility Flat File
- Facility Table A
- Facility Table B
- Facility Table C

A credit facility is a type of loan made in business or corporate ... ing revolving credit, term loans, committed facilities, letters of credit and most retail credit accounts

Tag 1 | Tag 1 | Tag 1

★★★★★ 583

**Critical Data Element: Credit Facility**

Author: John Smith | Last Modified: 10/31/2016, 5:42PM | Contained In

A credit facility is a type of loan made in business or corporate finance context, including revolving credit, term loans, committed facilities, letters of credit and most retail credit accounts

Tag 1 | Tag 1

★☆☆☆☆ 224

**Table Column: Facility_typ**

Author: John Smith | Last Modified: 10/31/2016, 5:42PM | Contained In

A credit facility is a type of loan made in business or corporate finance context, including revolving credit, term loans, committed facilities, letters of credit and most retail credit accounts.

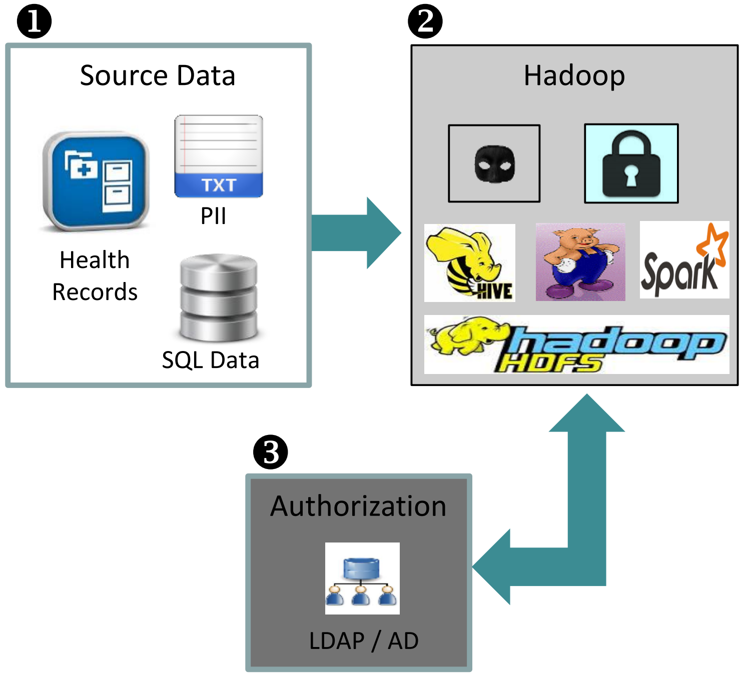
Tag 1

★★☆☆☆ 179

**Report: Credit Facility Report**

Author: John Smith | Last Modified: 10/31/2016, 5:42PM | Contained In

# Masking & Encryption

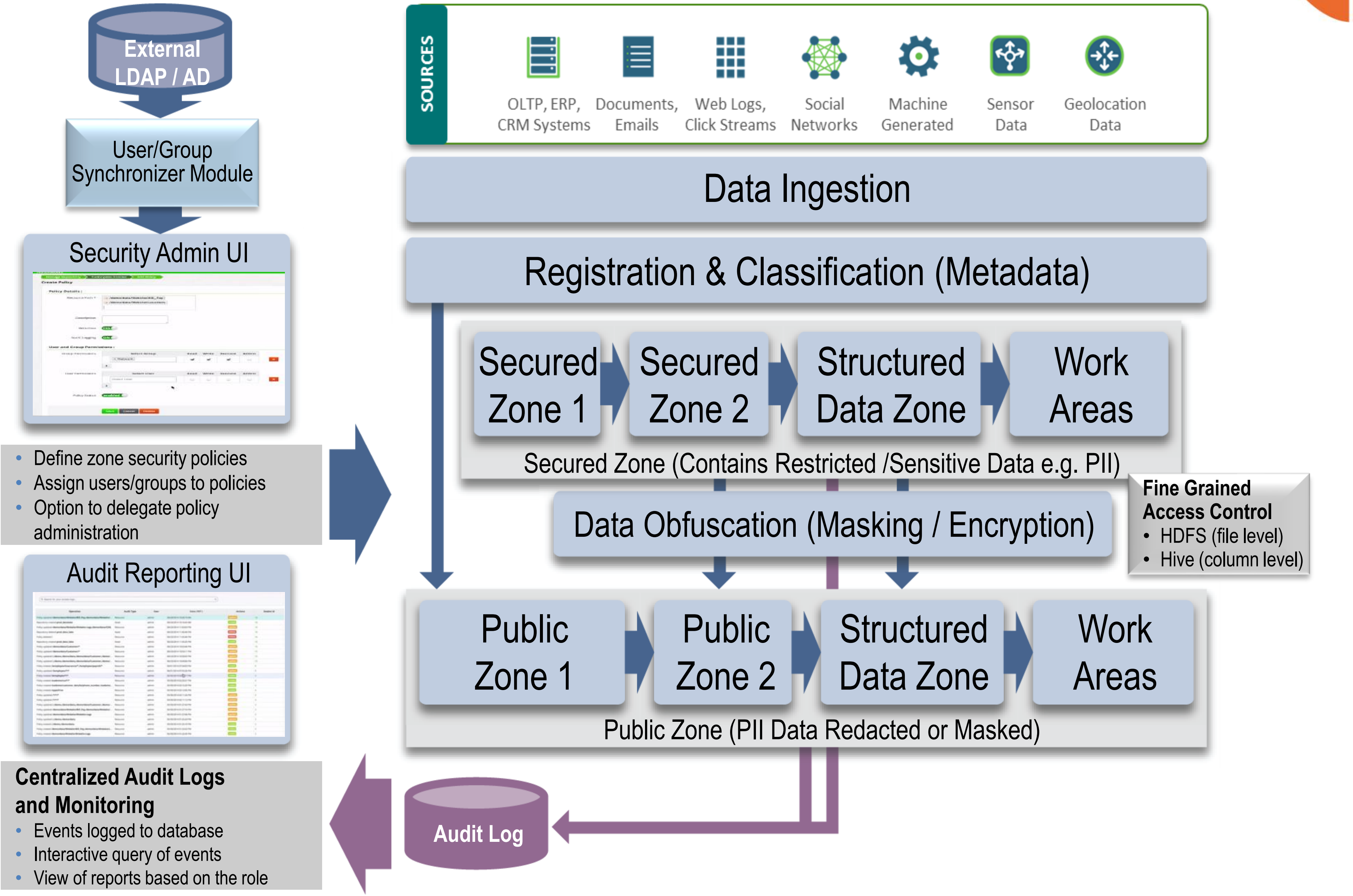


❶ Source data in database, data warehouse and various other structures

❷ Healthcare firm requires both masking (one-way) and encryption/ decryption for various sensitive data elements

❸ Leveraging Active Directory and LDAP, organization controls which users can see what degree of sensitive data– all 100% transparent and automatic to users

# Security & Audit



- Define zone security policies
- Assign users/groups to policies
- Option to delegate policy administration

**Centralized Audit Logs and Monitoring**

- Events logged to database
- Interactive query of events
- View of reports based on the role

**Fine Grained Access Control**

- HDFS (file level)
- Hive (column level)

# Users & Governance

**The Zones within the architecture service different user groups within the analytical community**

Analytical Users

Data

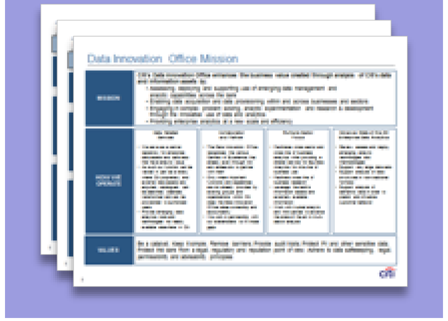
| Zones | Landing | Raw | Refined | Publish |
|---|---|---|---|---|
| **Users within Zones:**<br>• **Actor**<br>• **Tools** | Users:<br>• IT | Users:<br>• Data Scientists<br>• IT<br>• Data stewards<br>Tools: R, Python, SAS | Users:<br>• Data Scientists<br>• Business Power Users<br>Tools: R, Python, SAS | Users:<br>• Business Users<br>• Operational Systems<br>Tools: Qlik/Tableau, Reporting, Service Bus |
| **Data (within zones)**<br>• **Structure** | • Data remains in "as is" form without change<br>• Transient storage; data is removed after ingestion | • As identical as possible to the source data<br>• Data profiled & quality assessed<br>• Metadata augmented with data-specific information derived from discovery, profiling and quality checks<br>• Metadata enriched with business rules & context | • Data cleansed, aggregated, conformed, curated, remediated, or otherwise manipulated according to defined processes and rules<br>• Enriched data driven by business outcomes or analytic needs | • Data Certified for:<br>• Meet specific, defined and managed enterprise data management requirements<br>• High Quality and trusted<br>• Fit For Purpose/Operational Use<br>• Contextually relevant and accurate<br>• Governed by Business specific usage |
| **Data Management (between zones)**<br>• **Governance**<br>• **Security** | Security: Targeted user base, data access rules | Governance:<br>• Data Catalog<br>• Data is Profiled<br>• Quality is Measured<br>Security: Targeted user base, data access rules | Governance:<br>• Lineage Captured<br>• Enterprise Standards<br>• Data Quality Rules<br>• Enterprise Models<br>Security: Targeted user base, data access rules | Governance:<br>• Lineage Captured<br>• Modeled for a Specific Business Use<br>Security: Broad user base, access aligned to limited datasets |

**Governance and security is applied during data movement across zones and within the zones**

# Operating Models

An essential component of enterprise data strategy will be a detailed approach for supporting and operating the future state

## MISSION STATEMENT


Data Innovation Office Mission

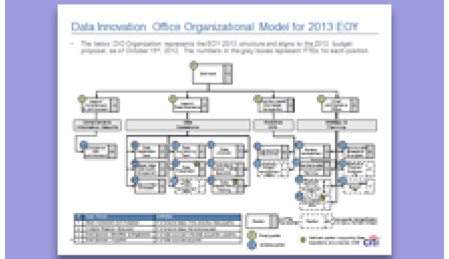
Mission statements and guiding principles for organizational change

## SERVICE MODEL

Service Overview

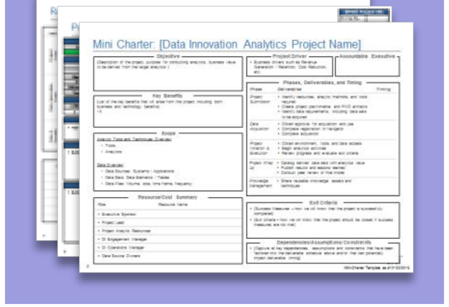
Identification and definition of all services to be offered

## ORGANIZATIONAL MODEL


Data Innovation Office Organizational Model for 2013 EOY

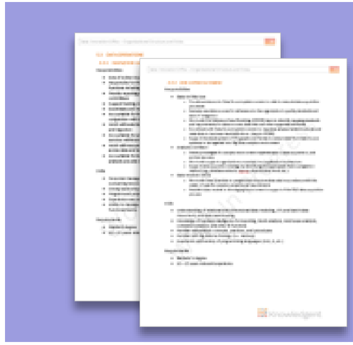
Organizational patterns for Management & Governance, Data, Analytics and Technology

## KEY ARTIFACT TEMPLATES


Mini Charter: [Data Innovation Analytics Project Name]

Standard templates to define new projects and track progress

## ROLES & RESPONSIBILITIES



Roles & Responsibilities for industry standard job descriptions and skill requirements

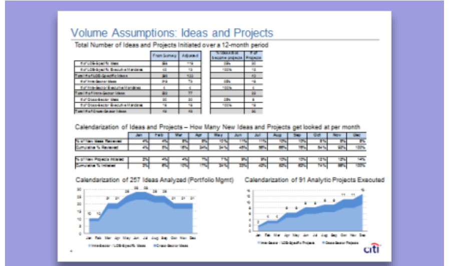
## PLAYBOOK

Data Innovations Data Analytics Program – Workflows & Day 1 Procedures

Volume 1 – Management & Governance

Playbooks that describe workflows for the services and procedures for production support activities

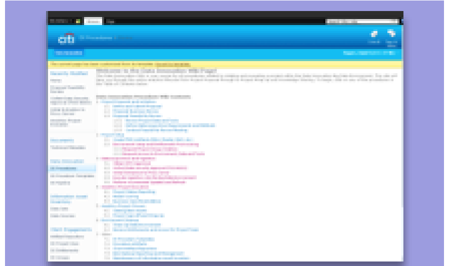
## BUDGET MODEL


Volume Assumptions: Ideas and Projects

Total Number of Ideas and Projects Initiated over a 12-month period

Templates for estimating and amortizing build and support costs
for hardware, software, etc. services/resources

## PROCEDURES WIKI



Online intranet/extranet communication and collaboration resources to support platform operations, BAU and break-fix operations

# Operating Framework on Data Lake Architecture

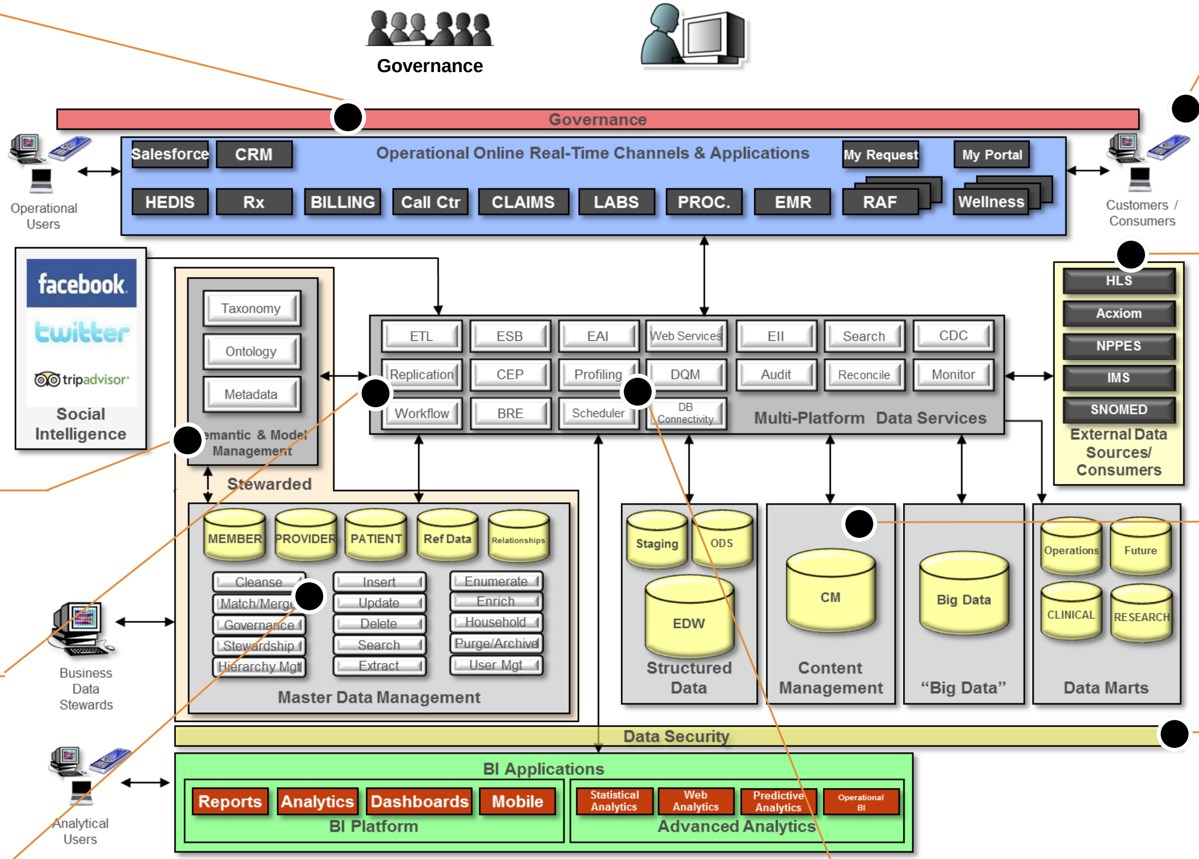
**Governance:**
Establish rules, policies and standards to protect, exploit and maximize the value of information in the organization

**Metadata:**
Define a business metadata strategy that is key in harmonizing information across disparate data sources and for consistent use of information by business users.

**Enterprise Models:**
Data need analysis, authoritative sources, standard data structures

**Data Integration:**
Evaluate data integration needs and make decisions around consistent use of EII, EAI,ETL

**Master Data Management:**
Provide a gold copy of reference data to the enterprise.



**Tools and Technologies:**
Standardizing tools and technologies based on best of breed tools.

**Data Quality:**
Implement on-going processes to measure and improve the timeliness, accuracy and consistency of the data.

**Presentation:**
Strategy to allow users to access information in a user-friendly manner.

**Data Access:**
Provide standard ways of sharing data with applications, business intelligence tools and downstream applications.

**Enterprise Content Mgmt:**
Provide a platform for delivery, storage and search for structured and un-structured data

**Security:**
Provide access to data based on roles using common technologies for access management and security management across different layers.

The End