ENTERPRISE DATA WORLD

OMNI ATLANTA HOTEL
ATLANTA, GA
APRIL 2-7, 2017

The Components of Data Lake Architecture

April Reeve

Sunday, April 2, 2017

2:30 PM to 5:45 PM

PRODUCED BY:

DATAVERSITY

DAMA International

www.dataversity.net www.dama.org

#EDW17

April Reeve

Data Stuff

- Thirty years doing data oriented stuff
- **Data Management disciplines – Data Integration, Data Governance, Data Modeling, Data Quality, Business Intelligence, Master Data Management, Data Conversion, Data Warehousing , Enterprise Content Management, Big Data Management**
- Currently **Director of Enterprise Information Strategy and Architecture** at **Celgene Corporation** in Summit, NJ
- Certifications –
  - **Certified Data Management Professional (DAMA CDMP)**
  - **Certified Data Governance and Stewardship Professional (ICCP DGSP)**
  - **Certified Business Intelligence Professional (TDWI CBIP)**
  - Certified in Enterprise Governance of IT (ISACA CEGIT)
  - Certified Information Systems Auditor (ISACA CISA)
- Masters degree in Financial Management (predictive modeling, risk management, derivatives, corporate finance)
- **Book “Managing Data in Motion – Data Integration Best Practice Techniques and Technologies”**
- **New chapters - Data Management Body of Knowledge (DMBoK) release 2 – Data Integration and Big Data**

Agenda

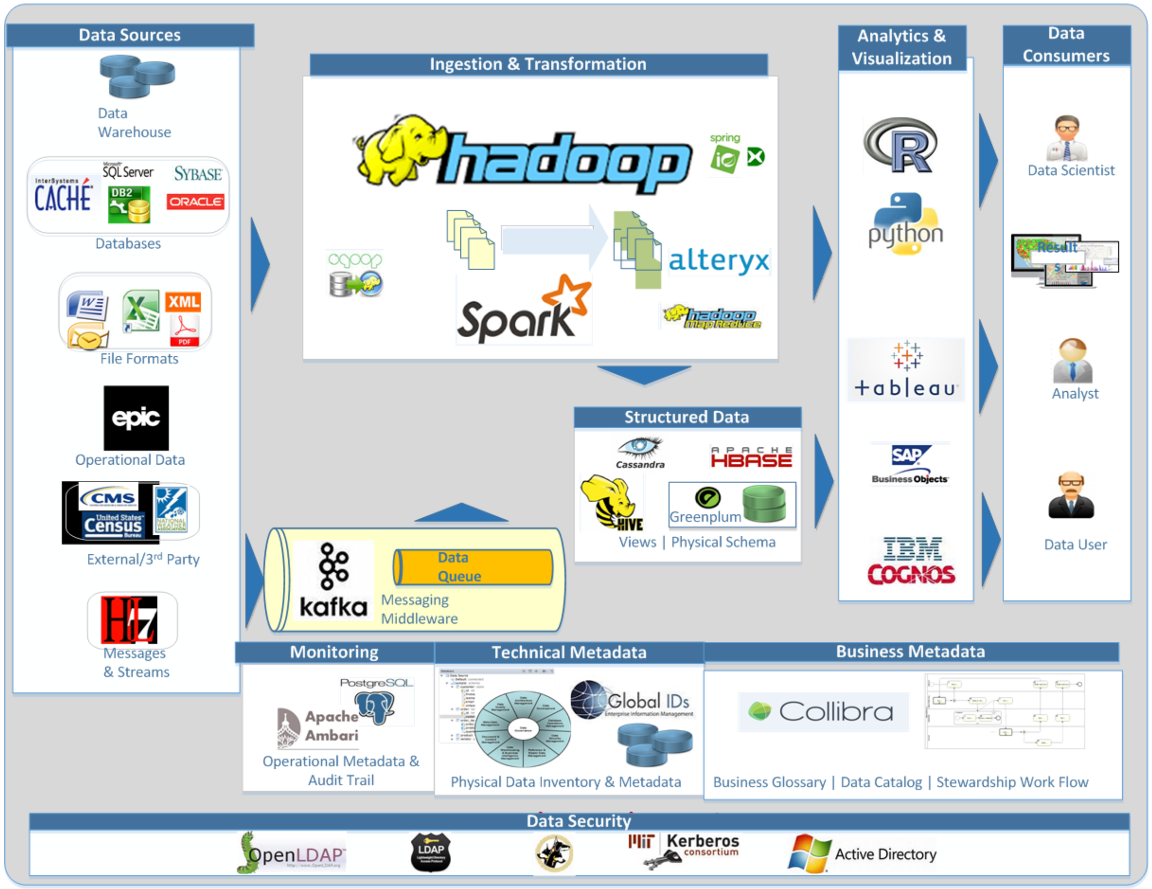
- The attributes of a Data Lake
- How Data lakes are different from Data Warehouses
- Architecture of a Data Scientist Sandbox
- Components Needed for Hadoop Data Governance
- Architecture of a Big Data Analytics Lake
- Architecture of a Real-Time Streaming Operation
- Recommendations in Implementing Data Lakes

The Data Lake

Key Attributes of a Data Lake

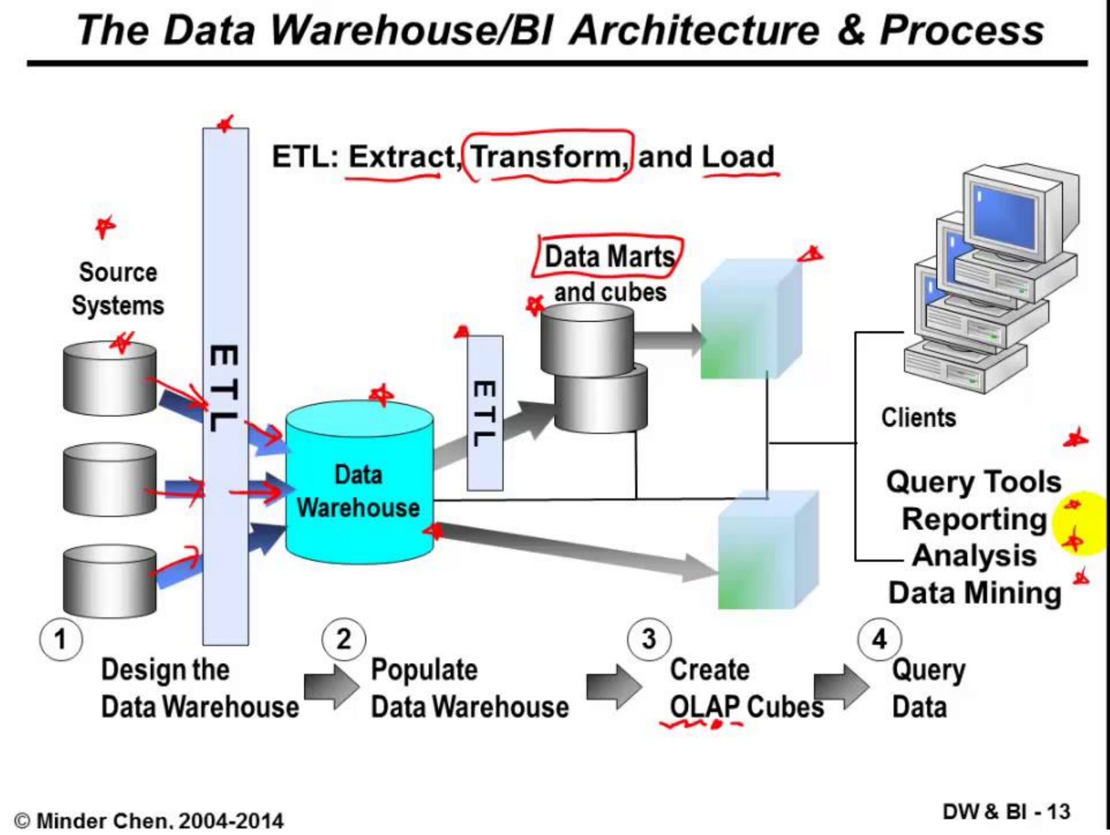
- **A single shared repository of data, typically stored within Distributed File System (DFS).** Hadoop data lakes preserve data in its original form and capture changes to data and contextual semantics throughout the data lifecycle. This approach is especially useful for compliance and internal auditing activities.
- **Includes orchestration and job scheduling capabilities (e.g., via YARN).** Workload execution is a prerequisite for enterprise Hadoop and YARN provides resource management and a central platform to deliver consistent operations, security and data governance tools across Hadoop clusters, ensuring analytic workflows have access to the data and the computing power they require.
- **Contains a set of applications or workflows to consume, process or act upon the data.** Data is preserved in its original form. Whether structured, unstructured or semi-structured, data is loaded and stored as-is.
    - From Datafloq – “What is a Data Lake and What are the Benefits?”

Example Big Data Architecture



How Data Lakes are different from Data Warehouses

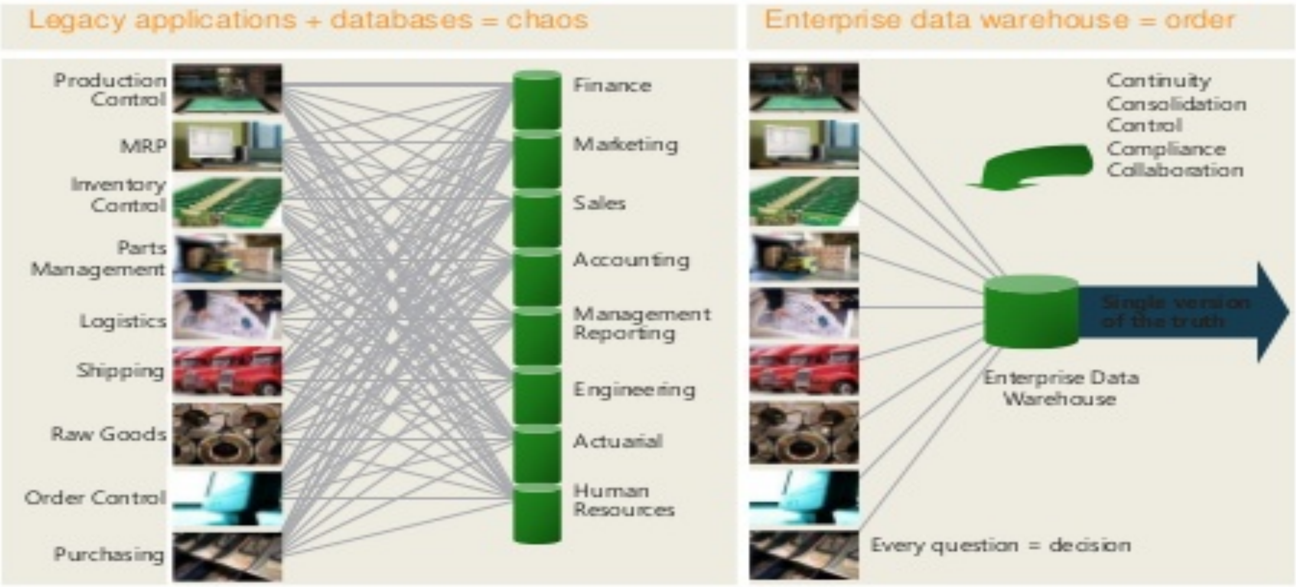
Traditional Data Warehouse Architecture



The Data Warehouse/BI Architecture & Process

1 Design the Data Warehouse → 2 Populate Data Warehouse → 3 Create OLAP Cubes → 4 Query Data

Why use a Data Warehouse?



Two purposes of data warehouse: 1) save time building reports; 2) slice in dice in ways you could not do before

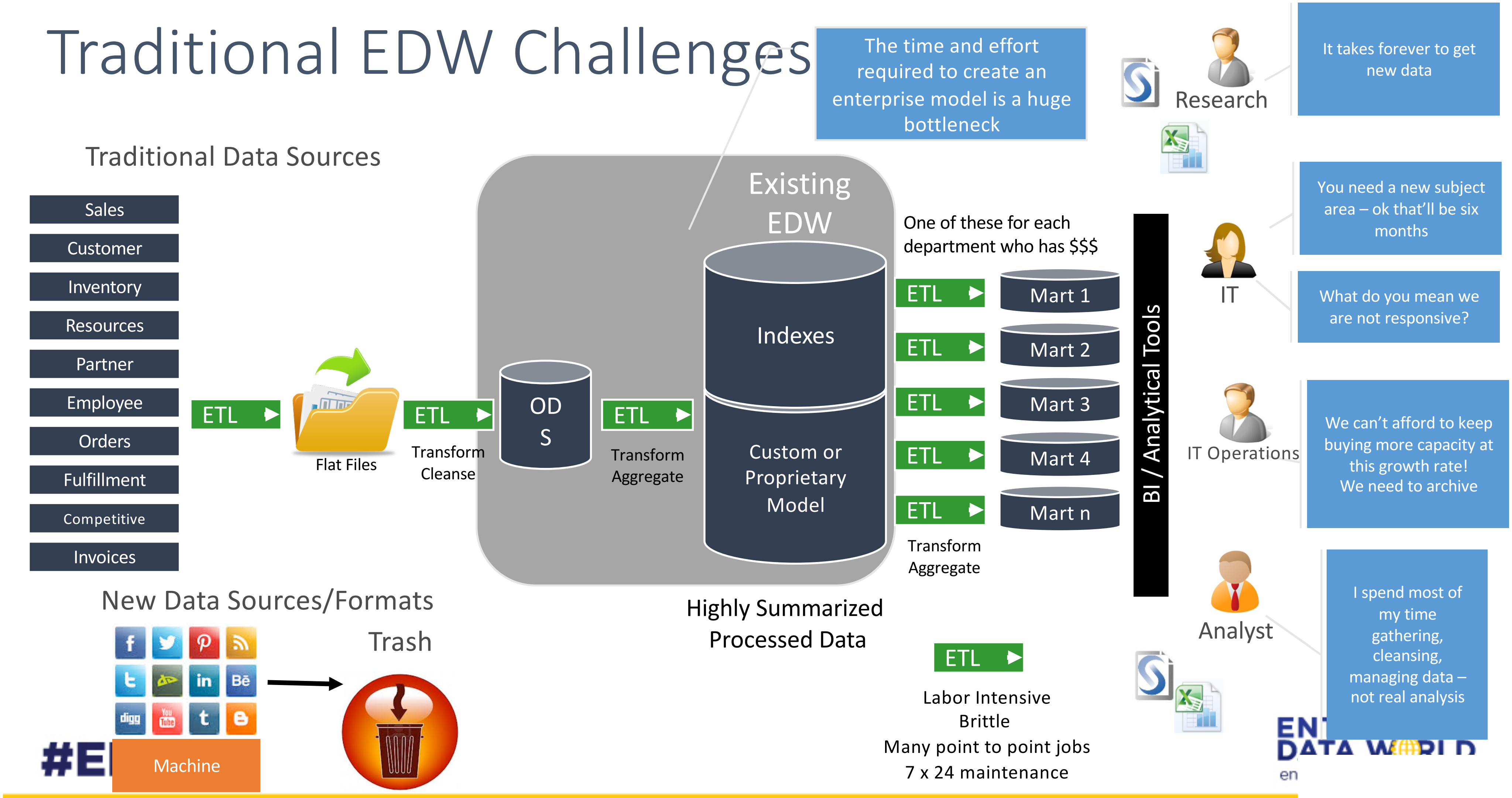
James Serra

, Big Data/Data Warehouse Evangelist at Microsoft

What are the benefits of Data Warehouses?

- Eliminate multiple access on same source data
- Reduce stress on production systems
- Optimized for read access
- Integrate many sources of data
- Keep historical records
- Model data differently than production
- Protect against source system upgrades
- Leverage Master Data Management, including hierarchies
- Self service analytics - No IT involvement required to create reports (?)
- Improve data quality problems in source systems
- One version of the truth

Traditional EDW Challenges



The time and effort required to create an enterprise model is a huge bottleneck

Traditional Data Sources

Sales
Customer
Inventory
Resources
Partner
Employee
Orders
Fulfillment
Competitive
Invoices

ETL

Flat Files

ETL

Transform
Cleanse

OD
S

ETL

Transform
Aggregate

Existing EDW

Indexes

Custom or Proprietary Model

Highly Summarized Processed Data

One of these for each department who has $$$

ETL – Mart 1
ETL – Mart 2
ETL – Mart 3
ETL – Mart 4
ETL – Mart n

Transform
Aggregate

BI / Analytical Tools

Research

It takes forever to get new data

IT

You need a new subject area – ok that'll be six months

What do you mean we are not responsive?

IT Operations

We can't afford to keep buying more capacity at this growth rate!
We need to archive

Analyst

I spend most of my time gathering, cleansing, managing data – not real analysis

New Data Sources/Formats

Machine

Trash

ETL

Labor Intensive
Brittle
Many point to point jobs
7 x 24 maintenance

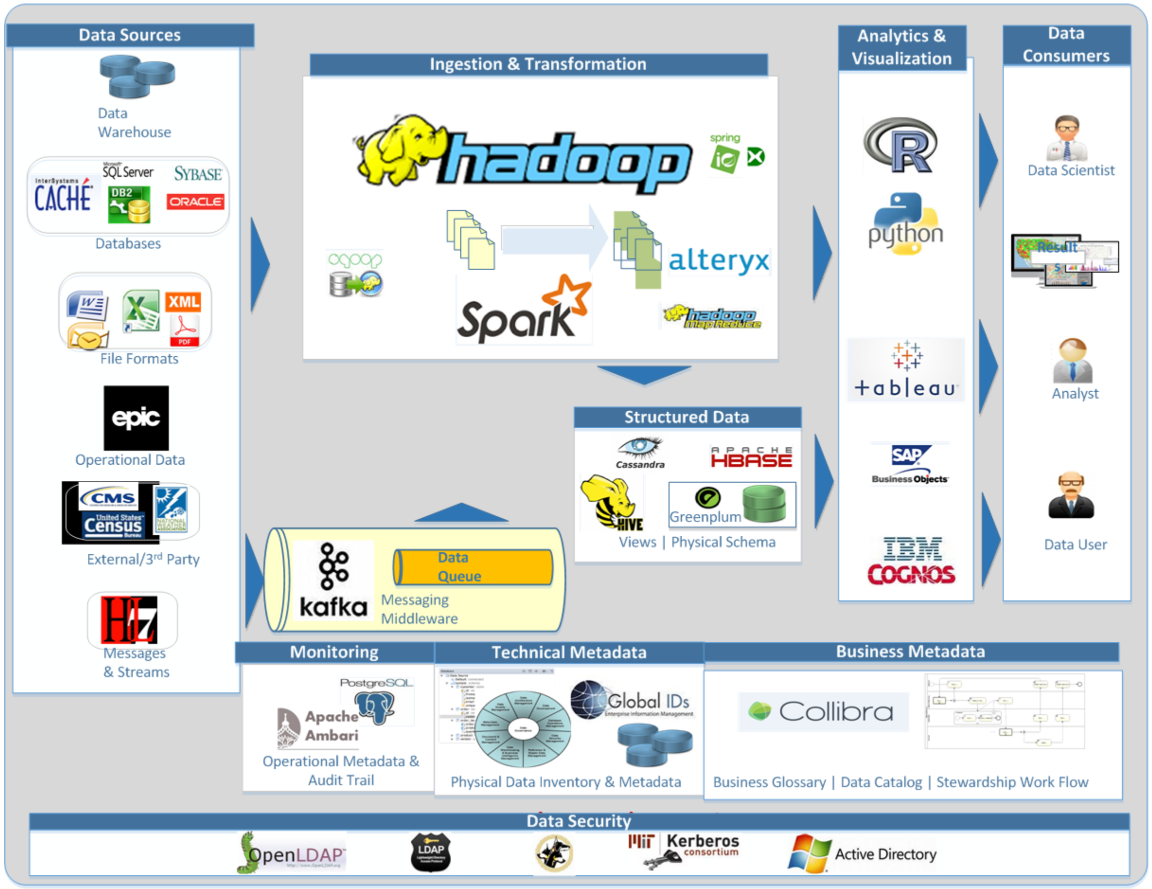
What are the problems with Data Warehouses?

- Time Consuming delivery
- Rigid model makes change complex
- No real solution for integration of unstructured data
- Expensive infrastructure and software
- Too time consuming to move data at volume
- High cost of redundant infrastructure
- Lost context and business meaning for sophisticated analytics
- Time to Integrate 3rd party and
  cloud based data with self defined data models

How do Data Lakes solve these problems?

- Enable highly skilled analysts to perform advanced analytics
- Allow data scientists to quickly get access to and assess new and external data sources, to determine if useful
- Commodity disk to keep large volumes of raw, detailed data at much lower cost
- Don’t move data to a central location (?)
- Access to raw, detailed data

Example Big Data Architecture



Problems with Data Lakes

- Resources to build/support new technologies are rare and expensive
- Resources to use new technologies and techniques are very expensive
- Governance (metadata, audit trail) and security (access) are not built in to new technology solutions and require advanced knowledge and effort to implement additional solutions and processes
- Not meant to support large numbers of users with limited analytics skills
- Doesn’t easily support data integration, real time processing, data update

Benefits of a Data Lake

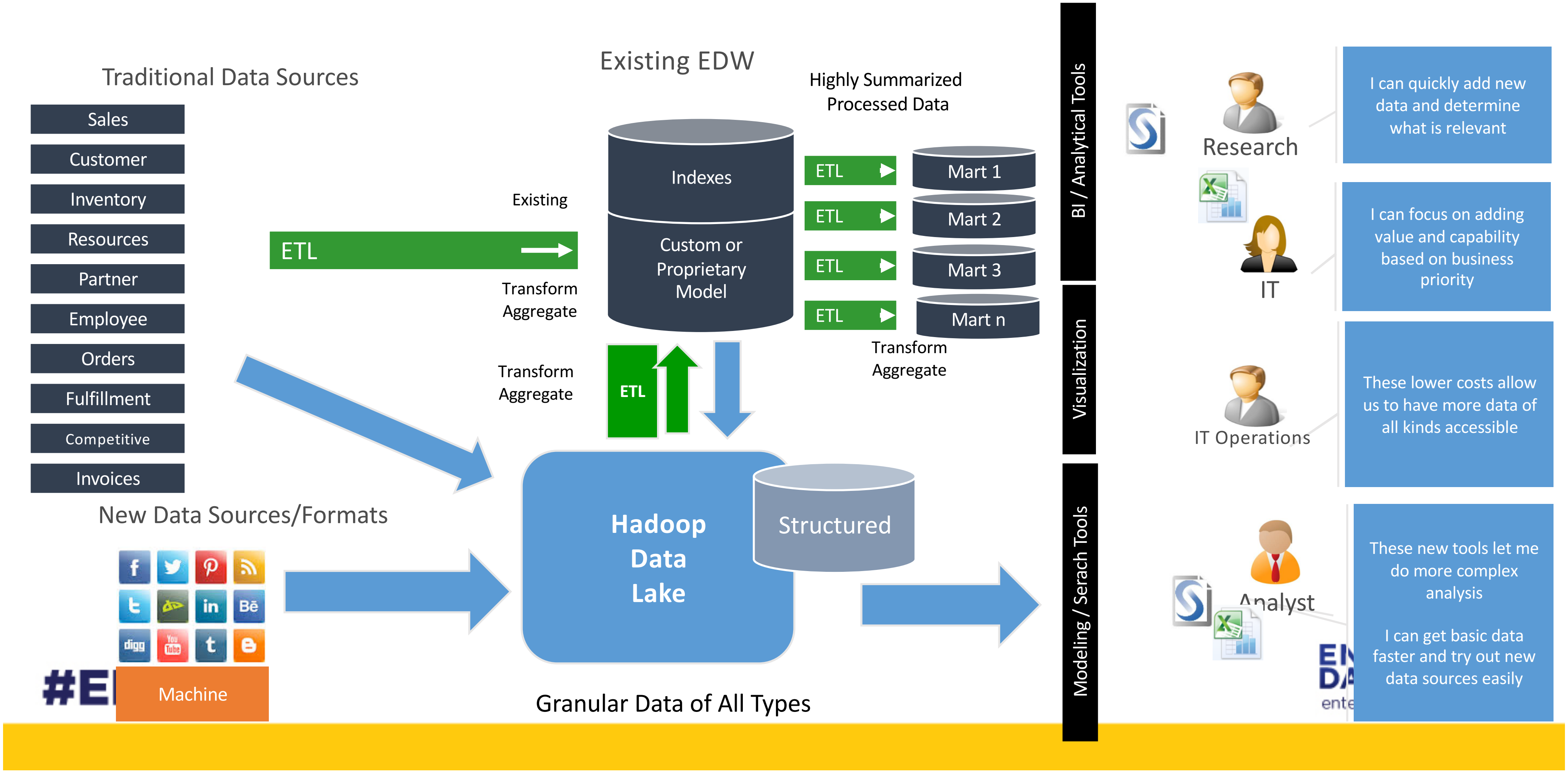
- Ability to store all types of structured and unstructured data in a data lake, from CRM data to social media posts
- More flexibility—you don't have to have all the answers up front
- Ability to store raw data—you can refine it as your understanding and insight improves

From Datafloq – "What is a Data Lake and What are the Benefits?"

- Ability to analyze very high volumes that are physically distributed
- Less expensive storage allows the maintenance of raw details
- Tools enabling complex analytics - predictive analytics and machine learning

Support of fast assessment of the value of a dataset and inclusion in analytics

How do Data Warehouses and Data Lakes work together?



Traditional Data Sources

Sales
Customer
Inventory
Resources
Partner
Employee
Orders
Fulfillment
Competitive
Invoices

New Data Sources/Formats

Machine

ETL

Existing

Transform
Aggregate

Existing EDW

Indexes

Custom or
Proprietary
Model

Highly Summarized
Processed Data

ETL Mart 1
ETL Mart 2
ETL Mart 3
ETL Mart n

Transform
Aggregate

Transform
Aggregate

ETL

Hadoop
Data
Lake

Structured

Granular Data of All Types

BI / Analytical Tools

Visualization

Modeling / Serach Tools

Research

I can quickly add new data and determine what is relevant

IT

I can focus on adding value and capability based on business priority

IT Operations

These lower costs allow us to have more data of all kinds accessible

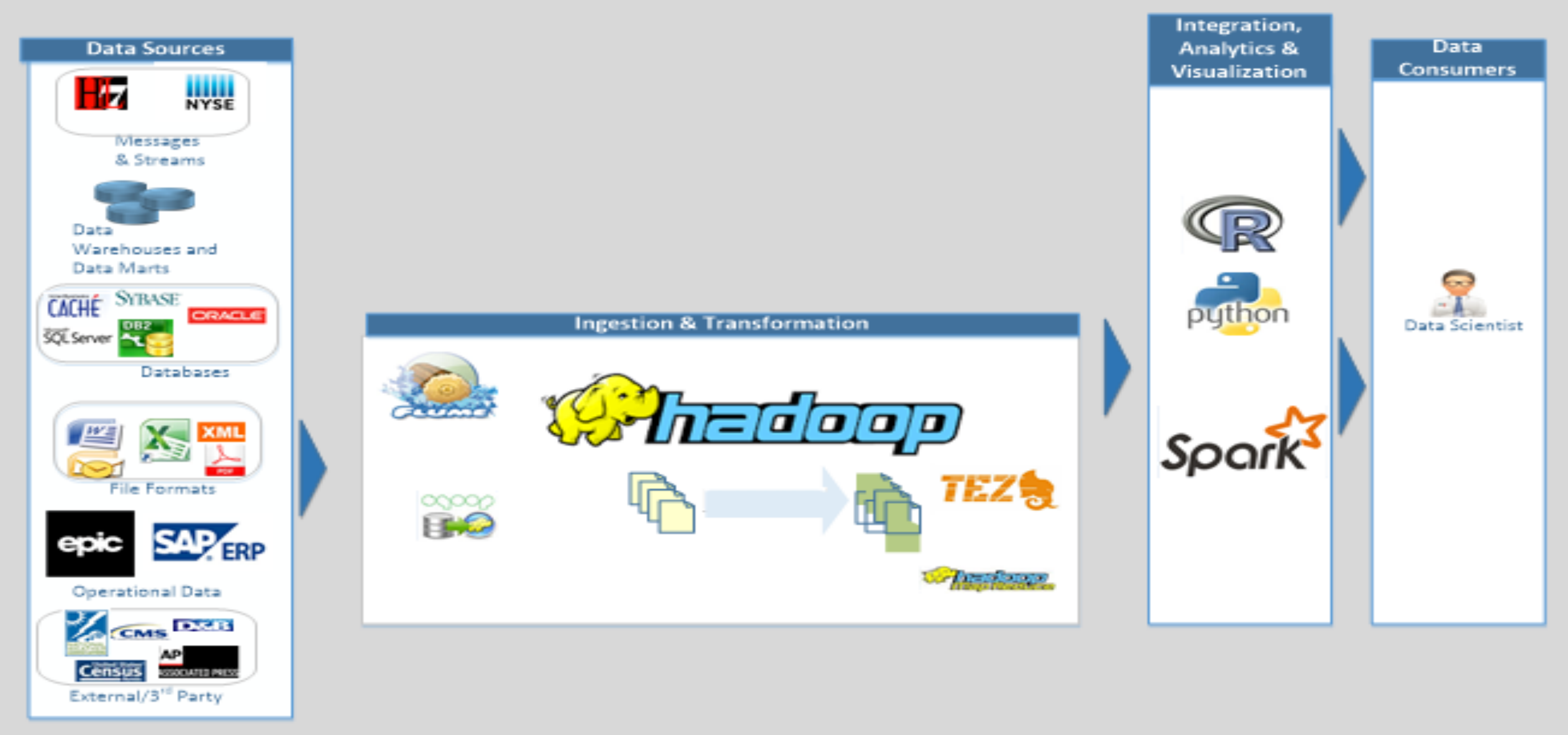
Analyst

These new tools let me do more complex analysis

I can get basic data faster and try out new data sources easily

Architecture of a Data Scientist Sandbox

Example Big Data Sandbox



Data Sources

Messages & Streams

Data Warehouses and Data Marts

Databases

File Formats

Operational Data

External/3rd Party

Ingestion & Transformation

Integration, Analytics & Visualization

Data Consumers

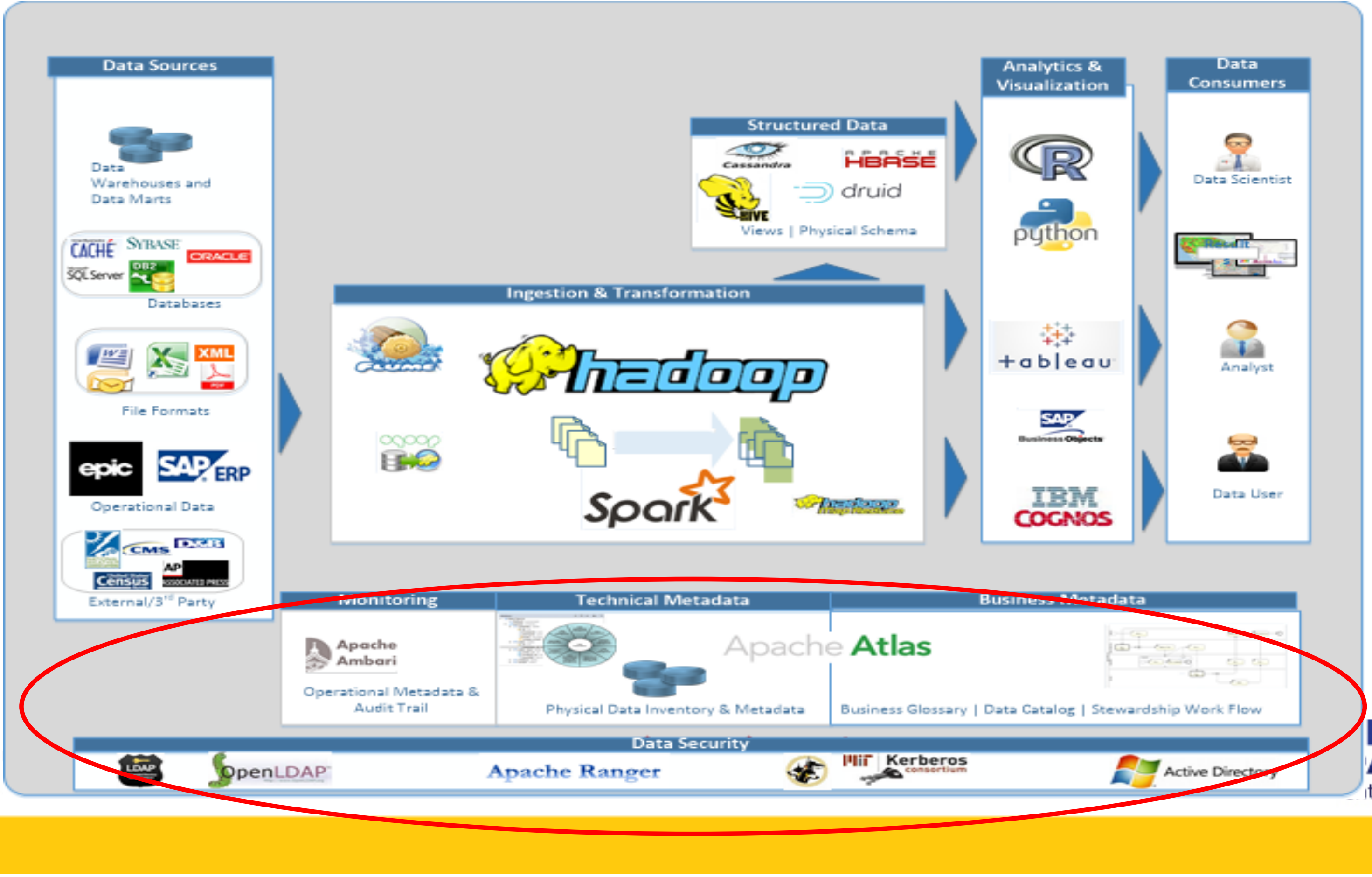
Data Scientist

Ad-Hoc Analysis Components

- Distributed data store for large data volumes of various types (i.e. Hadoop)
- Utilities for:
    - Data ingestion
    - Data transformation
    - Data analysis
    - Data Visualization
    - Reporting
- Audience – Data Scientists and Technical Resources

Components needed for Data Lake Governance

Example Data Lake Architecture – Batch Ingest



Data Sources

Data Warehouses and Data Marts

Databases

File Formats

Operational Data

External/3rd Party

Structured Data

Views | Physical Schema

Ingestion & Transformation

Analytics & Visualization

Data Consumers

Data Scientist

Analyst

Data User

Monitoring

Operational Metadata & Audit Trail

Technical Metadata

Physical Data Inventory & Metadata

Business Metadata

Business Glossary | Data Catalog | Stewardship Work Flow

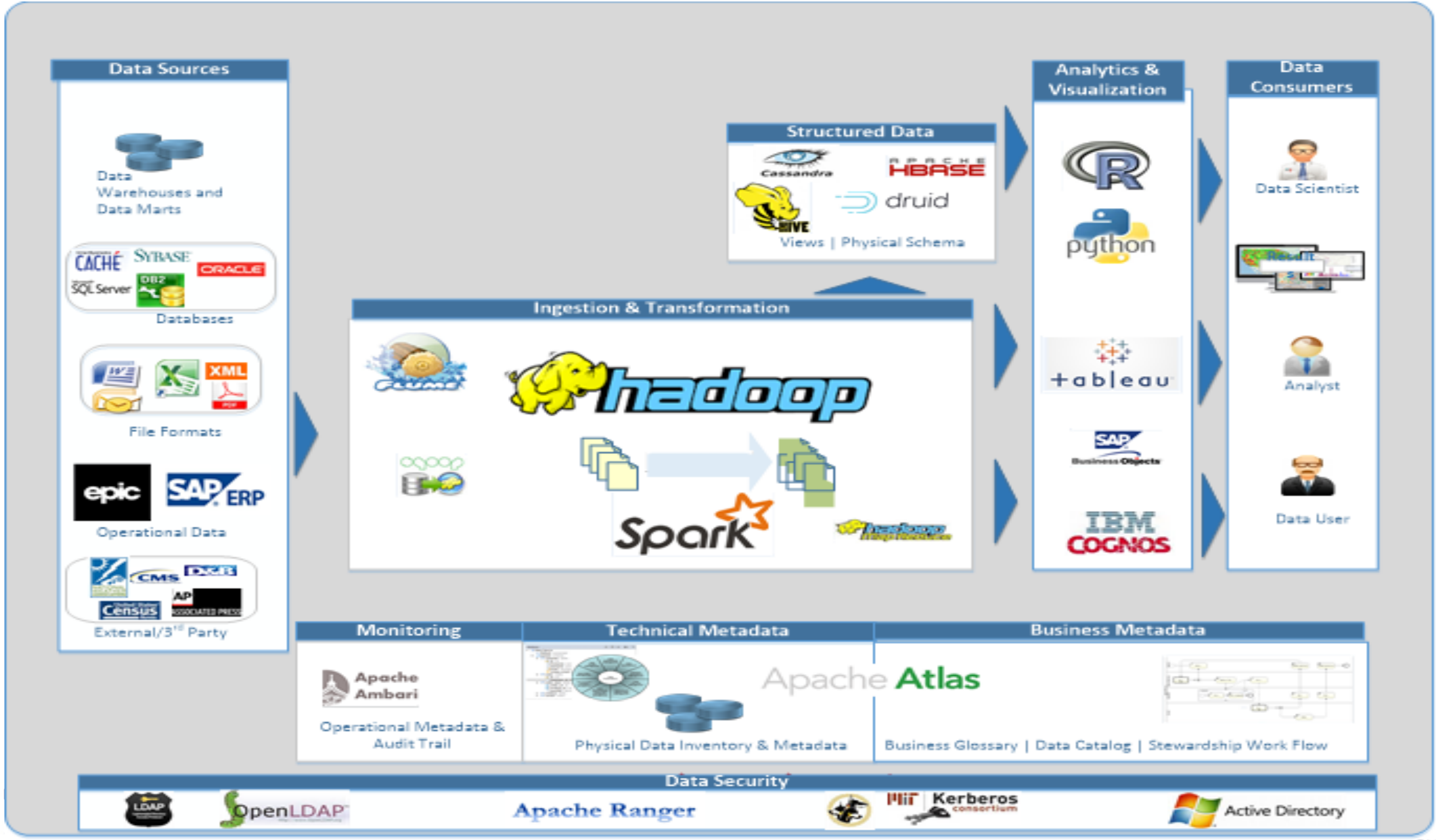
Data Security

Additional Big Data Governance Components

- Data security - authentication, access, encryption
- Metadata management
  - Technical metadata – physical format, structure, size, location (inventory)
  - Business metadata – meaning, acceptable use, rules, restrictions
  - Operational metadata – audit trail, when by whom was the data created, updated? Where did the data come from?

Architecture of a Big Data Analytics Lake

Example Big Data Architecture – with Batch Ingest

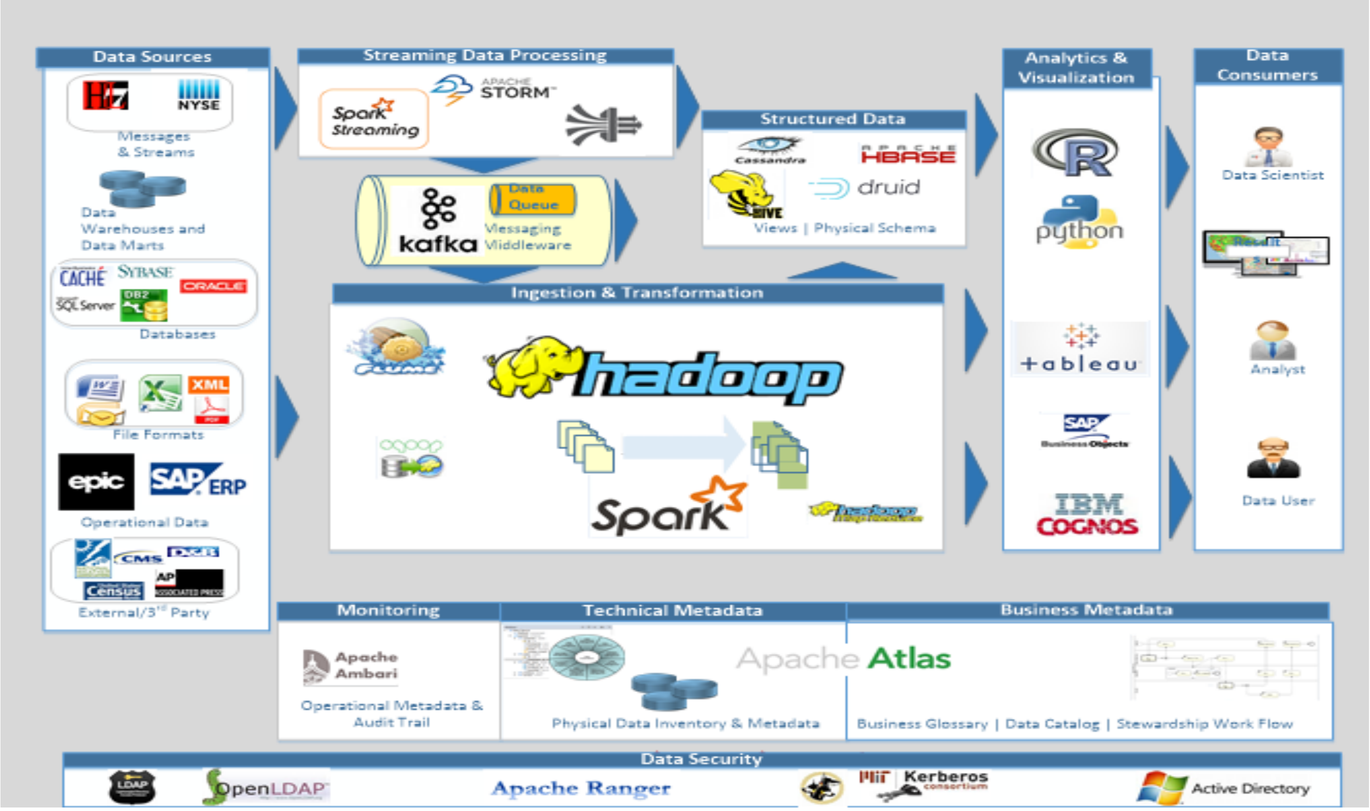


Operational Data Lake with Broader Audience Real-Time Data Access Components

- Data Pipelines and Orchestration – Yarn, Spark
- Enterprise Query and Reporting
  - Operational Reporting
  - Self Service Query and Reporting
- Real-time Data Query
  - Fast Query Database to store analytical results
    - Usually NoSQL key/value or columnar

Architecture of a Real-Time Streaming Operation

Example Real-Time Big Data Architecture



Additional Real-Time Data Components

- Real-time Data Ingestion
  - Data Streaming
    - Real-time data streams – events, data source with no end
  - Data Messaging
    - Used for processing incoming streams for later analysis or analysis of groups of events
  - In-memory database for extreme low latency requirements
    - For processing that doesn't have the time to write to disk

Recommendations in Implementing Data Lakes

Recommendations

- Use Big Data technology to support complex analysis of large volumes of data in a variety of technology and formats, such as predicting the behavior of customers and markets. Don't use to replace transaction processing solutions.

- Include additional components that support data security, technical format inventory documentation, and audit trail functionality in the Big Data architecture

- Gradually build up support for more complex use cases in order to develop experience supporting the new Big Data technology and add additional components as needed.
  - Start with an environment for complex ad-hoc data analysis by highly skilled analysis (Data Scientists).
  - Continue with use cases for operational production analysis and reporting against large volumes of data of various types.
  - Add additional use cases for real-time data query and then streaming data ingestion with appropriate additional technical components.

- Include change management: the selection, implementation, and support of the open source software components require adjustments to enterprise standards for commercial software selection and support contracts

April Reeve

- AprilReeve@sprintmail.com
- @Datagrrl on Twitter
- **Book - “Managing Data in Motion – Data Integration Best Practice Techniques and Technologies”**

#EDW17

ENTERPRISE DATA WORLD
enterprisedataworld.com