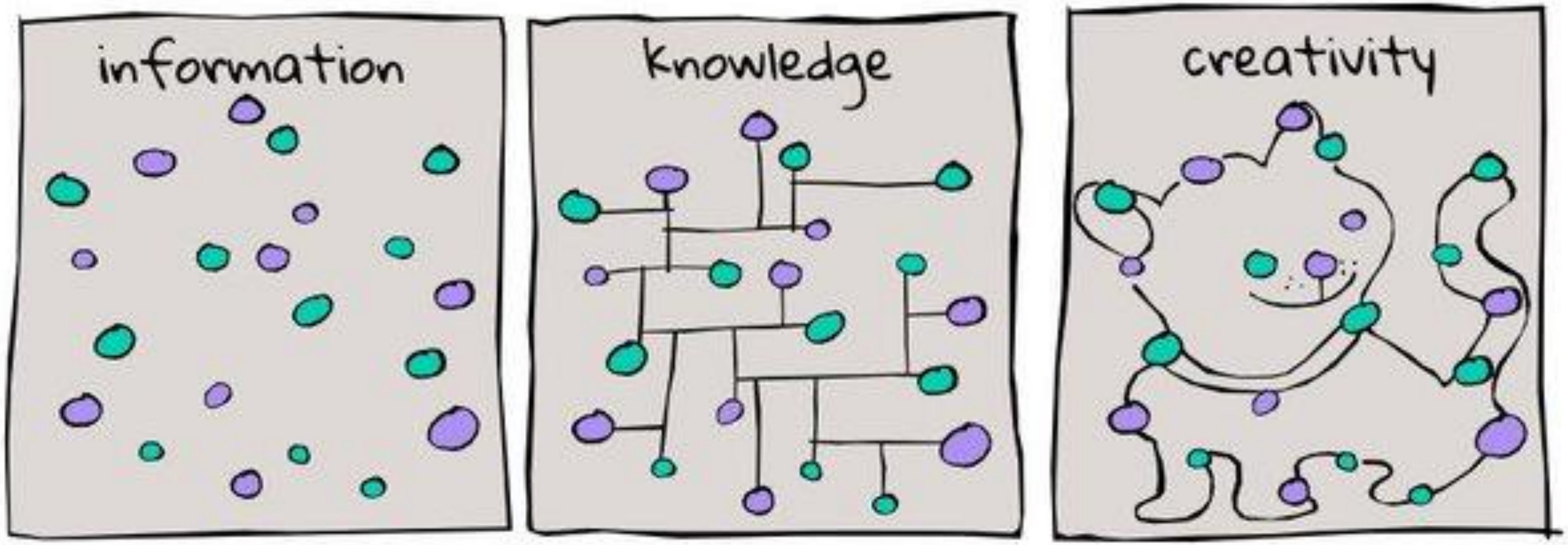
Big Data Analytics

Good Morning to all Big Data Analyst



Agenda

- Deploy Hadoop Distribution
- Reference Architecture
- Spark Intro : A Distributed Processing Engine
- Python Crash Course
- Spark DataFrame Basic (if time permit)

Installation of Hadoop Distribution

- **HDP (Hortonworks Data Platform)** on your Machine (Using Virtual Box Sandbox)

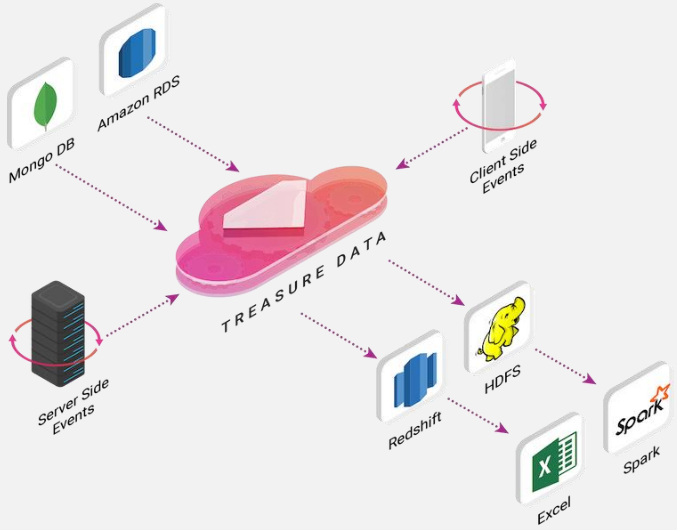
- **CM (Cloudera Manager)** on Cloud in Multi-node Cluster (Using GCP)

Ref: http://arif.works/hadoopadmin

What is Reference Architecture

A reference architecture shows which **functionality** is generally needed in a certain domain or the solve a certain class of problems. Also, how this functionality is divided and how **information flows** between the pieces (called the reference model).

It then maps this functionality onto software elements and the data flows between them.

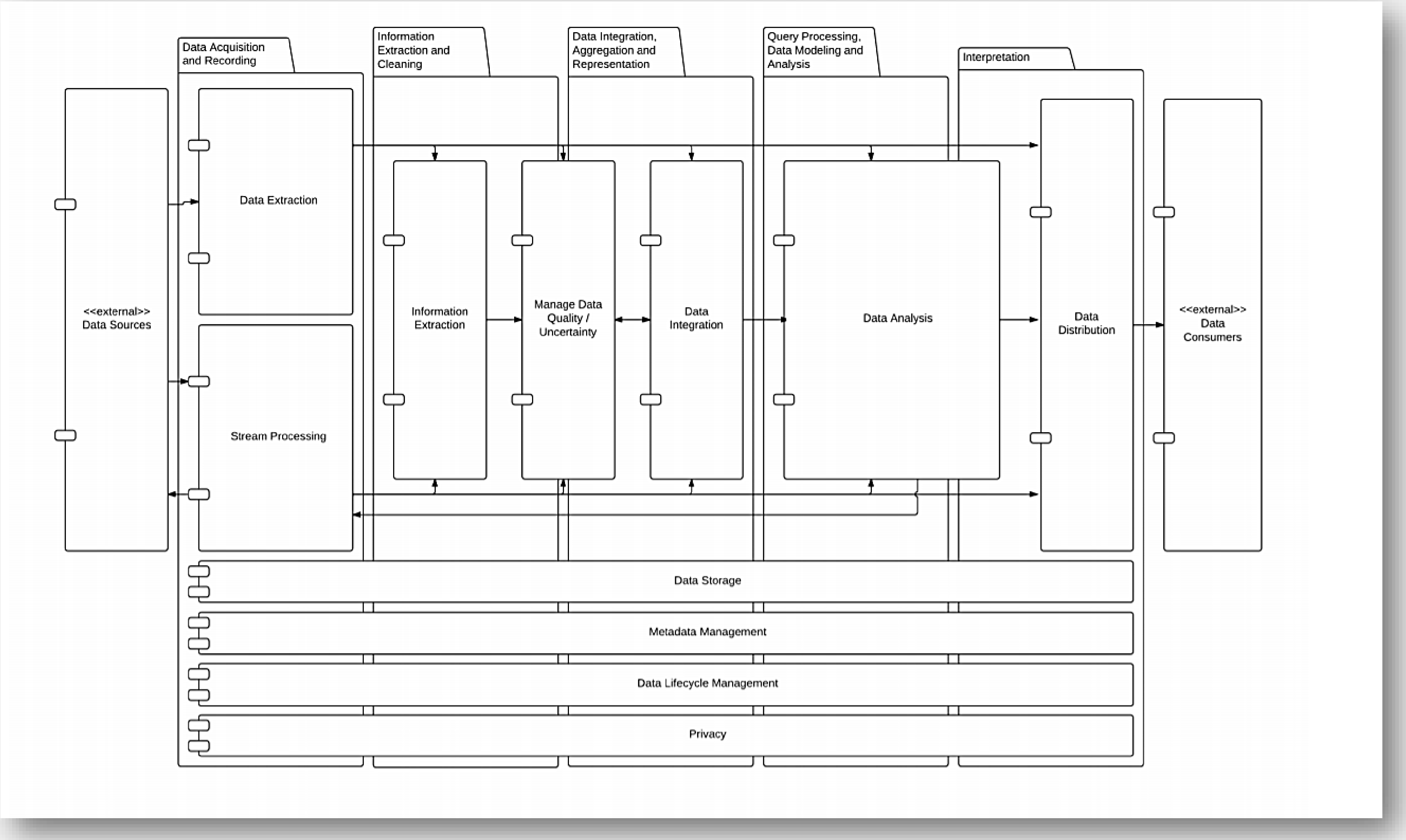


Reference Architecture: High-level functional view

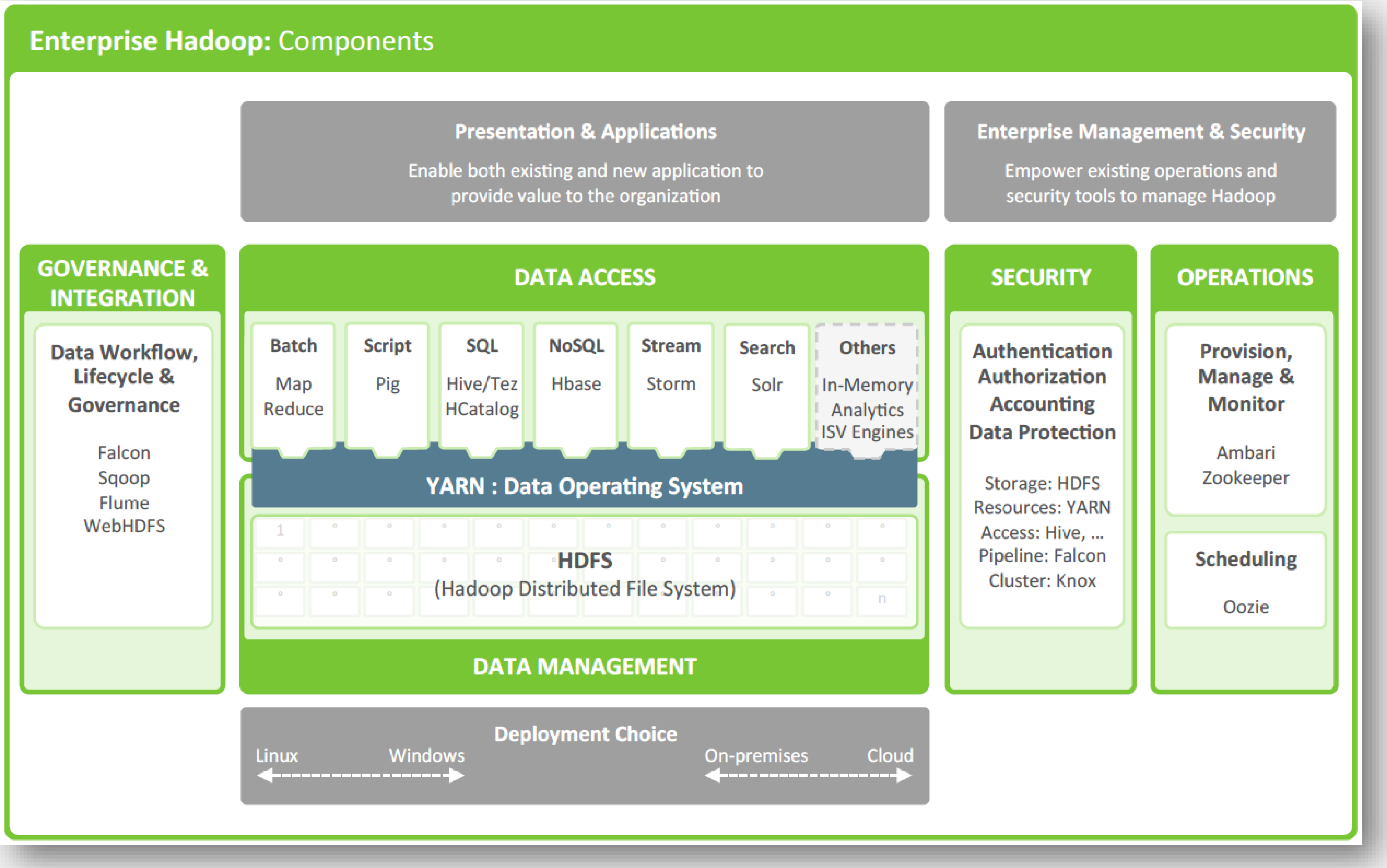
The analysis of ‘big data’ into distinct phases of a sequential processing **pipeline**:

- Acquisition / Recording,
- Extraction / Cleaning / Annotation,
- Integration / Aggregation / Representation,
- Analysis / Modeling, and
- Interpretation.

Reference Architecture: High-level functional view


Data Acquisition and Recording
Information Extraction and Cleaning
Data Integration, Aggregation and Representation
Query Processing, Data Modeling and Analysis
Interpretation
<<external>> Data Sources
Data Extraction
Stream Processing
Information Extraction
Manage Data Quality / Uncertainty
Data Integration
Data Analysis
Data Distribution
<<external>> Data Consumers
Data Storage
Metadata Management
Data Lifecycle Management
Privacy

Reference Architecture: High-level functional view



Enterprise Hadoop: Components

Presentation & Applications

Enable both existing and new application to provide value to the organization

Enterprise Management & Security

Empower existing operations and security tools to manage Hadoop

GOVERNANCE & INTEGRATION

Data Workflow, Lifecycle & Governance

Falcon
Sqoop
Flume
WebHDFS

DATA ACCESS

Batch Map Reduce

Script Pig

SQL Hive/Tez HCatalog

NoSQL Hbase

Stream Storm

Search Solr

Others In-Memory Analytics ISV Engines

YARN : Data Operating System

1

n

HDFS
(Hadoop Distributed File System)

DATA MANAGEMENT

SECURITY

Authentication Authorization Accounting Data Protection

Storage: HDFS
Resources: YARN
Access: Hive, …
Pipeline: Falcon
Cluster: Knox

OPERATIONS

Provision, Manage & Monitor

Ambari
Zookeeper

Scheduling

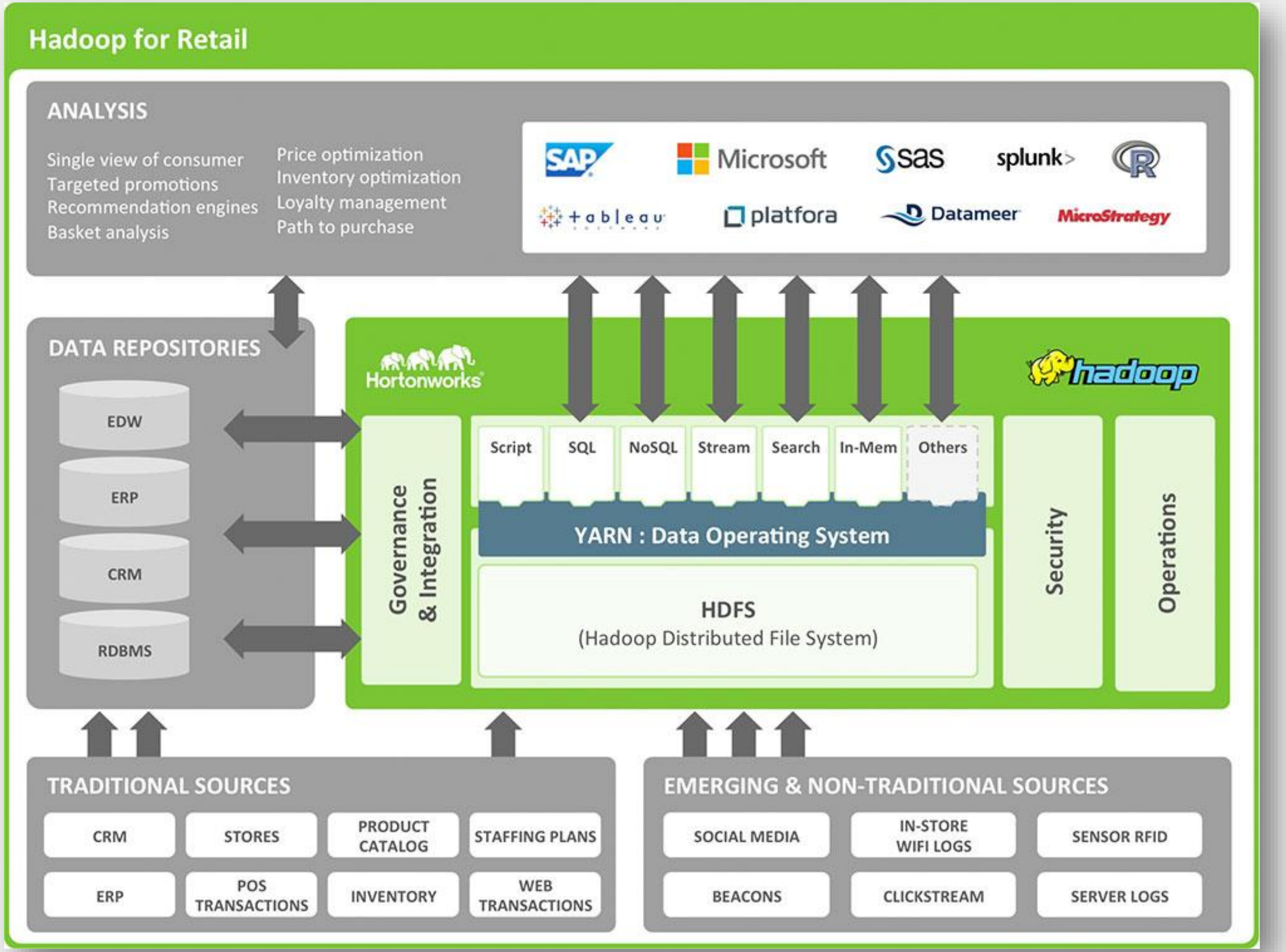
Oozie

Deployment Choice

Linux ⟷ Windows

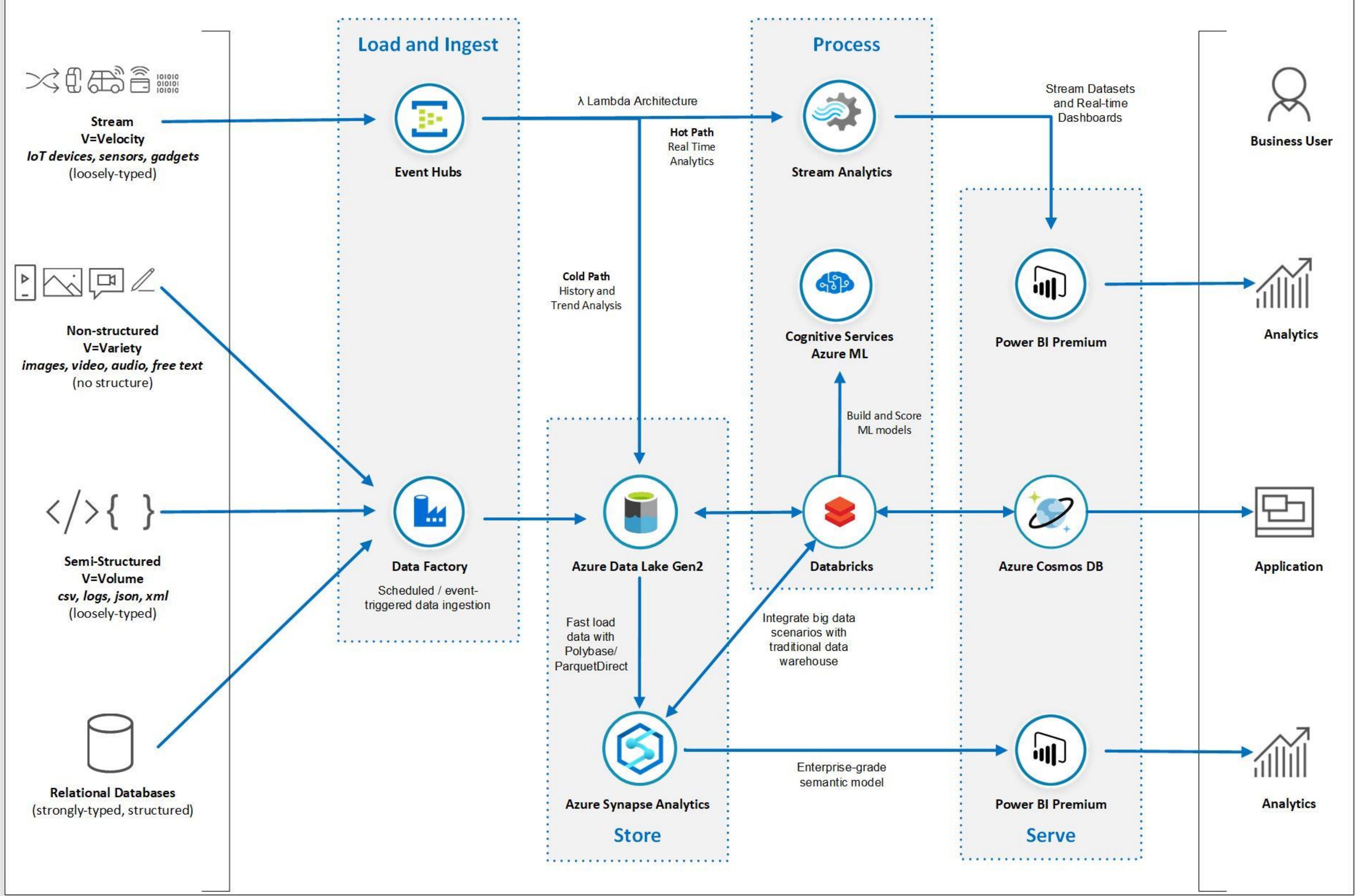
On-premises ⟷ Cloud

Reference Architecture: Retail (Example)



Reference Architecture: Modern Data Platform Azure

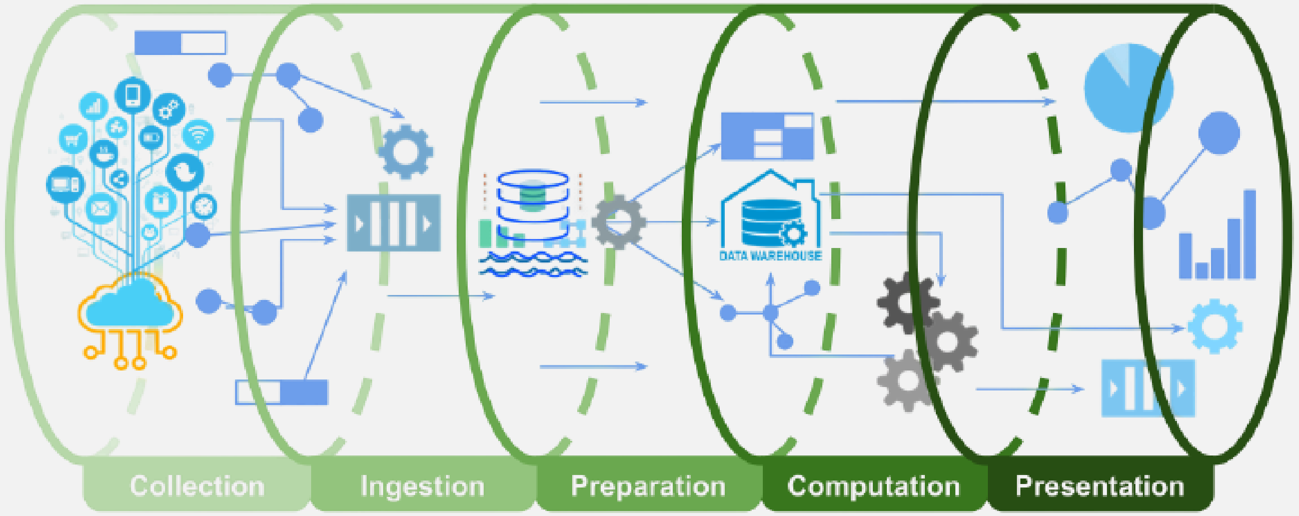
Modern Data Platform Reference Architecture



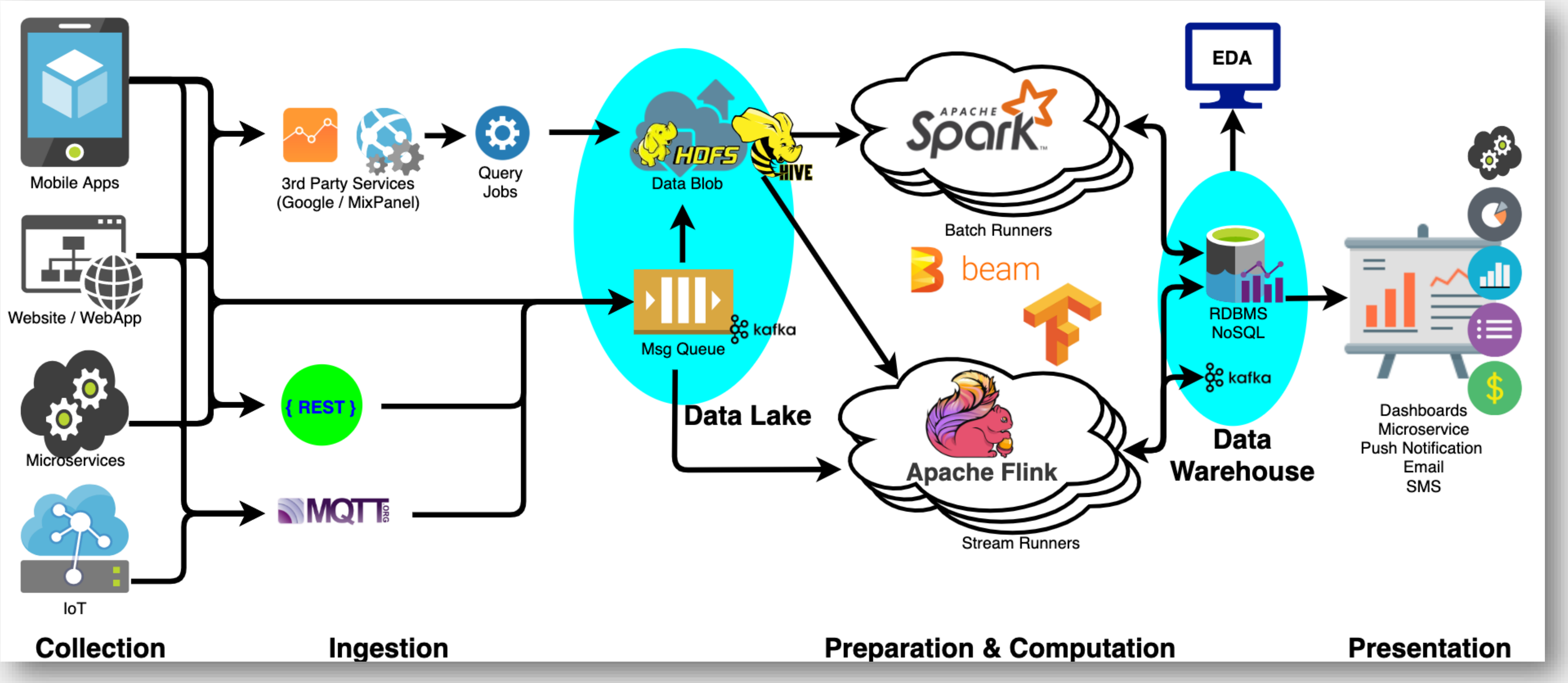
In summary,

Data Pipeline has five stages:

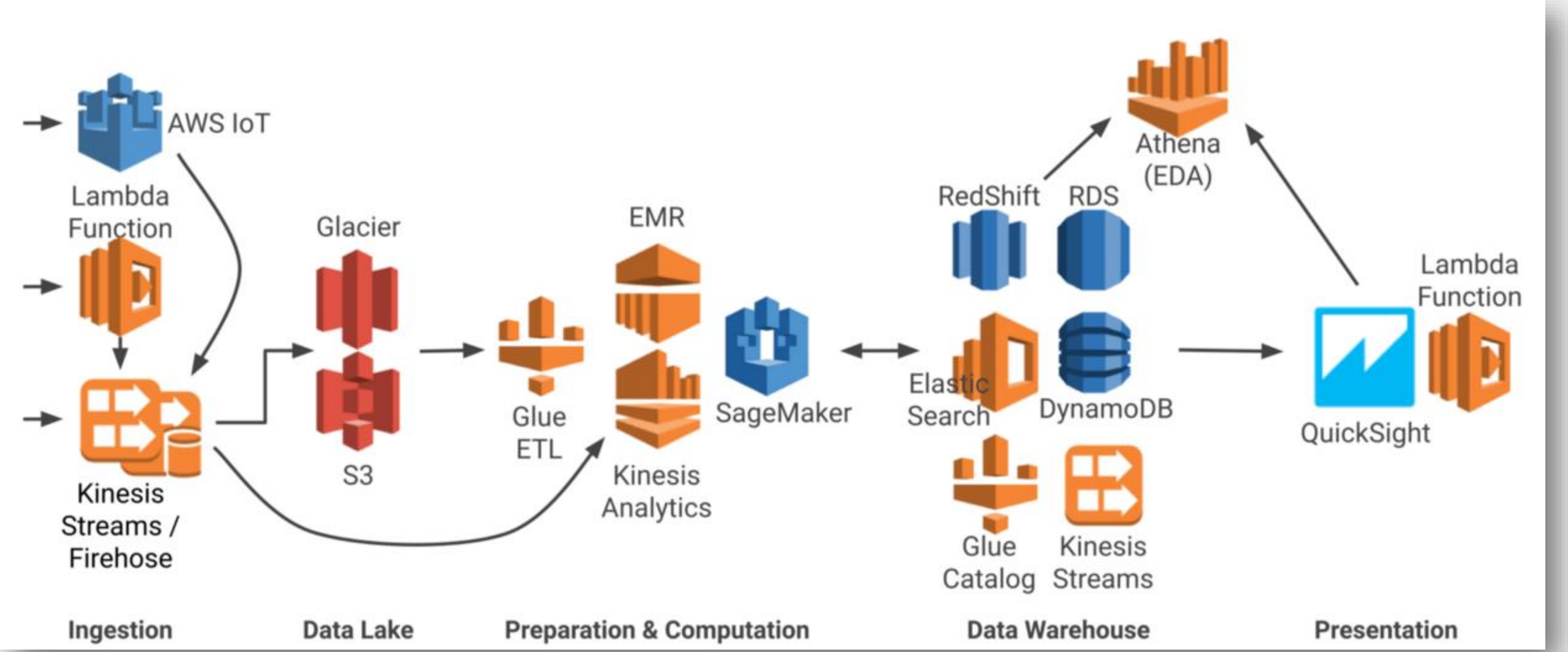
- Collection
- Ingestion
- Preparation
- Computation
- Presentation



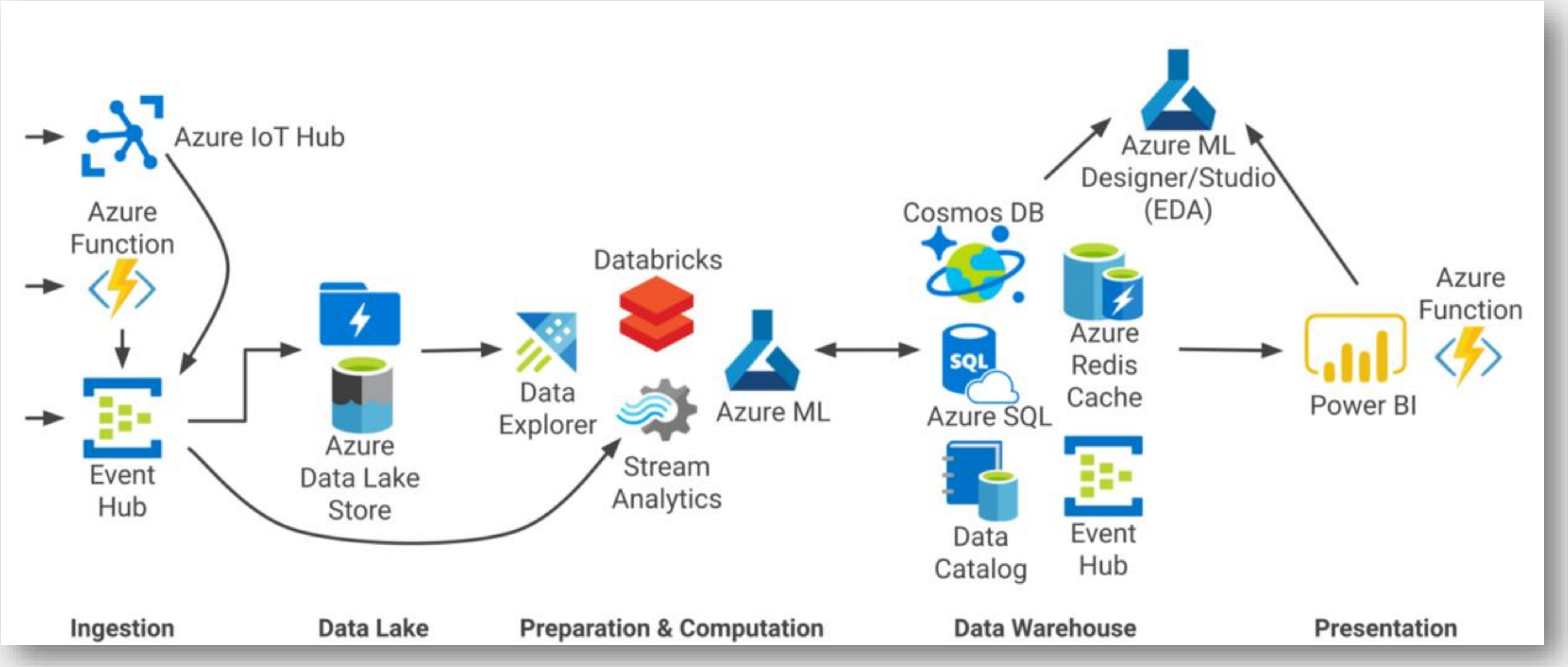
Big Data Architecture: Open Source Technologies



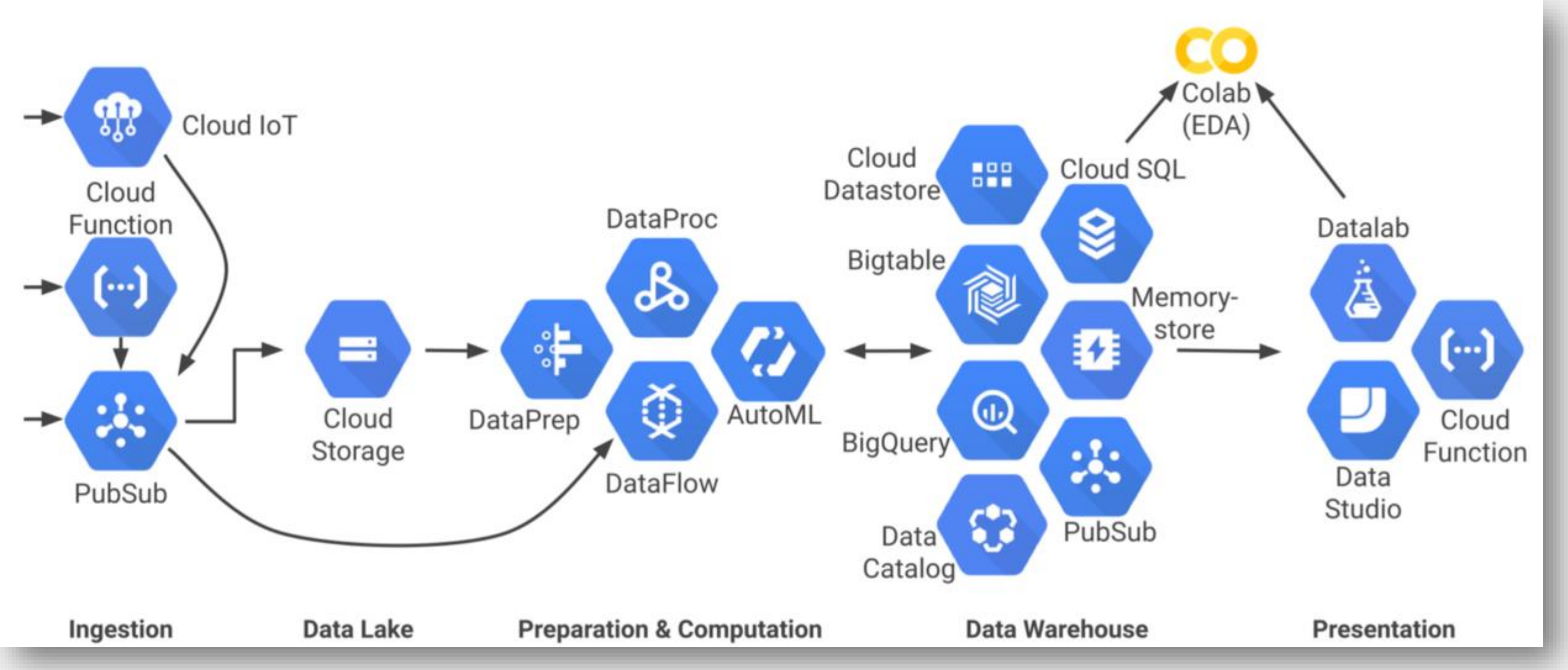
Big Data Architecture: Amazon Web Services



Big Data Architecture: Microsoft Azure



Big Data Architecture: Google Cloud Platform



Case Study on Big Data Architecture

Background

A multi-national company ABC provides financial services including investment services and insurance services using the online platform for financial products transactions. ABC has millions of clients worldwide to use their services and do online transactions. ABC has thousands of staffs to maintain the platform, serve the clients requests, collect and use the data to generate insights. However, there are no efficient systems to ingest the data and transform them into meaningful insights. The profit has dropped among this year and ABC wants to investigate the reasons from data point of view.

Case Study on Big Data Architecture

Needs and Requirements

ABC has requested to build a data warehouse and data streaming **pipeline** platform to utilize the data. In addition, ABC wants to use the data to analyze customer buying behavior trend and recommend them with personalized products.

Case Study on Big Data Architecture

Data Sources

After consolidating the needs and every kinds of data necessary, we have collected below three types of data:

1. Traditional enterprise data from operational systems related to customer touch points such as:

❑ Call centres.

❑ Branches/Brokerage units.

❑ Credit cards.

❑ Volatility measures that impact the clients' portfolios.

Case Study on Big Data Architecture

Data Sources

2. Financial business forecasts from various sources such as:

❑ Industry data.

❑ Trading data.

❑ Alerts about events (news, blogs, Twitter and other messaging feeds).

Case Study on Big Data Architecture

Data Sources

3. Other sources of data such as:

❑ Advertising response data.

❑ Social media data.

As far as we know, the data is a category of **structured**, **unstructured** and **semi-structured** data which increased the difficulty of building the big data system.

Case Study on Big Data Architecture

Key Questions To Think Through

1. What is the business challenge?
2. What is the need for data transactions: real-time vs. batch processing vs. transaction processing?
3. What is the data ingestion / storage challenge?
4. Among big data architecture, what tools and languages will be used?

Let’s keep these questions in mind and go through them one by one.

Case Study on Big Data Architecture

What is the business challenge?

ABC is facing the challenge of connecting efficiently with their customers without the prescriptive analytics platform to visualize their data and provide customer insights in terms of customer buying behaviors and customer engagement.

ABC also doesn’t have a good data ingestion / storage platform and architecture design to make use of their data. The inability of data modeling / analytics / architecture leads to the potential loss of customer satisfaction and further loss of their customers from business perspective.

Case Study on Big Data Architecture

What is the business challenge?

In summary, there are three main challenges:

1. Understand the transactions of customer data, financial data and other kinds of data across platforms, design the big data storage techniques (data formats, data store types and storage tools)
2. Determine data pipeline architecture in terms of real-time vs batch processing, tools and language selection
    i. Real-time vs batch processing
    ii. Tools and language selection
3. Build a machine learning platform to predict:
    i. How location impacts customer behavior in order to provide accurate financial products offering
    ii. Customer engagement based on their preferences to recommend more attractive business selling

Case Study on Big Data Architecture

What is the need for data transaction: real-time vs. batch processing vs. transaction processing?

To determine which method is the best, we have to understand what each means and best used case.

- ✓ **Real-time** processing, or streaming processing, means the data can be processed in milliseconds.
- ✓ **Batch processing** is collecting all data at specific time and process all at once in a regular manner.
- ✓ **Transactional processing**, or data store processing, means once data has been processed by streaming or batch computations, it needs to be stored in a way that can be quickly accessed by a data scientist.

Case Study on Big Data Architecture

What is the need for data transaction: real-time vs. batch processing vs. transaction processing?

Scenario: Given that Twitter data can be used to find reasons on customer dissatisfaction, the DS team wants to build a pipeline for sentiment analytics.

What type of processing should they choose?

Case Study on Big Data Architecture

What is the data ingestion / storage challenge?

The data described above comes from different sources, e.g., from databases, log files, online financial applications, social media networks and offline operational systems. These data are collected in databases, local disks, and cloud file systems (such as AWS S3) in structured and unstructured formats. We need to collect these data and integrate it into a data lake, transform and deliver it into a data warehouse.

What storage type to choose?

Case Study on Big Data Architecture

What is the data ingestion / storage challenge?

Scenario 1: User profiles storage

Every user has unique profile data such as user id, name, gender, and other identification attributes, as well as preferences such as language, time zone, which products the user has access to, and so on. In this case, we assume that each user has a unique key, and assume that some profiles data are optional to fill in, such as user age.

The optimal storage will be **key-value store** or column family store, providing the capability of in-memory processing. However, if the data is not growing, RDBMS could be an option but not optimal.

Case Study on Big Data Architecture

Among big data architecture, what tools and languages will be used?

There are thousands of tools in the market to evaluate and compare. Can you make a Reference Architecture with Open Source Ecosystem components?

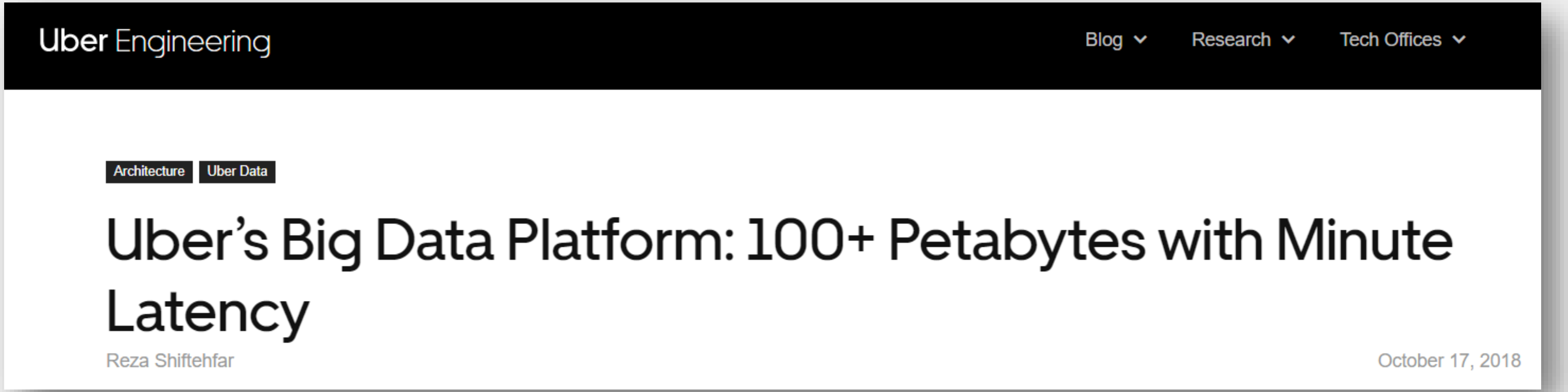
Just have a look how many ecosystem components are there?

http://bigdata.andreamostosi.name/

Case Study (Uber)

Transformation Journey towards Big Data Platform.

Please do read this article.



Uber's Big Data Platform: 100+ Petabytes with Minute Latency

Reza Shiftehfar

October 17, 2018

https://eng.uber.com/uber-big-data-platform/