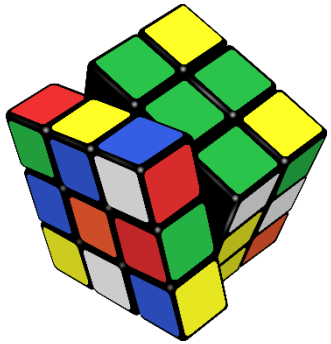


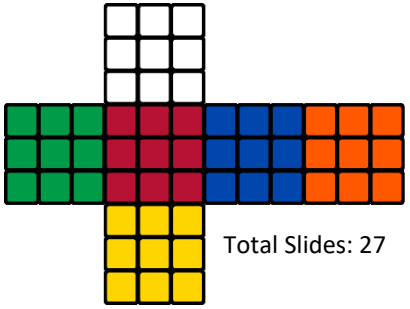
Big Data Engineering For Analytics

Modern Big Data Platform

Suria R Asai

(suria@nus.edu.sg)

NUS-ISS



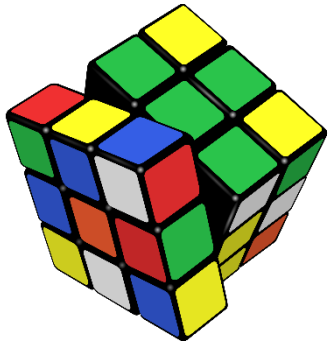
Total Slides: 27

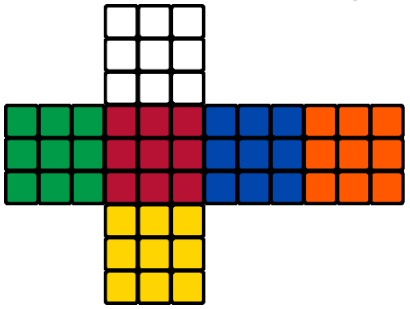
Learning Objectives

- *Define a modern Big Data platform*
- *Describe expectations from data*
- *Describe expectations from a platform*

Agenda

- Defining Modern Big Data Platform
  - *List components of Big Data platforms*
  - *Describe uses cases for Big Data*
- Concluding Remarks



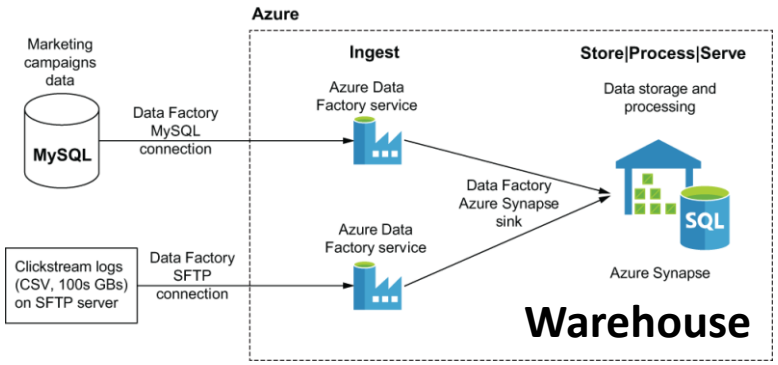


Modern Big Data Platform

Steven Hawking on the Universe:

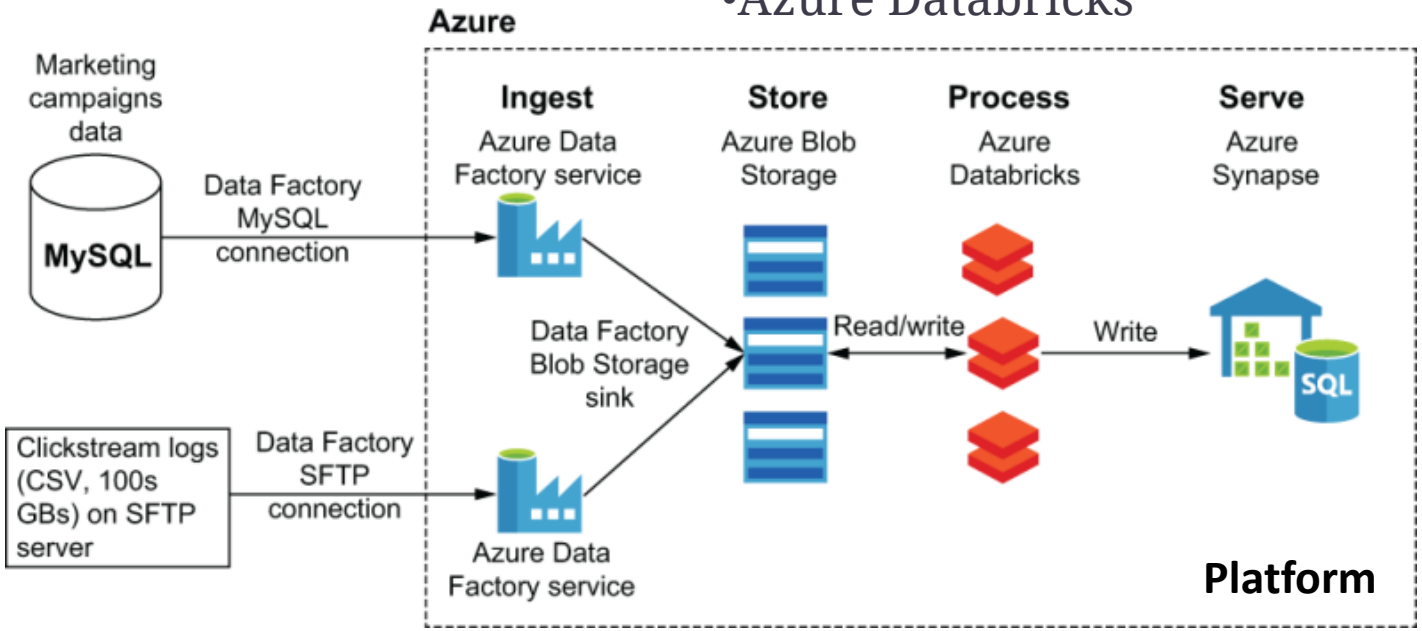
"It would not be much of a universe if it wasn't home to the people you love."

Cloud Data Warehouse Vs Platform - Example

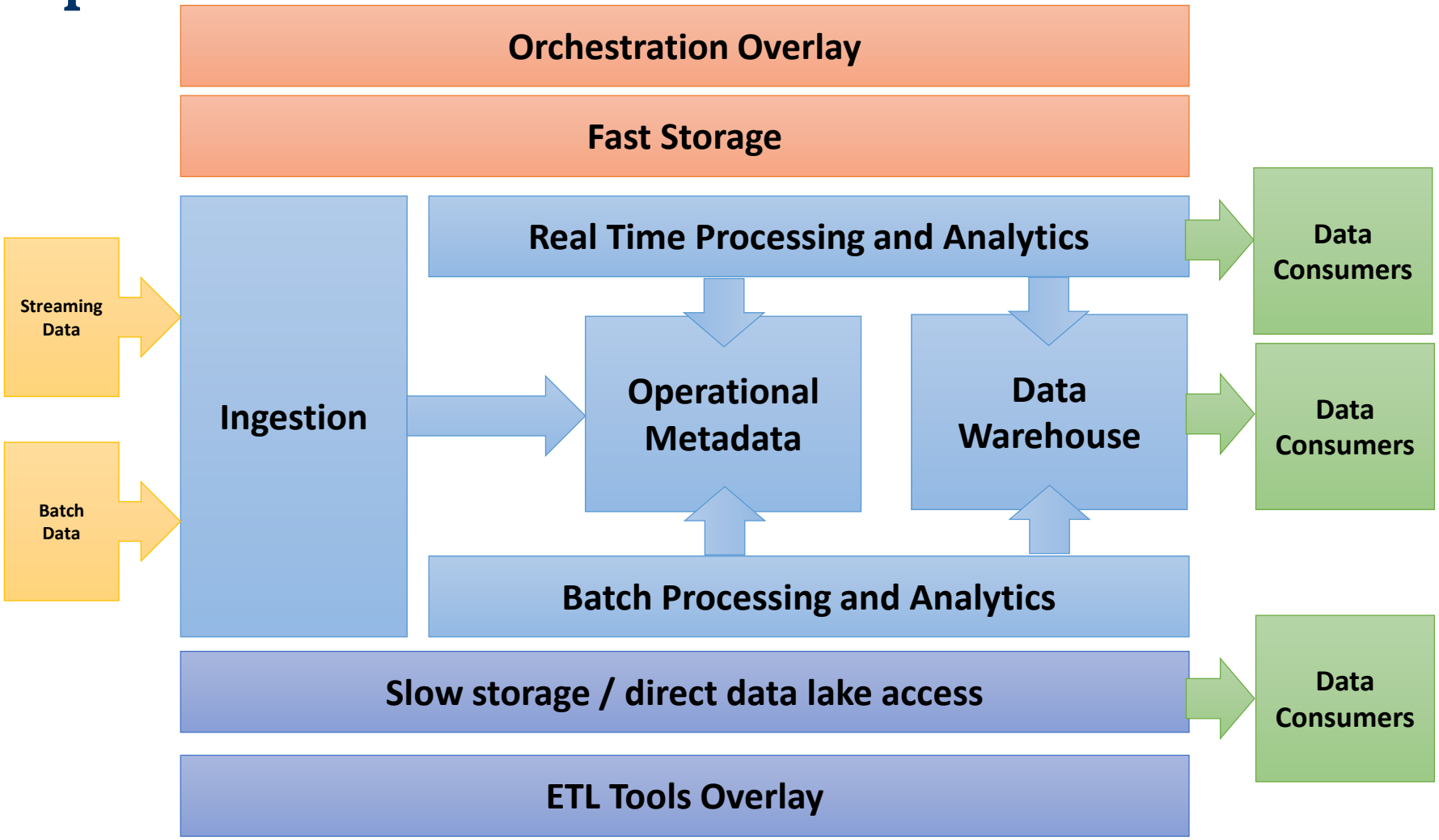


This example cloud data platform architecture consists of these Azure PaaS services:

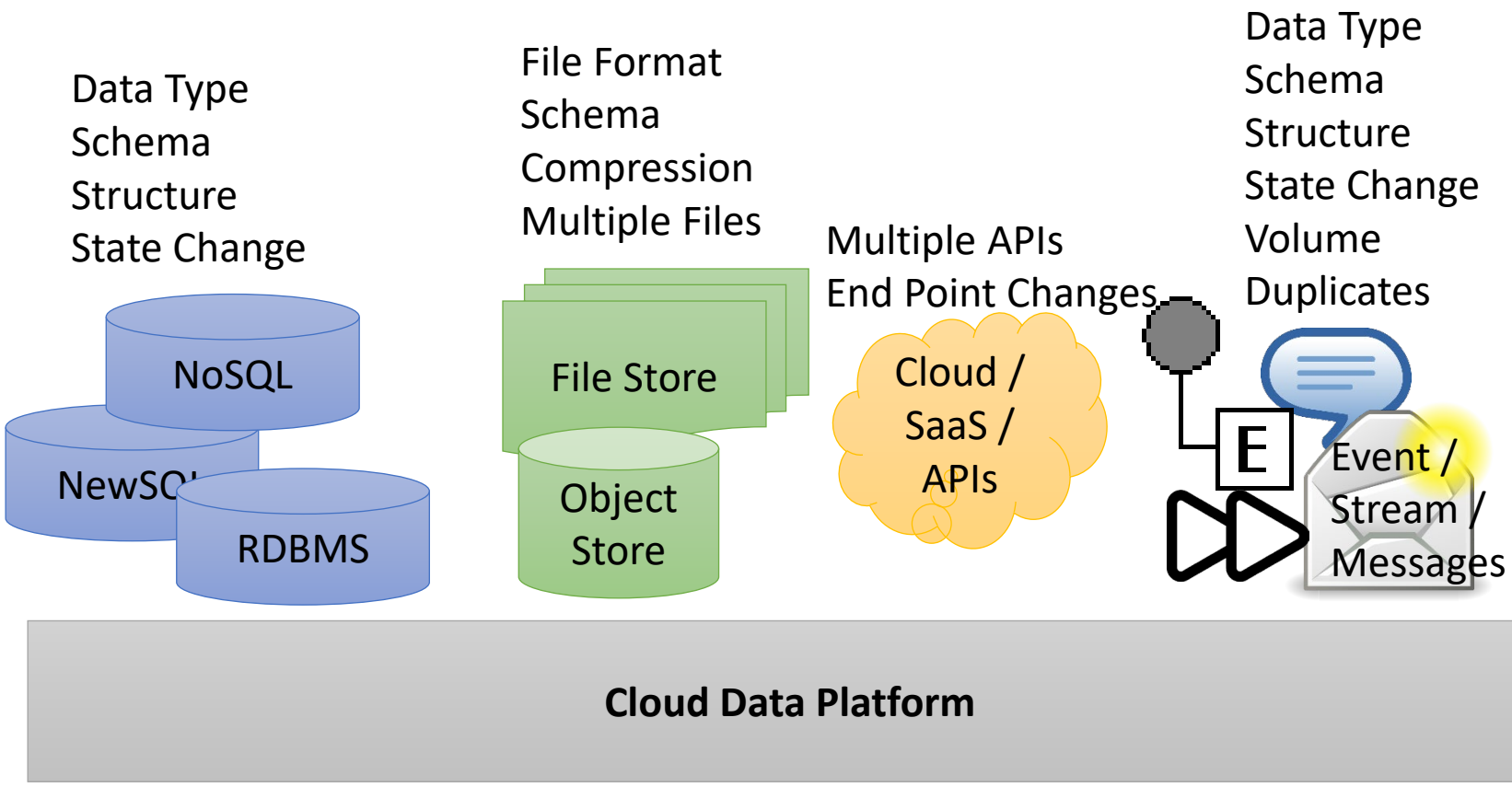
- Azure Data Factory
- Azure Blob Storage
- Azure Synapse
- Azure Databricks

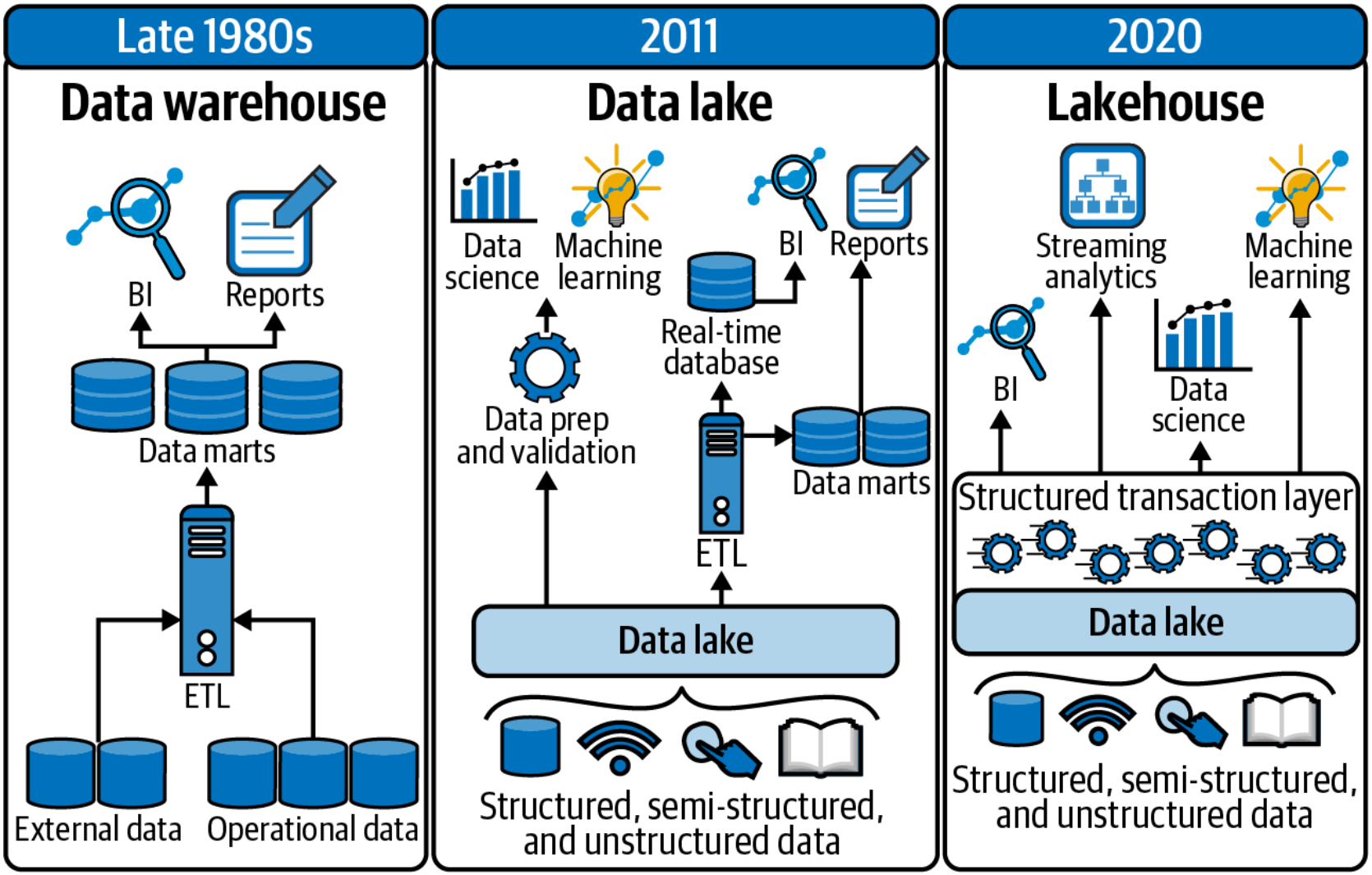


So what are the essential layers of data platform?



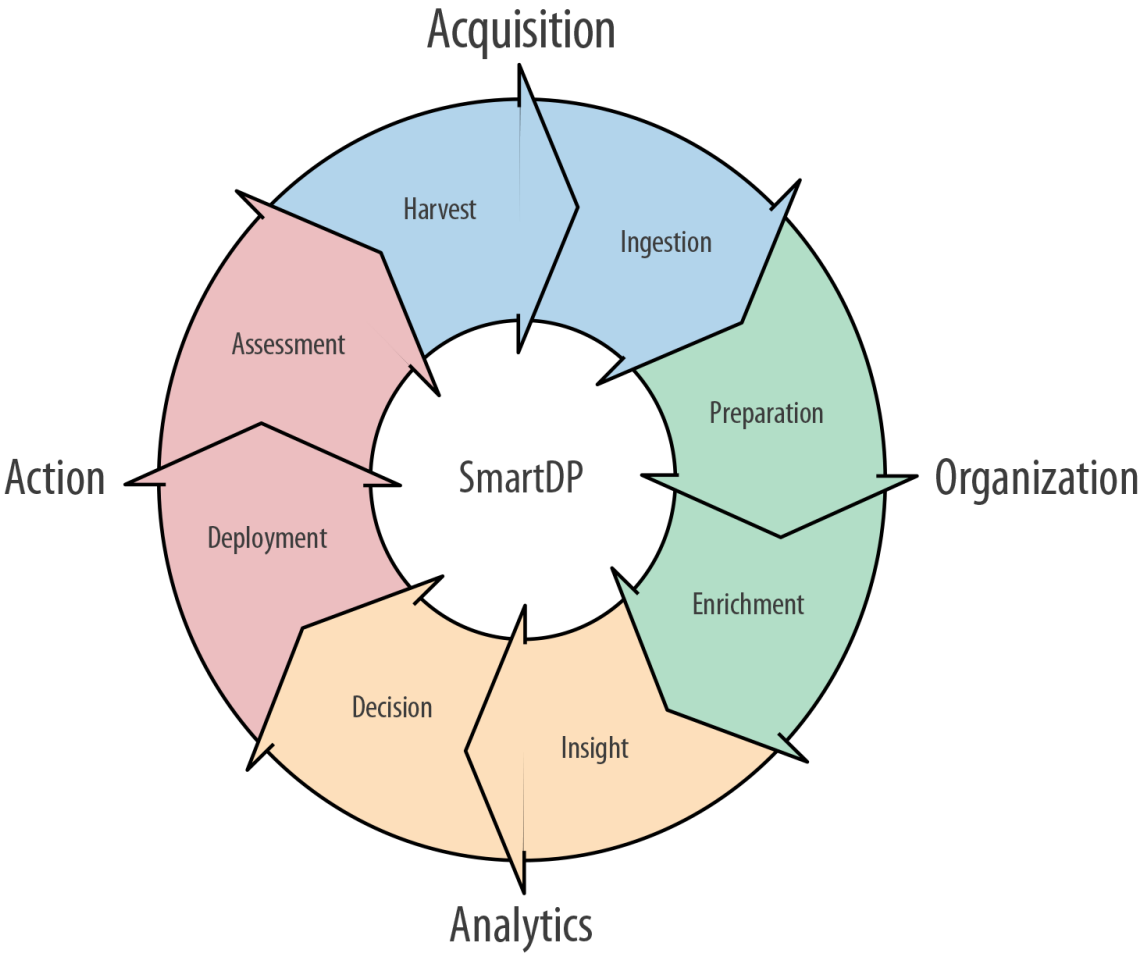
Ingestion Challenges





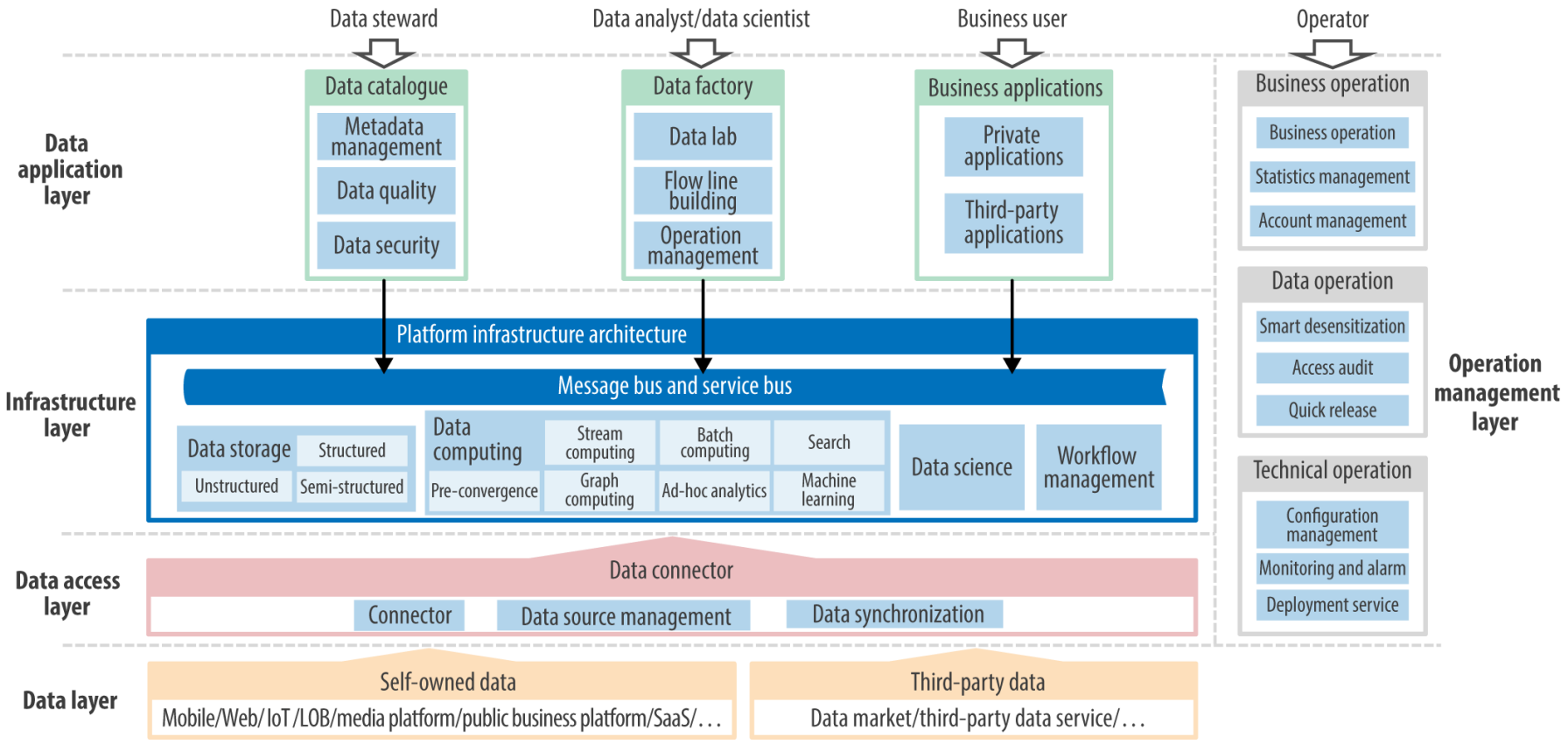
Reference: "The Modern Cloud Data Platform: Rise of the Lakehouse" Report from McKinsey Survey

Smart Data Platform (Processing)



Reference: Wenfeng Xiao - CTO - TalkingData

SmartDP reference architecture

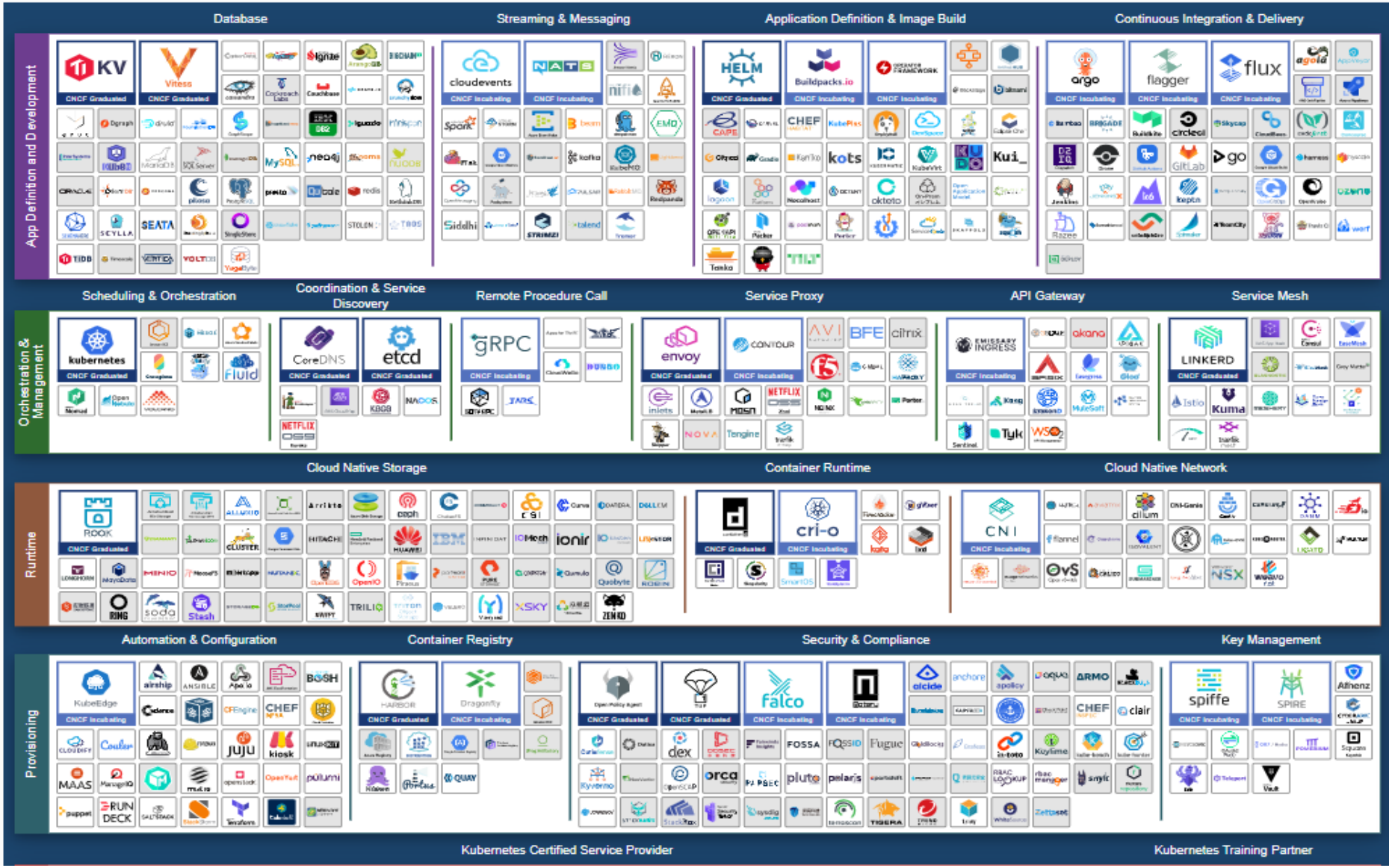


Reference: Wenfeng Xiao - CTO - TalkingData

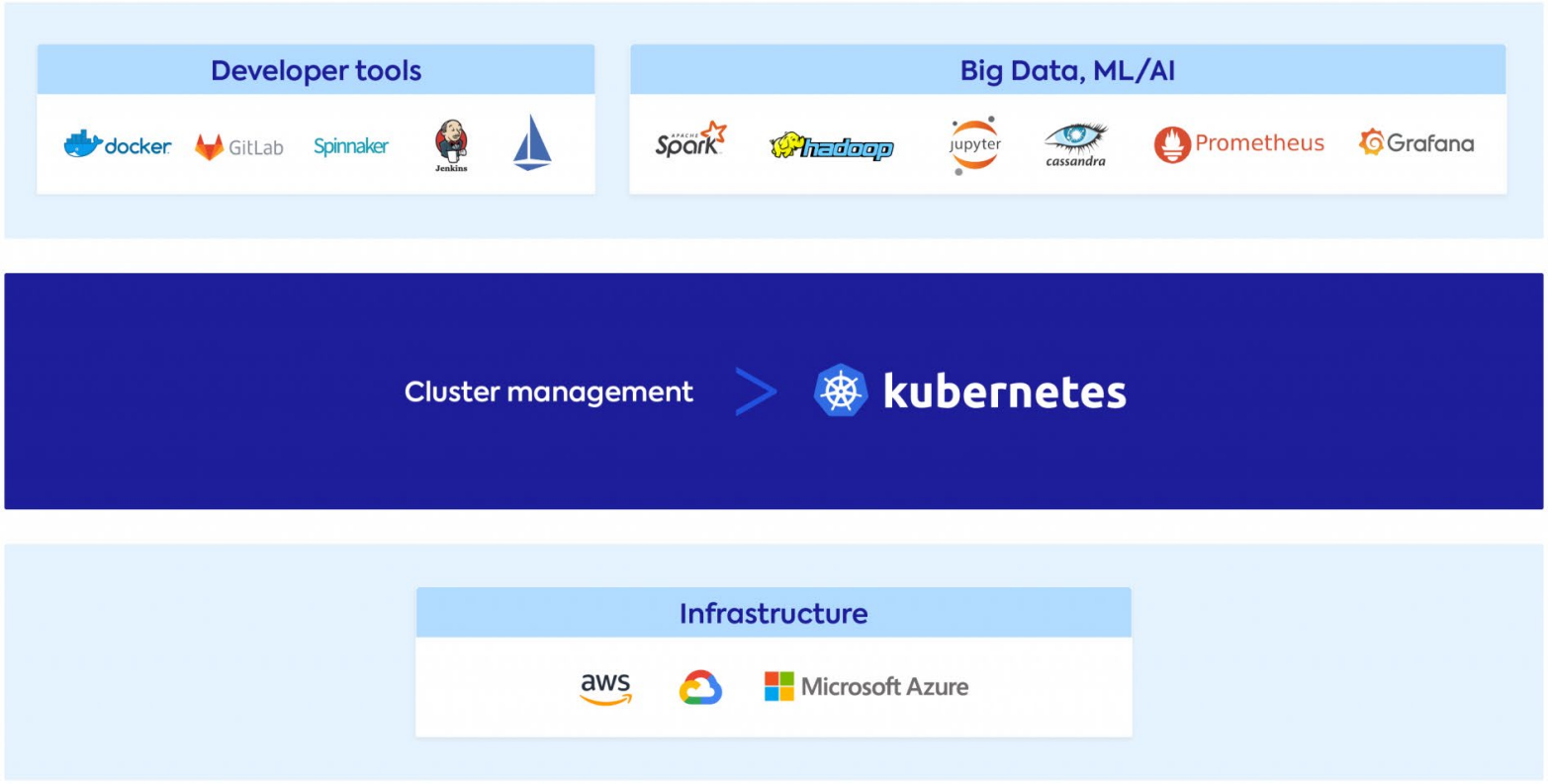
Should you use Kubernetes?

- In the past…
  - Big data workloads needed direct access to storage and network resources
  - Kubernetes schedulers didn't understand the specific needs of big data workloads
  - Support for monitoring in Kubernetes was too limited
  - Integration with big data software wasn't advanced enough to make it easy to operate software like Spark, Kafka, and similar in Kubernetes-managed containers.
- Now, what has changed?
  - Kubernetes's architecture and capabilities have always made it appealing for deploying and operating scalable distributed applications on large-scale infrastructure
  - Kubernetes and its ecosystem emerged to unlock data intensive application loads.

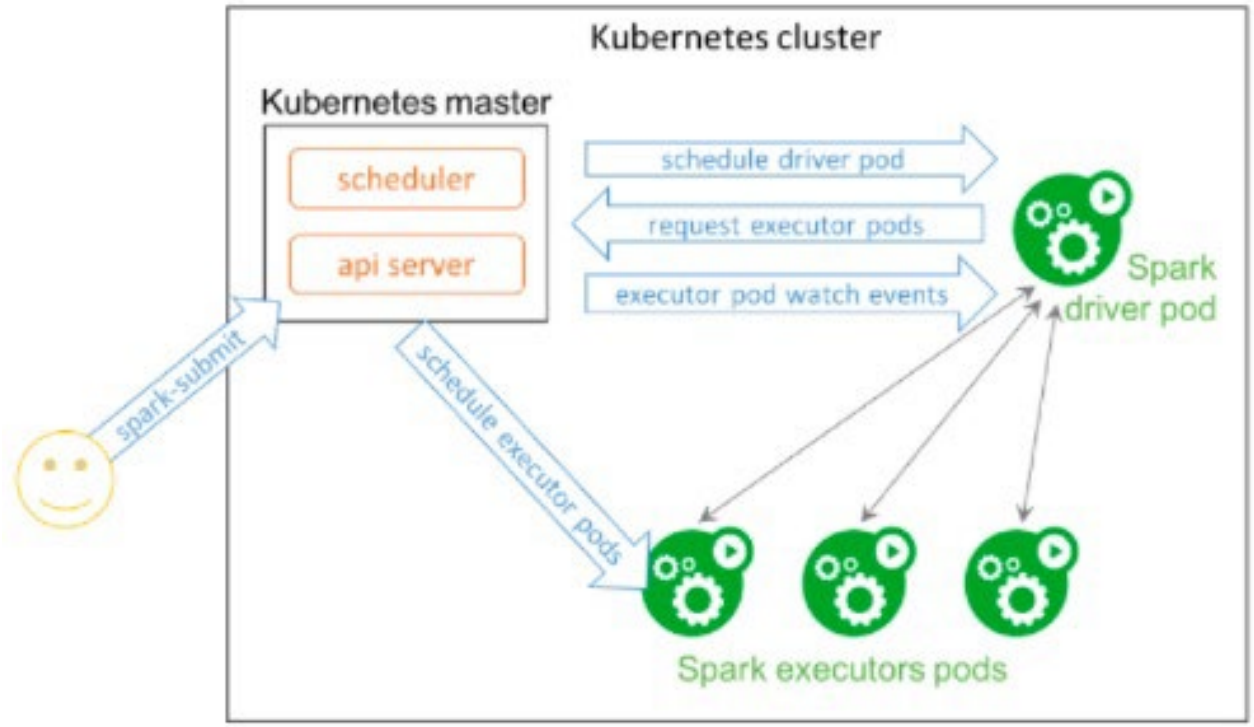
https://landscape.cncf.io/



Kubernetes for Big Data



K8s Cluster

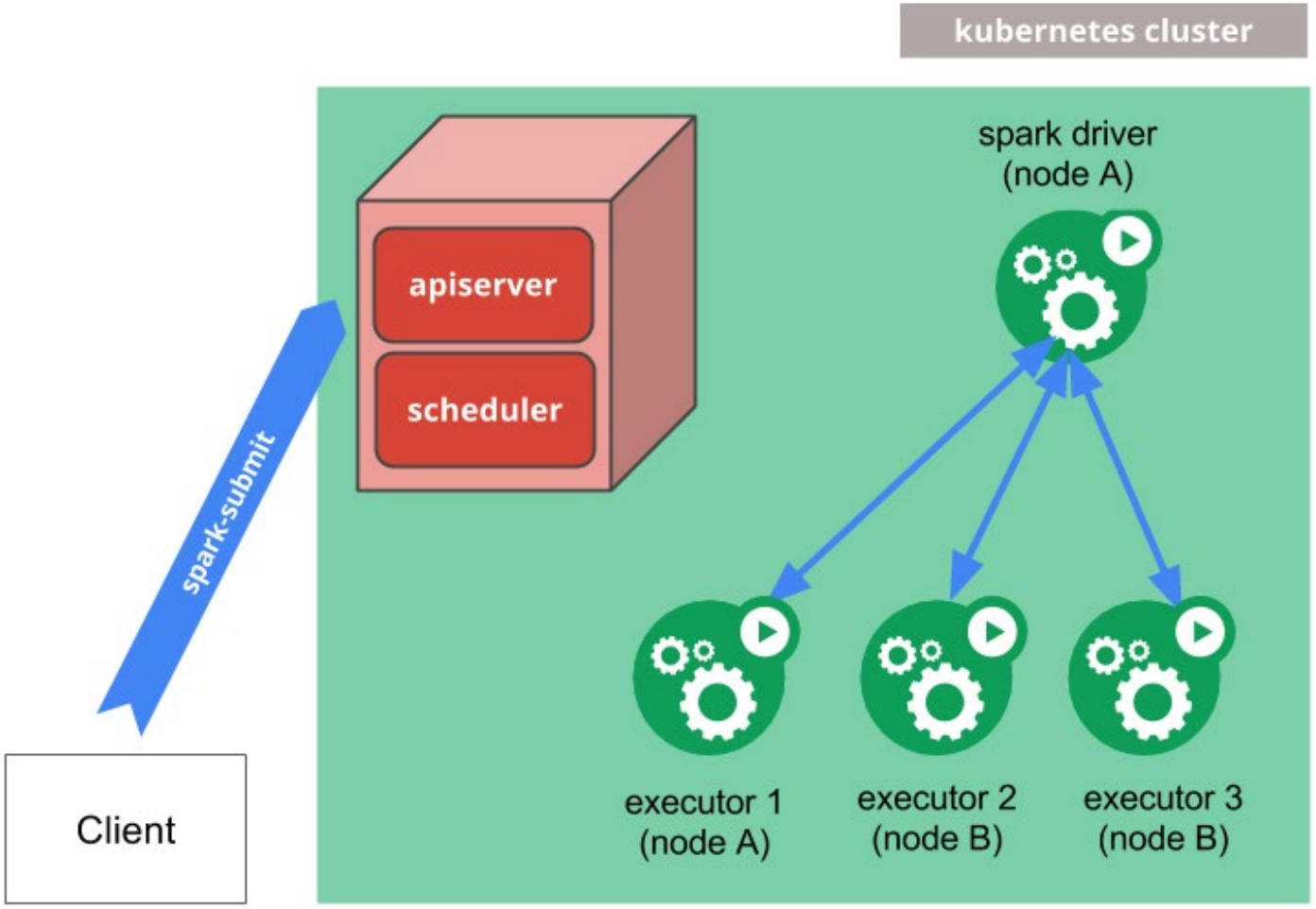


Apache Spark running natively in a Kubernetes cluster

Pre requisites

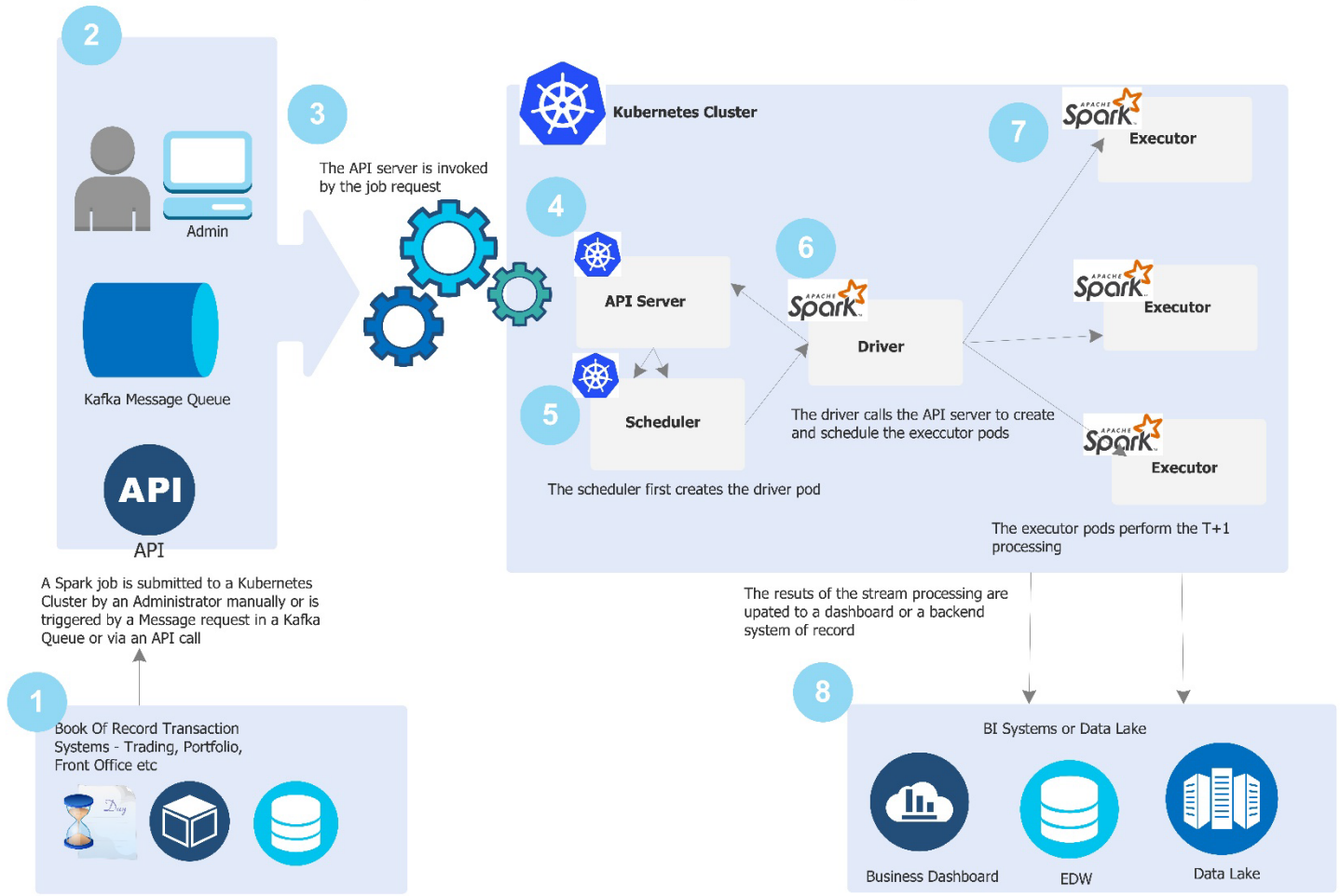
- A runnable distribution of Spark 2.3 or above.
  - A running Kubernetes cluster at version >= 1.6 with access configured to it using kubectl. If you do not already have a working Kubernetes cluster, you may set up a test cluster on your local machine using minikube.
- Using the latest release of minikube with the DNS addon enabled.
  - Be aware that the default minikube configuration is not enough for running Spark applications. We recommend 3 CPUs and 4g of memory to be able to start a simple Spark application with a single executor.
  - You must have appropriate permissions to list, create, edit and delete pods in your cluster. You can verify that you can list these resources by running kubectl auth can-i <list|create|edit|delete> pods.
  - The service account credentials used by the driver pods must be allowed to create pods, services and configmaps.

How it works?

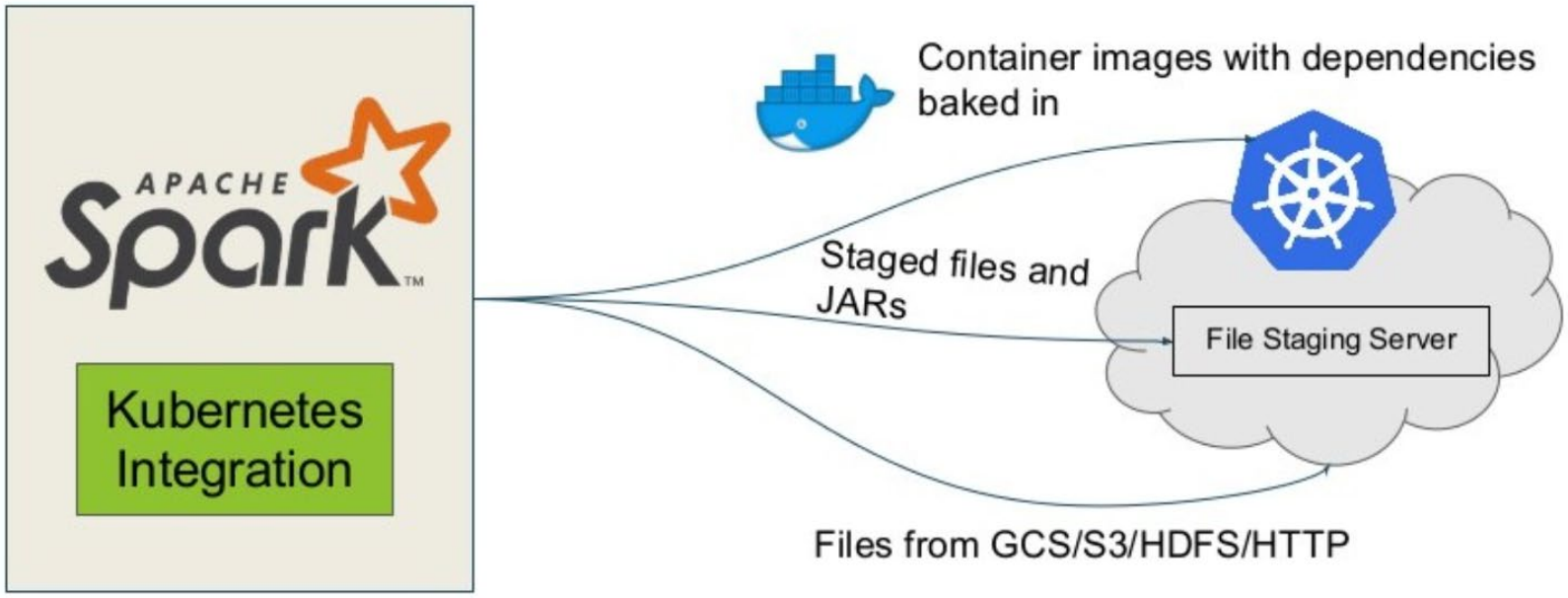


https://spark.apache.org/docs/latest/running-on-kubernetes.html

Spark and Kubernetes Integration



Dependencies



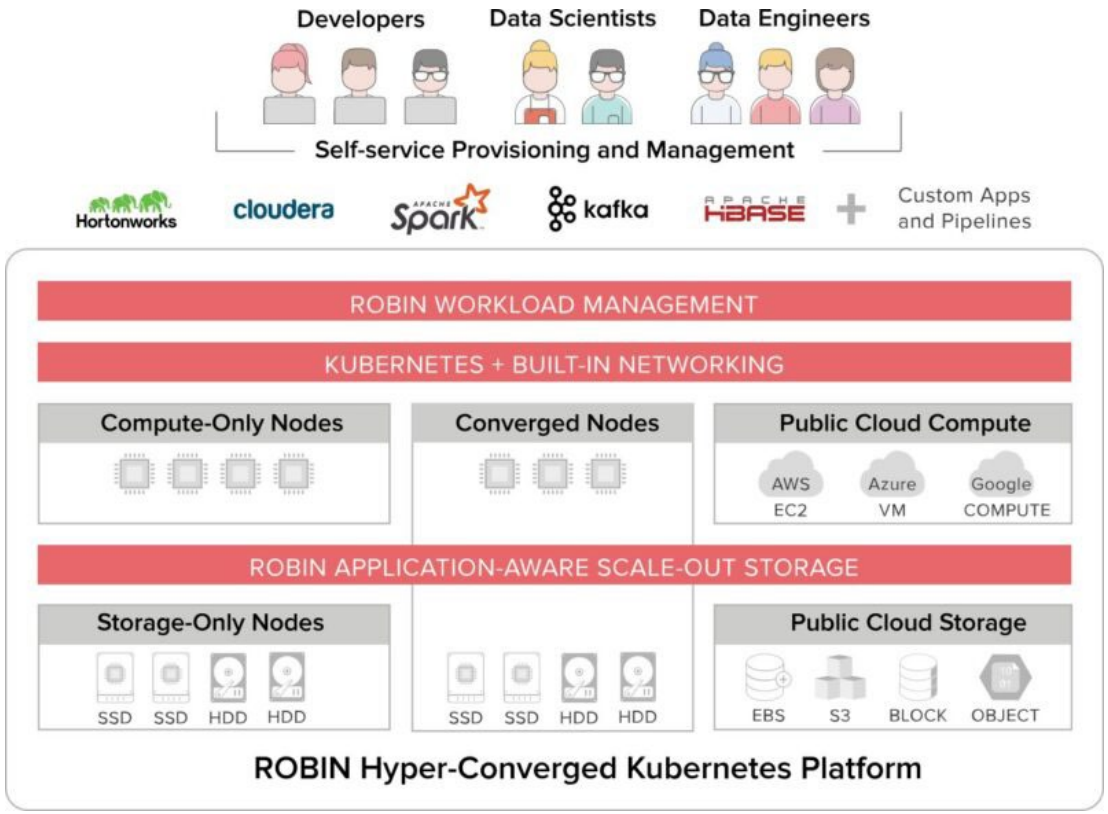
Several ways of running Spark Jobs along with their dependencies on Kubernetes

Data Platform Companies

- Alpine Data Labs
- Anaconda (fka Continuum Analytics)
- Angoss
- Civis Analytics
- Databricks
- Dataiku
- Datawatch
- Domino Data Lab
- IBM
- Indico
- Knime
- Mathworks
- Mortar
- Prevision
- RapidMiner
- Rubikloud
- SAS
- Sense
- TIBCO
- Yhat

Platform Consulting Examples - 1

- Robin (https://www.robin.io/)



Platform Consulting Examples - 2

- Xenonstack (https://www.xenonstack.com/)

Explainable Artificial Intelligence (XAI) Principles and ModelOps Best Practices

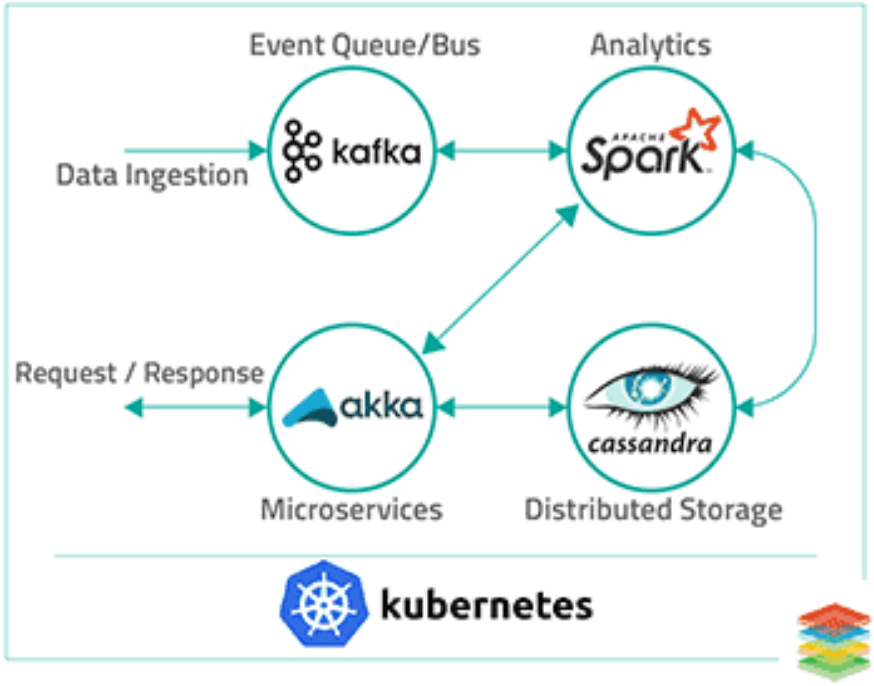
Streamline organization's capabilities for Managing and Deploying Machine Learning Models, and build AI-enabled Cloud Solutions

- Model Visualization Solutions
- Principles for Explainable AI Development
- Enabling Explainability for ML
- Streamline ML Lifecycle
- Self-Service Capabilities
- Continuous Model Solutions

Xenontstack

Data Analytics Stack On Kubernetes

Spark	Micro-batch processing
Kubernetes	Cluster management
Akka	Event processing
Cassandra	Persistence
Kafka	Event transport

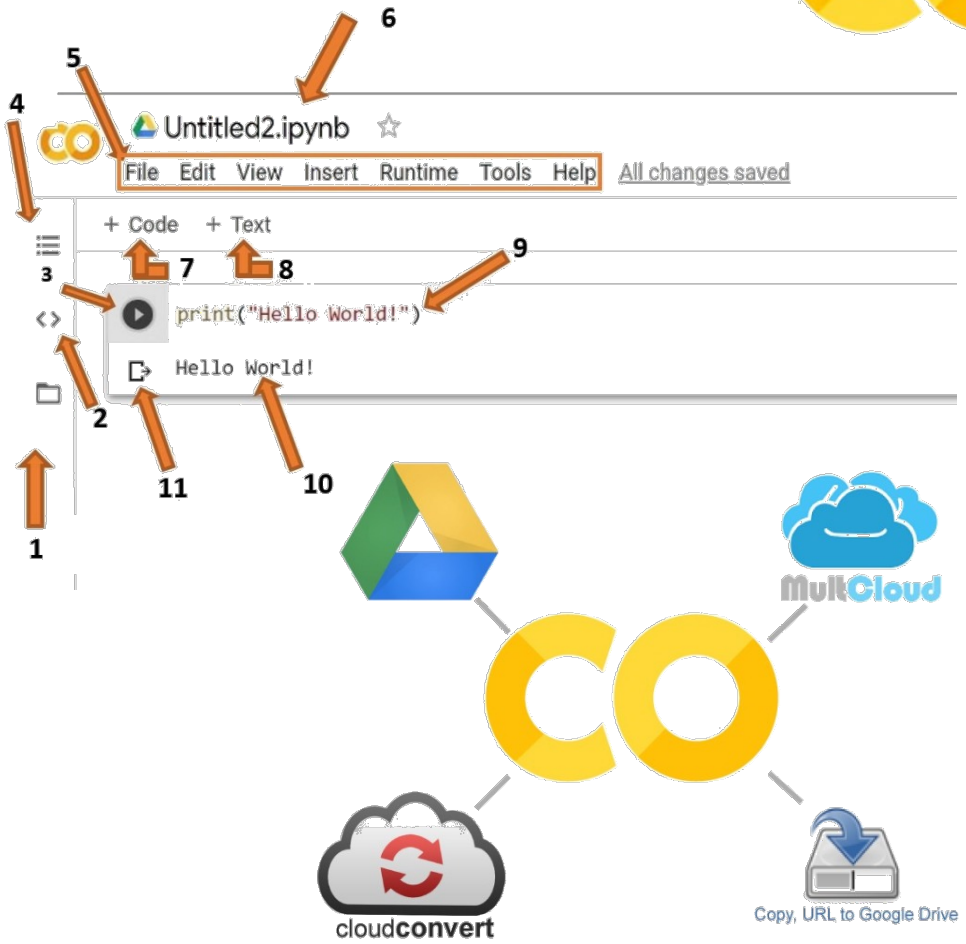


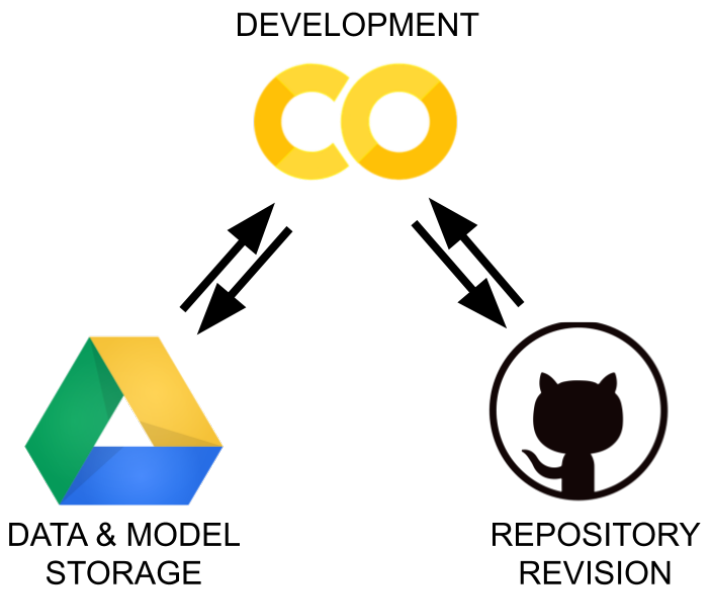
Platform Consulting Examples – 3.

- Cognitree (https://cognitree.com/)

Google Colab

Google colab

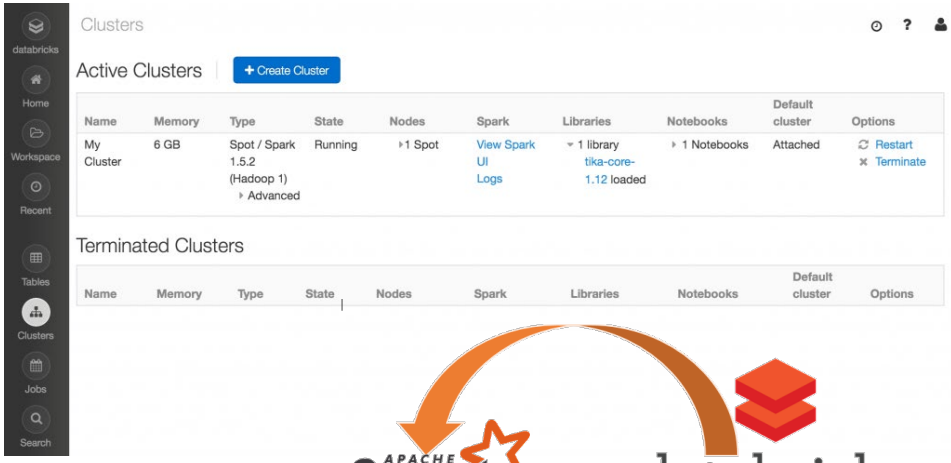




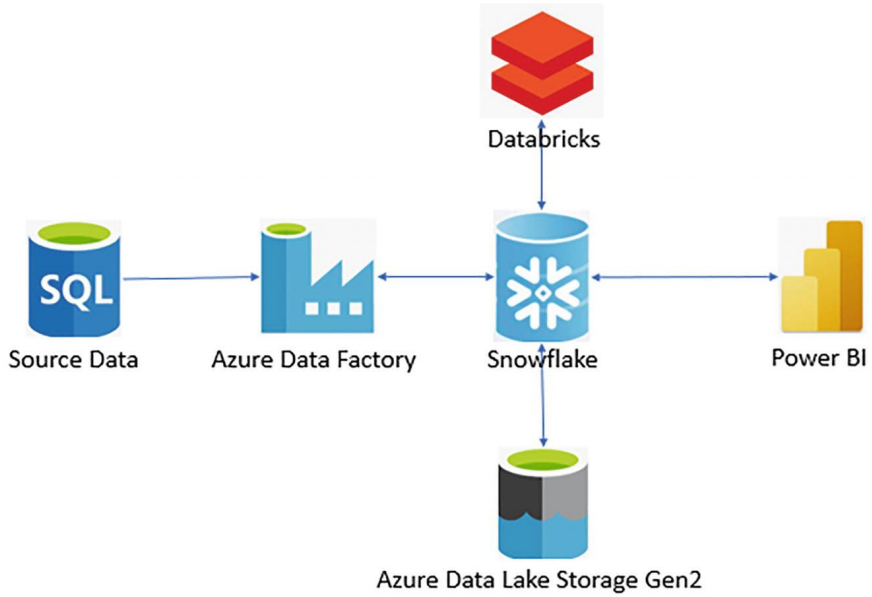
Databricks

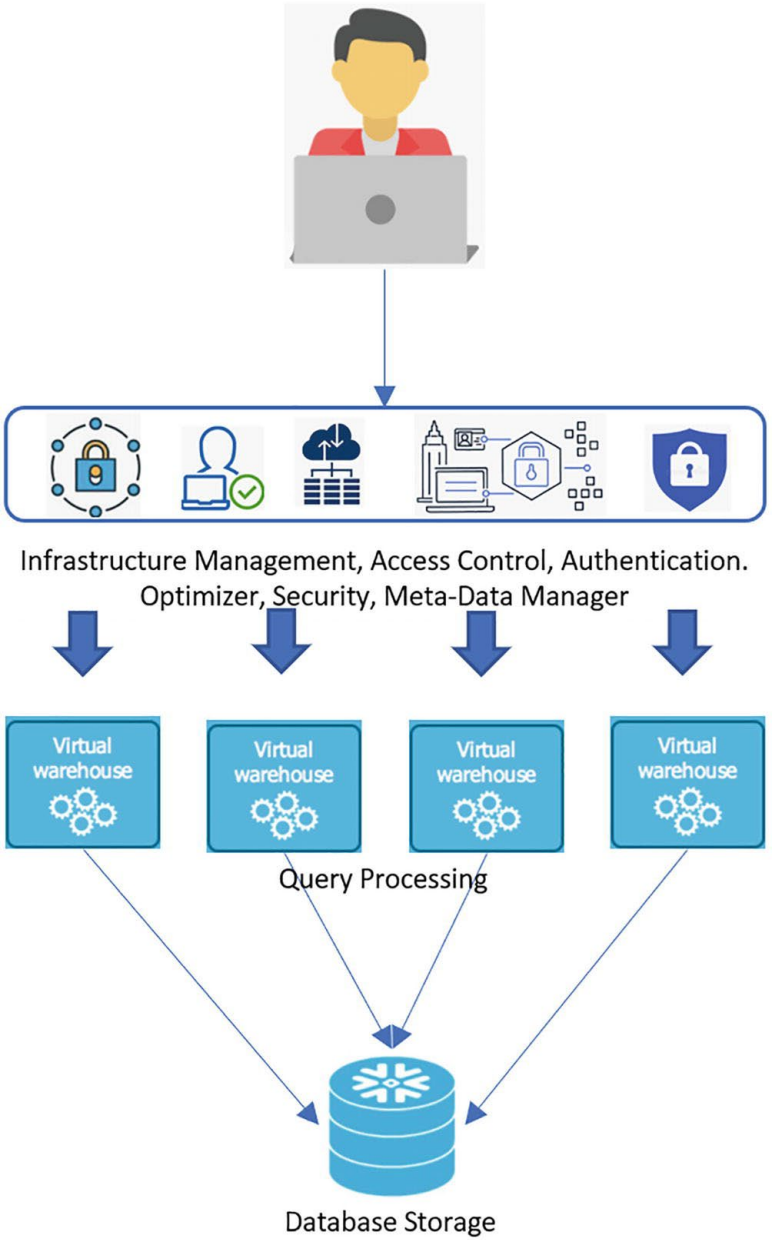
databricks

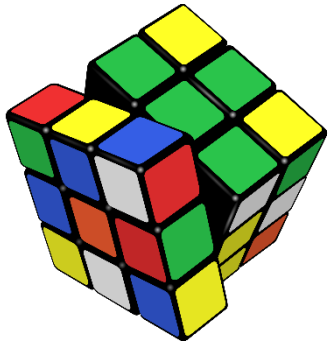
Community Edition

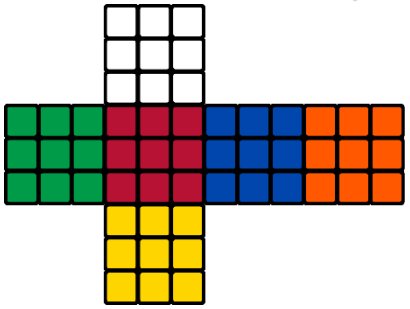


Snowflake









Concluding Remarks

"What is the one sentence summary of how you change the world? Always ***Work hard on something uncomfortably exciting!"***

~Larry Page

Essential Points

- Kubernetes requires users to supply images that can be deployed into containers within pods.
  - The images are built to be run in a container runtime environment that Kubernetes supports.
- Spark Delight – is a Spark UI Replacement
- DropWizard library helps producing detailed metrics and time series calculations
- Kubernetes Security applies to Spark Jobs also.