Intro to Spark & DeltaLake in Synapse

// BRIAN BØNK / MAY 2022

Fellowmind

Spark

Simple.
Fast.
Scalable.
Unified.

Key features

Batch/streaming data

Unify the processing of your data in batches and real-time streaming, using your preferred language: Python, SQL, Scala, Java or R.

SQL analytics

Execute fast, distributed ANSI SQL queries for dashboarding and ad-hoc reporting. Runs faster than most data warehouses.

Data science at scale

Perform Exploratory Data Analysis (EDA) on petabyte-scale data without having to resort to downsampling

Machine learning

Train machine learning algorithms on a laptop and use the same code to scale to fault-tolerant clusters of thousands of machines.

Image source: Apache Spark™ - Unified Engine for large-scale data analytics

Delta Lake

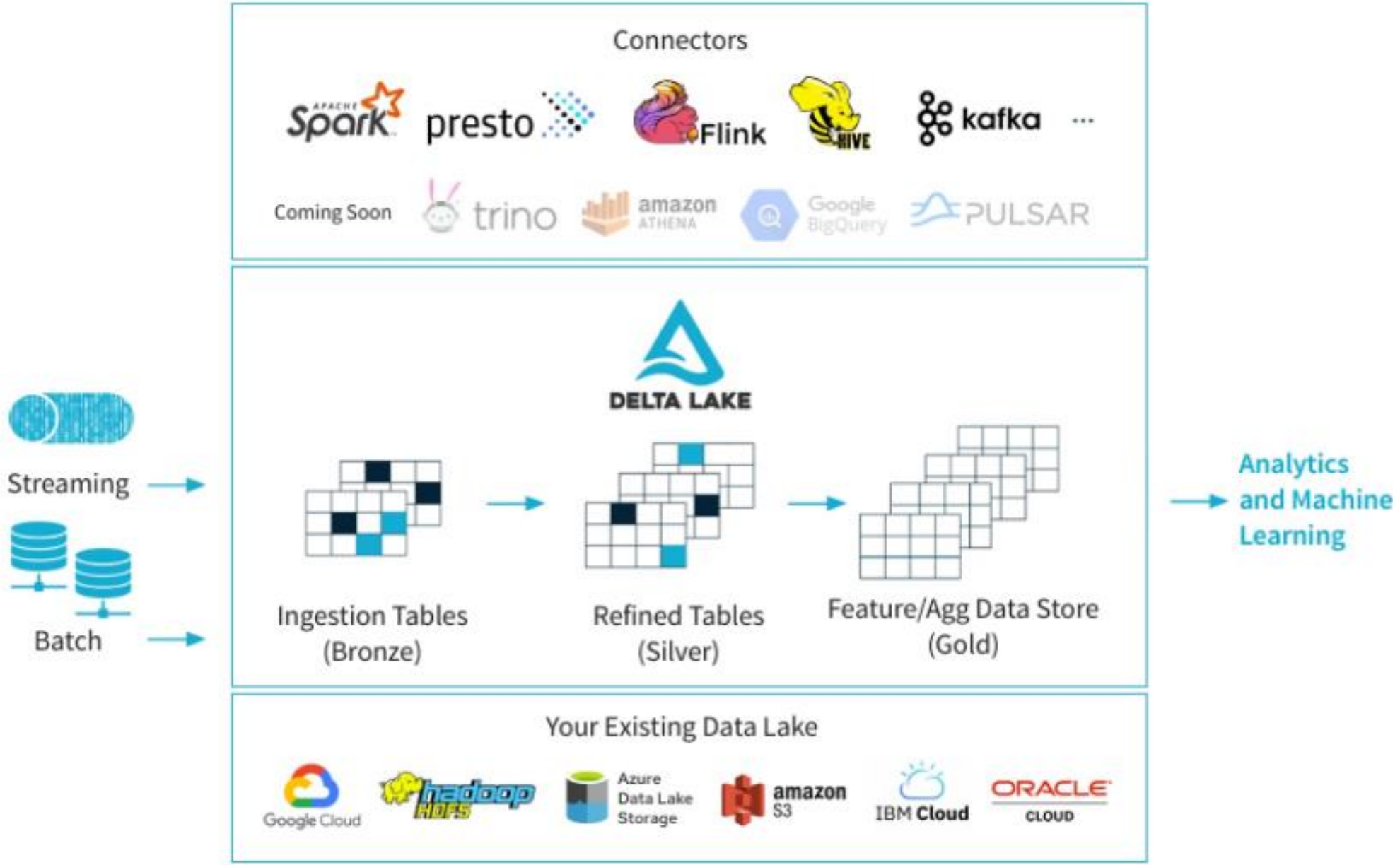
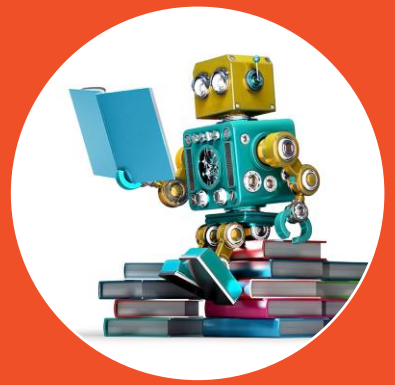


Image source: Delta Lake - Reliable Data Lakes at Scale

Scenarioes



Machine Learning and AI

Data Lake House

Scalable computing

Fill in the blanks

Delta Lake

Image sources: Data Lake vs Warehouse vs Data Lake House | XenonStack

Data Lakehouse

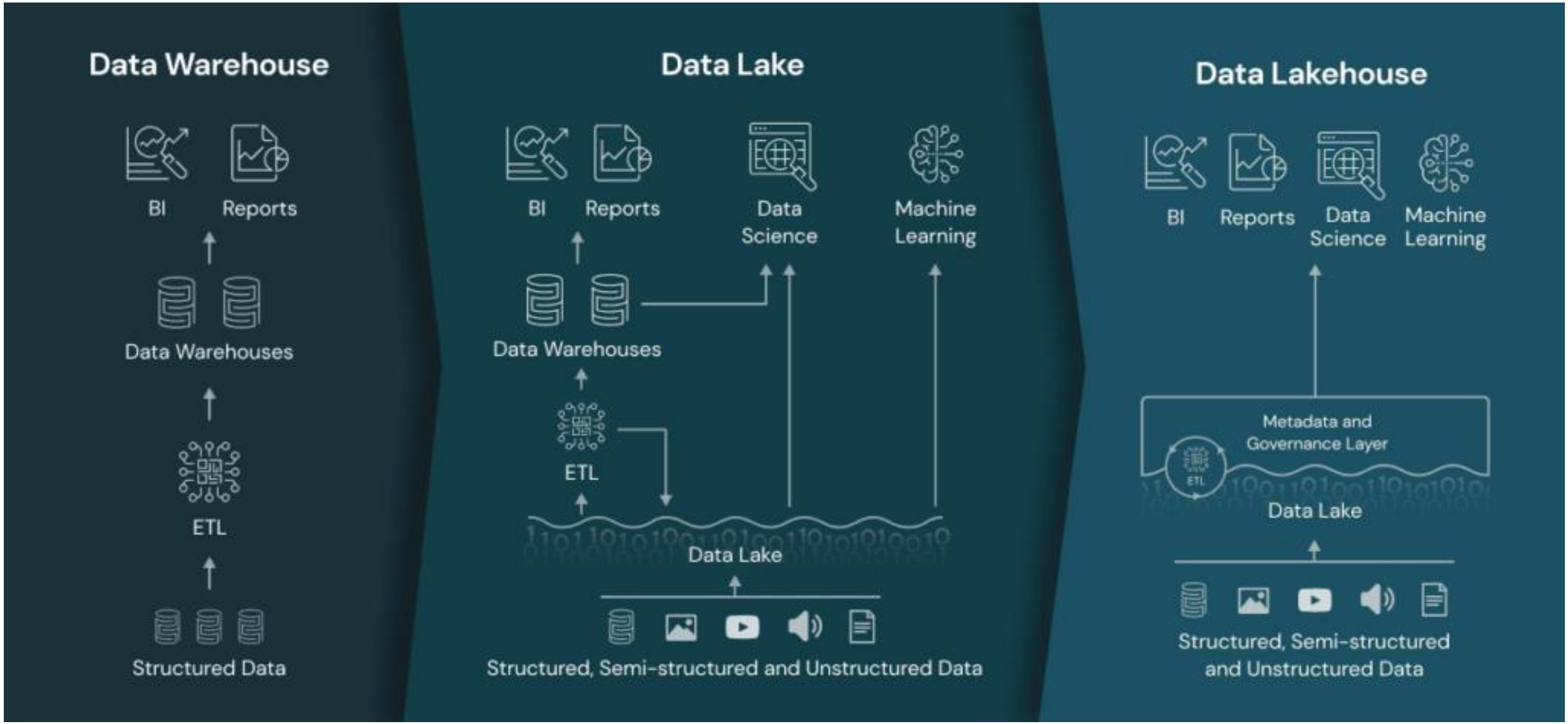
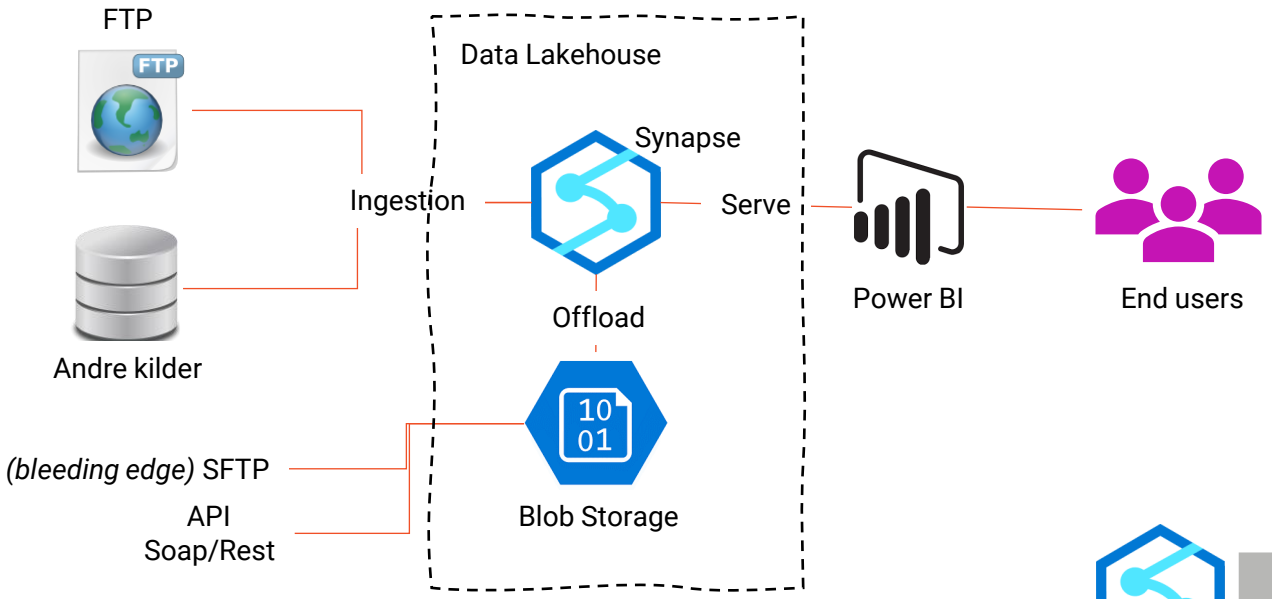
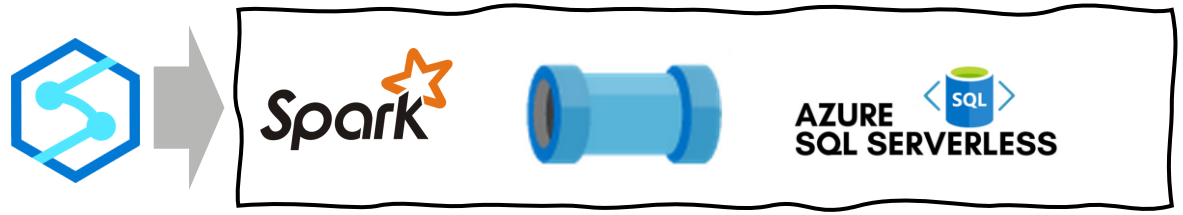
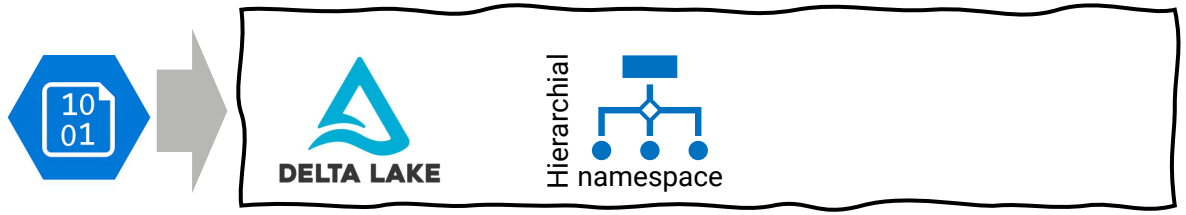


Image sources: What Is a Lakehouse? - The Databricks Blog

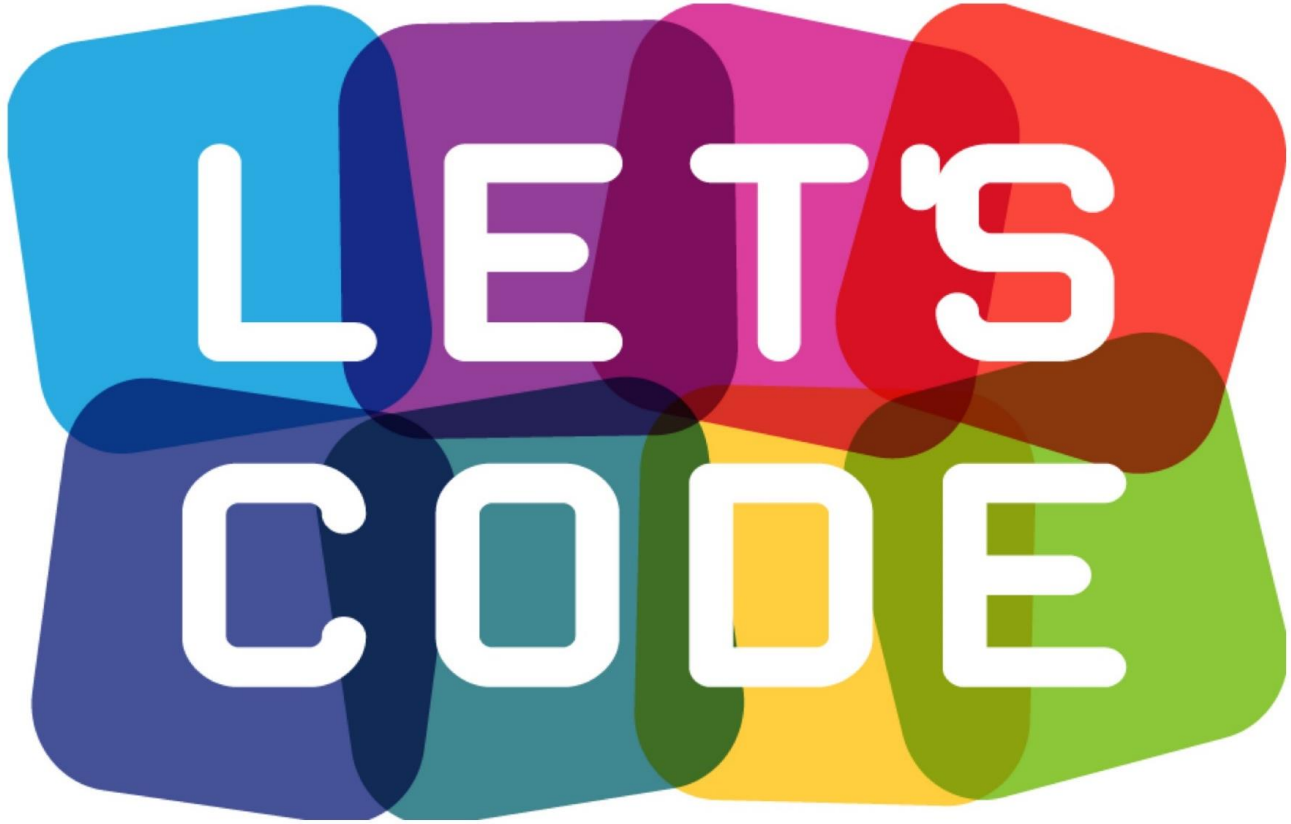
Example of simple architecture



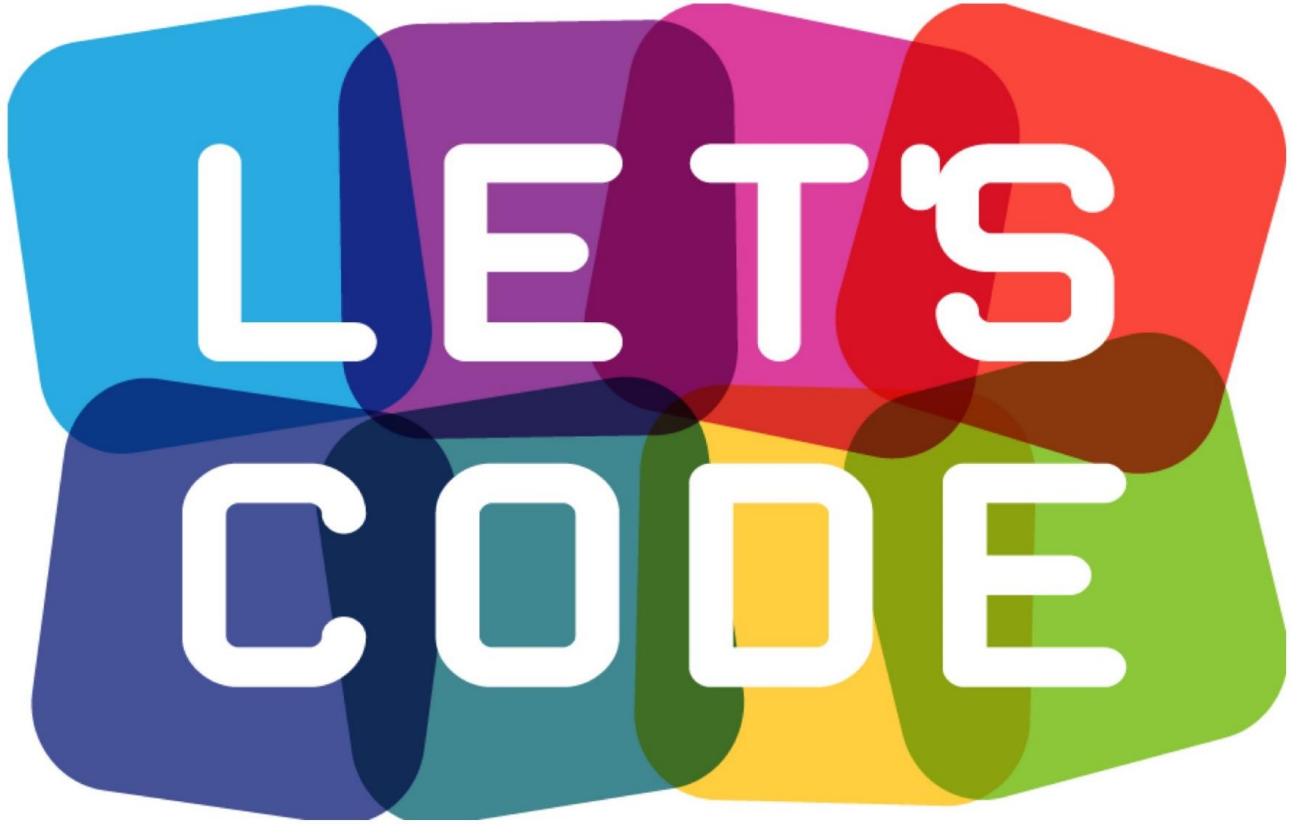




FORMAT = 'DELTA'


LET'S
CODE

This Photo by Unknown Author is licensed under CC BY-SA



All slides and notebooks here: http://fmdk.io/msbip70