

Data Platform

Data Platform is a lambda architecture demo for data processing and analysis



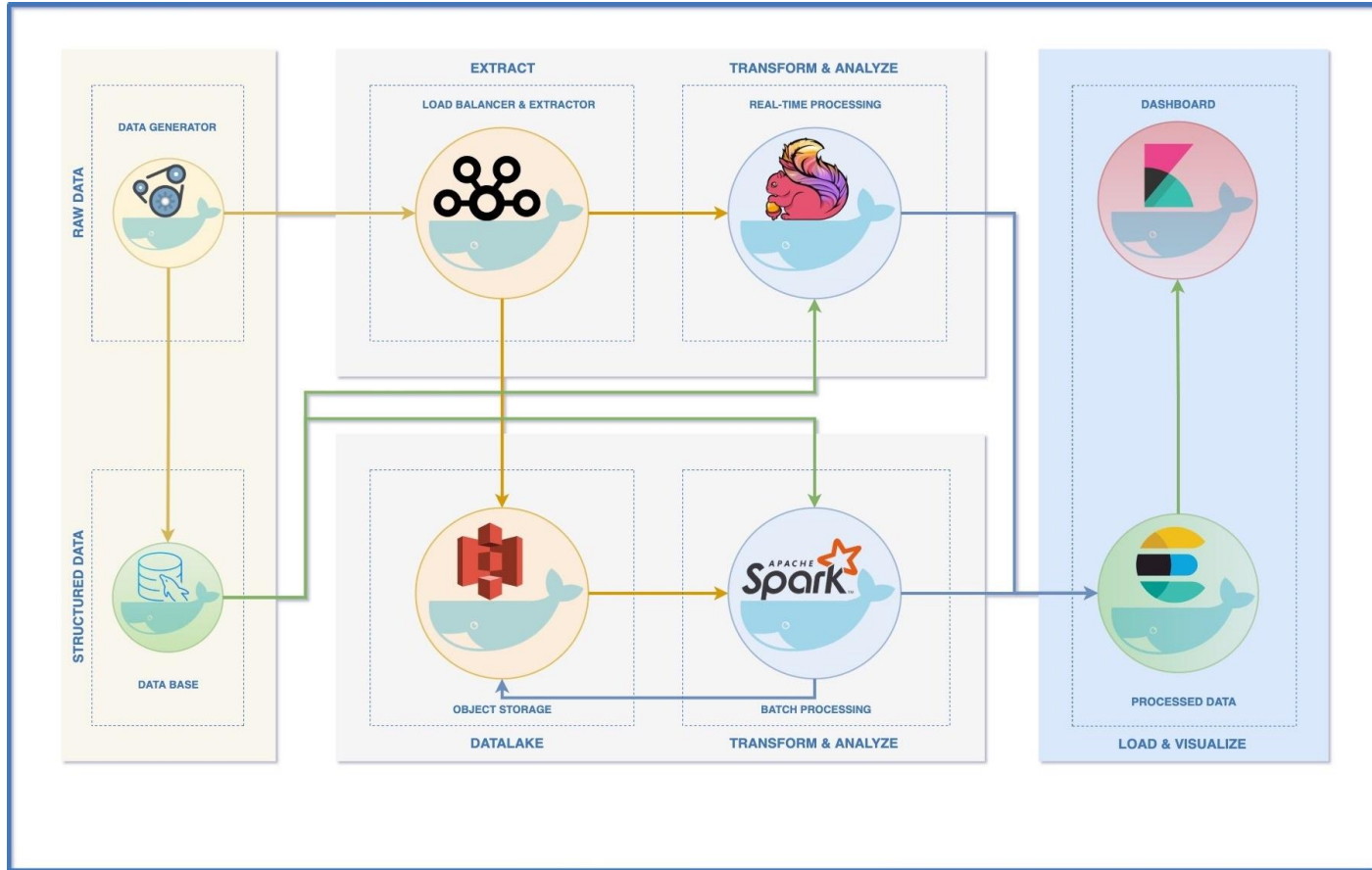


Overview

Data Platform is a lambda architecture demo for data processing and analysis.

The objective of this project is to demonstrate an analytical platform model that is capable of processing data in real time and batch and visualizing the result using different types of technologies..

Architecture



Components



01
Data Generator



02
MySQL



03
Apache Kafka



04
AWS S3



05
Apache Flink



06
Apache Spark



07
Elasticsearch



08
Kibana



Data-Flow

Data Generator:

It is a process developed in Python that reads a soccer game event dataset and sends the events to Kafka every 0.5 seconds. In addition, it reads some datasets (data of the matches, types of events, etc.), and saves them in MySQL.

MySQL:

MySQL stores the data received from the Data Generator to consume it later in the Spark and Flink process.

Apache Kafka:

It receives the events in a topic to consume from the Flink process and stores them in S3 using Kafka connect.



AWS S3 (localstack):

S3 stores the data received from Kafka connect in a bucket to consume it later in batch processes (Spark).

Apache Flink:

The Flink process consumes Kafka topic and MySQL tables (queries every 30 min) to analyze events in real-time and inserts them into Elasticsearch indexes.

Apache Spark (Pending to implement):

Spark reads data stored in S3 and MySQL to analyze in batch mode and stores in Elasticsearch indexes.

Elasticsearch:

Elasticsearch stores the data analyzed in real-time and batch mode.

Kibana:

Visualizes Elasticsearch data in a Dashboard.



 Thank You
For Your Attention



**| Any
Questions?**

© 2015 Pearson Education, Inc. All rights reserved.