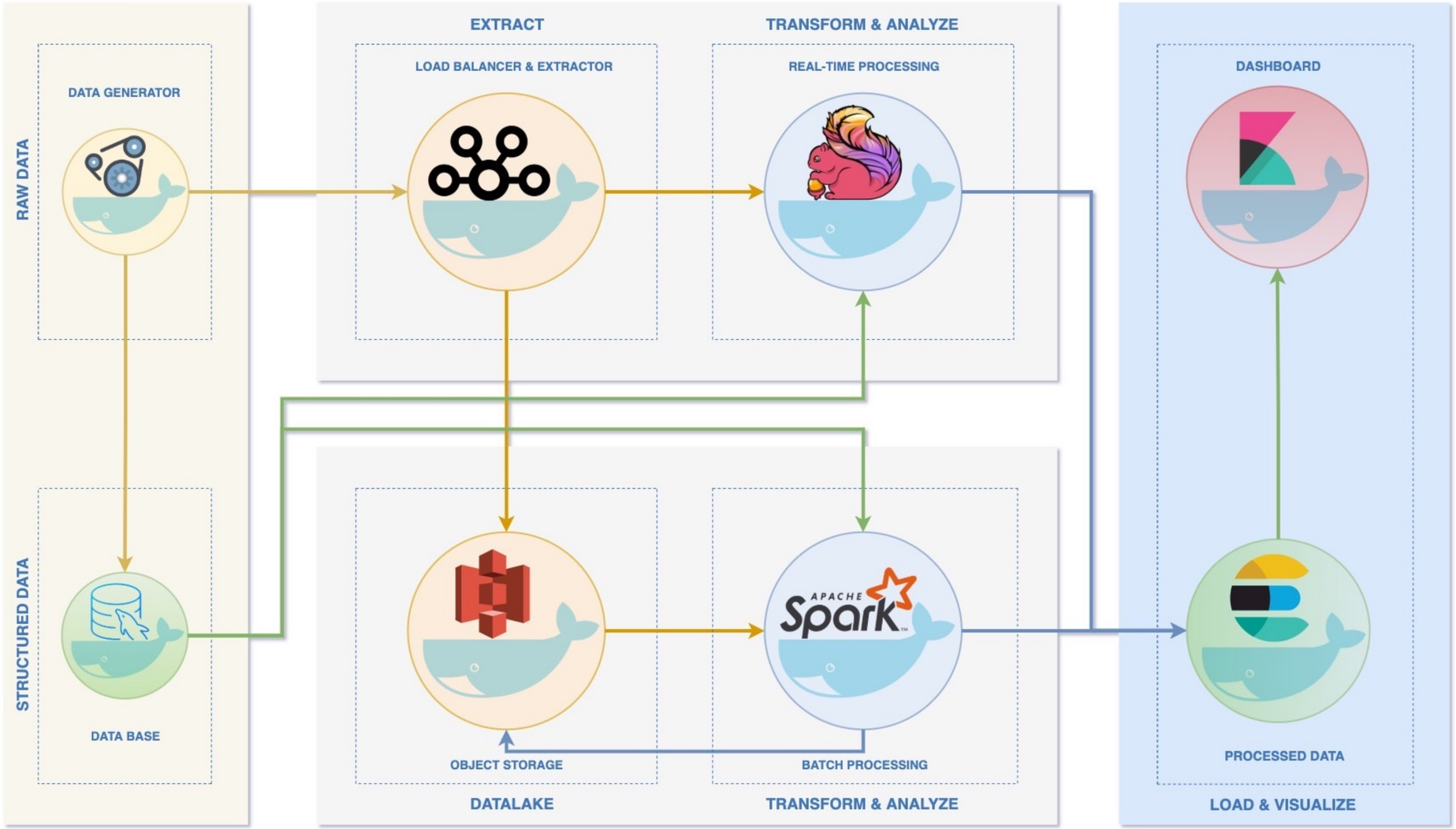

INFODIUM

¡Collect, analyze, repeat!

Infodium (info + podium) 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..

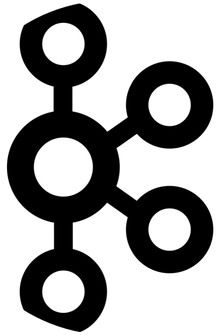
ARCHITECTURE



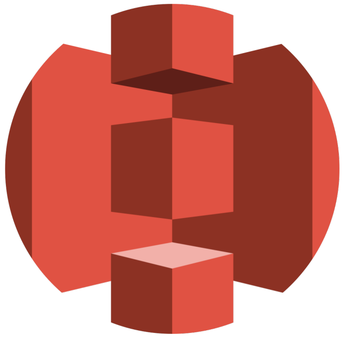
COMPONENTS

01
Data Generator

02
MySql



03
Apache Kafka



04
AWS S3

05
Apache Flink

06
Apache Spark

07
Elasticsearch

08
Kibana

DATA FLOW

01

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 it in MySql.

02

MySql

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

03

Apache Kafka

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

04

AWS S3

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

05

Apache Flink

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

06

Apache Spark (Pending to implement)

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

07

Elasticsearch

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

08

Kibana

Visualizes Elasticsearch data in a Dashboard.