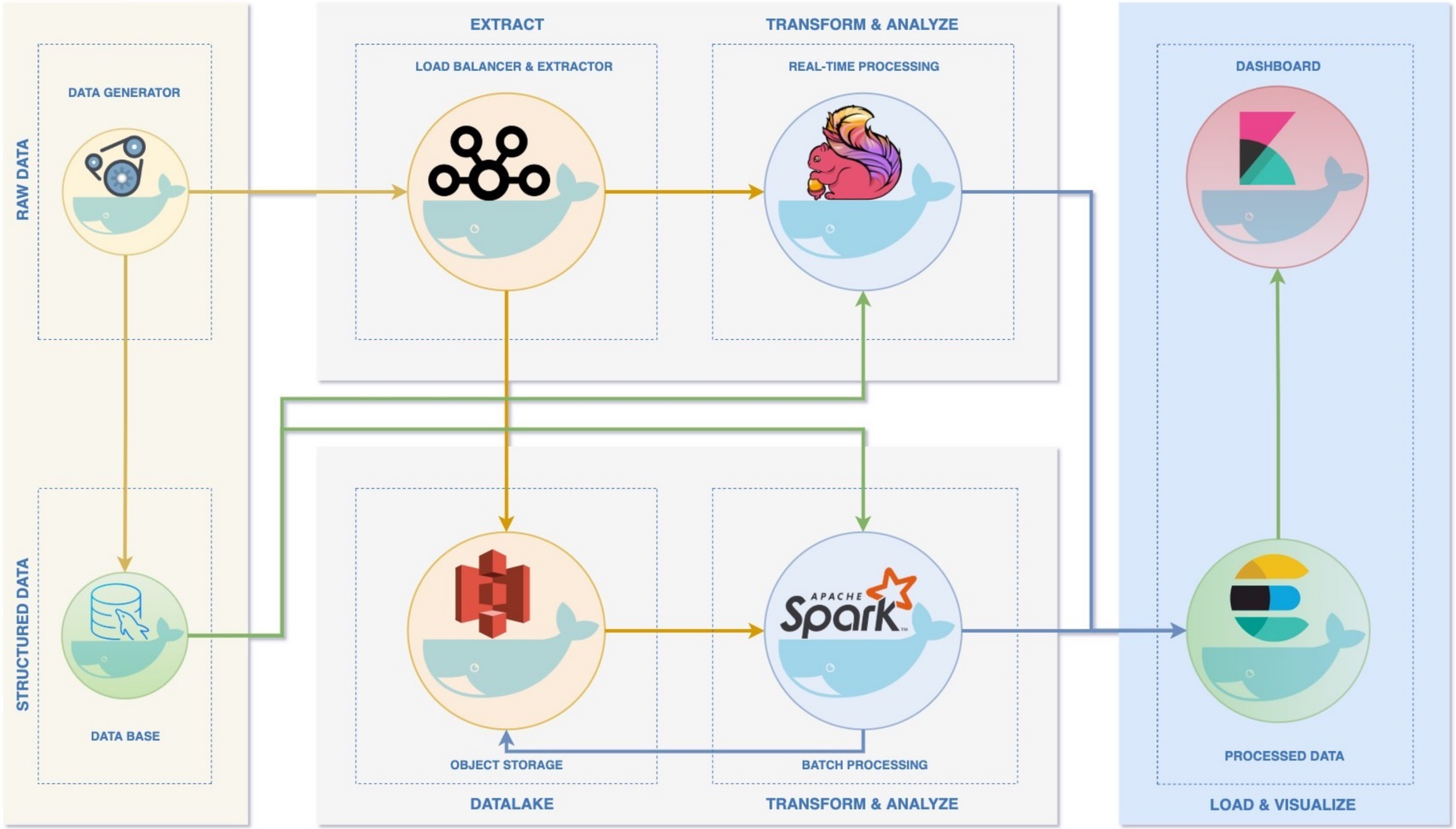

INFODIUM

¡Collect, analyze, repeat!

Infodium (info + podium) es una demo de arquitectura lambda para procesamiento y análisis de datos.

El objetivo de este proyecto es demostrar un modelo de plataforma analítica que es capaz de procesar datos en tiempo real y batch para visualizar el resultado utilizando diferentes tipos de tecnologías.

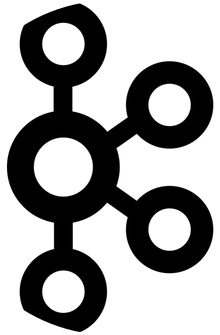
ARQUITECTURA



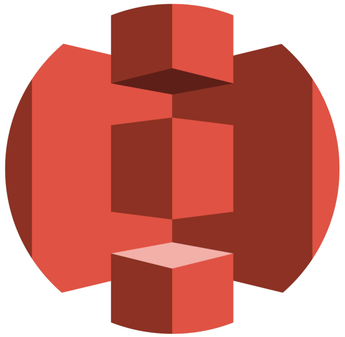
COMPONENTES

01
Data Generator

02
MySql



03
Apache Kafka



04
AWS S3

05
Apache Flink

06
Apache Spark

07
Elasticsearch

08
Kibana

FLUJO DE DATOS

01

Data Generator

Es un proceso desarrollado en Python que lee un dataset de eventos de partidos de fútbol y envía los eventos cada 0.5 segundos a Kafka. Además, lee unos datasets (datos de los partidos, tipos de eventos, etc.), y lo guarda en MySql.

02

MySql

Almacena los datos recibidos del Data Generator para poder consumir posteriormente en el proceso de Spark y Flink.

03

Apache Kafka

Recibe los eventos en un topic para poder consumir desde el proceso de Flink y guarda los eventos en S3 utilizando Kafka connect.

04

AWS S3

Almacena los datos recibidos de Kafka connect en un bucket para consumirlos en procesos batch (Spark).

05

Apache Flink

El proceso Flink consume de topic de Kafka y MySql (consulta cada 30 min) para analizar los eventos en tiempo real e inserta en índices de Elasticsearch.

06

Apache Spark (Pendiente de implementar)

Spark lee los datos guardados en S3 y MySql para analizar en modo batch y guardar en índices de Elasticsearch.

07

Elasticsearch

Almacena los datos analizados en tiempo real y modo batch.

08

Kibana

Visualiza los datos de Elasticsearch en un Dashboard.