Escuela de Datos Vivos

Curso Data Engineer: Creando un pipeline de datos

Módulo E - Clase 11

Agenda

- Arquitectura
- Exámen final
- ML con Pablo Casas

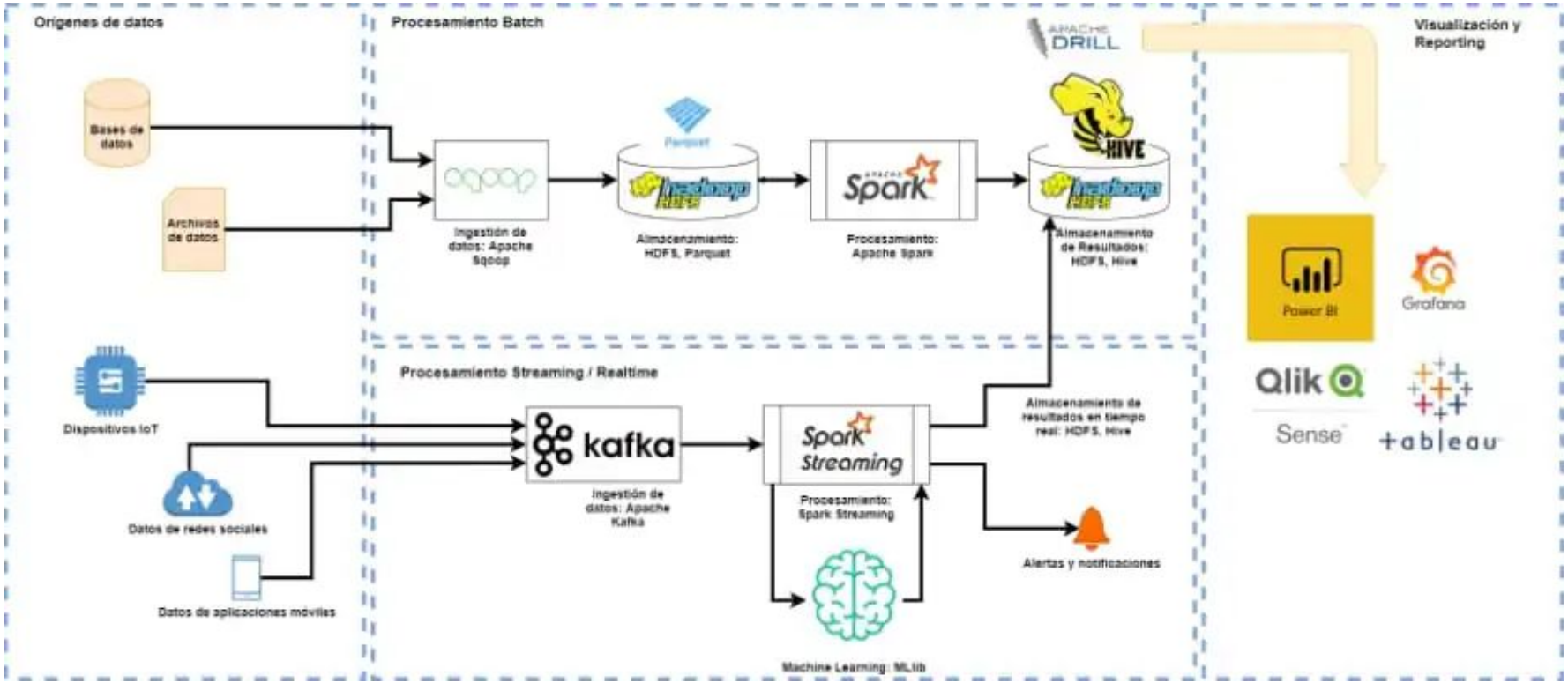
Arquitectura de Datos

Arquitectura Lambda

La Arquitectura Lambda, surgió en el año 2012 y se atribuye a Nathan Marz. La definió en base a su experiencia en sistemas de tratamiento de datos distribuidos durante su etapa como empleado en las empresas Backtype y Twitter.

Su objetivo era tener un **sistema robusto tolerante a fallos, tanto humanos como de hardware, que fuera linealmente escalable y que permitiese realizar escrituras y lecturas con baja latencia**.

Arquitectura Lambda



Arquitectura Kappa

El término Arquitectura Kappa, fue introducido en 2014 por Jay Kreps en su artículo Questioning the Lambda Architecture.

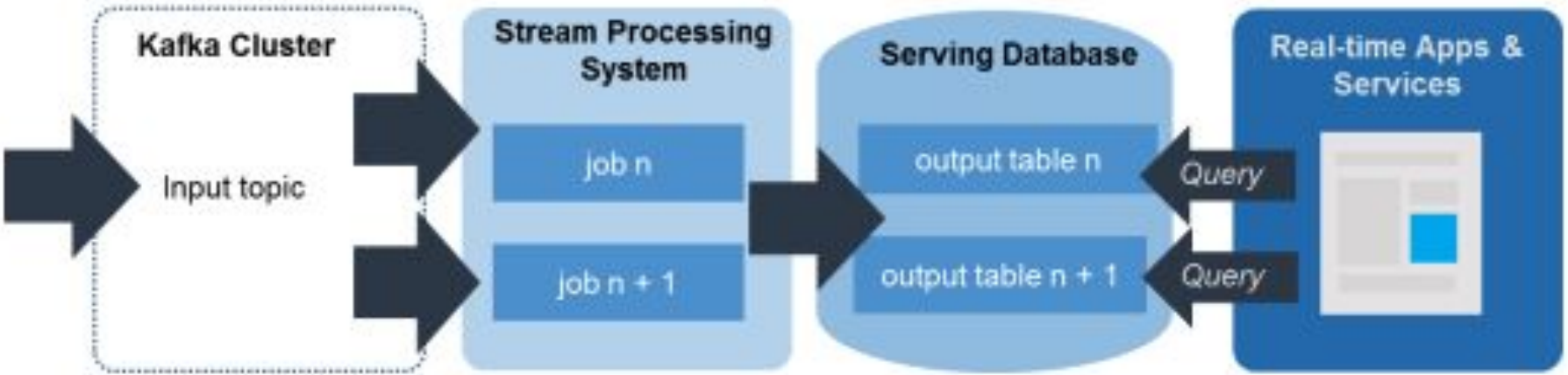
En él señala los **posibles puntos "débiles" de la Arquitectura Lambda** y cómo solucionarlos mediante una evolución. Su propuesta consiste en **eliminar la capa batch** dejando solamente la capa de streaming.

Esta capa, a diferencia de la de tipo batch, no tiene un comienzo ni un fin desde un punto de vista temporal y **está continuamente procesando nuevos datos** a medida que van llegando.

Como un **proceso batch se puede entender como un stream acotado**, podríamos decir que el procesamiento batch es un subconjunto del procesamiento en streaming.

Esta evolución consiste en una simplificación de la Arquitectura Lambda, en la que se elimina la capa batch y **todo el procesamiento se realiza en una sola capa denominada de tiempo real o Real-time Layer**, dando soporte a procesamientos tanto batch como en tiempo real.

Arquitectura Lambda



Práctica Arquitectura

Ejercicio de arquitectura

Te acaban de contratar en una empresa de la industria minera como Data Engineer/Data Architect para delinear su arquitectura y sugerir qué herramientas deberían utilizar para ingestar la data, procesar la información, almacenarla en un datawarehouse, orquestar y realizar Dashboards que ayuden a la toma de decisiones basadas en datos.

Luego de realizar algunas reuniones con el team de analitica de la empresa pudimos relevar:

Sistema ERP: SAP con una base de datos Oracle

Sistema de Producción: App desarrollada "in house" con una base de datos Postgres.

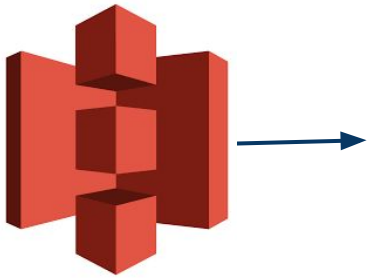
Fuentes externas: un proveedor que realiza algunas mediciones de la calidad de las rocas le deja todos sus análisis en un bucket de AWS S3 con archivos Avro.

Mediciones en tiempo real: Utilizan +100 sensores de mediciones de vibración por toda la mina para detectar movimiento del suelo y se podrían utilizar para predecir posibles derrumbes.

Desarrollar una arquitectura, que sea escalable, robusta, que sea orquestada automáticamente, que contemple seguridad, calidad, linaje del dato, que sea utilizada para procesar tanto información batch como información en tiempo real.

Exámen final

Ejercicio 1



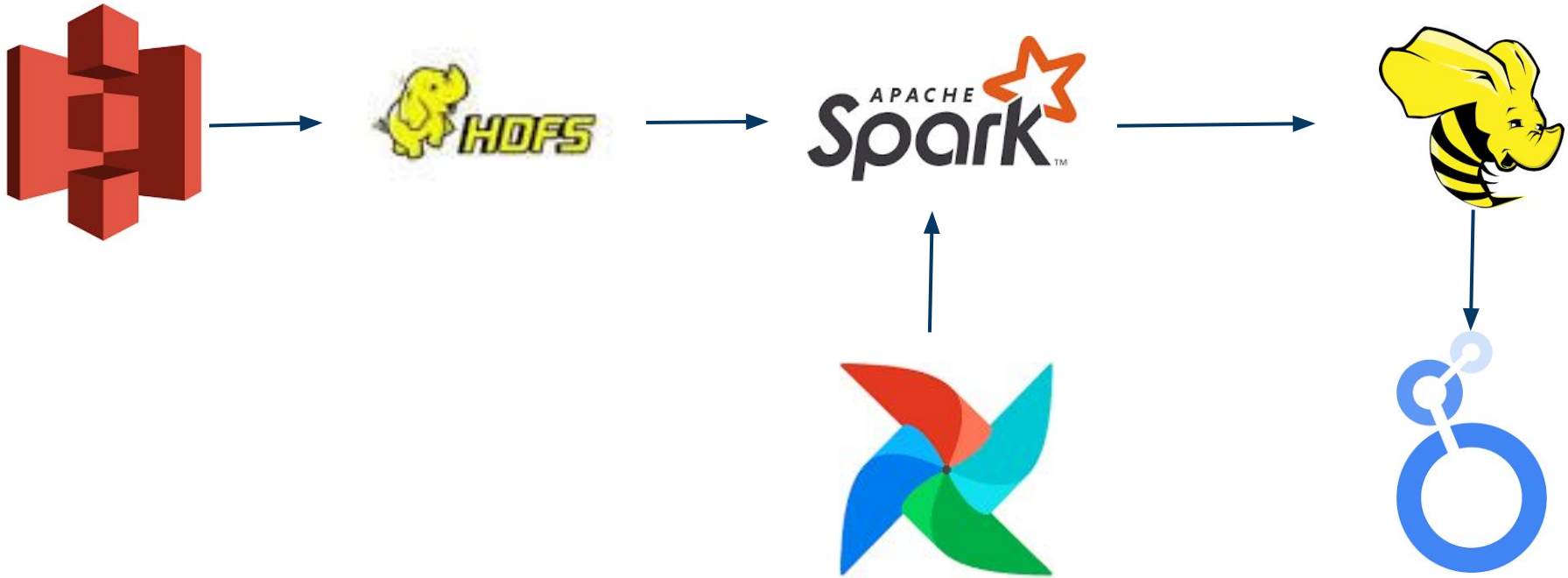
Ejercicio 1

Ejercicio 1

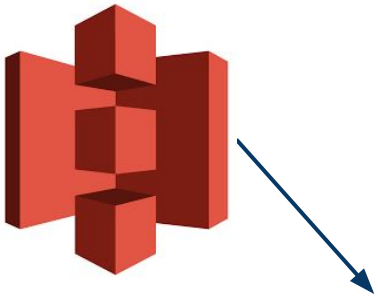
Ejercicio 1

Ejercicio 1

Ejercicio 1



Ejercicio 2



Ejercicio 2

Ejercicio 2

Ejercicio 2

Ejercicio 2

Ejercicio 2

Ejercicio 2

Ejercicios

- Ponerse en ror de DE
- Elaborar conclusiones
- Elaborar recomendaciones de negocio
- Proponer diferentes arquitecturas (si aplica)

Preguntas ?

Gracias