Engenharia de Dados

03/02/2021

Programa de Aceleração Banco Inter

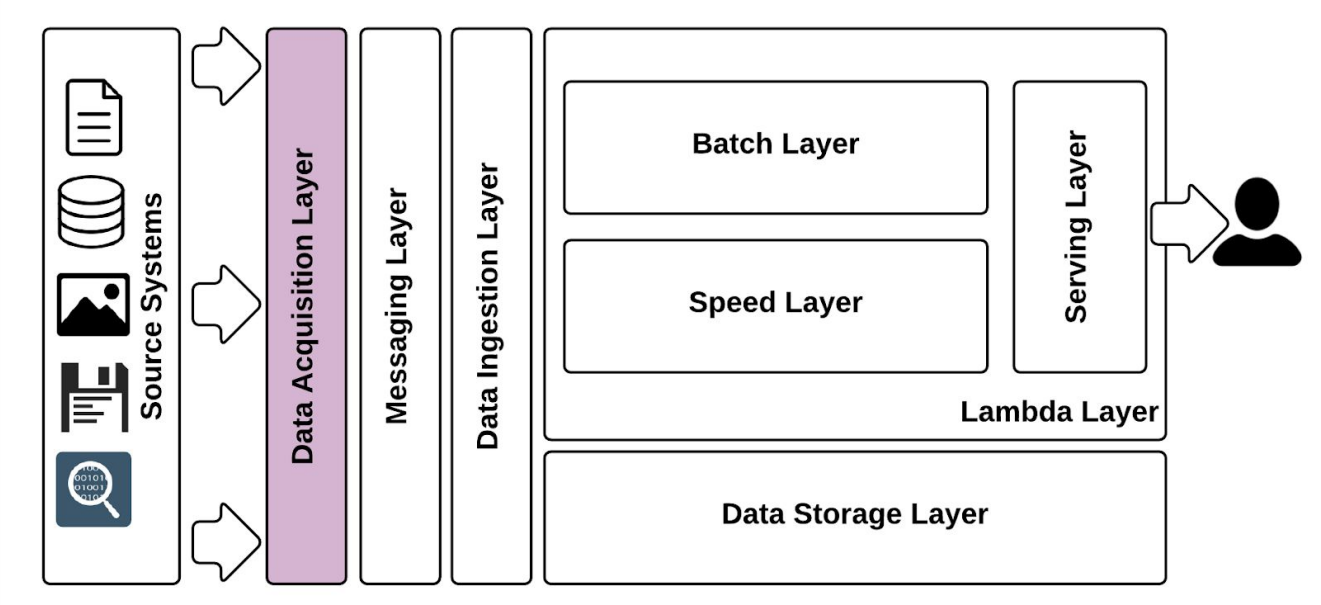
Ingestão de Dados por Batch e Streaming

datasprints

Agenda

1. Camadas de um Data Lake;
2. Ingestão de Dados;
3. Tecnologias para Ingestão de Dados;
4. Ingestão de Dados por Batch;
5. Tecnologias e práticas para trabalhar com Batch;
6. Ingestão de Dados por Streaming;
7. Tecnologias e práticas para trabalhar com Streaming;

Camadas de um Data Lake



datasprints

Ingestão de Dados no Data Lake

- Adquirir dados de diversas fontes;
- Armazenamento centralizado;
- Desafios:
    - Fontes variadas;
    - Tecnologias variadas;
        - Arquivos, Rest API, Banco Relacional e não relacional....
    - Tipos de tecnologias para ingerir os dados;
    - Batch vs Streaming.

Ingestão de Dados no Data Lake

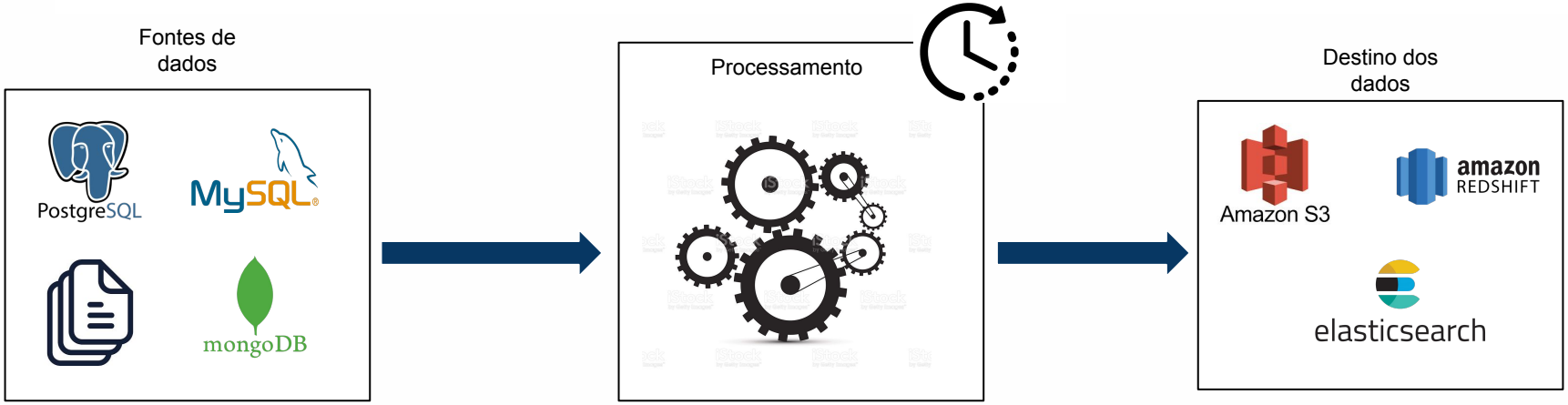
Como funciona?

Quais tecnologias utilizar?

datasprints

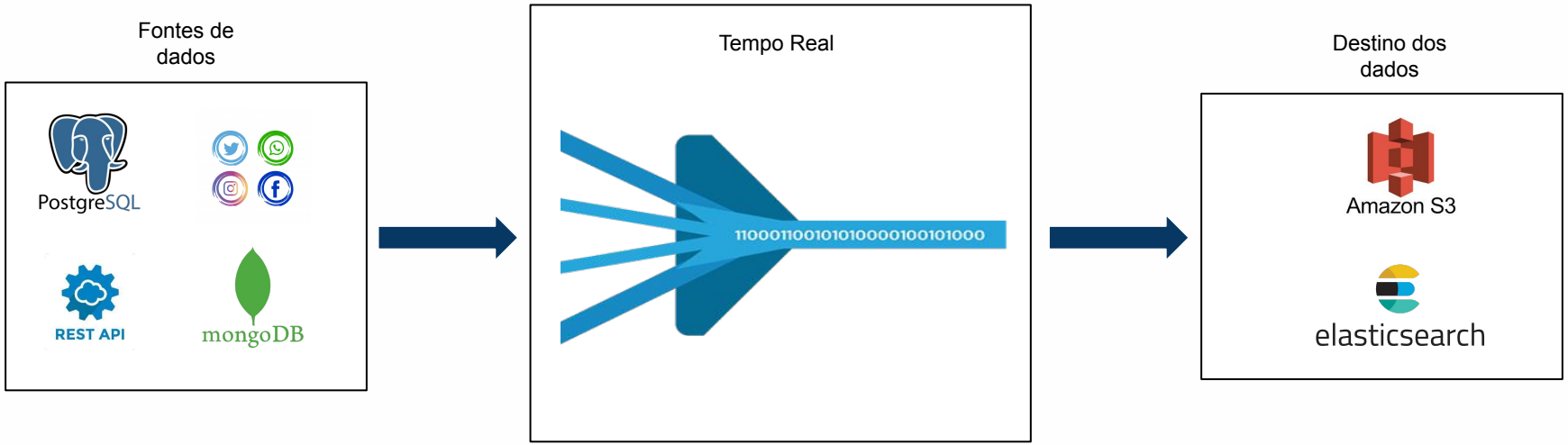
Ingestão de Dados por Batch

Como funciona?



Ingestão de Dados por Streaming

Como funciona?



Ingestão por Batch

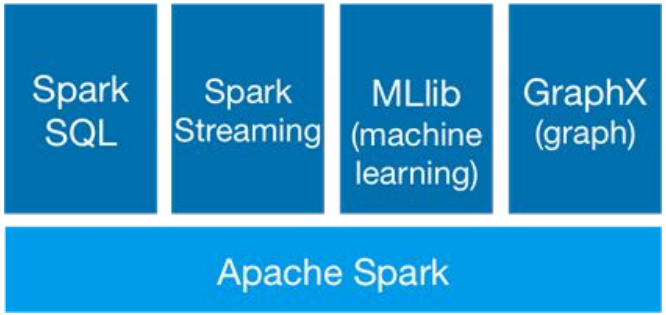
Python e SQL

- Criação de processos como ETL ou ELT;
- Agendamento de Jobs por outras ferramentas:
  - Airflow, Cron...
- Utilização de lib's como:
  - SQLAlchemy;
  - Pandas;
  - NumPy;
  - Outras...

Apache Spark

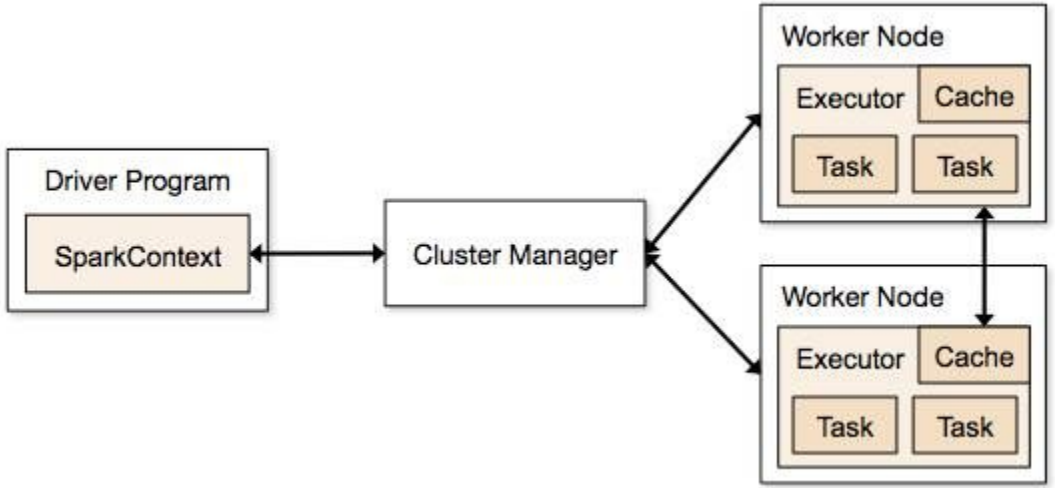
- Framework para processamento de dados por MapReduce;
- Suporta grande quantidade de dados;
- O core é escrito em Java;

- Conjunto de bibliotecas:



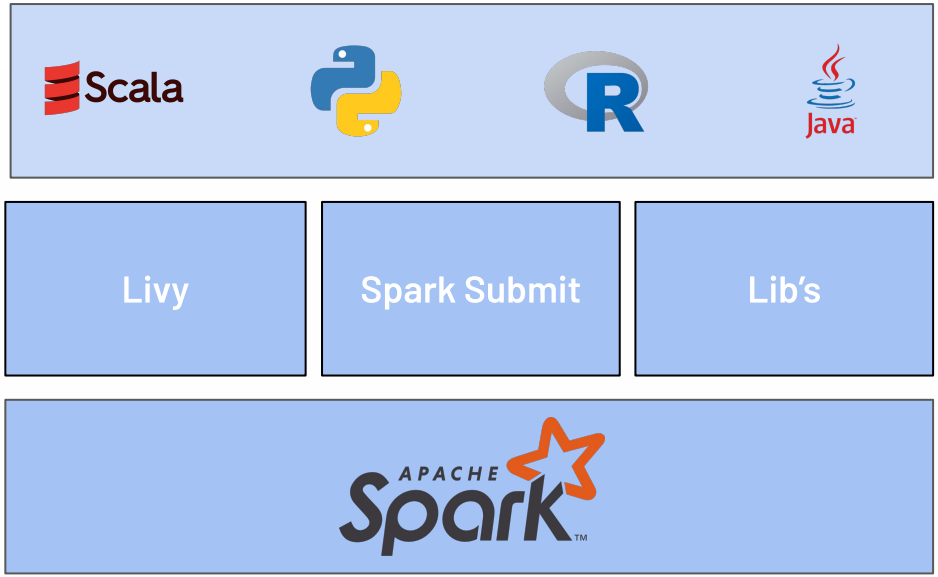
https://www.devmedia.com.br/introducao-ao-apache-spark/34178

Apache Spark - Arquitetura



https://www.devmedia.com.br/introducao-ao-apache-spark/34178

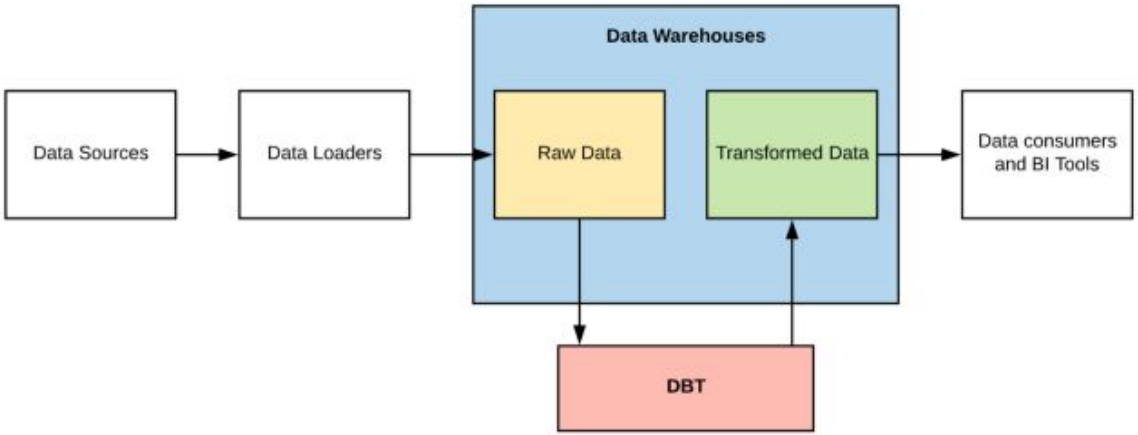
Apache Spark - Arquitetura



DBT (Data Build Tool)

- Ferramenta moderna de ETL;
- Controle de fluxo para Stack's de DW;
- Funcionalidades:
  - CLI e UI;
  - Data Lineage;
  - Interface para Python;
  - ETL em SQL:
    - Modularização;
    - Tabelas temporárias;

DBT - Arquitetura



https://medium.com/the-telegraph-engineering/dbt-a-new-way-to-handle-data-transformation-at-the-telegraph-868ce3964eb4

AWS - Serviços de Bach de dados

- Serviços relacionados:
  - EC2:
    - Serviço de computação;
  - EMR:
    - Elastic MapReduce:
      - Hadoop e Spark;
  - ECS:
    - Serviços de containers;

Ingestão por Streaming

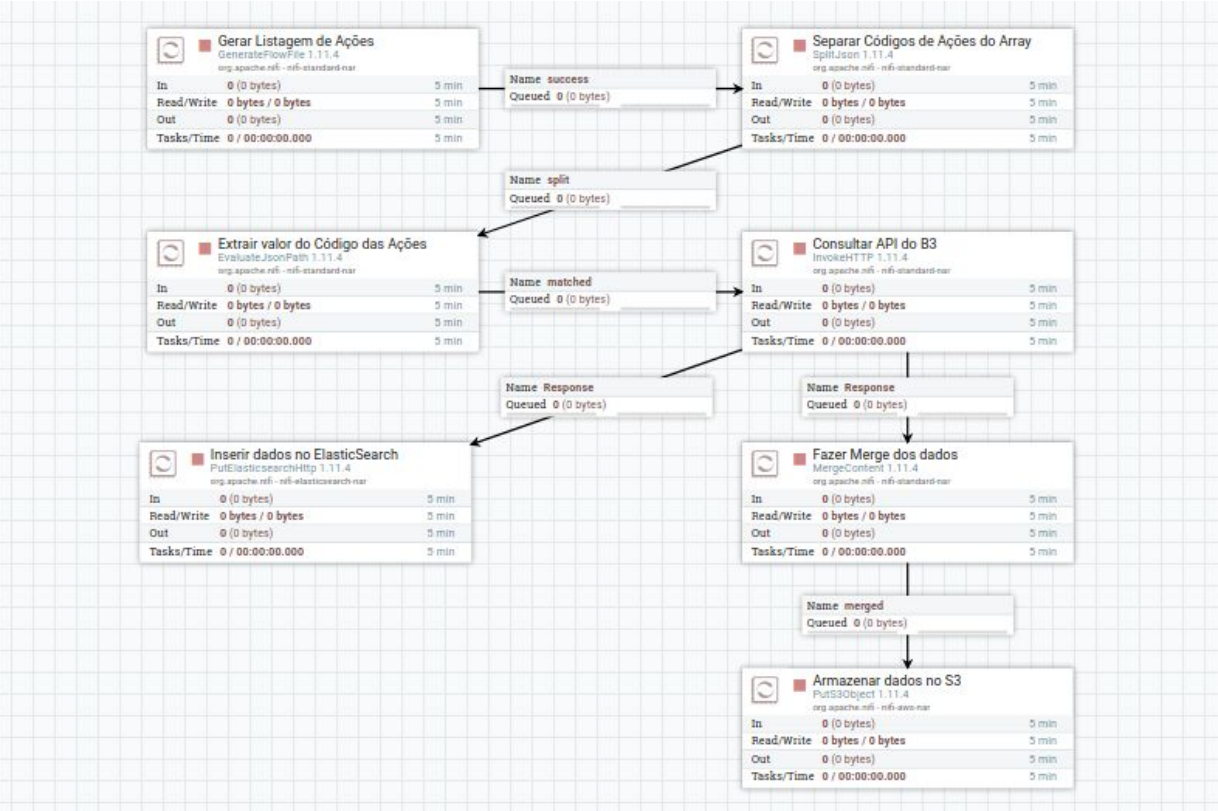
Apache NiFi

- Open Source;
- Criação de pipeline de dados;
- Orquestração, monitoramento e linhagem dos dados;
- Interface Web e Drag and drop;

Apache NiFi

- FlowFiles;
- Processor;
- Group Processors;
- Queues;
- Connections;

Apache NiFi - UI

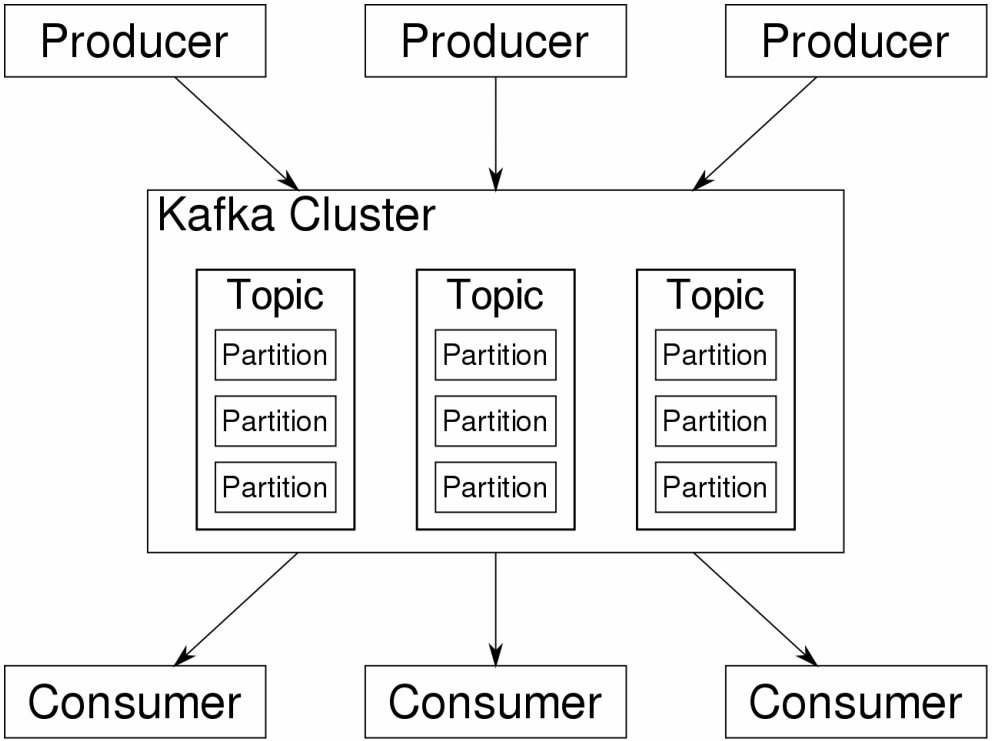


datasprints

Apache Kafka

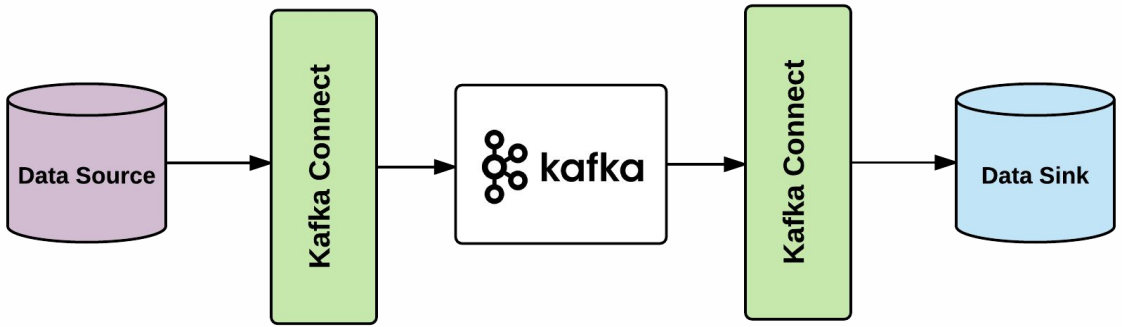
- Ferramenta open source para Streaming de dados;
- Conceitos:
  - Topics e partitions;
  - Brokers;
  - Producers;
  - Consumers;
  - Zookeeper;
  - Kafka Connectors.

Apache Kafka - Arquitetura



https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781787281349/7/ch07lvl1sec80/kafka

Kafka Connect - Arquitetura



https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781787281349/7/ch07lvl1sec80/kafka-connect

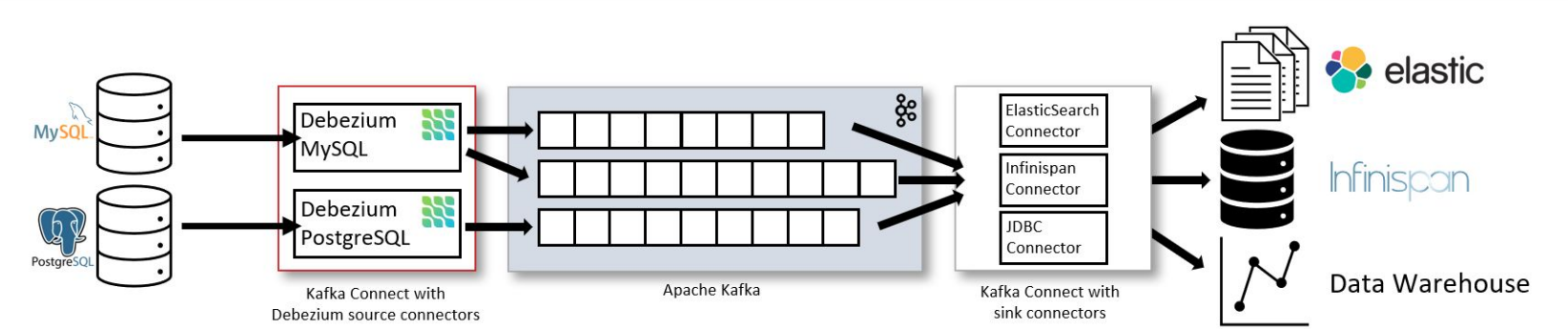
Debezium

- Ferramenta para capturar alterações em bases de dados;
- Formato distribuído;
- Baseado em Kafka e Kafka Connect;
- Conectores disponíveis:
  - Postgres;
  - MySQL;
  - Cassandra;
  - Oracle;
  - E outros...

debezium

datasprints

Debezium - Arquitetura



https://debezium.io/documentation/reference/architecture.html

datasprints

AWS - Serviços de Streaming

- Serviços relacionados:
  - Kinesis Stream:
    - Serviço para integrar com suas aplicações.
  - Kinesis Firehouse:
    - Serviço gerenciado pela AWS.
  - EMR:
    - Possível executar o Spark Streaming.
  - EC2:
    - Instalação de serviços próprios de stream.

Atividades propostas

1. Utilização do Debezium para gravar no S3 (padrão como avro);
    a. Desafio I: gravar como parquet no S3.

Obrigado!

datasprints