Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1. REAL E NEAR REAL – TIME ETL

PROF. CARLOS BARBOSA

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1. REAL E NEAR REAL – TIME ETL

PROF. CARLOS BARBOSA

Event Stream [ES]

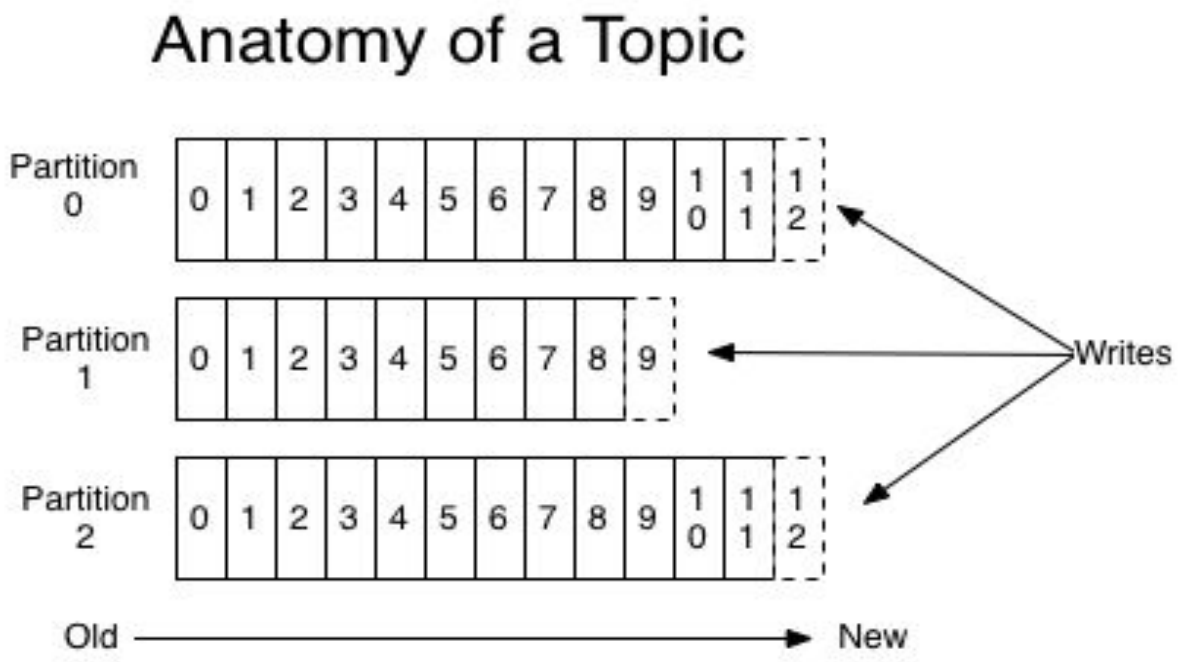
- Event Stream [ES] = Representação de Dados de forma unbounded
- Unbounded = Infinito e que sempre cresce
- Transações Financeiras, IoT, Ecommerce, Live Streaming

Características de um Event Stream

- Ordenação;
- Imutabilidade;
- Replayble.

Ordenação [ES]

- Dados Ordenados;
- Ordenados por natureza pelo tempo;
- Sistema bancário.



Saldo Atual = 20

Extrato + 20

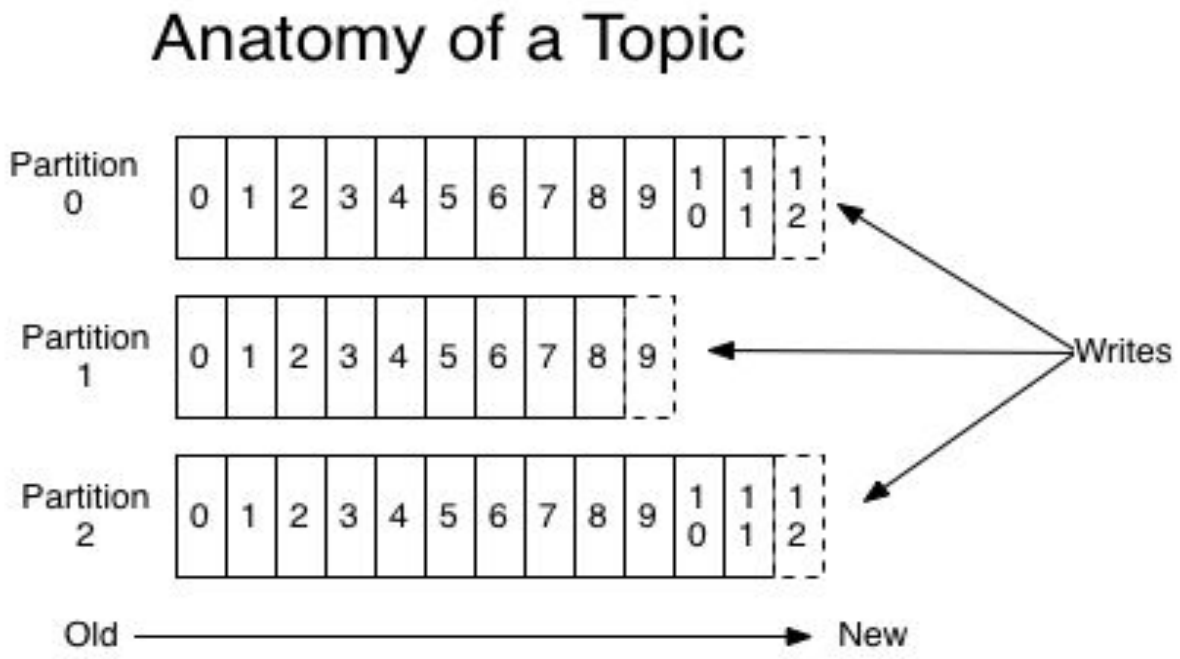
Devolução − 20

Imutabilidade [ES]

- Dados Imutáveis;
- Não é possível modificar o evento;
- Append de todos os eventos.

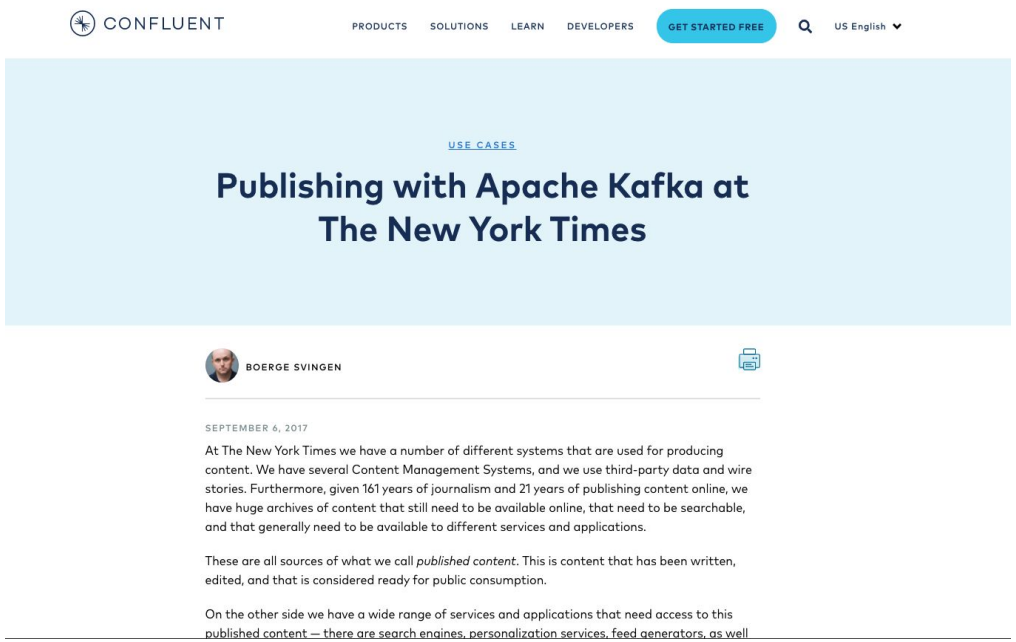
UPDATES

DELETES



Replayble [ES]

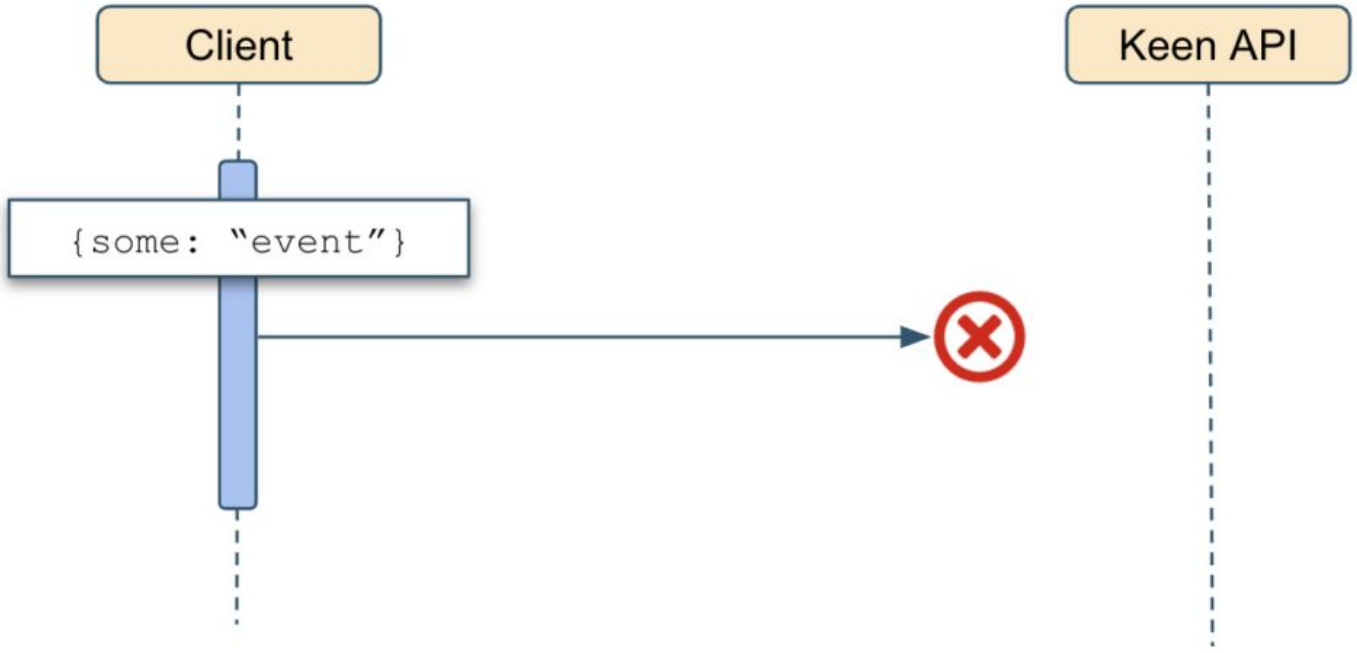
- Eventos Reutilizáveis;
- Retenção de Logs [7 Dias];
- Pode expandir para o infinito;
- Armazenamento histórico dos Dados.



https://www.confluent.io/blog/publishing-apache-kafka-new-york-times/

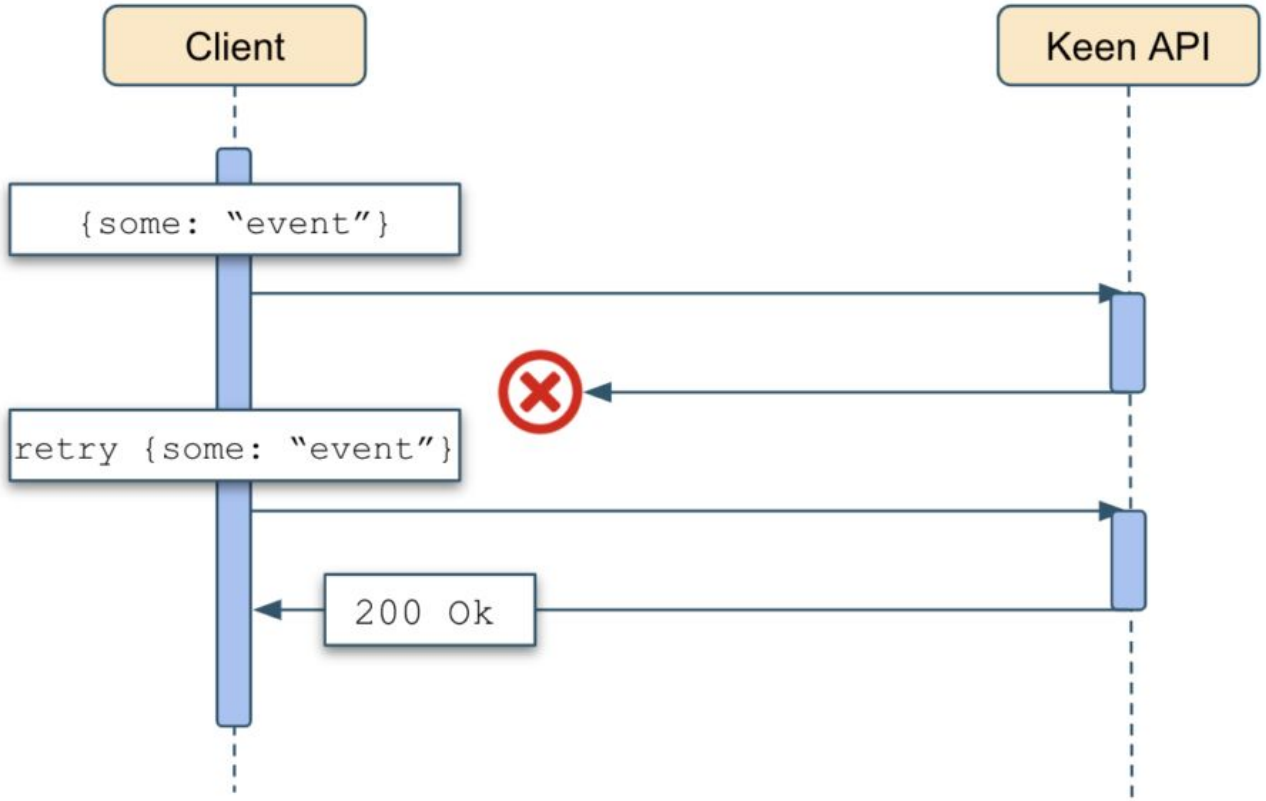
Delivery Message Guarantees

- At [Most] Once
    - Dados podem ser perdidos.



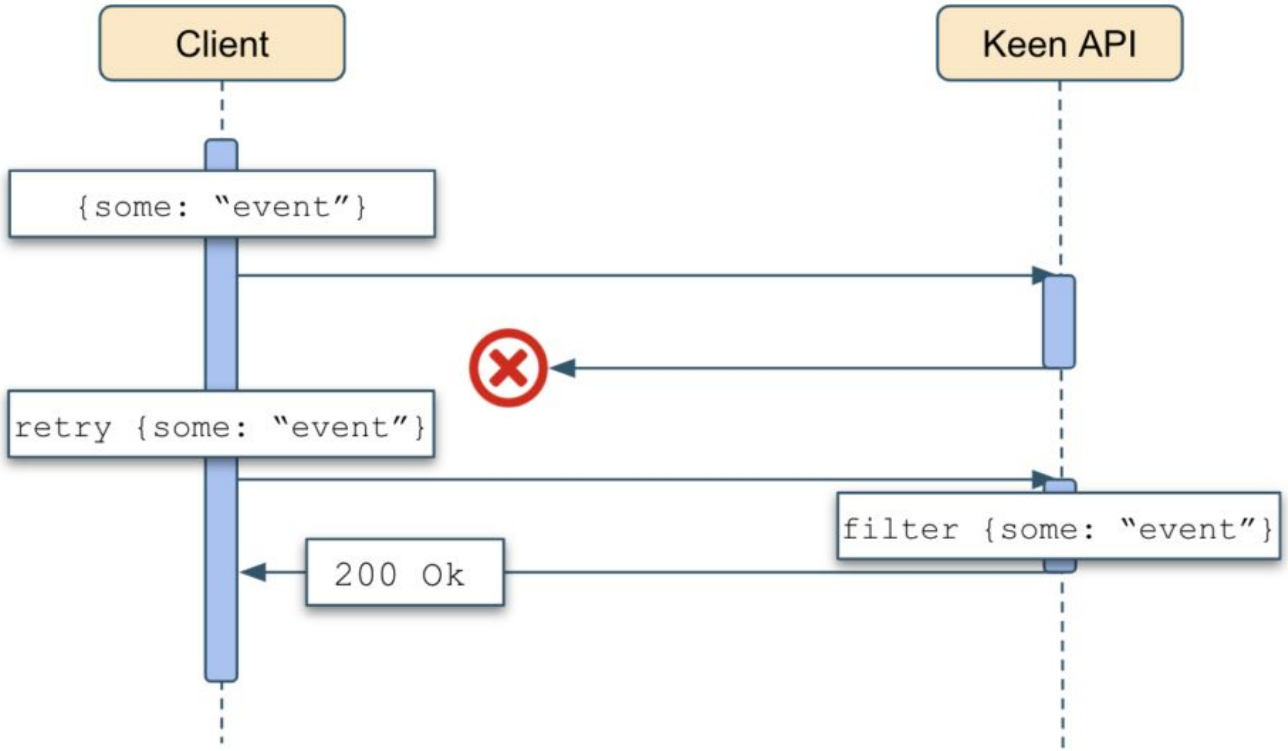
Delivery Message Guarantees

- At [Least] Once
  - Dados podem ser duplicados.



Delivery Message Guarantees

- [Exacly] Once Semantics



- Agora que entendemos o que é um Event Stream, iremos compreender como processar esses eventos.

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1.1 STREAM PROCESSING

PROF. CARLOS BARBOSA

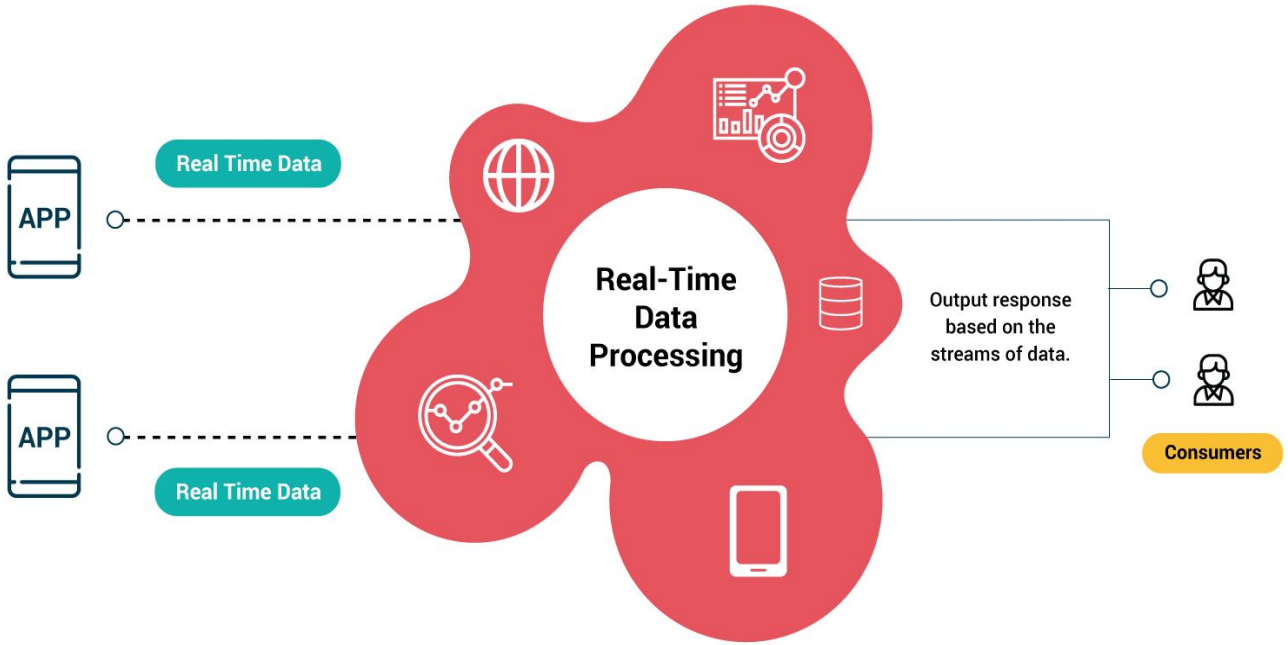
Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1.1 STREAM PROCESSING

PROF. CARLOS BARBOSA

Event Stream Processing [ESP]

- Capacidade de armazenar eventos em segundos ou milissegundos;
- Capacidade de receber eventos em segundos ou milissegundos;
- Disponibilização de Streams após eventos processados com baixa latência.



Engines Event Stream Processing

- ❑ Kinesis Data Streams
- ❑ Kinesis Data Firehose
- ❑ Kinesis Data Analytics
- ❑ AWS Lambda
- ❑ Amazon EMR

Engines Event Stream Processing

- Synapse Analytics
- Stream Analytics
- Azure Functions
- HdInsight

Engines Event Stream Processing

- Dataflow
- Google Cloud Functions
- Dataproc

Engines Event Stream Processing

- ❑ Kafka Strimzi Operator
- ❑ KsqlDB
- ❑ Apache Spark Structure Streaming
- ❑ Apache Flink
- ❑ Apache Storm
- ❑ Apache Beam

StreamingSQL [Engines]

- Flink SQL [2016]
- Structured Streaming [2016]
- KsqlDB [2017]
- Beam SQL [2017]

Python & Streaming [Engines]

- Apache Flink [Table API]
- Apache Spark [Pyspark API]
- Faust [FaustAPI]

- Agora que entendemos o que é um Stream Processing, iremos nos aprofundar na tecnologia open source mais utilizada no planeta para event streaming, a famosa Kafka.

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1.3 INTRODUÇÃO AO APACHE KAFKA

PROF. CARLOS BARBOSA

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1.3 INTRODUÇÃO AO APACHE

PROF. CARLOS BARBOSA

O que é o Apache Kafka?

- **É uma plataforma de streaming de eventos;**
- Sistema distribuído baseado em servidor e cliente;
- Comunicação de baixa latência;
- Comunicação por protocolo TCP.

O que é o Apache Kafka?

- **Capaz de ler eventos**;
- Grava eventos em tópicos;
- Banco de dados atômico;
- A segunda maior tecnologia open source do planeta.

Conceitos e Terminologia

- No Kafka, os eventos possuem chave, valor e carimbo de data.

  - Event key: "Alice"
  - Event value: "Made a payment of $200 to Bob"
  - Event timestamp: "Jun. 25, 2020 at 2:06 p.m."

Broker

- Um cluster Kafka é composto por um ou vários brokers;
- Responsável por receber os eventos dos producers e gravá-los em disco.

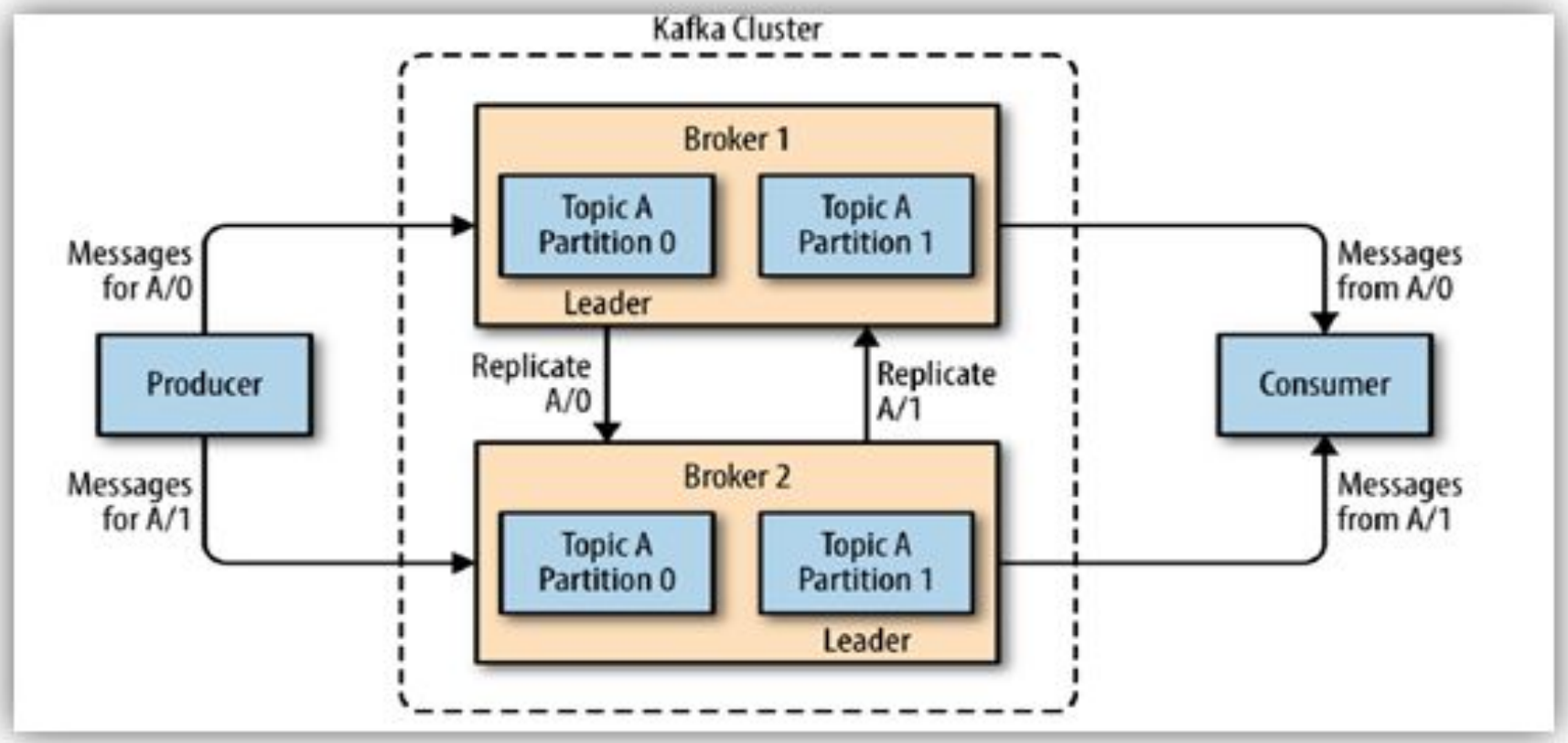
Producers

- Responsável por enviar dados ao Kafka.

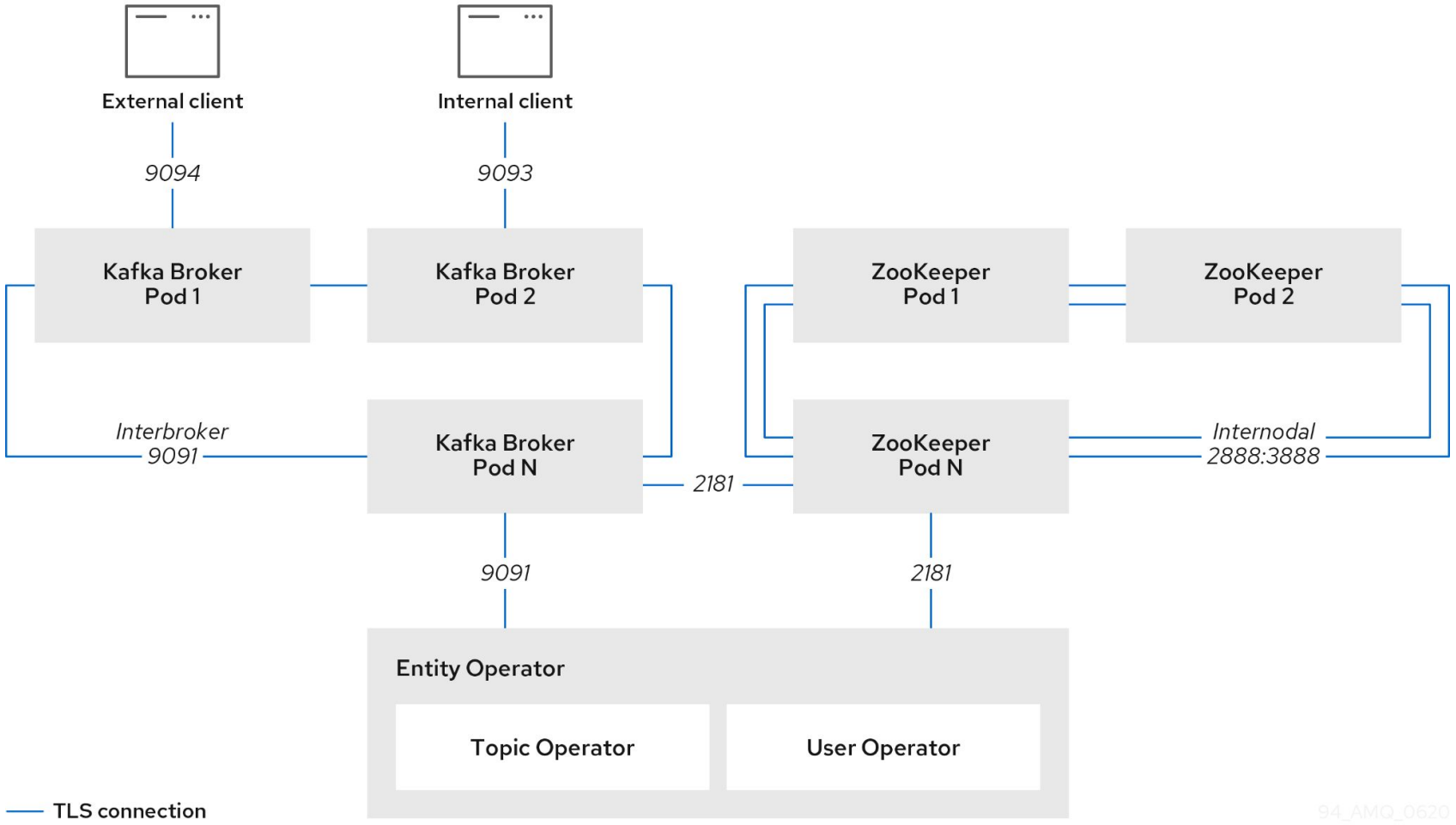
Consumers

- Responsável por ler ou processar eventos recebidos pelos producers.

Arquitetura do Kafka



Kafka no Kubernetes [k8s]



- Vejamos agora as principais tecnologias utilizadas para interagir com o Kafka.

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1.4 STREAM PROCESSING [KsqlDB E SPARK STREAMING]

PROF. CARLOS BARBOSA

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 1.4 STREAM PROCESSING [KsqlDB E SPARK STREAMING]

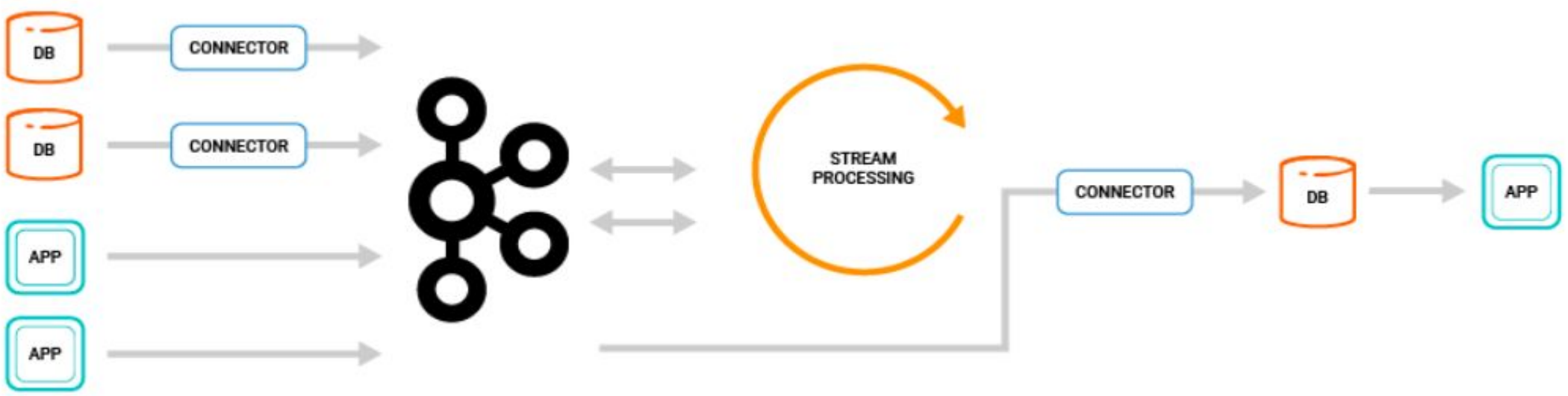
PROF. CARLOS BARBOSA

Stream Processing [KsqlDB]

- **Transformação contínua de fluxo de dados;**
- Ele permite que você pegue tópicos existentes do Apache Kafka ® e, em seguida, filtre, processe e crie novos tópicos derivados;
- Sendo possível reunir muitas fontes de dados dos clientes para criar um "perfil de cliente unificado".

Por que isso é importante?

- Facilitação na forma de realizar processamento em tempo real.



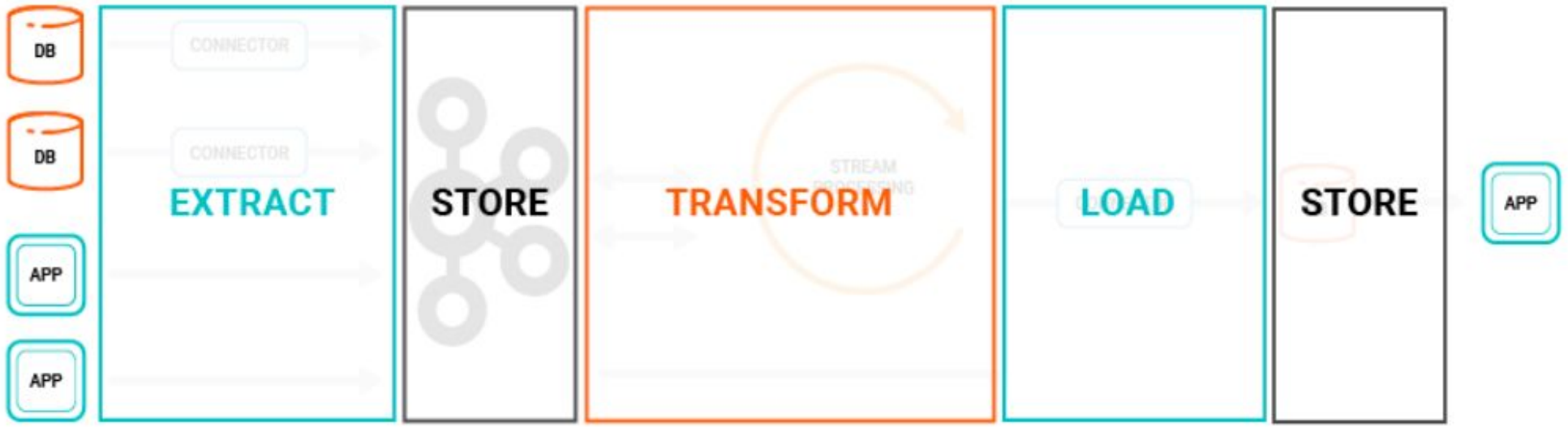
Stream processing [KsqlDB]

- Provavelmente, os dados com que você deseja trabalhar ainda estão em Batch;
- O Kafka Connect tem um grande ecossistema de conectores open source;
- KsqlDB lhe permite controlar e executar diretamente conectores desenvolvidos para funcionar com o Kafka Connect.

```
CREATE SOURCE CONNECTOR rider_profiles WITH (
  'connector.class'          = 'io.confluent.connect.jdbc.JdbcSourceConnector',
  'connection.url'           = 'jdbc:postgresql://postgres:5432/postgres',
  'key'                      = 'profile_id',
  ...
 );
```

Por que isso é importante?

- Em outras palavras, você tem as etapas modularizadas, de forma que cada uma delas seja tratada por sistemas diferentes:



Por que isso é importante?

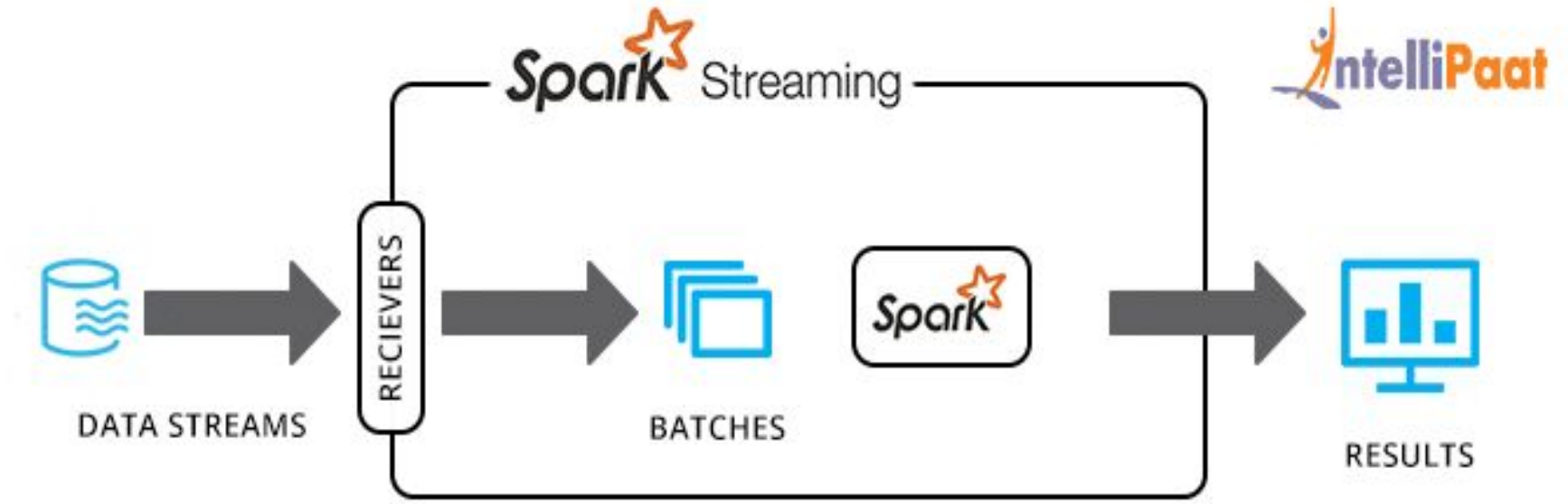
- ksqlDB nos permite simplificar a forma de processamento utilizando SQL.

Stream processing [Spark Streaming]

- Permite o processamento escalável, de alto rendimento e tolerante a falhas de fluxos in real time;

- Os dados podem ser ingeridos a partir de muitas fontes e processados por meio de algoritmos complexos.

Como funciona o Spark Streaming?



- Agora que entendemos as principais ferramentas utilizadas para o stream processing, iremos desenhar uma arquitetura orientada a serviços.

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 2. ARQUITETURA ORIENTADA A EVENTOS

PROF. CARLOS BARBOSA

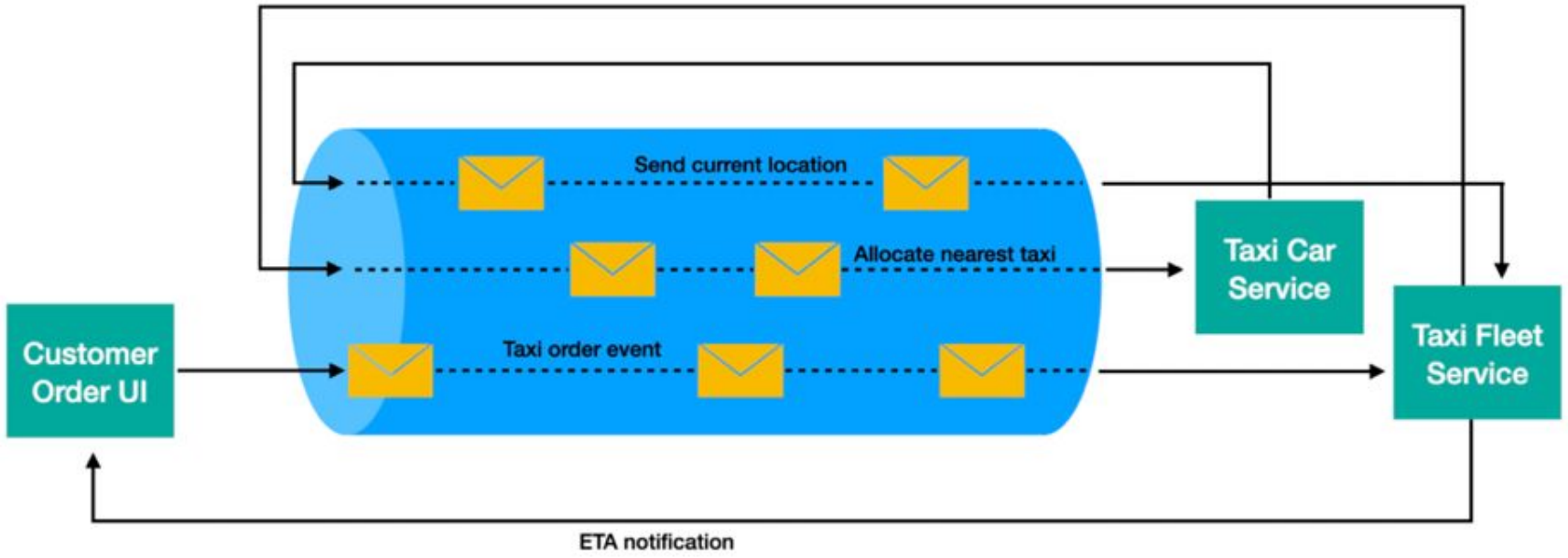
Processamento de Fluxos Contínuos de Dados

CAPÍTULO 2. ARQUITETURA ORIENTADA A EVENTOS

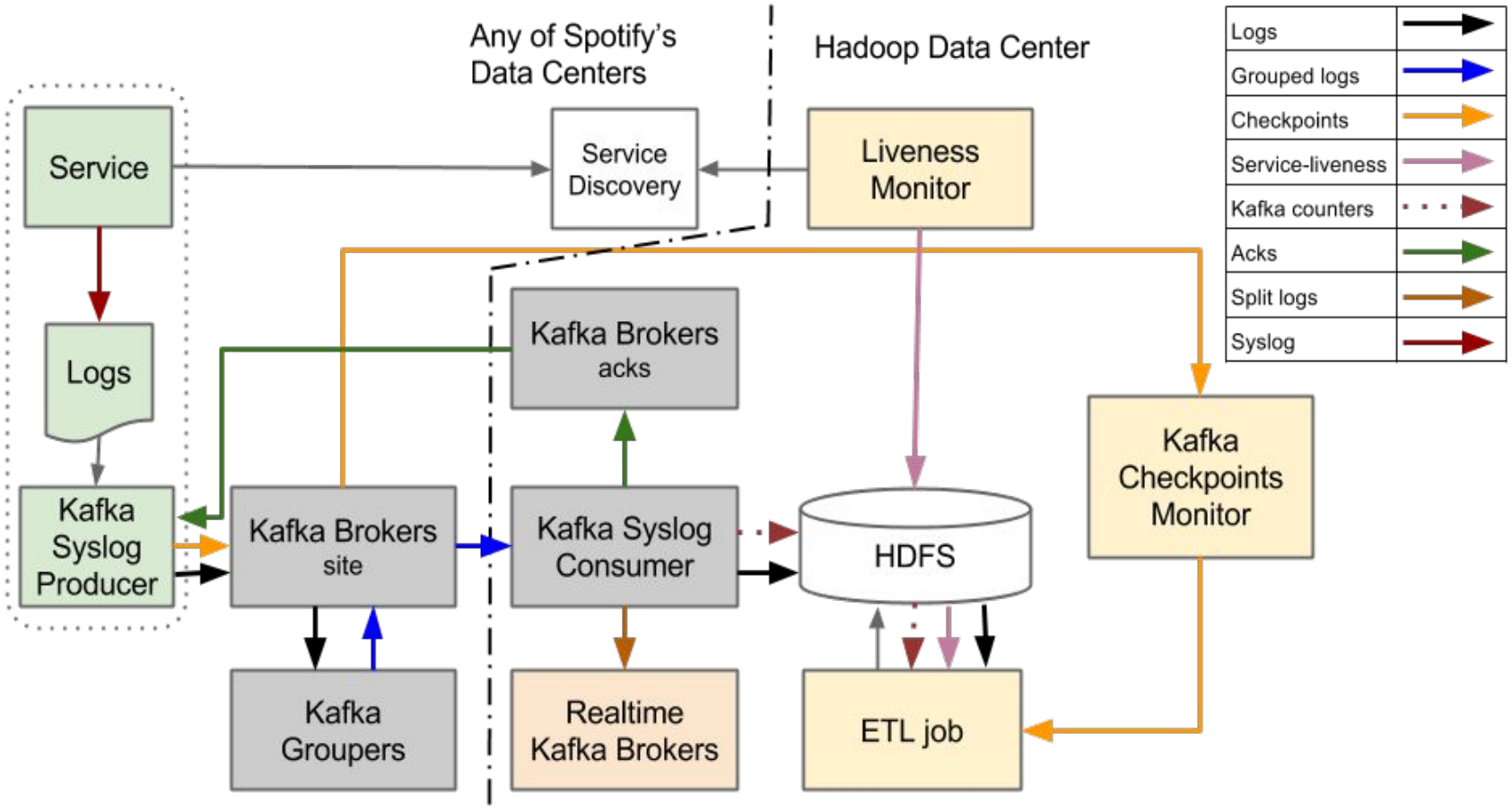
PROF. CARLOS BARBOSA

O que é uma Arquitetura Orientada a Eventos?

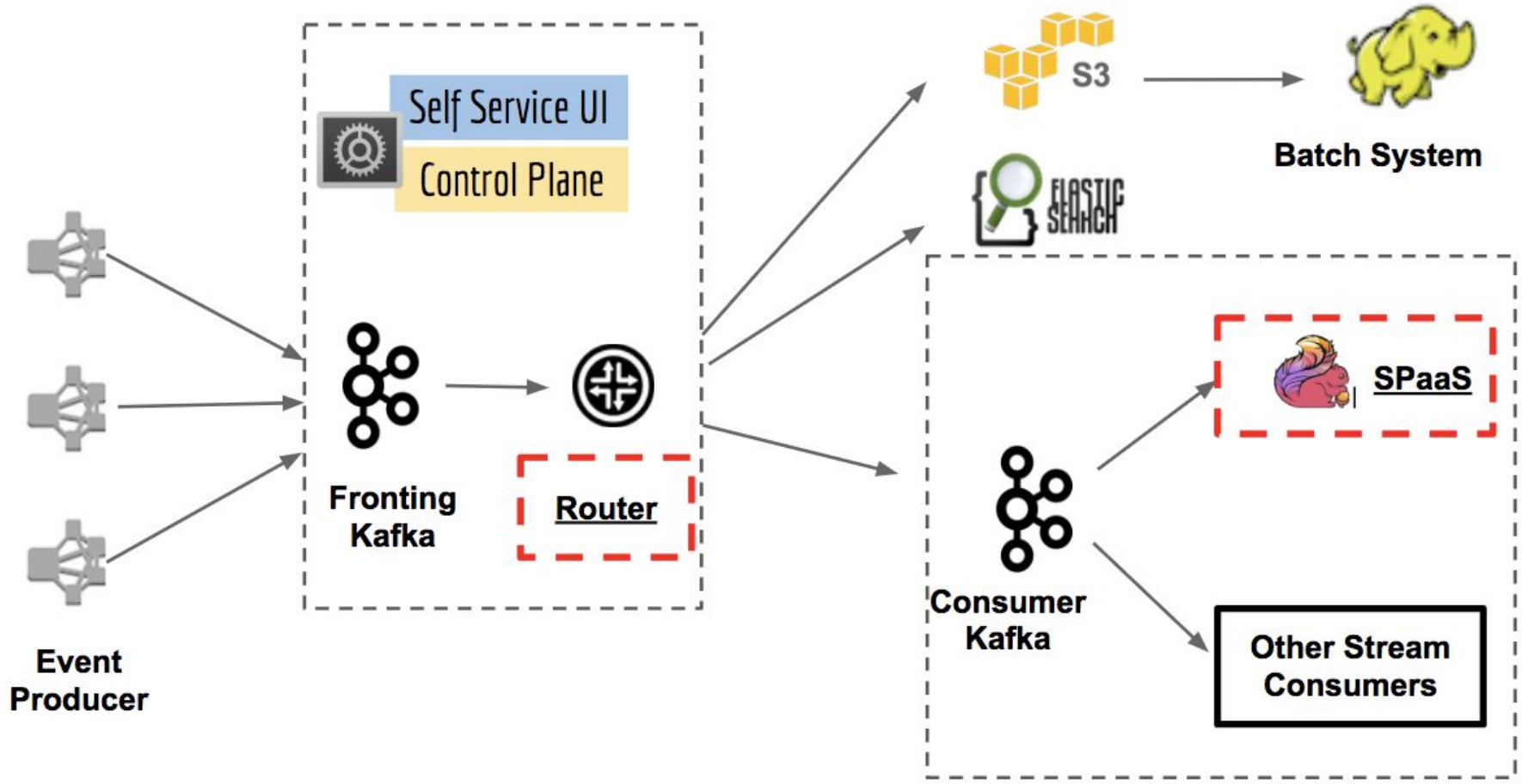
- A arquitetura orientada a eventos se refere a um sistema de microsserviços que são fracamente acoplados e trocam informações entre si, por meio da produção e do consumo de eventos;
- Fonte única de verdade com um registro de eventos imutáveis.



Spotify Delivery Stream

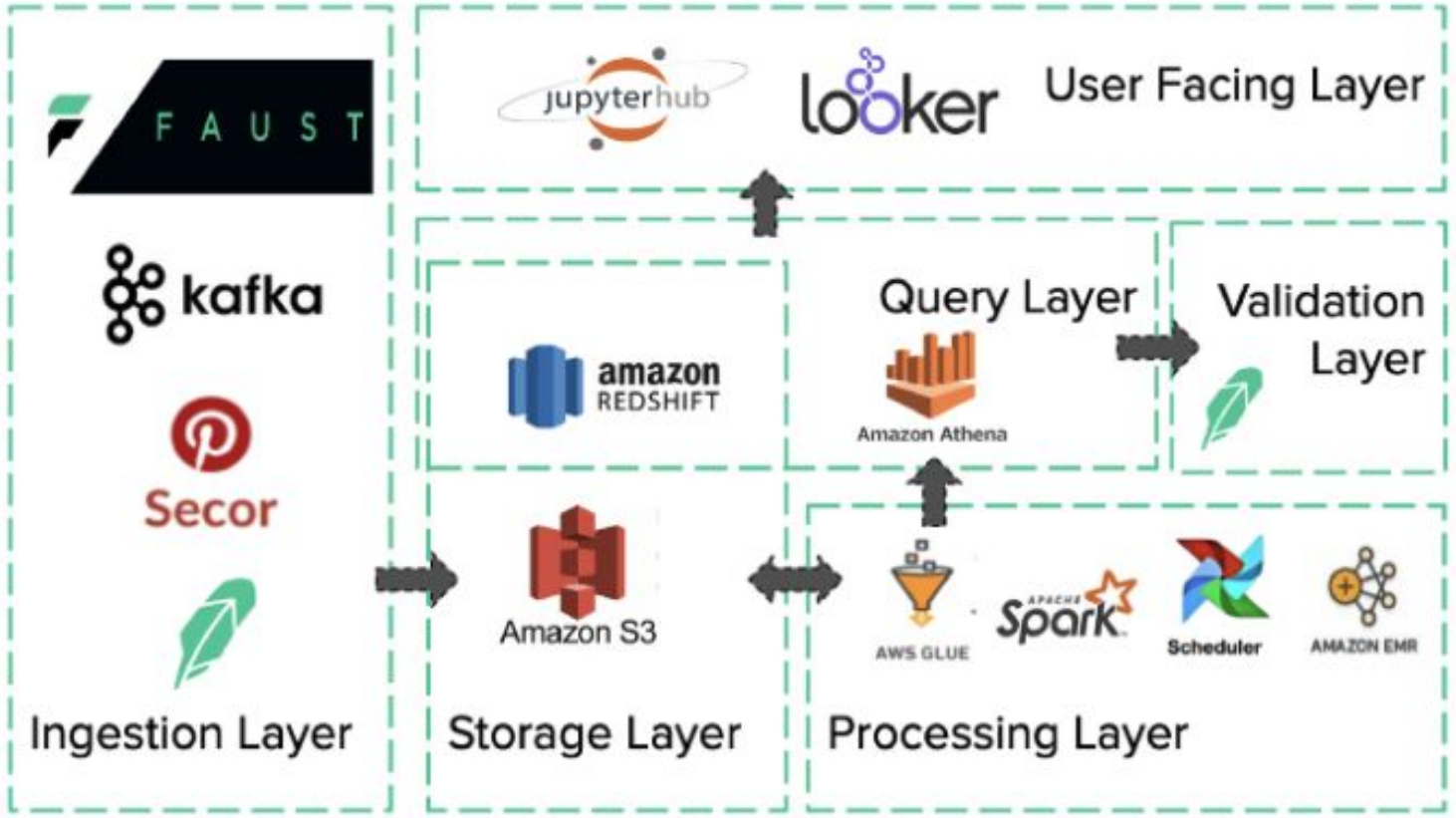


Netflix Delivery Stream



Robinhood Delivery Stream

Robinhood



Architecture Overview

- Agora que entendemos como desenvolver arquiteturas orientadas a eventos, iremos analisar os padrões a serem adotados.

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 3.1 DATA LAKE, DATA LAKE HOUSE

PROF. CARLOS BARBOSA

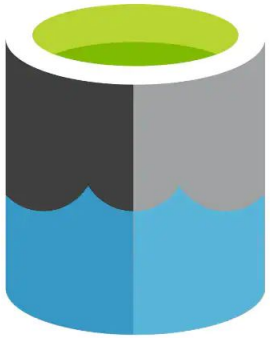
Processamento de Fluxos Contínuos de Dados

CAPÍTULO 3.1 DATA LAKE, DATA LAKE HOUSE

PROF. CARLOS BARBOSA

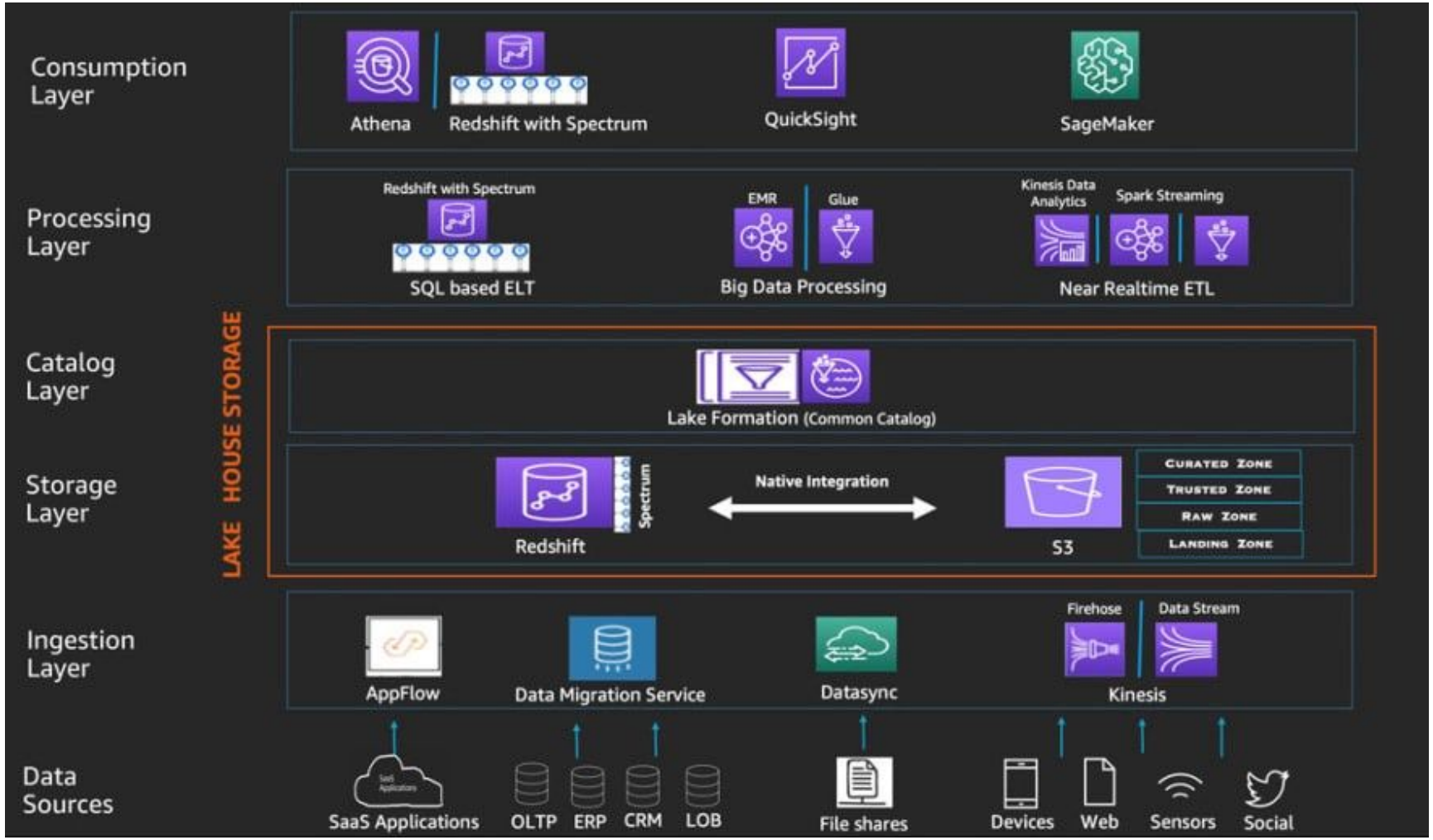
Architecture Data Lake

- Armazenamento de arquivos
  - Armazenamento dos dados em formato parquet;
  - Leitura de diversos tipos de dados [XSLX, CSV, TXT, TSV, PNG];
  - Operações de escrita [Overwrite, Append].
- Processamento de dados
  - Compressão Snappy [Até 70% de compressão do dado];
  - Dificuldade [Escrita lenta para grandes volumes];
  - Cenário limitado a [Batch e microbatch].



- Storage Layer
  - Data Storage [S3, Blob Storage, Azure Data lake gen 1, gen 2];
  - Redshift.

Architecture Data Lake

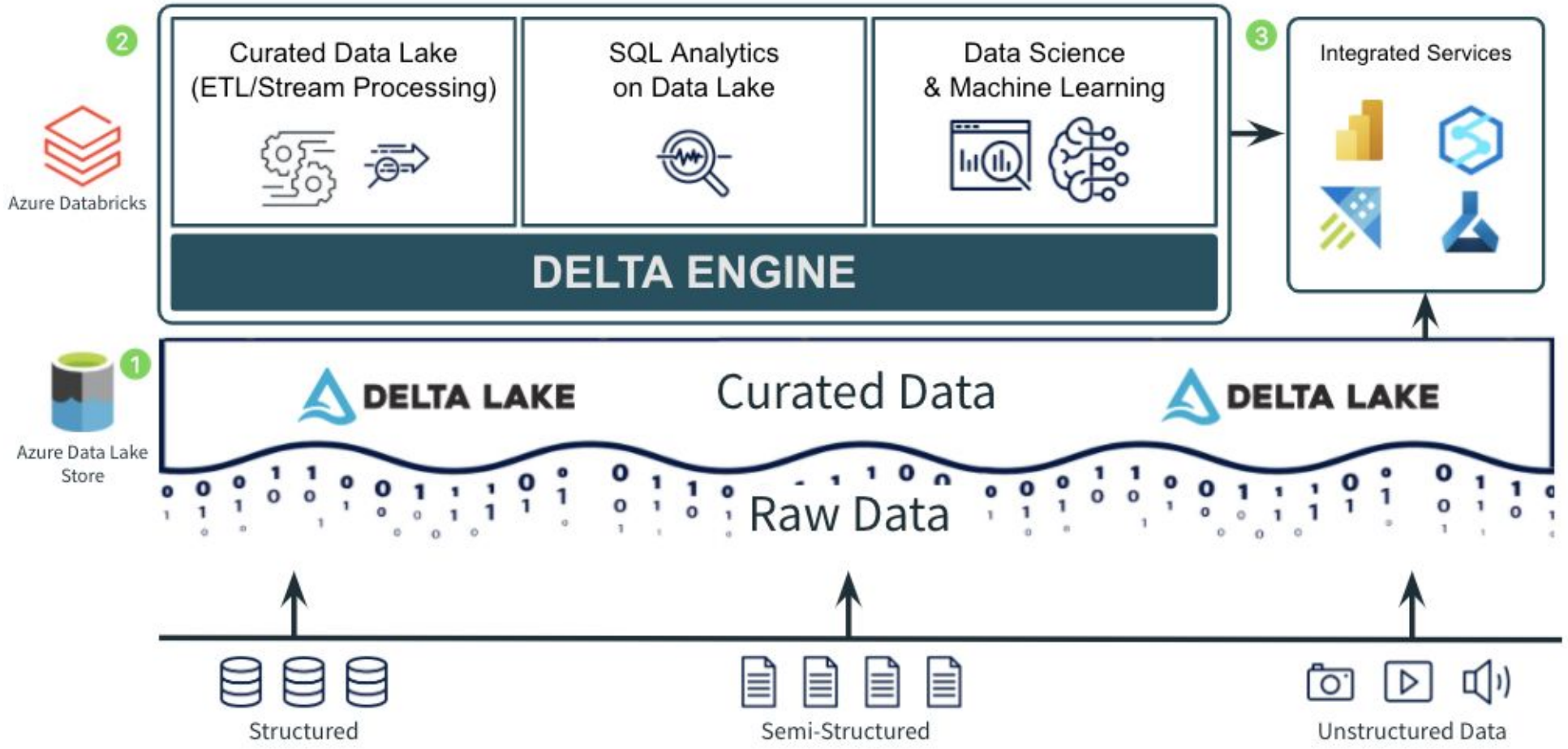


Architecture Data Lakehouse

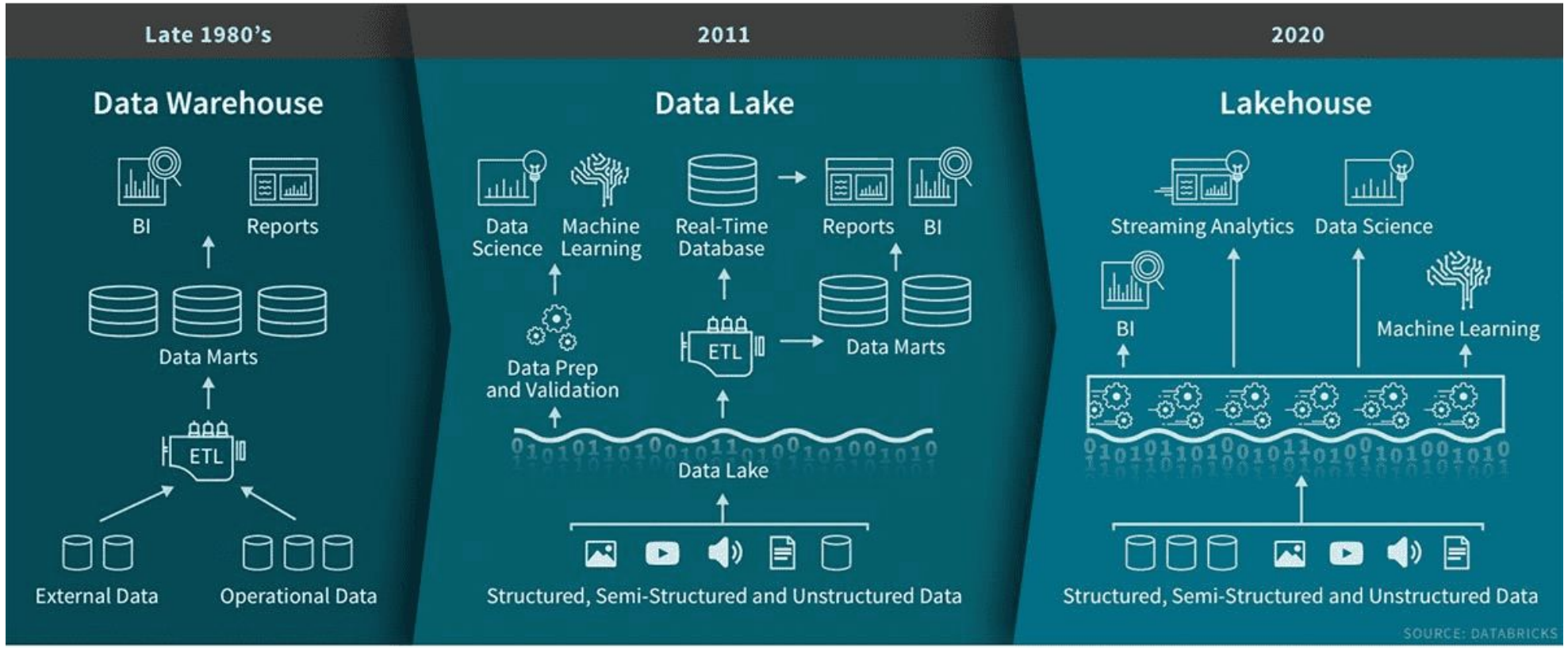
- Armazenamento de arquivos
  - Upserts;
  - Time Travel;
  - Schema Evolution;
  - Tabelas Delta [Bronze, Silver, Gold].
- Teorema ACID
  - Atomicidade;
  - Consistência;
  - Isolamento;
  - Durabilidade.

- Processamento de dados
  - Fácil de fazer real time analytics;
  - Cenário Lambda [Batch e Streaming];
  - 10x mais rápido na escrita.
- Consumo de dados
  - Ferramentas analíticas plugam diretamente nas tabelas delta [Metabase, Redash, PowerBI].

Architecture Data Lakehouse



Evolução Histórica



- Até então não era possível ter uma estrutura histórica com dados em tempo real, será mesmo?

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 4. CUSTOS DE ARQUITETURAS DE PROJETOS DE DADOS

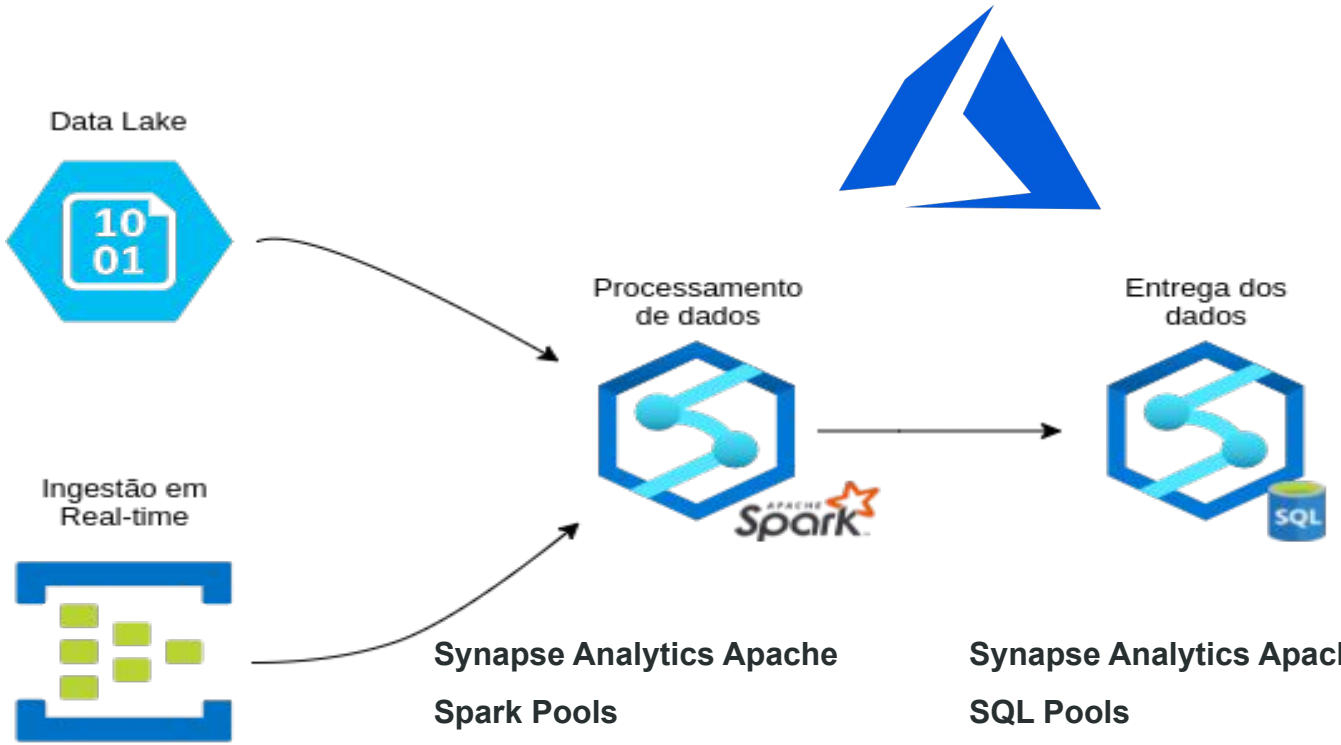
PROF. CARLOS BARBOSA

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 4. CUSTOS DE ARQUITETURAS DE PROJETOS DE DADOS

PROF. CARLOS BARBOSA

Pipeline de dados no Azure



Azure Blob Storage

- Região: EastUS2
- Tier: Premium
- Redundância: LRS
- Capacidade: 1 TB
- Custo Mensal: **R$ 847**

Azure Event Hubs

- Região: EastUS2
- Tier: Standard with Capture
- Ingress: 1 Milhão
- Throughput: 3 MB
- Custo Mensal: **R$ 1.567**

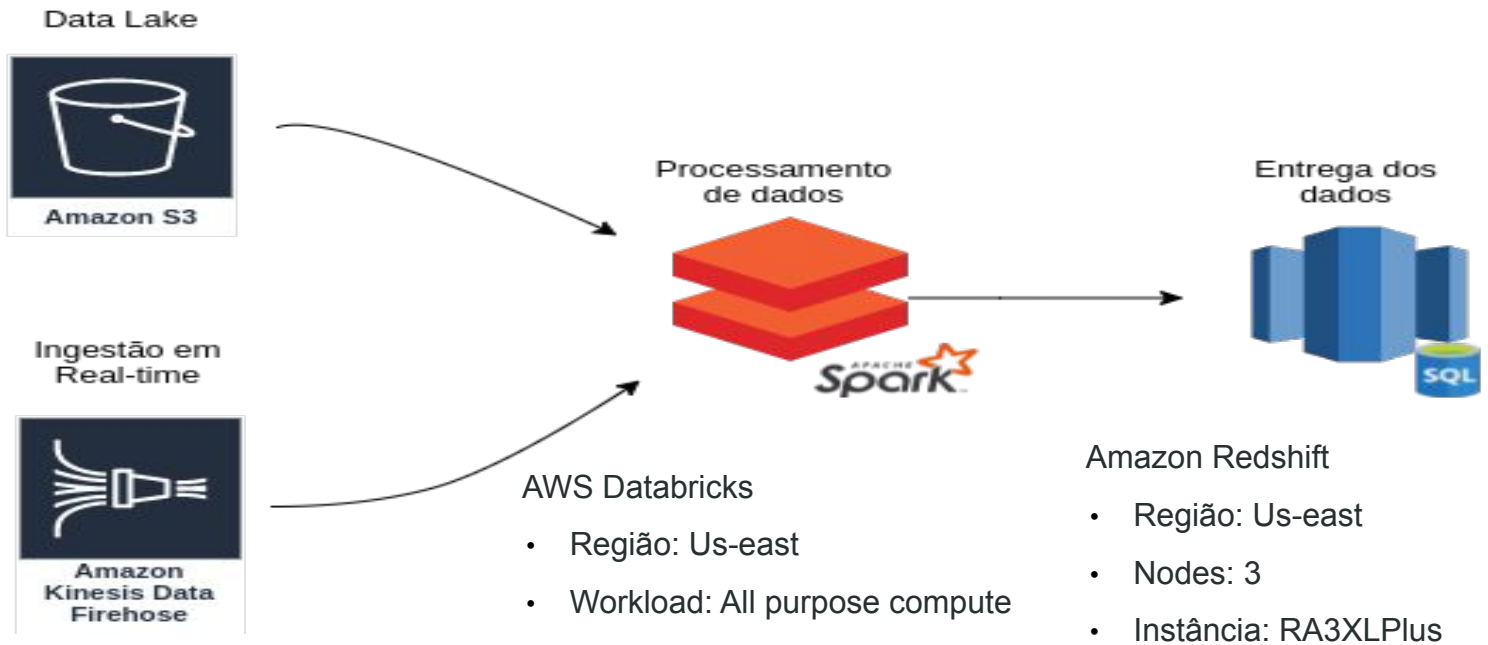
Synapse Analytics Apache Spark Pools

- Região: EastUS2
- Instância: Small
- Spec: 4 vCPUs & 32 GB RAM
- VMs: 3
- Custo Mensal: **R$ 6.556**

Synapse Analytics Apache Dedicated SQL Pools

- Região: EastUS2
- Tier: Compute Optmized Gen2
- DWU: 300
- Armazenamento: 1 TB
- Custo Mensal: **R$ 14.580**

Pipeline de dados na Amazon



Amazon S3

- Região: us-east-1
- Capacidade: 1 TB
- Escaneado: 100 GB
- Custo Mensal: **R$ 170**

Amazon Kinesis Firehose

- Região: Us-east
- Unidades: Milhares
- Quantidade de registros: 1 Milhão
- Tamanho do registro: 10 KB
- Conversão: Parquet & ORC
- Custo Mensal: **R$ 2.464**

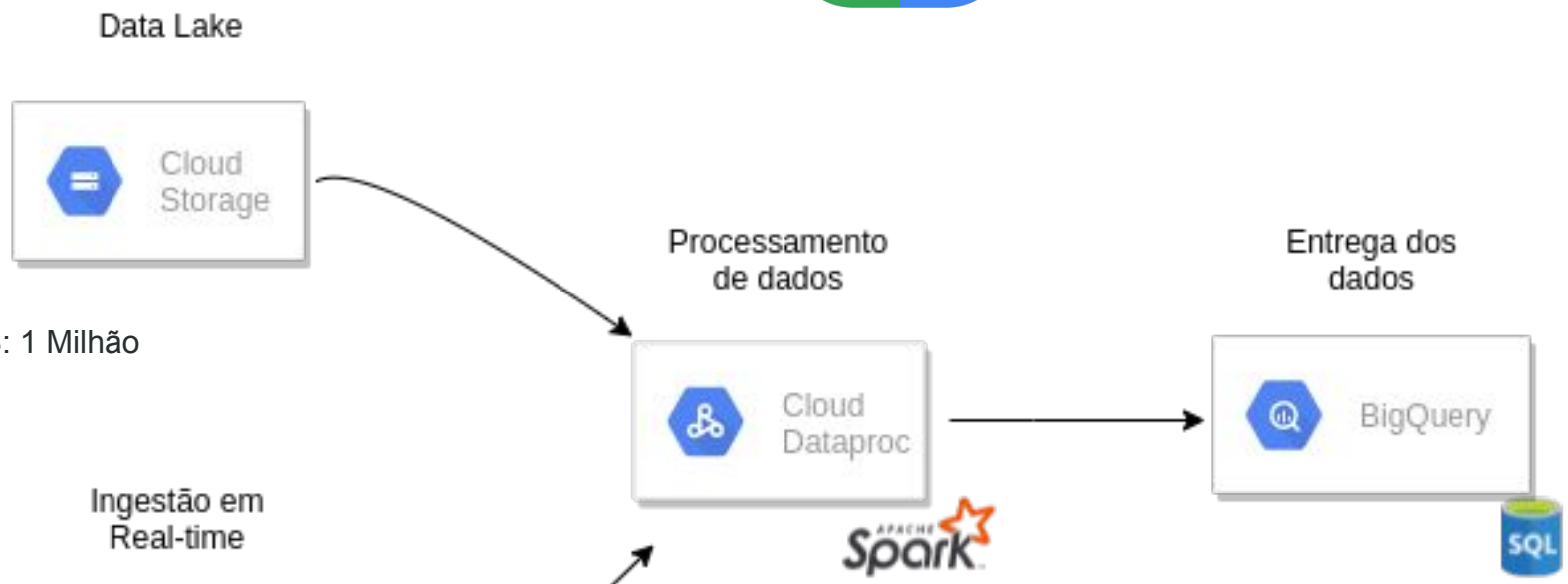
AWS Databricks

- Região: Us-east
- Workload: All purpose compute
- Tier: Enterprise
- Instância: Me.Xlarge
- Spec: 4 vCPUs & 16 GB RAM
- VMs: 3
- Custo Mensal: **R$ 5.791**

Amazon Redshift

- Região: Us-east
- Nodes: 3
- Instância: RA3XLPlus
- Spec: 4 vCPUs & 32 GiB
- Backup: 100 Gb
- Spectrum: 100 Gb
- Armazenamento: 1TB
- Custo Mensal: **R$ 13.238**

Pipeline de dados na Google Cloud



Google Cloud Storage

- Região: us-east-1
- Capacidade: 1 TB
- Operações Classe A e B: 1 Milhão
- Custo Mensal: **R$ 143**

Google Cloud Pub/Sub

- Região: Us-east
- Volume de mensagens: 100 Gb
- Retido: 10 GB
- Snapshot: 10 GB
- Custo Mensal: **R$ 193**

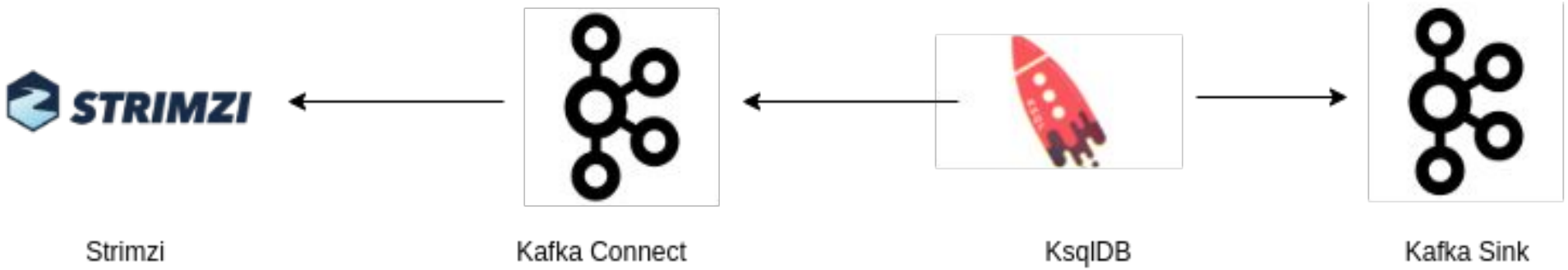
Google DataProc

- Região: Us-east
- VM-Class: Regular
- Instância: N1-Standard-4
- Spec: 4 vCPUs & 15 GB RAM
- VMs: 3
- Armazenamento: 500 Gb SSD
- Custo Mensal: **R$ 2.156**

Google BigQuery

- Região: Us-east
- Armazenamento: 1TB
- Inserts Streaming: 100 GB
- Queries: 2 TB
- Custo Mensal: **R$ 170**

Pipeline de dados no Kubernetes



EC2

- Vms: 8
- Instância: M5.xLarge
- Memória: 16 GB RAM
- Custo Mensal: **R$ 3.795**

EKS

- Unidades: 3
- Custo Mensal: **R$ 1.138**

Amazon FSX for Lustre

- Armazenamento: 2 TB SSD
- Throughput: 50 MB/s
- Custo Mensal: **R$ 1.492**

Pipeline de dados no Kubernetes

Custo da Pipeline na Amazon AWS

- Camada de Armazenamento: **R$ 2.634**
- Camada de Processamento: **R$ 5.791**
- Camada de Entrega de Dados: **R$ 13.238**
- Custo Mensal Total: **R$ 21.663**

Custo da Pipeline no Kubernetes

- Máquinas virtuais: **R$ 3.795**
- Cluster: **R$ 1.138**
- Armazenamento: **R$ 1.492**
- Custo Mensal Total: **R$ 6.425**

Custo da Pipeline no Microsoft Azure

- Camada de Armazenamento: **R$ 2.414**
- Camada de Processamento: **R$ 6.556**
- Camada de Entrega de Dados: **R$ 14.580**
- Custo Mensal Total: **R$ 23.550**

Custo da Pipeline na Google GCP

- Camada de Armazenamento: **R$ 336**
- Camada de Processamento: **R$ 2.156**
- Camada de Entrega de Dados: **R$ 170**
- Custo Mensal Total: **R$ 2.662**

Processamento de Fluxos Contínuos de Dados

CAPÍTULO 5. NOVOS STORAGES PARA ARMAZENAR DADOS EM TEMPO REAL

PROF. CARLOS BARBOSA

Processamento de Fluxos Contínuos de Dados

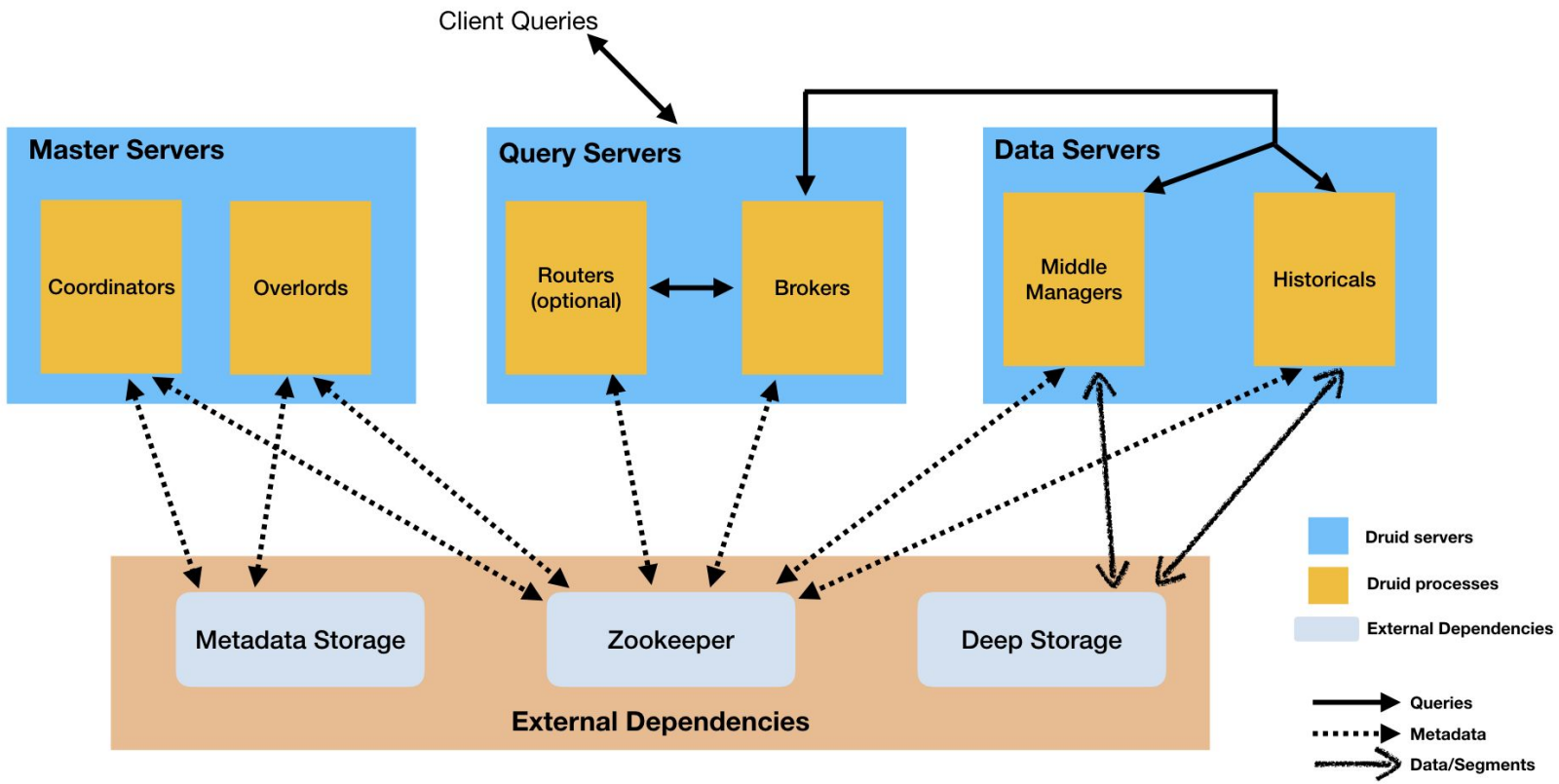
CAPÍTULO 5. NOVOS STORAGES PARA ARMAZENAR DADOS EM TEMPO REAL

PROF. CARLOS BARBOSA

Apache Druid

- Um banco de dados analítico rápido e moderno;
- Fácil de integrar com pipelines existentes;
- Consultas rápidas e consistentes em alta simultaneidade;
- Integração nativa com Kafka.

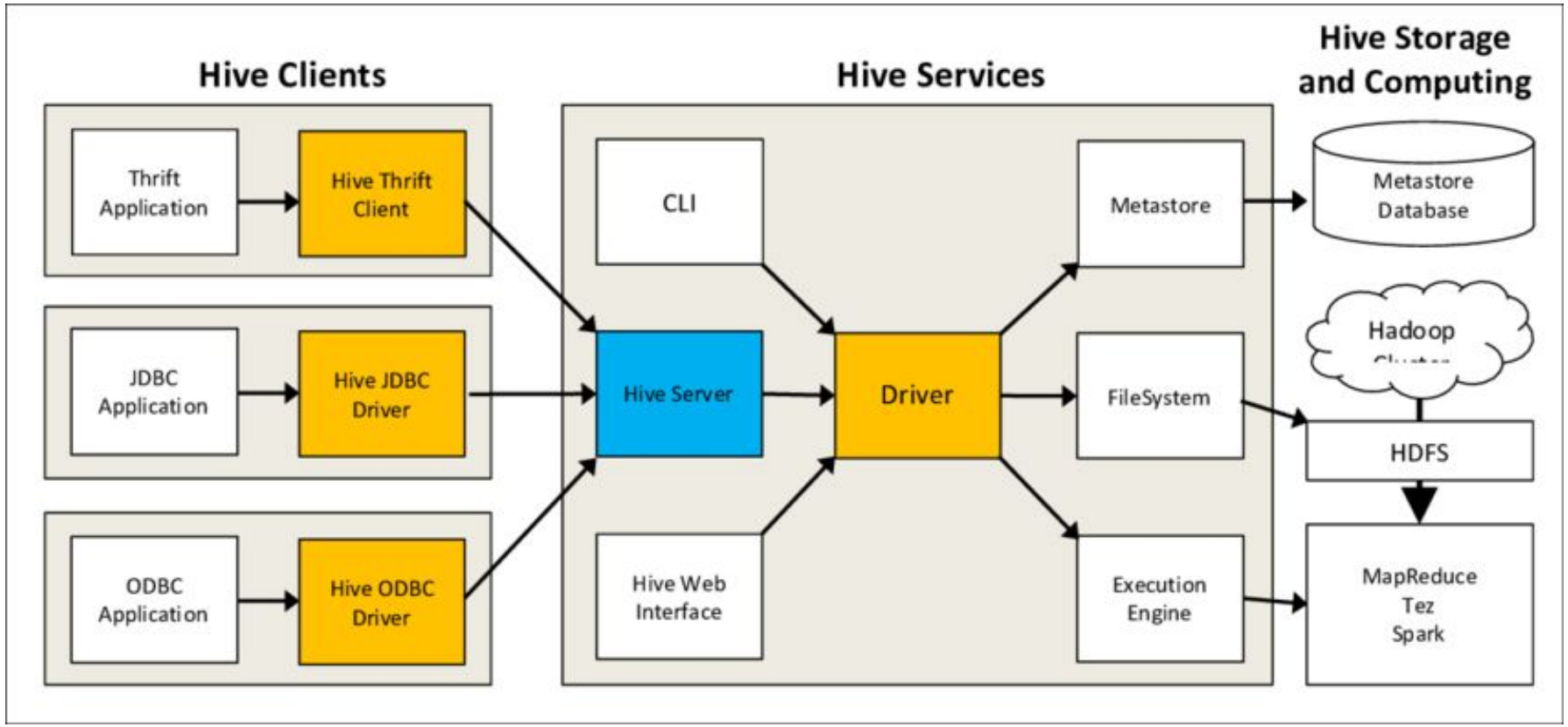
Apache Druid



Apache Hive

- Fácil de usar;
- Baixo custo;
- Big Data Scale;
- Integração nativa com HDFS.

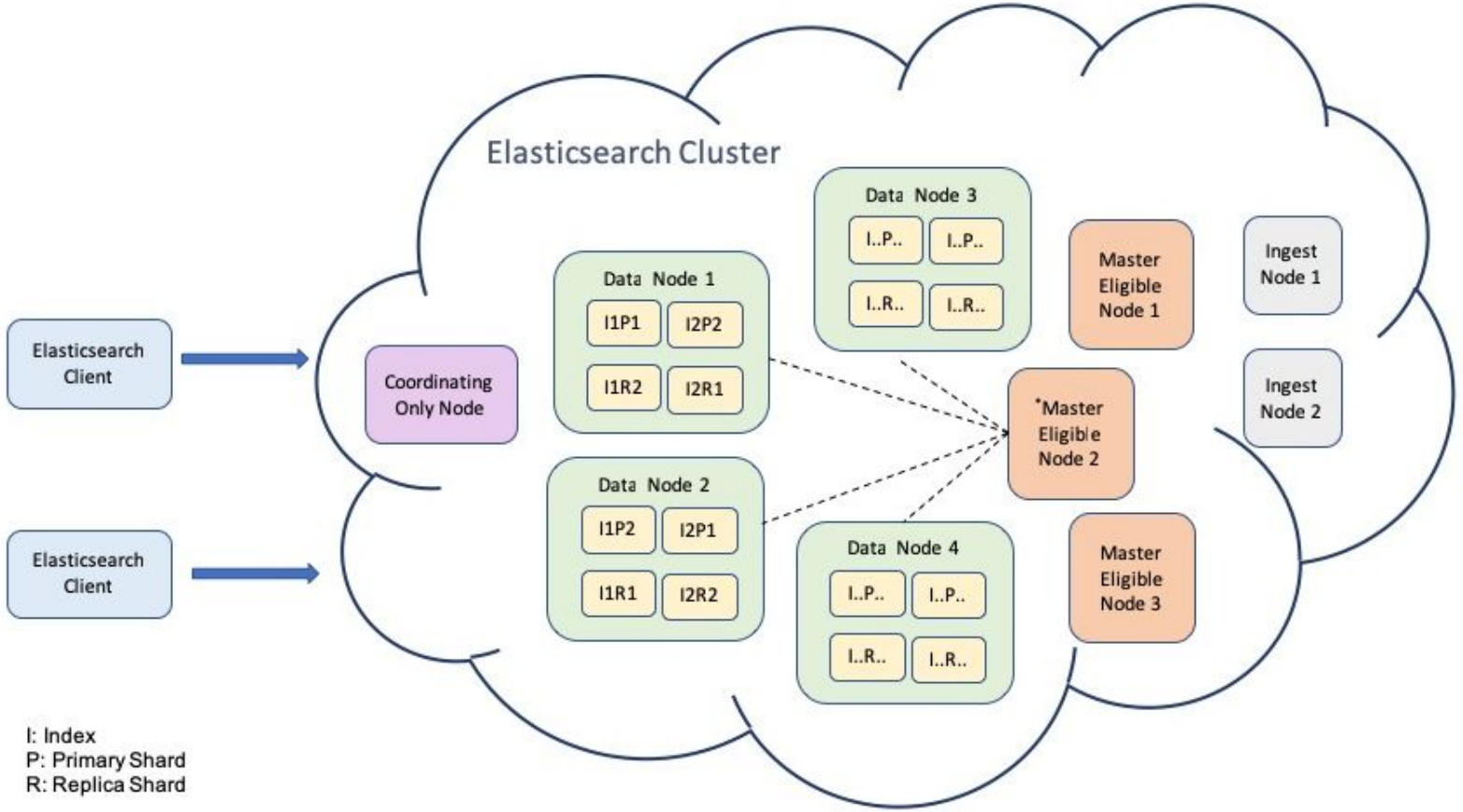
Apache Hive


Hive Clients
Hive Services
Hive Storage and Computing
Thrift Application
Hive Thrift Client
JDBC Application
Hive JDBC Driver
ODBC Application
Hive ODBC Driver
CLI
Hive Server
Hive Web Interface
Driver
Metastore
FileSystem
Execution Engine
Metastore Database
Hadoop
HDFS
MapReduce
Tez
Spark

ElasticSearch

- Orientado a documentos;
- Restful API;
- Escalável;
- Ótimo para pesquisa.

ElasticSearch



Apache Pinot

- Desenvolvido com a intenção de atender a consultas analíticas de baixa latência para aplicativos voltados ao cliente;
- Escalável horizontalmente;
- Dados imutáveis;
- Mudanças de configuração dinâmica.

Apache Pinot