Processamento e Análise de Dados em Tempo Real com Python, Kafka e ElasticSearch

DataTrain
meetup

Cícero
Moura

Desenvolvedor Full Stack na Máxima Tech, Pós-Graduando em IoT, Big Data e Machine Learning. Entusiasta de Tecnologia Assistiva.

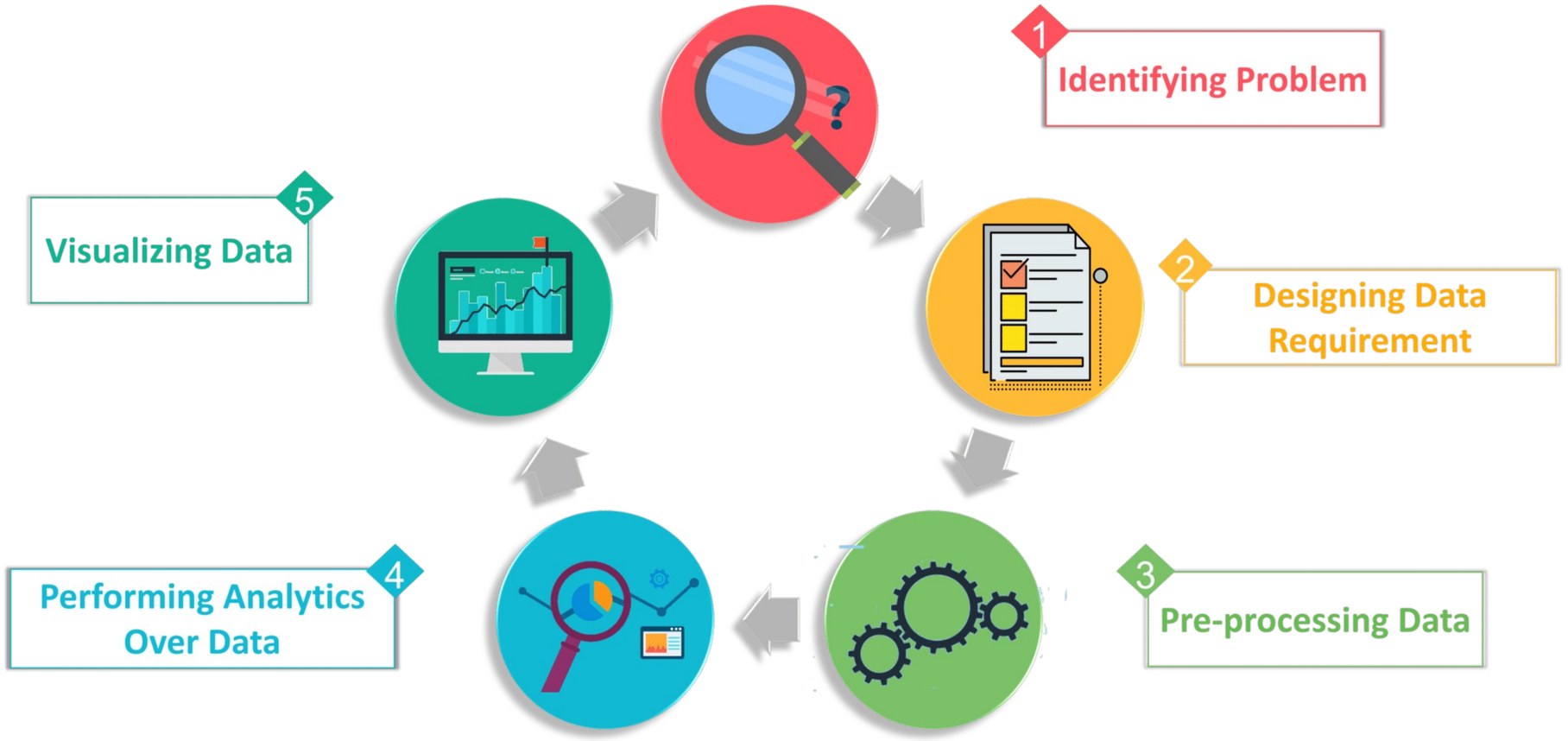
Sobre o que vamos conversar?

- Conceitos;
- Arquiteturas de Big Data;
- Python;
- Apache Kafka;
- ElasticSeach;
- Análise de Dados em “Tempo Real”.

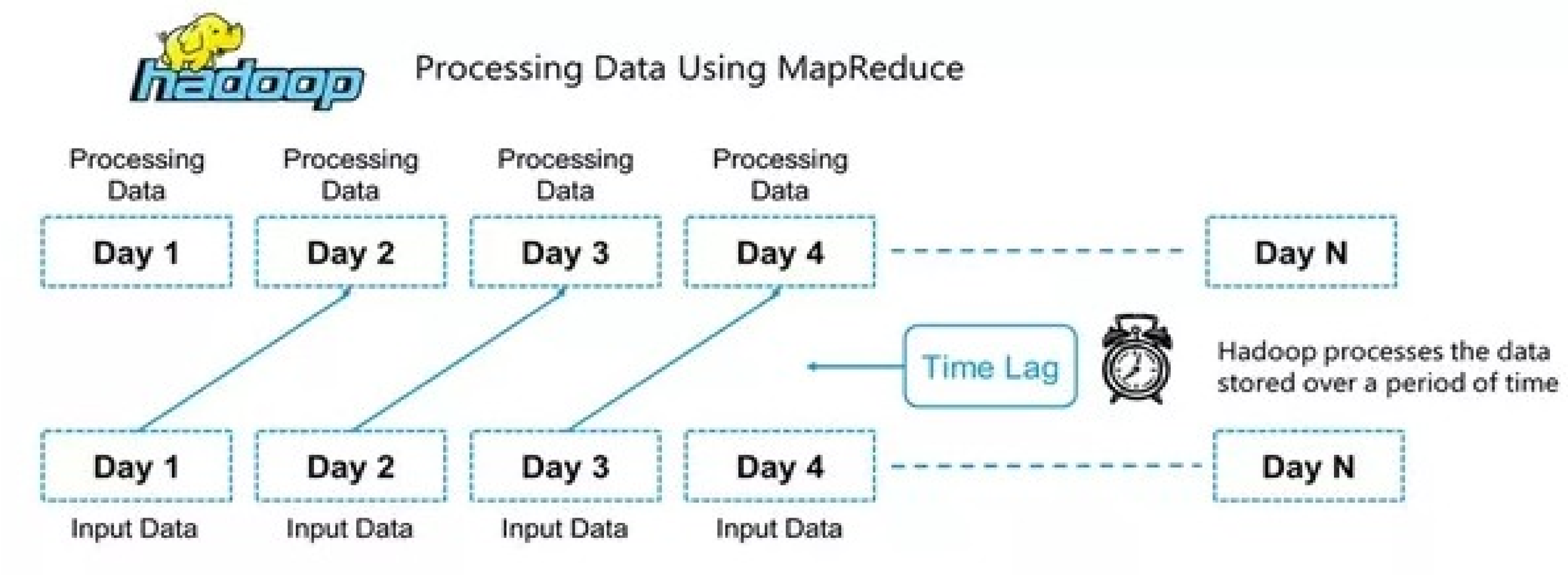
Big Data?

Conceitos

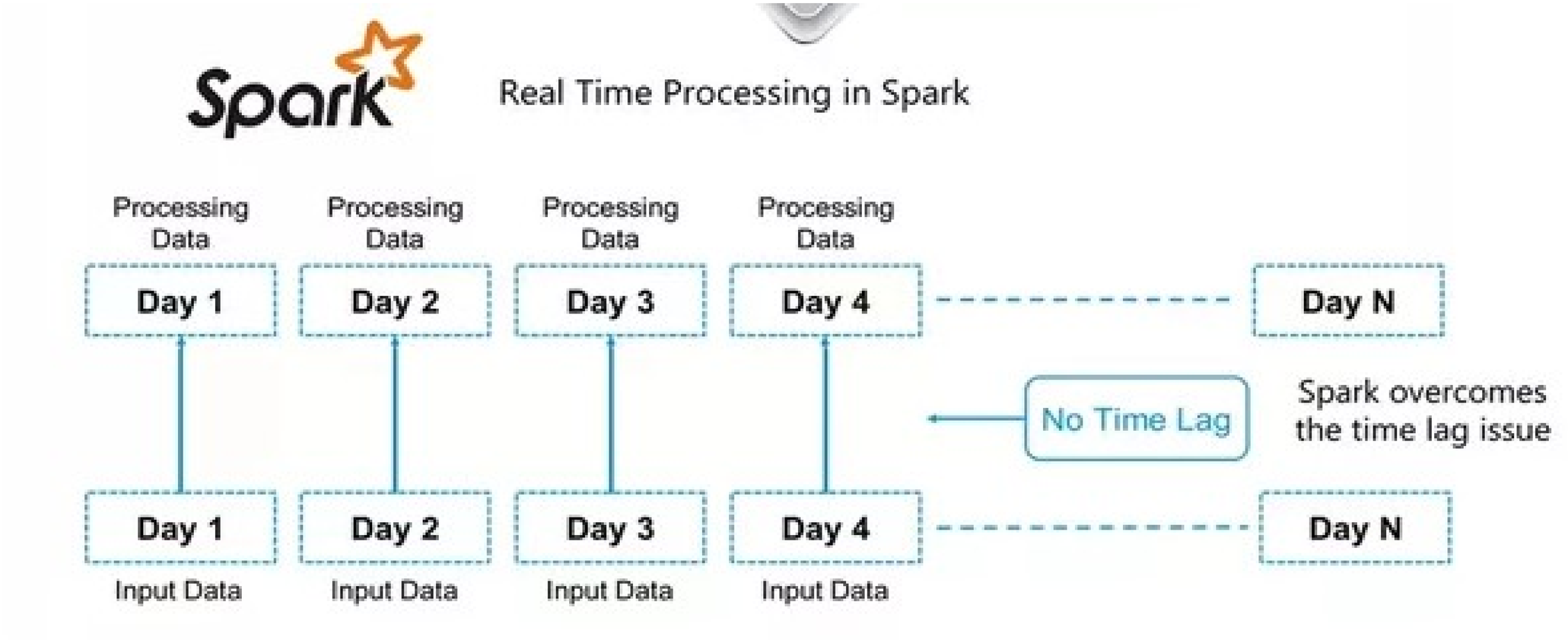
Processamento e Análise de Dados



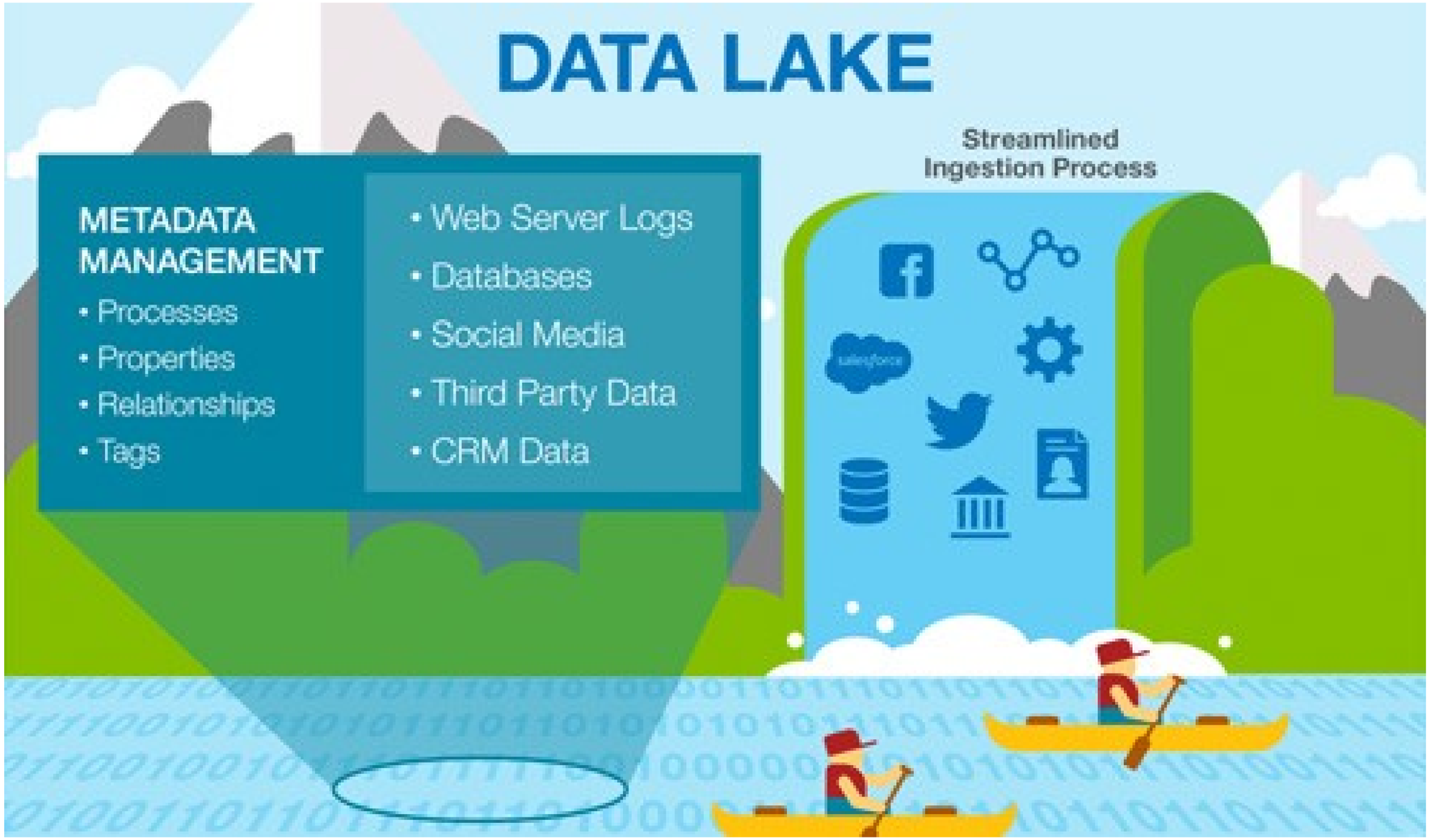
Processamento por Batch



Processamento por Streaming

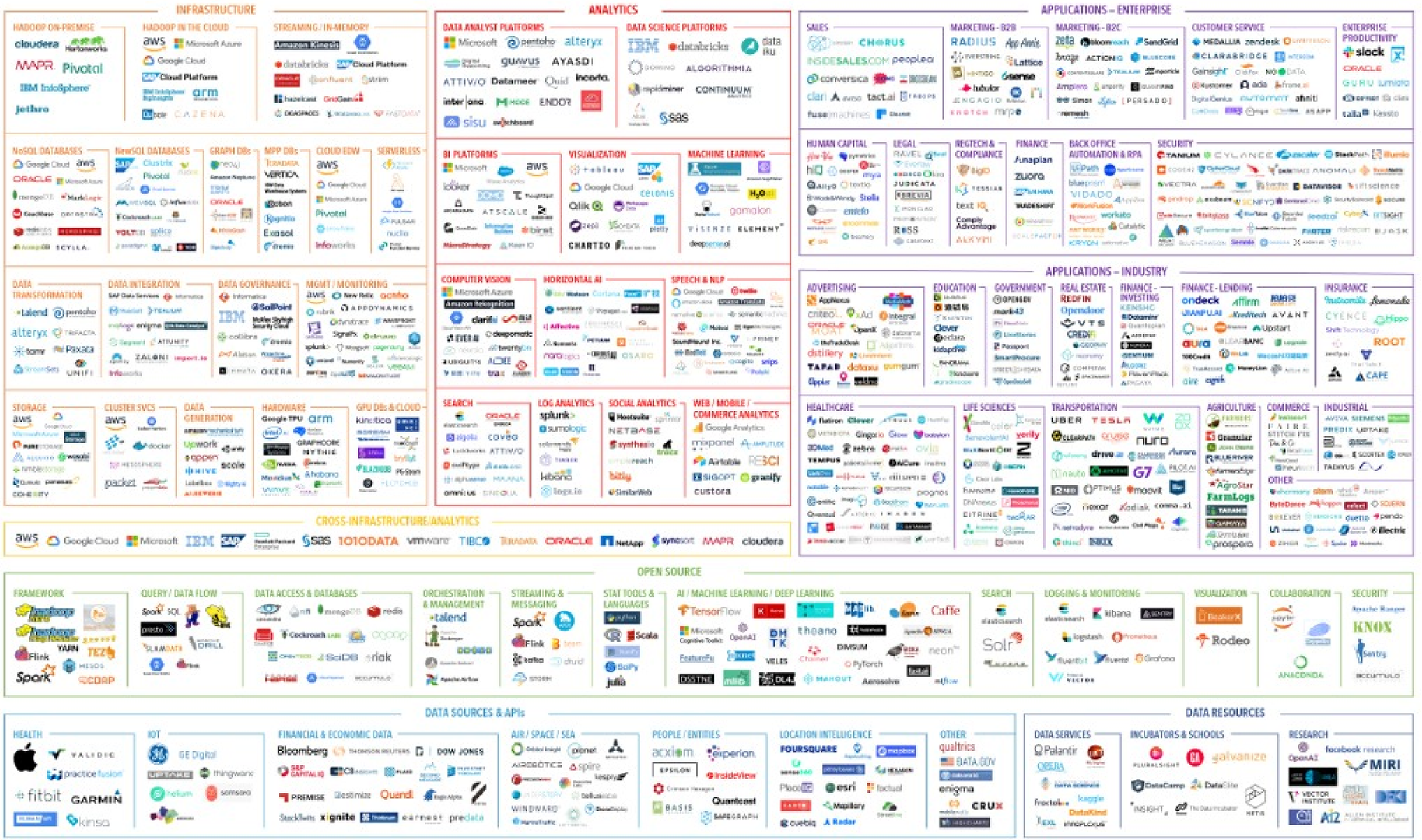


Data Lake

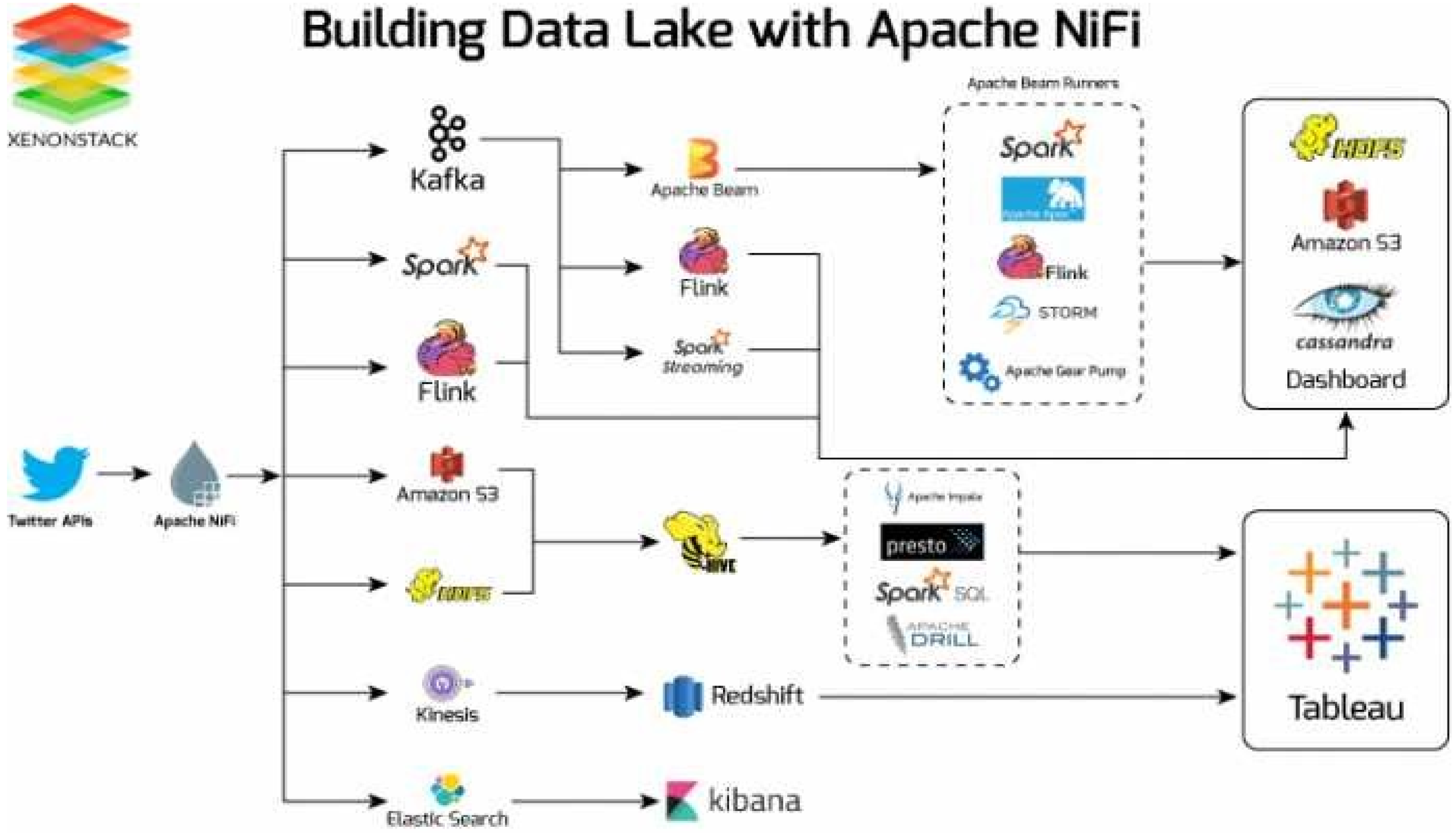

DATA LAKE
Streamlined
Ingestion Process
METADATA
MANAGEMENT
• Processes
• Properties
• Relationships
• Tags
• Web Server Logs
• Databases
• Social Media
• Third Party Data
• CRM Data

E como fazer tudo isso?

DATA & AI LANDSCAPE 2019



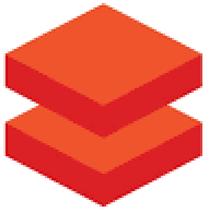
Algumas Ferramentas



Plataformas em nuvem

Google Cloud

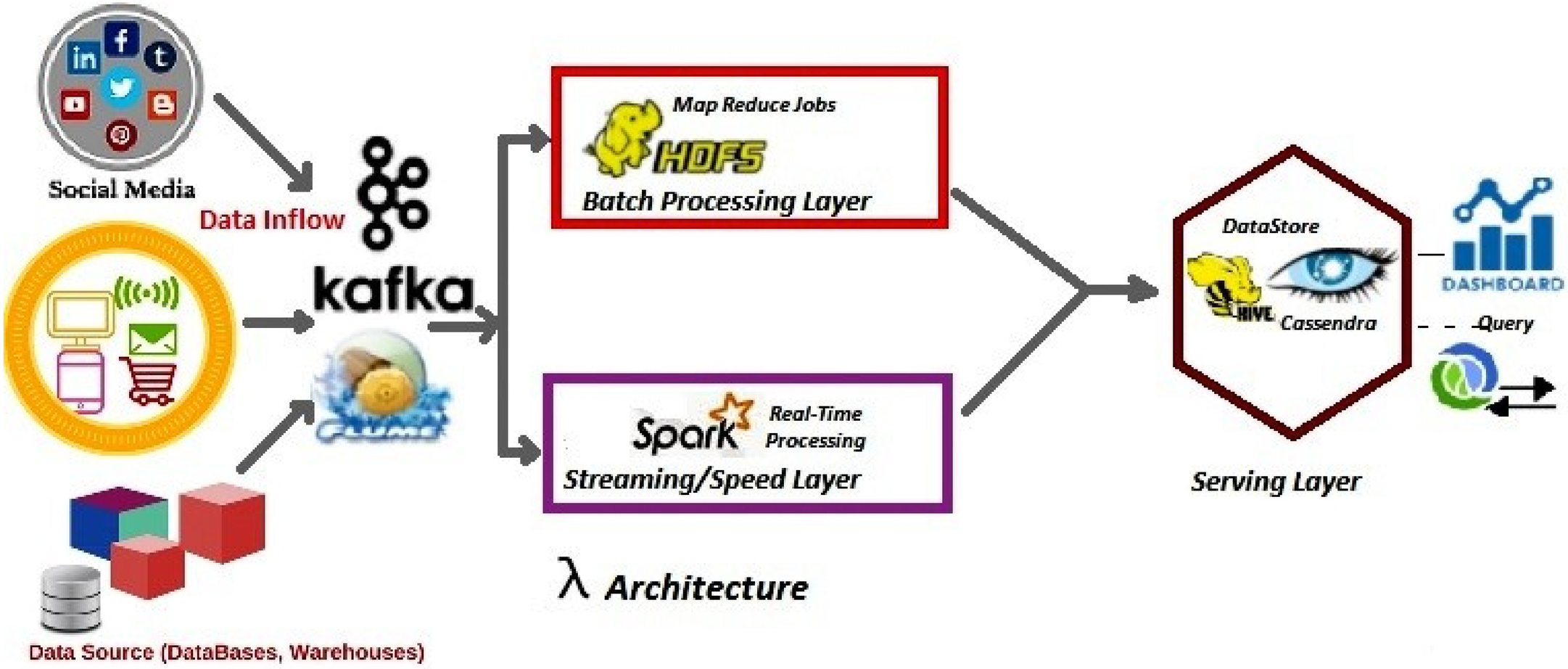
aws



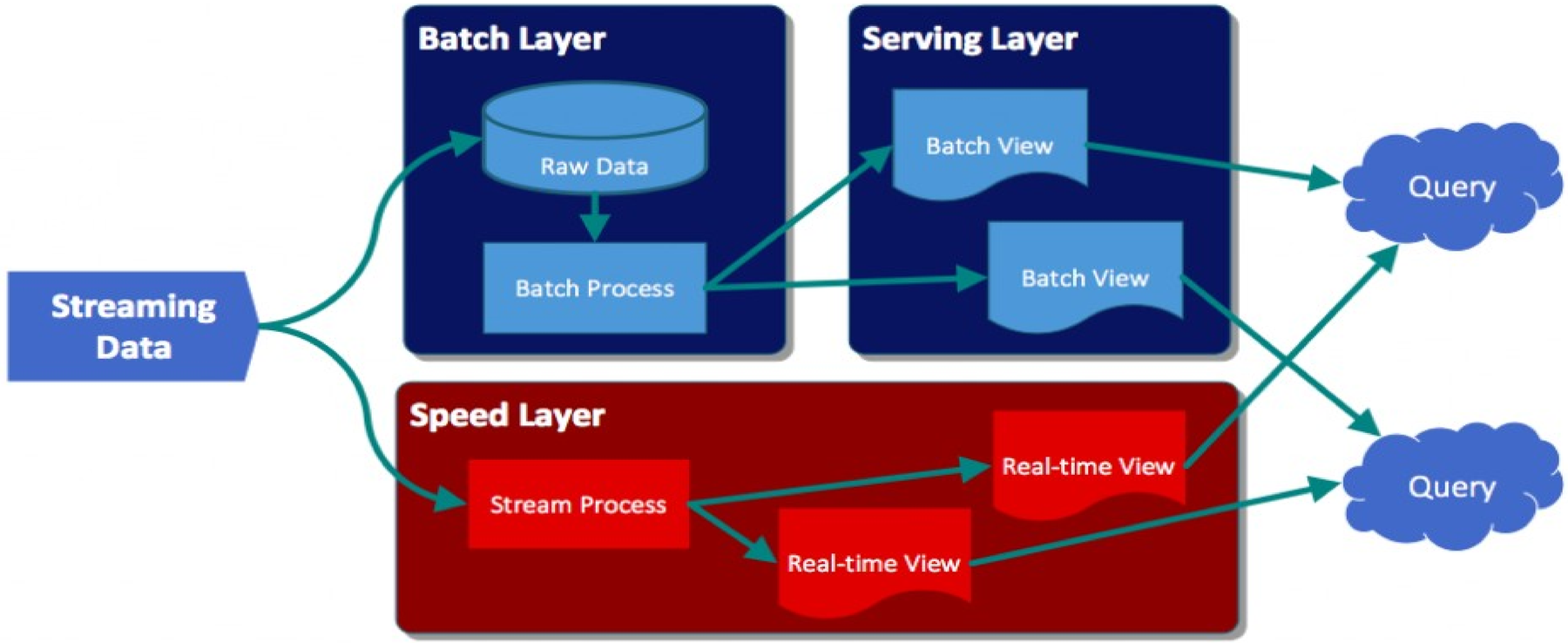
databricks

Arquiteturas

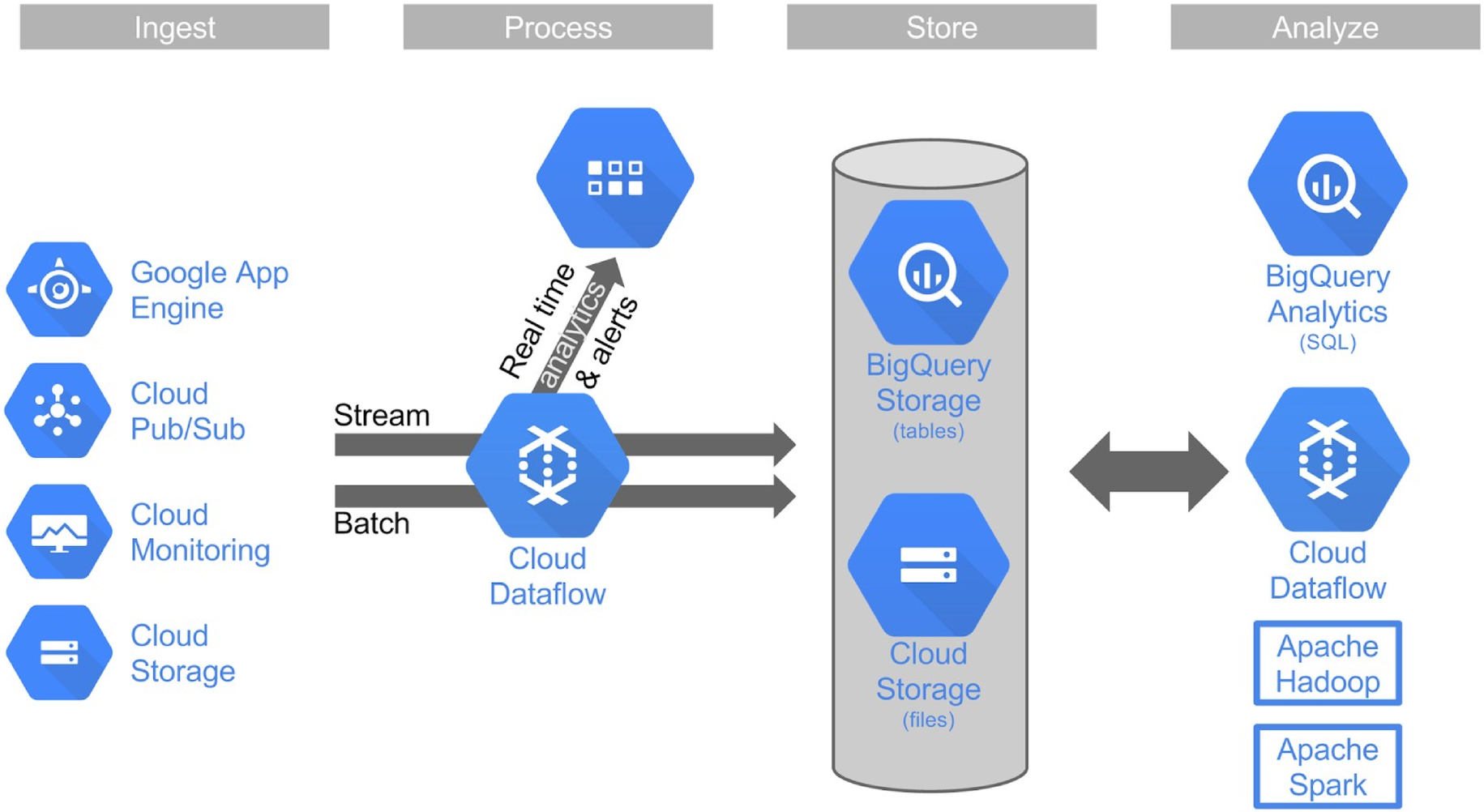
Lambda


Social Media
Data Inflow
kafka
Map Reduce Jobs
HDFS
Batch Processing Layer
Spark
Real-Time Processing
Streaming/Speed Layer
λ Architecture
DataStore
HIVE
Cassendra
DASHBOARD
Query
Serving Layer
Data Source (DataBases, Warehouses)

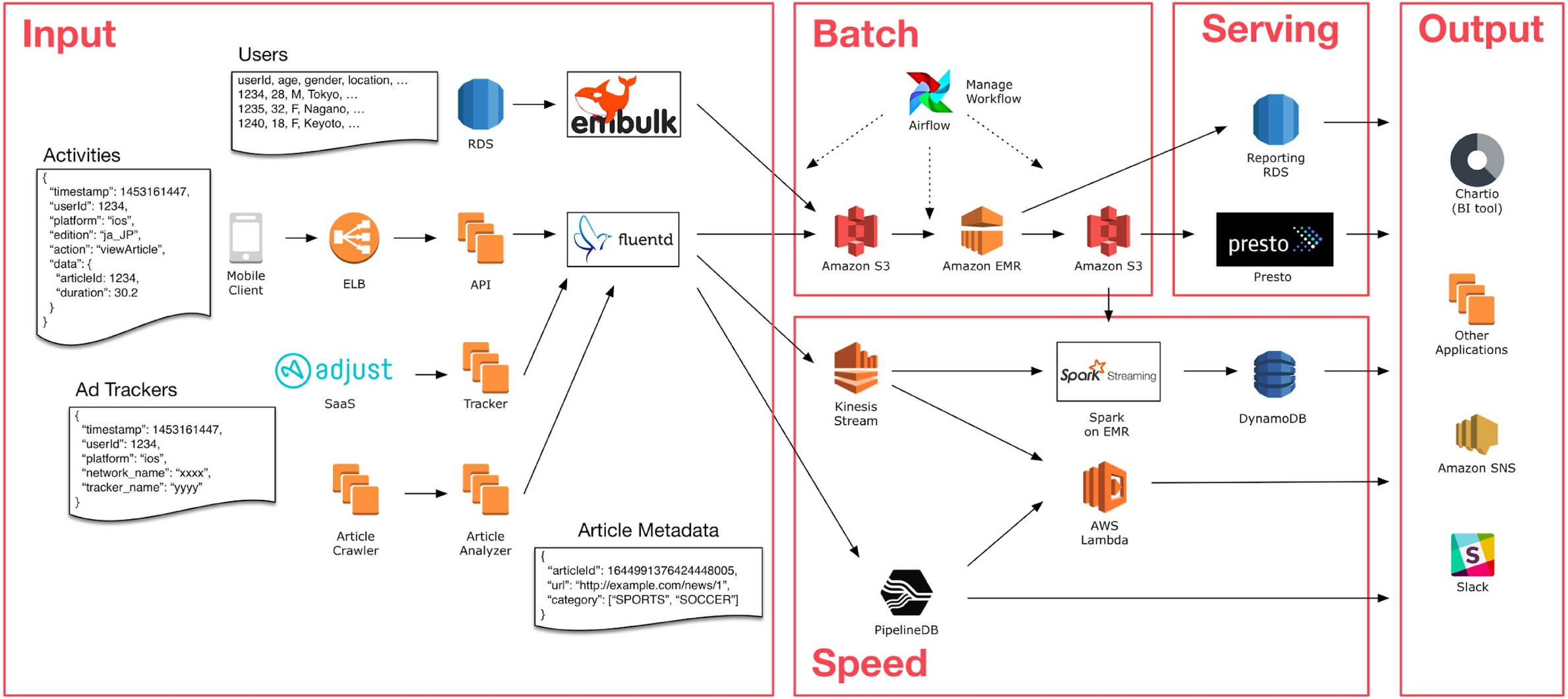
Kappa


Batch Layer
Raw Data
Batch Process
Serving Layer
Batch View
Batch View
Streaming Data
Speed Layer
Stream Process
Real-time View
Real-time View
Query
Query

Plataforma Google Cloud

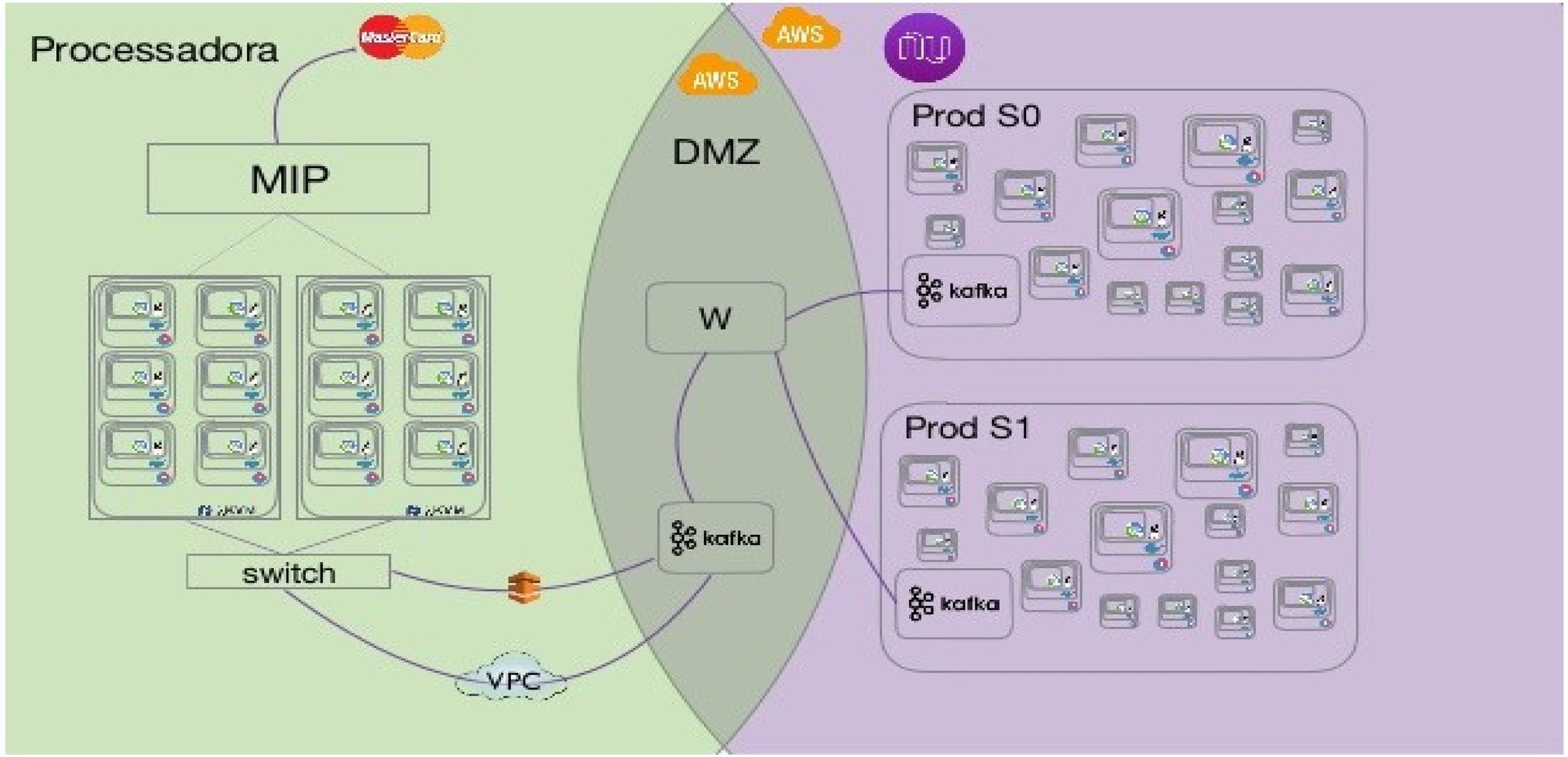


Plataforma AWS

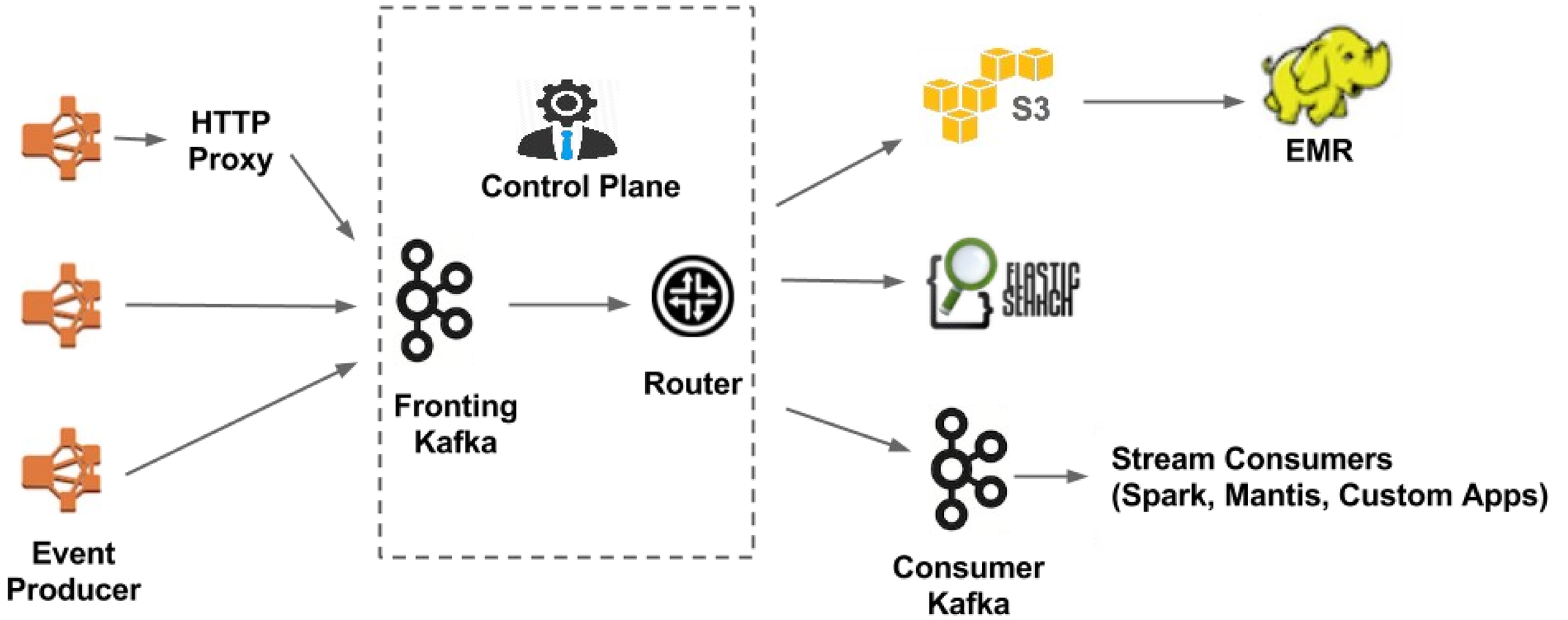


Cases

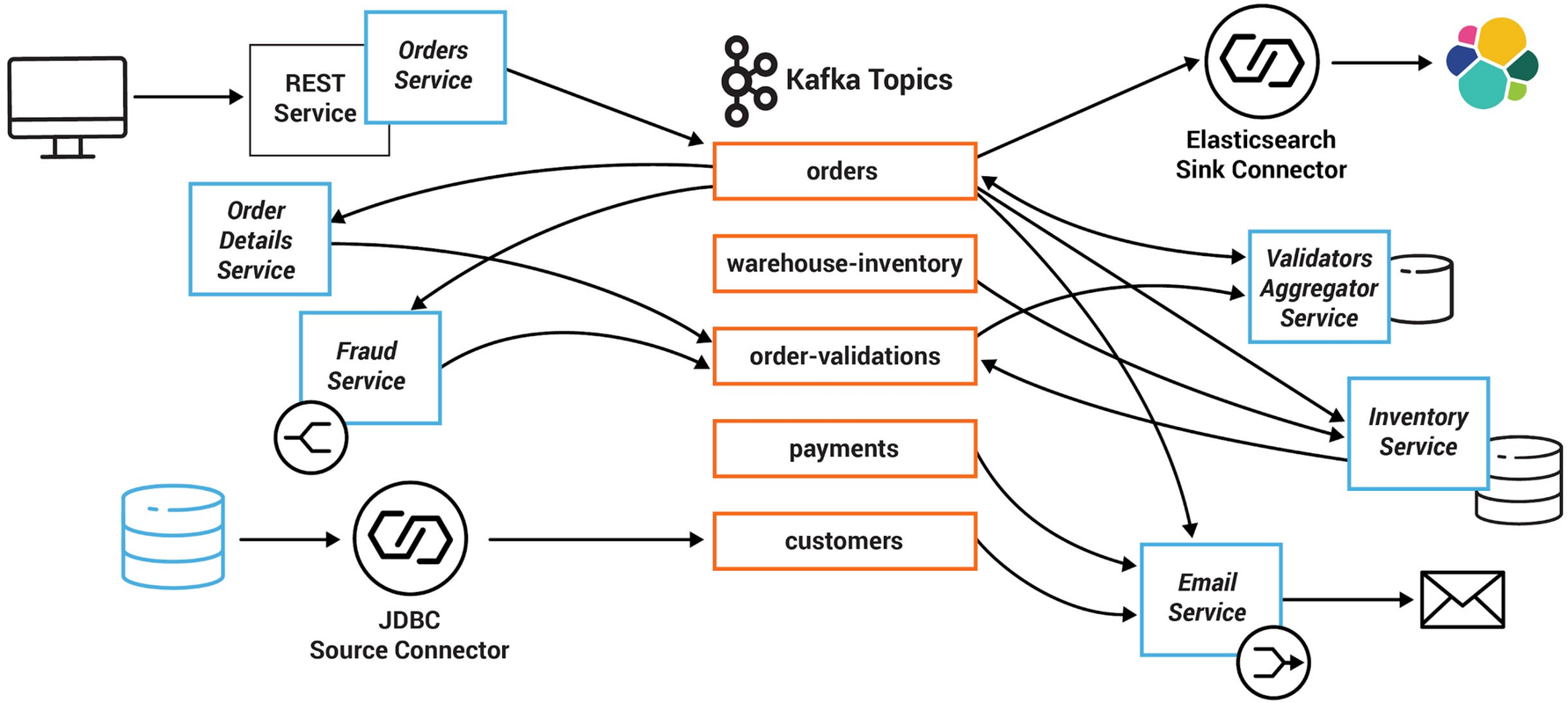
Pagamentos Nubank



Pipeline Netflix



Pipeline com Microserviços

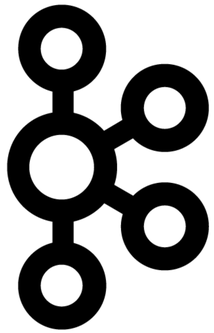


Hands-on

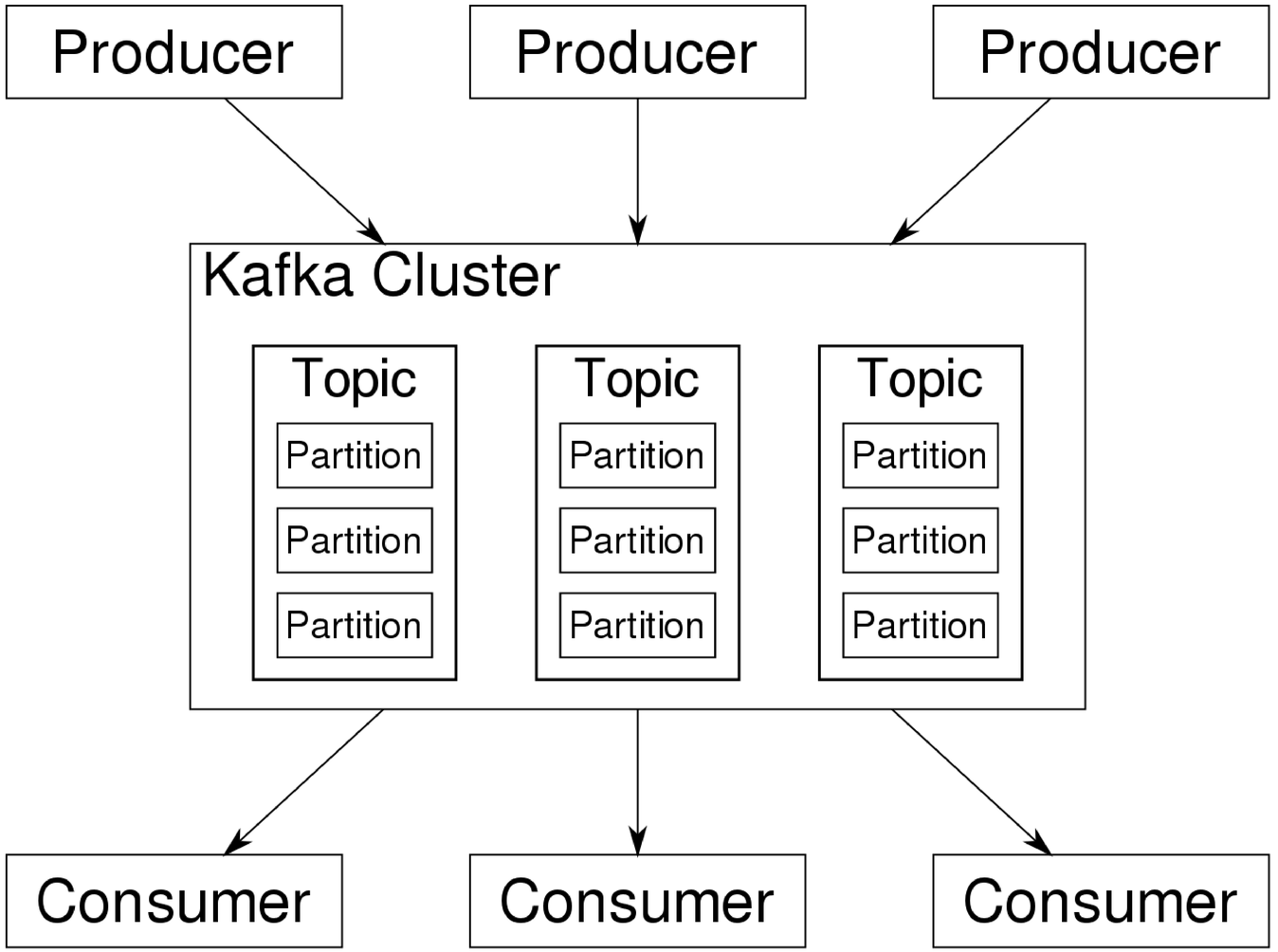
Ferramentas

Apache Kafka: Conceitos

- Topics e partitions;
  Brokes;
  Producers;
  Consumers;
  Zookeper;
  Kafka Connectors.



Apache Kafka: Conceitos



ElasticSearch

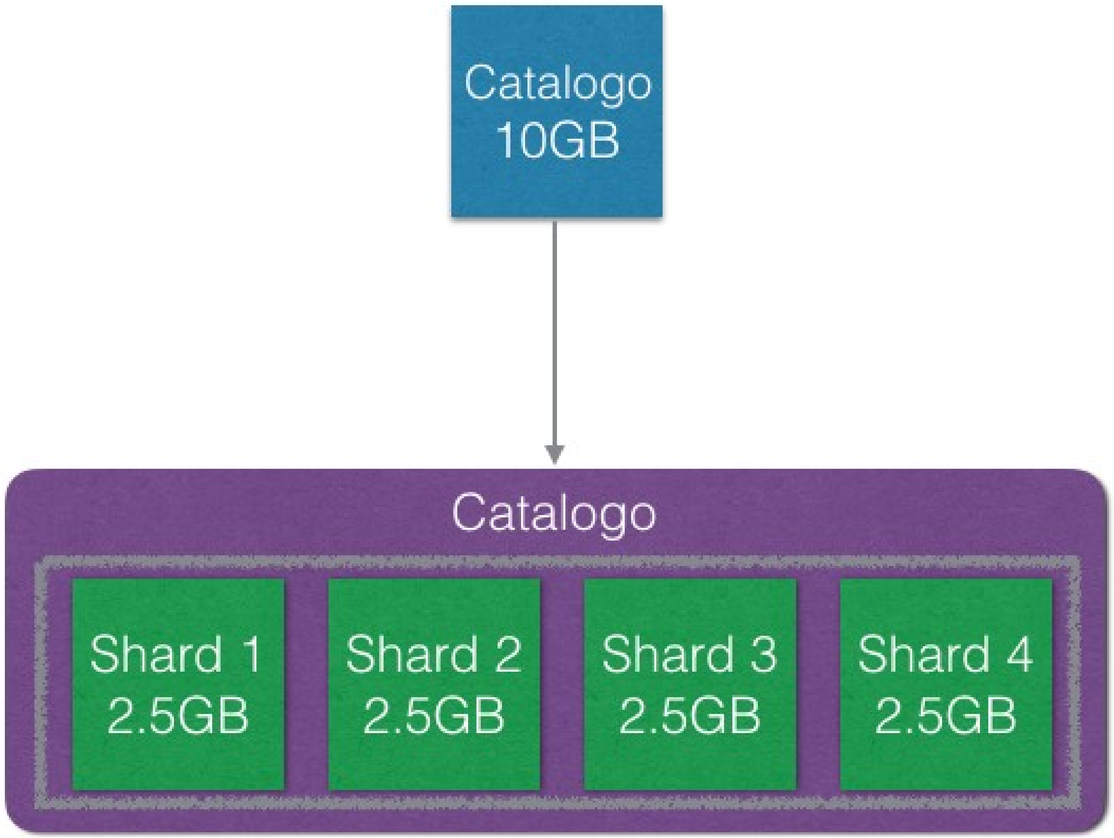
- Full text search;
  Ferramenta de buscas opensource;
  Escável e tolerante a falhas;
  Possui API’s;
  Banco no formato JSON.

ElasticSearch: Conceitos

Banco Relacional	ElasticSearch
Instância	Index
Tabela	Type
Schema	Mapping
Tupla	Documento
Coluna	Atributo

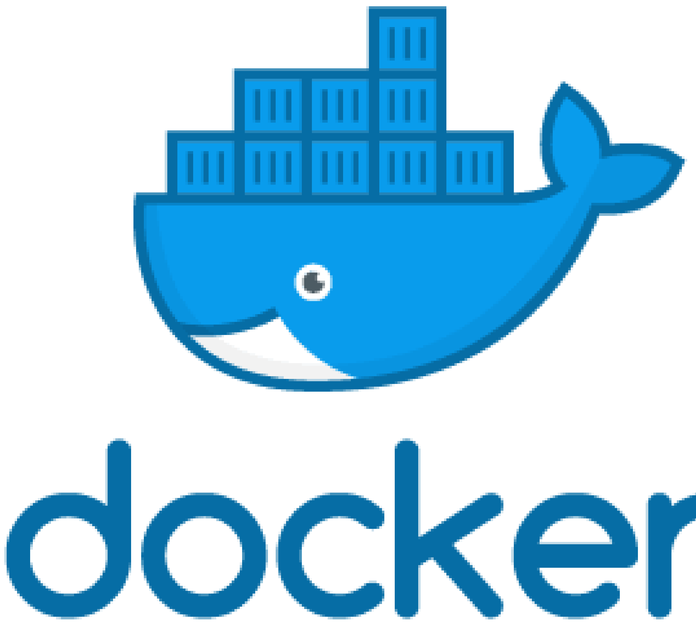
ElasticSearch: Conceitos

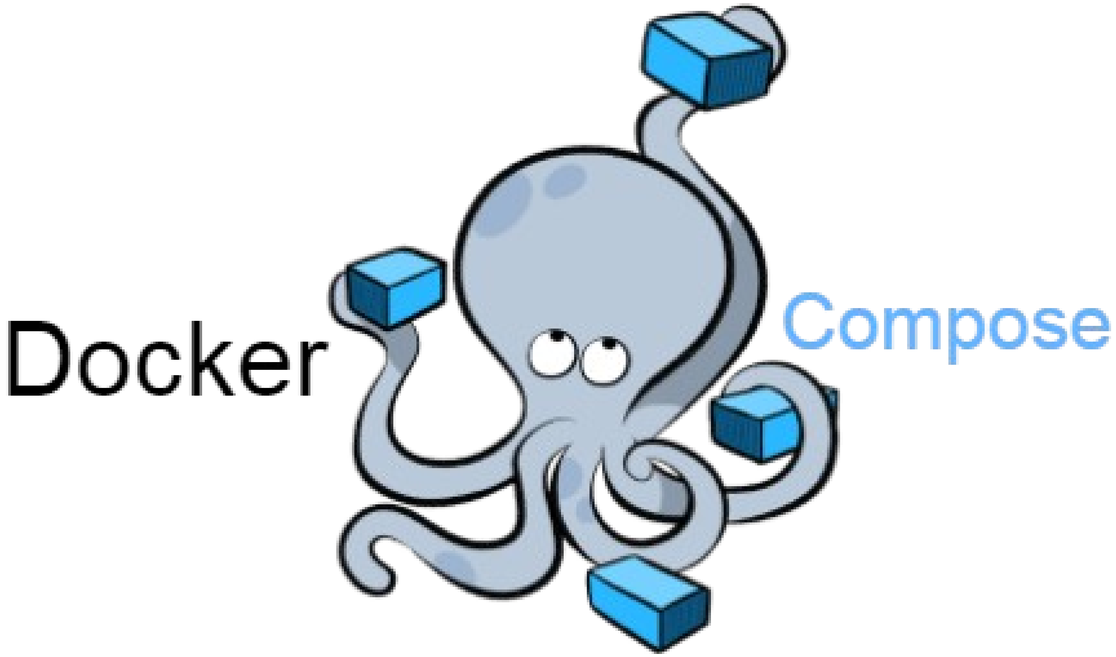
- Cluster;
  Réplicas.



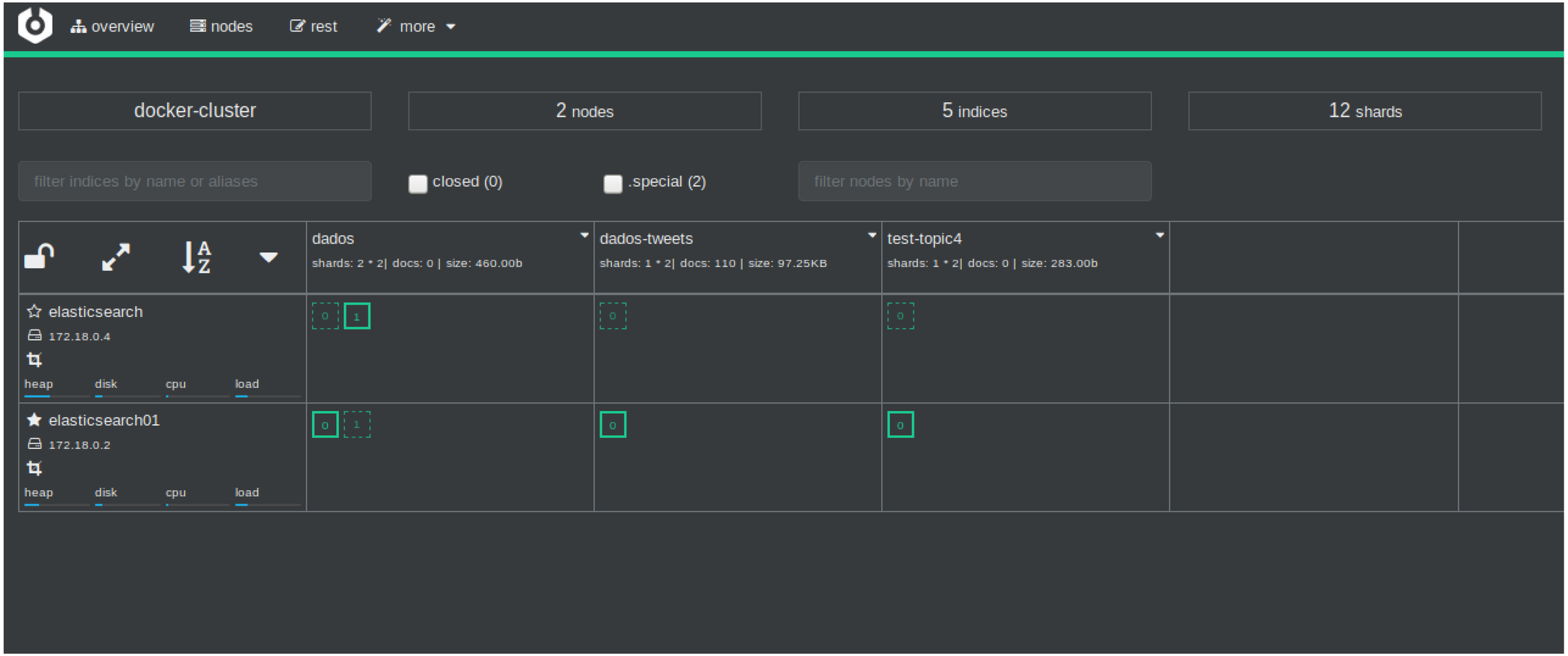
Show me the code!

Ambiente com Docker e Docker Compose

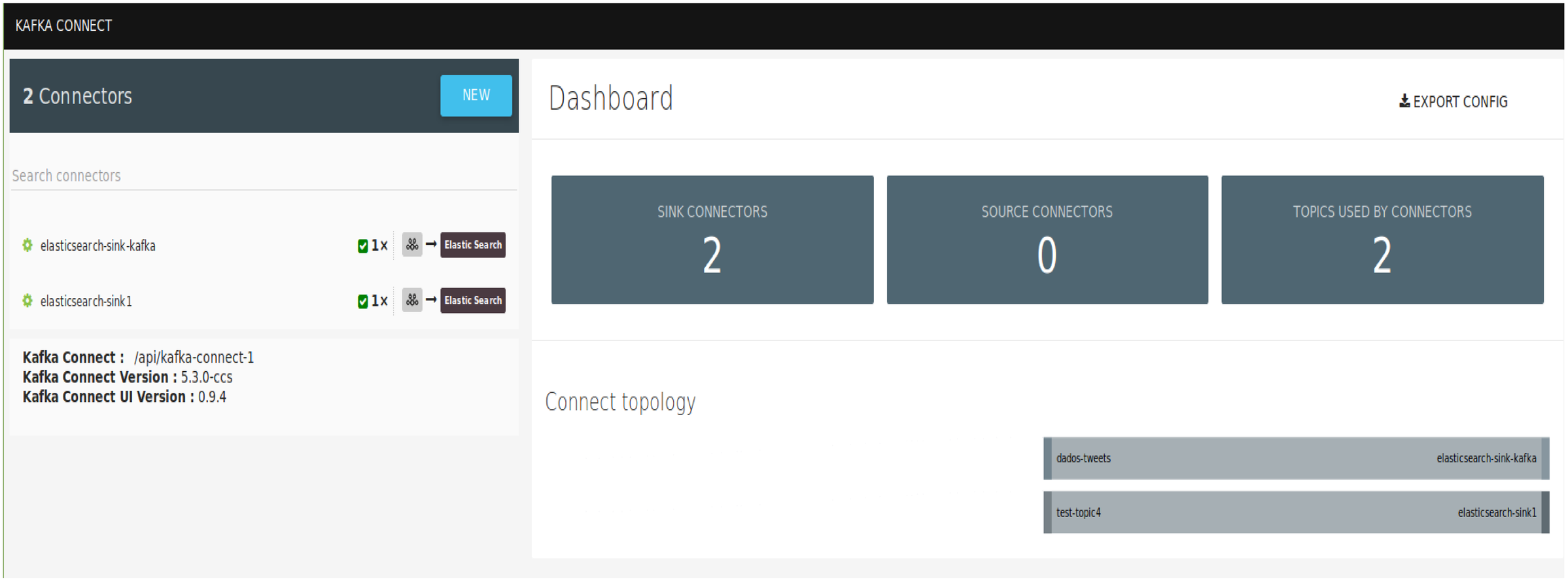




Cluster Elastic Search + Cerebro



Kafka + Connector + Connector UI



Kafka Connector

```
{
  "connector.class": "io.confluent.connect.elasticsearch.ElasticsearchSinkConnector",
  "type.name": "kafka-connect",
  "key.converter.schemas.enable": "false",
  "tasks.max": "1",
  "topics": "dados-tweets",
  "name": "elasticsearch-sink-kafka",
  "value.converter.schemas.enable": "false",
  "key.ignore": "true",
  "connection.url": " http://10.158.0.5:9200",
  "value.converter": "org.apache.kafka.connect.json.JsonConverter",
  "key.converter": "org.apache.kafka.connect.storage.StringConverter",
  "schema.ignore": "true"
}
```

Jupyter Notebook

In [59]:
```
#geração da nuvem de palavras em tempo real
frases = ''
for messagem in consumer:
    texto = json.loads(messagem.value.decode('utf-8'))
    frases = frases + texto['tweet']
    clear_output()
    wordcloud = WordCloud(max_font_size=100, width = 1520, height = 535).generate(frases)
    plt.figure(figsize=(16,9))
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.show()
```

Código Completo da Talk

https://bit.ly/2VSW06W

Obrigado!

/cicerojmm

/in/cicero-moura

/cicerojmm

/cicerojmm