

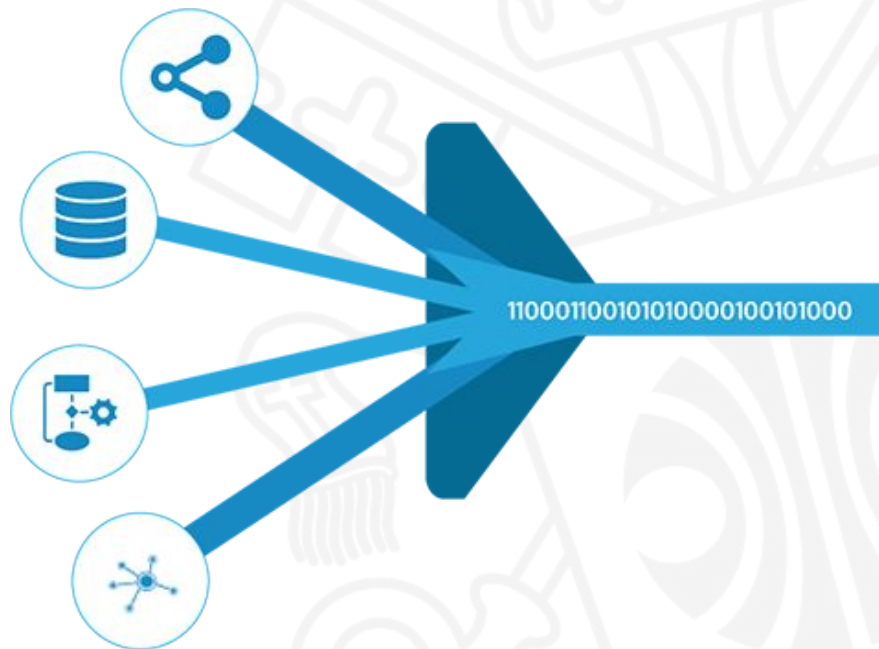
Integração de Dados

Fábio Jardim

Introdução

Ingestão de Dados

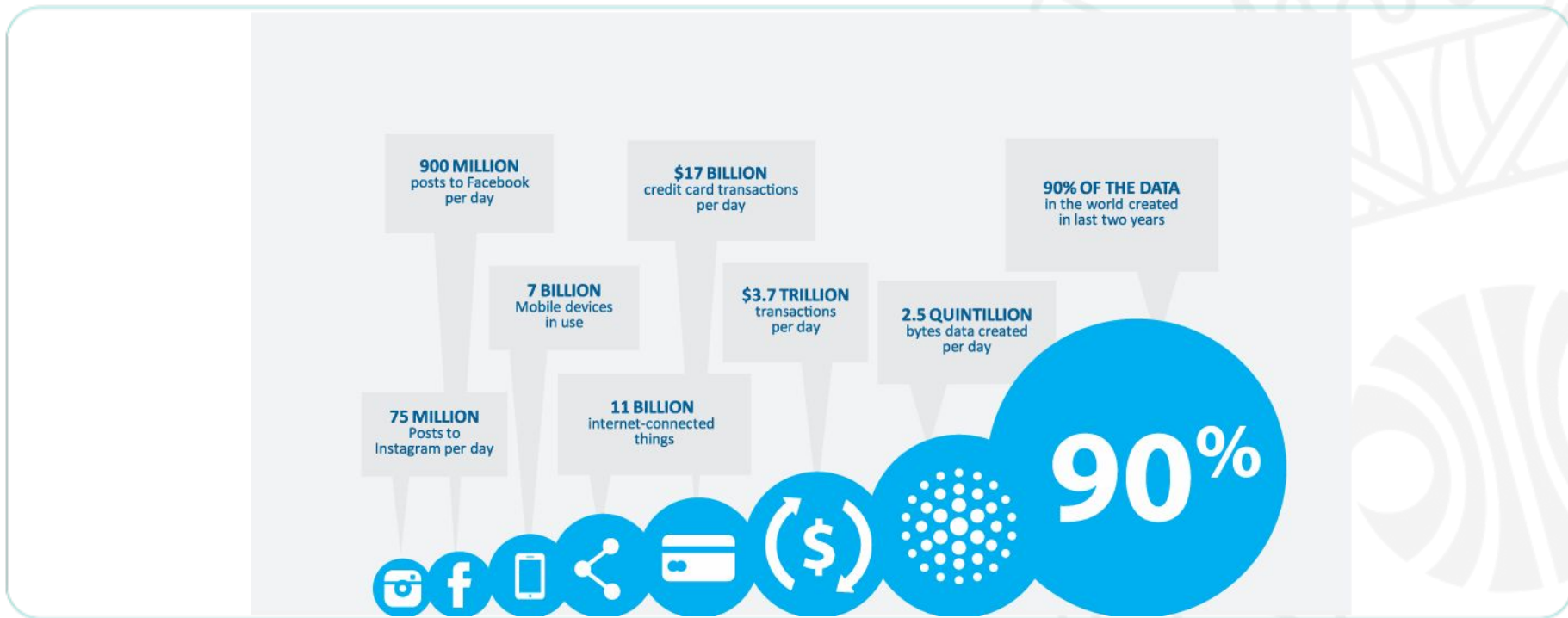
- Transporte de dados para um armazenamento centralizado
- DW, Data Mart, Data Lake, Database, etc...
- Deve suportar fontes diversas
- Camada central de um ambiente de dados
- Deve ser escalável



Velocidade dos Dados



Velocidade dos Dados



<http://aoife.dbsdatapoints.com/tag/data-evolution/>

Tipos de Dados

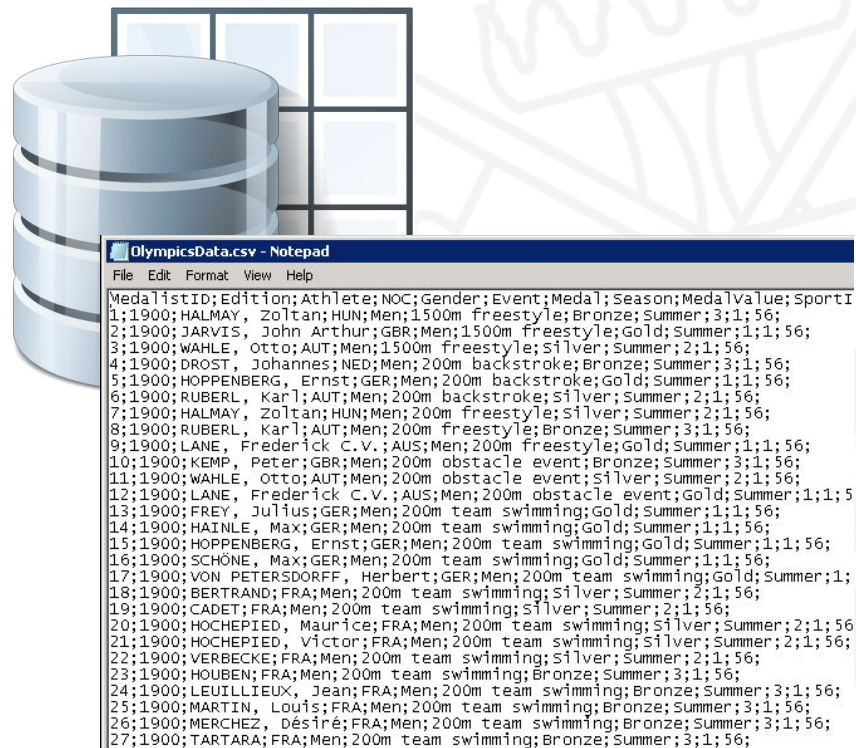
Tipos e Formatos de Dados

- Novos modelos de dados
- Text (XML, JSON, TXT, CSV)
- Vídeo e imagens
- Dados geoespaciais
- Compactados
- Serializados
- Streaming



Dados Estruturados

- Muito utilizado
- Dados organizados
- Segue uma padronização
- Abordagem linha e coluna
- Vários formatos
- Banco de dados e CSV



Dados Não Estruturados

- Maior parte dos dados atuais
- Não são em linhas e colunas
- Não possuem modelo de dados
- Mais difícil de analisar
- IA possibilita processamento mais fácil
- Vídeos, fotos, textos, etc...



Dados Semi Estruturados

- Fica entre estruturado e não estruturado
- Estrutura flexível
- Existência de “metadados”
- Organizado
- Fácil entendimento
- Muito utilizado na web. Ex:JSON

```
{  
  "orders": [  
    {  
      "orderno": "748745375",  
      "date": "June 30, 2088 1:54:23 AM",  
      "trackingno": "TN0039291",  
      "custid": "11045",  
      "customer": [  
        {  
          "custid": "11045",  
          "fname": "Sue",  
          "lname": "Hatfield",  
          "address": "1409 Silver Street",  
          "city": "Ashland",  
          "state": "NE",  
          "zip": "68003"  
        }  
      ]  
    }  
  ]  
}
```

Fontes de Dados

Fontes de Dados

- Dados internos
- Dados de parceiros
- Dados externos
- Social
- Dados públicos
- IoT



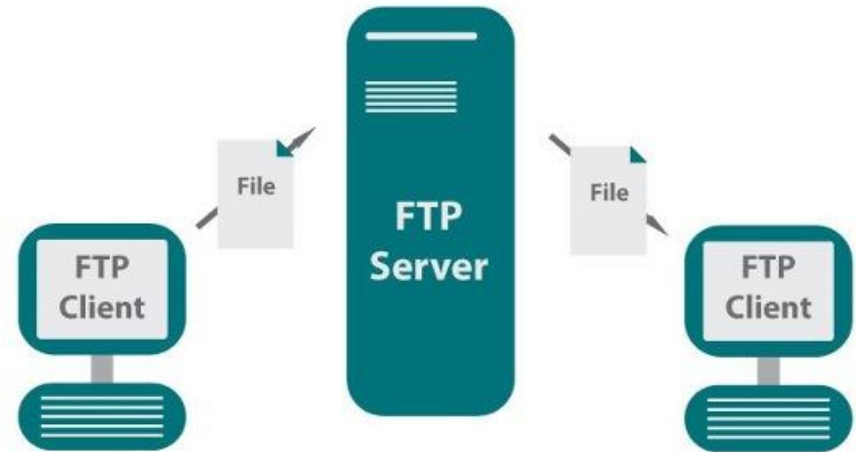
API

- Rest API
- Representational State Transfer
- Integração por rede
- Requisição HTTP
- Baixa curva de aprendizagem
- Informações de uma ferramenta para outra

{ REST }

FTP

- File Transfer Protocol
- Ainda muito utilizado
- Transferência de arquivos
- Geralmente arquivo texto
- Seguro (SFTP)
- Legado
- Interno ou externo



File Server

- Arquivos na rede
- Interno
- Compartilhamento temporário
- Abordagem antiga
- ERP > File Server
- Arquivo texto



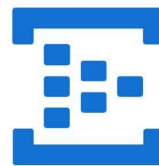
Database

- Banco de dados relacional
- NoSQL
- Batch ou Real Time
- Várias abordagens
- Fácil integração
- Pode ter overhead



Mensageria

- Arquitetura orientada a eventos
- Assíncrono
- Real Time
- IoT
- Data Flow
- Vários formatos
- Frameworks diversos



Azure Event Hubs



Google Cloud
Pub/Sub



Amazon Kinesis

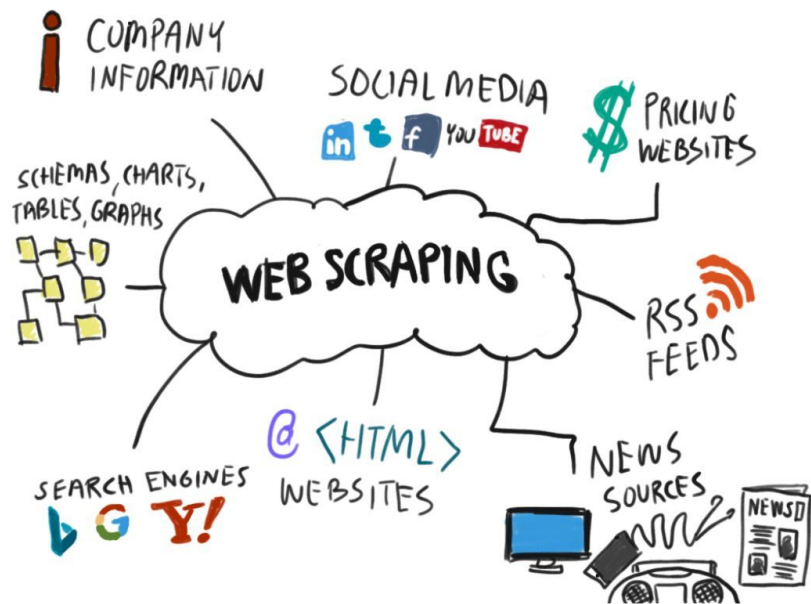


APACHE
ACTIVEMQ



Dados Semi Estruturados

- Crawling, coleta de dados na Web
- Técnica de extração de dados, “Raspagem” de sites
- Ferramenta automatizadas
- Simulação de navegação
- Prática polêmica
- CAPTCHA veio para dificultar



Tipos de Ingestão

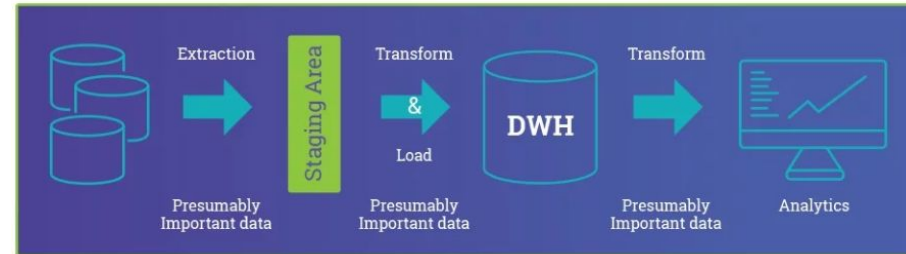
Ingestão Batch

- Dados estáticos
- Cargas agendadas
- Grande fluxo de dados
- Arquivos ou banco de dados
- Muito utilizado para BI
- Também utilizado em Big Data



ETL

- Extract, Transform e Load
- Amplamente difundido
- Abordagem tradicional
- Fácil utilização
- Cargas batch
- Processamento massivo
- Pouco flexível depois do Load



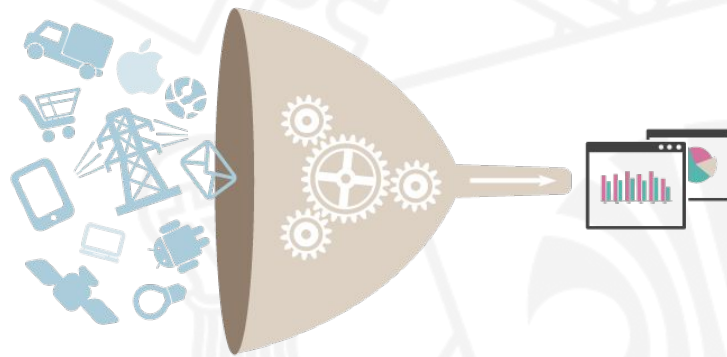
Fontes de Dados

- Mudança do fluxo de dados
- Maior flexibilidade
- Dados brutos
- Alternativa moderna ao ETL
- Diminuição de tempo de carga



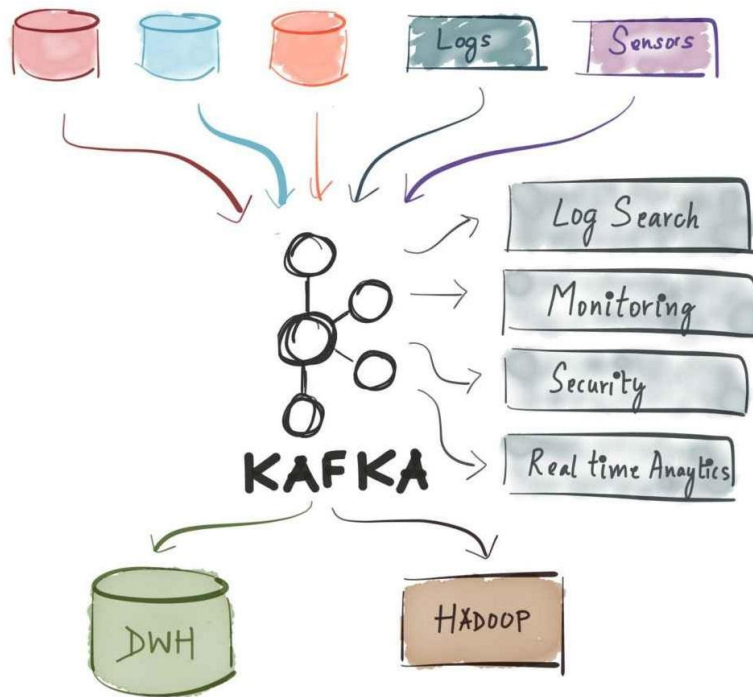
Ingestão Real Time

- Dados em tempo real
- Permite tomada de decisão rápida
- Fluxo de dados contínuo
- Espalhados em um período de tempo
- Streaming
- Orientado a eventos

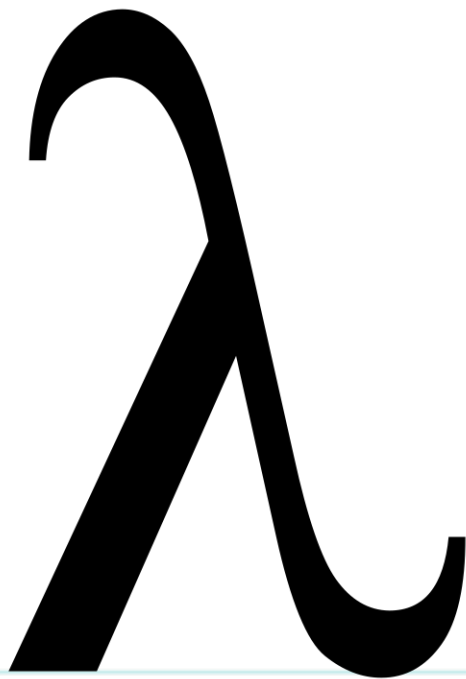


Arquitetura Baseada em Eventos

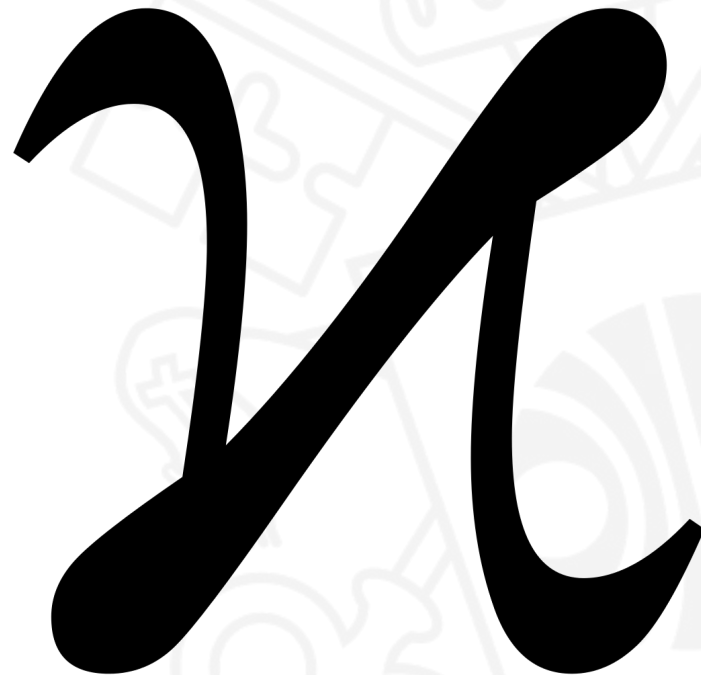
- Todo dado se torna um evento/mensagem
- Assíncrono
- Distribuição de dados
- Sistema de mensageria
- Altamente escalável
- Mudança de paradigma



Arquiteturas de Ingestão

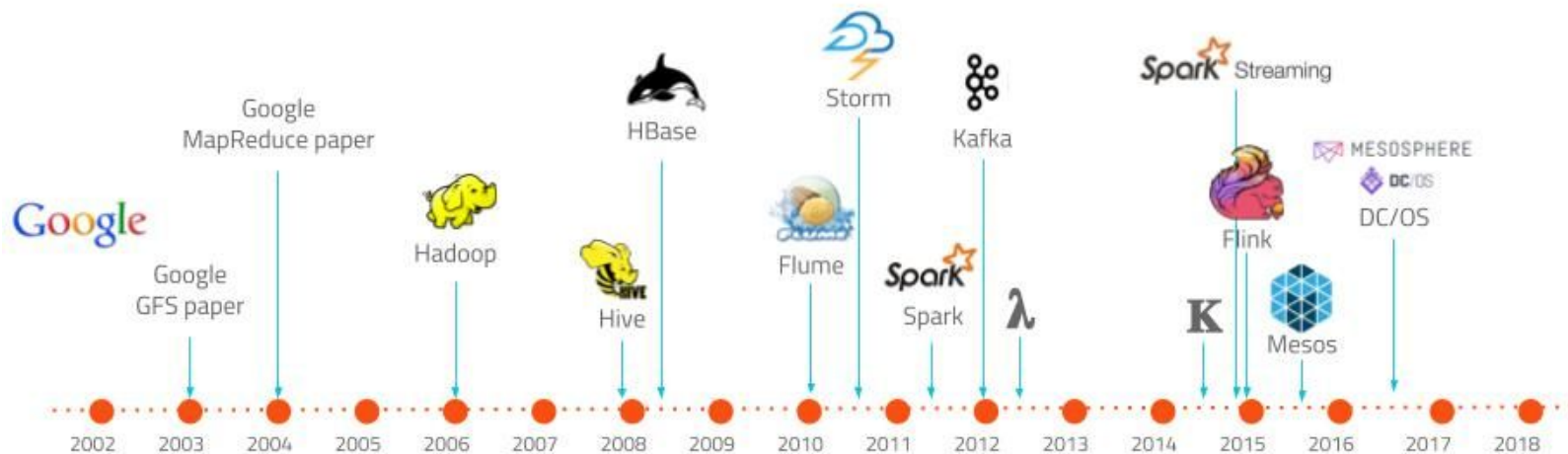


A large, bold, black stylized lambda symbol (λ) is centered within a light blue rounded rectangular frame. The symbol has a thick, flowing, cursive-like appearance.



A large, bold, black stylized xi symbol (ξ) is centered within a light blue rounded rectangular frame. The symbol has a thick, flowing, cursive-like appearance, similar to the lambda symbol but with a more complex, multi-stroke structure.

Evolução das Arquiteturas de Ingestão

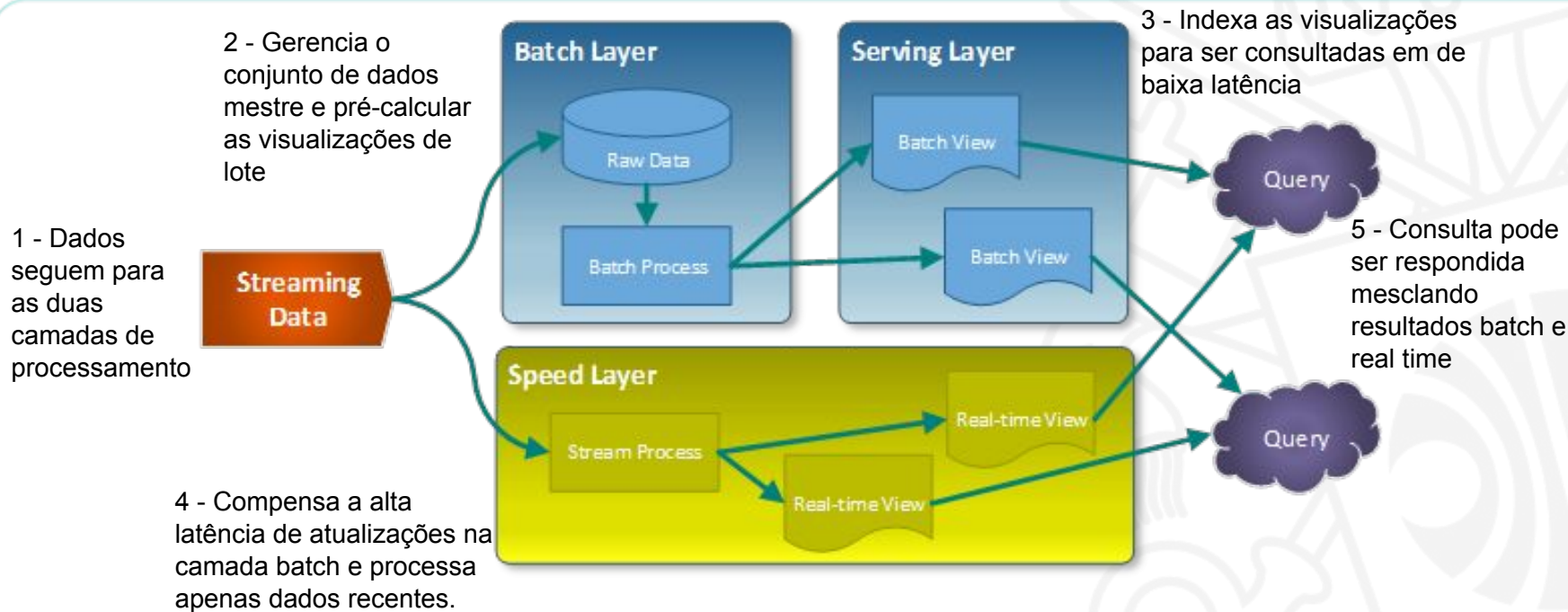


Arquitetura Lambda

- Batch e Streaming
- Tolerante a falha e escalável
- Dois fluxos de dados
- Alto custo de manutenção
- Alto custo de processamento
- Lógica duplicada

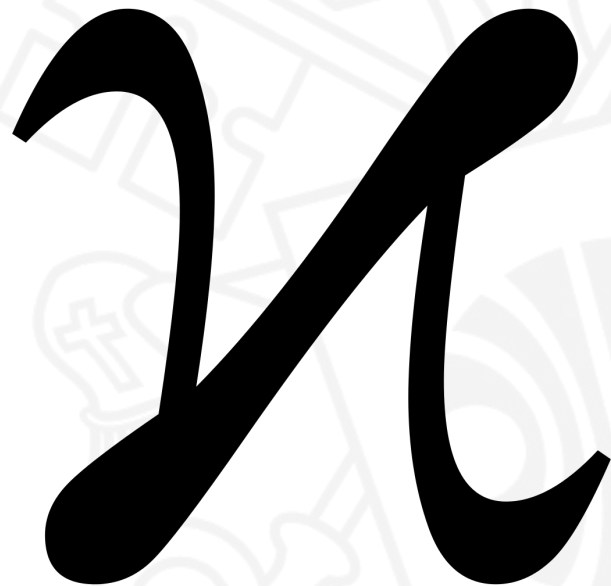


Lambda

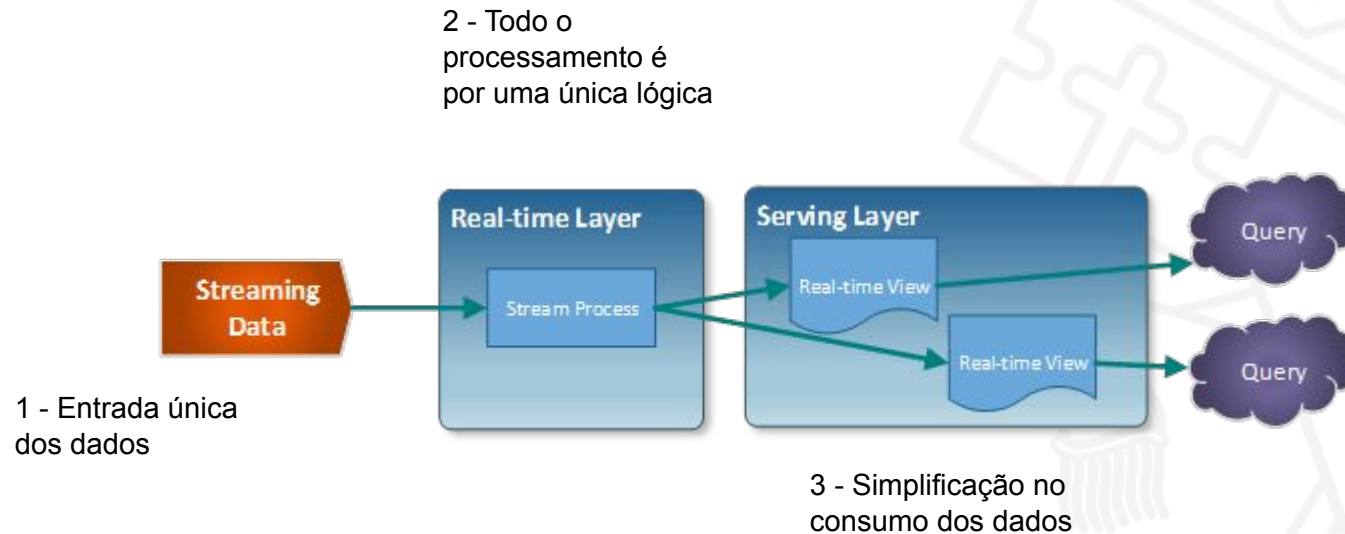


Arquitetura Kappa

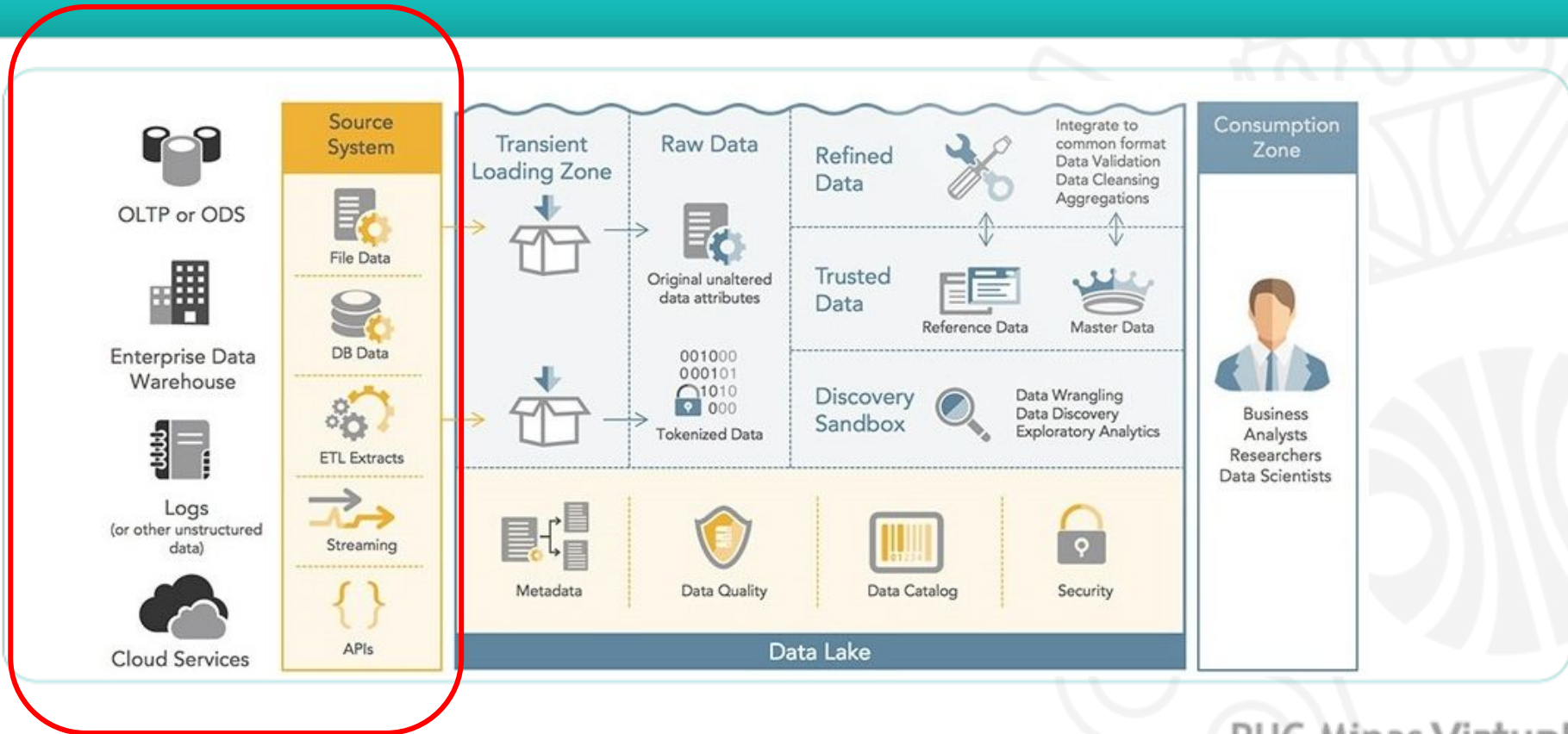
- Único fluxo de dados
- Dados históricos processados em pipeline real time
- Reprocessamento a partir da entrada de dados
- Simplificação de código, gestão e infraestrutura



Kappa



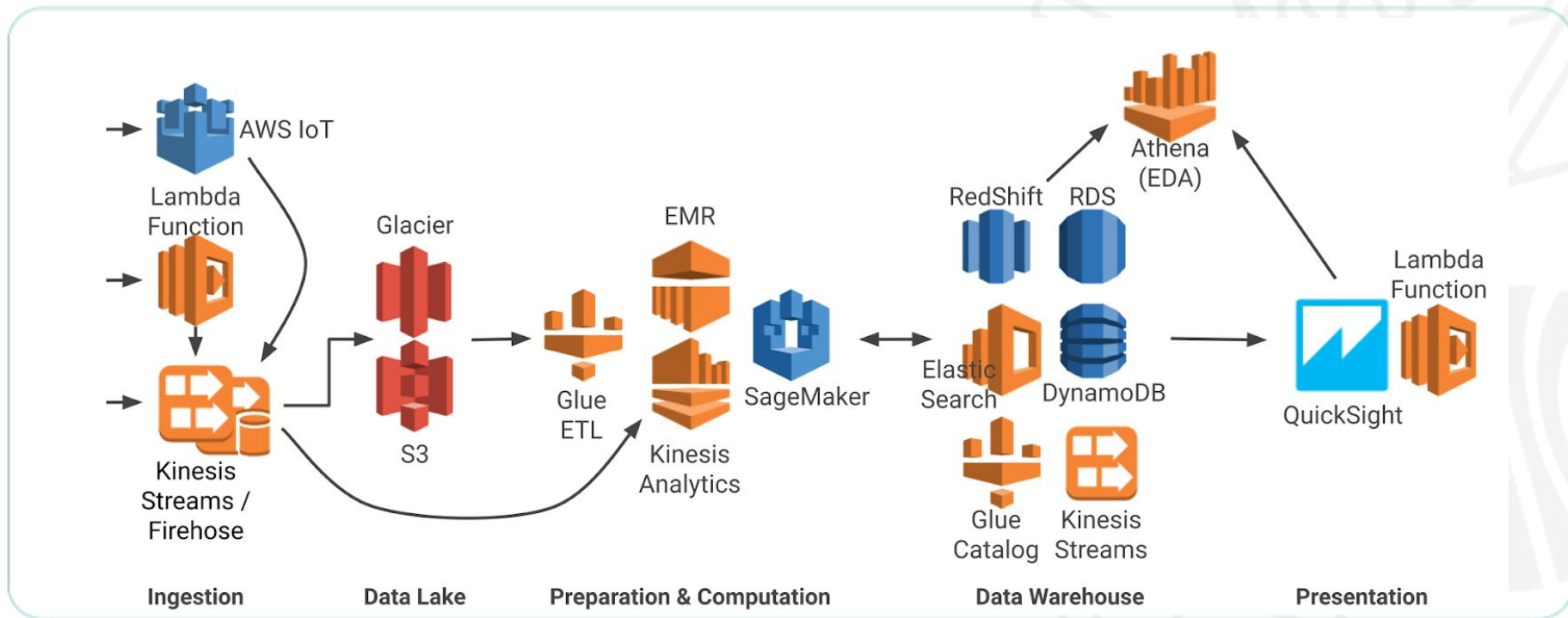
Ingestão em um Data Lake



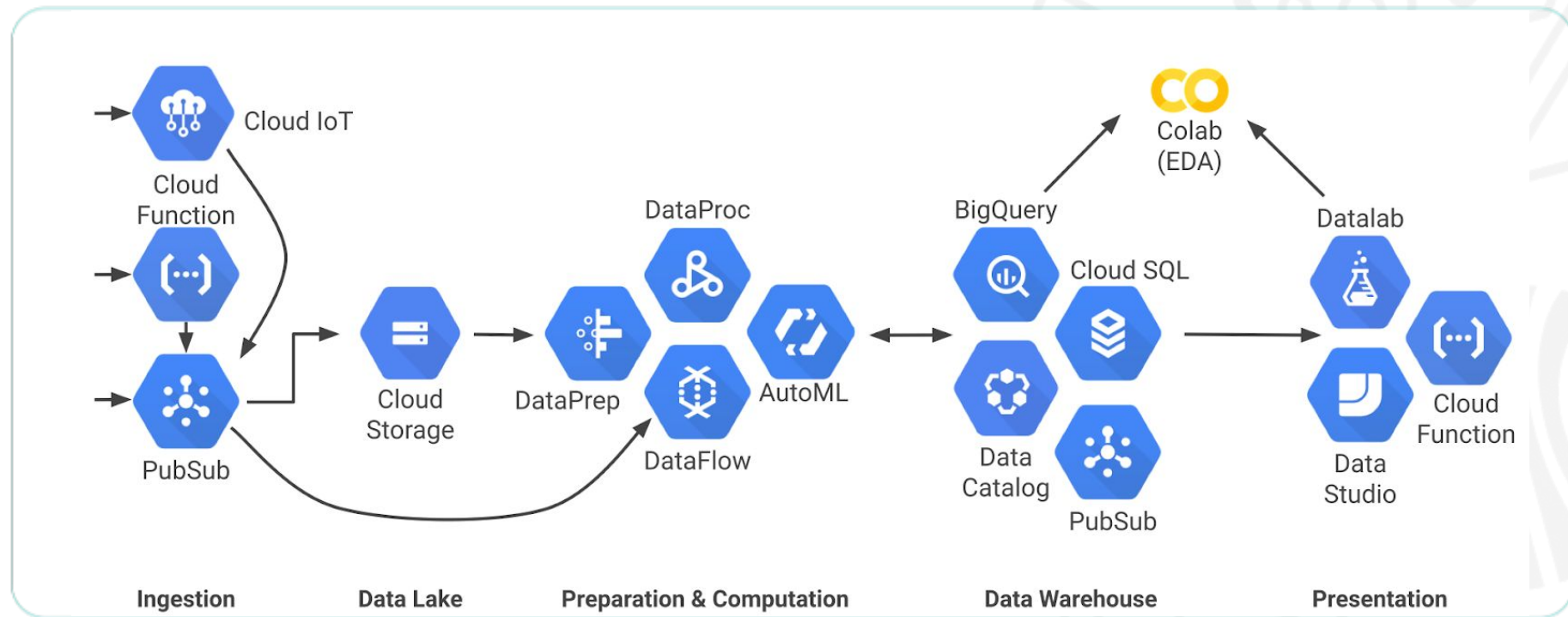
Onde trabalhar?



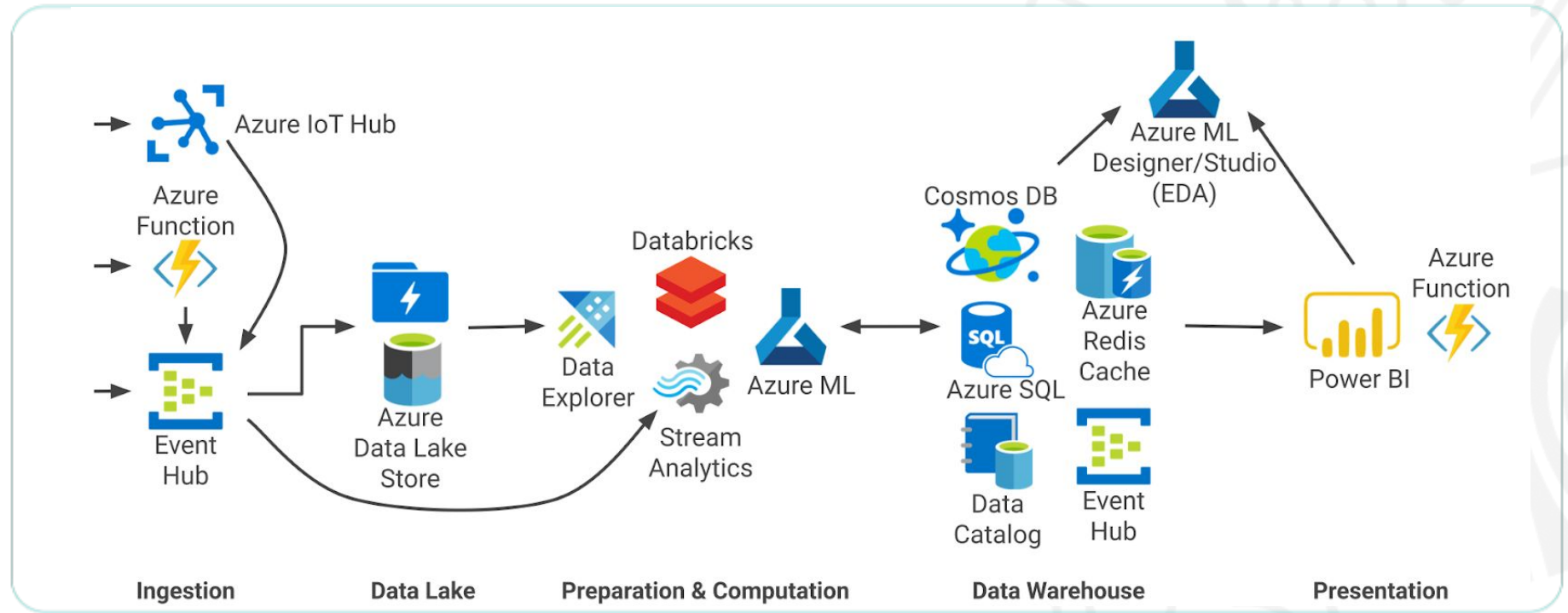
AWS



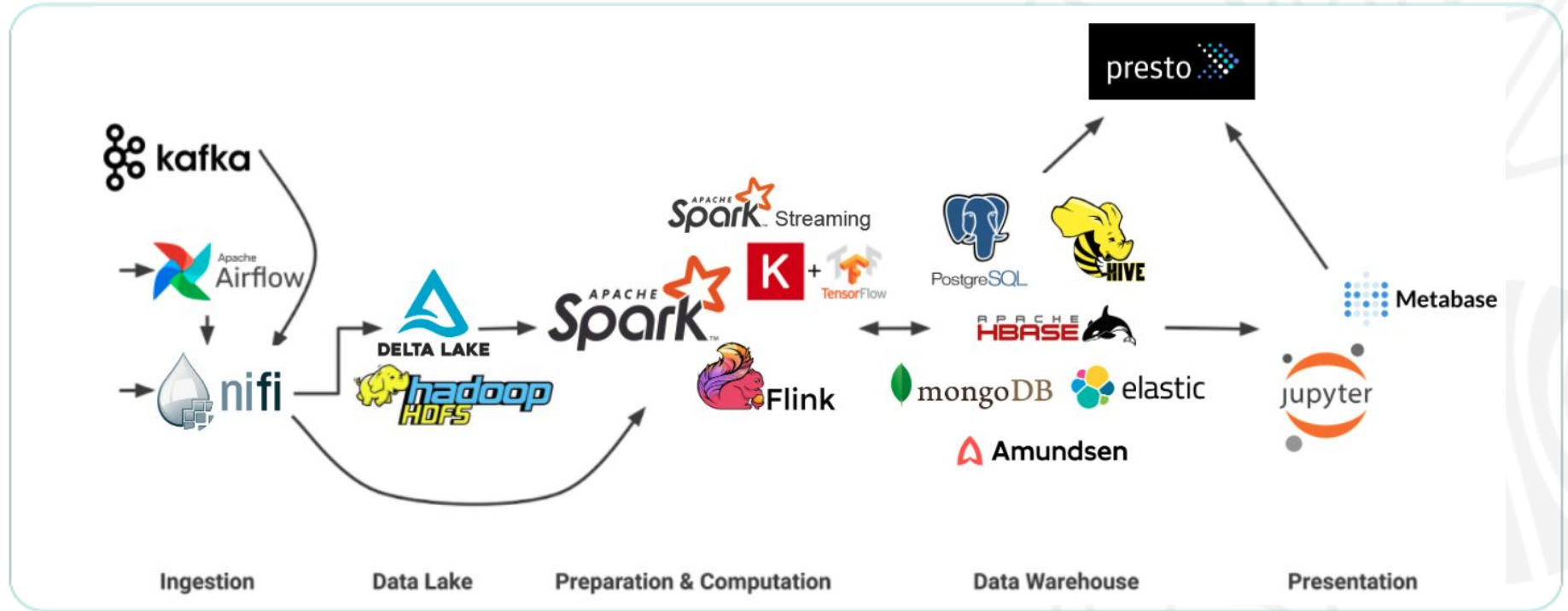
GCP



Azure



Open Source



Open Source



Flink



beam



kafka



STORM



nifi

ETL



talend

● ● ● ● ● ● ● ● ● ●
informatica
powercenter

ORACLE®

DATA INTEGRATOR



Microsoft®
SQL Server®

Integration Services



pentaho



**InfoSphere
DataStage**



PUC Minas
Virtual