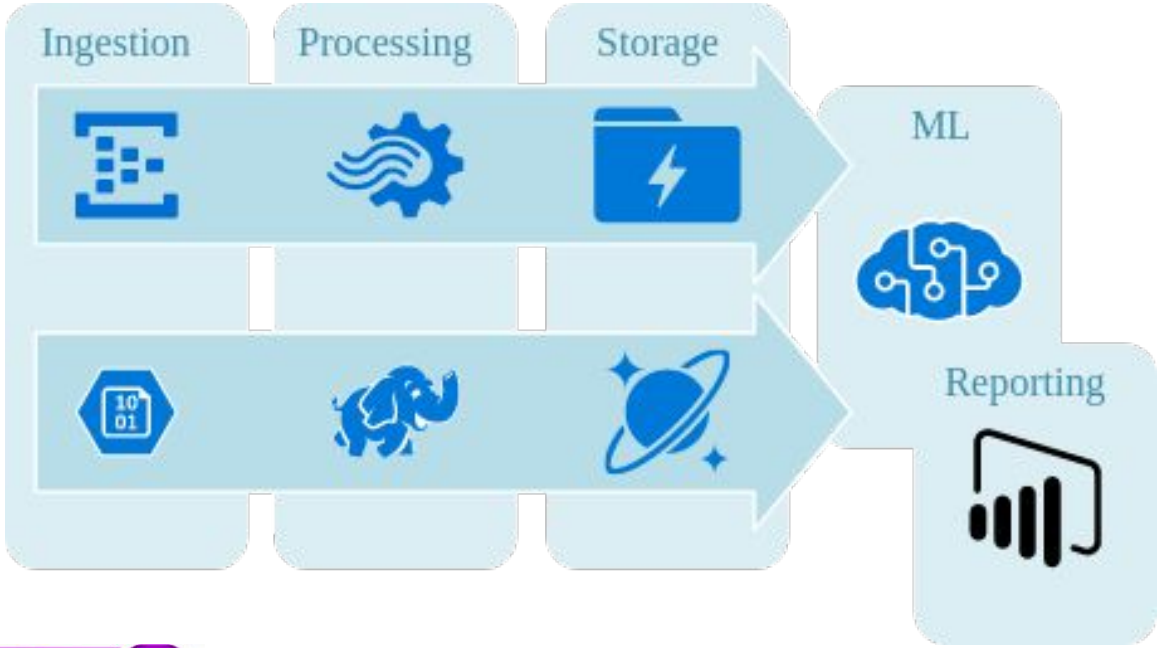
ARQUITETURA E PLATAFORMA DE DADOS



Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis
Módulo: BIG DATA COM PYTHON - Professor: André Morais
Turma 2019-2020

Fonte da Ilustração
https://docs.microsoft.com/pt-br/azure/architecture/data-guide/

O que é uma Plataforma de Dados?

- Plataforma de dados é uma arquitetura de Big Data projetada para lidar com a ingestão, processamento e análise de dados, sendo eles grandes ou complexos para banco de dados tradicionais.

- Possui a capacidade de processar dados em lote ou em tempo real, também conhecido com Arquitetura Lambda.

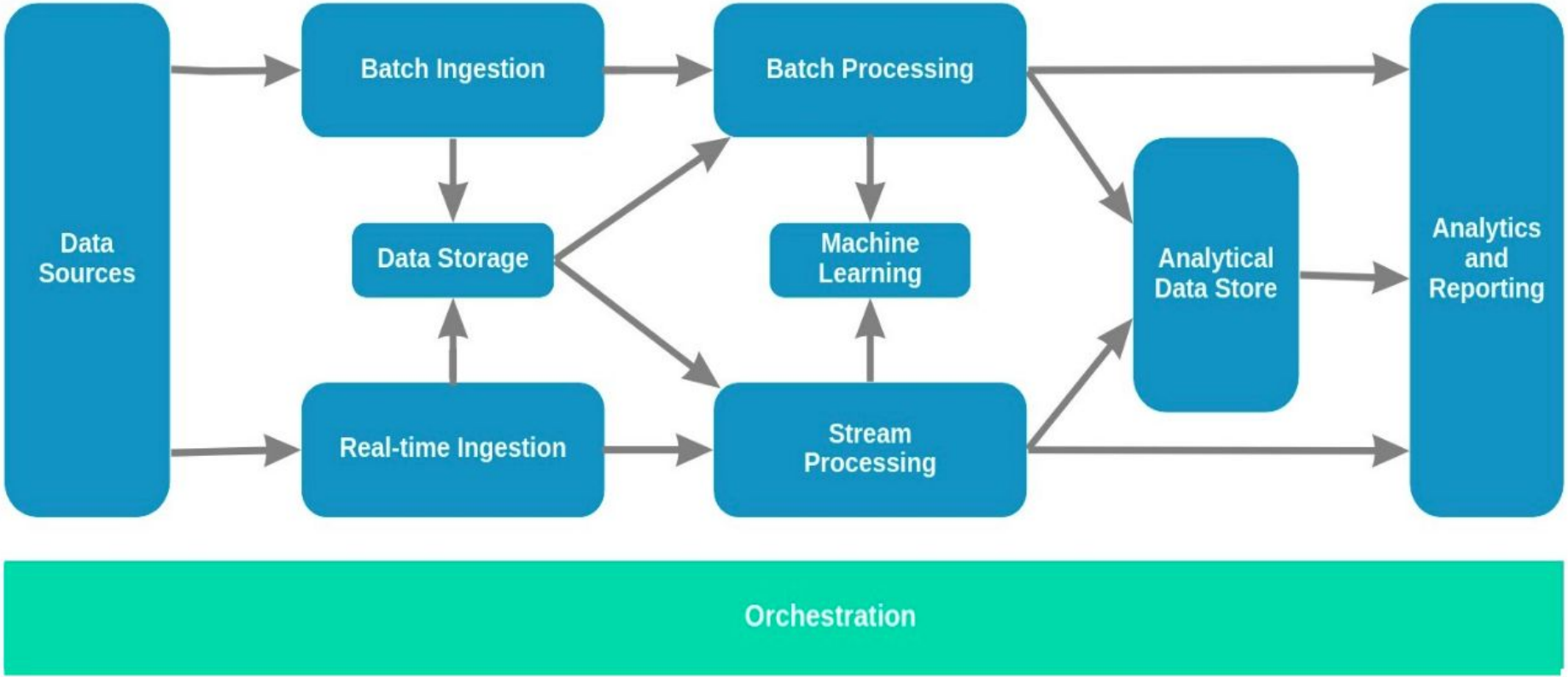
- São projetadas para trabalhar com pipelines de dados, que define o fluxo dos dados em uma solução. Isto é, como são processados, armazenados e consumidos pelo próximo componente do pipeline.

Pipeline de Dados

Pipeline de dados é um conjunto de etapas para capturar, processar e analisar dados para obter valiosos insights. Utiliza ferramentas de processamento para mover dados, transformados ou não, de um sistema “source” para outro “target”. O Pipeline se limita a um contexto, e uma plataforma de dados pode suportar uma infinidade de pipelines.

Fonte da Ilustração:
https://www.astera.com/type/blog/data-pipeline-architecture/

Plataforma de Dados - Arquitetura conceitual



Data Sources

São inúmeras origens de dados que podem ser tratadas em processos de Bigdata Analytics. Os dados podem estar espalhadas em ambientes internos ou externos à corporação, e se apresentam em formatos estruturados, semi-estruturados e não estruturados. Como exemplo, podemos destacar alguns tipos de dados como: **sociais, de sensores e máquinas, e os transacionais.**

Camada de Ingestão

A ingestão de dados é definida como o processo de absorver dados dos “data sources” e transferi-los através um pipeline de dados. Podem ser armazenados e processados dependendo do objetivo. A Ingestão pode ocorrer em tempo real, em lotes ou uma combinação de ambos (arquitetura lambda):

- **Ingestão de dados em tempo real**

Também conhecida como streaming de dados, é útil quando os dados coletados são sensíveis ao tempo. Eles são extraídos, processados e armazenados assim que são gerados. E possibilita a tomada de decisões em tempo real.

- **Ingestão de dados em lote**

Na ingestão em lotes ou batch, os dados são movidos em massa por intervalos agendados de forma recorrente. Essa abordagem é benéfica para processos repetíveis, cuja análise sobre o histórico seja relevante. Pode ser realizado de forma full ou em checkpoints.

Algumas Tecnologias utilizadas na Ingestão de Dados

JDBC Drive

Camada de Data Storage
Datalake

- O armazenamento em “data storages” geralmente refere-se a volumes que crescem exponencialmente em escala de terabyte ou petabyte.

- Diferentemente dos analytical storages, os data storages precisam estar preparados para receber dados em vários formatos através dos processos de ingestão, em sua forma bruta.

- Na busca pela geração de valor, o conceito Datalake tem evoluído. Podemos segmentar o armazenamento dos dados em “zonas” (Zone). Os dados passam por pipelines de estruturação e enriquecimento e vão migrando da Zona.

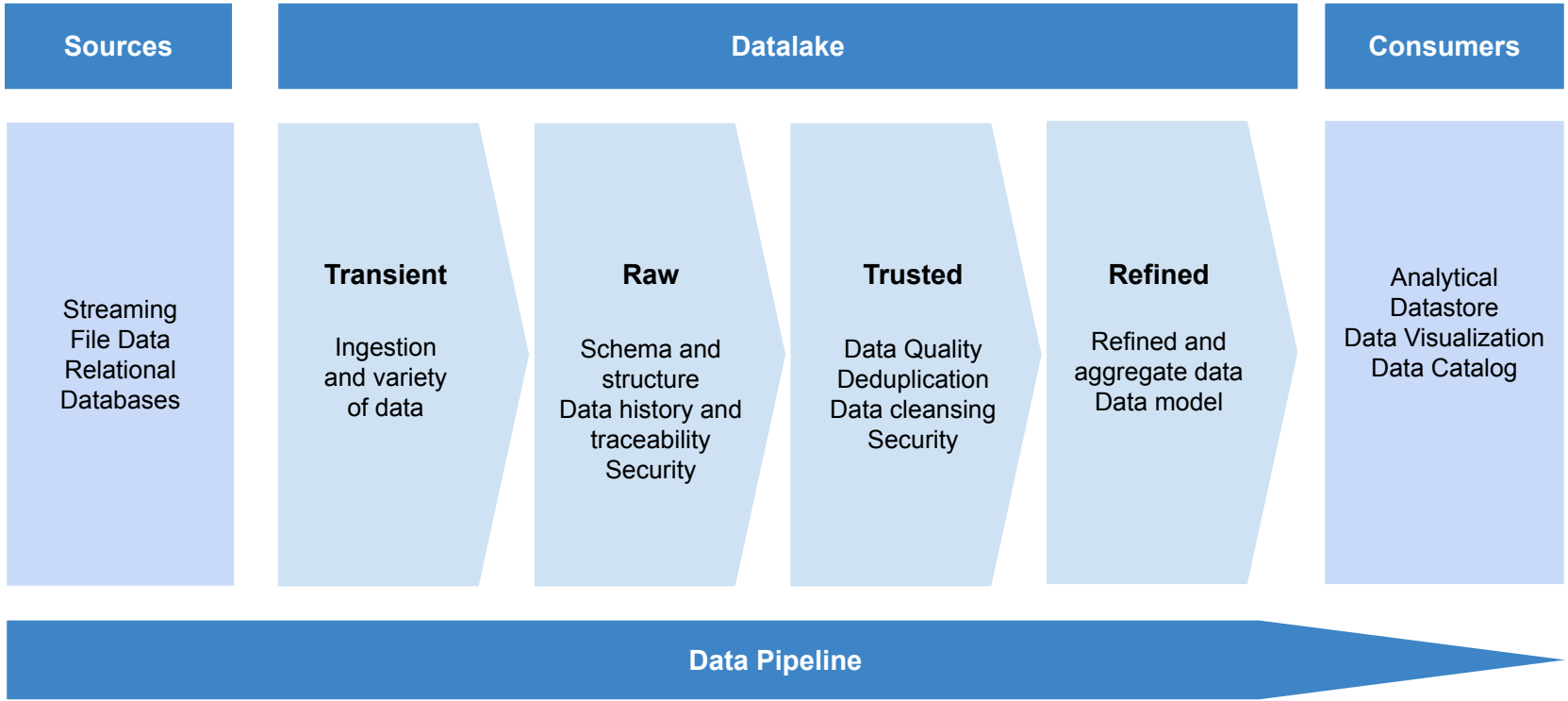
Zonas do Datalake

Zonas são estruturas lógicas que são parte do ciclo de vida, qualidade e governança do dado dentro do ambiente. O ciclo segue um padrão determinado de ELT (Extract, Load and Transform):

- O dado é gerado em seu sistema de origem, em uma infinidade de fontes;
- É capturado através do processo de ingestão em formatos diversos;
- Passa por um processo de estruturação e padronização;
- E pode ser usado em processos de analytics gerando informações de nichos de negócio;

A implementação de Zonas podem variar no mercado. Veremos um padrão que pode servir de referência para possíveis arquiteturas.

Zonas do Datalake



Zona Transient

É a camada de entrada e ingestão dos dados no Datalake. Pode receber dados de vários formatos e fontes, sendo o início da governança e mapeamento dos sources.

Raw

Schema and structure
Data history and traceability
Security

Zona Raw

Os dados ainda estão em seu estado bruto, porém geralmente com schema definido e formato estruturado. Forma-se um histórico de todos os eventos ocorridos. Em alguns caso é definido como a primeira zona, a entrada dos dados no Lake. Dados confidenciais já podem ser tratados.

Zona Trusted

Essa camada torna-se a “fonte da verdade" dentro de seu contexto. Geralmente passam por processos de qualidade dos dados, higienização e de-duplicações de registros. Com isso, estarão disponíveis para processos de refinamento e geração de informações pertinentes ao negócio.

Refined

Refined and aggregate data
Data model

Zona Refined

A camada Refined é uma zona especializada, cujo dado tratado e enriquecido está ligado a nichos de negócio. Muitas vezes com regras específicas aplicadas. Onde as aplicações irão consumir. As informações geralmente são disponibilizadas em bancos de dados analíticos, ou bancos relacionais, onde podem ser disponibilizados em APIs e visualização.

Algumas Tecnologias utilizadas em Datalakes

Camada de Processamento

Essa camada é primordial na geração de valor de uma plataforma, considerando os grandes desafios do bigdata e com a agilidade que o negócio necessita.

- **Processamento em tempo real -** Processamento através de streaming de dados são necessários para tomadas de decisões no momento em que o evento ocorre, são sensíveis ao tempo.

- **Processamento em Batch -** Processamento em lote precisam de ambientes escaláveis e aplicações distribuídas para lidar com o volume, tratamento sobre o histórico de dados e cruzamentos com uma infinidade de outras informações.

- **Machine Learning** - Precisa atender as necessidades de modelagem, treinamento e predição baseados em Machine Learning, dando autonomia aos analistas e cientistas de dados os acessos e recursos necessários para o trabalho.

Tecnologias utilizadas no Processamento e Análise

Camada de Analytical Data Store

Analytical Data Store são bases de dados especializadas e otimizadas para fornecer menores tempos de resposta e análises avançadas. São escaláveis e geralmente colunares, possibilitando gravação, leitura e compactação de dados com eficiência em disco, a fim de acelerar o tempo de resposta de uma consulta. Possui escalabilidade horizontal, compatibilidade com SQL e funcionalidade analítica avançada.

Camada de Visualização

A visualização de dados está diretamente relacionada à geração de valor para tomada de decisão. Permite aos gestores que as análises sejam feitas visualmente, para que possam compreender conceitos difíceis ou identificar novos padrões. Com a visualização interativa, é possível analisar uma informação sobre vários ângulos.

Orquestração de Pipelines

Grandes soluções de dados consiste em operações de processamento de dados repetidas, agendadas e implementadas em fluxos. Um orquestrador de pipelines é uma ferramenta que possibilita automatizar estes fluxos de trabalho. Realiza tarefas como agendamento de jobs, execução dos fluxos e coordenação das dependências entre tarefas.

Resumo das possíveis tecnologias úteis por camada

Ingestão

Data Store

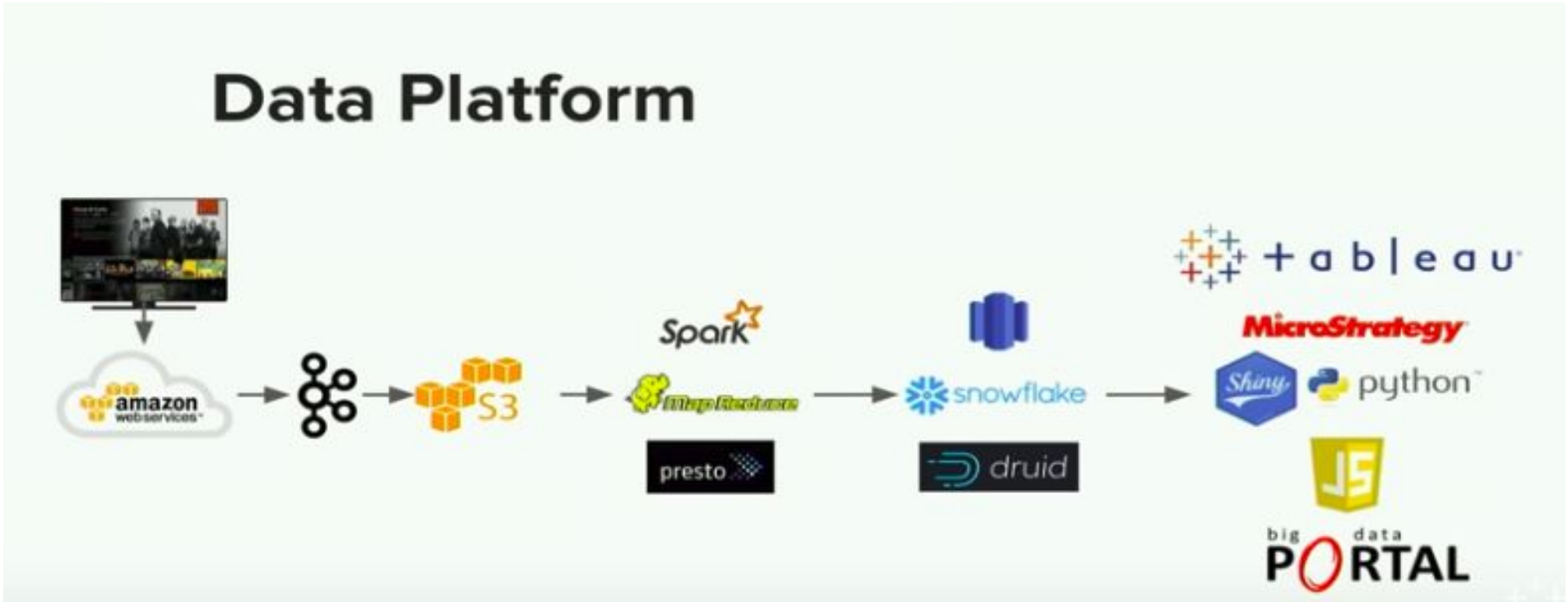
Processo/Análise

Analytical Store

Visualização

CASE NETFLIX

NETFLIX



Enfim o que faz a Engenharia de Dados

- Lida com grandes volumes de dados
- Criação de pipelines confiáveis
- Combinação de múplicas fontes de dados
- Criação de arquitetura e soluções distribuídas e escaláveis
- Colaboração com os cientistas de dados e equipe de analytics para solução de problemas de negócio
- Organização, governança e democratização dos dados

Mãos na massa, ou melhor no código \o/