Uni-FACEF
Centro Universitário Municipal de Franca

PÓS GRADUAÇÃO Uni-FACEF

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis

Turma 2021-2022

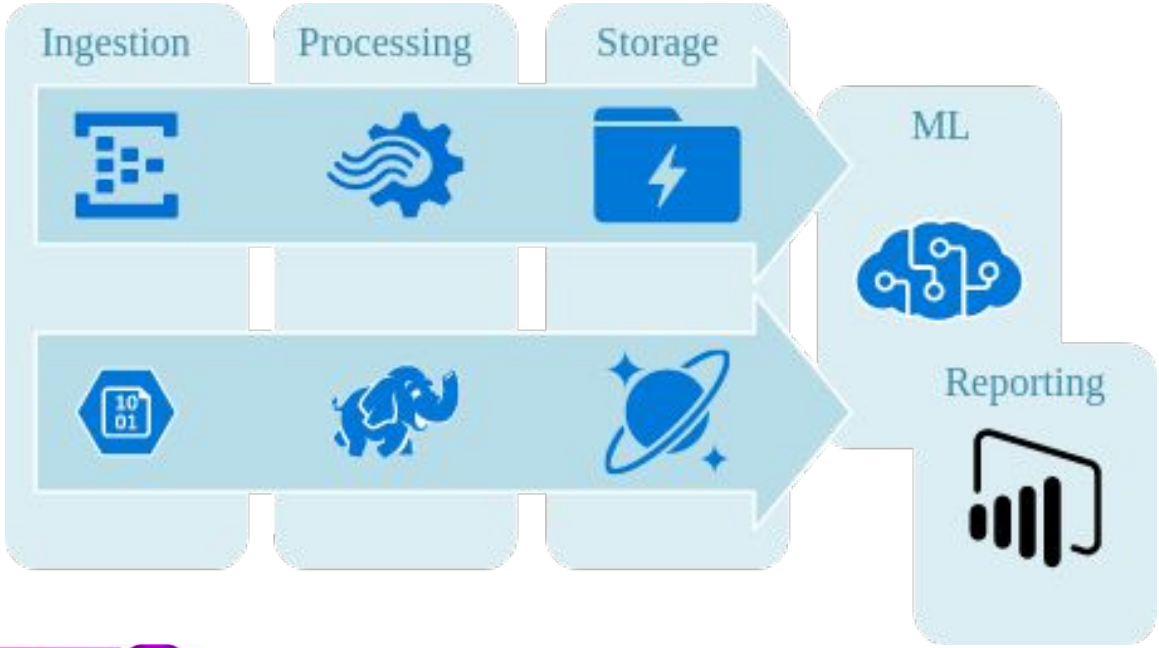
Big Data com Python

André Morais

andre.morais@luizalabs.com

09/2022

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 2021-2022

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

Arquitetura O que é 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**.

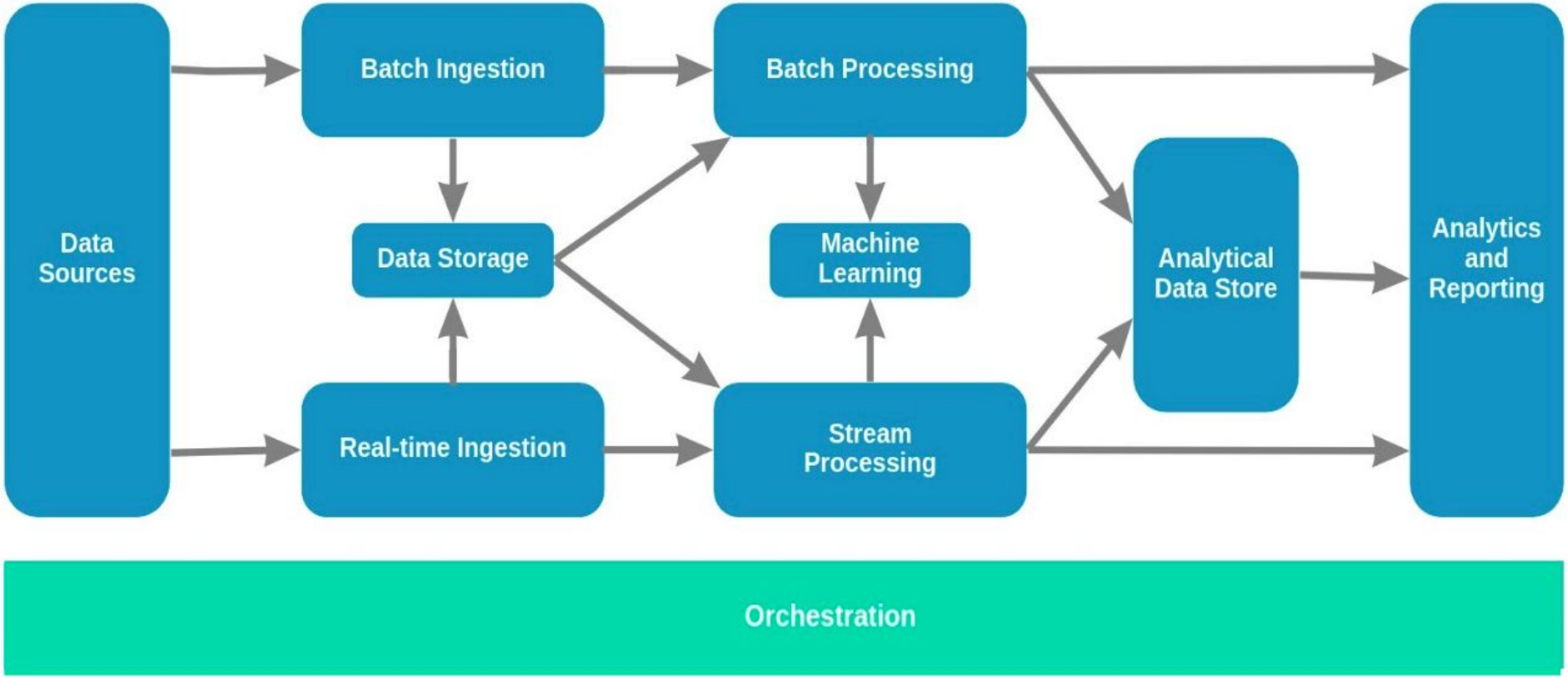
- São projetadas para trabalhar com **pipelines**, que definem o **fluxo dos dados** de um determinado contexto ou produto. Isto é, como serão orquestrados as etapas de ingestão, processamento e armazenamento dos dados.

Plataforma de dados 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 de **diferentes times**.

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

Plataforma de dados Arquitetura conceitual



Plataforma de dados Data Sources

São inúmeras **origens de dados** que podem ser tratadas em processos de Big Data Analytics.

Os dados podem estar espalhados em ambientes **internos** ou **externos** à corporação, e se apresentam em **formatos** estruturados, semi-estruturados e não estruturados.

Exemplo de tipos de dados como: **sociais**, de **sensores e máquinas**, e os **transacionais**.

Plataforma de dados Camada de Ingestão

Processo de extração de dados dos **sources** e transferi-los através um pipeline de dados. Podem ser armazenados e processados dependendo do objetivo e podem 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. Possibilita tomada de decisões em tempo real.

- **Ingestão de dados em lote ou batch:**

Dados são movidos em massa por intervalos agendados de forma recorrente. É benéfica para processos repetíveis, cuja análise sobre o histórico seja relevante. Pode ser realizado de forma full ou em checkpoints.

Plataforma de dados Tecnologias para Ingestão de dados

JDBC Drive

Plataforma de dados

Camada de Data Storage - Data Lake

- O armazenamento em **data storages** 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.
- Podemos segmentar o **armazenamento dos dados em "zonas"** (Zone). Os dados passam por pipelines de estruturação e enriquecimento e vão migrando de Zona.

Plataforma de dados Tecnologias para Data Storages

Plataforma de dados Zonas do Data Lake

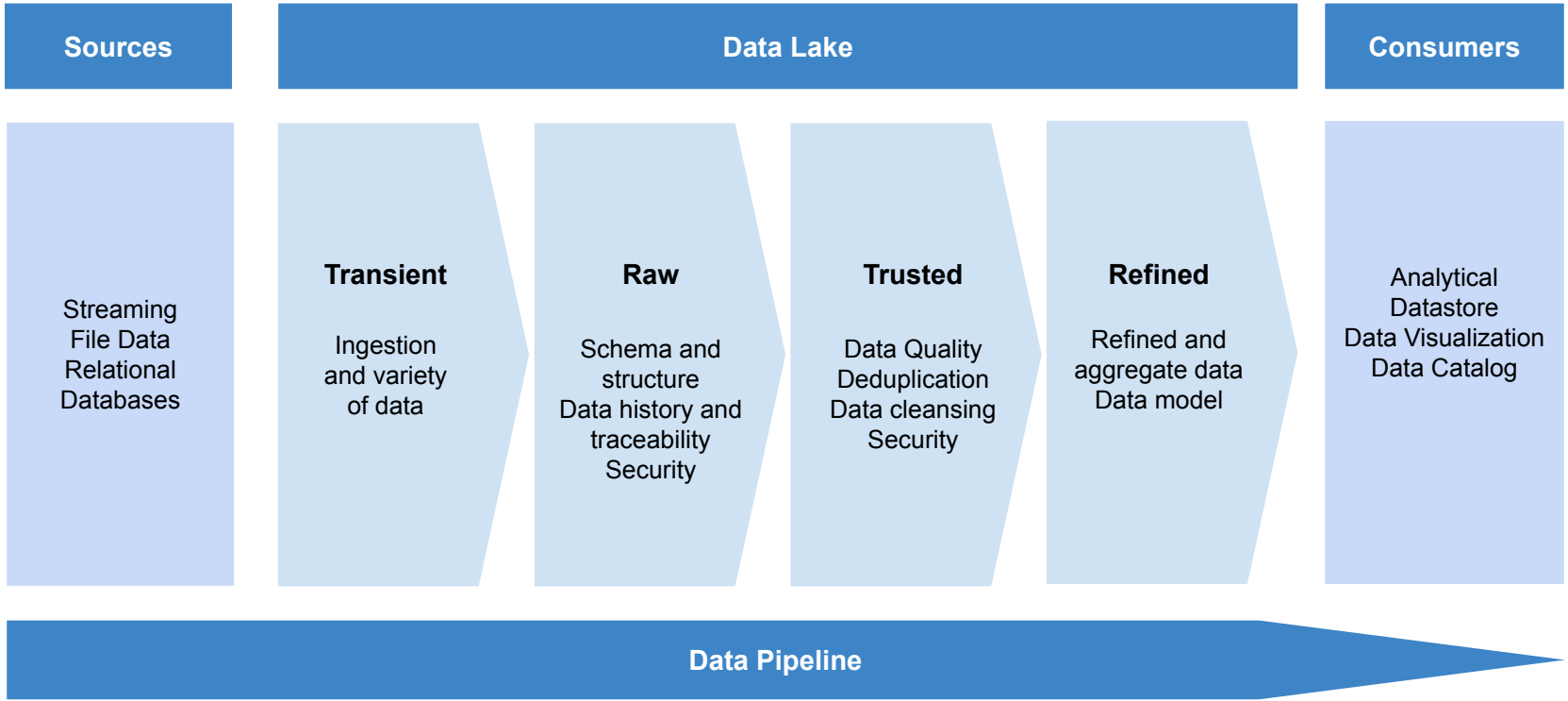
Zonas são estruturas lógicas, parte do ciclo de vida, qualidade e governança do dado dentro do ambiente.

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**
- Pode ser usado para **analytics** gerando informações de nichos de negócio

A implementação de Zonas pode variar de empresa para empresa.

Plataforma de dados Zonas do Data Lake

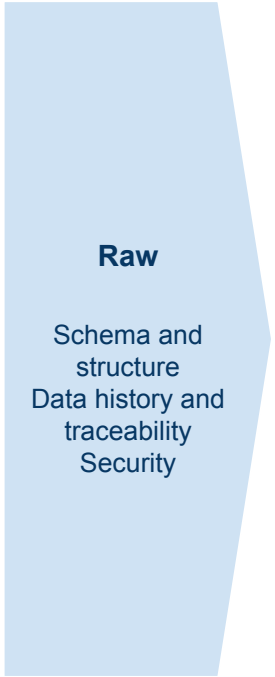


Plataforma de dados Zonas do Data Lake

Zona Transient

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

Plataforma de dados Zonas do Data Lake



Zona Raw

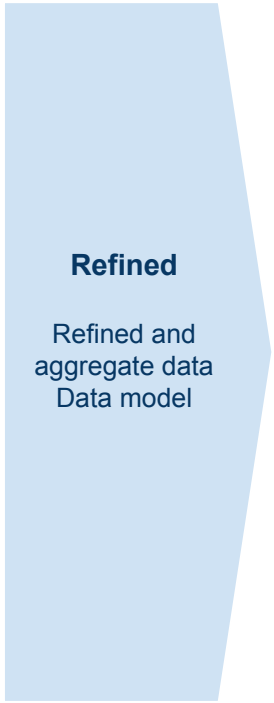
Os dados ainda estão em seu estado bruto, porém geralmente com schema definido e formato estruturado. Forma-se um histórico de todos os eventos ocorridos. Em alguns casos é definido como a primeira zona, a entrada dos dados no Lake. Dados confidenciais já podem ser tratados.

Plataforma de dados Zonas do Data Lake

Zona Trusted / Safe

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. É a camada de democratização dos dados. A Safe é a trusted para dados sensíveis.

Plataforma de dados Zonas do Data Lake



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, cientistas e analistas irão consumir. As informações geralmente são disponibilizadas em bancos de dados analíticos para camada de Dataviz, mensagerias e bancos relacionais, onde podem ser disponibilizados em APIs e aplicações.

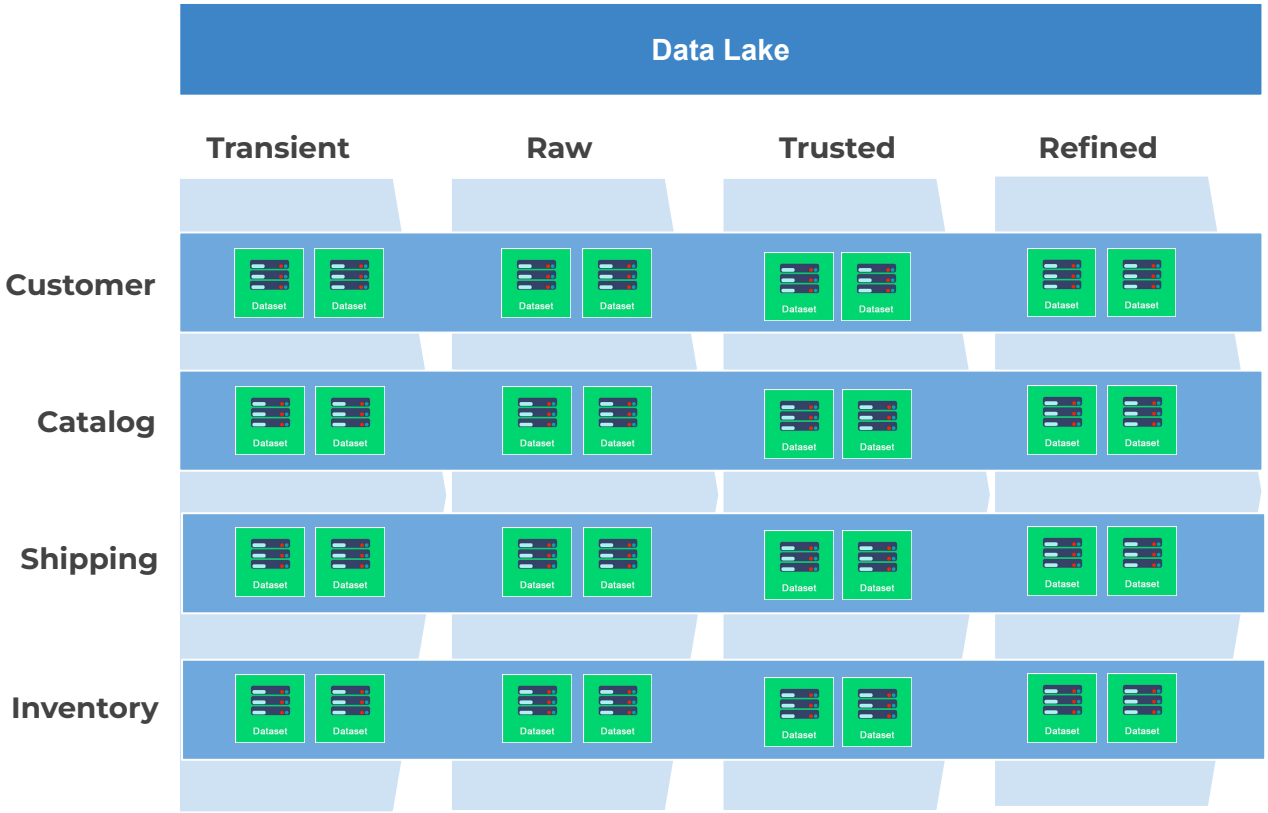
Plataforma de dados Namespaces do Data Lake

Namespaces são estruturas lógicas que possibilitam uma melhor classificação e **governança** dos dados dos **ownerships** dos **domínios** e/ou **subdomínios** dentro da organização.

Isso quer dizer que um **namespace** tem um dono **responsável** por aquele **contexto** ou produto de **dados**.

É através dos **namespaces** e **zonas** que todo controle de **segurança** e **acesso** aos dados podem ser **implementados**, assim como os owners users ou schemas de um banco de dados.

Plataforma de dados Namespaces do Data Lake



Plataforma de dados Camada de Processamento

Essa camada é primordial na geração de valor de uma plataforma, considerando os grandes desafios do big data 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 precisa 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.

Plataforma de dados

Tecnologias de Processamento e Análise

Plataforma de dados 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**.

Plataforma de dados 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**.

Plataforma de dados Orquestração de Pipelines

Grandes soluções de dados consistem 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.

Plataforma de dados Orquestração de Pipelines

DAG - Directed Acyclic Graph

Fonte da imagem:
https://umair-iftikhar.medium.com/data-pipeline-solution-part-2-af2f0a2ddd8d

Plataforma de dados Resumo das tecnologias

Ingestão

Data Store

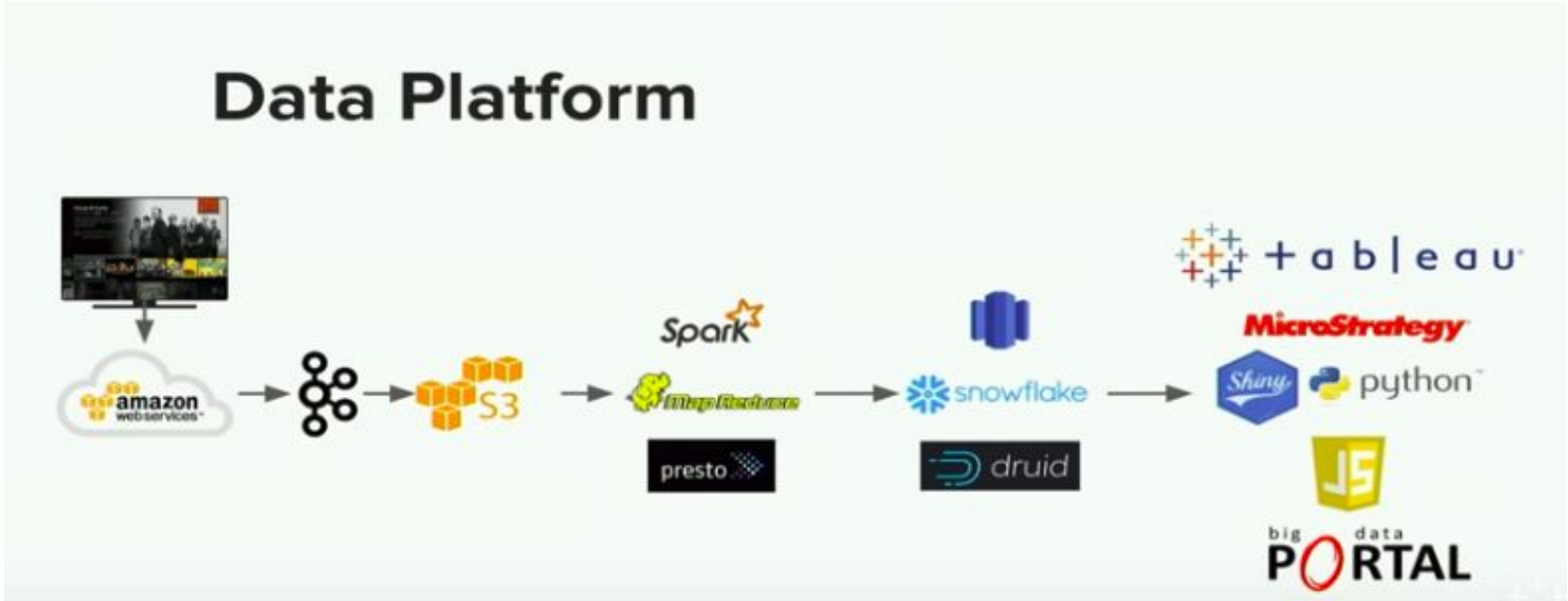
Processo/Análise

Analytical Store

Visualização

CASE NETFLIX

NETFLIX



Plataforma de dados Governança dos Dados

- **Autoria**: Catalogação dos Domínios, subdomínios e ownerships dos dados. Responsáveis pelo contexto dos dados
- **Acessibilidade**: Controle dos acessos aos dados. Os dados sensíveis devem estar protegidos com políticas bem definidas
- **Segurança**: Mecanismos e políticas que garantem os acessos. Garantir que a liberação siga um fluxo de aprovações
- **Qualidade**: Garantir de que os dados estejam estruturados e com qualidade, ranqueados pelos usuários na plataforma
- **Conhecimento**: Garantir que o dado agregue valor para a organização, e descontinuá-lo caso não seja mais relevante

Fonte da imagem:
https://blog.magrathealabs.com/projetando-para-prote%C3%A7%C3%A3o-de-dados-98db5fa96996

Plataforma de dados Características e benefícios

- Lida com grandes volumes de dados de uma organização
- Lida com um infinidade de pipelines de diversos domínios
- Combinação de múltiplas fontes de dados de diversos formatos
- Arquitetura e soluções distribuídas e escaláveis
- Ferramentas e insumo para cientistas de dados e equipe de analytics para solução de problemas de negócio
- Organização, governança e democratização dos dados
- Autonomia aos ownerships de domínios para criação de seus produtos de dados
- Tomada de decisão baseada em dados, cultura Data Driven.

No próximo Capítulo veremos:

- Case Magalu

“Os dados são importantes ativos para as empresas, pois são a bases para construção de Conhecimento”