Pipelines ETL

Aplicação de conceitos de DW para a construção de pipelines de extração, transformação e carregamento de dados

Carlos Baqueta

Contextualização

Contextualização

Quem sou eu? *(momento filosófico)*

- Estudante de Ciências de Computação (último semestre, uhuuul);
- Estagiário em *Business Intelligence,* atuando principalmente como Engenheiro de Dados, na Arquivei.

Quem é a Arquivei?

- Empresa que atua no mercado B2B e tem como principal produto o Arquivei.

O que é o Arquivei?

- É um serviço de gestão de documentos fiscais (NFe, CTe, <sopa_de_letrinhas>...).

Qual era o problema?

Qual era o problema?

Os gerentes e sócios desejavam ter uma visão de alguns indicadores de desempenho, do inglês, *Key Performance Indicators* - KPIs), tais como:

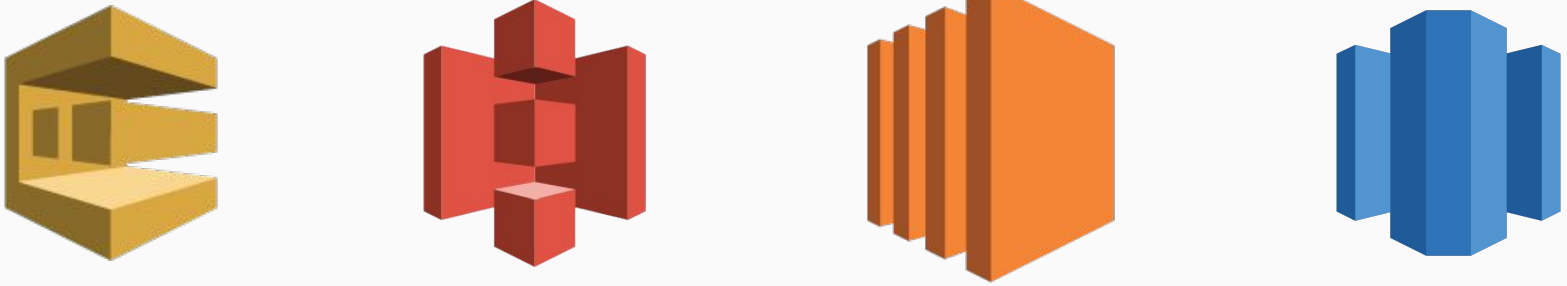
- Logins
- Leads
- Criação de Contas
- Churns
- Receita
- etc

Solução: criação do DW da empresa (*EDW*)

Solução 1: Kimball + AWS

O que é Amazon Web Services (AWS)?

AWS oferece serviços de infraestrutura de TI



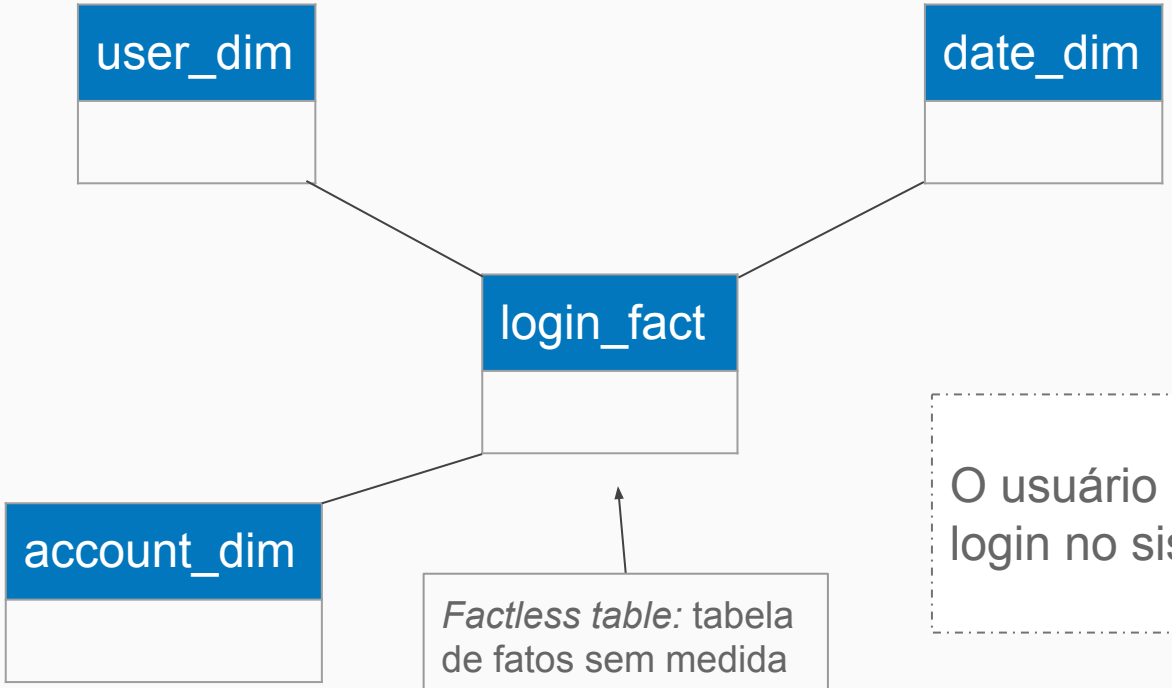
Solução 1: Kimball + AWS

Principais características da primeira solução:

Modelagem dimensional	Esquema estrela
Literatura base	*The Data Warehouse Toolkit,* Kimball and Ross. 3 edição.
Arquitetura do pipeline ETL	Réplica/S3 --(**E**)--> stg1 --(**T**)--> stg2 --(**L**)--> Redshift **stg:** staging area. Local de armazenamento temporário de dados no pipeline ETL.

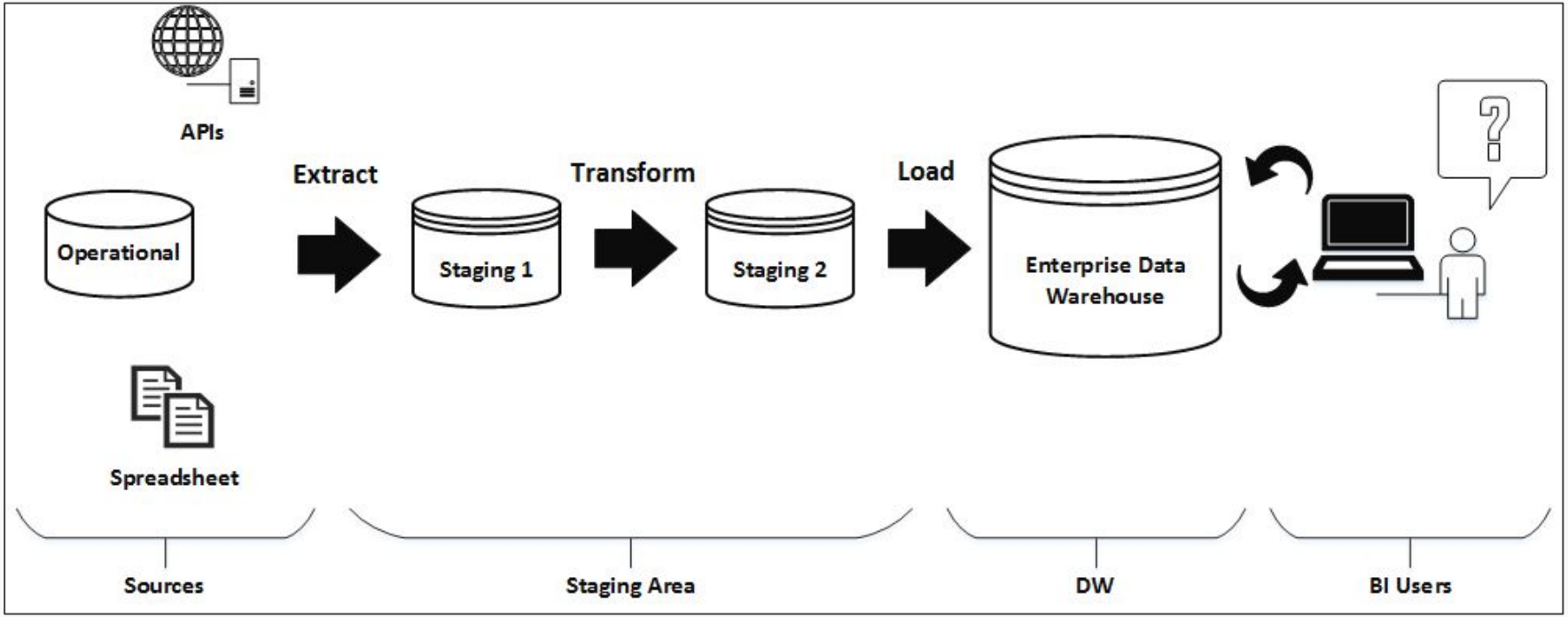
Solução 1: Kimball + AWS

Exemplo do esquema estrela para o evento de processo de negócio *login*



Solução 1: Kimball + AWS

Arquitetura da solução:



Solução 1: Kimball + AWS - Desafios

Um dos principais desafios no processo de ETL é a identificação de instâncias advindas de diferentes provedores de informação.

- Ex.: Considere que temos a tabela de dimensão *user_dim* no banco de dados analítico e que seus dados advêm de 2 provedores:

Provedor 1

id	name	cpf	email
1	Anna	x	y
...	...	...	...

Provedor 2

user_id	name	cpf	email
...	...	...	...
54	Anna	x	y

Como saber que esses 2 registros referem-se a mesma instância?

Solução 1: Kimball + AWS - Desafios

Segundo problema: como identificar as instâncias dentro do DW?

- Ex.: Considere que temos a tabela de dimensão *user_dim* no banco de dados analítico e que seus dados advêm apenas da tabela *user* de um provedor de informação
    - Podemos utilizar como id da tabela *user_dim* o mesmo id da tabela *user*?

Solução 1: Kimball + AWS - Implementação

Detalhes sobre a implementação da primeira solução:

Linguagem de programação	Python3.4
Bibliotecas úteis	petl , psycopg2
Orquestração de tarefas	Airflow (https://airflow.incubator.apache.org/)

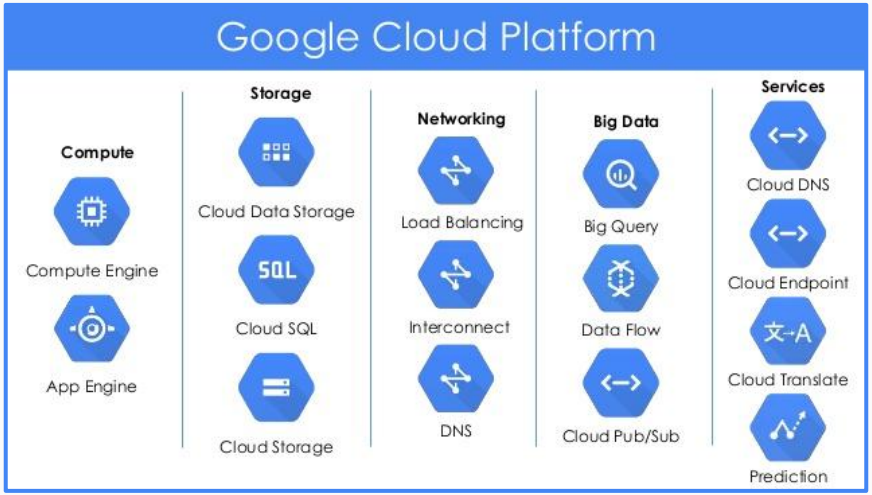
Solução 1: Kimball + AWS - Desvantagens da solução

- ❏ Gerenciamento da infraestrutura;
- ❏ Gerenciamento de *dockers*;
- ❏ Gerenciamento do Airflow (gerenciador de tarefas);
- ❏ Demora na implementação dos pipelines (Também por falta de experiência)

Solução 2: Google Cloud Platform + Apache Beam

O que é Google Cloud Platform (GCP)?

Plataforma que provê acesso a **infraestrutura**, **análise de dados** e **aprendizado de máquina** da Google.



Solução 2: Google Cloud Platform (GCP) + Apache Beam

Principais componentes da seção *Big Data* de serviços da GCP:

Pub/Sub	serviço de mensagem a partir do qual aplicações independentes podem enviar e receber mensagens. Conceito de *topic* e *subscription*.
Dataflow	serviço de processamento de dados que oferece suporte tanto a execução de pipelines *batch* quanto *stream*. Completamente gerenciado pelo Google.
BigQuery	*EDW* que possibilita a análises de dados em larga escala com alto desempenho.

O que é Apache Beam?

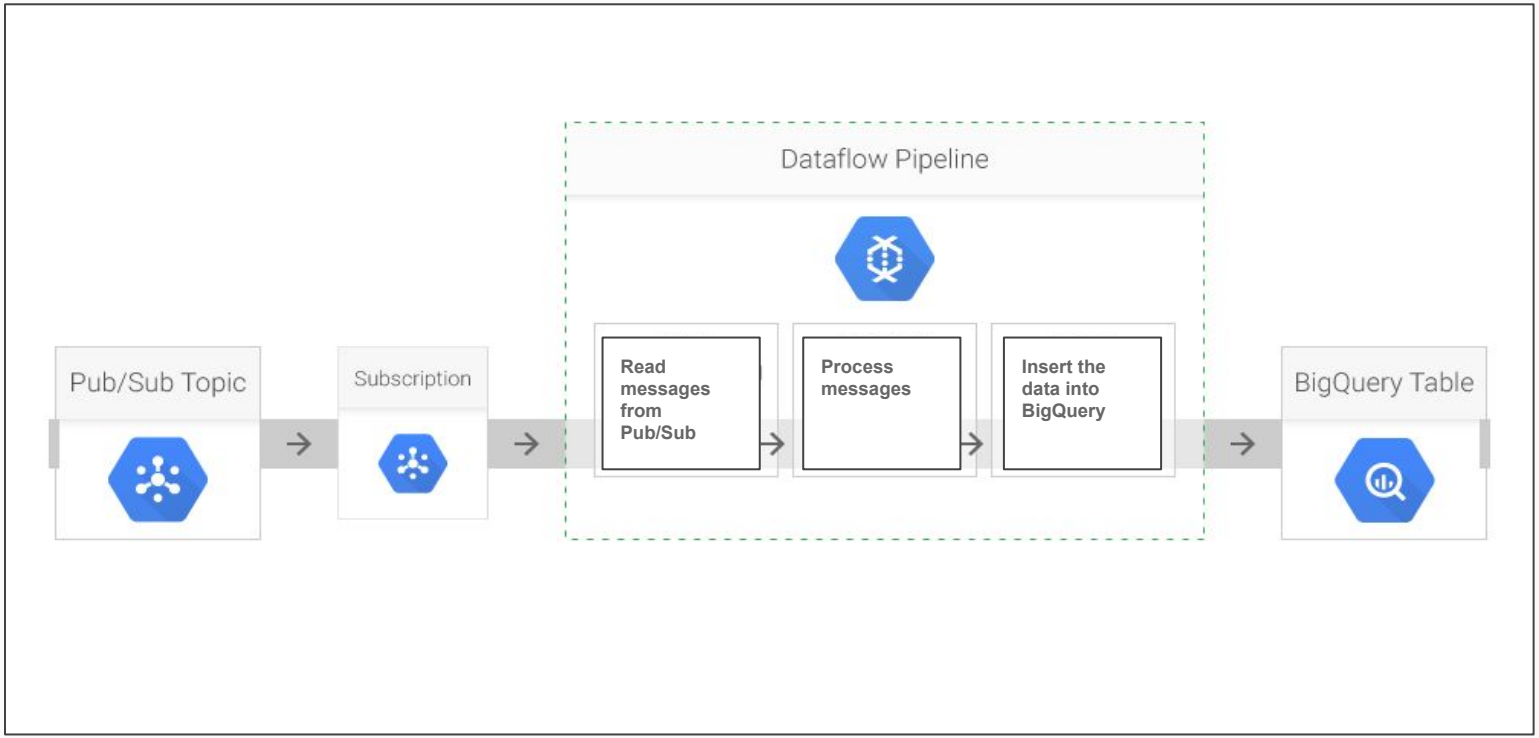
Apache Beam é um projeto open source que provê um **modelo unificado** para a **definição de pipelines** de processamento de dados em paralelo tanto para a execução em *batch* quanto para a execução em *stream*

Uma vez definido o pipeline utilizando **Beam SDKs** é possível submetê-lo como um job para ser executado no **Google**

fonte: https://beam.apache.org/get-started/beam-overview/

Solução 2: Google Cloud Platform (GCP) + Apache Beam

Arquitetura da solução:



Solução 2: Google Cloud Platform (GCP) + Apache Beam

PInput -> PTransform -> PCollection -> … -> POutput

Perguntas?

Obrigado!