Arquitetura e processamento de dados

Meetup 25/05/2022

Slides e código

https://github.com/arquivei/data-meetup-beam

Ordem da apresentação

- Contextualização do problema
- O que é o Apache Beam
- Básico da tecnologia
- O que é um pipeline
- Exemplos
- Casos de uso da Arquivei
- Mão na massa

Problema

- Big Data
- Computação distribuída
- Como movemos dados de maneira **confiável** e **eficiente**?
- 3Vs: volume, velocidade e variedade

Problema

- Migração de um banco de dados que não suporta mais sua aplicação
- Consumo massivo de dados (IoT)
- Tratamento dos dados para ML

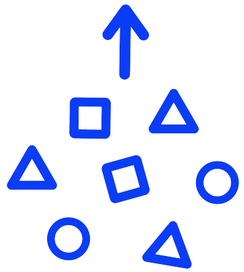
Esses problemas conseguem ser abstraídos como **pipelines de processamento de dados**

Apache Beam

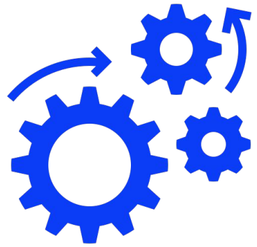
Modelo de programação **open source** para construir **pipelines de processamento de dados**, tanto em **batch** quanto em **streaming**

Apache Beam

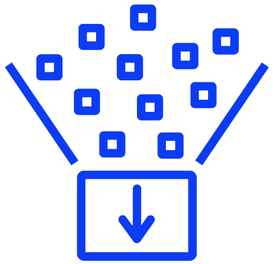
Usado na etapa de processamento de dados



Ingestão

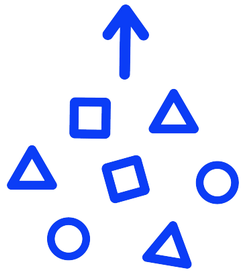


Processamento

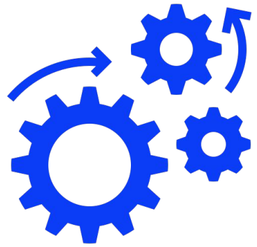


Escrita

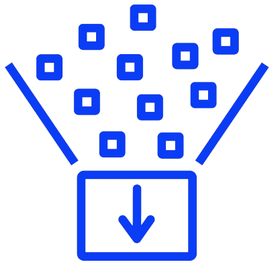
Apache Beam



Leitura dos dados on-premises ou na nuvem



Definição da lógica de negócios em batch ou em streaming



Escrita do resultado nos destinos mais populares

Apache Beam

Os 3 passos básicos para construir um pipeline com Apache Beam:

1. Escolha um runner

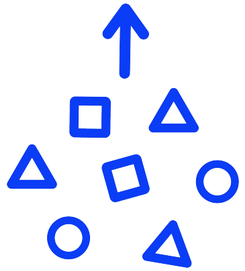
Dataflow

Flink

2. Escolha uma SDK

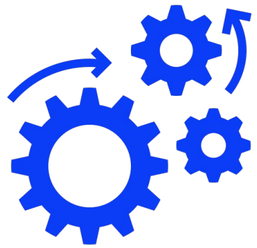
3. Use conectores e transformações para seu caso de uso

Conectores de entrada



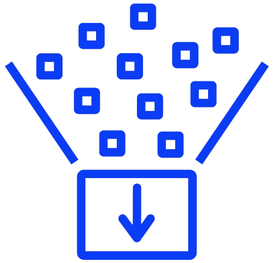
- Sistemas de arquivos
- Serviços de mensageria
- Banco de dados
- ...

Operações



- ParDo
- GroupByKey
- Flatten
- Partition
- ...

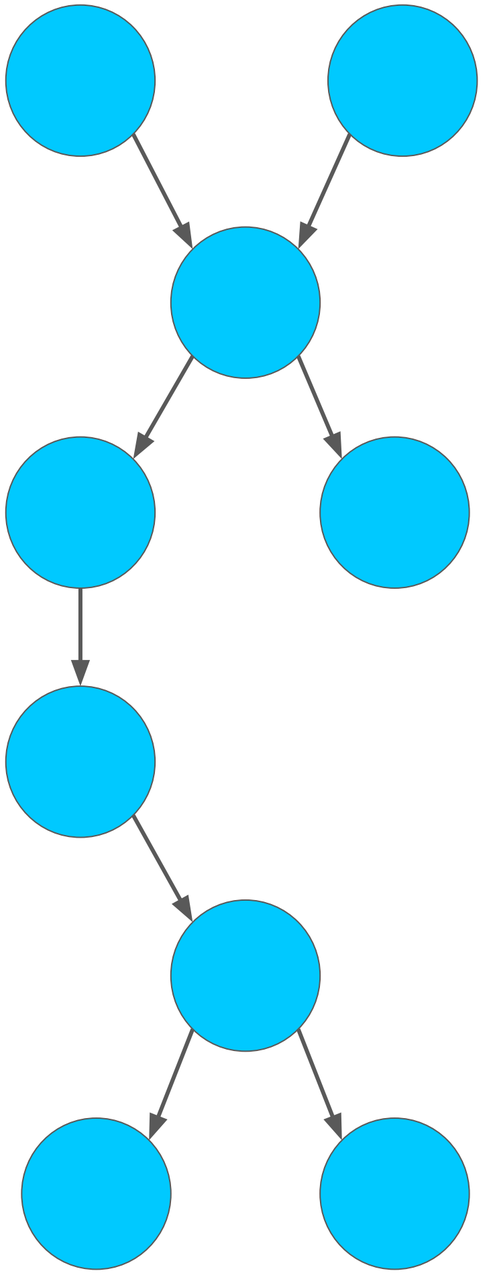
Conectores de saída



- Sistemas de arquivos
- Serviços de mensageria
- Banco de dados
- ...

O que é um pipeline?

- **DAG** (Directed Acyclic Graph - Grafo Acíclico Direcionado) de transformações de dados aplicado a uma ou mais coleções de dados
- Pode conter **múltiplas** entradas e destinos
- Suas operações (**PTransforms**) podem tanto ler quanto gerar **PColletions**



PCollection

- Coleção imutável de valores
- Pode ser limitada (batch) ou ilimitada (streaming)
- PTransforms usam PCollections como inputs e outputs

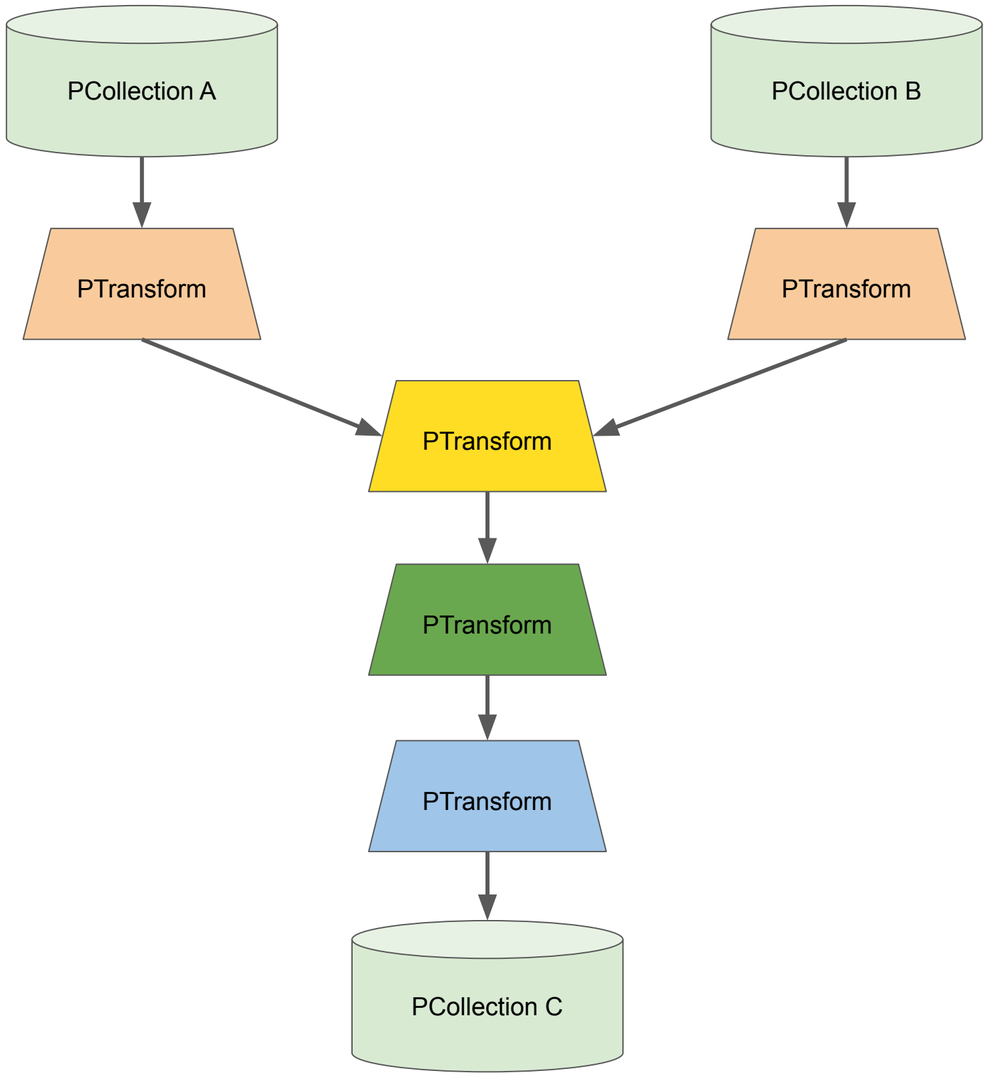
e.g.

- **Banco de dados** → PCollection[**Documento / Entrada**]
- **Arquivo de texto** → PCollection[**String (linha)**]
- **Mensageria** → PCollection[**Mensagem**]

PTransform

- Operação em um pipeline
- Uma função aplicada para cada elemento de uma ou mais PCollections
- Dependendo do runner, vários workers de um cluster podem executar o código em paralelo para gerar os elementos da PCollection final produzida pela transformação

PCollections e PTransforms

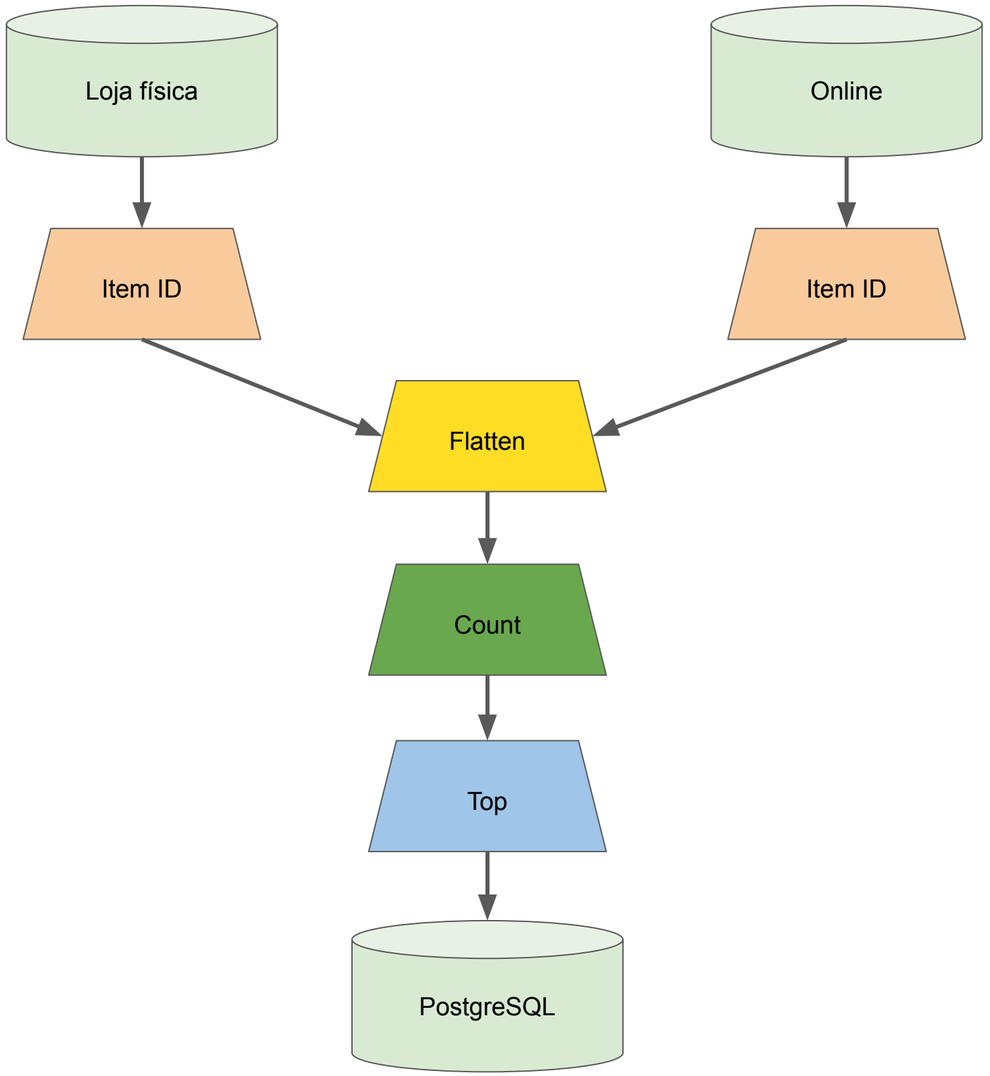


Exemplo

Quais são os produtos mais vendidos de uma loja?

- Alto volume de dados para serem analisados
- Duas lojas: uma física e outra online
- Problema pode ser abstraído em um pipeline com duas entradas de dados

Exemplo



Leitura das vendas na loja física e online

Extrair o ID do produto

Flatten para uma view unificada

Para cada ID, contar o número de vendas

Selecionar os itens com o maior número

Escrever os resultados em uma tabela

Casos de uso da Arquivei

NFSe

- https://arquivei.com.br/consulta-nota-fiscal-de-servico-nfse#nfse-coverage

Bundle parser

- https://medium.com/engenharia-arquivei/libertando-dados-massivos-presos-em-uma-aplica%C3%A7%C3%A3o-web-9954e590ada2

Pipeline genérico

- https://medium.com/engenharia-arquivei/processando-eventos-gen%C3%A9ricos-em-streaming-usando-bigquery-e-dataflow-394b80d8a182

arquivei

Dúvidas?

Mão na massa com Léo