↳ Introdução a pipeline de dados

Camila Stenico

TDC Connections 2021

Objetivo

Fundamentos de uma pipeline de dados

O que é uma pipeline de dados?

Etapas de um **ETL**

Trabalhando com **dataframe**

Introdução ao Apache Airflow (introdução mesmo)

Tem código? Sim!

Pipeline de dados:

Fluxo de dados entre diferentes sistemas

Conjunto de ações que extraem dados de diferentes fontes, definindo como, quais e quando os dados são movidos, transformados e carregados

Jornada dos dados

Bancos de dados

Arquivos

Mensageria

Sensores

APIs

Fontes de dados

Jornada dos dados

Bancos de dados

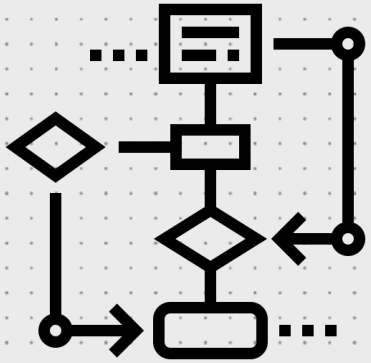
Arquivos

Mensageria

Sensores

APIs

Ingestão



Fontes de dados

Pipeline de dados

Jornada dos dados

Bancos de dados

Arquivos

Mensageria

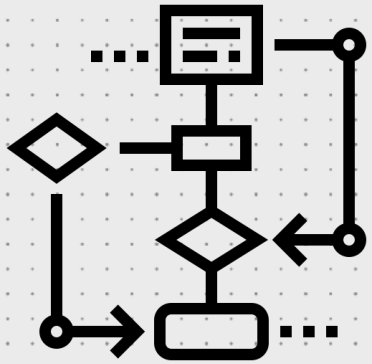
Sensores

APIs

Fontes de dados

Ingestão

Integração de dados
Integridade de dados
Transformações



Pipeline de dados

Jornada dos dados

Bancos de dados

Arquivos

Mensageria

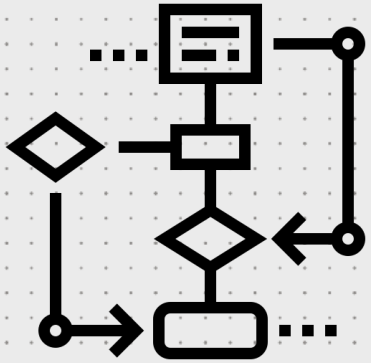
Sensores

APIs

Fontes de dados

Ingestão

Integração de dados
Integridade de dados
Transformações



Pipeline de dados

Storage

Data lakes

Data warehouses

Jornada dos dados

Bancos de dados

Arquivos

Mensageria

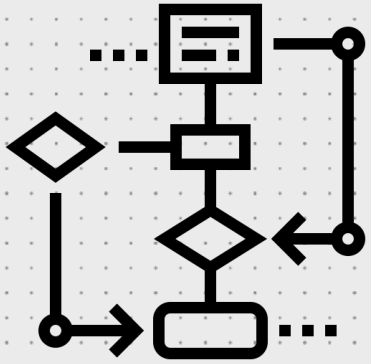
Sensores

APIs

Fontes de dados

Ingestão

Integração de dados
Integridade de dados
Transformações



Pipeline de dados

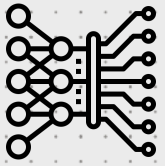
Storage

Data lakes

Data warehouses

Visualização

Business Intelligence



Ciência de dados

Consumidores de dados

↳Desafio

Dados + Spotify

- Objetivo: base de músicas ouvidas
- Aplicações: análise de comportamento do usuário

Receita:

- Python 3
- Pandas
- SQLite3
- Apache Airflow
- Conta no Spotify

Pegue seu OAuth token aqui:
https://developer.spotify.com/console/get-recently-played/

↳Automatização == Airflow

Apache Airflow:

Gerenciar um ETL com várias camadas de dependências

Scheduling de jobs

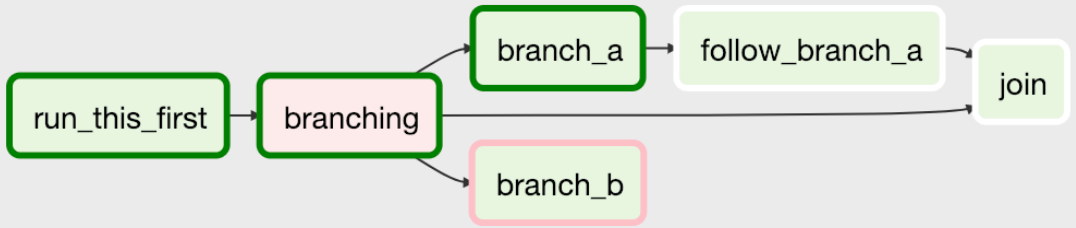
Execução de forma distribuída

Monitorar todos os passos da sua pipeline

Open source

PS: Usar cron não é adequado!

Pipelines são representadas como DAGs: grafos acíclicos



Concluindo...

Próximos passos:

- Boas práticas
- Mais usuários
- Mais fontes de dados

Evolução de pipelines conforme objetivos:

- Inclusão de outras ferramentas de processamento
- Etapas de armazenamento

↳Referências

Documentações:

https://airflow.apache.org/docs/apache-airflow/stable/index.html

https://pypi.org/project/pyspark/

https://pandas.pydata.org/

Saiba mais:

https://medium.com/the-data-experience/building-a-data-pipeline-from-scratch-32b712cfb1db

https://itnext.io/big-data-pipeline-recipe-c416c1782908

https://blogs.informatica.com/2019/08/20/data-processing-pipeline-patterns/

https://towardsdatascience.com/data-engineering-how-to-build-a-gmail-data-pipeline-on-apache-airflow-ce2cfd1f9282

https://medium.com/@itunpredictable/apache-airflow-on-docker-for-complete-beginners-cf76cf7b2c9a

https://www.analyticsvidhya.com/blog/2016/10/spark-dataframe-and-operations/

https://www.youtube.com/watch?v=dvviIUKwH7o

Código disponível em:

https://gitHub.com/cstenico/tdc-data-pipeline

Obrigada!

linkedin.com/camila-stenico

GitHub.com/cstenico

CONSUMER TEAM