Data Engineering

Roadmap

- Introdução
- Engenharia de Dados
- Coleta e Aquisição
- Crawlers/Scrapers
- Filas
- ETL em Python
- Docker
- Kubernetes
- Cloud Function
- ETL no BigQuery

Introdução

Eu

Arquivei

- Documentos Fiscais
  - Sefaz
  - XMLs
- Inteligência Fiscal
- +80.000 empresas
- +800 milhões de DFes

Time de DataEng

- Responsável por DWs
  - DFes
  - BI
  - Integrações
  - Mensageria
  - Governança
- Time independente
  - Gestão de Infra
  - Gestão de Custo
  - Priorizamos tecnologias gerenciadas

Engenharia de Dados

Libertando dados

Engenharia de Dados

- No mercado:
  - Junto com BI e DataSci
  - Modelagem de BD
  - Foco em Dados não em Sistemas
- Principal atividade:
  - Extração
  - Transformação
  - Carregamento
- Cargo ou Função?

Engenharia de Dados neste curso

- Responsável por:
  - Coleta e aquisição de dados
  - ETL
  - Padronizações
- Skills
  - Processamento massivo
  - Sistemas distribuídos

Coleta e Aquisição

Entendendo fontes

Aquisição e extração

- Pode ser ativo ou passivo
- Ativo
  - APIs
  - Scrapers/Crawlers
- Passivo
  - Filas
  - API

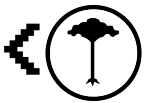
Ativo

- Entender as Fontes
  - Conhecer a doc de uma API
  - Saber fazer um crawler/scraper
  - Saber interagir com um banco
  - Limitações de escalabilidade da fonte
- Padronização por conta do Eng de Dados
- Entender lógicas de negócio
- Possivelmente não data-driven

Passivo

- Baseado em contratos
  - Maior controle de quem recebe
  - Quem envia se preocupa com o dado
- Pode ser reaproveitado para várias finalidades

Crawlers/Scrapers



Crawler/Scraper 101

- Scrapy
  - Passar por vários sites
- Selenium
  - Lib de QA
  - Permite interação com a página

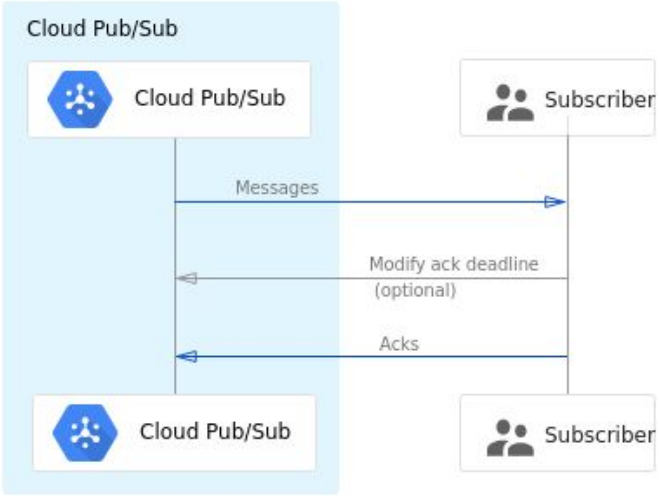
Filas

Conceitos

- Stream
  - Fluxo de dados / Pedaço de informação
- Streaming
  - Processamento contínuo
  - Banco de dados “infinito”
- Queue
  - Sequência ordenada
- Mensageria
  - Comunicação entre serviços
- Replicação de Dados
  - Processo complicado de cópia

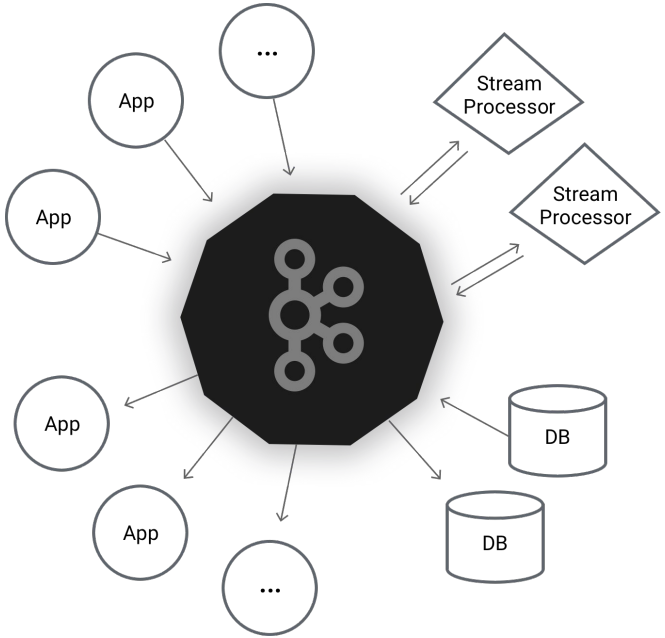
Pub/Sub

- Padrão de mensageria
- Publisher
- Subscriber
- Organizado em tópicos
- Gerenciado: Google Pub/Sub



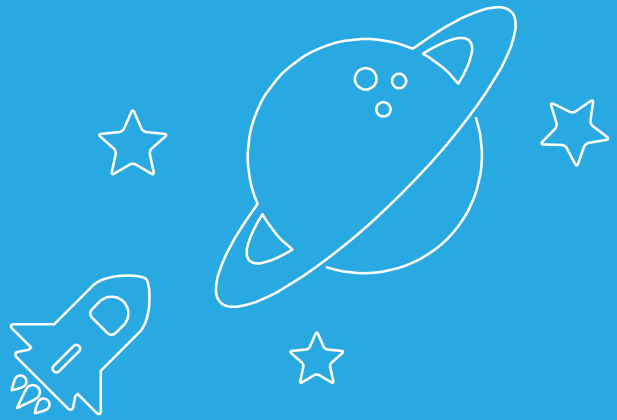
Kafka

- Ecossistema de ferramentas
  - Kafka
  - Kafka Streams
  - Kafka Connect

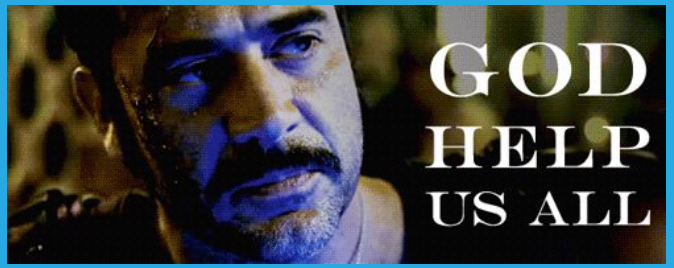


Kafka

- Tópicos
- Partições
- Grupos
- Streaming:
  - processamento + armazenamento



Demo



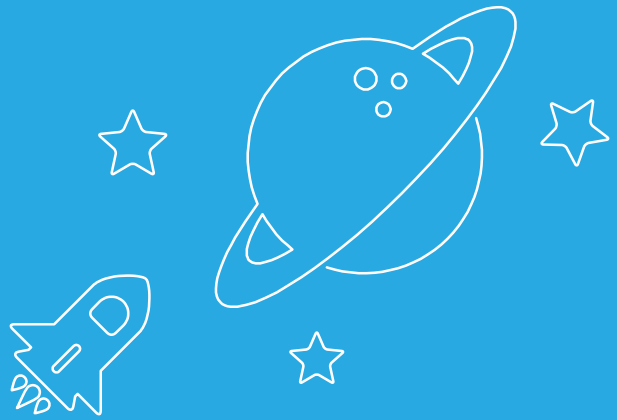
ETL 101

Python

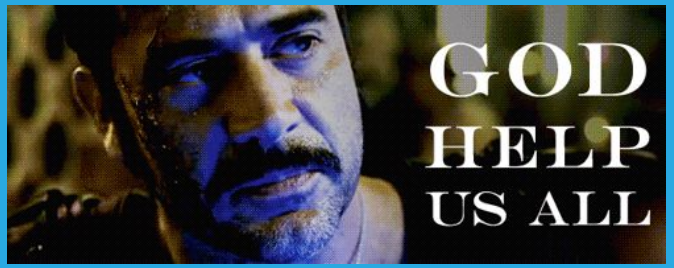
- Pandas
- PETL

PETL

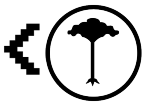
- Versão simplificada da Pandas
- Mais focado em ETL
  - Leitura de arquivos
  - Manipulação de dados



Demo

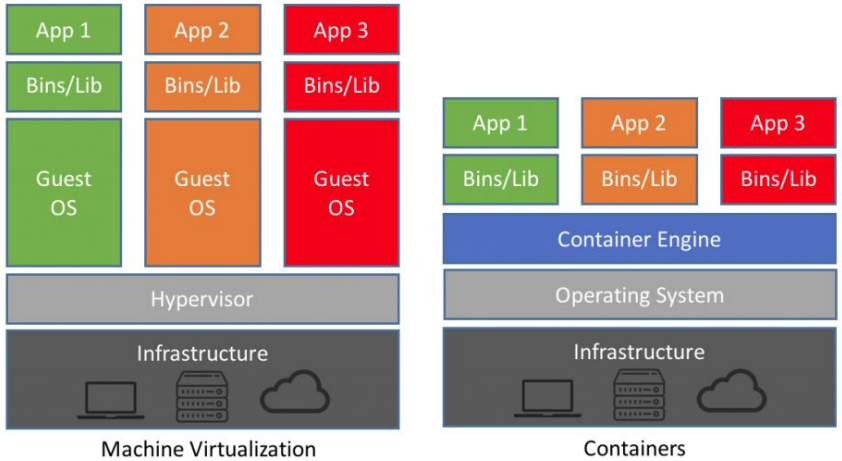


Docker



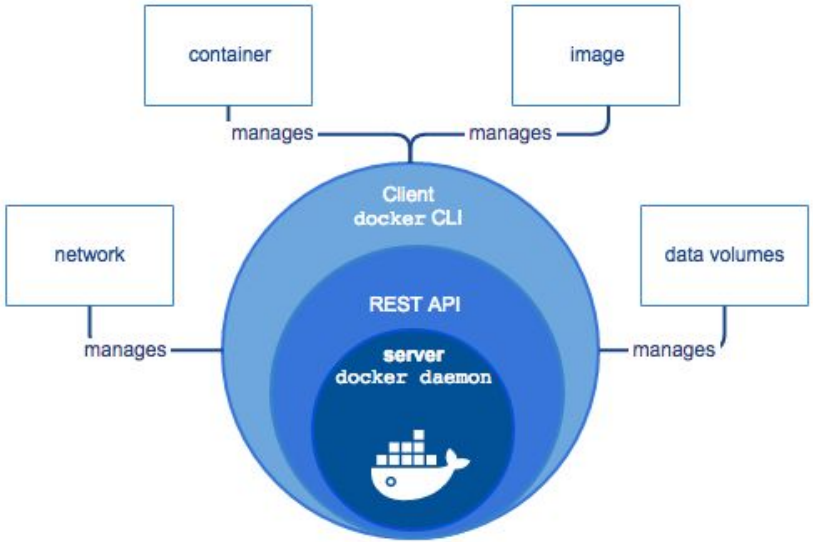
Containers

- Jeito de empacotar código
- Abstração do SO
- VM vs Container:



Conceitos

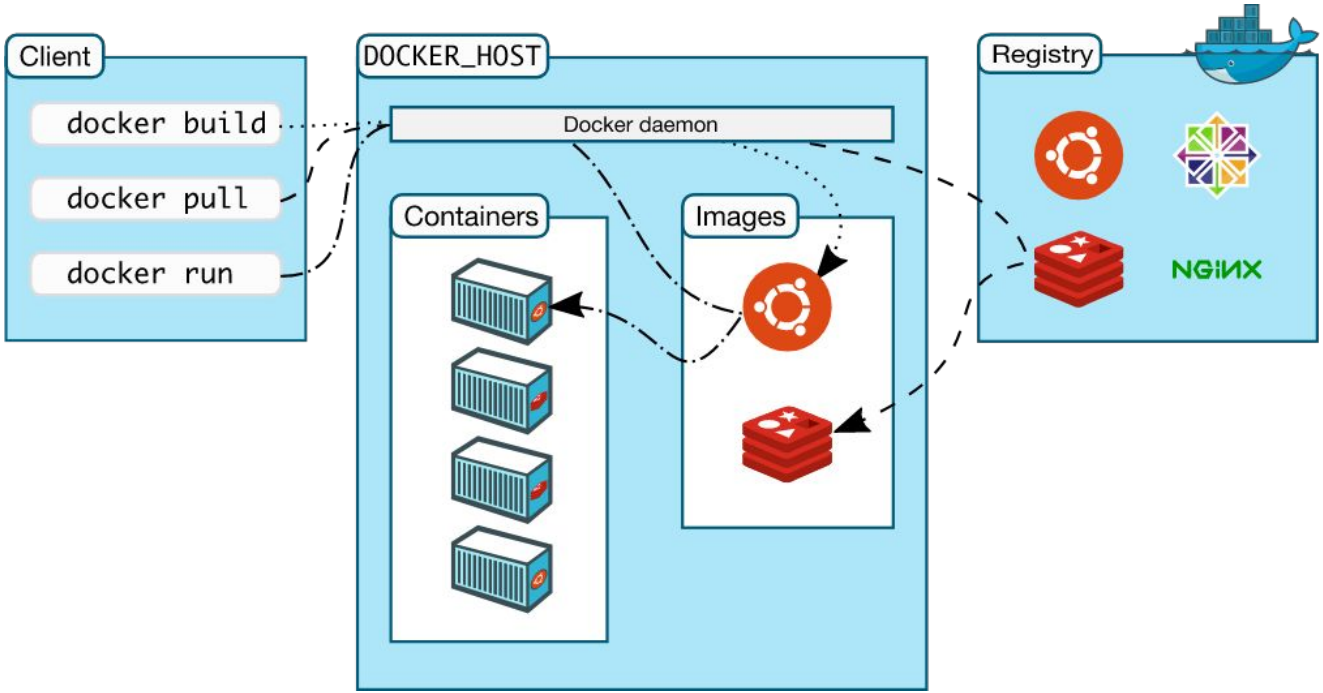
- Docker Engine
  - client-server



Conceitos

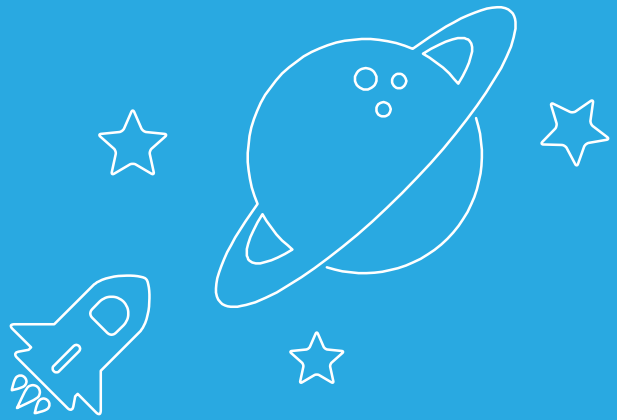
- Docker Daemon
- Docker Client
- Docker Registry
  - Imagens: intruções para criar um objeto
- Docker Objects
  - Container
  - Network
  - Volume

Conceitos

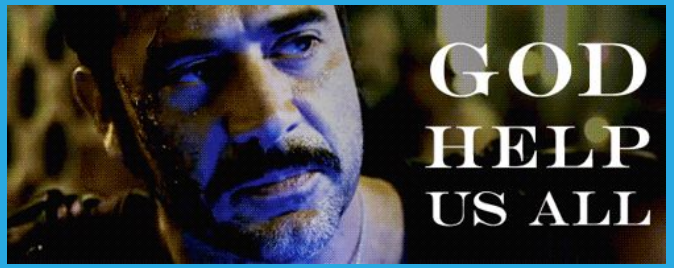

Client
docker build
docker pull
docker run
DOCKER_HOST
Docker daemon
Containers
Images
Registry
NGINX

Docker Compose

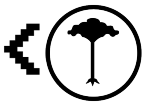
- Multi container
- Yaml
  - Define vários serviços



Demo

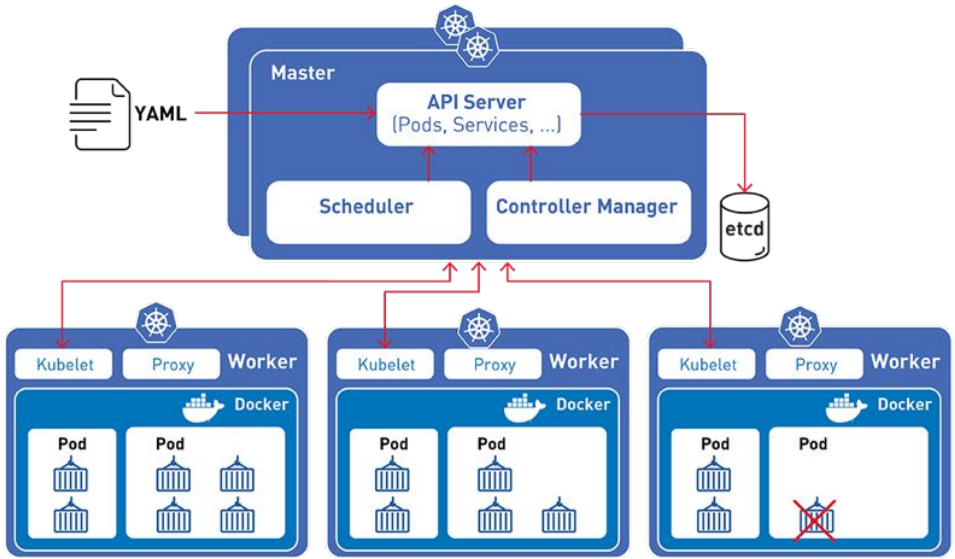


Kubernetes



Orquestração de Containers

- Gerenciador de containers distribuído
- Possui serviços gerenciados

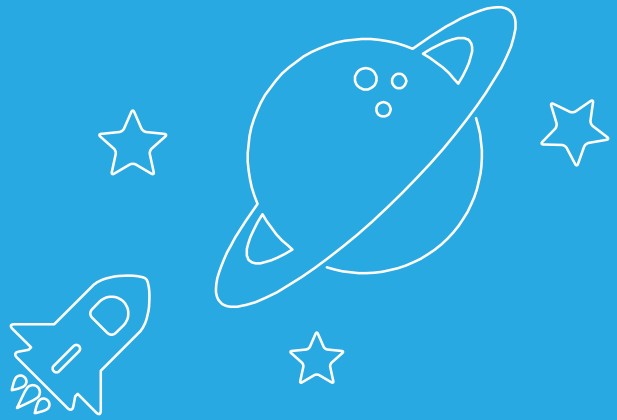


Objetos

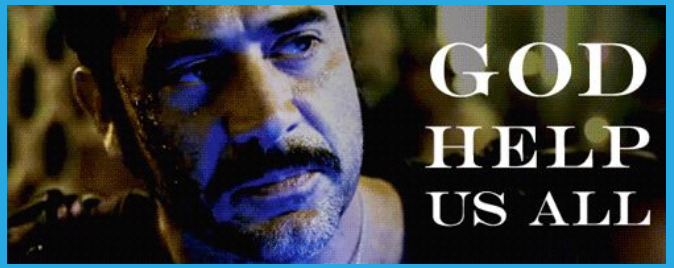
- Pod
- Service
- Volume
- Namespace

Abstrações

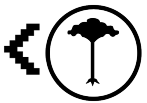
- Deployment
- Job
- DaemonSet
- StatefulSet



Demo



BigQuery UDF



UDF

- User Defined Functions
- Função JavaScript que transforma os dados
- ETL no BigQuery

That’s all folks