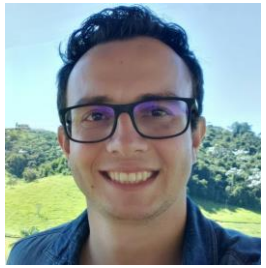
Spark – Big Data Processing

Aula 1

Semantix®

All about data

Quem sou eu?



Rodrigo Augusto Rebouças

Engenheiro de dados da Semantix
Instrutor do Semantix Academy

Contatos
rodrigo.augusto@semantix.com.br
linkedin.com/in/rodrigo-reboucas

RODRIGO.A.REBOUCAS

Semantix®
All about data

Ementa

- Projetos com Jupyter Notebooks - Python
- Operações com RDD
- Operações com Dataframe
- Operações com Dataset
- IDE – Python e Scala
- Struct Streaming - Kafka;
- Spark Streaming - Kafka;
- Otimizações e Tunning

Revisão

Spark – Básico (Big Data Foundations)

- spark-shell – Scala
- DataFrame
  - Transformação
  - Ação
  - Schemas
  - Join
- Spark SQL Queries
- API Catalog - Scala

Python vs Scala - Revisão

- Diferenças Scala para Python
  - DataFrame (Não precisa declarar variável/constante)
    - Transformação
    - Ação
    - ~~Schemas~~
    - ~~Join~~
  - Spark SQL Queries (Não precisa declarar variável/constante)
  - API Catalog

Preparar Ambiente de Desenvolvimento

- Instalação

Semantix®
All about data

Preparação Ambiente – Instalação Docker e Docker-compose

- Instalação
  - Docker: https://docs.docker.com/get-docker/
  - Docker Compose: https://docs.docker.com/compose/install/
  - SO
    - Windows
      - Docker Desktop (Hyper-V ou WSL2)
      - Docker Toolbox (VirtualBox)
    - Linux
      - Docker Engine
      - Docker Compose
    - Mac
      - Docker Desktop

Baixar Cluster de Big Data

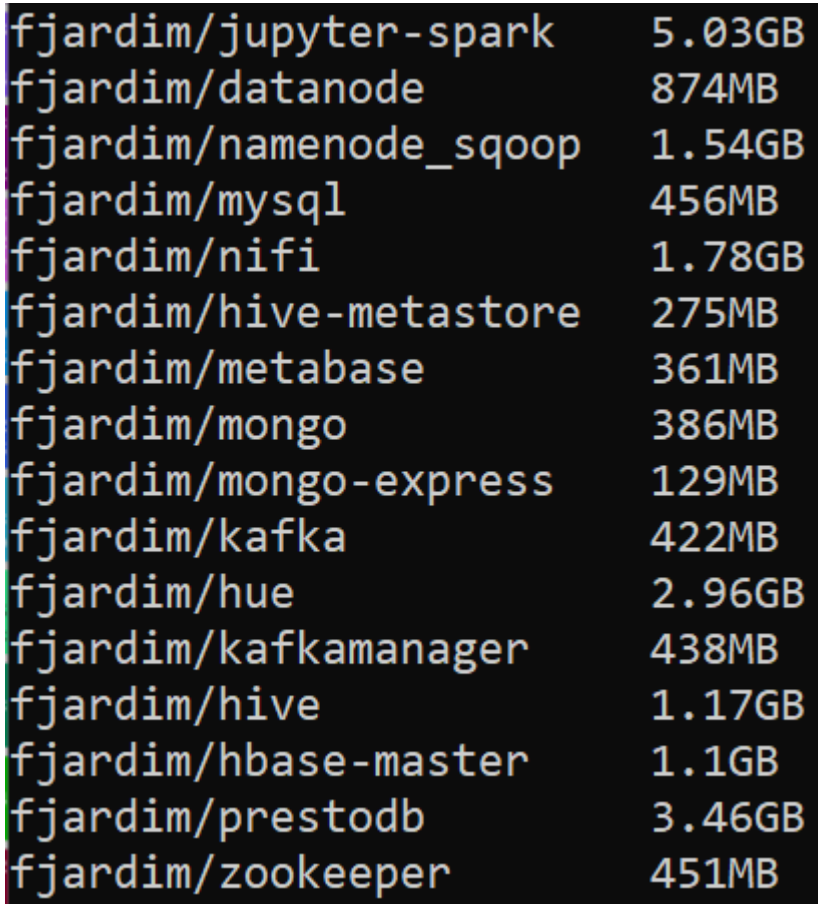
- Baixar conteudo do Cluster

  git clone https://github.com/rodrigo-reboucas/docker-bigdata.git spark
- Baixar as imagens

  docker-compose -f docker-compose-completo.yml pull
- Listar as imagens

  docker image ls
- Iniciar todos os serviços

  docker-compose -f docker-compose-completo.yml up –d



Semantix®
All about data

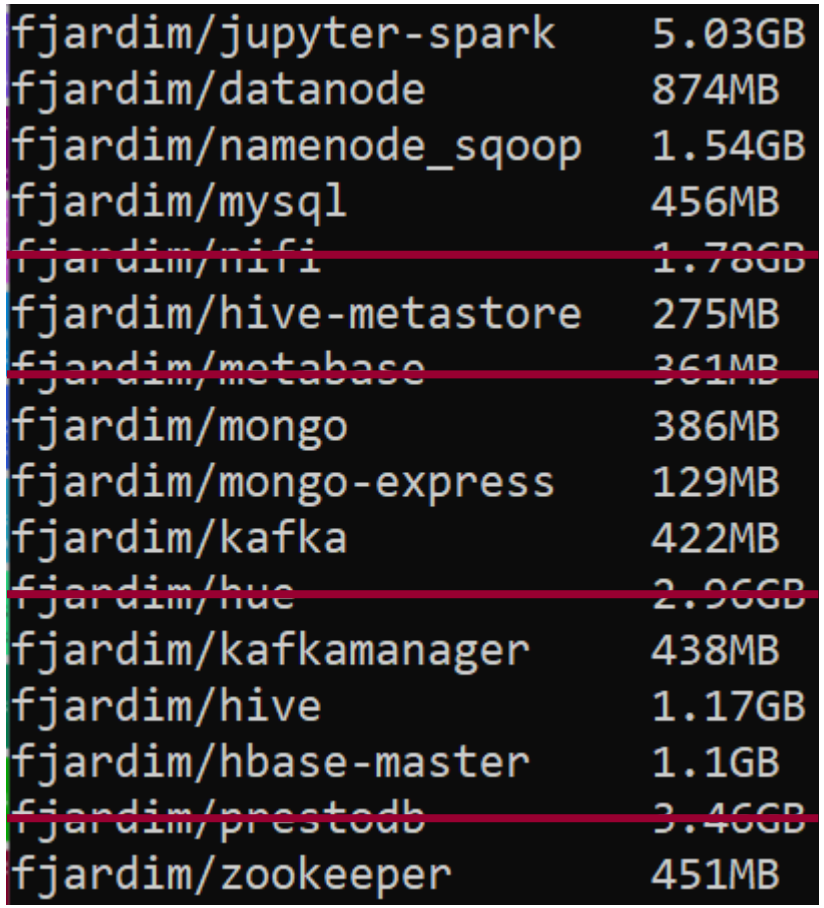
Baixar Cluster de Big Data - Parcial

- Arquivo: docker-compose-parcial.yml
  - Remover ~8GB
- Baixar as imagens

  docker-compose -f docker-compose-parcial.yml pull
- Listar as imagens

  docker image ls
- Iniciar todos os serviços

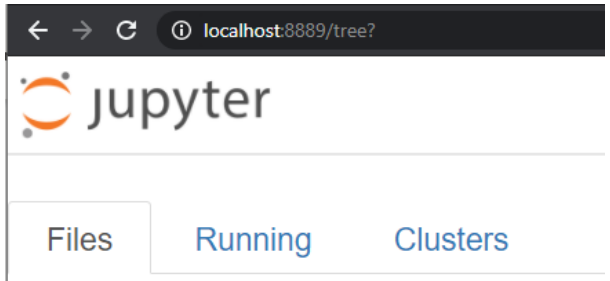
  docker-compose -f docker-compose-parcial.yml up –d



Semantix®
All about data

Exercícios - Instalação de Ambiente

1. Instalação do docker e docker-compose
2. Executar os seguintes comandos, para baixar as imagens do Cluster de Big Data:
   - git clone https://github.com/rodrigo-reboucas/docker-bigdata.git spark
   - cd spark
   - docker-compose –f docker-compose-parcial.yml pull
3. Iniciar o cluster Hadoop através do docker-compose
   - docker-compose –f docker-compose-parcial.yml up -d
4. Listas as imagens em execução
5. Verificar os logs dos containers do docker-compose em execução
6. Verificar os logs do container jupyter-spark
7. Acessar pelo browser o Jupyter, através do link:
   - http://localhost:8889



Semantix®
All about data

Semantix®

All about data

contato@semantix.com.br

www.semantix.com.br