



GERÊNCIA DE INFRAESTRUTURA PARA BIG DATA

Marcos Takeshi – Aula 01

Pós-Graduação em
Ciência de Dados e Inteligência Artificial

Ementa da disciplina

Introdução à arquitetura para Big Data Analytics. Visão geral sobre Infraestrutura de armazenamento de dados para Big Data. Visão geral sobre Infraestrutura de computação e de rede para Big Data. Tópicos sobre virtualização e computação em nuvem. Plataformas de Big Data na nuvem: HDFS, Hadoop e MapReduce. Estudos de caso com Spark.

Professores

MARCOS TAKESHI

Professor Convidado

Especialista em Big Data na Semantix, que atua em diversos projetos de empresas do setor financeiro, telecom, varejo e saúde. Realiza análises de arquiteturas, infraestruturas, ambientes, sistemas e ferramentas big data, visando o correto funcionamento e performance. Formado em engenharia eletrônica pela Escola de Engenharia Mauá, pós-graduado em Administração de Empresas pela FGV-SP, MBA em Big Data na FIAP, e empreendedorismo no Babson College.

TIAGO COELHO FERRETO

Professor PUCRS

É professor adjunto da Pontifícia Universidade Católica do Rio Grande do Sul. Possui Doutorado em Ciência da Computação pela PUCRS (2010) com Doutorado sanduíche na Technische Universität Berlin, Alemanha (2007-2008). Tem experiência na área de Ciência da Computação, com ênfase em Redes de Computadores, atuando principalmente nos seguintes temas: computação em nuvem, grades computacionais, virtualização, processamento de alto desempenho e gerência de infraestrutura de TI.

Encontros e resumo da disciplina

AULA 1

Para ser um profissional de Data Science é necessário ter paciência e construir um bom Network.

Empresas tem grande interesse em processar os dados e deles extrair informação com a finalidade de monetizar.

É bom estar no meio de pessoas que saibam mais do que você, sempre você tem que estar no meio de pessoas melhores.

MARCOS TAKESHI
Professor Convidado

AULA 2

O Spark possibilita a obtenção de resultados imediatos.

É importante você saber e conseguir atuar em mais de uma frente.

Certificações podem mostrar que você tem conhecimento do assunto.

MARCOS TAKESHI
Professor Convidado

AULA 3

Nos últimos anos a gente tem, a cada ano, um novo software auxiliando no processamento de grandes volumes de dados.

Além de armazenar e processar, eu tenho que conseguir extrair valor.

O Hadoop como a principal ferramenta para trabalhar com grandes volumes de dados.

TIAGO COELHO FERRETO
Professor PUCRS

AULA 4

A redundância garante a persistência da informação.

O HDFS é a principal fonte de dados de entrada e saída do Hadoop.

Como utilizar as aplicações Sqoop e Flume.

TIAGO COELHO FERRETO
Professor PUCRS

AULA 5

MapReduce uma solução de escalonamento e capacidade de processamento.

Hadoop Streaming como implementação de funções Map e Reduce em linguagens diferentes de Java.

O Pig como linguagem alternativa para programar MapReduce.

TIAGO COELHO FERRETO
Professor PUCRS

AULA 6

O Hive trabalha com a linguagem SQL com interações através de linhas de comando em formato shell.

O Spark tem como benefícios uma melhor performance, extensibilidade e melhor suporte para outros cenários.

O componente principal do Spark é o RDD (Resilient Distributed Dataset).

TIAGO COELHO FERRETO
Professor PUCRS

Gerência de Infraestrutura para Big Data



Marcos Takeshi

Passatempos

desenho

computadores
eletrônicos

videogames

Formação

eletrotécnica

engenharia eletrônica

administração de empresas

Projetos

internet banking (PF & PJ)

wap banking

email banking

publishing

plataforma de impressão

armazenamento de documentos

sistema de empréstimos

sistema de financiamentos

sistema de comissões

Formação+

eletrotécnica

engenharia eletrônica

administração de empresas

mba big data

empreendedorismo

Paciência

MBA
(início)

Abril
2015



candidatura

Fevereiro
2016

Março
2016

MBA
(fim)

Junho
2016

extensão

Julho
2016

contratação

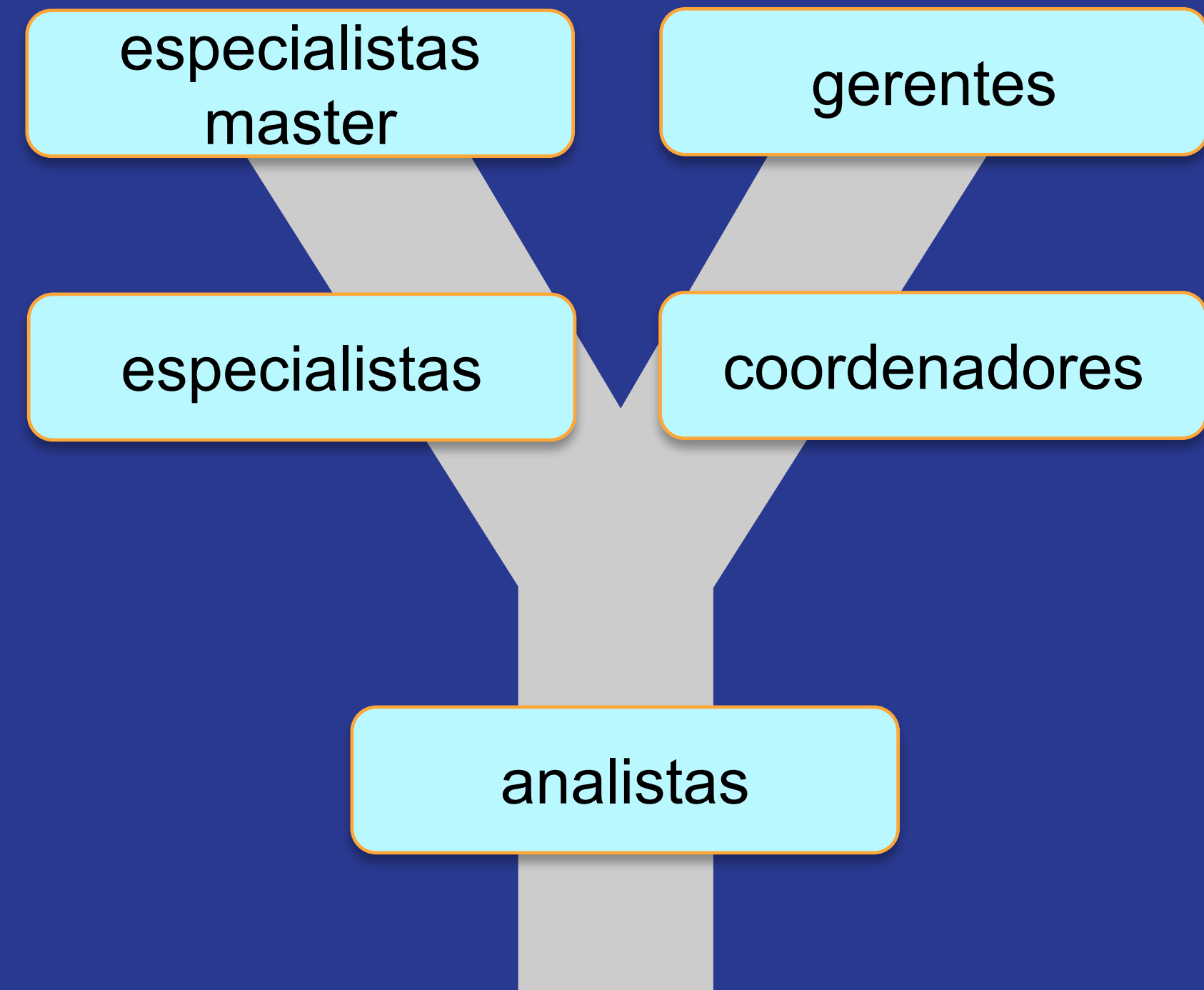
Setembro
2016

Big Data Master Specialist

aws certified big data specialty
cloudera spark & hadoop developer
cloudera data analyst
elastic certified engineer
cloudera instructor

Master Specialist

- Perfil Técnico
- Referência Técnica
- Compartilha conhecimento
- Atua em diversos papéis
- Mobilidade entre projetos
- Mentor da equipe

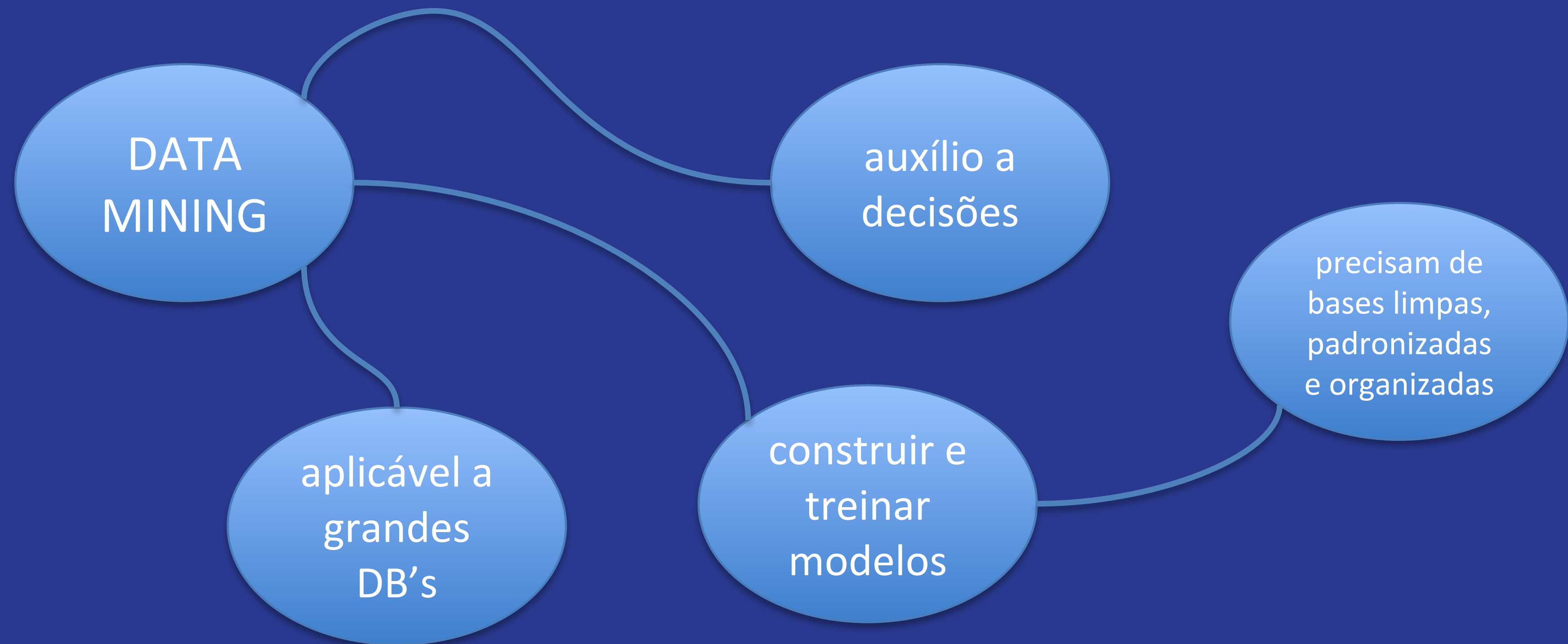


DATA is the new OIL

we need to find it
extract it
refine it
distribute it
and monetize it

Data Mining

“Uso de técnicas, preferencialmente automáticas, de exploração de grandes quantidade de dados de forma a encontrar padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nú pelo ser humano” – (Carvalho, 2001)




Dado: valor sem significado

Informação: dado com significado

Conhecimento: informação estruturada e contextualizada





data: 04/06/2019
hora: 23:28:43
temperatura: 15 °C
latitude: -12.1316898
longitude: -77.0303418

ETL

Empresas têm grande interesse em:

processar dados, extrair informações, fazer previsões, agrupar pessoas por diversos fatores, etc.

Aplicando Conhecimento

Dados > Informação > Conhecimento > Sabedoria

Business Intelligence

Queries (select) / Relatórios

Ferramentas de Visualização / Seleções Especiais

BSC / KPI / Performance Management

Business Analytics

Data Mining
Modelos de Segmentação
Experimentação
Simulações
Risk Management
Modelos Preditivos

Volume de Dados

Com meios de comunicação mais rápida e utilização intensa de sites e aplicativos, a quantidade de dados aumentou.

As tecnologias e recursos para armazenamento e processamento também evoluíram.



Big Data

Big Data

É um termo bastante utilizado atualmente e diretamente ligado ao volume de dados a ser capturado, processado e analisado

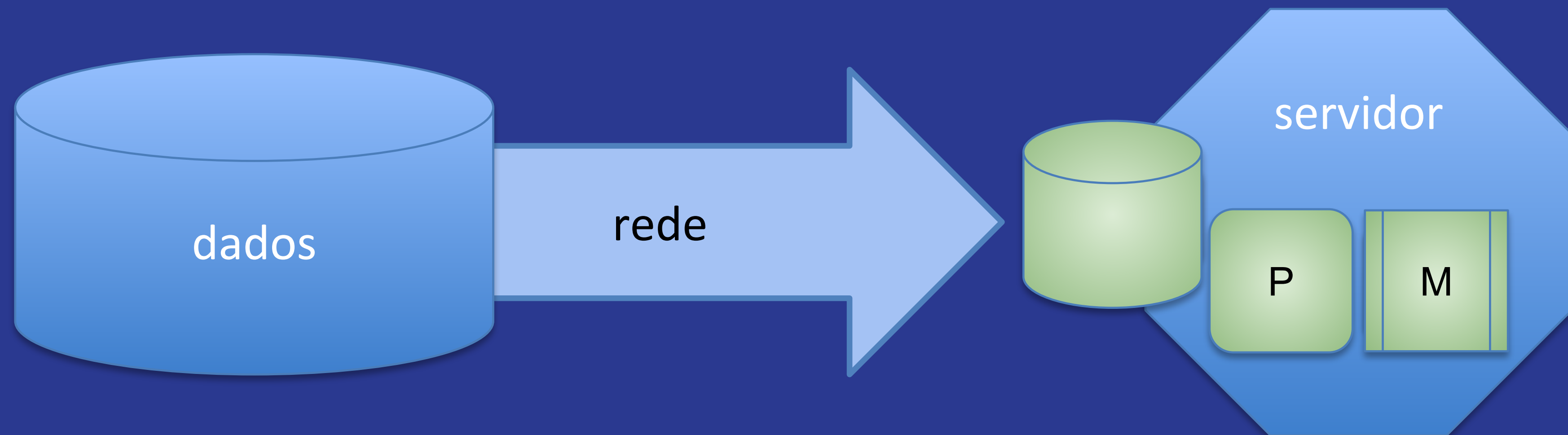
Data Science

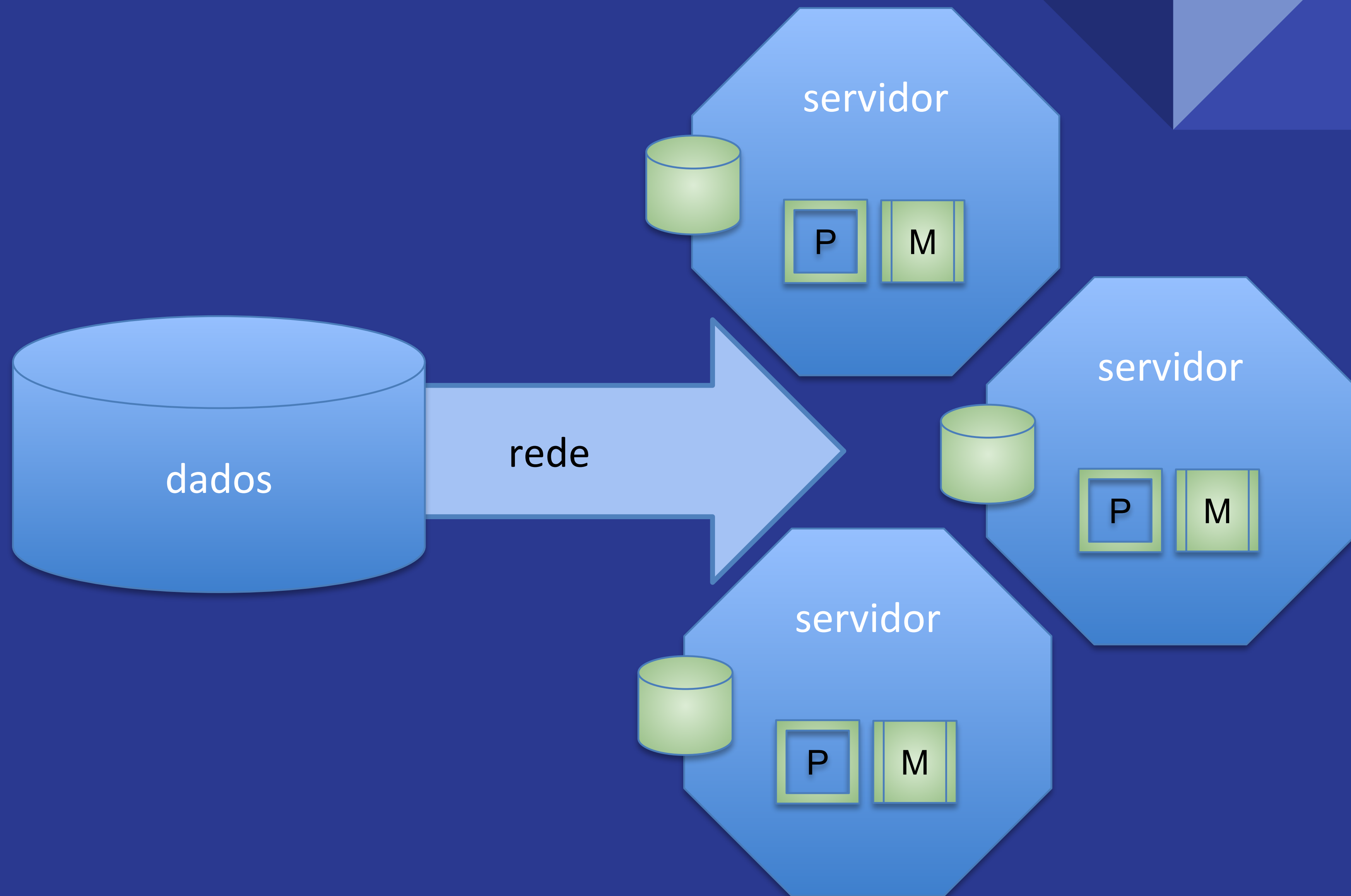
Prática que envolve métodos científicos, processos e sistemas para extrair conhecimento de dados

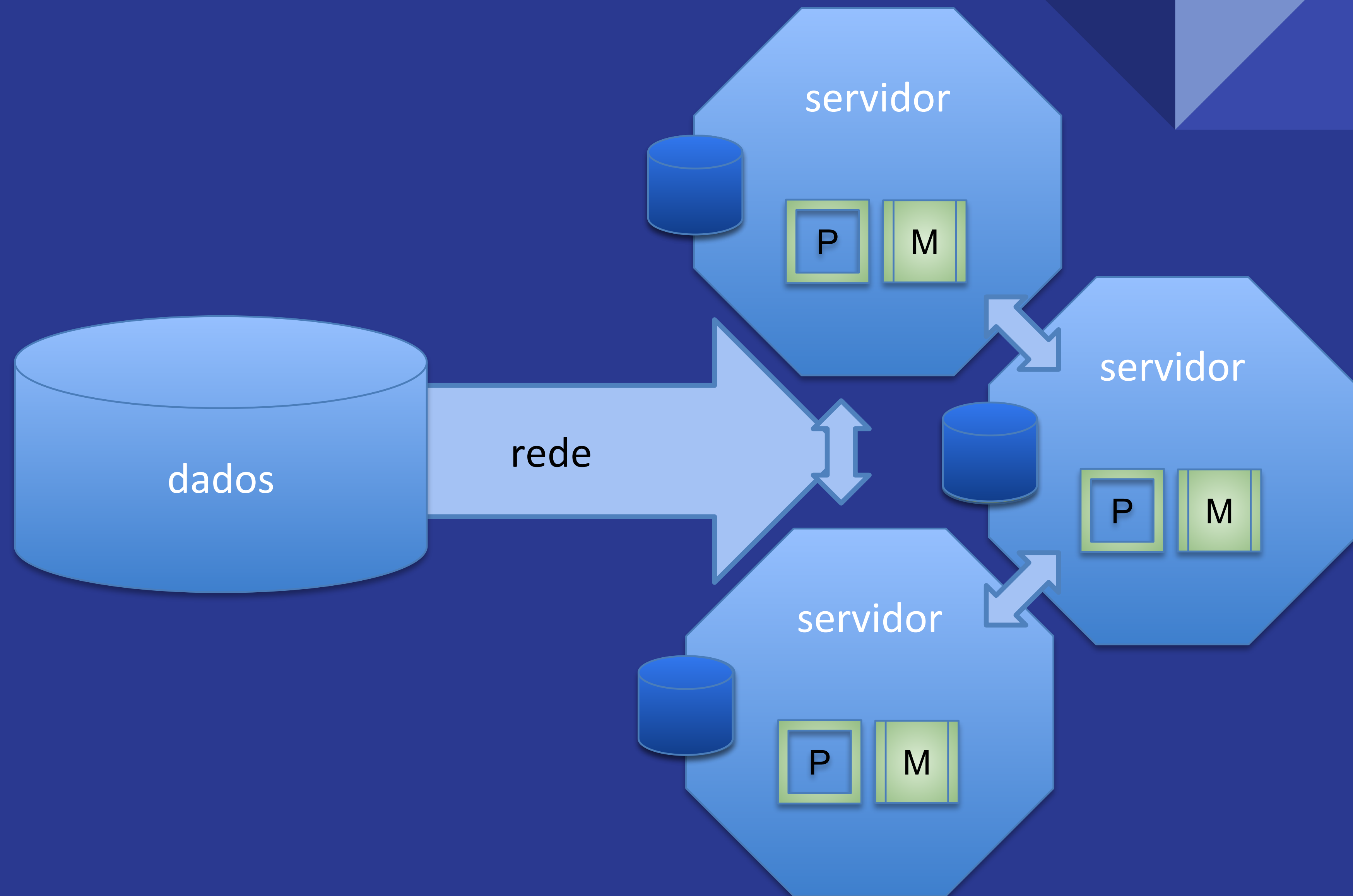
Recursos

Processador
Memória
Disco

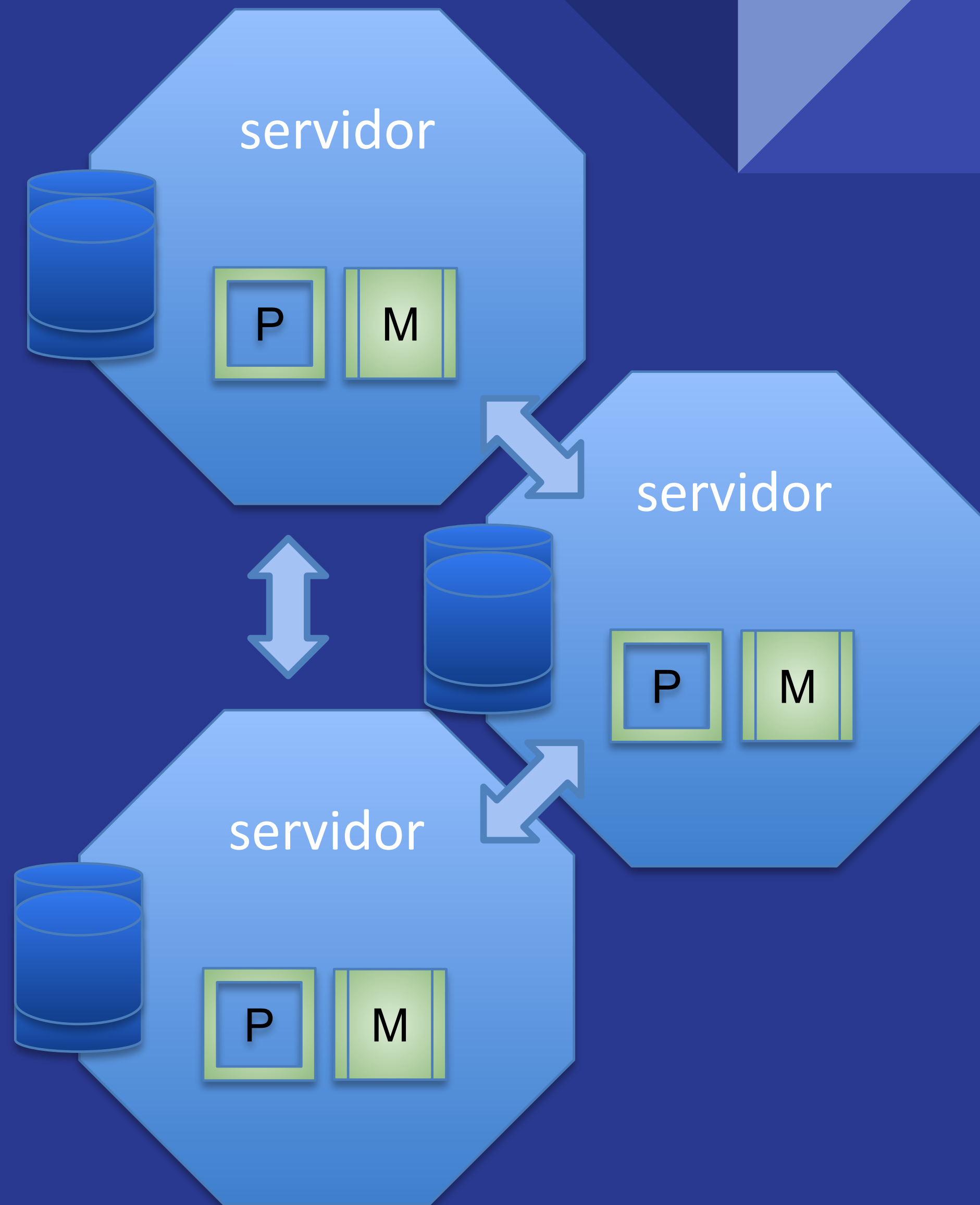
Processamento







Arquitetura Distribuída



Configurações (On premises)

2 x 8 cores = 16 cores
96Gb RAM
12 x HDD 4Tb

2 x 10 cores = 20 cores
256Gb RAM
12 x HDD 6Tb

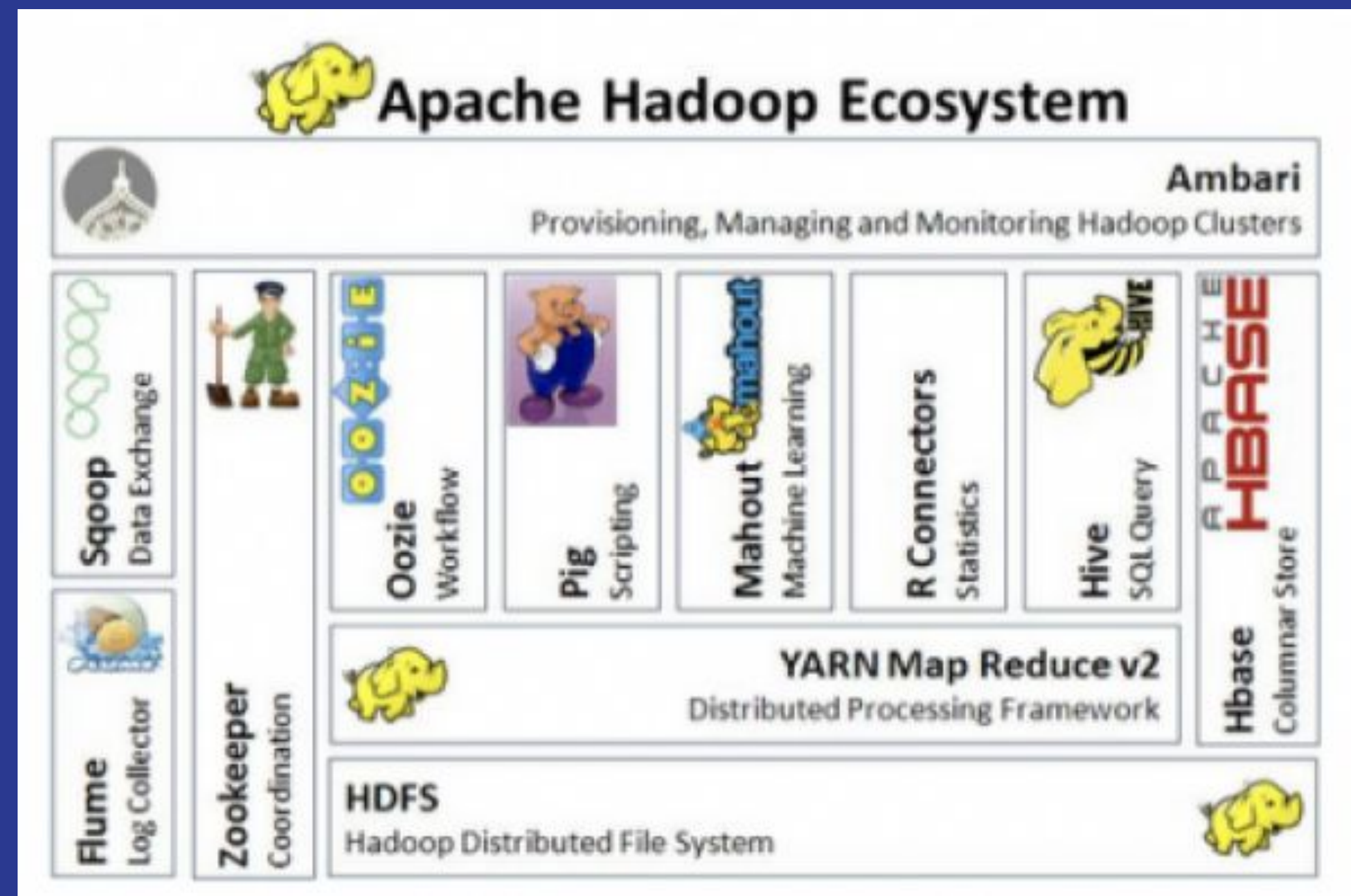


Hadoop

Hadoop

Framework que auxilia no processamento distribuído de datasets em clusters e que utiliza modelos simples de programação

Ecosystem



Ecosystem

Oozie

Pig

Mahout

Ambari

Flume

Sqoop

Zookeeper

HBase

Solr

Yarn

Hive

HDFS

Sistema Operacional

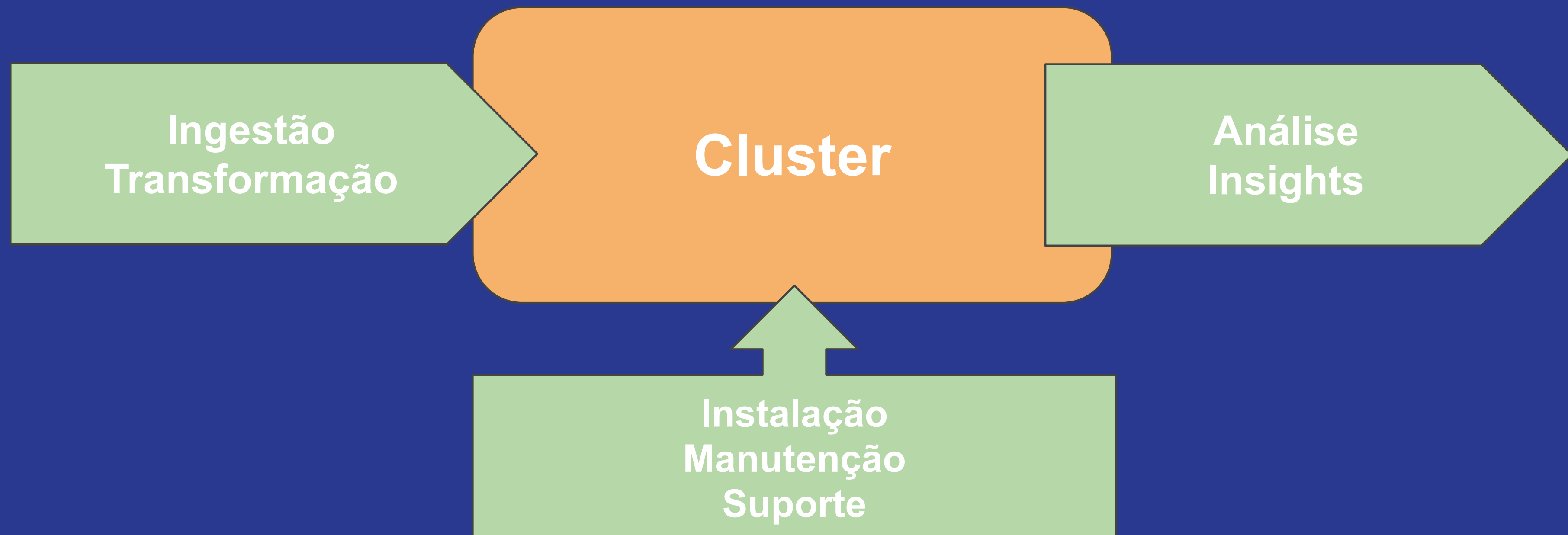
LINUX



CentOS / RedHat



Atividades





Docker

Instalação

<https://www.docker.com/products/docker-desktop>

Imagem

<https://hub.docker.com/>

busca: bde2020/hive



bde2020/hive

By [bde2020](#) • Updated 2 years ago

Docker container for Apache Hive with hiveserver2

Container

Linux

x86-64

Docker pull

> docker pull bde2020/hive

Docker Pull Command

```
docker pull bde2020/hive
```



Imagem

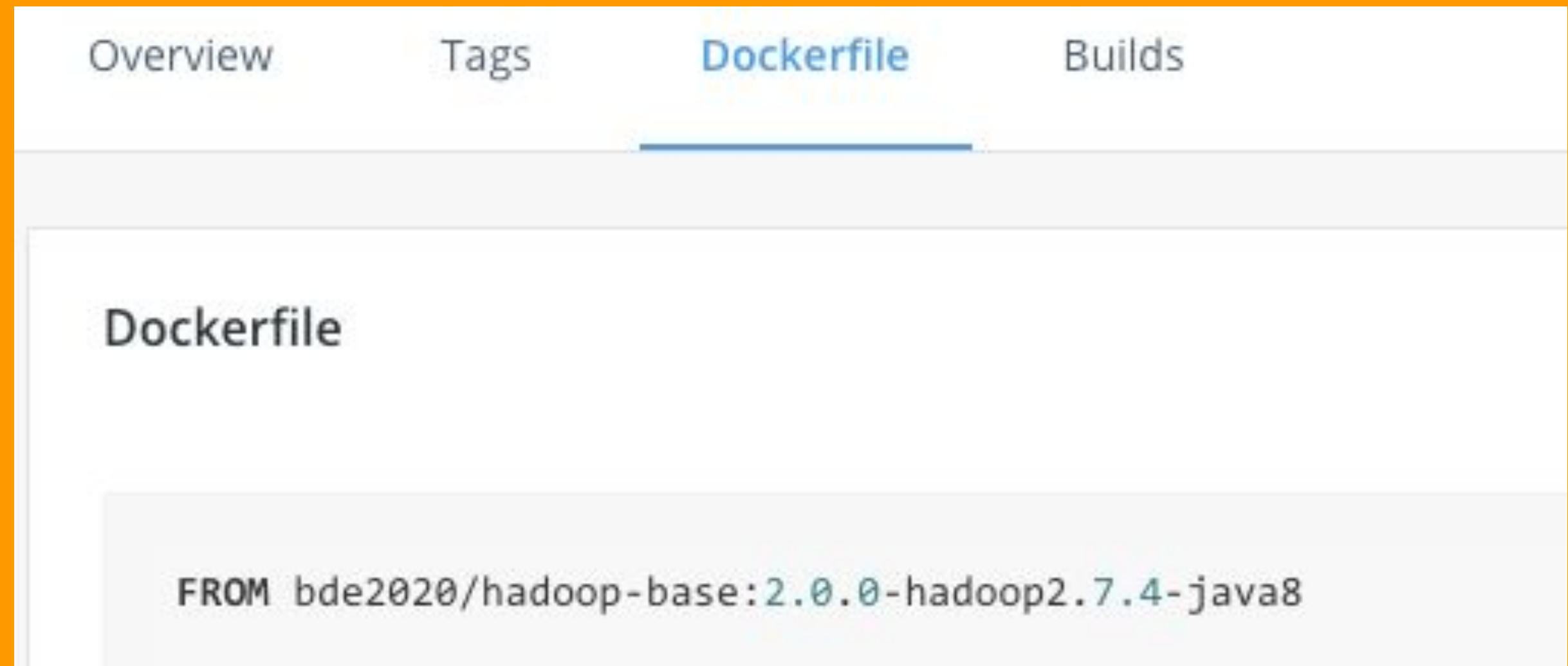
> docker image ls

```
TAKMBPi715:hive takeshi$ docker image ls
```

REPOSITORY	TAG	IMAGE ID	CREATED	SIZE
bde2020/hive-metastore-postgresql	2.3.0	7ab9e8f93813	9 months ago	275MB
bde2020/hive	latest	a65dc394c508	2 years ago	1.17GB
bde2020/hadoop-datanode	2.0.0-hadoop2.7.4-java8	d96116df9f46	2 years ago	874MB
bde2020/hadoop-namenode	2.0.0-hadoop2.7.4-java8	23d8c9a8ce60	2 years ago	874MB
bde2020/hive	2.3.2-postgresql-metastore	87f5c9f4e2df	2 years ago	1.17GB
shawnzhu/prestodb	0.181	7cc5e6c14cc8	3 years ago	3.46GB

Arquivo Dockerfile

> docker build -t "hive:hive" .



Container Start

> docker-compose up -d



Containers

> docker container ps

```
TAKMBPi715:hive takeshi$ docker container ps
```

CONTAINER ID	IMAGE	COMMAND	CREATED
175938e8a7c1	bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8	"/entrypoint.sh /run..."	33 hours ago
f5a97eb33704	bde2020/hive:2.3.2-postgresql-metastore	"entrypoint.sh /bin/..."	33 hours ago
2f58c0526d55	bde2020/hive-metastore-postgresql:2.3.0	"/docker-entrypoint..."	33 hours ago
56913a59d7c5	bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8	"/entrypoint.sh /run..."	33 hours ago
df17e8969457	bde2020/hive:2.3.2-postgresql-metastore	"entrypoint.sh /opt/..."	33 hours ago
3633e3db6bd6	shawnzhu/prestodb:0.181	"./bin/launcher run"	33 hours ago

Acessando o Container

> docker-compose exec hive-server bash

```
TAKMBPi715:hive takeshi$ docker-compose exec hive-server bash
root@f5a97eb33704:/opt#
root@f5a97eb33704:/opt# pwd
/opt
root@f5a97eb33704:/opt# ls -lha
total 28K
drwxr-xr-x 1 root  root  4.0K Feb  5  2018 .
drwxr-xr-x 1 root  root  4.0K Dec 12 12:53 ..
drwxr-xr-x 1 20415 input 4.0K Feb  5  2018 hadoop-2.7.4
drwxr-xr-x 1 root  root  4.0K Feb  5  2018 hive
root@f5a97eb33704:/opt#
```



Professionais

DevOps

unix/linux
docker/ puppet/ terraform/ kubernetes
virtualização
integração contínua
redes, roteadores, switches
servidores, containers, cloud ...



DevOps

Data Engineer

programação scala, python, java
spark
HDFS
hive, kudu, impala, cassandra
flume/kafka ...



Data
Engineer

Data Scientist

estatística

R, Jupyter, Zeppelin

regressões, time series, clusterização

visualização

grafos

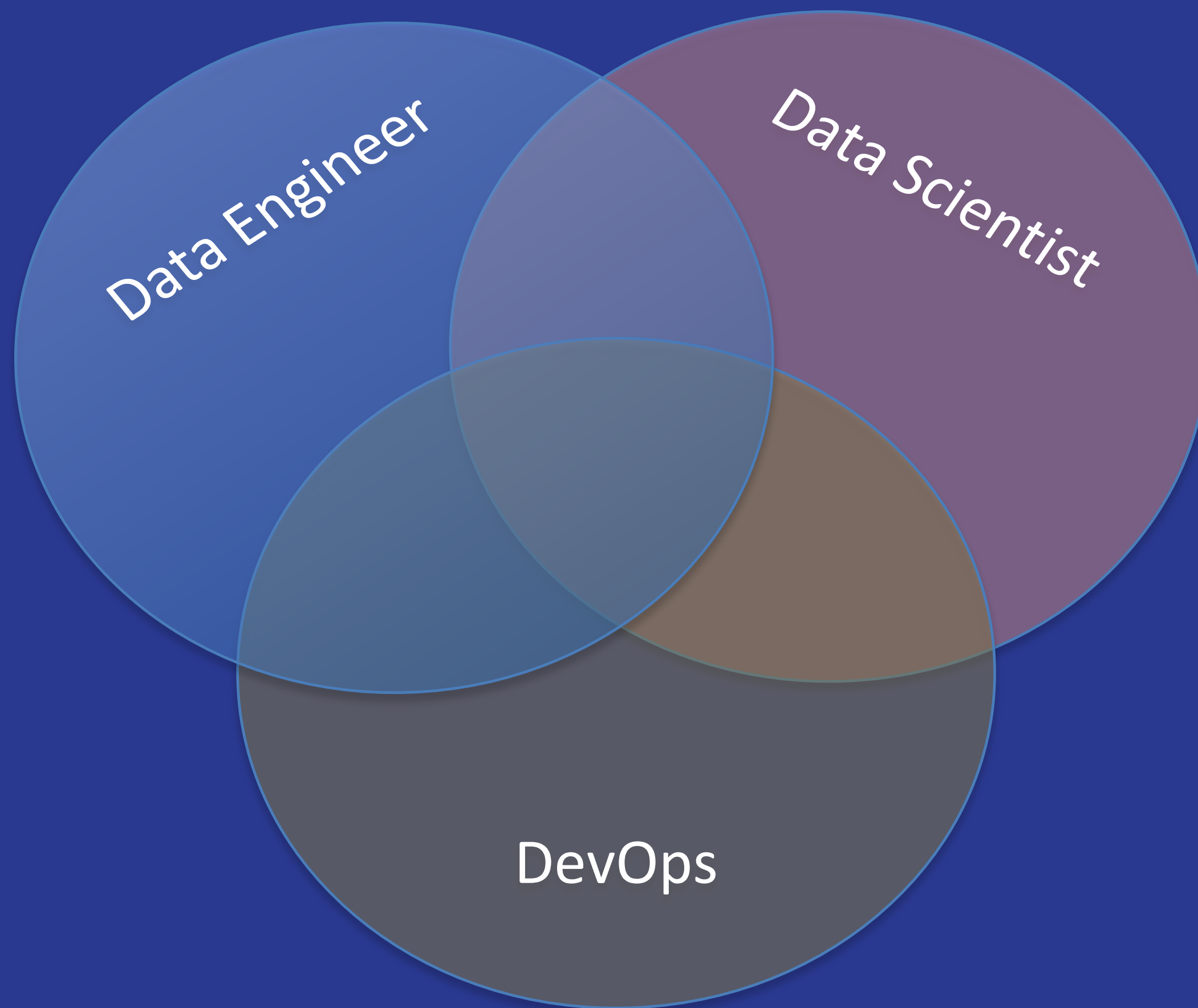
árvore de decisão, random forest, gradient boosting ...



Data
Scientist

Conhecimentos

HDFS
Flume
Kafka
Hive
Impala
Spark
Scala
Python
Linux
:



HDFS
Jupyter
Zeppelin
R
Python
Scala
Spark
Linux
:

Hadoop | HDFS | Linux | Python | Java (UDF) | ..



Linux

Comandos

> docker-compose exec hive-server bash

> pwd

> ls -lha

> cd <diretório>

> mkdir / rmdir

> cp <origem> <destino>

> rm <arquivo>

Prática

Verificar existência do arquivo kv1.txt (diretório /opt/hive/examples/files/)

Criar diretório com seu nome dentro do diretório /opt

Copiar arquivo kv1.txt para o seu diretório com o nome kv2.txt

Verificar o conteúdo do arquivo kv2.txt

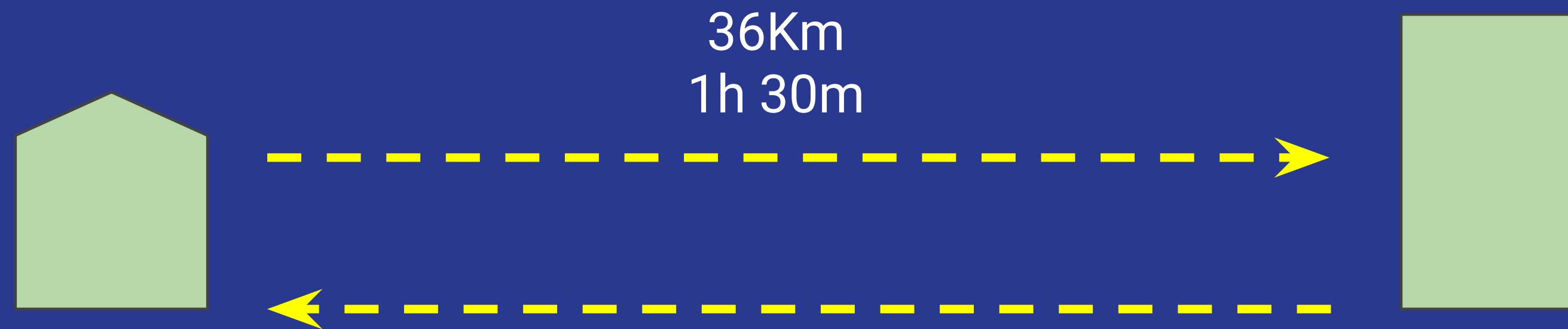
Alterar permissões do arquivo kv2.txt

Case 1

Financeiro

Aprendizado

Financeiro



9 meses
ambiente corporativo
certificação hadoop + aws developer



HDFS

Comandos

- > hdfs
- > hdfs dfs
- > hdfs dfs -ls /
- > hdfs dfs -mkdir <diretório>
- > hdfs dfs -put <arquivo> <diretório>
- > hdfs dfs -get <arquivo> <diretório>
- > hdfs dfs -cat <arquivo>
- > hdfs dfs -cp <origem> <destino>

Prática

Criar diretório com seu nome dentro do diretório /user no HDFS

Colocar o arquivo kv2.txt (da prática anterior) no diretório do HDFS com nome kv3.txt

Verificar o conteúdo do arquivo kv3.txt (no HDFS)

Criar diretório "teste" dentro do seu diretório

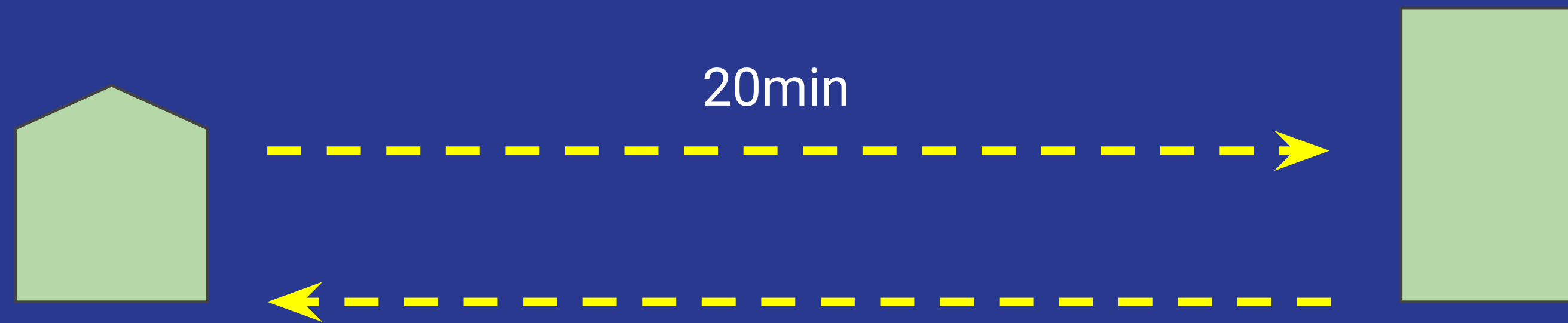
Copiar o arquivo kv3.txt para o diretório "teste"

Case 2

Marketing

Python, Pyspark + Algoritmo

Marketing



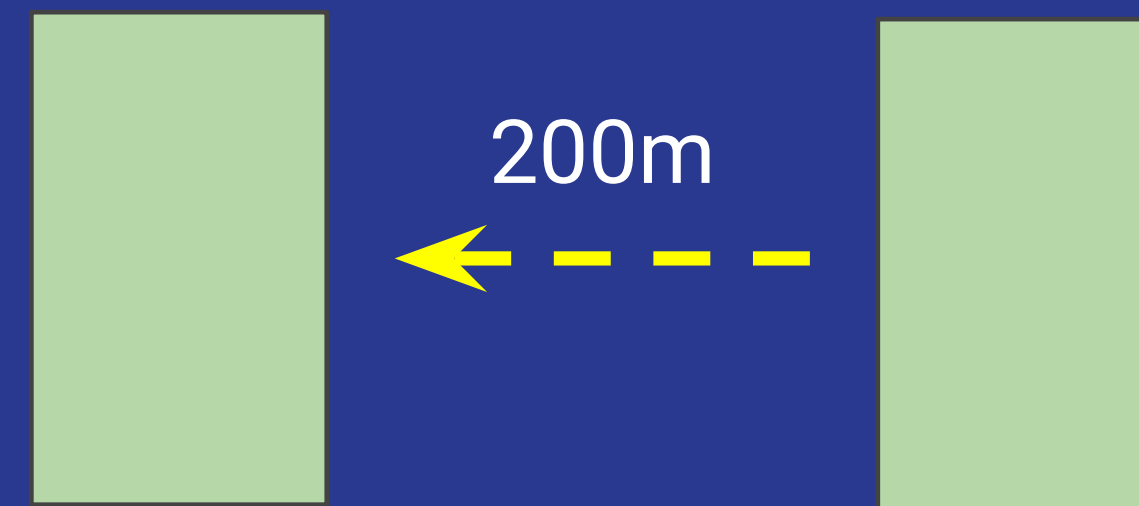
3 meses
ambiente descontraído
clusters exterior + cloud (API + algoritmo)
mentoria

Varejo

Java + Impala

Varejo

1 a 2 semanas
à noite
debug (aplicação e processos)
contrato





Hive

Comandos

- > /opt/hive/bin/beeline -u jdbc:hive2://localhost:10000
- > show databases;
- > show tables;
- > create table indicadores (cod int, valor string);
- > show create table <tabela>;
- > load data local inpath '/opt/nome/kv2.txt' overwrite into table indicadores;
- > select * from indicadores;
- > select count(*) from indicadores;

Prática

Criar e verificar criação da tabela indicadores

Verificar location no HDFS

Carregar arquivo kv2.txt na tabela indicadores

Verificar location no HDFS

Apagar a tabela indicadores

Verificar location no HDFS

Case 3



Telecom

Consultoria + Tuning

Telecom

3 meses

ambiente corporativo

contato com arquiteto, gerente e diretor

maior cluster Latam

mentoria

treinamento exterior



Hive

Comandos

- > /opt/hive/bin/beeline -u jdbc:hive2://localhost:10000
- > show databases;
- > show tables;
- > create external table indicadores2 (cod int, valor string);
- > show create table <tabela>;
- > load data inpath '/opt/nome/teste/kv3.txt' into table indicadores2;
- > select * from indicadores;
- > select count(*) from indicadores;

Prática

Criar e verificar criação da tabela não gerenciada chamada indicadores2

Verificar location no HDFS

Carregar arquivo kv3.txt (HDFS) na tabela indicadores2

Verificar location no HDFS

Apagar a tabela indicadores2

Verificar location no HDFS

Case 4



Financeiro

Performance

Financeiro

3 meses + 2 meses
ambiente corporativo

processamento da base de
clientes de todos os sistemas
para geração de relatório onde
o tempo de processamento
estava entre 3h e 4h (cada)

após adaptação de códigos e
ajustes na chamada da execução,
tempo de processamento entre
10min e 15 min.

Produto

AWS & Spark

Produto

Algoritmo Machine Learning
execução em 8h ...

Após ajuste: 7min



Spark

O que é

Spark é um framework open source
criado em 2009
para processamento paralelo.

É uma ferramenta que complementa
o Hadoop, e não o substitui.

Finalidade

melhorar performance do processamento distribuído, evitando gravação em disco (I/O limitado), mantendo informações em memória (velocidade alta) e oferecendo flexibilidade na programação

Preparação

- JAVA_HOME
- SPARK_HOME
- HADOOP_HOME (windows)
winutils.exe
- PATH
JAVA_HOME/bin
SPARK_HOME/bin

APIs



8+



2.11+



2.7+/3.4+

Componentes

Spark SQL (Shark)

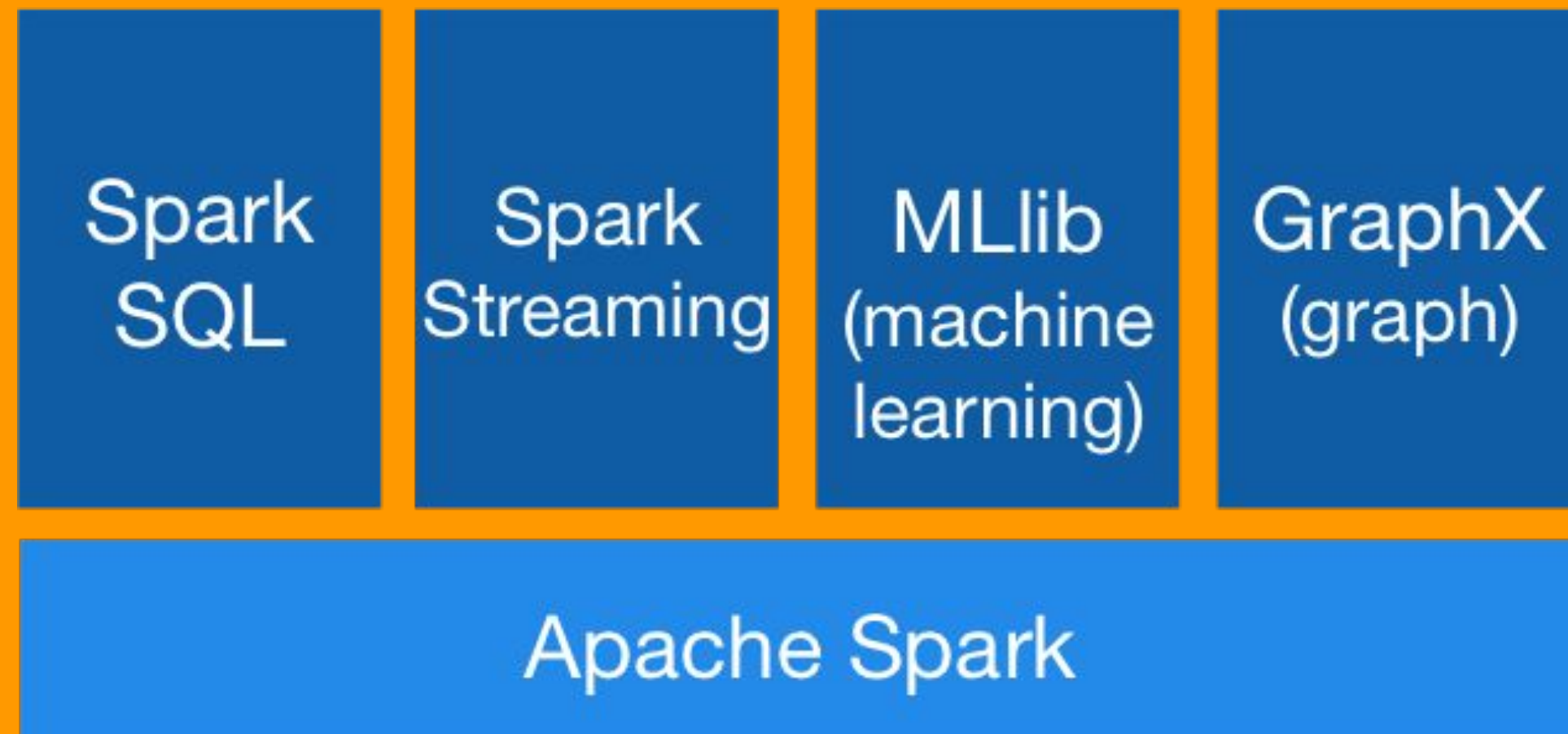
Spark Streaming

MLlib

GraphX

SparkR

Spark (Core)



Spark Core

Mais de 80 operadores para construir aplicativos

Scala, Python, R

Funções básicas e gerenciamento de memória

HDFS, File System, S3 (Amazon)

Hadoop, Mesos, Standalone, Cloud

<http://hadoop.apache.org/>

<http://spark.apache.org/>

Fluxo Spark

Aplicações spark começam com leitura de dados

Terminam com gravação dos resultados

HDFS/ Hive/ JDBC/ S3/ Hbase/ Cassandra/ ElasticSearch

CSV/ JSON/ Parquet/ ORC

Spark Shell

Shell REPL

Read – Eval – Print Loop

spark-shell

pyspark

```
Spark context Web UI available at http://10.1.1.179:4040
Spark context available as 'sc' (master = local[*], app id = local-1607972834964).
Spark session available as 'spark'.
Welcome to

      /---\         /---\
     /   \       /   \
    /     \     /     \
   /       \   /       \
  /         \ /         \
 /           \|         \|
/_-----/_ . _--/\_/_/_/_/_/_/_/_/_/_/_/_/_/_/_ version 2.2.0
 /_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_
  /_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_/_

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_162)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

PUCRS online  **UOL** edtech_