PUCRS online

APRESENTAÇÃO
DE APOIO

GERÊNCIA DE INFRAESTRUTURA PARA BIG DATA

Marcos Takeshi – Aula 02

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).

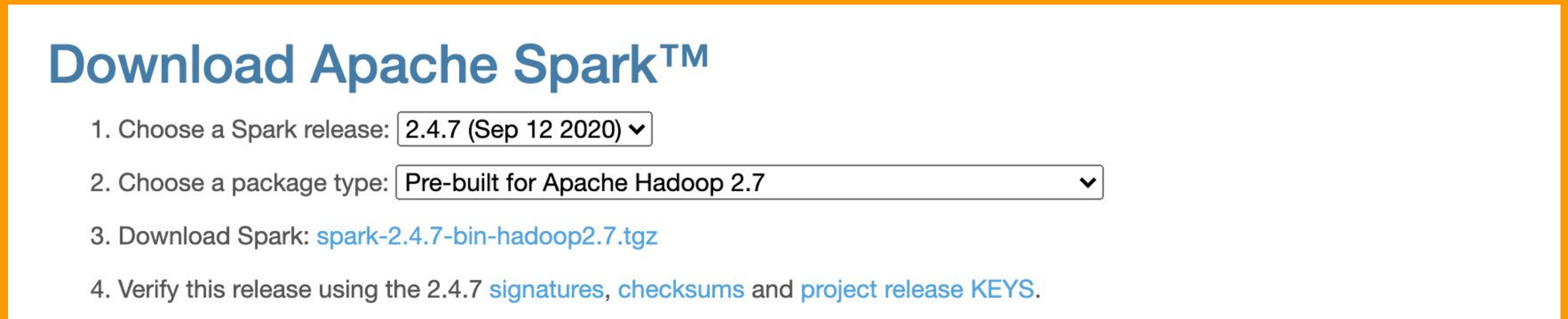
TIAGO COELHO FERRETO
Professor PUCRS

Instalação

> apt-get update

> apt-get install wget vim

Download do pacote
https://spark.apache.org/downloads.html



Instalação (cont)

> wget https://downloads.apache.org/spark/spark-2.4.7/spark-2.4.7-bin-hadoop2.7.tgz

> tar zxvf spark-2.4.7-bin-hadoop2.7.tgz

Copiar arquivo hive-site.xml para /conf do spark

Incluir spark no PATH

Comandos

> spark-shell

```
scala> spark.sql("show databases")
scala> spark.sql("show databases").show
scala> spark.sql("use default")
scala> spark.sql("show tables")
scala> spark.sql("select * from indicadores2").show
scala> spark.sql("select count(*) from indicadores2").show
```

Case 5

Governo

Elasticsearch

Governo

1 semana + 1 semana
Elasticsearch
API indexação PDF
Plugin leitura HTTP

Spark - parte 2

Declaração de Variáveis

```
val exemplo1 = "PUC"

exemplo1 += "RS"


var exemplo1 = "PUC"

exemplo1 += "RS"
```

Prática

```
scala> val numero = 1
scala> val b1:Byte = 127
scala> val c1: Char = 'm'
scala> val d1 = 123.45
scala> val long1 = 789L
scala> val f1 = 123.45f
scala> val bol1 = 5 > 1
```

Prática

```
scala> def verificaMultiplicador(num:Int): Int  = {
            if    (num      %    2     ==   0)   {
                        return num * 2
            }     else {
                        return num * 3
            }
        }
scala> verificaMultiplicador(10)
```

Prática

Leitura do arquivo kv3.txt no spark-shell

Visualização do conteúdo carregado no spark-shell

Gravação do conteúdo na location da tabela indicadores2

Consulta da tabela indicadores2 pelo hive

Case 6

Cosméticos

AWS Cloud Formation + Spark + Algoritmo

Cosméticos

2 semanas x 4
AWS Cloud Formation
Spark + Algoritmo

Financeiro

Arquitetura

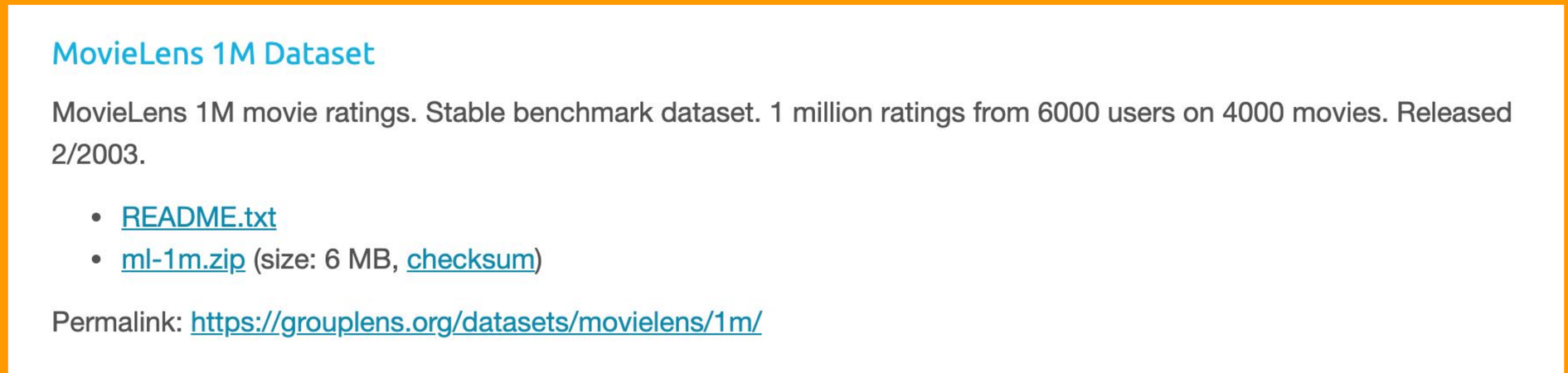
Financeiro

8 meses
3 projetos como arquiteto
Spark

Spark - parte 3

Dataset

Download https://grouplens.org/datasets/movielens/1m/



MovieLens 1M Dataset

MovieLens 1M movie ratings. Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

Permalink: https://grouplens.org/datasets/movielens/1m/

Prática

Carga e Tratamento dos dados

Exemplo com Spark SQL

Exemplo com Spark Core

Gravação de arquivo com resultado

Case 7

Pagamentos

Infraestrutura (Nifi, Stream Processor & certificados)

Spark - parte 4

Comandos

> coalesce

> repartition

Prática

Carga dos dados

Reparticionamento do dados utilizando coalesce

Gravação de arquivo com resultado

Reparticionamento do dados utilizando repartition

Gravação de arquivo com resultado

Outros Cases

Telecom Elasticsearch + Spark

Telecom Spark

Pagamentos Spark + Java

Spark - parte 5

Prática

Construir um código .scala

Recebendo argumentos

Executar o código scala via spark-shell

Spark-submit

```
spark-submit  --master yarn  --deploy-mode cluster
--driver-memory 2g  --executor-cores 2  --executor-memory 4g
--num-executors 20
--conf spark.yarn.driver.memoryOverhead=200m
--conf spark.yarn.executor.memoryOverhead=400m
--files config.properties --name movieRating
--class sparkMovieRating  <args>
```

Certificações

Cloudera

CCA Spark and Hadoop Developer
CCA Data Analyst
CCA Administrator
CCA HDP Administrator Exam

CCP Data Engineer

AWS

Cloud Practitioner

Solutions Architect Associate
SysOps Administrator Associate
Developer Associate

Solutions Architect Professional
DevOps Engineer Professional

Advanced Networking Specialty
Security Specialty
Machine Learning Specialty
Alexa Skill Builder Specialty
Data Analytics Specialty
Database Specialty

GCP

Associate Cloud Engineer

Cloud Architect
Cloud Developer
Data Engineer

Cloud DevOps Engineer
Cloud Security Engineer
Cloud Network Engineer
Collaboration Engineer
Machine Learning Engineer

Azure

MC Azure
MC Azure AI
MC Azure Data

MC Azure Solutions Architect
MC Azure DevOps Engineer

MC Azure Administrator
MC Azure Developer
MC Azure Security Engineer
MC Azure AI Engineer
MC Azure Data Scientist
MC Azure Data Engineer
MC Azure Database Administrator

Elastic

Elastic Certified Engineer

Elastic Certified Analyst

Por onde começar

40 certificaçõe$

Elastic Certified Engineer
Engenheiro de Dados - Data Engineer
Arquiteto - Architect Professional
Administrador - SysOps / DevOps
Cientista - Machine Learning / IA

Atitudes

Fala e Faz

Fala e Não Faz

Não Fala e Não Faz

Não Fala e Faz

"When intentions go public"

Faixas Salariais

estagiário

analista junior

analista pleno

analista senior

coordenador

gerente

Carreira

Projetos

A participação em projetos não garante conhecimento abrangente

Variáveis

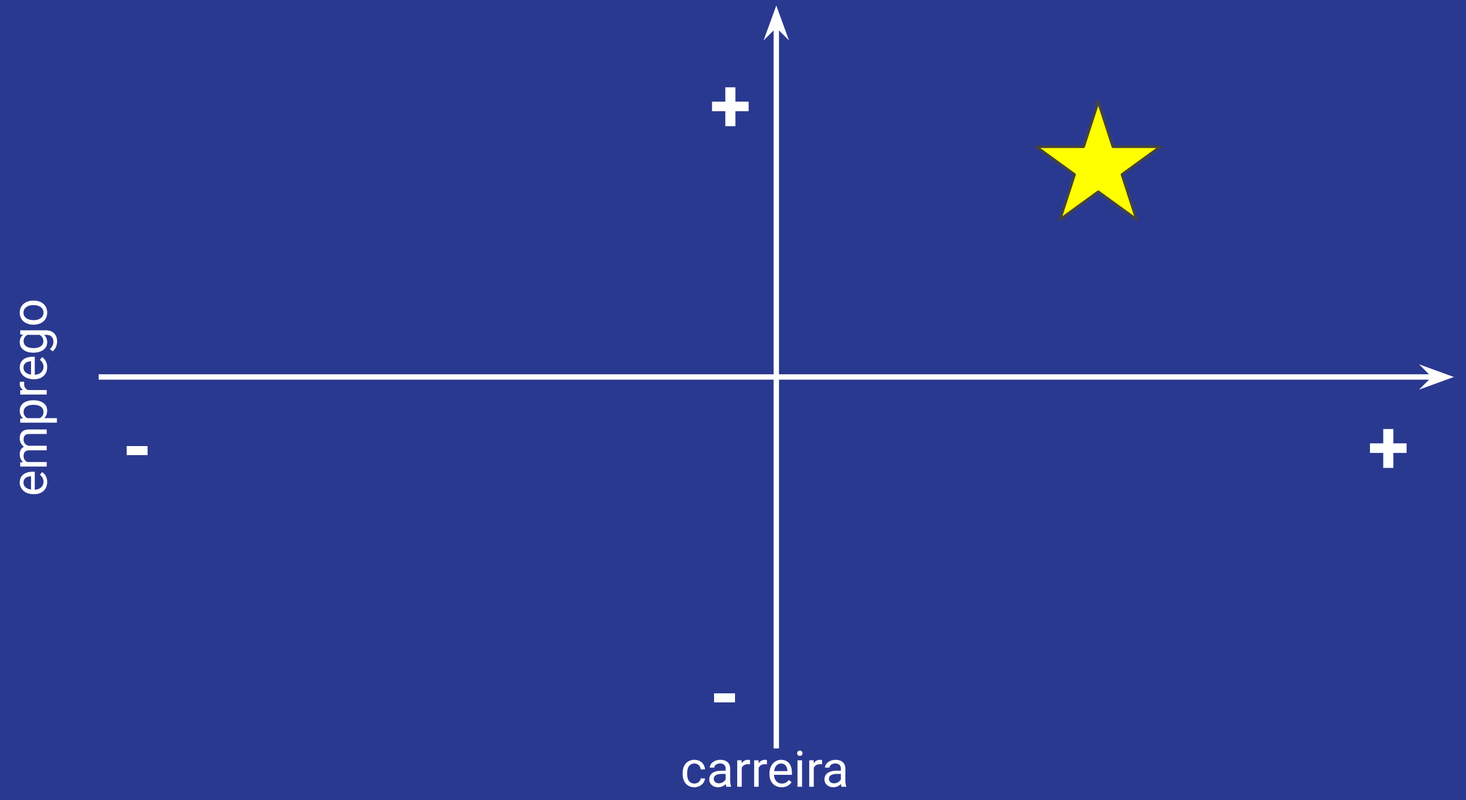
Muitos projetos
Complexidades distintas
Diferentes importâncias
Diversos clientes
Diferentes prioridades
Diversos profissionais

Prepare-se

Não espere um projeto para aprender sobre uma nova ferramenta ou aprofundar-se nela.

Crie projetos pessoais e aplique seu conhecimento.

Carreira x Emprego



Ikigai

alegria, mas sem prosperidade

satisfação, mas sentimento de inutilidade

RAZÃO DE VIVER

entusiasmo, mas sentimento de incerteza

confortável, mas sentimento de vazio

Um pouco todo dia

$$1^{365} = 1$$

$$1{,}01^{365} = 37{,}78$$

Gerencie

Seja protagonista da sua carreira

Mensagem

Construa sua sorte

PUCRS online  UOL edtech_