

# GERÊNCIA DE INFRAESTRUTURA PARA BIG DATA

---

Com Marcos Takeshi e Tiago Coelho Ferreto



É interessante olhar os dados sob as multifacetadas possíveis e estar ciente das **limitações dessas visualizações**.

*Renan Xavier Cortes*



# Conheça o livro da disciplina

## CONHEÇA SEUS PROFESSORES 3

*Conheça os professores da disciplina.*

## EMENTA DA DISCIPLINA 4

*Veja a descrição da ementa da disciplina.*

## BIBLIOGRAFIA BÁSICA 5

*Veja as referências principais de leitura da disciplina.*

## O QUE COMPÕE O MAPA DA AULA? 6

*Confira como funciona o mapa da aula.*

## MAPA DA AULA 7

*Veja as principais ideias e ensinamentos vistos ao longo da aula.*

## ARTIGOS 37

*Links de artigos científicos, informativos e vídeos sugeridos.*

## RESUMO DA DISCIPLINA 38

*Relembre os principais conceitos da disciplina.*

## AVALIAÇÃO 40

*Veja as informações sobre o teste da disciplina.*

# *Conheça seus 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.



# *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.

# Bibliografia básica

As publicações destacadas têm acesso gratuito pela **Biblioteca da PUCRS**.

## Bibliografia básica:

GUPTA, Sumit et al. Real-Time Big Data Analytics. Packt Publishing Ltd, 2016.

**RAJ, Pethuru (Ed.). Handbook of research on cloud infrastructures for big data analytics. IGI Global, 2014.**

RYZA, S.; LASERSON, U.; OWEN, S.; WILLS, J. Advanced Analytics with Spark: Patterns for Learning from Data at Scale. O'Reilly Media, 2017.

**HAN, J. & KAMBER, M. . San Francisco, CA : Morgan Kaufmann, 2001. 550 p. Data mining: concepts and techniques.**

## Bibliografia complementar:

LIU, B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2011.

TAN, P.; STEINBACH, M.; KUMAR, V. Introduction to data mining. Boston, Addison-Wesley, 2014.

WHITE, Tom. Hadoop: The Definitive Guide, 4th Edition. Storage and Analysis at Internet Scale. O'Reilly Media. 2015.

WITTEN, I.; FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. San Francisco (CA): Morgan Kaufmann, 2000. 371 p.

# O que compõe o Mapa da Aula?

## FUNDAMENTOS

Conteúdos essenciais sem os quais você pode ter dificuldade em compreender a matéria. Especialmente importante para alunos de outras áreas, ou que precisam relembrar assuntos e conceitos. Se você estiver por dentro dos conceitos básicos dessa disciplina, pode tranquilamente pular os fundamentos.



## CURIOSIDADES

Curiosidades, fatos e peculiaridades que ampliam o seu conhecimento e conectam a assuntos do cotidiano e vida profissional.



## DESTAQUES

Frases dos professores, que resumem sua visão sobre um assunto ou situação.



## ENTRETENIMENTO

Inserções de conteúdos da equipe de design educacional para tornar a sua experiência mais agradável e significar o conhecimento da aula.



## LEITURAS INDICADAS

A jornada de aprendizagem não termina ao fim de uma disciplina. Ela segue até onde a sua curiosidade alcança. Aqui você encontra uma lista de indicações de leitura. São artigos e textos já publicados sobre temas abordados em aula, que poderão ser indicações dos professores ou da equipe de conteúdo da PUCRS Online.



## PALAVRAS-CHAVE

Significado de termos técnicos ou palavras específicas do campo da disciplina.

## VÍDEOS

Assista novamente aos conteúdos expostos pelos professores em vídeo. Aqui você também poderá encontrar vídeos mencionados em sala de aula.

Lembre-se que a diversificação de estímulos sensoriais na hora do estudo otimiza seu aprendizado.

## CASE

Neste item você relembra o case analisado em aula pelo professor.

## MOMENTO DINÂMICA

Aqui você encontra a descrição detalhada da dinâmica realizada pelo professor em sala de aula com os alunos.

# Mapa da Aula

Os tempos marcam os principais momentos das videoaulas.

## AULA 1 • PARTE 1

00:26

### Experiência do professor

#### PALAVRAS-CHAVE

**Internet Banking:** ambiente bancário na internet, onde é possível realizar diversas operações antes feitas nas agências físicas.

**WAP:** ambiente de aplicações e um conjunto de protocolos de comunicação para aparelhos sem fio.

**Memory Leak:** ou vazamento de memória, quando uma porção de memória, alocada para uma operação, não é liberada quando não mais necessária.



04:11

O professor Marcos Takeshi apresenta sua trajetória em desenvolvimento de sistemas e Big Data. Iniciando sua formação no ramo da eletrotécnica, não demorou muito para que descobrisse sua paixão pelos sistemas, participando de projetos focados no internet banking, wap banking, e-mail banking e publishing.

10:58

### O que é necessário para entrar na área?

#### EXERCÍCIO DE FIXAÇÃO

Segundo o professor Marcos Takeshi, quais os dois fatores necessários para se tornar um Data Scientist?



10:58

Para ser um profissional de Data Science é necessário ter paciência e buscar se profissionalizar para estar preparado para as oportunidades que se apresentarem ao longo do tempo. Além disso é importante expandir suas relações e ter um bom networking.

Oportunidades e networking.

Paciência e conhecimento em Python.

Paciência e Networking.

Nenhuma das anteriores.



17:06

#### PALAVRAS-CHAVE

**Hard Skills:** também chamadas de habilidades técnicas, são quaisquer habilidades relacionadas a uma tarefa ou situação específica.

**Soft Skills:** são um conjunto de traços de personalidade produtivos que caracterizam os relacionamentos de uma pessoa em um ambiente social.

## Dados são o novo petróleo

Assim como o petróleo, os dados são encontrados, extraídos, refinados, distribuídos e, por fim, monetizados. O dado em si não tem significado algum, quando é atribuído um significado a este, ele se torna uma informação. Eles podem ser aplicados em diversas áreas como: ETL, Business Intelligence, Business Analytics, entre outras.

Vista a importância dos dados no mundo atual, populariza-se o processo de Data Mining, que auxilia na tomada de decisões, além de construir e treinar de modelos a partir dos dados. Para isso, é necessário que as bases sejam limpas, padronizadas e organizadas. O Data Mining é aplicado a grandes bancos de dados.

## PALAVRAS-CHAVE

### Luis Alfredo Vidal De Carvalho:

Autor do livro “Datamining - A Mineracao De Dados No Marketing, Medicina, Economia, Engenharia e Administração”.

## EXERCÍCIO DE FIXAÇÃO

Em que é utilizado o Data Mining?

No armazenamento os dados em silos.

No gerenciamento de Data Warehouse.

Na atribuição de significado aos dados.

No auxílio de decisões, construção e treinamento de modelos.

17:50

17:50



## EXERCÍCIO DE FIXAÇÃO

Em que área os dados são aplicáveis?

ETL.

Business Intelligence.

Business Analytics.

Todas as anteriores.

18:16



18:24



## FUNDAMENTO I

### Data Mining

Data Mining é o processo de exploração dos dados na busca por padrões consistentes, a fim de identificar padrões e relacionamentos sistemáticos entre variáveis. Utiliza diversas técnicas da estatística, recuperação de informação, inteligência artificial e reconhecimento de padrões.

São alguns tipos de informações que podem ser obtidas através da mineração de dados:

- Associações: ocorrências ligadas a um único evento;
- Sequências: eventos ligados ao longo do tempo;
- Classificação: reconhece modelos que descrevem o grupo ao qual o item pertence por meio do exame dos itens já classificados e pela inferência de um conjunto de regras;
- Aglomeração: funciona de maneira semelhante a classificação quando ainda não foram definidos grupos;
- Prognósticos: embora todas essas aplicações envolvam previsões, os prognósticos as utilizam de modo diferente.

**“** Empresas têm grande interesse em processar os dados e deles extrair informação [...] com a finalidade de monetizar. **”**

20:54



21:03

## PALAVRAS-CHAVE

**BSC:** Indicadores Balanceados de Desempenho, é uma metodologia de medição e gestão de desempenho.

**KPI:** são ferramentas de gestão que buscam medir o consequente nível de desempenho e sucesso de uma organização ou de um determinado processo.



22:40

25:10

## PALAVRAS-CHAVE

### ETL – Extração Transformação e Carga:

são softwares que têm como função a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e, por fim, o carregamento dos dados geralmente para um Data Mart e/ou Data Warehouse.

## Processamento de Big Data

Ao longo das últimas décadas a quantidade de dados produzidos têm crescido de forma exponencial. O termo Big Data foi então cunhado para nomear essa quantidade crescente e não estruturada de dados que são gerados a cada segundo.

O processamento desses dados ocorre da seguinte forma: os dados armazenados em uma rede de servidores, provido de discos e memória, são divididos em partes menores as quais são armazenadas em servidores diferentes - a chamada arquitetura distribuída.

## AULA 1 • PARTE 2

### Ecossistema Hadoop

O Hadoop é um framework que auxilia no processamento distribuído de datasets em clusters, e que utiliza modelos simples de programação. Ele é composto por diversas ferramentas, mas, dentre estas, podemos destacar: Oozie, Pig, Mahout, Ambari, Flume, Sqoop, Zookeeper, HBase, Solr, Yarn, Hive e HDFS.

00:00



## EXERCÍCIO DE FIXAÇÃO

Assinale a alternativa que corresponde a um framework que auxilia no processamento distribuído de datasets em clusters:

Hadoop.

Pig.

Hive.

Sqoop.

## Sistema Operacional e Docker

O principal sistema operacional para utilização das ferramentas do ecossistema Hadoop, podendo ainda ser utilizado os sistemas CentOS ou RedHat.

O professor Marcos Takeshi demonstra a instalação do Docker, um gerenciador de container utilizado para emular outros sistemas operacionais.

07:13



## SITE INDICADO

Docker é um conjunto de produtos de plataforma como serviço (PaaS) que usam virtualização de nível de sistema operacional para entregar software em pacotes chamados contêineres. Acesse o site clicando [aqui](#).

## Profissionais

São os principais profissionais presentes em Big Data:

- DevOps: infraestrutura e cloud;
- Data Engineer: programador;
- Data Scientist: especialistas analíticos.

19:32

## AULA 1 • PARTE 3

### PALAVRAS-CHAVE

#### Hadoop Distributed File System (HDFS)

**(HDFS):** é um sistema escalável baseado em Java que armazena dados em diversas máquinas, sem organização prévia.

00:44 07:57



### Exercício prático

O professor Marcos Takeshi realiza um exercício prático a fim de demonstrar a utilização dos comandos de Linux no ambiente Docker.

### PALAVRAS-CHAVE

**Hive:** é um software de Data Warehouse desenvolvido em cima do Apache Hadoop para consulta e análise de dados.

**Sparky:** é um framework de código aberto para computação distribuída.

08:14 14:30



### Case: financeiro

O professor Marcos Takeshi conta a sua experiência em um case com um cliente da área financeira, onde iniciou seu aprendizado no mundo corporativo, contando com a experiência de colegas com maior conhecimento.

14:59



15:46



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



## HDSF

O HDFS é a ferramenta principal de armazenamento no Hadoop. O professor Marcos Takeshi explica as funções dos principais comandos e demonstra na prática a utilização desses.

15:46



26:03



## EXERCÍCIO DE FIXAÇÃO

O HDFS é uma ferramenta de:

Análise de dados.

Armazenamento.

Comunicação de servidores.

Gerenciador de container.

**“** É importante saber listar só uma parte do arquivo. **”**

## Case: Marketing

Após seu sucesso no projeto anterior, o professor Marcos Takeshi é indicado para um novo cliente, agora em um projeto de marketing, no desenvolvimento de um aplicativo para coleta de dados de compras. Sua função foi a melhoria de uma API, processamento e estudo dos dados capturados pelo aplicativo.

28:49



## PALAVRAS-CHAVE

**API:** Interface de Programação de Aplicações, é um conjunto de rotinas e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem envolver-se em detalhes da implementação do software.

**KNN (K-Nearest Neighbors Algorithm):** é um método de classificação não paramétrico, usado para classificação e regressão.

## HIVE I

O professor Marcos Takeshi explica as funções dos principais comandos do HIVE e demonstra na prática a utilização desses. Pela característica do HIVE de executar um bypass na autenticação do cluster, possibilitando que qualquer usuário possa acessar os arquivos que estão no HDFS. A fim de evitar acessos indevidos, é possível aplicar um alias apontando para o bline, que vai permitir ou barrar o acesso.

41:49

## AULA 1 • PARTE 4

09:11



• 09:35

### FUNDAMENTO II

#### Tabelas gerenciadas e não gerenciadas

Cada tabela, no SQL do Spark, tem informações de metadados que armazenam o esquema e os dados em si.

Uma tabela gerenciada é uma tabela SQL do Spark para a qual esse gerencia os dados e os metadados.

Já em uma tabela não gerenciada, é possível permitir que o Spark SQL gerencie os metadados, enquanto o usuário controla o local dos dados.

#### Case: Telecom

O professor Marcos Takeshi comenta a sua experiência em uma empresa Telecom, onde participou de um processo de turning, identificando o que já havia sido desenvolvido, prestando consultoria, identificando falhas e auxiliando a equipe na aplicação de soluções.

#### Hive II

O professor Marcos Takeshi retoma os principais comandos do HIVE e demonstra na prática a utilização desses de forma mais detalhada.

### PALAVRAS-CHAVE

**Map Reduce:** modelo de programação desenhado para processar grandes volumes de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes.

**Impala:** mecanismo SQL de processamento paralelo maciço (MPP) de código aberto.

• 16:18

23:41

#### Case: ambiente corporativo

Em um novo ambiente, o professor Marcos Takeshi inicia um projeto com base em performance, buscando o processamento de uma base de clientes de todos os sistemas de um banco, a fim de gerar relatórios cruzados com tempo de processamento menor que o anterior - de 3 a 4 horas cada. Como resultado do trabalho com a equipe, o tempo de processamento foi reduzido para algo entre 10 a 15 minutos.



27:12

33:42

#### Spark I



42:19

O Spark é um framework para processamento paralelo que complementa o Hadoop. Ele tem como finalidade uma melhor performance de processamento distribuído, oferecendo flexibilidade de programação. O fluxo do Spark começa com a leitura dos dados e termina com a gravação dos resultados.

### PALAVRAS-CHAVE

**Compressão:** ato de reduzir o espaço ocupado por dados em um determinado dispositivo.

## PALAVRAS-CHAVE



42:49

**IntelliJ:** ambiente de desenvolvimento integrado escrito em Java para a criação de software de computador.

**Read-eval-print loop (REPL):** é um ambiente de programação interativo e simples que recebe entradas de um único usuário, executando-as e retornando o resultado ao mesmo.

## AULA 2 • PARTE 1

06:43

### Exercício I

## PALAVRAS-CHAVE



11:12

**Dataframe:** pode ser entendido como uma tabela de uma base de dados, onde cada linha corresponde a um registro.

**Elasticsearch:** mecanismo de busca e análise de dados distribuído, gratuito e aberto para todos os tipos de dados, estruturados e não estruturados.

14:54

### Spark II

O Spark possibilita uma interação facilitada. Ou seja, com a inserção de comandos, você pode obter o resultado de imediato. O professor Marcos Takeshi demonstra como realizar a declaração de variáveis e funções.

## PALAVRAS-CHAVE



15:23



15:53

## EXERCÍCIO DE FIXAÇÃO

Dentro do Spark, ao declarar uma variável que não pode ser mutável, qual o comando utilizado?

String.

Val.

Var.

Scala.

## Case: Cosméticos e Financeiro

Outro projeto em que o professor Marcos Takeshi se envolveu durante sua trajetória foi em uma empresa estrangeira de cosméticos. O projeto já estava em andamento e o papel do professor foi melhorar o desempenho do produto.

Depois disso, novamente na área financeira, Marcos Takeshi participa de um novo projeto como arquiteto de software.

27:05

27:16



## PALAVRAS-CHAVE

**AWS Cloudformation:** aplicação para modelar uma coleção de recursos relacionados da AWS e de terceiros, provisioná-la com rapidez e consistência e gerenciar todo o seu ciclo de vida mediante o tratamento da infraestrutura como código.

## AULA 2 • PARTE 2

## PALAVRAS-CHAVE

00:07



## PALAVRAS-CHAVE

**Dataset:** é uma coleção de dados normalmente tabulados.

## Exercício II

O professor Marcos Takeshi realiza um exercício prático de carga e tratamento de dados de um dataset, utilizando o Spark SQL.

01:18

## Exercício II: continuação

O professor Marcos Takeshi realiza um exercício prático de carga e tratamento de dados de um dataset, utilizando o Spark Core.

## Case: pagamentos

Em uma nova empresa, no setor de pagamentos, o professor Marcos Takeshi participa de um novo projeto como Devops, realizando a manutenção e suporte de um cluster de ferramentas.

14:26

27:33

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

## PALAVRAS-CHAVE

**Apache NiFi:** projeto de software da Apache Software Foundation projetado para automatizar o fluxo de dados entre sistemas de software.

**Stream processing:** é um modelo de processamento que possibilita que aplicações possam explorar limitadas formas de processamento paralelo.

28:23



30:49



## EXERCÍCIO DE FIXAÇÃO

Qual a diferença entre os comandos coalesce e repartition?

O coalesce trafega mais informação na rede.

O repartition trafega menos informação na rede.

O repartition não se importa com o tráfego de informação na rede.

Nenhuma das anteriores.

32:40



36:50

## Exercício III

O professor Marcos Takeshi realiza um exercício prático demonstrando a utilização dos comandos coalesce e repartition, importantes no processamento distribuído. Neste exercício o professor executa a carga de dados, reparticionamento e gravação de arquivos com resultado.

## AULA 2 • PARTE 3

### Exercício IV

O professor Marcos Takeshi realiza um exercício prático construindo um código scala, recebendo argumentos e executando por meio de spark-shell, semelhante a execução em um cluster real. Após a execução, o professor apresenta a função spark-submit, que tem como função submeter para o cluster.

01:40



01:45

“É interessante ter o conhecimento de outras ferramentas.”

12:36



## EXERCÍCIO DE FIXAÇÃO

Define a memória que será utilizada por cada executor:

**Executor.memory.**

**Driver.memory.**

**Config.memory.**

**Class.memory.**

### Certificações

Uma forma de atestar seu conhecimento na área é por meio de certificações, muitas vezes oferecidas pelas empresas desenvolvedoras das aplicações utilizadas pelo cientista de dados. O professor Marcos Takeshi demonstra algumas das certificações que considera mais importantes disponíveis no mercado.

16:01



## Por onde começar?

As certificações são inúmeras, e podem ser um grande investimento. A fim de orientar o caminho do profissional que está ingressando nessa área, o professor Marcos Takeshi sugere algumas certificações como ponto de partida.

• 25:43



É interessante você ter uma certificação para mostrar que você tem conhecimento no assunto.



29:12



## LEITURAS INDICADAS

### **When Intentions Go Public: Does Social Reality Widen the Intention-Behavior Gap?**

O artigo sugerido pelo professor pode ser acessado clicando [aqui](#).

## Salário e carreira

O professor Marcos Takeshi comenta brevemente sobre os níveis de profissionais da área e suas respectivas faixas salariais.

Esteja sempre atento a sua carreira. Em muitos casos, sua participação em um projeto pode ser limitada, não garantindo um conhecimento abrangente, pois esses podem ser de complexidades distintas, diferentes importâncias, diversos clientes, diferentes prioridades e diversos profissionais.

Portanto, não espere um novo projeto para aprender uma nova ferramenta ou se aprofundar nela. Crie projetos pessoais e aplique seu conhecimento.

• 30:33

35:49



## CURIOSIDADE

### **Ikigai**

A palavra japonesa Ikigai significa “razão de viver”, “objeto de prazer para viver” ou “força motriz para viver”. De acordo com a cultura japonesa, todos tem um Ikigai, e descobrir qual é o seu requer uma profunda busca de si mesmo. A partir dessa busca, seria possível obter satisfação e significado na vida.

38:21



38:25



## PALAVRAS-CHAVE

**Utopia:** definido como lugar ou estado ideal, de completa felicidade e harmonia entre os indivíduos.

“  
O importante é você estar sempre atento e fazer um pouco todo dia.”

## AULA 3 • PARTE 1

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

02:14



07:22



### CURIOSIDADE

#### SETI@home

É um experimento científico, baseado na UC Berkeley, que usa computadores conectados à Internet no Search for Extraterrestrial Intelligence (SETI).

Saiba mais clicando [aqui](#).

### FUNDAMENTO I

#### Banco de dados relacional e não-relacional

Um banco de dados relacional é uma coleção de dados correlacionados entre si, organizados como conjuntos de tabelas. Neste modelo, cada linha da tabela é um registro com uma identificação única, a chave primária.

Já um banco de dados não-relacional é qualquer banco de dados que não segue o modelo relacional fornecido pelos sistemas tradicionais de gerenciamento de bancos de dados relacionais (SGBDR). Pode ser de quatro tipos: Key-value stores, Graph stores, Column stores, e Document stores.

11:25



08:46

#### A era dos dados

Por meio do vídeo “Exploração no limite da grande quantidade de dados”, o professor Tiago Coelho Ferreto busca demonstrar como uma grande quantidade de dados é processada e entendida pelos físicos do CERN, e os desafios enfrentados por eles.

Na atualidade, os dados são gerados de forma exponencial. O professor traz algumas estatísticas e informações que ilustram esse cenário.

11:40



### PALAVRAS-CHAVE

**Tim Smith:** Chefe do Grupo de Colaboração, Dispositivos e Aplicativos do CERN, na Suíça.

**Cada vez mais a gente gera mais dados e temos mais informação para processar.**

19:51



### PALAVRAS-CHAVE

**New York Stock Exchange:** na bolsa de valores nova-iorquina são transacionadas ações das maiores empresas estadunidenses.

## PALAVRAS-CHAVE



20:51

**Internet Archive:** é uma organização sem fins lucrativos dedicada a manter um arquivo multimídia de informações.

**Radio-Frequency IDentification (RFID):** é um método de identificação automática através de sinais de rádio, recuperando e armazenando dados remotamente através de dispositivos denominados etiquetas RFID.

22:29



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

## O que é big data?

Big data são dados em grande volume, em relação ao sistema de processamento, variados entre dados estruturados e não estruturados, de diferentes padrões a serem analisados.

São alguns fatos sobre o Big Data:

- As abordagens tradicionais de armazenamento e processamento não funcionam;
- Os dados podem conter informações valiosas, sendo necessário que sejam processados em um curto espaço de tempo;
- Essas informações podem ter diversos usos;
- A abordagem tradicional não conclui a análise dentro do prazo determinado.

22:57

29:10

## Dimensões de Big Data

A definição das dimensões do Big Data evoluiu com o tempo, sendo então caracterizadas pelos 6V's:

- Volume: em big data, é uma quantidade que não pode ser coletada, armazenada e processada por meio de abordagens tradicionais;
- Velocidade: taxa em que os dados são gerados;
- Variedade: classificados entre estruturados e não estruturados;
- Veracidade: incerteza sobre os dados, devido à baixa qualidade ou ruído nos dados;
- Variabilidade: falta de consistência ou padrões fixos nos dados;
- Valor: se vale ou não a pena o investimento em infraestrutura para o armazenamento destes.

## EXERCÍCIO DE FIXAÇÃO

Considerando os 6 V's do Big Data, correlacione:

- |                  |   |
|------------------|---|
| 1) Volume        | ( ) Em big data, é uma quantidade que não pode ser coletada, armazenada e processada por meio de abordagens tradicionais. |
| 2) Velocidade    | ( ) Falta de consistência ou padrões fixos nos dados.   |
| 3) Variedade     | ( ) Classificados entre estruturados e não estruturados.  |
| 4) Veracidade    | ( ) Incerteza sobre os dados, devido à baixa qualidade ou ruído nos dados.  |
| 5) Variabilidade | ( ) Se vale ou não a pena o investimento em infraestrutura para o armazenamento destes.                                   |
| 6) Valor         | ( ) Taxa em que os dados são gerados.   |

Assinale a alternativa que corresponde a sequência correta:

5; 6; 2; 1; 3; 4.

4; 6; 1; 5; 3; 2.

6; 1; 5; 3; 2; 4.

1; 5; 3; 4; 6; 2.



29:10

40:22

### Abordagens tradicionais

Nas abordagens tradicionais, os trabalhos em lote são programados para migrar dados para Data Warehouses em um período de dia, semana ou mês. Os dados nestas abordagens são estruturados e passam por um ciclo de análise, a fim de criar conjuntos de dados e extrair informações significativas.

A ingestão no Data Warehouse são realizadas por meio de operações ETL, pegando dados brutos e processando-os para relatórios e análise.



41:17

### PALAVRAS-CHAVE

**Business Intelligence**: é um conjunto de teorias, metodologias, processos, técnicas e estruturas que busca transformar grandes quantidades de dados em informações importantes para uma boa gestão da empresa ou negócio.

**Data Warehouse**: também conhecido como armazém de dados ou ainda depósito de dados, são estruturas utilizadas para armazenar informações relativas às atividades de uma organização em bancos de dados, de forma consolidada.



41:52

## EXERCÍCIO DE FIXAÇÃO

É um problema da abordagem tradicional, no processamento de Big Data, EXCETO:

Latência.

Fontes limitadas.

Valor baixo dos dados.

Escala limitada.

## EXERCÍCIO DE FIXAÇÃO

Qual a solução para os desafios da abordagem tradicional?

Data Warehouse.

Cluster Computing.

Business Intelligence.

Internet Archive.

44:46



43:16



## PALAVRAS-CHAVE

**Scale-out:** escalonamento horizontal, significa adicionar mais nós a (ou remover nós de) um sistema, como adicionar um novo computador para uma aplicação de software distribuída.

## AULA 3 • PARTE 2

00:00



## PALAVRAS-CHAVE

**Hardware de commodity:** dispositivo ou componente de dispositivo que é relativamente barato, amplamente disponível e mais ou menos intercambiável com outro hardware de seu tipo.

## Hadoop

07:26

O Hadoop é a principal ferramenta criada para trabalhar com grandes volumes de dados. Sua história inicia em 2003, com o lançamento do artigo “The Google File System”, da empresa Google, explicando o funcionamento do seu sistema de arquivos. Em 2004, com um novo artigo, “MapReduce: Simplified Data Processing on Large Clusters”, apresenta uma solução de processamento em grande escala.

## Nuvem x Infraestrutura local

### Nuvem

Evita os custos de inicialização, podendo aumentar sob demanda e pagando por uso;

Não há garantias sobre onde estão os dados, como estão sendo gerenciados ou quem pode acessá-los.

Requer pessoal, porém com foco no problema;

Recursos adicionados sob demanda.

### On-promises

Grande custo inicial. Necessária a configuração de servidores de alta tecnologia, rede e armazenamento;

Sensação de maior segurança. Controle de acessos dos dados, quando são usados e com que finalidade;

Requer pessoal para gerenciar a implementação, configuração e suporte;

Requer avaliação adicional da capacidade necessária para adquirir hardware. Dependendo da análise de demanda de capacidade, pode haver uma grande quantidade de recursos ociosos.

10:21



## Componentes e cenário do Hadoop

O Hadoop é baseado em alguns componentes: Hadoop Common, HDFS, YARN e MapReduce. Além disso, ele possui um ecossistema que inclui outros projetos que interagem ou se integram com o sistema: Ingestão de dados, Análise de dados, Bancos de dados NoSQL, Segurança, Gestão, Pesquisa, Aprendizado de máquina, Coordenação de fluxo de trabalho, entre outros.

Criado em 2008, pela Cloudera, o Hadoop segue recebendo novas soluções por outras empresas como a MapR e HortonWorks. Empresas essas que têm como modelo de negócio o Hadoop.

## PALAVRAS-CHAVE

**Doug Cutting:** é um designer de software, defensor e criador da tecnologia de pesquisa de código aberto. Fundador da Lucene e cofundador da Nutch e Apache Hadoop.

**Web Crawler:** também conhecido como rastreador web, software que navega pela rede mundial de uma forma metódica e automatizada.

18:48



## EXERCÍCIO DE FIXAÇÃO

É o sistema de arquivos distribuído para armazenamento de dados do Hadoop:

Hadoop Common.

Hadoop YARN.

Hadoop MapReduce.

HDFS.

21:10



## PALAVRAS-CHAVE

**Cloudera:** empresa com sede nos Estados Unidos que fornece uma nuvem de dados corporativos.

**Hadoop Distributed File System (HDFS):** sistema escalável baseado em Java que armazena dados em diversas máquinas, sem organização prévia.

## Casos de uso

Um dos usos mais comuns do Hadoop nas empresas é no descarregamento de data warehouses, no armazenamento de longo prazo e rotinas ETL.

Outras formas de uso são: no armazenamento de eventos e processamento de eventos complexos e análise avançada.

25:50



26:32

## PALAVRAS-CHAVE

**Data Lake:** sistema ou repositório de dados armazenados em seu formato bruto, geralmente objetos blobs (Binary Large OBject) ou arquivos.

## AULA 3 • PARTE 3

### Arquitetura do cluster Hadoop

Em um cluster Hadoop, utiliza-se os clusters HDFS e YARN configurados nos mesmos nós, seus objetivos principais são:

- YARN: Processamento;
- HDFS: Armazenamento.

01:31



### EXERCÍCIO DE FIXAÇÃO

Qual a função do NameNode?

Gerenciar o armazenamento.

Gerenciar os metadados.

Replicar os blocos.

Nenhuma das anteriores.

### Modos de implantação

O Hadoop oferece suporte a três modos de implantação:

- LocalJobRunner: executa uma aplicação localmente em uma única JVM, permite integração da aplicação com um IDE para realizar testes de unidade, depuração e rastreamento e usa o sistema de arquivos local ao invés de HDFS;
- Pseudo-Distribuído: todos os daemons do Hadoop são executados em JVMs separados em um único host, simula como um cluster completo funcionaria, útil para simular interações entre componentes em um cluster real com uma única máquina;
- Cluster totalmente distribuído: executa daemons mestre e escravo em máquinas distintas, modo de implantação típico para sistemas de produção.

10:05



### EXERCÍCIO DE FIXAÇÃO

Referente a um cluster totalmente distribuído:

É modo de implantação típico para sistemas de produção.

Todos os daemons do Hadoop são executados em JVMs.

Executa uma aplicação localmente em uma única JVM.

Usa o sistema de arquivos local ao invés de HDFS.

### Exemplos de configurações

O professor Tiago Coelho Ferreto apresenta exemplos de configurações dos clusters de grandes empresas como: Alibaba, Ebay, Facebook, Last FM, Linkedin, Spotify e Yahoo.

16:07



### PALAVRAS-CHAVE

**Spanning Tree Protocol (STP):** é um protocolo de rede que cria uma topologia lógica sem loop para redes Ethernet.

20:07



## PALAVRAS-CHAVE

### Outras informações importantes

O professor Tiago Coelho Ferreto explica os requisitos de software necessários para utilização do Hadoop e suas aplicações e como instalar o Hadoop On-premises. Traz também o funcionamento do Hadoop em nuvem, serviço fornecido por grandes empresas.

20:38

**Domain Name System (DNS):** é um sistema para computadores de nomenclatura hierárquica e descentralizada, serviços ou outros recursos conectados à Internet ou a uma rede privada.

## AULA 3 • PARTE 4

00:00



00:33

### Criando um container e instalando as dependências

O professor Tiago Coelho Ferreto demonstra como criar um contêiner e instalar as dependências do Hadoop, como o Java, e, por fim, instalando o próprio Hadoop.

## PALAVRAS-CHAVE

**Ubuntu:** é um sistema operacional de código aberto, construído a partir do núcleo Linux, baseado no Debian e utiliza GNOME como ambiente de desktop de sua mais recente versão com suporte de longo prazo (LTS).

05:16



15:38

## SITE INDICADO

A documentação do Hadoop pode ser acessada no site do software. Acesse o site clicando [aqui](#).

### Configurando e importando I

O professor Tiago Coelho Ferreto demonstra como configurar as variáveis de ambiente, parâmetros especificados que podem ser passados para a aplicação que está sendo executada, semelhantes a parâmetros de sistema.

26:54

### Configurando e importando II

O professor Tiago Coelho Ferreto demonstra como configurar o HDFS, especificando os locais dos arquivos, demais propriedades, parâmetros e variáveis, além dos tamanhos de blocos. O próximo passo é a configuração do YARN.

### Configurando e importando III

Após as configurações necessárias, o próximo passo é a inicialização dos daemons. Por fim, o professor demonstra como verificar os processos que estão sendo executados.

37:51

## AULA 4 • PARTE 1

00:16

YARN I

### PALAVRAS-CHAVE

**Single Point of Failure (SPOF):** ponto único de falha ou ponto crítico de falha designa um local num sistema informático que, caso falhe, provoca a falha de todo o sistema.



03:30

### EXERCÍCIO DE FIXAÇÃO

O que acontece caso o ResourceManager do YARN falhar?

Tarefas em execução no NM serão consideradas como falha.

Se houver um RN standby, este irá assumir a posição do RN e continuar a execução das aplicações. Caso não houver, nenhuma nova aplicação pode ser iniciada.

Após falhar quatro vezes a aplicação é considerada como falha.

Todas as anteriores.



10:11

14:21

YARN II

- Ao executar o YARN no modo de cluster, cancelar a aplicação ou sair de invocação não mata o aplicativo. Para isso utiliza-se o comando `yarn`.
- Outro ponto importante é a agregação de logs, que são movidos para HDFS para armazenamento de longo prazo, e podem ser acessados usando o comando `yarn`.
- O escalonamento de aplicações, no YARN, permite que várias aplicações sejam executadas simultaneamente, compartilhado recursos de memória e computação distribuída do cluster. São suas políticas de escalonamento: `FIFO Scheduler`, `Fair Scheduler` e `Capacity Scheduler`.

### PALAVRAS-CHAVE

**MRJobHistory Server:** permite que o usuário obtenha o status dos aplicativos concluídos.



17:32

26:47

Demonstração YARN I

O professor Tiago Coelho Ferreto demonstra de forma prática a utilização do cluster YARN por meio do exemplo de uma aplicação de geração e contagem de palavras.

## AULA 4 • PARTE 2

### Demonstração YARN II

O professor Tiago Coelho Ferreto segue com a demonstração da utilização do YARN, aplicando os comandos vistos em aula.

### EXERCÍCIO DE FIXAÇÃO

Como é o sistema de arquivos do HDFS?

Físico.

Virtual.

Misto.

Nenhuma das anteriores.

00:00

09:20

### HDFS I

O HDFS é a principal fonte de dados de entrada e saída do Hadoop para operações de processamento de dados. Suas características são: escalável, tolerante a falhas, usa hardware comum, suporta alta concorrência e fornece a alta demanda por largura de banda em relação ao acesso aleatório de baixa latência.



12:29

12:47



### PALAVRAS-CHAVE

**Overlay filesystem:** permite que uma árvore de diretório seja sobreposta em outra árvore de diretório somente leitura.



14:28

17:39

### EXERCÍCIO DE FIXAÇÃO

Qual a vantagem dos blocos distribuídos do HDFS?

Os dados não podem ser atualizados após escritos.

Aumento das oportunidades de localização dos dados.

Não há nenhum tipo de compartilhamento e processamento paralelo de dados.

Fornece tolerância a falhas.

### HDFS II

Para recuperar falhas de um DataNode ou bloco, cada objeto do HDFS tem um fator de replicação. O NameNode obtém inventários de bloco regulares de cada DataNode no cluster. Ele também recebe heartbeats regulares para verificar a saúde do DataNode.

Quando o NameNode detecta que um bloco não tem o número certo de réplicas, ele instrui um DataNode a replicar esse bloco para outro nó.

**“**A redundância garante a persistência da informação.**”**

### HDFS III

Para controle de acesso o HDFS tem ACLs associados para definição do proprietário do objeto e as permissões, e utiliza uma máscara de permissões de acesso (Unix). Porém a segurança do HDFS é considerada fraca. Para solucionar este problema, é aconselhável a utilização de métodos de segurança adicionais em clusters de produção.

26:04

18:00

27:39

35:03

27:39

### PALAVRAS-CHAVE

**Cache:** dispositivo de acesso rápido, interno a um sistema, que serve de intermediário entre um operador de um processo e o dispositivo de armazenamento ao qual esse operador acede.

**Snapshots:** é o estado de um sistema em um determinado ponto no tempo.

39:04

43:09

### HDFS VI

O professor Tiago Coelho Ferreto traz algumas questões mais avançadas do HDFS, para garantir a persistência e tolerância a falhas:

- Rack Awareness;
- Alta disponibilidade;
- Federação;
- Cache;
- Snapshots.

## AULA 4 • PARTE 3

### HDFS na prática I

O professor Tiago Coelho Ferreto realiza uma demonstração da utilização do HDFS, aplicando os conhecimentos vistos em aula de armazenamento e recuperação de dados.

10:16

00:00

### HDFS na prática II

O professor Tiago Coelho Ferreto segue com a demonstração da utilização do HDFS, por meio do exercício prático.

## PALAVRAS-CHAVE

**Recursivo:** processo que pode ser repetido de modo infinito.



16:42

19:32

## HDFS na prática III

O professor Tiago Coelho Ferreto segue com a demonstração da utilização do HDFS, utilizando um MapReduce.

## HDFS na prática VI

30:06

O professor Tiago Coelho Ferreto finaliza o exercício prático do HDFS demonstrando o descomissionamento dos nodos, ou seja, o desligamento dos nodos para manutenção.

## AULA 4 • PARTE 4

### EXERCÍCIO DE FIXAÇÃO

É uma característica do Flume:

Pode ser escalado horizontalmente.

Suporta um grande conjunto de fontes e tipos de destinos.

Fornece um processo eficiente para ingerir dados de log de vários servidores em um armazenamento centralizado.

Todas as alternativas.



02:12

## Flume

O Flume é um projeto do ecossistema Hadoop, desenvolvido originalmente pela Cloudera, e tem como objetivo capturar, transformar e ingerir dados em HDFS usando um ou mais agentes. Usado geralmente na captura de arquivos de log ou weblogs de um servidor web e encaminhamento para HDFS à medida que são gerados. Além disso, busca superar as desvantagens do comando put do HDFS e transferir dados de streaming dos geradores de dados para sistemas de armazenamento centralizado com menos atraso.

## PALAVRAS-CHAVE

**HBase:** é um banco de dados Hadoop, que tem como objetivo hospedar tabelas muito grandes.



06:00



09:52

## PALAVRAS-CHAVE

**Timestamp:** é uma cadeia de caracteres denotando a hora ou data que certo evento ocorreu.

14:45

**Sqoop e API's****Praticando Flume**

O professor Tiago Coelho Ferreto demonstra uma prática da utilização do Flume, realizando a instalação, criação de um agente e como disparar e capturar dados que serão enviados ao HDFS.

• 24:02

Desenvolvido originalmente pela Cloudera, o Sqoop é um projeto Apache que também possui integração com HIVE, com objetivo de pegar dados de um banco de dados relacional e ingerir esses dados em arquivos no HDFS. Ele também pode ser utilizado para enviar dados do Hadoop para um banco de dados relacional.

API's também podem ser utilizadas para operações de host e guest, como o WEBHDFS e o HTTFS.

**Praticando Sqoop I**

O professor Tiago Coelho Ferreto demonstra uma prática da utilização do Sqoop, realizando o envio de dados de um banco de dados relacional para o HDFS.

• 33:02

35:24

**PALAVRAS-CHAVE**

**MySQL:** é um serviço de banco de dados totalmente gerenciado para implantar aplicativos nativos da nuvem.

**Praticando Sqoop II**

O professor Tiago Coelho Ferreto demonstra uma prática da utilização do Sqoop, realizando listagem banco de dados existentes no MySQL e, por fim, a importação.

• 42:12

**AULA 5 • PARTE 1**

00:00

**MapReduce**

O MapReduce, um dos principais componentes do Hadoop, é baseado no artigo da Google “MapReduce: Simplified Data Processing on Large Clusters”, de 2004. A motivação para a criação do MapReduce parte das limitações na abordagem de escalonamento para aumentar a capacidade de processamento.

Visando tratar esses problemas, surgem as metas de projeto para o MapReduce:

- Paralelização e distribuição automáticas;
- Tolerância a falhas;
- Escalonamento de entrada/saída;
- Status e monitoramento.

## EXERCÍCIO DE FIXAÇÃO

Quais são as duas fases de processamento implementadas pelo desenvolvedor no MapReduce?

Shuffle and Sort.

Fase de mapeamento e fase de redução.

Fase de mapeamento e Shuffle.

Nenhuma das alternativas.



09:23



17:55

## EXERCÍCIO DE FIXAÇÃO

Em MapReduce, qual etapa a saída de cada tarefa Map é enviada para uma tarefa Reduce destino?

Fase de mapeamento.

Fase de redução.

Shuffle and Sort.

Nenhuma das alternativas.

## Tolerância a falhas e funções de otimização

Quando uma tarefa Map falha, ela é automaticamente reprogramada pelo processo mestre para outro nó, de preferência um nó que possui uma cópia dos mesmos blocos, mantendo a localidade dos dados. Além disso, uma tarefa pode falhar e ser reprogramada quatro vezes antes de ser considerado como falha.

Se uma tarefa Reduce falhar, ela também pode ser reprogramada e seus dados de entrada reabastecidos.

Ainda existem outras funções de otimização que podem ser utilizadas, como a Função Combiner, frequentemente igual a função reduce, porém executada no mesmo nó da tarefa Map.

21:23



23:46

## PALAVRAS-CHAVE

**Comutatividade:** é uma propriedade de operações binárias, ou de ordem mais alta, em que a ordem dos operandos não altera o resultado.

27:51

## Implementação Mapreduce

Desde o início do Hadoop, o MapReduce está presente, já o YARN surge posteriormente. Criando assim, duas possíveis formas de implementação: MapReduce cluster framework (MR1 ou MapReduce versão 1) – Hadoop 1; e YARN (MR2) – Hadoop 2.

O Hadoop 2 foi criado devido as desvantagens do MR1, que não funciona com programas não-MapReduce, possui escalabilidade limitada e uso ineficiente da capacidade de processamento.

## EXERCÍCIO DE FIXAÇÃO

Em relação a Fase Map, é possível afirmar que:

As tarefas Map são agendadas de acordo com InputSplits.

‘Expecifica como os dados são extraídos de um arquivo.’

Os dados são transferidos fisicamente entre os nós.

Possui limitações na abordagem de escalonamento.

30:34

34:55

### Java MapReduce API

Todos os dados no Hadoop devem ser serializáveis, processo amplamente presente na API Java MapReduce. Para isso, por meio do InputFormats, o MapReduce utiliza a informação de um método de entrada, a fim de definir o esquema de visualização da informação. Ou seja, como os dados são extraídos de um arquivo.

## AULA 5 • PARTE 2

## EXERCÍCIO DE FIXAÇÃO

Qual dos componentes do MapReduce é responsável por enviar a aplicação e sua configuração (Job) ao ResourceManager?

Mapper.

Reducer.

Java.

Drive.

03:14

04:08

### Componentes de um MapReduce

Um MapReduce típico contém os seguintes componentes:

- Driver: código executado no cliente que configura e inicia a aplicação MApReduce;
- Mapper: classe Java que contém o método map;
- Redutor: classe Java que contém o método reduce.

## EXERCÍCIO DE FIXAÇÃO

Qual dos componentes do MapReduce é executado sobre uma partição e respectivas chaves ordenadas?

Mapper.

Reducer.

Java.

Drive.

08:06

10:55

## Exemplo de código

O professor Tiago Coelho Ferreto demonstra um exemplo de código de Word Count, uma aplicação de introdução do programador ao mundo do MapReduce, aplicando suas funções básicas.

21:21

**Hadoop Streaming****Mrjob**

O Mrjob é um framework específico para criação de aplicações MapReduce utilizando Python, que busca solucionar algumas das suas limitações. Ele permite a escrita de tarefas MapReduce com várias etapas, o teste em máquina local e a execução em um cluster Hadoop ou Hadoop na nuvem.

25:30

O Hadoop Streaming permite a implementação de funções Map e Reduce em linguagens diferentes de Java, utilizando entrada e saída das fases pela entrada padrão (STDIN) e saída padrão (STDOUT). Sua vantagem é que não se faz necessária a utilização de Java, simplificando a implementação. Apesar disso, possui algumas limitações, uma vez que o código se torna mais artesanal.

**AULA 5 • PARTE 3**

00:46

**Prática de MapReduce I****Prática de MapReduce II**

O professor Tiago Coelho Ferreto analisa o resultado da aplicação MapReduce, e chama atenção para o tamanho do arquivo de entrada, em relação a ajustes e ganhos, a fim de se beneficiar do paralelismo. O professor segue a demonstração utilizando o Hadoop Streaming.

10:44

O professor Tiago Coelho Ferreto realiza a prática do exercício de Wordcount em Java MapReduce API, no cluster YARN, já configurado anteriormente.

20:21

**Prática de MapReduce III****Padrões de MapReduce para processamento de dados I**

O professor Tiago Coelho Ferreto apresenta os padrões de MapReduce para processamento de dados, utilizando três datasets distintos. Para iniciar é necessário descompactar os datasets para o diretório.

Em seguida o professor apresenta o padrão count, para contar palavras ou ocorrências. O segundo exemplo de padrão é o Max Value, que verifica o valor máximo.

29:28

39:51

**Padrões de MapReduce para processamento de dados II**

O professor Tiago Coelho Ferreto segue com o exemplo demonstrando o padrão Average, que calcula uma média, e o Top N, que resulta em uma ordenação da qual podem ser extraídos os valores iniciais com maior ocorrência, por exemplo, montando um rank.

Outro padrão é o Filter, utilizado para realizar filtragens, retornando somente as informações triviais. Para valores distintos, utiliza-se o padrão District, agrupando os valores pela chave.

O padrão Binning possibilita o agrupamento em categorias nas entradas. Para criação de índice invertido, utiliza-se o padrão Inverted index.

## Padrões de MapReduce para processamento de dados III

50:00

O professor Tiago Coelho Ferreto finaliza a demonstração apresentando os padrões sort, para ordenação crescente ou decrescente dos valores, e o padrão join, para combinação de dados de mais de uma tabela.

## AULA 5 • PARTE 4

### Instruções

São instruções do Pig:

- Filter: filtra tuplas de uma bag;
- Foreach: equivale a uma operação Map em um job MR;
- Order by: ordena as tuplas por um determinado campo;
- Describe: inspeciona o esquema de uma bag;
- Illustrate: além de apresentar o esquema, apresenta seus predecessores e dados de amostra;
- Funções integradas: normalmente operam em um campo e são usados com o operador FOREACH;
- Group: permite agrupar registros por um campo específico;
- Cogroup: permite agrupar itens de várias bags;
- Join: combina registros de duas bags com base em um campo comum.

01:18  
13:04

### Pig

Iniciado no Yahoo!, em 2006, tem como objetivo fornecer uma linguagem alternativa para programar MapReduce, onde o interpretador (Grunt) recebe instruções na linguagem PigLatin, transformando-as em tarefas MR, e enviando-as ao cluster, monitorando seu progresso e retornando os resultados para o console ou os salva em HDFS.

16:16



### PALAVRAS-CHAVE

**Sequência de Fibonacci:** é uma sequência de números inteiros, começando normalmente por 0 e 1, na qual cada termo subsequente corresponde à soma dos dois anteriores.

### Pig na prática I

O professor Tiago Coelho Ferreto realiza na prática a utilização do Pig, demonstrando a sua instalação, descompactação e configuração das variáveis de ambiente.

### Pig na prática II

O professor Tiago Coelho Ferreto segue com a prática da utilização do Pig, aplicando as instruções vistos em aula: Filter, Foreach, Order by, Describe, Illustrate, Funções integradas, Group, Cogroup e Join.

29:14  
47:03

## AULA 6 • PARTE 1

00:29

Hive

### EXERCÍCIO DE FIXAÇÃO

Assinale a alternativa que apresenta uma diferença do Hive em relação a uma plataforma de banco de dados convencional:

Possui rollbacks.

Não possui chaves primárias.

Updates implementados através de um grão grosso.

Nenhuma das anteriores.



03:41

O Hive é um framework que trabalha com a linguagem SQL, similar ao Pig, para trabalhos em Hadoop. Ele analisa e consulta em HiveQL e gera operações Java Mr, enviados ao cluster Hadoop. Ao final do processo, o Hive monitora e retorna resultados ao cliente.

05:13



### PALAVRAS-CHAVE

**Rollbacks:** possibilita reverter edições rapidamente, sejam elas edições próprias ou edições de outros usuários, e bloquear vândalos.

### EXERCÍCIO DE FIXAÇÃO

Qual o grande diferencial do Hive?

Metastore.

Rollbacks.

Chaves primárias.

A linguagem SQL.



06:36



07:34

### PALAVRAS-CHAVE

**HCatalog:** é uma ferramenta que permite acessar tabelas metastore Hive no Pig, Spark SQL e/ou aplicativos MapReduce personalizados

07:46

### Interações e introduções Hive

As interações com o Hive são através de linhas de comando em formato shell, comum para realização de consultas e carregamento de dados. O HiveServer2, uma segunda versão do Hive, busca solucionar a limitação, possibilitando multi sessões para vários clientes.

O Hive também fornece suporte à execução não interativa ou em lote. Além disso, os objetos Hive consistem basicamente em bancos de dados e tabelas.

### Analizando dados com Hive

A HiveQL é baseada na especificação SQL-92, com algumas funções adicionais específicas do Hive. Suas instruções podem abranger várias linhas e são encerradas por um ponto e vírgula. Comentários de linha única são suportados usando o hífen duplo. E a semântica SQL típica, como listas de colunas e cláusulas WHERE, são totalmente suportadas.

18:06

## PALAVRAS-CHAVE

22:44



26:08

**Partições:** é uma divisão do espaço de um disco rígido (SCSI ou SATA). Cada partição pode conter um sistema de arquivos diferente.

**Buckets:** é mais comumente um tipo de buffer de dados ou um tipo de documento no qual os dados são divididos em regiões.

## PALAVRAS-CHAVE

**Struct:** também conhecidas como registros, definem tipos de dados que agrupam variáveis sob um mesmo tipo de dado.

## AULA 6 • PARTE 2

00:45

### Utilizando o HIVE I

#### Utilizando o HIVE II

O professor Tiago Coelho Ferreto inicia o exercício prático do framework Hive, aplicando as instruções vistas em aula.

08:24

O professor Tiago Coelho Ferreto demonstra como baixar, descompactar, instalar e configurar as variáveis de ambiente e path do Hive, carregando também o banco de dados que será utilizado no exemplo.

15:15

### Utilizando o HIVE III

O professor Tiago Coelho Ferreto segue com o exercício no Hive, apresentando os resultados das instruções utilizadas.

## AULA 6 • PARTE 3

### Spark I

Sendo uma das ferramentas de Big Data mais recentes e populares, o Spark foi iniciado em 2009 na universidade Berkeley RAD Lab, como alternativa ao MapReduce no Hadoop. Tem como benefícios uma melhor performance, extensibilidade e melhor suporte para outros cenários.

O Spark é escrito em Scala, permitindo que desenvolvedores criem rotinas complexas de processamento de dados multi-estágios. Implementa uma estrutura distribuída e tolerante a falhas na memória chamada RDD.

00:00

## PALAVRAS-CHAVE

**Apache:** é o servidor web livre criado em 1995 por Rob McCool.



## EXERCÍCIO DE FIXAÇÃO

Qual é o diferencial do Spark?

Utiliza SQL.

Tem sua própria biblioteca.

Tem vários módulos e trabalha com várias bibliotecas de processamento de dados e machine learning.

Todas as anteriores.

## Spark I

O Spark suporta diversos sistemas de entrada e saída, como:

- HDFS;
- Sistemas de arquivos locais ou de rede;
- Armazenamento de objetos, como Amazon S3 ou Ceph;
- Sistemas de banco de dados relacional;
- Bancos de dados NoSQL, incluindo Apache Cassandra, HBase e outros;
- Sistemas de mensagens, como Kafka.

08:40



## EXERCÍCIO DE FIXAÇÃO

É um gerenciador de cluster suportado pelo Spark:

Standalone.

Kubernetes.

Hadoop YARN.

Todas as anteriores.

## PALAVRAS-CHAVE

**Directed Acyclic Graph (DAG):** é um grafo orientado sem ciclos dirigidos, ou seja, ele consiste em vértices e arestas, com cada aresta direcionada de um vértice a outro, de forma que seguir essas direções nunca formará um loop fechado.



14:48

## Componente principal RDD

O componente principal do Spark é o RDD (Resilient Distributed Dataset), a forma como o Spark abstrai a informação, existindo desde o carregamento dos dados até seu resultado, suportando diversos tipos de dados, e armazenado em memória.

São características do RDD: Resiliente, Distribuído, Conjunto de dados, Sem compartilhamento, Imutabilidade.

## EXERCÍCIO DE FIXAÇÃO

Considerando as características do RDD, assinale a alternativa INCORRETA:

Pode ser reconstruído se um nó for perdido.

RDD's não são registros.

Os dados são divididos em partições e distribuídos como coleções de objetos na memória entre nós do cluster.

RDDs não podem ser atualizados após instanciados e preenchidos com dados.

19:56

15:24

### Como utilizar o Spark

O Spark trabalha com dois tipos de operações:

Transformações: operações realizadas contra RDDs resultando em novos RDDs;

Ações: operações que produzem saída de um RDD ou salvam o conteúdo de um RDD em um sistema de arquivos.

## AULA 6 • PARTE 4

00:00

• 03:57

### Extensões Spark

O Spark possui várias extensões:

- SparkSQL: fornece abstração semelhante a SQL para Spark;
- Spark Streaming: permite o processamento de fluxos de dados;
- SparkR: mecanismo de execução com a linguagem R;
- MLLib: biblioteca de aprendizado de máquina integrada ao Spark;
- GraphX: processamento de grafos com Spark.

### Spark na prática I

O professor Tiago Coelho Ferreto realiza um exercício prático de Spark, instalando, configurando e disparando uma aplicação, a fim de rever os conceitos vistos em aula.

### Spark na prática II

O professor Tiago Coelho Ferreto finaliza a demonstração do Spark utilizando a shell do pyspark executando um word count como exemplo, disparando a aplicação no cluster Hadoop.

• 14:12

19:38



## PALAVRAS-CHAVE

**Pyspark:** é uma API Python que ajuda a fazer a interface com conjuntos de dados distribuídos resilientes (RDDs) na linguagem de programação Apache Spark e Python.

# Artigos

Nesta página, você encontra links de artigos científicos, informativos e vídeos sugeridos pelo professor PUCRS.

## **ARTIGO CIENTÍFICO**

---

[Apache hadoop: conceitos teóricos e práticos, evolução e novas possibilidades](#)

[Apache spark: a unified engine for big data processing](#)

## **ARTIGO INFORMATIVO**

---

[Hadoop: o que é e qual sua relação com big data](#)

## **VÍDEO**

---

[Exploração no limite da grande quantidade de dados - Tim Smith](#)

[The Human Face Of Big Data | Trailer | PBS](#)

# Resumo da disciplina

Veja nesta página, um resumo dos principais conceitos vistos ao longo da disciplina.

## AULA 1

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



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

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



## AULA 2

O Spark possibilita a obtenção de resultados imediatos.



Certificações podem mostrar que você tem conhecimento do assunto.



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



## AULA 3

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



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



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



## 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.



## 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.



## 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).



# Avaliação

**Veja as instruções para realizar a avaliação da disciplina.**

Já está disponível o teste online da disciplina. O prazo para realização é de **dois meses a partir da data de lançamento das aulas**.

Lembre-se que cada disciplina possui uma avaliação online.  
A nota mínima para aprovação é 6.

Fique tranquilo! Caso você perca o prazo do teste online, ficará aberto o teste de recuperação, que pode ser realizado até o final do seu curso. A única diferença é que a nota máxima atribuída na recuperação é 8.

**Pós-Graduação em**  
Ciência de Dados e Inteligência Artificial