ibmec.br

Inovação e as Bases de IA, Data Science e Big Data

Aula 4

Professor: Marcius Linhares

Ibmec

Conteúdo

Nº AULA	DATA	CONTEÚDO
1	27/08/2024 (terça-feira)	Introdução à Inovação, Business Intelligence e Tecnologia
2	28/08/2024 (quarta-feira)	Fundamentos de Inteligência Artificial e Soluções Emergentes
3	03/09/2024 (terça-feira)	Fundamentos de Data Science
4	04/09/2024 (quarta-feira)	Fundamentos de Big Data
5	10/09/2024 (terça-feira)	Integração de IA, Data Science e Big Data na Inovação
6	11/09/2024 (quarta-feira)	Tendências, Ética e o Futuro da IA e Big Data / Inicio Trabalho em Grupo

1. Introdução ao Big Data

O que é Big Data?

Big Data é um conjunto de dados extremamente amplo que exige ferramentas especiais para serem processados em tempo hábil. De forma mais simples, é a análise de grandes volumes de dados com o objetivo de gerar valor para o negócio. Esse conceito ganhou força nos últimos anos devido à popularização de smartphones, sensores, mídias sociais e serviços que tentam entregar soluções mais adequadas aos usuários.

Aula 4

Os 3 Vs do Big Data:

Volume: A quantidade massiva de dados gerados constantemente. Um exemplo é que 1 minuto de vídeo no YouTube pode ocupar aproximadamente 45 MB.

Velocidade: A rapidez com que os dados são gerados e processados, como cliques de usuários em sites de e-commerce ou a localização contínua dos smartphones.

Variedade: A diversidade dos formatos de dados, que podem ser estruturados (armazenados em tabelas de banco de dados), semiestruturados (dados em formato XML ou JSON) e não estruturados (como arquivos PDF, vídeos, ou e-mails).

A variedade de dados é grande, e podem ser classificadas em três grupos:

Estruturados: dados geralmente armazenados em tabelas de bancos de dados.

Semiestruturados: dados que, apesar de não estarem em bancos de dados, possuem marcações ou *tags* que dão a indicação semântica dos dados, geralmente, disponibilizados em formato xml ou json.

Não estruturados: são arquivos diversos encontrados nos computadores, como pdf, doc, xls, txt, *e-mail*, imagem, vídeo, por exemplo. Esse tipo de dado não era muito explorado pelas soluções tecnológicas.

Exemplo Arquivo XML

```
<cliente>
    <id>123</id>
    <nome>Maria Silva</nome>
    <email>maria.silva@email.com</email>
</cliente>
```

Exemplo Arquivo JSON

```json
{
  "id": 123,
  "nome": "Maria Silva",
  "email": "maria.silva@email.com"
}
```

2. Funcionalidades de Big Data

Processamento Distribuído (Hadoop e MapReduce)

Hadoop é uma das plataformas mais populares para processamento distribuído de grandes volumes de dados. Ele permite que clusters de computadores comuns sejam usados para armazenar e processar dados, dividindo a tarefa entre várias máquinas e tornando o processo mais eficiente e acessível financeiramente.

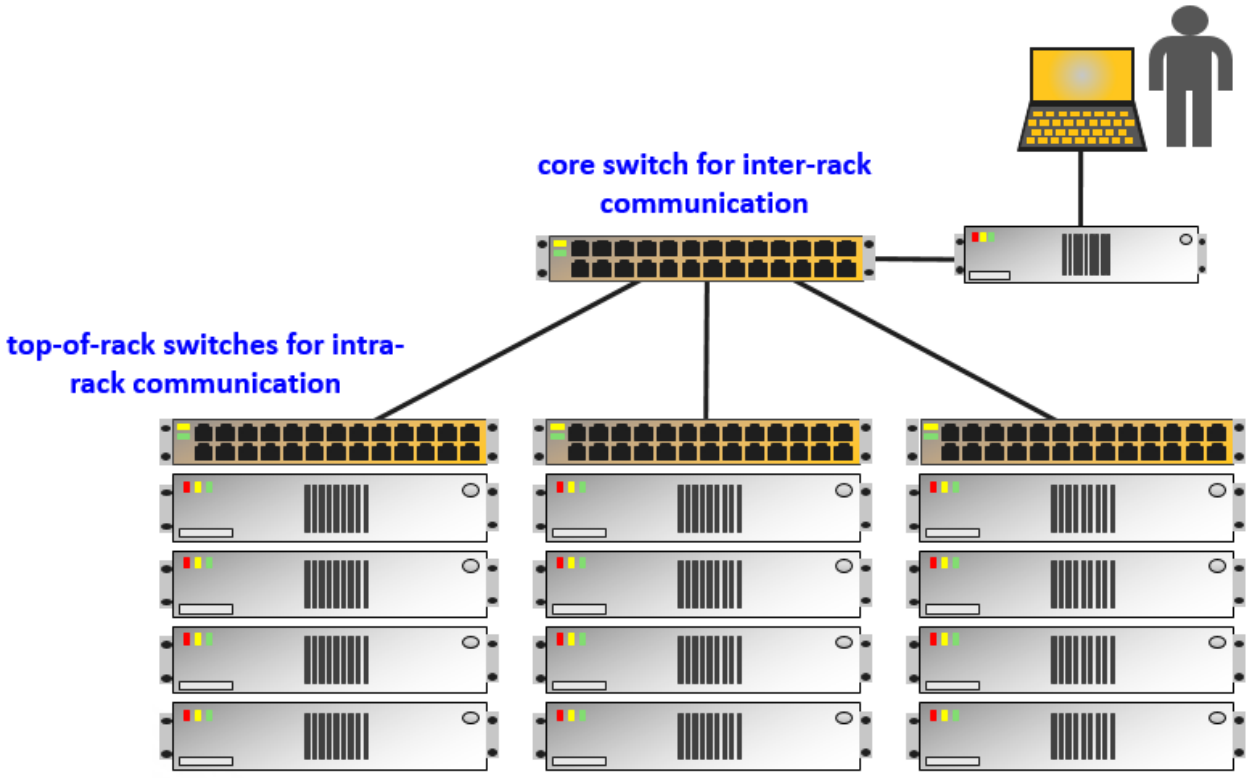
Exemplos Práticos:

Comércio Eletrônico: Plataformas online coletam cada clique dos usuários para oferecer recomendações personalizadas, criar campanhas de marketing direcionadas e melhorar a experiência do cliente.

Ministério Público: Nos sistemas judiciais, Big Data é usado para processar grandes volumes de documentos judiciais e detectar padrões em processos, como identificar em quais cidades certas ações têm mais chance de sucesso.

Aula 4

Cluster



Clusters é o nome dado ao conjunto de computadores que trabalham de forma sincronizada para funcionar como um único computador.

Aula 4

HDFS – Hadoop File System

Hadoop Distributed File System (HDFS), é responsável por gerenciar o disco das máquinas que compõem o Cluster. HDFS também serve para leitura e gravação dos dados.

HDFS – Características

Tolerância a falhas e recuperação automática;

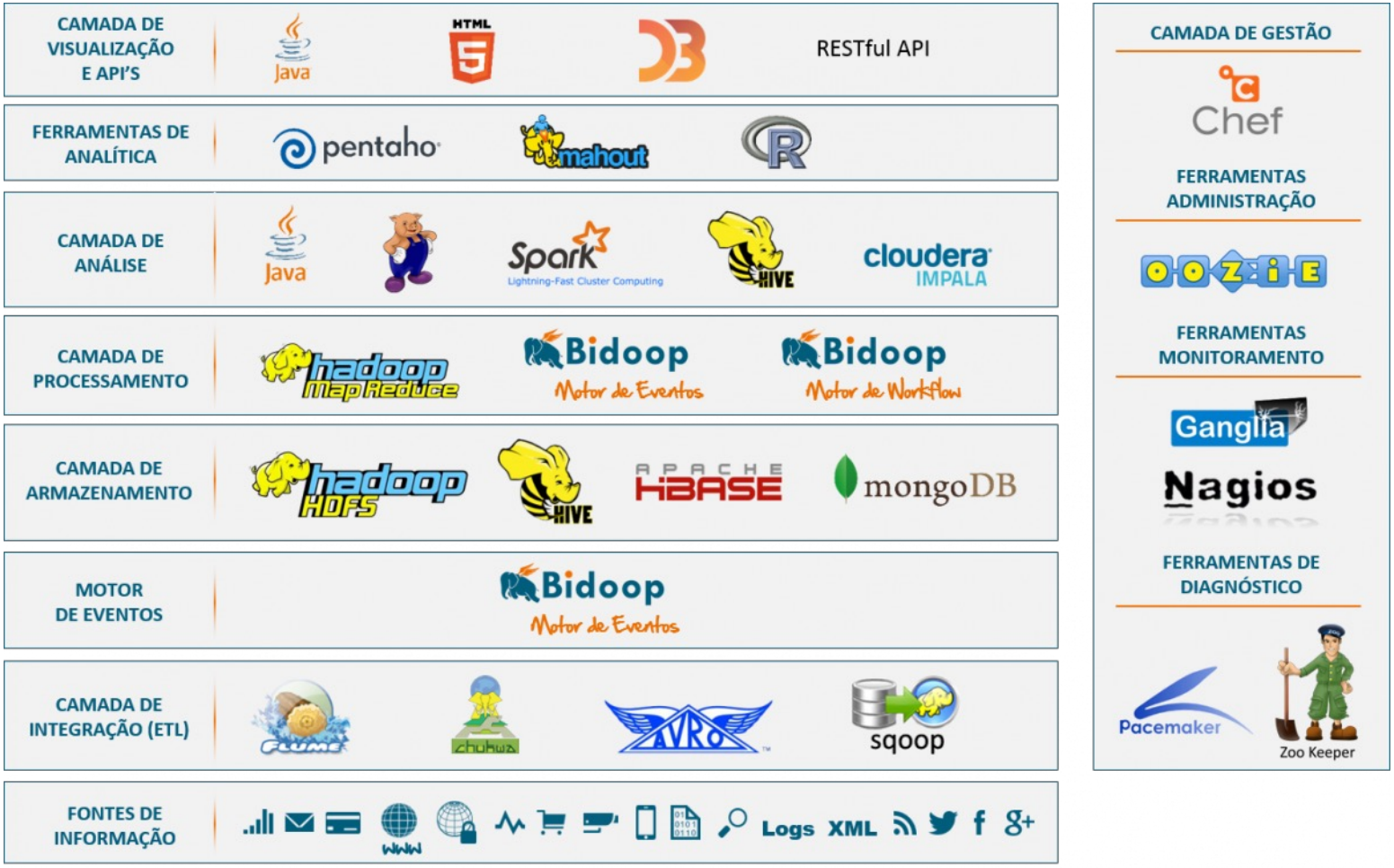
Portabilidade entre hardware e sistemas iguais;

Escalabilidade para armazenar grande volume de dados;

Confiabilidade, através de diversas cópias de dados.

Aula 4

Arquitetura Hadoop



Aula 4

ZooKeeper

ZooKeeper é um serviço de coordenação distribuída para gerenciar grandes conjuntos de Clusters.

Oozie

Apache Oozie é um sistema de agendamento de WorkFlow, usado para gerenciar principalmente os Jobs de MapReduce.

Pig

Apache Pig, é uma linguagem de procedimentos de alto nível para consultar grandes conjuntos de dados semiestruturados usando Hadoop e a Plataforma MapReduce

Aula 4

Sqoop

Apache Sqoop, é um projeto do ecossistema Hadoop, cuja responsabilidade é importar e exportar dados do banco de dados de dados relacionais.

Spark

Apache Spark, é uma ferramenta Big Data para o processamento de grandes conjuntos de dados. Foi desenvolvido para substituir o MapReduce, pois processa 100x mais rápido que o MapReduce.

HBase

Apache Hbase, é um banco de Dados não relacionais, projetado para trabalhar com grande conjunto de dados (Big Data). É o banco de dados oficial do hadoop.

Kafka

Apache Kafka, é foi desenvolvido pelo Linkedin e liberado como projeto Open-source em 2011. O Apache Kafka é um sistema para gerenciamento de fluxo de dados em tempo real, gerados a partir de websites, aplicações e sensores.

Exemplos utilização Apache Kafka

1. Monitoramento de Transações Bancárias em Tempo Real

Cenário: Um banco quer monitorar as transações de seus clientes em tempo real para detectar atividades suspeitas.

Principais Serviços de Big Data na AWS

Amazon S3 (Simple Storage Service)

Função: Armazenamento de dados escalável e seguro. Ideal para armazenar grandes volumes de dados não estruturados.

Aplicação: Armazenamento de *data lakes*, onde os dados podem ser armazenados de forma bruta e acessados por outros serviços para análise e processamento.

Vantagens: Alta durabilidade (99,999999999%) e integração com outros serviços da AWS.

Amazon EMR (Elastic MapReduce)

Função: Plataforma gerenciada para processar grandes volumes de dados usando ferramentas de código aberto, como Hadoop, Spark, HBase e Presto.

Aplicação: Processamento de grandes volumes de dados para tarefas como análise de logs, movimentação de dados de ETL, e processamento de grandes datasets de maneira paralela.

Vantagens: Escalabilidade automática e otimização de custos em comparação com soluções tradicionais on-premises.

Amazon Kinesis

Função: Plataforma de streaming de dados em tempo real.

Aplicação: Processar e analisar dados em tempo real, como logs de transações financeiras, cliques em websites e dados de sensores IoT.

Vantagens: Permite responder a eventos assim que ocorrem, possibilitando análises em tempo real e insights instantâneos.

AWS Glue

Função: Serviço de ETL (Extração, Transformação e Carregamento) totalmente gerenciado.

Aplicação: Automatiza o processo de descoberta de dados, transforma esses dados e os move entre data lakes e data warehouses.

Vantagens: Simplicidade na criação de fluxos de dados complexos, com interface de arrastar e soltar para usuários menos técnicos.

Amazon Redshift

Função: Data warehouse altamente escalável e gerenciado.

Aplicação: Executar consultas SQL em grandes volumes de dados, ideal para relatórios empresariais e análises complexas.

Vantagens: Capacidade de processamento em massa de dados estruturados, com integração com ferramentas de BI como Tableau e Power BI.

Google Cloud Platform (GCP), há diversas funcionalidades voltadas para Big Data, oferecendo soluções robustas e escaláveis para processamento, armazenamento e análise de grandes volumes de dados.

BigQuery

Funcionalidade: BigQuery é um data warehouse totalmente gerenciado e sem servidor, otimizado para análise de grandes volumes de dados. Ele permite consultas SQL rápidas em datasets grandes com baixa latência.

Cloud Dataflow

Funcionalidade: Cloud Dataflow é um serviço gerenciado para o processamento de dados em modo *streaming* e *batch*. Ele usa o modelo Apache Beam para transformar e enriquecer dados de forma escalável.

Cloud Storage

Funcionalidade: Cloud Storage oferece armazenamento de objetos escalável e seguro, projetado para armazenar grandes volumes de dados. Ele é ideal para armazenar *data lakes* e dados não estruturados.

Cloud Composer

Funcionalidade: Cloud Composer é um serviço gerenciado de orquestração de fluxo de trabalho baseado no Apache Airflow. Ele permite a automação e o gerenciamento de pipelines de dados.

Aula 4

AI Platform

Funcionalidade: O AI Platform permite que você desenvolva, treine e implante modelos de aprendizado de máquina de forma escalável. Ele oferece um ambiente gerenciado para treinamentos com TensorFlow e outras bibliotecas de machine learning.

Google Vertex AI

Funcionalidade: O Vertex AI é uma plataforma unificada para construção, implantação e gestão de modelos de aprendizado de máquina (ML) e inteligência artificial (IA). Ele permite o treinamento de modelos customizados, implantação de pipelines de ML e integração com outras ferramentas de Big Data no Google Cloud.

Microsoft Azure e exemplos de uso prático para cada uma, com foco em como essas ferramentas podem ajudar a resolver desafios de *Big Data* e oferecer insights em tempo real.

Azure Synapse Analytics

Funcionalidade: Azure Synapse Analytics é uma plataforma de análise integrada que combina data warehousing e Big Data para ingerir, preparar, gerenciar e servir dados para a análise.

Azure Data Lake Storage

Funcionalidade: Azure Data Lake Storage é um serviço de armazenamento escalável e seguro, projetado para armazenar grandes volumes de dados não estruturados e estruturados, facilitando a criação de *data lakes*.

Azure Databricks

Funcionalidade: Azure Databricks é uma plataforma de análise baseada no Apache Spark, otimizada para processamento rápido de Big Data. Ele permite o desenvolvimento colaborativo de ciência de dados, machine learning e engenharia de dados.

Azure Stream Analytics

Funcionalidade: Azure Stream Analytics é um serviço de processamento de fluxo de dados em tempo real. Ele permite que você realize consultas SQL em fluxos de dados em tempo real e gere insights instantâneos.

Azure HDInsight

Funcionalidade: Azure HDInsight é um serviço gerenciado que permite executar clusters Hadoop, Spark, Kafka, HBase e outras tecnologias de Big Data. Ele oferece uma plataforma escalável para processamento de grandes volumes de dados.

Azure Data Factory

Funcionalidade: Azure Data Factory é um serviço de integração de dados que permite a criação, orquestração e automação de pipelines de dados para ETL (Extração, Transformação e Carga).

Azure Cosmos DB

Funcionalidade: Azure Cosmos DB é um banco de dados NoSQL distribuído globalmente e altamente escalável. Ele oferece latência baixa e alta disponibilidade, sendo ideal para cenários que exigem rápida recuperação de dados.

Azure Machine Learning

Funcionalidade: Azure Machine Learning é um serviço que permite criar, treinar e implantar modelos de machine learning em escala, com suporte para ferramentas e frameworks como TensorFlow, PyTorch e Scikit-learn.

Azure Cognitive Services

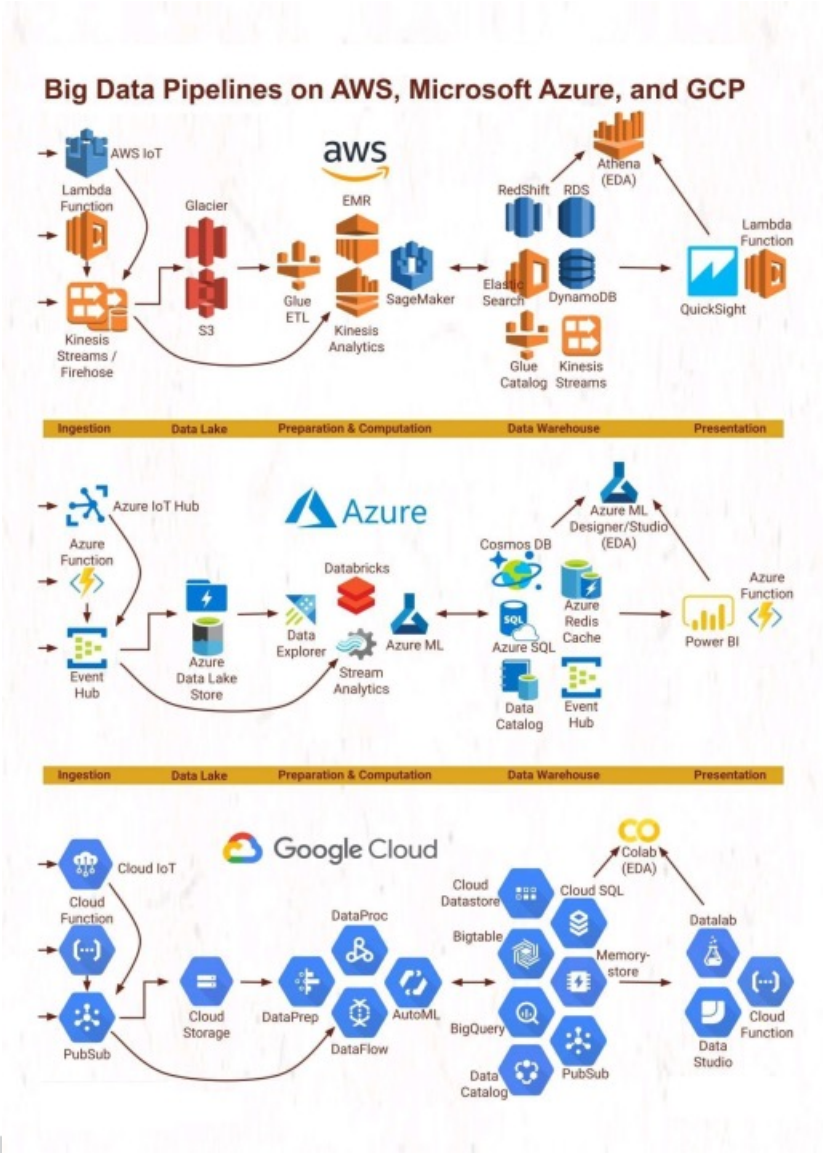
Funcionalidade: Azure Cognitive Services oferece APIs baseadas em inteligência artificial para adicionar capacidades como visão computacional, reconhecimento de fala e análise de sentimentos às aplicações.

Azure Blob Storage

Funcionalidade: Azure Blob Storage é um serviço de armazenamento de objetos escalável e econômico, projetado para armazenar grandes volumes de dados não estruturados, como imagens, vídeos e documentos.

Aula 4

Aula 4



IBMEC.BR

/IBMEC

IBMEC

@IBMEC_OFICIAL

@IBMEC

Ibmec