Uma imagem contendo Forma

Descrição gerada automaticamente

Nome: Paulo Gurgel Tudini

RM: 564418

Checkpoint-1 -1TSCP

Arquitetura de Big Data e Integração de Dados

1. (1,5) Escreva o comando hdfs completo para ler um arquivo no **sistema operacional** e copiar para o Hadoop HDFS.

O arquivo(**origem**) encontra-se no sistema operacional local no caminho abaixo:

/home/arquivo.csv

O arquivo deverá ser copiado para o destino no hdfs:

/inboud

Comando para Ler e copiar: hdfs dfs -get /home/arquivo.csv /inboud

1. (1,5) Descreva o que você entende como Big Data, para que serve esta área de conhecimento e sua utilização.

O Big Data é um termo utilizado para uma área de dados em que sua escalabilidade é fundamental e essencial para o seu funcionamento, ela serve para fazer a manipulação de uma grande quantidade de dados ao mesmo tempo e de diversos tipos, sejam eles estruturados, semiestruturados ou não estruturados.

1. (1,5) Explique ou defina com as suas palavras o que é um Cluster de Computadores e sua aplicabilidade.

O Cluster de Computadores é um conjunto de computadores, onde um computador é o principal, porém juntos, eles funcionam como um só, sendo aplicável para gerenciamento de dados, gerenciamento de arquivos e recursos.

1. (1,5) Explique, Dados Estruturados, Dados Semiestruturados e Dados Não Estruturados. (Dê exemplos)

**Dados Estruturados** possuem uma estrutura organizada e definida, como por exemplo: Arquivos CSV, Planilhas e Banco de Dados Relacional.

**Dados Semiestruturados** possuem uma estrutura não tão organizada como os dados estruturados, porém partem do pressuposto de uma estrutura parcialmente definida, como por exemplo: E-mail.

**Dados Não Estruturados** não possuem uma estrutura organizada ou uma definição de como os dados são inseridos, como por exemplo: Vídeo, Imagens, Conversas no Whatsapp etc.

**Múltipla Escolha (1 ponto)**

1. No ecossistema Hadoop quem é o responsável por gerenciar os recursos disponíveis:
2. O Sistema Operacional
3. O HDFS
4. O Yarn
5. O usuário
6. Dentre os principais V’s que definem o termo Big Data, temos: **Valor** **e Variedade**.

São definidos como: (**Escolha 2 respostas** **corretas)**

1. Valor: O quanto custa para armazenarmos os dados
2. Variedade: O dado varia de um dia para o outro
3. Variedade: Diferentes formatos de dados e estruturas
4. Valor: Os gastos que as empresas têm com Big Data
5. Valor: O retorno desses dados para os negócios e sociedade
6. Variedade: A variedade de profissionais em Big Data
7. Na Arquitetura Hadoop temos o Master Node e o Worker Node. Existem 3 componentes básicos que formam um **Worker Node**, são eles:
8. DataNode, NameNode, JobTracker.
9. DataNode, NodeManager, MapReduce.
10. NameNode, ResourceManager, JobTracker.
11. DataNode, ResourceManager, MapReduce.
12. Ao executar o comando hdfs dfs -ls /ingest

Obtivemos o seguinte resultado:

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

O que tem dentro do diretório /ingest?

1. Um arquivo
2. Um diretório
3. Está vazio
4. Dois arquivos