Mestrado em Engenharia Informática - Bases de Dados Avançadas
Ano letivo: 2022/2023 - 1.º semestre
Trabalho Teórico/Prático II – Enunciado

Introdução

O trabalho é realizado em grupo, com a distribuição realizada para este 1.º semestre.

A submissão do trabalho será feita na página da disciplina no Moodle até quarta-feira, 15 de fevereiro às 23h59. A apresentação do trabalho será feita no dia 18 de fevereiro (sábado).

Entregáveis:

1. Relatório.
   No relatório deve documentar todos os passos seguidos, decisões arquiteturais tomadas, e justificação para as mesmas.

2. Código.

Enunciado

Uma empresa que comercializa bicicletas numa cadeia de lojas de mobilidade sustentável, está a mudar o seu sistema de dados de bases de dados relacionais on-premise, para Big Data.

Além dos dados novos pretende integrar no seu novo sistema os dados históricos de vendas, que foram exportados dos sistemas atuais em formato CSV.

O armazenamento de dados será feito em HDFS, e será utilizado o Hive para gestão de dados.

Na elaboração do seu projeto deverá considerar os seguintes pedidos:

1. Efetue a modelação dos dados de forma que seja possível depois realizar análises de dados de forma complexa utilizando a linguagem de consulta em Hive;

2. Escolha o melhor formato em termos de armazenamento de dados em Hive de forma a garantir um bom desempenho do sistema;
   Para justificar a escolha deve indique quais os formatos de ficheiros em Hive, características, principais vantagens de cada um deles, e exemplos de em que tipo de projetos ou dados devem ser utilizados por serem mais adequados.

3. Implemente o modelo de dados em Hive, e nas suas decisões de arquitetura tenha em consideração a melhor forma de garantir a melhor performance possível;
   Por exemplo relativo a particionamento, managed ou external, etc...

   A implementação dos objectos em Hive (ex.º tabelas) deve ser sempre feita através de código.

4. Efetue o upload de dados para as tabelas do modelo de dados criado;

   Esta tarefa deve ser feita através de código.

5. Implemente as seguintes consultas em Hive.

   Crie uma consultaque retorne os produtos em que o preço de venda (ListPrice) seja maior que o preço médio unitário (preço unitário – UnitPrice).

   Crie uma consulta que retorne todos os produtos com um preço de venda de 100 ou maior, que foram vendidos por menos de 100.

   Crie uma consulta que retorne o custo (StandardCost), preço de venda (ListPrice), e preço de venda médio, para cada produto

   Crie uma consulta que retorne os produtos que tenham um preço médio de venda menor que o de custo.

   Crie uma consulta que permita listar a quantidade média de vendas por produto, em quantidade por produto, apresentando o ID do produto (ProductID), o nome do produto (tabela Product, campo Name), a categoria do produto (tabela Product Category, campo Name) e a quantidade vendida (tabela SalesOrderDetail, OrderQTY). A lista deve ser ordenada do produto mais vendido para o menos vendido.

Crie uma consulta que permite listar o volume de vendas por cliente, ou seja o número de compras efetuadas, apresentando o nome do cliente (tabela Customer, campo CompanyName) e número de compras do mesmo.

6. Efetue a implementação de algumas análises de dados em Zeppelin.

Para cada uma das consultas deve apresentar código, plano de execução, detalhes de execução. Caso faça otimizações deve indicar o processo seguido.

Bom trabalho,

Hélder Quintela