

Curso 03 - Exercícios de fixação

Quinzena 02 - Prof. Jose Fernando Rodrigues Junior

Material de apoio: arquivo Curso03-Quinzena02-ExerciciosFixacao-SUPORTE.ipynb

1) A base do ecossistema Hadoop é a combinação de um sistema de armazenamento distribuído, o HDFS, com um sistema de processamento distribuído, o Hadoop MapReduce. Este ecossistema consegue ser altamente escalável pois:

- (a) Novos nós de processamento e armazenamento podem ser facilmente inseridos no sistema; tais nós podem ser computadores *commodity* de baixo custo;
- (b) Seus nós de processamento são projetados para receber mais memória e processadores mais rápidos, ampliando a capacidade de processamento de todo o sistema;
- (c) Os nós de processamento de um cluster Hadoop funcionam sempre com frequências de processamento acima do estipulado pelo fabricante (*overclock*), bastando aumentar o *clock* sempre que houver mais demanda.
- (d) Sistemas Hadoop são sempre configurados com redes de alto desempenho, de modo que, havendo necessidade de mais processamento, o algoritmo MapReduce negocia automaticamente serviços adicionais na nuvem.

2) A prática de Business Intelligence se refere a um conjunto de softwares e serviços que visam trazer maior entendimento sobre os dados de uma empresa/instituição, norteados por ações, e promovendo o acompanhamento dos negócios com relação ao mercado/demanda. Dentre as classes de ferramentas que suportam o Business Intelligence, pode-se citar:

- (a) Tecnologia Data Warehouse, ferramentas de Online Transactional Processing, ambientes integrados de programação, e aplicações voltadas ao versionamento de software.
- (b) Tecnologia Data Warehouse, editores de texto avançados, softwares de gerenciamento de equipe, gerenciamento de recursos, e inteligência artificial.
- (c) Ferramentas de Online Analytical Processing, editores gráficos, software para edição e transmissão de vídeo, e para aprendizado de máquina.
- (d) Tecnologia Data Warehouse, ferramentas de Online Analytical Processing, softwares de visualização, estatística, aprendizado de máquina, e inteligência artificial.

3) Um data warehouse caracteriza-se como uma consolidação dos dados produzidos em bases de dados operacionais, muitas vezes heterogêneas; data warehouses são especializados para muita leitura e pouca escrita, e para propiciar o chamado *Online Analytical Processing* (OLAP), o qual prevê:

- (a) Análise de dados básica, incluindo contagens, somas, médias, máximo, mínimo, e ordenação.
- (b) Análise de dados estatística, incluindo distribuição de dados e teste de hipóteses.

(c) Análise de dados avançada, incluindo aprendizado de máquina nas modalidades de classificação e regressão.

(d) Aprendizado de máquina avançado, incluindo arquiteturas de redes neurais visando inteligência artificial.

4) Dentre as ferramentas do ecossistema Hadoop que permitem a criação e o uso analítico de data warehouses sobre o Apache Hive, destacam-se:

(a) O arcabouço de armazenamento e processamento distribuído Apache Hadoop, e a suíte de aplicativos Apache OpenOffice, em especial suas planilhas de cálculo Calc.

(b) O arcabouço OLAP Apache Kylin, o servidor web Apache HTTP Server, o servidor web Apache Tomcat, especializado em tecnologias Java, e o servidor de e-mail Apache JAMES.

(c) O arcabouço de armazenamento e processamento distribuído Apache Hadoop, o banco de dados Apache HBase, a ferramenta de carregamento de dados Apache Sqoop, e o arcabouço OLAP Apache Kylin.

(d) O banco de dados Apache HBase, a ferramenta de carregamento de dados Apache Sqoop, e o motor de busca textual Apache Lucene.

5) O Apache Hive é um projeto de código aberto que implementa uma infraestrutura de data warehouse. Dentre as afirmações abaixo, indique aquela que não é correta com relação às suas características:

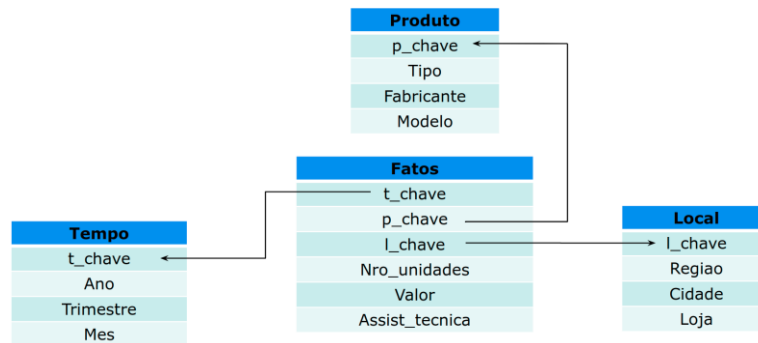
(a) Embora funcione como uma base de dados robusta, o Apache Hive não é capaz de substituir um sistema de banco de dados operacional (OLTP) devido à sua alta latência e ineficiência em acessar dados de maneira aleatória.

(b) Sendo um projeto de código aberto, o Hive é distribuído unicamente pela fundação Apache, sua mantenedora.

(c) Graças à infraestrutura Apache Hadoop, o Apache Hive é altamente escalável por meio da adição de nós de processamento e armazenamento, os quais não necessitam ser computadores especializados de alto custo.

(d) Trata-se de uma camada de software capaz de converter comandos SQL em jobs MapReduce, os quais são executados sobre uma infraestrutura computacional abstraída pelo arcabouço Apache Hadoop.

6) Considerando o esquema estrela a seguir:



Escreva consultas que computem a média de Nro_unidades:

⇒ OPCIONALMENTE: use o script “Curso03-Quinzena02-ExerciciosFixacao-SUPORTE.ipynb” para criar o esquema deste exercício com dados de teste.

a) Escreva a correspondente consulta OLAP para computar a média de Nro_unidades vendidas considerando a maior granularidade possível nas 3 dimensões:

```
SELECT      Tempo.Ano,      Produto.Tipo,      Local.Regiao,      AVG(Nro_unidades)      AS      MEDIA
FROM          Fatos,          Tempo,          Produto,          Local
WHERE          Fatos.t_chave=Tempo.t_chave      AND      Fatos.p_chave=Produto.p_chave      AND
Fatos.l_chave=Local.l_chave
GROUP        BY          Tempo.Ano,          Produto.Tipo,          Local.Regiao
ORDER BY MEDIA DESC -- ordenacao opcional
```

b) Agora refaça a consulta aplicando uma operação de Drill Down na dimensão Local:

```
SELECT      Tempo.Ano,      Produto.Tipo,      Local.Cidade,      AVG(Nro_unidades)      AS      MEDIA
FROM          Fatos,          Tempo,          Produto,          Local
WHERE          Fatos.t_chave=Tempo.t_chave      AND      Fatos.p_chave=Produto.p_chave      AND
Fatos.l_chave=Local.l_chave
GROUP        BY          Tempo.Ano,          Produto.Tipo,          Local.Cidade
ORDER BY MEDIA DESC -- ordenacao opcional
```

c) Sobre a consulta do item b), aplica uma operação de Drill Down na dimensão Tempo:

```
SELECT      Tempo.Ano,      Tempo.Trimestre,      Produto.Tipo,      Local.Cidade,      AVG(Nro_unidades)      AS      MEDIA
FROM          Fatos,          Tempo,          Produto,          Local
WHERE          Fatos.t_chave=Tempo.t_chave      AND      Fatos.p_chave=Produto.p_chave      AND
Fatos.l_chave=Local.l_chave
GROUP        BY          Tempo.Ano,          Tempo.Trimestre,          Produto.Tipo,          Local.Cidade
ORDER BY MEDIA DESC -- ordenacao opcional
```

⇒ Note que no item b), bastou um único atributo (Local.Cidade) para se fazer o Drill Down na dimensão Local; já a dimensão Tempo necessitou de dois atributos (Tempo.Ano, Tempo.Trimestre). Isto se deve à semântica dos dados. Na dimensão tempo, um trimestre é unicamente identificado apenas se for acompanhado do ano a qual pertence; já uma cidade não precisa ser acompanhada de uma região para ser unicamente identificada. Caso a granularidade acima de Cidade fosse Estado, aí sim, seriam necessários dois atributos.

d) Sobre a consulta do item c), aplique uma operação de slicing na dimensão Tempo considerando os anos entre 2014 e 2017:

```
SELECT Tempo.Ano, Tempo.Trimestre, Produto.Tipo, Local.Cidade, AVG(Nro_unidades)
AS
MEDIA
FROM      Fatos,           Tempo,           Produto,           Local
WHERE     Fatos.t_chave=Tempo.t_chave      AND     Fatos.p_chave=Produto.p_chave      AND
Fatos.l_chave=Local.l_chave      AND     Tempo.Ano      BETWEEN      2014      AND      2017
GROUP    BY      Tempo.Ano,      Tempo.Trimestre,      Produto.Tipo,      Local.Cidade
ORDER BY MEDIA DESC -- ordenacao opcional
```

e) Usando a sintaxe SQL ROLAP, expresse uma consulta ROLAP tal que as agregações realizadas inclua a agregação do item a):

```
SELECT  Tempo.Ano,  Produto.Tipo,  Local.Regiao,  AVG(Nro_unidades)  AS  MEDIA
FROM      Fatos,           Tempo,           Produto,           Local
WHERE     Fatos.t_chave=Tempo.t_chave      AND     Fatos.p_chave=Produto.p_chave      AND
Fatos.l_chave=Local.l_chave
GROUP    BY      CUBE (Tempo.Ano,           Produto.Tipo,           Local.Regiao)
ORDER BY Tempo.Ano, Produto.Tipo, Local.Regiao, MEDIA -- ordenacao opcional
```