Orientações para elaboração do trabalho1 de Tecnologias Big Data

2º semestre 2022 – Prof. Bianca

Atualizado em 12/10/2022

ETL de dados usando ferramentas analíticas da AWS

Implementação de um **ETL/pipeline de dados**, na AWS e utilizar a linguagem SQL para fazer análise dos dados. Exibir os resultados mais relevantes. Trabalho pode ser feito em dupla e deve ser entregue até 06/nov.

1. Organização do conteúdo do trabalho

Os trabalhos devem ter os seguintes tópicos:

a) Introdução: apresentar uma visão geral da solução, com o diagrama de arquitetura AWS implementada

b) Infraestrutura (região, buckets, recursos (número e configuração das máquinas))

c) Metadados (descrição dos dados, colunas, tabelas, etc)

d) Os scripts de 10 consultas SQL para análise de dados. Esses scripts devem incluir comandos de consulta úteis em análise de dados, tais como:

1. Junções e suas variações (join, left join, right join)
2. Agregações (Group by, having, max, min, avg, sum)
3. Subconsultas e Funções (not in, when, date_format, concat, matemáticas)
4. Ordernações (order by, limit)
5. funcões analíticas (partition, rank, etc)

Obs: Apresentar as consultas SQL **mais relevantes** do trabalho com destaques para a sintaxe dos comandos utilizados e efeitos do comando no BD. Não repetir comandos, por exemplo, não pode usar as mesmas funções e operadores e só mudar o nome das tabelas. Consultas simples usadas apenas para apresentar uma amostra dos dados não são consideradas na contagem.

e) Visualizações das consultas mais relevantes (máximo 3).

2. Tema

O tema do banco de dados é de livre escolha dos participantes. Entretanto, o banco de dados deve ter de 2 a 4 "*data source*"(csv, json, xml).

Usar uma ou mais das ferramentas analíticas: athena, glue, redshift ou pipeline de dados.

3. Forma de entrega e apresentação

Um relatório contendo o diagrama da infraestrutura AWS utilizada e a descrição dos passos utilizados (**não precisa ser tão detalhado quanto um tutorial**). Os scripts SQL utilizados e evidências dos resultados gerados (até 5 linhas, como nos laboratórios de data analytics).**Pode postar no GIT ou Google Drive e só enviar o link pelo moodle**

A infraestrutura utilizada, juntamente com os dados deve ficar acessível no **learner labs ou no GIT** até **15/novembro** para avaliação. Quem optar por fazer em outro ambiente deve fazer uma apresentação ou demonstração em vídeo.

Contéudo	Pontuação
Introdução	1 (texto)
Metadados	1 (esquema ou DER)
Consultas SQL	5 (10 consultas)
Visualização dos resultados	1 (3 visualizações)
Infraestrutura	1 (Diagrama de arquitetura)
Dados (volume, qualidade, fonte)	1 (texto)

Exemplos:

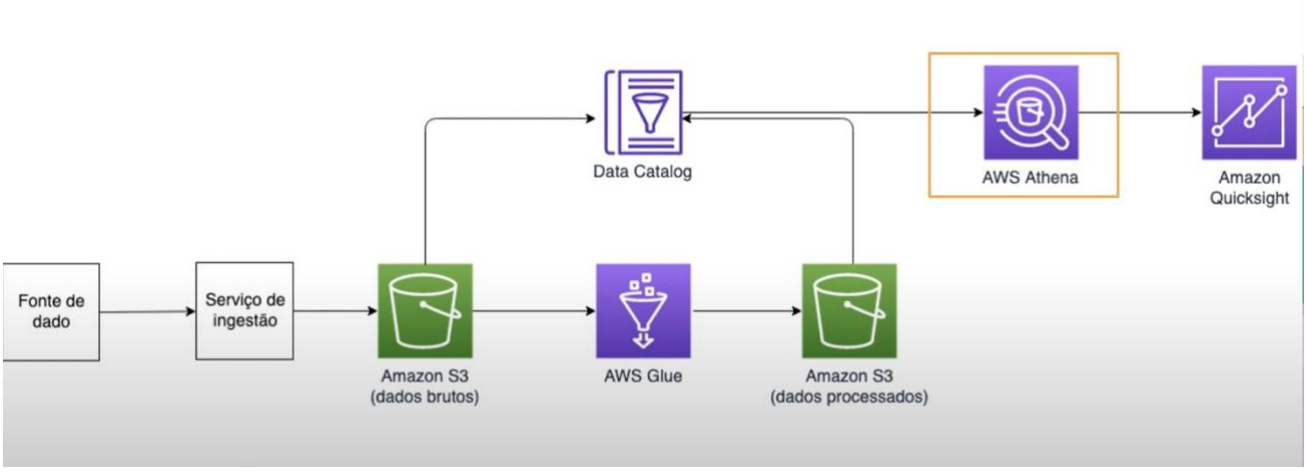
a) Os melhores exemplos são:
   1. Tutoriais do AWS Data Analytics
   2. Demos da série AWS LATAM transformando dados em insights

b) Um exemplo do Kaggle que apresenta uma **boa documentação** e **fundamentação do SQL**
https://www.kaggle.com/biancapedrosa/data-analysis-using-sql

c) AWS Blog tem bons exemplos
   1. Automate ETL-jobs betgen RDS for SQL Server anda Azure managed SQL using AWS glue Studio
   2. Enable self-service visual data integration and analysis for fund performance using AWS Glue Studio and Amazon QuickSight

d) Exemplos de trabalhos de ex-alunos, hospedados no GIT
   1. NBAstats
   2. FootbalDatabase

Exemplo de Diagrama de infraestrutura/Worflow de trabalho



No site visual paradigm tem *templates* para diagramas AWS