escola britânica de artes criativas & tecnologia

Analista de Dados

Base de dados & Linguagem SQL

Módulo | SQL: Base de dados & Linguagem SQL

Caderno de Aula

Professor André Perez

Tópicos

1. Introdução ao Google Colab;
2. Bases de dados relacionais;
3. Introdução ao SQL;
4. Introdução ao AWS Console;
5. Armazenamento de dados com AWS S3;
6. Computação em SQL com AWS Athena.

Aulas

1. Introdução ao Google Colab

> Ferramenta web autogerênciada de cadernos (*notebooks*).

1.1. Ferramenta web

- Crie uma conta Google em gmail.com;
- Acesse o Google Colab através do endereço colab.research.google.com.

1.2. Autogerênciada

- A Google provisiona uma máquina virtual para você;
- A máquina virtual dura no máximo 12h.

1.3. Cadernos (*notebooks*)

Um **caderno** é um documento *web* composto por um conjunto de elementos (células) de texto e código:

- Células de **texto** podem ser editados com o editor da ferramenta, HTML ou Markdown;
- ~Células de **código** são exclusivamente para a linguagem de programação Python~.

2. Bases de Dados Relacionais

2.1. Físico

Conceitos relacionados ao armazenamento físico dos dados.

- Sistema Gerenciador de Base de Dados (SGBD)

  > Software que gerencia o armazenamento físico de dados.

Alguns exemplos:

- MySQL;
- PostgreSQL;
- MariaDB;
- Oracle

2.2. Lógico

Conceitos relacionados a organização lógica dos dados.

- Tabela

  > Estrutura lógica e tabular de dados organizados em linhas e colunas.

- Base de Dado

  > Conjunto lógico de tabelas.

3. Introdução ao SQL

3.1. Definição

A linguagem de consulta estruturada, do inglês *structured query language* ou SQL, é uma linguagem de programação declarativa para interação com os dados armazenados nas tabelas de uma base de dados. Existe uma versão padrão do SQL registrada no ANSI, mas cada SGBD tem a sua própria versão.

Os códigos SQL são conhecidos como *queries* e são divididas em dois grandes grupos:

- **DDL**: Linguagem de definição de dados;
- **DML**: Linguagem de manipulação de dados.

Para exemplificar esses conceitos, vamos utilizar a tabela abaixo, com dados de clientes de uma instituição financeira.

id	idade	sexo	dependentes	escolaridade	tipo_cartao	limite_credito	valor_transa
768805383	45	M	3	ensino medio	blue	12.691,51	1.144,90
818770008	49	F	5	mestrado	blue	8.256,96	1.291,45
713982108	51	M	3	mestrado	blue	3.418,56	1.887,72

3.2. DDL

Linguagem de definição de dados, do inglês *Data Definition Language* ou DDL são instruções para criar/excluir/alterar tabelas e inserir/remover/atualizar dados.

Query 1: Criar/excluir a tabela de `clientes`.

```
CREATE TABLE clientes (
  id BIGINT,
  idade BIGINT,
  sexo STRING,
  dependentes BIGINT,
  escolaridade STRING,
  tipo_cartao STRING,
  limite_credito DOUBLE,
  valor_transacoes_12m DOUBLE,
  qtd_transacoes_12m BIGINT
);
```

Query 2: Excluir a tabela de `clientes`.

```
DROP TABLE clientes;
```

Query 3: Inserir os dados na tabela de `clientes`.

```
INSERT INTO clientes VALUES (768805383, 45, 'M', 3, 'ensino medio',
'blue', 12691.51, 1144.90, 42);
INSERT INTO clientes VALUES (818770008, 49, 'F', 5, 'mestrado',
'blue', 8256.96, 1291.45, 33);
INSERT INTO clientes VALUES (713982108, 51, 'M', 3, 'mestrado',
'blue', 3418.56, 1887.72, 20);
```

Query 4: Remover os dados das mulheres na tabela de `clientes`.

```
DELETE FROM clientes WHERE sexo = 'F';
```

3.3. DML

Linguagem de manipulação de dados, do inglês *Data Manipulation Language* ou DML são instruções para manipular (selecionar, filtrar, agregar, limitar, etc.) os dados armazenados em tabelas.

Query 1: Selecionar o id, a idade e o limite de crédito dos clientes homens da tabela de `clientes`, ordenados por idade de maneira decrescente.

```
SELECT id, idade, limite_credito FROM clientes WHERE sexo = 'M'
ORDER BY idade DESC;
```

id	idade	limite_credito
713982108	51	3418.56
768805383	45	12691.51

Query 2: Selecionar a média da idade dos clientes por sexo da tabela de `clientes`.

```
SELECT sexo, AVG(idade) AS "media_idade_por_sexo" FROM clientes
GROUP BY sexo;
```

sexo	media_idade_por_sexo
M	48
F	49

4. Introdução ao AWS Console

A Amazon Web Service (AWS) é uma plataforma de computação em nuvem. Ela oferece uma série de serviços de computação, armanezamento de dados, etc. Vamos utilizar dois serviços da plataforma para construir o nosso SGBD e aprender o SQL:

- AWS S3: Serviço de armazenamento de dados;
- AWS Athena: Serviços de computação em SQL.

> **Atenção:** Todo serviço utilizado tem um custo associado! Se a sua conta da AWS é nova, você é elegível a diversos serviços gratuitos (por tempo/quota de uso, etc.). Confira nesse link a oferta atual (elas podem mudar ao longo do tempo).

Para criar a sua conta na plataforma, basta acessar este link.

5. Armazenamento de dados com AWS S3

5.1. Definição

O AWS S3 é um serviço de armazenameto distribuido e sem servidor que atua como um repositório de dados. O serviço é inspirado no famoso projeto *open source* Apache Hadoop.

5.2. Funcionamento

1. Um *bucket* é uma partição lógica de dados, como uma pasta do seu computador;
2. Um objeto é um dado que você armazena dentro de um um *bucket*;
3. Outros serviços da plataforma podem acessar os dados armazenados.

5.3. Definição de Preço

O AWS S3 cobra por volume de dados armazenado. O preço atual é complexo, mas inicia-se com 0,0405 USD por *gigabyte* (GB) armazenado (0,21 BRL aproximadamente). Você deve sempre consultar o preço na página web do serviço (link).

6. Computação em SQL com o AWS Athena

6.1. Definição

O AWS Athena é um serviço de computação distribuida e sem servidor que atua como um motor de consulta (*query engine*). Por trás dos panos, ele implementa na infrastrutura computacional da AWS um projeto *open source* chamado Presto.

6.2. Funcionamento

1. Transforma uma *query* em código;
2. Processa os arquivos armazenados no AWS S3 com o código gerado;
3. Retorna os resultados no console.

6.3. Definição de Preço

O AWS Athena cobra por consulta. O preço atual é de 9,00 USD por *terabyte* escaneado (47,46 BRL aproximadamente). Você deve sempre consultar o preço na página web do serviço (link).