INSTITUTO SUPERIOR DE TECNOLOGIAS AVANÇADAS

Ciência de Dados

Licenciatura Engenharia Informática

2º Semestre – 2021/2022

Ricardo Jesus Ferreira

ricardojesus.ferreira@my.istec.pt

Trabalhar com Dados

- Técnicas de Manipulação de Dados
- Métodos de Otimização

Manipulação de Dados - *Filtering*

- Processo de seleção de um *subset* de dados para posterior análise
- É extraído uma parte dos dados mantendo-se a restante informação armazenada

```
SELECT * FROM
     table_name
WHERE
     column_one = value_one
OR
     column_two > value_two
AND
     column_six < value_six
```

Manipulação de Dados - *Sorting*

- Processo de ordenação de *query*
- A ordenação é importante para a visualização e análise

```
SELECT * FROM
     table_name
WHERE
     condition
ORDER BY
     column ASC|DESC
```

Manipulação de Dados - Datas

Função	SQL	Python
Data/Hora atual	NOW(), CURDATE(), CURTIME()	now(), today()
Dia, mês, ano	EXTRACT()	datetime.days(), datetime.month(), datetime.year(),
Expressão Data	DATE()	date(YYYY-MM-DD)
Data em diferentes formatos	DATE_FORMAT()	strptime()
Adição de determinado intervalo	DATE_ADD()	datetime.now()+timedelta()
Subtração de determinado intervalo	DATE_SUB()	now()-timedelta()
Diferença entre datas	DATEDIFF()	(datetime.strptime(a) - datetime.strptime(b)).days()

Manipulação de Dados - Lógica

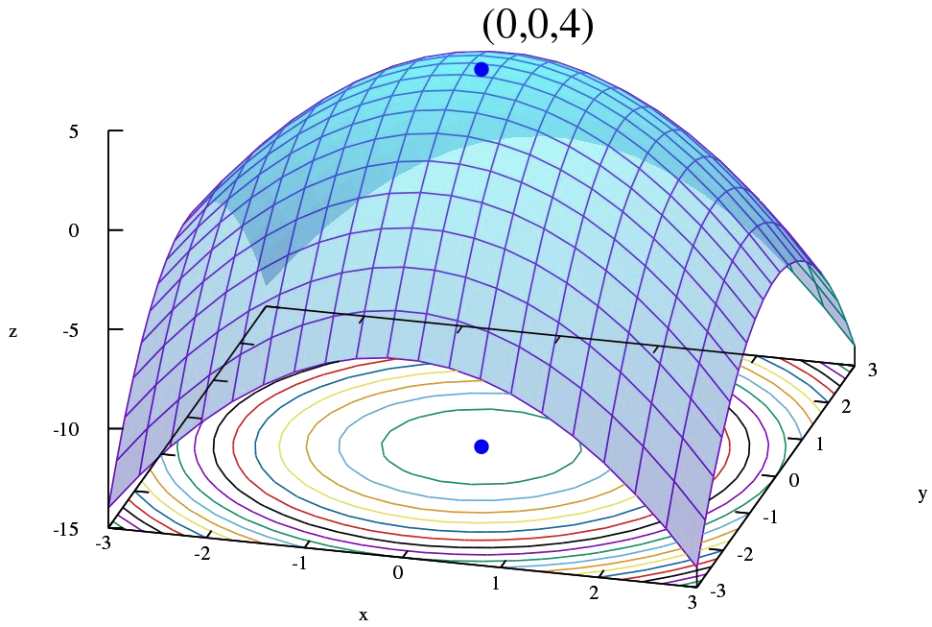
Função	SQL	Python
AND	A AND B	A and B
OR	A OR B	A or B
NOT	NOT A, NOT B	!A, not(B)
IN	IN A, IN B	in A, in B
EXISTS	A EXISTS, B EXISTS	A exists, B exists

Manipulação de Dados – Agregação

Função	SQL	Python
COUNT	COUNT()	count()
SUM	SUM()	sum()
MIN	MIN()	min()
MAX	MAX()	max()
MEAN	AVG()	sum()/len()

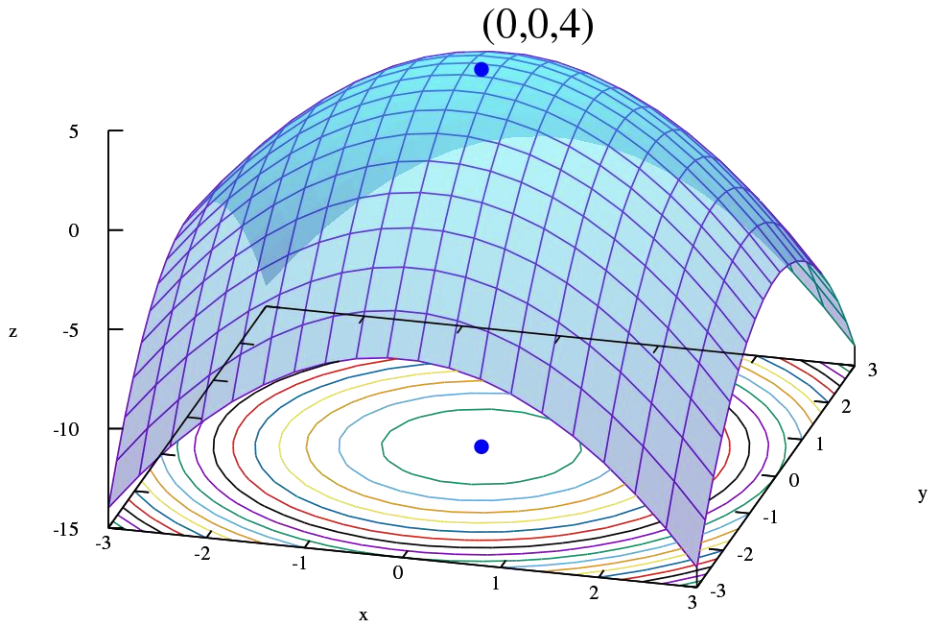
Métodos de Otimização

- *Query Optimization* – processo para melhorar a arquitetura e implementação de uma *query*
- Consome tempo na elaboração, mas depois poupa tempo na execução
- É necessário ter uma visão holística da BD



Parametrização

- Processo que substitui múltiplas *queries* com uma variável dinâmica
- O procedimento irá ficar armazenado em memoria, sendo posteriormente utilizado
- Ajuda na gestão da memoria do servidor



Parametrização

```
SELECT id
FROM customer
WHERE country=“Portugal”


SELECT id
FROM customer
WHERE country=“USA”
```

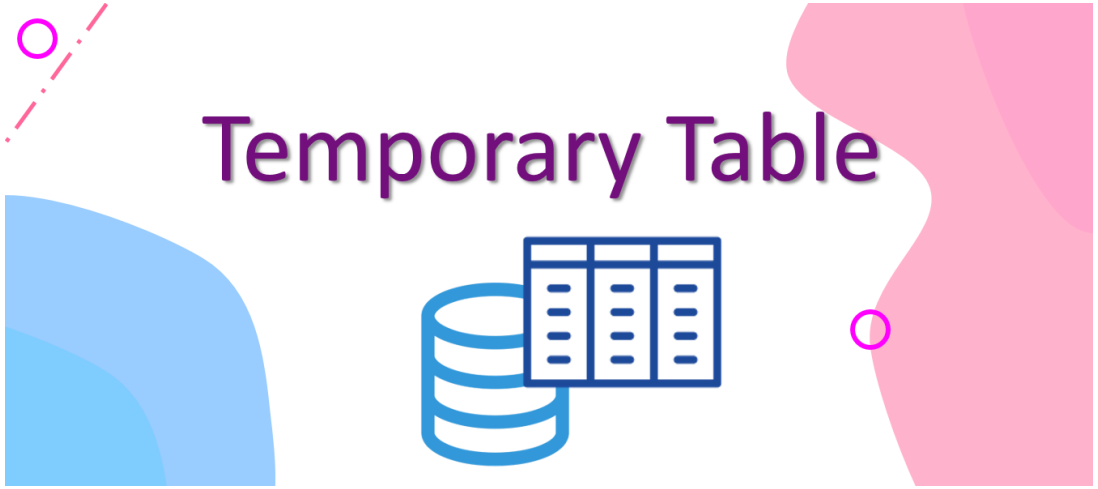
```
CREATE  PROCEDURE  get_country
      @country STRING
AS
BEGIN
      SELECT id
      FROM customer
      WHERE country=@country
END
```

Indexing

- Os indexes são utilizados para aumentar a velocidade de uma *query*

- Podem ser utilizados no SELECT, INSERT, DELETE

Tabelas temporárias

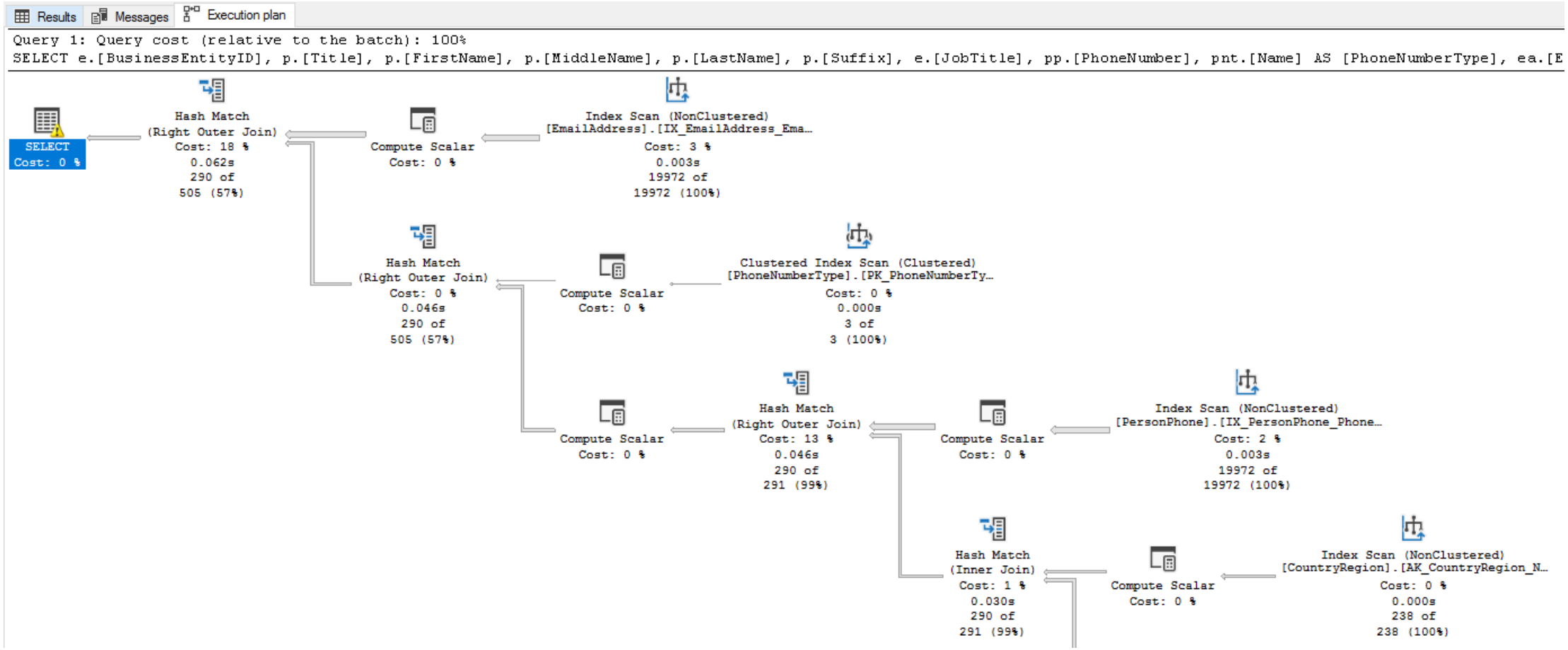


- Usadas para armazenar dados por tempo limitado
- Podem ser feitas todas as operações que se fazem numa tabela permanente
- Eliminadas quando terminamos a ligação
- Uteis para transformação de grandes *datasets*

Plano de Execução

- É um plano detalhado de como devem ser realizadas as instruções SQL
- Permite que os administradores e programadores usem o sistema com uma maior eficácia
- Possui informações das tabelas, de que forma são acedidas e quais os indexes a utilizar
- As etapas incluídas no plano de execução estão ordenadas por níveis. O cumprimento desta hierarquia irá resultar numa resposta muito mais optimizada

Plano de Execução



Bibliografia

- B. Gomez,(2020) “Resolviendo problemas de Big Data”, Alfaomega.
- D. Insua, (2019)“Big data: Conceptos, tecnologías y aplicaciones”, CSIC.
- H. Jones, (2019)“Analítica de datos”, HJ,.
- J. Somed, (2020)“Big Data Analytics”, JLC.
- D. Petković (2020)“Microsoft® SQL Server® 2019 A Beginner’s Guide - Seventh Edition”, McGraw Hill.