

Prova Técnica – Spark Itaú

== Teórico ==

1. O que é um RDD?
2. Qual a diferença entre DataFrame e RDD?
3. Quando usar DataFrame e quando usar RDD?

== Prático ==

4. Crie 2 DataFrames e Salve-os como CSV, salvando com o delimitador ";" (Dica: default é sempre ",", vírgula), sobrescrevendo caso já haja algum arquivo:

DataFrame1:

Pessoa

Campos (Nome, Idade, Profissao)

Dados:

Marta, 33, 1
João Vitor, 24, 2
Pedro, -1, null
Marta, 24, 4
Alberto, 30, 8

DataFrame2

Profissao

Campos (Id, Descricao, Status)

Dados:

1, Médico, true
2, Engenheiro, true
3, Desembargador, false
4, Informático, false
5, Enfermeiro(a), false
6, Policial, true
7, CEO, true
8, Piloto, true
9, Gerente de Vendas, true
10, Atendente, true

Path de destino:

/FileStore/tables/CSV/Pessoa/

/FileStore/tables/CSV/Profissao/

5. Leia os dois arquivos CSV anteriores e salve-os em PARQUET:

6. Realize um Join e grave um em novo arquivo Parquet consolidado somente as pessoas que TEM profissão correspondente:

Dados:

Pessoa_Profissao

- Campos:

Nome, Idade, DescricaoProfissao, StatusProfissao

- Path Destino:

/FileStore/tables/CSV/PessoaProfissaoConsolidado/

- Dica:

* Exiba os dados

```
import org.apache.spark.sql.functions.{_}
```