Atividade Spark

- iniciar o cluster

-> docker-compose up –d

Habilitar a porta para ter acesso o jupyter notebook

1 – Criar dataframes das tabelas que estão no hive

2 – Criar um dataframe (df_pedidos) esta dataframe deve ser a união de pedido e item_pedido;

1. Ver a quantidade de pedidos?
2. Quantidade de produtos e agrupa-los por pedido
3. Quantidade de pedidos por cliente
4. Quantidade de pedidos por parceiro
5. Quantidade de pedido por filial

3 – Juntar criar df_filial que deverá ser a junção das tabelas filial, cidade e estado;

1. Juntar com o df_pedidos
2. Ver a quantidade de pedidos por estado
3. Top 10 filial que mais vendeu

4 – Criar o dataframe df_stage juntando todas as bases do nosso modelo relacional