Teste de evidência Big Data

1: Para começar os códigos no console, iniciamos com a criação da pasta principal chamada axistech utilizando o hdfs.

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir axistech
```

2: Utilizamos o wget para baixar o link do github que está o dataset.

```
[cloudera@quickstart ~]$ wget https://raw.githubusercontent.com/Gianinao/challen
ge-csv/refs/heads/main/supermarket_sales.csv
```

3: Agora verificáramos se tanto a pasta quanto arquivo csv foi para o hdfs.

```
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - cloudera supergroup          0 2025-04-15 17:54 axistech
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/supermarket_sales.csv /user/cloudera/axistech/
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/axistech
Found 1 items
-rw-r--r--   1 cloudera supergroup     114780 2025-04-15 17:59 /user/cloudera/axistech/supermarket_sales.csv
```

4: Aqui vamos utilizar o put para puxar o arquivo csv pro hdfs no destino da pasta criada que é o axistech.

```
[cloudera@quickstart ~]$ hdfs dfs -put /home/cloudera/supermarket_sales.csv /user/cloudera/axistech/
```

5: Agora nós vamos fazer um arquivo. pig para colocar todos os códigos que iram fazer o processo de ETL do nosso dataset, sejam eles os clientes, os produtos e as vendas.

```
  GNU nano 2.0.9                    File: ETL.pig                         Modified

raw_data = LOAD '/user/axistech/supermarket_sales.csv'
  USING PigStorage(';')
  AS (f1:chararray, f2:chararray, f3:chararray, f4:chararray, f5:chararray,
      f6:chararray, f7:chararray, f8:chararray, f9:chararray, f10:chararray,
      f11:chararray, f12:chararray, f13:chararray, f14:chararray, f15:chararray,
      f16:chararray, f17:chararray);

dados = FILTER raw_data BY f1 != 'Invoice ID';

clientes = FOREACH dados GENERATE f1 AS invoice_id, f4 AS customer_type, f5 AS gender;
STORE clientes INTO '/user/axistech/clientes' USING PigStorage(';');

produtos = FOREACH dados GENERATE f1 AS invoice_id, f6 AS product_line, (float)f7 AS unit_price, (int)f8 AS quantity;
STORE produtos INTO '/user/axistech/produtos' USING PigStorage(';');

vendas = FOREACH dados GENERATE f1 AS invoice_id, (float)f10 AS total, f11 AS date, f13 AS payment;
STORE vendas INTO '/user/axistech/vendas' USING PigStorage(';');
```

6: Feito esse código nós vamos rodar o arquivo ETL.pig onde está todos eles e mostramos que foi executado com sucesso.

```
[cloudera@quickstart ~]$ pig -x mapreduce ETL.pig
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.13.0 0.12.0-cdh5.13.0        cloudera        2025-04-15 19:08:33     2025-04-15 19:08:53     FILTER

Success!
```

7: Após criado estes datasets tratados pelos códigos nossos realocamos as análises feitas pra dentro dele.

```
cloudera@quickstart ~]$ hdfs dfs -get /user/axistech/clientes RM560454_clientes.csv
cloudera@quickstart ~]$ hdfs dfs -get /user/axistech/produtos RM560454_produtos.csv
cloudera@quickstart ~]$ hdfs dfs -get /user/axistech/vendas RM560454_vendas.csv
```

8: Fizemos o head em cada um dos arquivos tratados:

```
[cloudera@quickstart ~]$ head RM560454_clientes.csv/part-m-00000
750-67-8428;Member;Female
226-31-3081;Normal;Female
631-41-3108;Normal;Male
123-19-1176;Member;Male
873-73-7910;Normal;Male
699-14-3026;Normal;Male
355-53-5943;Member;Female
815-22-5665;Normal;Female
665-32-9167;Member;Female
692-92-5582;Member;Female
[cloudera@quickstart ~]$ head RM560454_produtos.csv/part-m-00000
750-67-8428;Health and beauty;74.69;7
226-31-3081;Electronic accessories;15.28;5
631-41-3108;Home and lifestyle;46.33;7
123-19-1176;Health and beauty;58.22;8
873-73-7910;Sports and travel;86.31;7
699-14-3026;Electronic accessories;85.39;7
355-53-5943;Electronic accessories;68.84;6
815-22-5665;Home and lifestyle;73.56;10
665-32-9167;Health and beauty;36.26;2
692-92-5582;Food and beverages;54.84;3
[cloudera@quickstart ~]$ head RM560454_vendas.csv/part-m-00000
750-67-8428;;01/05/2019;Ewallet
226-31-3081;80.22;03/08/2019;Cash
631-41-3108;;03/03/2019;Credit card
123-19-1176;489.048;1/27/2019;Ewallet
873-73-7910;;02/08/2019;Ewallet
699-14-3026;;3/25/2019;Ewallet
355-53-5943;433.692;2/25/2019;Ewallet
815-22-5665;772.38;2/24/2019;Ewallet
665-32-9167;76.146;01/10/2019;Credit card
692-92-5582;172.746;2/20/2019;Credit card
```

9: Agora vamos conectar no mysql usando o banco de dados criado do grupo:

```
mysql> CREATE DATABASE AxisTech;
Query OK, 1 row affected (0.00 sec)

mysql> USE AxisTech;
Database changed
mysql>
```

10: Inserir as tabelas pro database AxisTech:

```
mysql> USE AxisTech;
Database changed
mysql> CREATE TABLE RM560454_T_CLIENTES (
    ->     INVOICE_ID VARCHAR(20) PRIMARY KEY,
    ->     CUSTOMER_TYPE VARCHAR(20),
    ->     GENDER VARCHAR(10)
    -> );
Query OK, 0 rows affected (0.02 sec)

mysql> CREATE TABLE RM560454_T_PRODUTOS (
    ->     INVOICE_ID VARCHAR(20),
    ->     PRODUCT_LINE VARCHAR(50),
    ->     UNIT_PRICE DECIMAL(10,2),
    ->     QUANTITY INT,
    ->     FOREIGN KEY (INVOICE_ID) REFERENCES RM560454_T_CLIENTES(INVOICE_ID)
    -> );
Query OK, 0 rows affected (0.02 sec)

mysql> CREATE TABLE RM560454_T_VENDAS (
    ->     INVOICE_ID VARCHAR(20),
    ->     TOTAL DECIMAL(10,3),
    ->     DATA DATE,
    ->     PAYMENT VARCHAR(20),
    ->     FOREIGN KEY (INVOICE_ID) REFERENCES RM560454_T_CLIENTES(INVOICE_ID)
    -> );
Query OK, 0 rows affected (0.01 sec)
```

11: Utilizamos o sqoop export para exportar a tabela T_CLIENTES:

```
[cloudera@quickstart ~]$ sqoop export \
> --connect jdbc:mysql://localhost/AxisTech \
> --username root \
> --password cloudera \
> --table RM560454_T_CLIENTES \
> --export-dir /user/axistech/clientes \
> --input-fields-terminated-by ';' \
> --columns INVOICE_ID,CUSTOMER_TYPE,GENDER
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
25/04/15 20:28:21 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
25/04/15 20:28:21 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
25/04/15 20:28:21 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
25/04/15 20:28:21 INFO tool.CodeGenTool: Beginning code generation
25/04/15 20:28:21 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `RM560454_T_CLIENTES` AS t LIMIT 1
25/04/15 20:28:21 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `RM560454_T_CLIENTES` AS t LIMIT 1
25/04/15 20:28:21 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/6e676732c68640af6db7a3f8d611d11d/RM560454_T_CLIENTES.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
25/04/15 20:28:22 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/6e676732c68640af6db7a3f8d611d11d/RM560454_T_CLIENTES.jar
25/04/15 20:28:22 INFO mapreduce.ExportJobBase: Beginning export of RM560454_T_CLIENTES
25/04/15 20:28:22 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
25/04/15 20:28:23 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
25/04/15 20:28:23 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
25/04/15 20:28:23 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
25/04/15 20:28:23 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
25/04/15 20:28:23 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/04/15 20:28:26 INFO input.FileInputFormat: Total input paths to process : 1
25/04/15 20:28:26 INFO input.FileInputFormat: Total input paths to process : 1
25/04/15 20:28:26 INFO mapreduce.JobSubmitter: number of splits:4
25/04/15 20:28:26 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
25/04/15 20:28:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744762569685_0008
25/04/15 20:28:27 INFO impl.YarnClientImpl: Submitted application application_1744762569685_0008
25/04/15 20:28:27 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1744762569685_0008/
25/04/15 20:28:27 INFO mapreduce.Job: Running job: job_1744762569685_0008
25/04/15 20:28:33 INFO mapreduce.Job: Job job_1744762569685_0008 running in uber mode : false
25/04/15 20:28:33 INFO mapreduce.Job:  map 0% reduce 0%
```

```
                Total megabyte-milliseconds taken by all map tasks=19147776
        Map-Reduce Framework
                Map input records=1000
                Map output records=1000
                Input split bytes=676
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=411
                CPU time spent (ms)=2730
                Physical memory (bytes) snapshot=838701056
                Virtual memory (bytes) snapshot=6323429376
                Total committed heap usage (bytes)=956301312
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
25/04/15 20:28:44 INFO mapreduce.ExportJobBase: Transferred 40.1436 KB in 20.821 seconds (1.928 KB/sec)
25/04/15 20:28:44 INFO mapreduce.ExportJobBase: Exported 1000 records.
[cloudera@quickstart ~]$
```

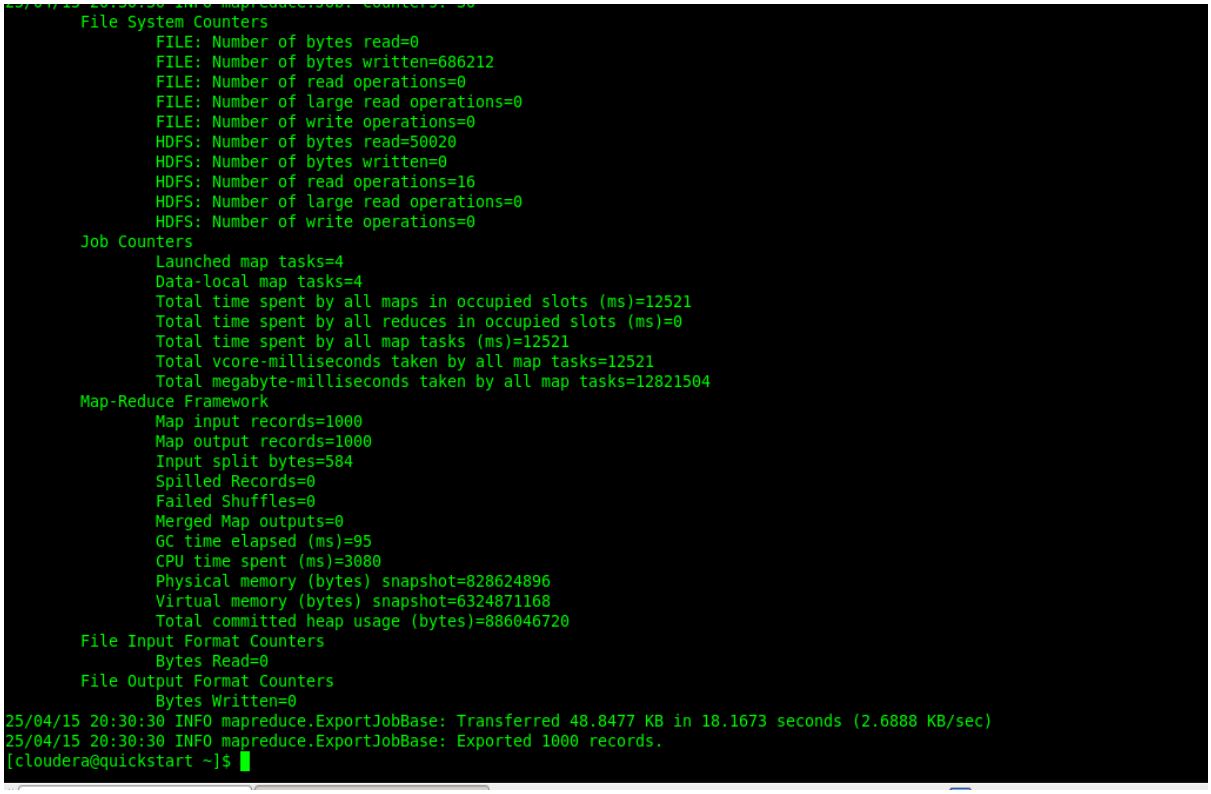
12: Utilizamos o sqoop export para exportar a tabela T_PRODUTOS:

```
[cloudera@quickstart ~]$ sqoop export \
> --connect jdbc:mysql://localhost/AxisTech \
> --username root \
> --password cloudera \
> --table RM560454_T_PRODUTOS \
> --export-dir /user/axistech/produtos \
> --input-fields-terminated-by ';' \
> --columns INVOICE_ID,PRODUCT_LINE,UNIT_PRICE,QUANTITY
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
25/04/15 20:29:22 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
25/04/15 20:29:22 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
25/04/15 20:29:22 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
25/04/15 20:29:22 INFO tool.CodeGenTool: Beginning code generation
25/04/15 20:29:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `RM560454_T_PRODUTOS` AS t LIMIT 1
25/04/15 20:29:23 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `RM560454_T_PRODUTOS` AS t LIMIT 1
25/04/15 20:29:23 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/24fc5f5fa6996f51eb82c0f083255a89/RM560454_T_PRODUTOS.java uses or overrides a deprecated API.
```

```
                Total megabyte-milliseconds taken by all map tasks=16490496
        Map-Reduce Framework
                Map input records=1000
                Map output records=1000
                Input split bytes=676
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=343
                CPU time spent (ms)=2840
                Physical memory (bytes) snapshot=832204800
                Virtual memory (bytes) snapshot=6285799424
                Total committed heap usage (bytes)=888143872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
25/04/15 20:29:42 INFO mapreduce.ExportJobBase: Transferred 55.3096 KB in 17.8754 seconds (3.0942 KB/sec)
25/04/15 20:29:42 INFO mapreduce.ExportJobBase: Exported 1000 records.
[cloudera@quickstart ~]$
```

13: Utilizamos o sqoop export para exportar a tabela T_VENDAS:

```
[cloudera@quickstart ~]$ sqoop export \
> --connect jdbc:mysql://localhost/AxisTech \
> --username root \
> --password cloudera \
> --table RM560454_T_VENDAS \
> --export-dir /user/axistech/vendas \
> --input-fields-terminated-by ';' \
> --columns INVOICE_ID,TOTAL,DATA,PAYMENT
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
25/04/15 20:30:10 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
25/04/15 20:30:10 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
25/04/15 20:30:10 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
25/04/15 20:30:10 INFO tool.CodeGenTool: Beginning code generation
25/04/15 20:30:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `RM560454_T_VENDAS` AS t LIMIT 1
25/04/15 20:30:10 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `RM560454_T_VENDAS` AS t LIMIT 1
25/04/15 20:30:10 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/029a6dbf9033ec642813da47b4341315/RM560454_T_VENDAS.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
25/04/15 20:30:11 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/029a6dbf9033ec642813da47b4341315/RM560454_T_VENDAS.jar
25/04/15 20:30:11 INFO mapreduce.ExportJobBase: Beginning export of RM560454_T_VENDAS
25/04/15 20:30:11 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
25/04/15 20:30:12 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
25/04/15 20:30:12 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
25/04/15 20:30:12 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
25/04/15 20:30:12 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
25/04/15 20:30:12 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
25/04/15 20:30:15 INFO input.FileInputFormat: Total input paths to process : 1
25/04/15 20:30:15 INFO input.FileInputFormat: Total input paths to process : 1
25/04/15 20:30:15 INFO mapreduce.JobSubmitter: number of splits:4
25/04/15 20:30:15 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
25/04/15 20:30:15 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1744762569685_0010
25/04/15 20:30:16 INFO impl.YarnClientImpl: Submitted application application_1744762569685_0010
25/04/15 20:30:16 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1744762569685_0010/
25/04/15 20:30:16 INFO mapreduce.Job: Running job: job_1744762569685_0010
25/04/15 20:30:21 INFO mapreduce.Job: Job job_1744762569685_0010 running in uber mode : false
25/04/15 20:30:21 INFO mapreduce.Job:  map 0% reduce 0%
```



14: Consultar a tabela T_CLIENTES:

```
mysql> SELECT * FROM RM560454_T_CLIENTES LIMIT 5;
+-------------+---------------+--------+
| INVOICE_ID  | CUSTOMER_TYPE | GENDER |
+-------------+---------------+--------+
| 585-86-8361 | Normal        | Female |
| 807-14-7833 | Member        | Female |
| 775-72-1988 | Normal        | Male   |
| 288-38-3758 | Member        | Female |
| 652-43-6591 | Normal        | Female |
+-------------+---------------+--------+
5 rows in set (0.00 sec)
```

15: Consultar a tabela T_PRODUTOS:

```
mysql> SELECT * FROM RM560454_T_PRODUTOS LIMIT 5;
+-------------+------------------------+------------+----------+
| INVOICE_ID  | PRODUCT_LINE           | UNIT_PRICE | QUANTITY |
+-------------+------------------------+------------+----------+
| 585-86-8361 | Food and beverages     |      27.28 |        5 |
| 807-14-7833 | Electronic accessories |      17.42 |       10 |
| 775-72-1988 | Home and lifestyle     |      73.28 |        5 |
| 288-38-3758 | Fashion accessories    |      84.87 |        3 |
| 652-43-6591 | Fashion accessories    |      97.29 |        8 |
+-------------+------------------------+------------+----------+
5 rows in set (0.00 sec)
```

16: Consultar a tabela T_VENDAS:

```
mysql> SELECT * FROM RM560454_T_VENDAS LIMIT 5;
+-------------+---------+------------+-------------+
| INVOICE_ID  | TOTAL   | DATA       | PAYMENT     |
+-------------+---------+------------+-------------+
| 750-67-8428 |    NULL | 01/05/2019 | Ewallet     |
| 226-31-3081 |   80.22 | 03/08/2019 | Cash        |
| 631-41-3108 |    NULL | 03/03/2019 | Credit card |
| 123-19-1176 | 489.048 | 1/27/2019  | Ewallet     |
| 373-73-7910 |    NULL | 02/08/2019 | Ewallet     |
+-------------+---------+------------+-------------+
5 rows in set (0.00 sec)
```