**FIAP - FACULDADE DE INFORMÁTICA E ADMINISTRAÇÃO PAULISTA**
**DATA SCIENCE**

Fábio Pereira de Lima – RM98803
Giovanna Cardoso Satorres – RM99944
Giullia Bianca Rocha Souza – RM552108
Gustavo Semenuk – RM550472
Rafael Luiz Custódio Guimarães - RM98304

# SPRINT 3: PIG PARA O CHALLENGE MINSAIT

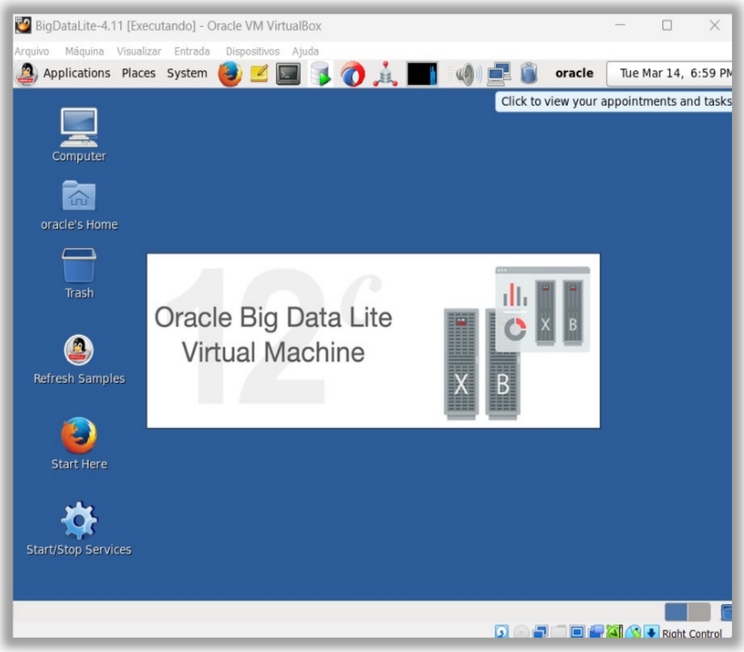
BIG DATA ARCHITECTURE & DATA INTEGRATION

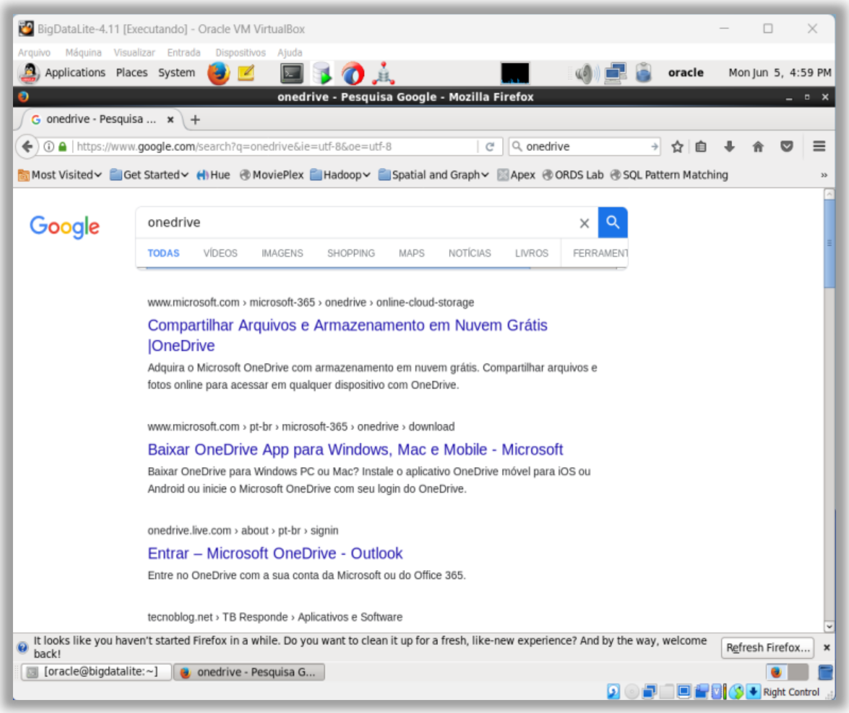
São Paulo
2023

# ★ Copiando o arquivo CSV para a máquina virtual Big Data

- Primeiramente, fizemos o upload do arquivo CSV para o OneDrive, e após isso, iniciamos a máquina virtual.

- Na VM, abrimos o navegador Mozilla Firefox e acessamos o OneDrive para realizar o upload do arquivo CSV. O arquivo está no diretório Downloads.


BigDataLite-4.11 [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
Applications Places System
oracle
Tue Mar 14, 6:59 PM
Click to view your appointments and tasks
Computer
oracle's Home
Trash
Refresh Samples
Start Here
Start/Stop Services
Oracle Big Data Lite
Virtual Machine
Right Control


BigDataLite-4.11 [Executando] - Oracle VM VirtualBox
Arquivo Máquina Visualizar Entrada Dispositivos Ajuda
Applications Places System
oracle
Mon Jun 5, 4:59 PM
onedrive - Pesquisa Google - Mozilla Firefox
onedrive - Pesquisa ...
https://www.google.com/search?q=onedrive&ie=utf-8&oe=utf-8
onedrive
Most Visited Get Started Hue MoviePlex Hadoop Spatial and Graph Apex ORDS Lab SQL Pattern Matching
Google
onedrive
TODAS VÍDEOS IMAGENS SHOPPING MAPS NOTÍCIAS LIVROS FERRAMENT
Refresh Firefox...
[oracle@bigdatalite:~]
onedrive - Pesquisa G...
Right Control

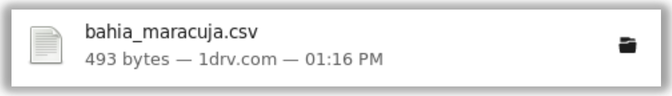
www.microsoft.com › microsoft-365 › onedrive › online-cloud-storage
Compartilhar Arquivos e Armazenamento em Nuvem Grátis |OneDrive
Adquira o Microsoft OneDrive com armazenamento em nuvem grátis. Compartilhar arquivos e fotos online para acessar em qualquer dispositivo com OneDrive.

www.microsoft.com › pt-br › microsoft-365 › onedrive › download
Baixar OneDrive App para Windows, Mac e Mobile - Microsoft
Baixar OneDrive para Windows PC ou Mac? Instale o aplicativo OneDrive móvel para iOS ou Android ou inicie o Microsoft OneDrive com seu login do OneDrive.

onedrive.live.com › about › pt-br › signin
Entrar – Microsoft OneDrive - Outlook
Entre no OneDrive com a sua conta da Microsoft ou do Office 365.

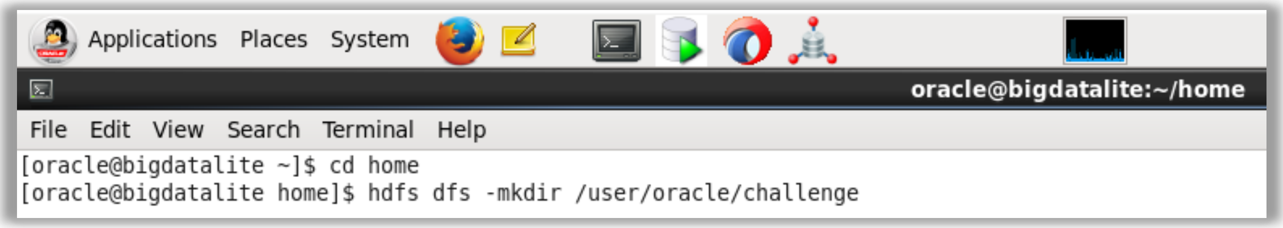
tecnoblog.net › TB Responde › Aplicativos e Software

It looks like you haven't started Firefox in a while. Do you want to clean it up for a fresh, like-new experience? And by the way, welcome back!


bahia_maracuja.csv
493 bytes — 1drv.com — 01:16 PM

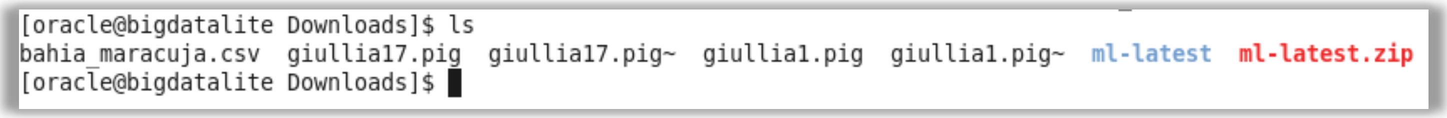
## ★ Copiando o arquivo de dados para o Hadoop

- Primeiramente, foi necessário criar um diretório no Hadoop. Para isso, abrimos o terminal da VM. Então, navegamos até o diretório ‘home’ e utilizamos o comando “**hdfs dfs -mkdir /user/oracle/challenge**”.



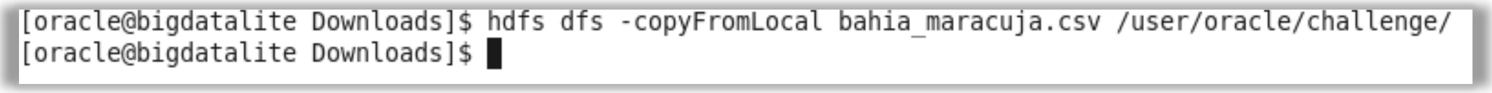
```
[oracle@bigdatalite ~]$ cd home
[oracle@bigdatalite home]$ hdfs dfs -mkdir /user/oracle/challenge
```

- Para verificar se o arquivo CSV estava mesmo no diretório ‘Downloads’, navegamos até esse diretório e executamos o comando “**ls**”.



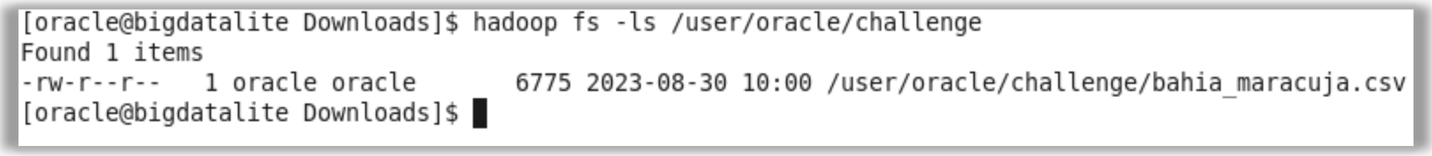
```
[oracle@bigdatalite Downloads]$ ls
bahia_maracuja.csv  giullia17.pig  giullia17.pig~  giullia1.pig  giullia1.pig~  ml-latest  ml-latest.zip
[oracle@bigdatalite Downloads]$ 
```

- O próximo passo foi copiar o arquivo CSV do ambiente local para o diretório recém-criado no Hadoop. Para isso, utilizamos o comando “**hdfs dfs -copyFromLocal bahia_maracuja.csv /user/oracle/challenge/**”.



```
[oracle@bigdatalite Downloads]$ hdfs dfs -copyFromLocal bahia_maracuja.csv /user/oracle/challenge/
[oracle@bigdatalite Downloads]$ 
```

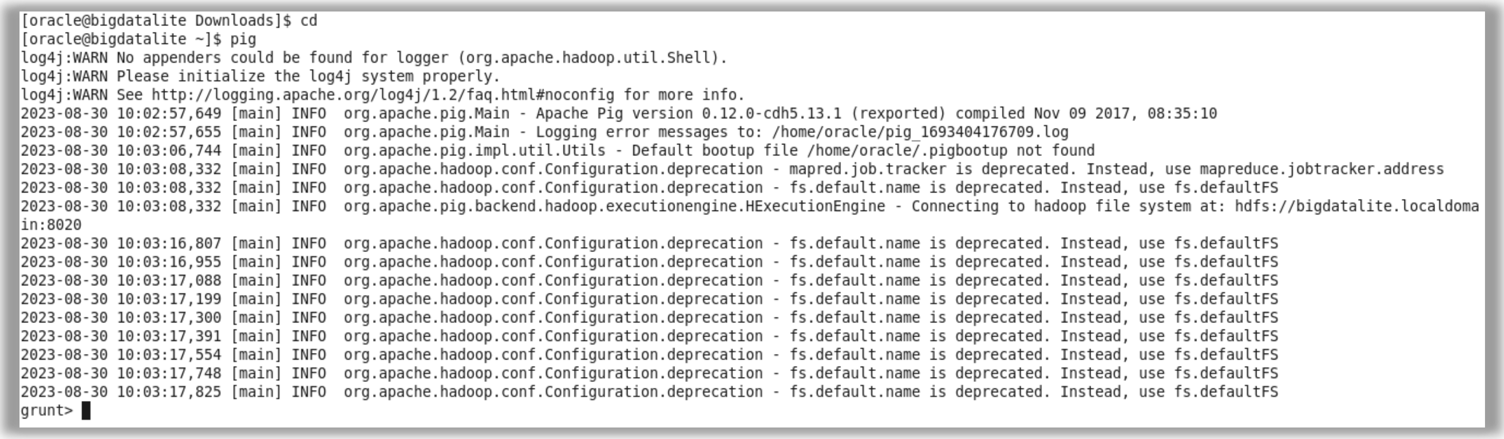
- Para garantir que o processo foi bem sucedido, executamos o comando “**hadoop fs -ls /user/oracle/challenge”, para listar o conteúdo do diretório Hadoop “Challenge**”.



```
[oracle@bigdatalite Downloads]$ hadoop fs -ls /user/oracle/challenge
Found 1 items
-rw-r--r--   1 oracle oracle       6775 2023-08-30 10:00 /user/oracle/challenge/bahia_maracuja.csv
[oracle@bigdatalite Downloads]$ 
```

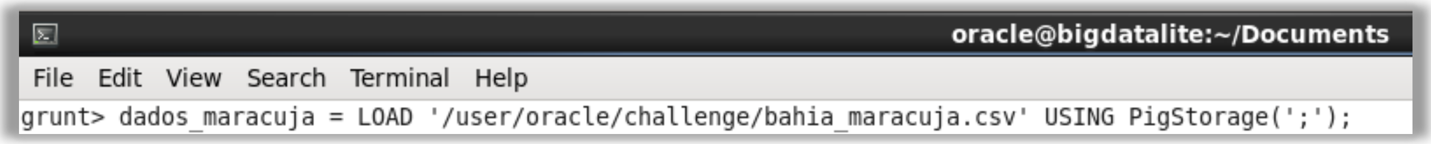
## ★ Carregando os dados do Hadoop para o PIG

- Para iniciar o PIG, executamos o comando “**pig**”.



```
[oracle@bigdatalite Downloads]$ cd
[oracle@bigdatalite ~]$ pig
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
2023-08-30 10:02:57,649 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.1 (rexported) compiled Nov 09 2017, 08:35:10
2023-08-30 10:02:57,655 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/oracle/pig_1693404176709.log
2023-08-30 10:03:06,744 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/oracle/.pigbootup not found
2023-08-30 10:03:08,332 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2023-08-30 10:03:08,332 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:08,332 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://bigdatalite.localdoma
in:8020
2023-08-30 10:03:16,807 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:16,955 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,088 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,199 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,300 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,391 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,554 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,748 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 10:03:17,825 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> █
```

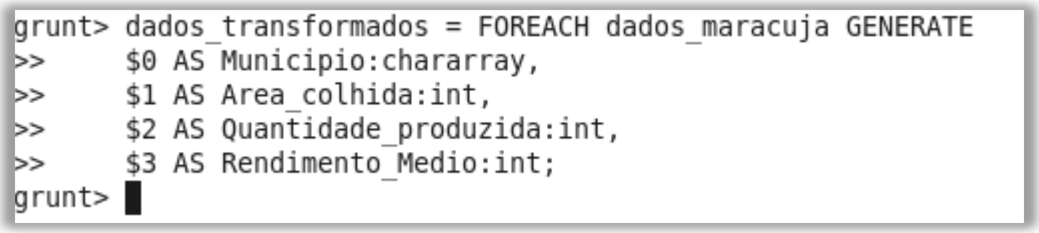
- Então, executamos o comando: “**dados_maracuja = LOAD '/user/oracle/challenge/bahia_maracuja.csv' USING PigStorage(';');**” para carregar os dados do arquivo CSV para o PIG.



```
grunt> dados_maracuja = LOAD '/user/oracle/challenge/bahia_maracuja.csv' USING PigStorage(';');
```

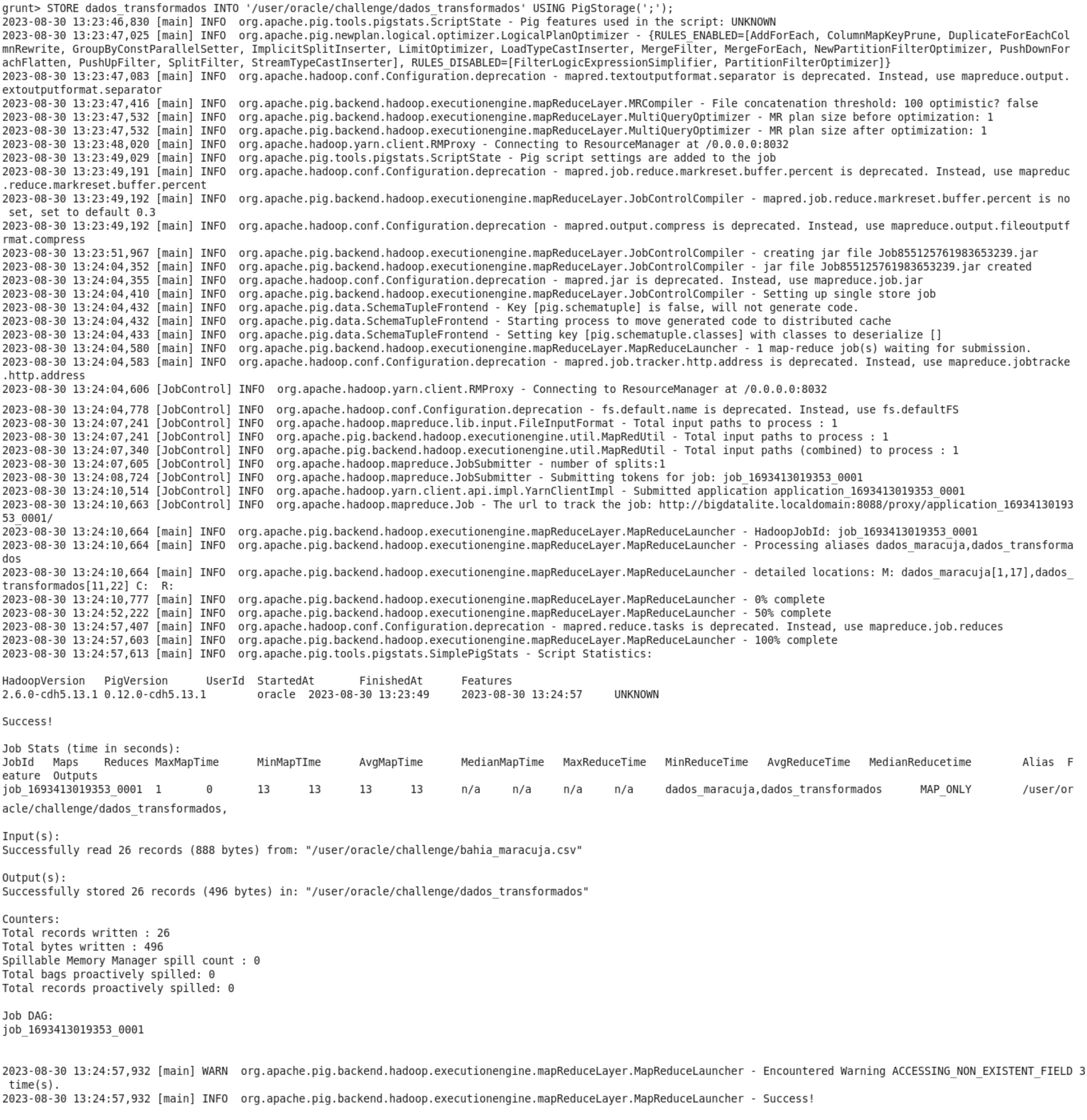
## ★ Transformando os dados utilizando PIG

- Uma vez que os dados foram carregados para o PIG, transformamos os dados de maneira a serem carregados no HIVE, posteriormente. As transformações foram realizadas conforme o comando “**dados_transformados = FOREACH dados_maracuja GENERATE $0 AS Municipio:chararray, $1 AS Area_colhida:int, $2 AS Quantidade_produzida:int, $3 AS Rendimento_Medio:int;**”.



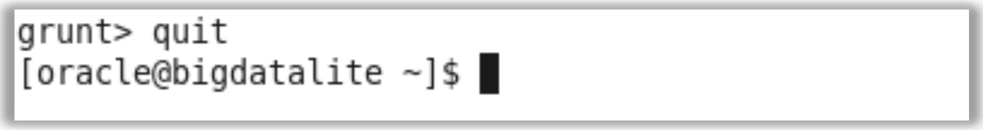
```
grunt> dados_transformados = FOREACH dados_maracuja GENERATE
>>     $0 AS Municipio:chararray,
>>     $1 AS Area_colhida:int,
>>     $2 AS Quantidade_produzida:int,
>>     $3 AS Rendimento_Medio:int;
grunt> █
```

- Após isso, executando o comando: "**STORE dados_transformados INTO '/user/oracle/challenge/dados_transformados' USING PigStorage(';');" armazenamos o resultado no diretório Hadoop "challenge**".



```
grunt> STORE dados_transformados INTO '/user/oracle/challenge/dados_transformados' USING PigStorage(';');
2023-08-30 13:23:46,830 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2023-08-30 13:23:47,025 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachCol
mnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDownFor
achFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2023-08-30 13:23:47,083 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.
extoutputformat.separator
2023-08-30 13:23:47,416 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2023-08-30 13:23:47,532 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2023-08-30 13:23:47,532 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2023-08-30 13:23:48,020 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-08-30 13:23:49,029 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2023-08-30 13:23:49,191 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.reduce.markreset.buffer.percent is deprecated. Instead, use mapreduc
.reduce.markreset.buffer.percent
2023-08-30 13:23:49,192 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is no
 set, set to default 0.3
2023-08-30 13:23:49,192 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.output.compress is deprecated. Instead, use mapreduce.output.fileoutputf
rmat.compress
2023-08-30 13:23:51,967 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - creating jar file Job855125761983653239.jar
2023-08-30 13:24:04,352 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - jar file Job855125761983653239.jar created
2023-08-30 13:24:04,355 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.jar is deprecated. Instead, use mapreduce.job.jar
2023-08-30 13:24:04,410 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2023-08-30 13:24:04,432 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2023-08-30 13:24:04,432 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2023-08-30 13:24:04,433 [main] INFO  org.apache.pig.data.SchemaTupleFrontend - Setting key [pig.schematuple.classes] with classes to deserialize []
2023-08-30 13:24:04,580 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2023-08-30 13:24:04,583 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker.http.address is deprecated. Instead, use mapreduce.jobtracke
.http.address
2023-08-30 13:24:04,606 [JobControl] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at /0.0.0.0:8032
2023-08-30 13:24:04,778 [JobControl] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-08-30 13:24:07,241 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2023-08-30 13:24:07,241 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2023-08-30 13:24:07,340 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2023-08-30 13:24:07,605 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2023-08-30 13:24:08,724 [JobControl] INFO  org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_1693413019353_0001
2023-08-30 13:24:10,514 [JobControl] INFO  org.apache.hadoop.yarn.client.api.impl.YarnClientImpl - Submitted application application_1693413019353_0001
2023-08-30 13:24:10,663 [JobControl] INFO  org.apache.hadoop.mapreduce.Job - The url to track the job: http://bigdatalite.localdomain:8088/proxy/application_16934130193
53_0001/
2023-08-30 13:24:10,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_1693413019353_0001
2023-08-30 13:24:10,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases dados_maracuja,dados_transforma
dos
2023-08-30 13:24:10,664 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: dados_maracuja[1,17],dados_
transformados[11,22] C:  R:
2023-08-30 13:24:10,777 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 0% complete
2023-08-30 13:24:52,222 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 50% complete
2023-08-30 13:24:57,407 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces
2023-08-30 13:24:57,603 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2023-08-30 13:24:57,613 [main] INFO  org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.6.0-cdh5.13.1 0.12.0-cdh5.13.1        oracle  2023-08-30 13:23:49     2023-08-30 13:24:57     UNKNOWN

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   F
eature  Outputs
job_1693413019353_0001  1       0       13      13      13      13      n/a     n/a     n/a     n/a     dados_maracuja,dados_transformados      MAP_ONLY        /user/or
acle/challenge/dados_transformados,

Input(s):
Successfully read 26 records (888 bytes) from: "/user/oracle/challenge/bahia_maracuja.csv"

Output(s):
Successfully stored 26 records (496 bytes) in: "/user/oracle/challenge/dados_transformados"

Counters:
Total records written : 26
Total bytes written : 496
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1693413019353_0001


2023-08-30 13:24:57,932 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FIELD 3
 time(s).
2023-08-30 13:24:57,932 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

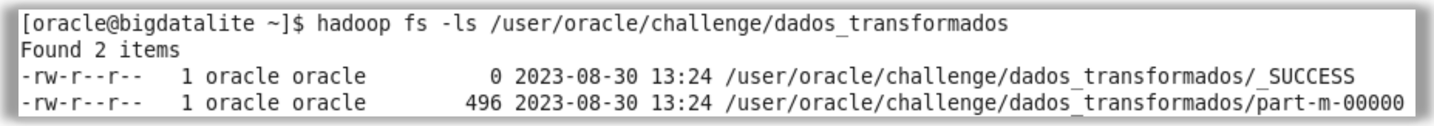
- Após essas etapas, saímos do PIG utilizando o comando “**quit**”.



```
grunt> quit
[oracle@bigdatalite ~]$ 
```

## ★ Copiando o arquivo gerado para o ambiente local

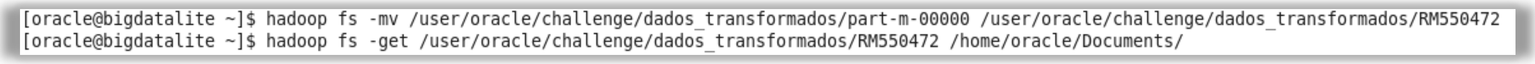
- O próximo passo foi renomear o arquivo resultante do processo anterior, como foi solicitado na tarefa. Então, executamos o comando “**hadoop fs -ls /user/oracle/challenge/dados_transformados**” para verificar o nome atual do arquivo.



```
[oracle@bigdatalite ~]$ hadoop fs -ls /user/oracle/challenge/dados_transformados
Found 2 items
-rw-r--r--   1 oracle oracle          0 2023-08-30 13:24 /user/oracle/challenge/dados_transformados/_SUCCESS
-rw-r--r--   1 oracle oracle        496 2023-08-30 13:24 /user/oracle/challenge/dados_transformados/part-m-00000
```

- Após verificado, renomeamos o arquivo para ‘RM550472’ utilizando o comando “**hadoop fs -mv /user/oracle/challenge/dados_transformados/part-m-00000 /user/oracle/challenge/dados_transformados/RM550472**”.

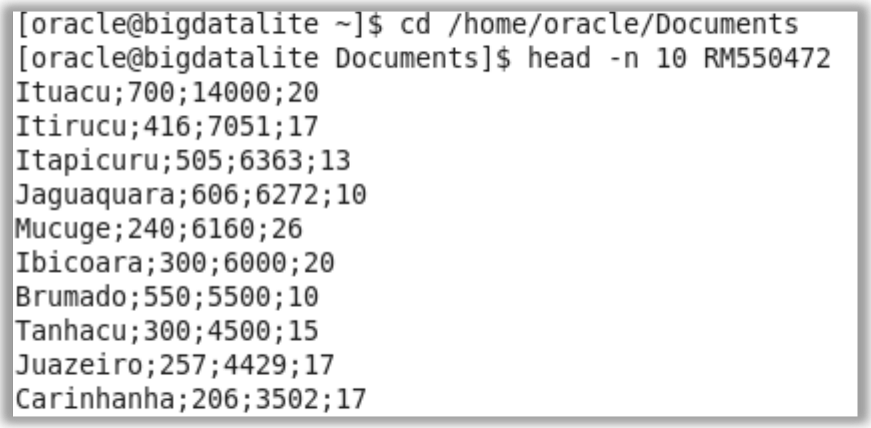
- Para transferi-lo para o ambiente local, mais precisamente para o diretório ‘Documents', executamos o comando “**hadoop fs -get /user/oracle/challenge/dados_transformados/RM550472 /home/oracle/Documents/**”.



```
[oracle@bigdatalite ~]$ hadoop fs -mv /user/oracle/challenge/dados_transformados/part-m-00000 /user/oracle/challenge/dados_transformados/RM550472
[oracle@bigdatalite ~]$ hadoop fs -get /user/oracle/challenge/dados_transformados/RM550472 /home/oracle/Documents/
```

## ★ Listando os dez primeiros registros do arquivo utilizando comandos do sistema operacional.

- Por fim, navegamos até o diretório ‘Documents’ e executamos o comando “**head -n 10 RM550472**” para listar os dez primeiros registros do arquivo 'RM550472’.



```
[oracle@bigdatalite ~]$ cd /home/oracle/Documents
[oracle@bigdatalite Documents]$ head -n 10 RM550472
Ituacu;700;14000;20
Itirucu;416;7051;17
Itapicuru;505;6363;13
Jaguaquara;606;6272;10
Mucuge;240;6160;26
Ibicoara;300;6000;20
Brumado;550;5500;10
Tanhacu;300;4500;15
Juazeiro;257;4429;17
Carinhanha;206;3502;17
```