

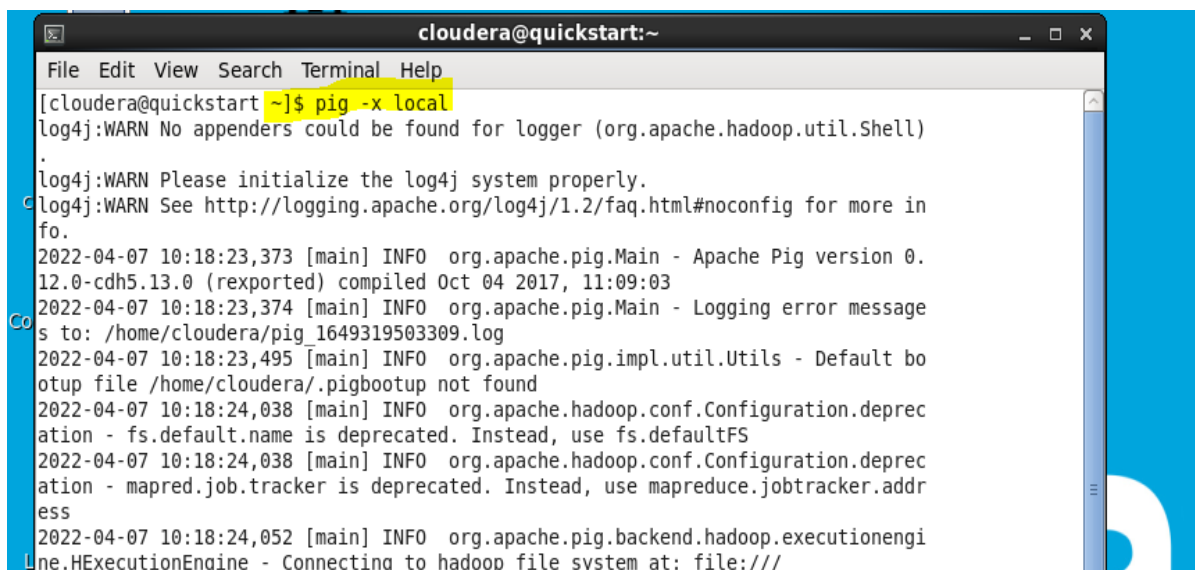
## PIG

PIG es una plataforma de alto nivel para crear programas MapReduce utilizados en hadoop

Pig Latin es un lenguaje de flujo de datos

### Ejercicios PIG

1. Copiar en local file sistema de la MV el fichero datos\_pig.txt en la ruta /home/cloudera/ejercicios/pigy abrir el fichero para revisar su contenido.

A screenshot of a terminal window titled 'cloudera@quickstart:~'. The terminal shows the command '[cloudera@quickstart ~]\$ pig -x local' being executed. The output includes several log messages from log4j and Apache Pig. The log4j messages are warnings about no appenders found and the need to initialize the system properly. The Apache Pig messages include version information (0.12.0-cdh5.13.0), logging error messages to a file, and deprecation warnings for 'fs.default.name' and 'mapred.job.tracker'. The terminal also shows the connection to the Hadoop file system at 'file:///'.

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ pig -x local  
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.  
2022-04-07 10:18:23,373 [main] INFO org.apache.pig.Main - Apache Pig version 0.12.0-cdh5.13.0 (rexported) compiled Oct 04 2017, 11:09:03  
2022-04-07 10:18:23,374 [main] INFO org.apache.pig.Main - Logging error message to: /home/cloudera/pig 1649319503309.log  
2022-04-07 10:18:23,495 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/cloudera/.pigbootstrap not found  
2022-04-07 10:18:24,038 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS  
2022-04-07 10:18:24,038 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2022-04-07 10:18:24,052 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
```

2. Arranca el Shell de Pig en modo local. “pig -x local”

`pig -x local`

3. Carga los datos en pigen una variable llamada “data”. Los nombres de las columnas deben ser (key, campana, fecha, tiempo, display, accion, cpc, pais, lugar). Los tipos

de las columnas deben ser chararray excepto accion y cpc que son int. A.

```
grunt> Data= LOAD 'datos_pig.txt' USING PigStorage(',') as (key:chararray, campana:chararray, fecha:chararray, tiempo:chararray, display:chararray, accion:int, cpc:int, pais:chararray, lugar:chararray)
>>
```

4. Usa el comando DESCRIBE para ver el esquema de la variable “data”

```
grunt> Data= LOAD '/home/cloudera/ejercicios/pig/datos_pig.txt' USING PigStorage(',') AS (key:chararray, campana:chararray, fecha:chararray, tiempo:chararray, display:chararray, accion:int, cpc:int, pais:chararray, lugar:chararray);
grunt> DESCRIBE
Data: {key: chararray, campana: chararray, fecha: chararray, tiempo: chararray, display: chararray, accion: int, cpc: int, pais: chararray, lugar: chararray}
grunt>
```

5. Selecciona las filas de “data” que provengan de USA.

```
grunt> FilasUSA = FILTER Data by pais == USA;
2022-04-07 11:28:19,535 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1025:
<line 4, column 34> Invalid field projection. Projected field [USA] does not exist in schema: key:chararray, campana:chararray, fecha:chararray, tiempo:chararray, display:chararray, accion:int, cpc:int, pais:chararray, lugar:chararray.
Details at logfile: /home/cloudera/pig.1649319503309.log
grunt> result = FILTER Data BY key MATCHES '^surf.*';
grunt>
```

6. Listar los datos que contengan en su key el sufijo surf:

```
grunt> FilasUSA = FILTER Data by pais == USA;
2022-04-07 11:28:19,535 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1025:
<line 4, column 34> Invalid field projection. Projected field [USA] does not exist in schema: key:chararray, campana:chararray, fecha:chararray, tiempo:chararray, display:chararray, accion:int, cpc:int, pais:chararray, lugar:chararray.
Details at logfile: /home/cloudera/pig.1649319503309.log
grunt> result = FILTER Data BY key MATCHES '^surf.*';
grunt>
```

7. Crear una variable llamada “ordenado” que contenga las columnas de data en el siguiente orden: (campana, fecha, tiempo, key, display, lugar, accion, cpc).

```
grunt> ordenado = FOREACH result GENERATE campana, fecha, tiempo, key, display, lugar, accion, cpc;
grunt> DUMP ordenado
```

8. Guarda el contenido de la variable “ordenado” en una carpeta en el local file system de tu MV llamada resultado en la ruta /home/cloudera/ejercicios/piga.

```

grunt> STORE ordenado INTO '/home/cloudera/ejercicios/pig/resultado';
2022-04-07 11:46:05,251 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2022-04-07 11:46:05,251 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEach,
umRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PushDown
rEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}
2022-04-07 11:46:05,265 [main] INFO org.apache.pig.newplan.logical.rules.ColumnPruneVisitor - Columns pruned for Data: $7
2022-04-07 11:46:05,269 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.outpu
textoutputformat.separator
2022-04-07 11:46:05,306 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2022-04-07 11:46:05,321 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-04-07 11:46:05,321 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-04-07 11:46:05,327 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initi
alized
2022-04-07 11:46:05,328 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-04-07 11:46:05,407 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is
t set, set to default 0.3
2022-04-07 11:46:05,426 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2022-04-07 11:46:05,435 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2022-04-07 11:46:05,435 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cache
2022-04-07 11:46:05,435 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.
cal.dir] with code temp directory: /tmp/1649324765426-0
2022-04-07 11:46:05,486 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2022-04-07 11:46:05,499 [JobControl] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already
initialized
2022-04-07 11:46:05,508 [JobControl] WARN org.apache.hadoop.mapreduce.JobResourceUploader - No job jar file set. User classes may not be found. See Job or Job#setJ
ar(String).
2022-04-07 11:46:05,531 [JobControl] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-04-07 11:46:05,531 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-04-07 11:46:05,534 [JobControl] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths (combined) to process : 1
2022-04-07 11:46:05,539 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - number of splits:1
2022-04-07 11:46:05,581 [JobControl] INFO org.apache.hadoop.mapreduce.JobSubmitter - Submitting tokens for job: job_local64104523_0004
2022-04-07 11:46:05,581 [JobControl] INFO org.apache.hadoop.mapreduce.Job - The url to track the job: http://localhost:8080/
2022-04-07 11:46:06,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - HadoopJobId: job_local64104523_0004
2022-04-07 11:46:06,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Processing aliases Data,ordenado,result
2022-04-07 11:46:06,016 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - detailed locations: M: Data[1,6],Data[-1,-1],re
sult[4,9],ordenado[6,11] C: R:

```

## 9.Comprobar el contenido de la carpeta.