SAKARYA ÜNİVERSİTESİ

BIG DATA

TOO BIG TO IGNORE

SÜMEYYE KAYNAK

OUTLINE

Apache Pig

MongoDB

APACHE PIG

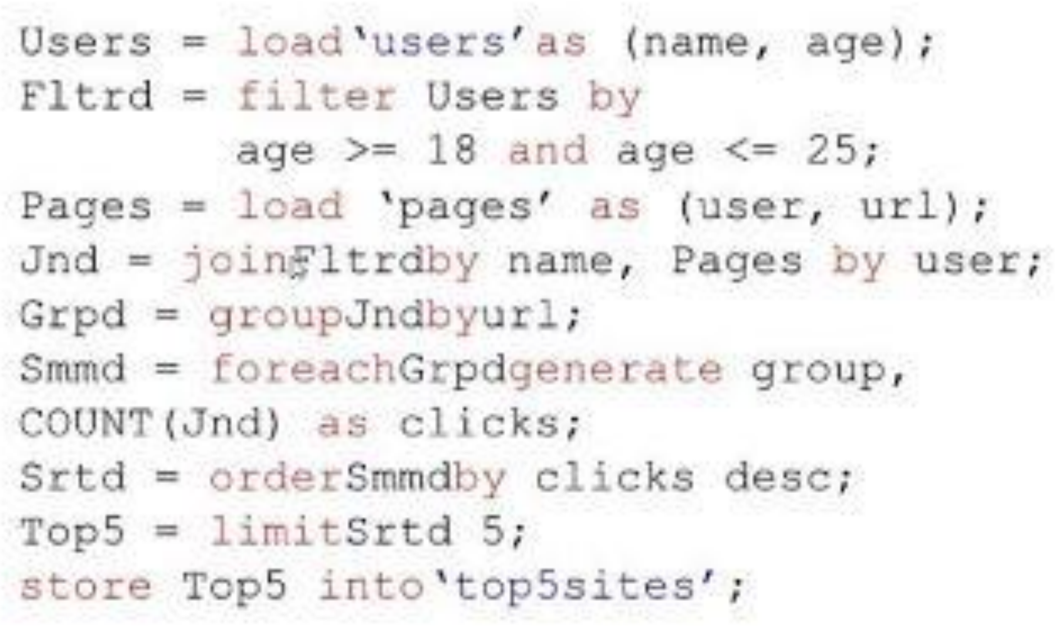
We can analyze big data with map-reduce.

- Map-reduce development methods:
    - Java, Phyton, Scala map-reduce
    - Apache Pig
    - Apache Hive

APACHE PIG

- Apache Pig has own programming language named pig Latin.

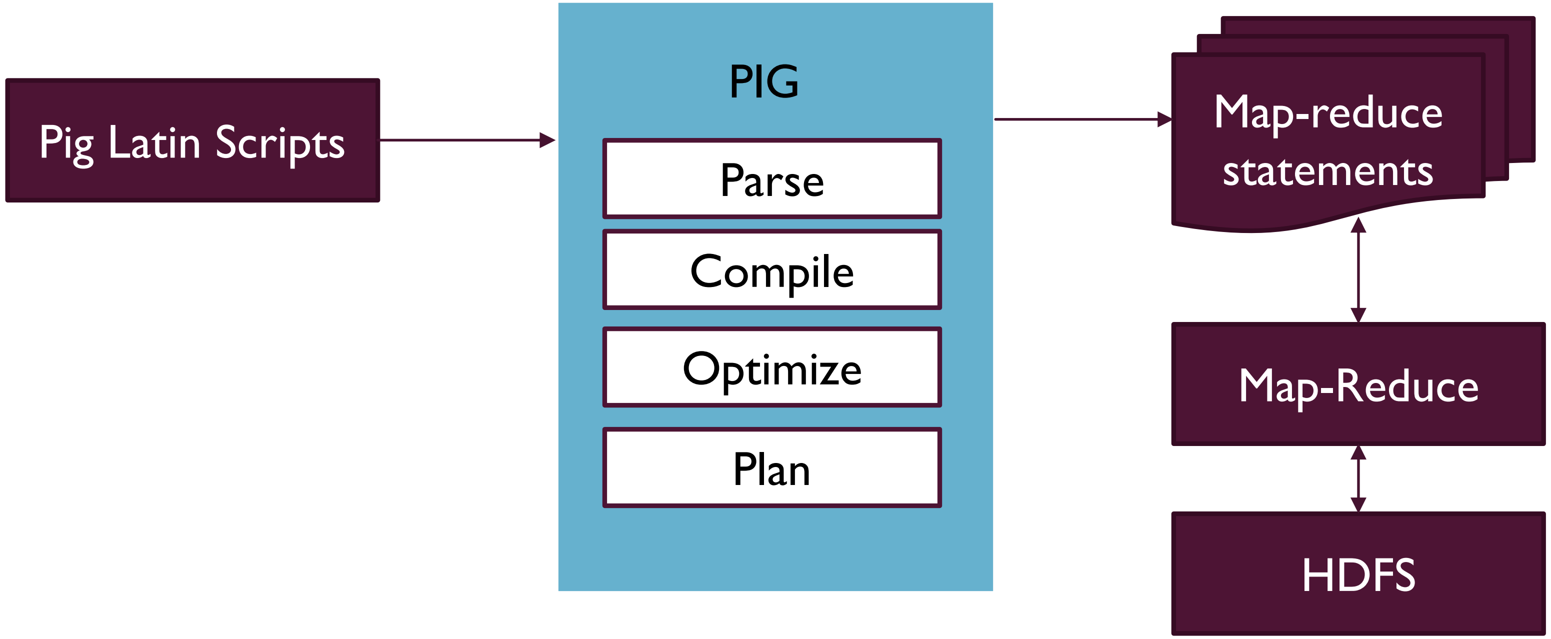
PIG LATIN



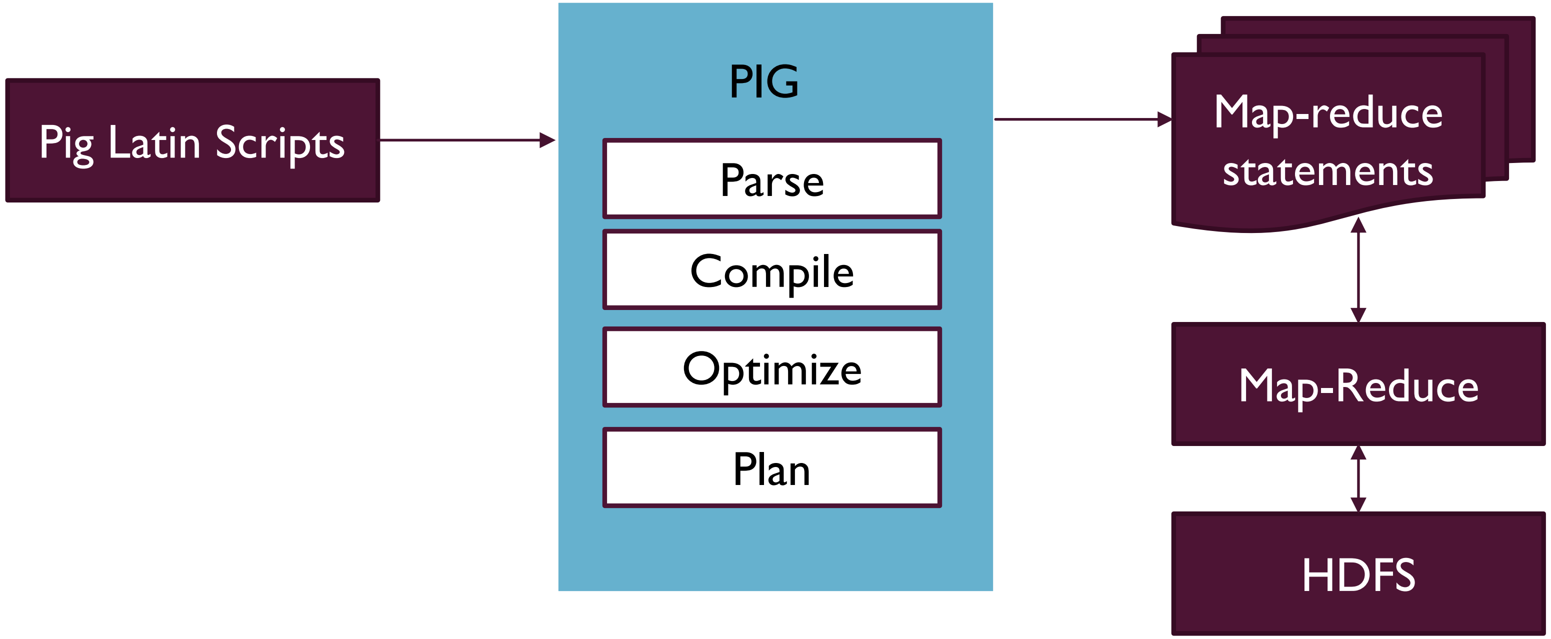
```
Users = load‘users’as (name, age);
Fltrd = filter Users by
        age >= 18 and age <= 25;
Pages = load ‘pages’ as (user, url);
Jnd = joinFltrdby name, Pages by user;
Grpd = groupJndbyurl;
Smmd = foreachGrpdgenerate group,
COUNT(Jnd) as clicks;
Srtd = orderSmmdby clicks desc;
Top5 = limitSrtd 5;
store Top5 into‘top5sites’;
```

- The “users” data in HDFS is loaded.
- The “users” data have name and age information.
- The “users” data is filtered. (Age greater than 18 and less than 25)
- The “pages” data in HDFS is loaded.
- The “pages” data have user and url information.
- The users and pages data are joined then grouped.
- Finally, the 5 most visited sites are selected.

PIG ARCHITECTURE

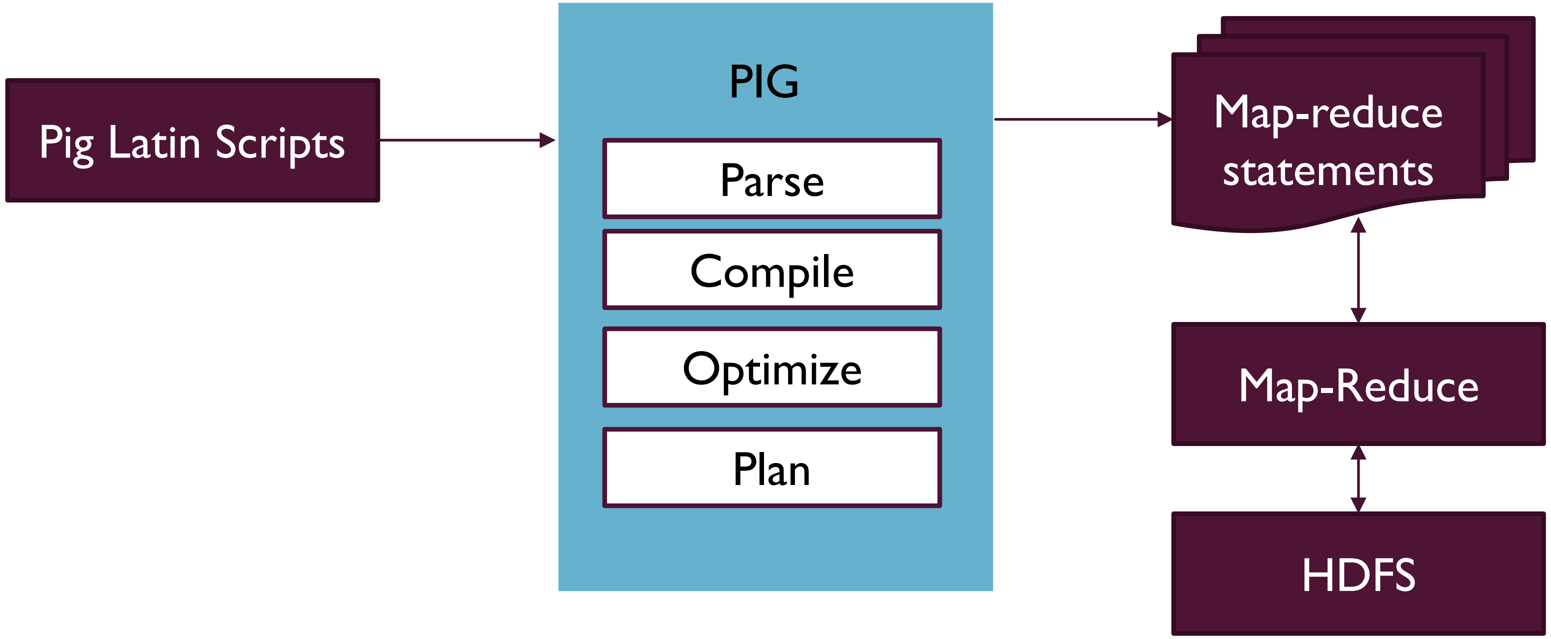


PIG ARCHITECTURE



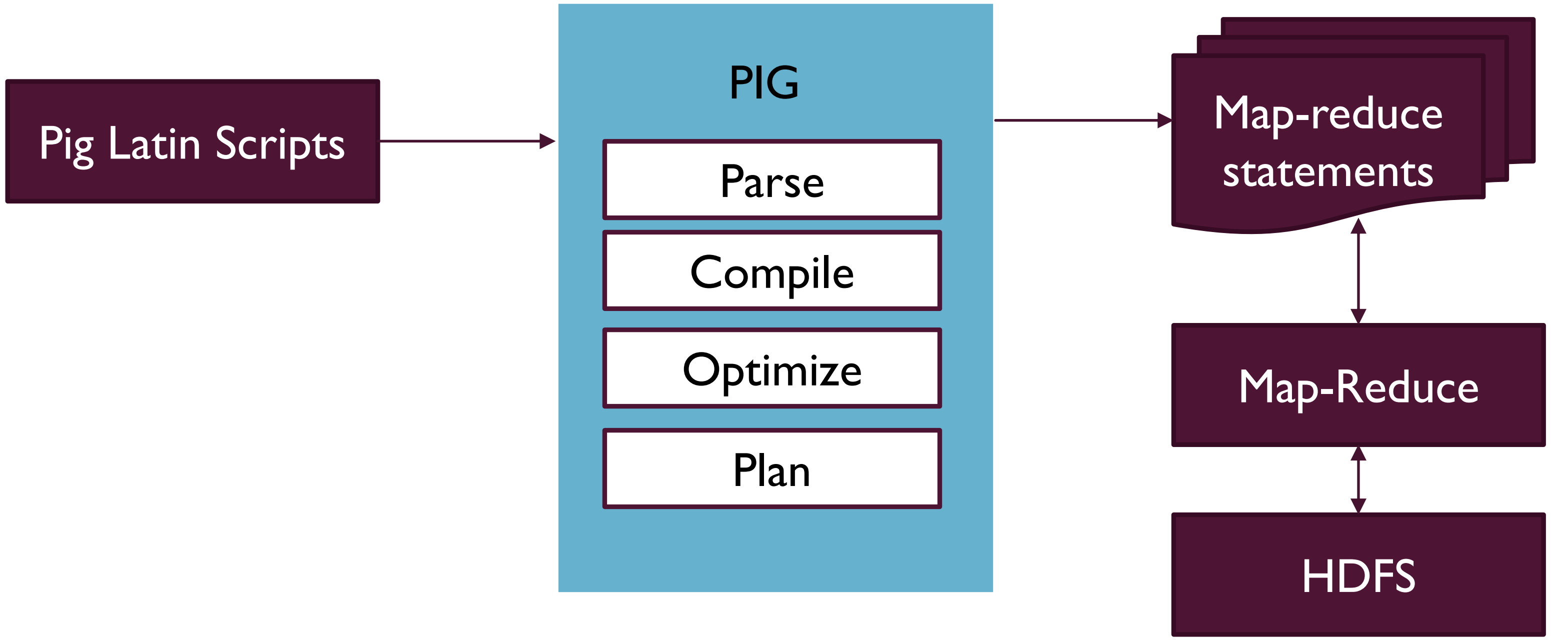
- Parse: Syntax check is done.

PIG ARCHITECTURE



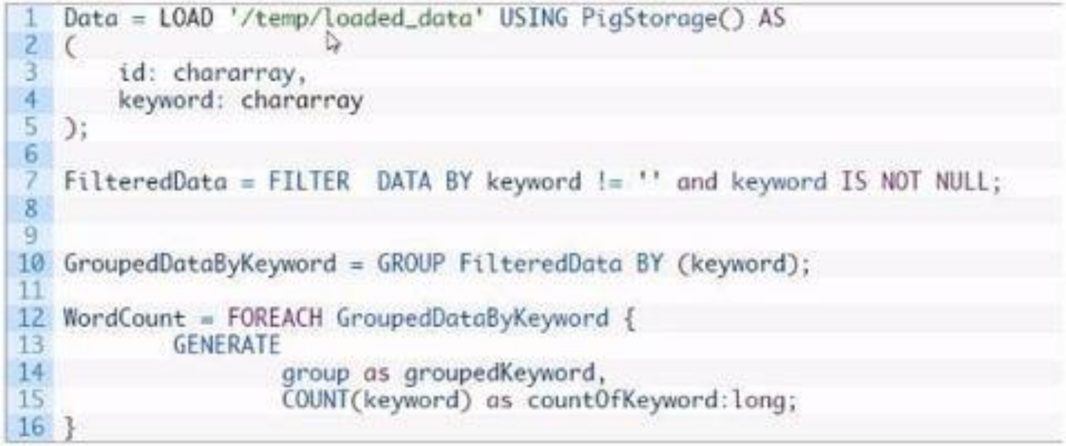
- Compile: The written codes are converted to map-reduce.

PIG ARCHITECTURE



- Optimize and plan: Optimization of the codes is done by Pig.

WORD COUNT APPLICATION



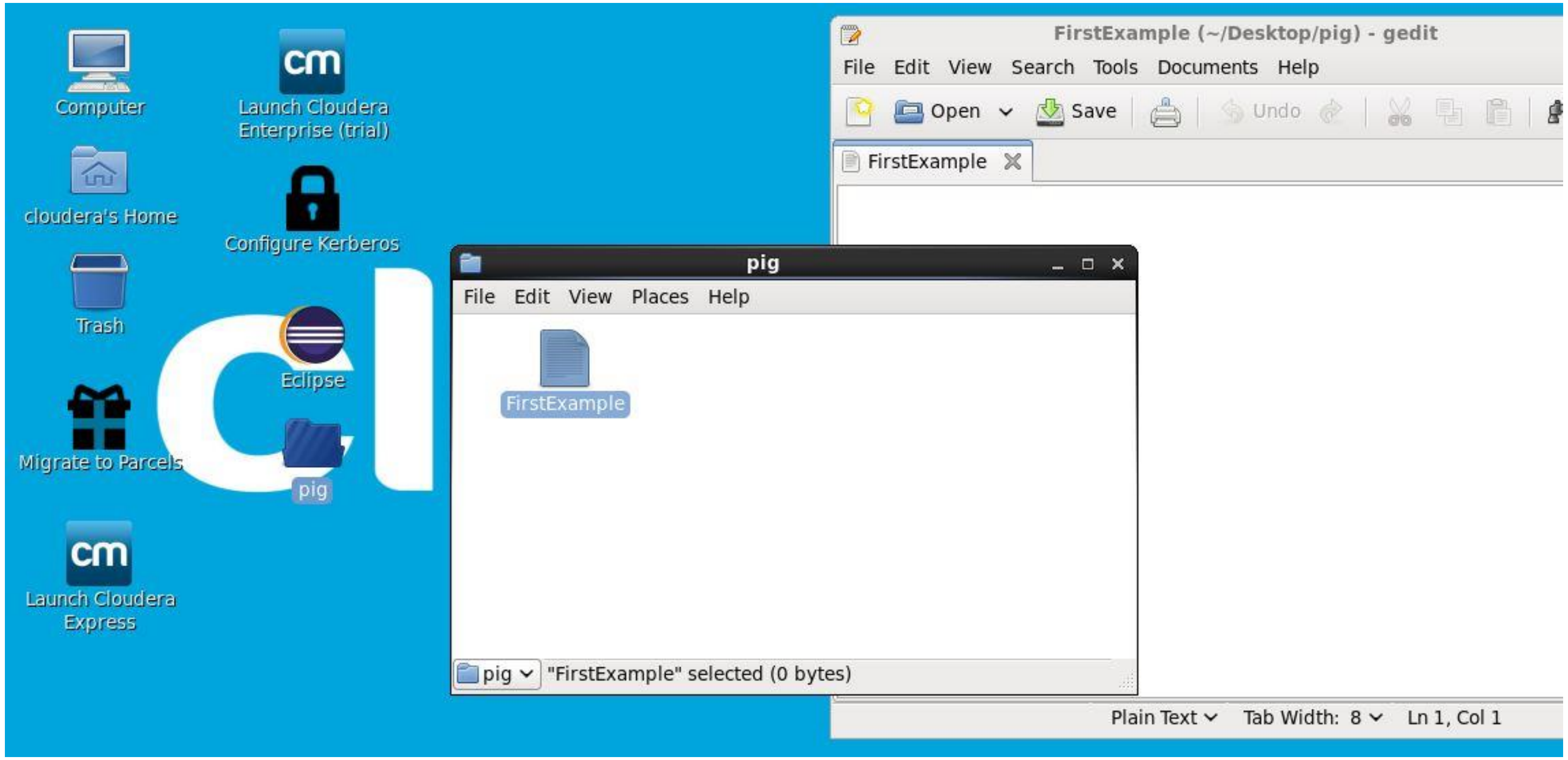
```
Data = LOAD '/temp/loaded_data' USING PigStorage() AS
(
    id: chararray,
    keyword: chararray
);

FilteredData = FILTER  DATA BY keyword != '' and keyword IS NOT NULL;


GroupedDataByKeyword = GROUP FilteredData BY (keyword);

WordCount = FOREACH GroupedDataByKeyword {
        GENERATE
                group as groupedKeyword,
                COUNT(keyword) as countOfKeyword:long;
}
```

- There is a file named “loaded_data” under the temp folder in HDFS.
- This file is loaded into the Data variable using the LOAD command.
- There are keyword and id fields in this file.
- The keyword field of the file is filtered.
- Grouping according to the keyword field has been performed.
- Counting words was done using the Foreach loop.

APACHE PIG ADVANTAGES

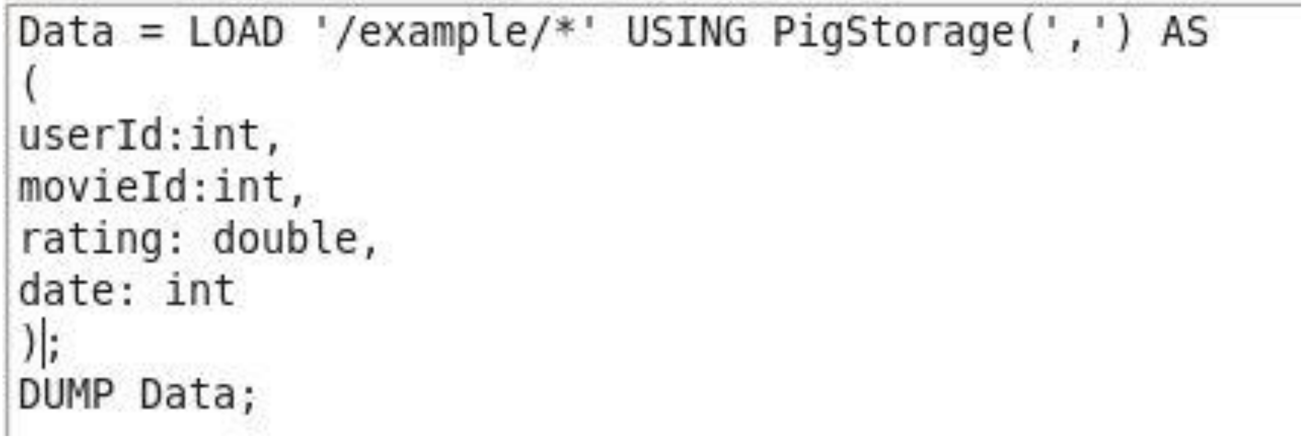
- It is simple to develop and learn to Apache Pig.
- It can easily perform analyzes on big data.
- It optimized the codes we write.
- It provides methods by which we can analyze data (filter, join).
- If needed, we can write libraries with javascript, java or python and use them in apache pig. This is named as UDF.

EXAMPLES

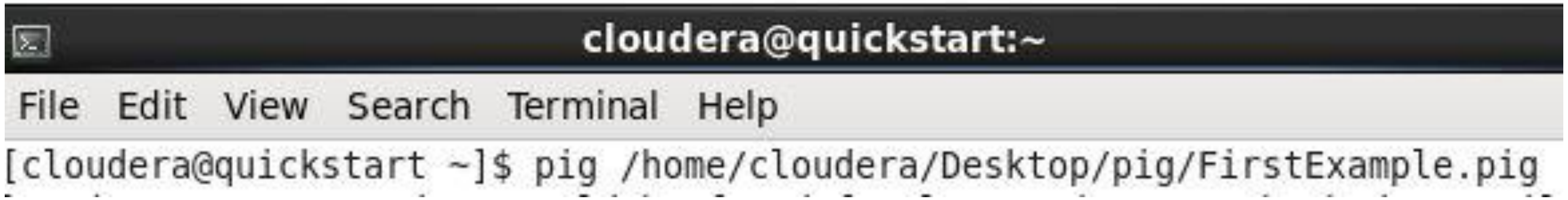


"FirstExample" selected (0 bytes)

EXAMPLES

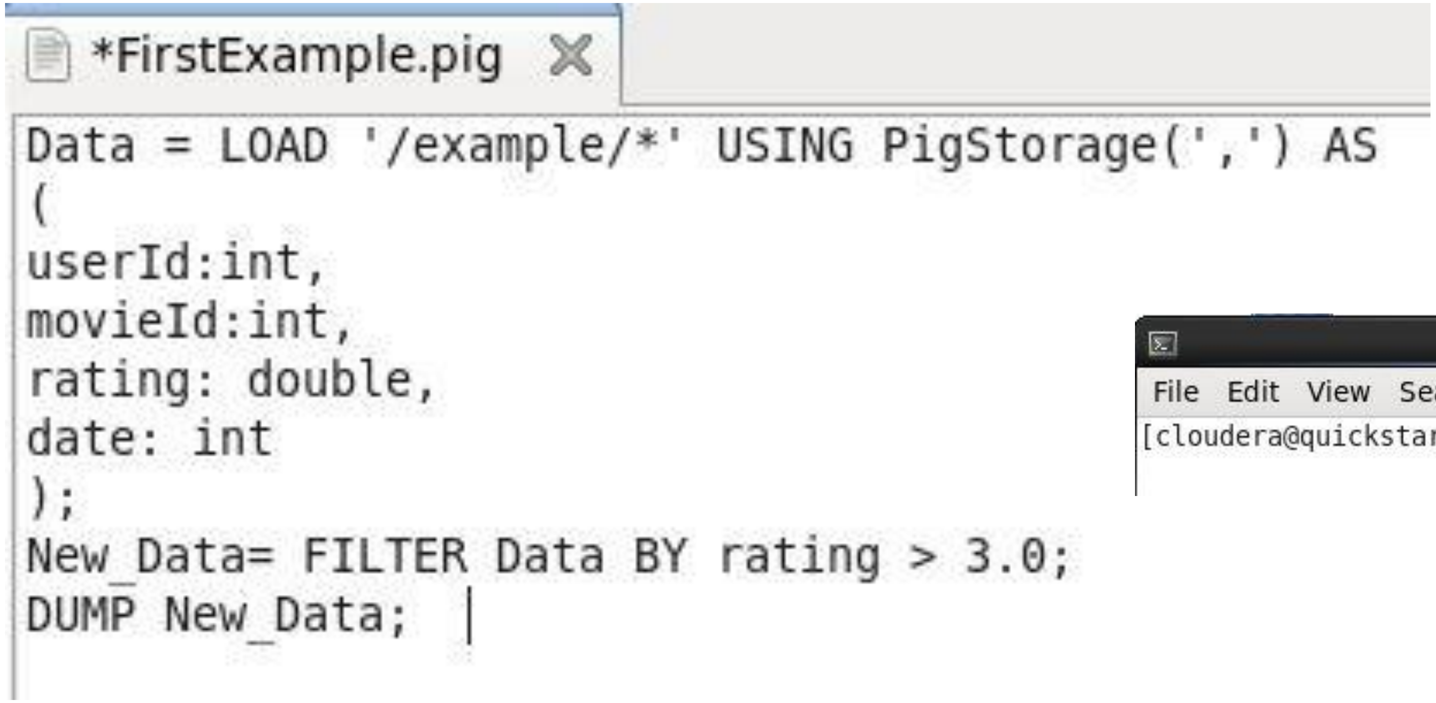


```
Data = LOAD '/example/*' USING PigStorage(',') AS
(
userId:int,
movieId:int,
rating: double,
date: int
);
DUMP Data;
```

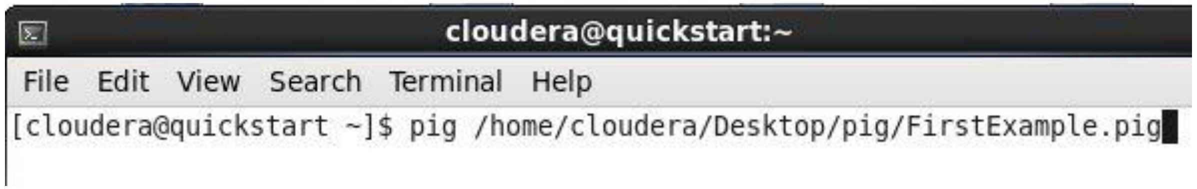


```
[cloudera@quickstart ~]$ pig /home/cloudera/Desktop/pig/FirstExample.pig
```

EXAMPLES

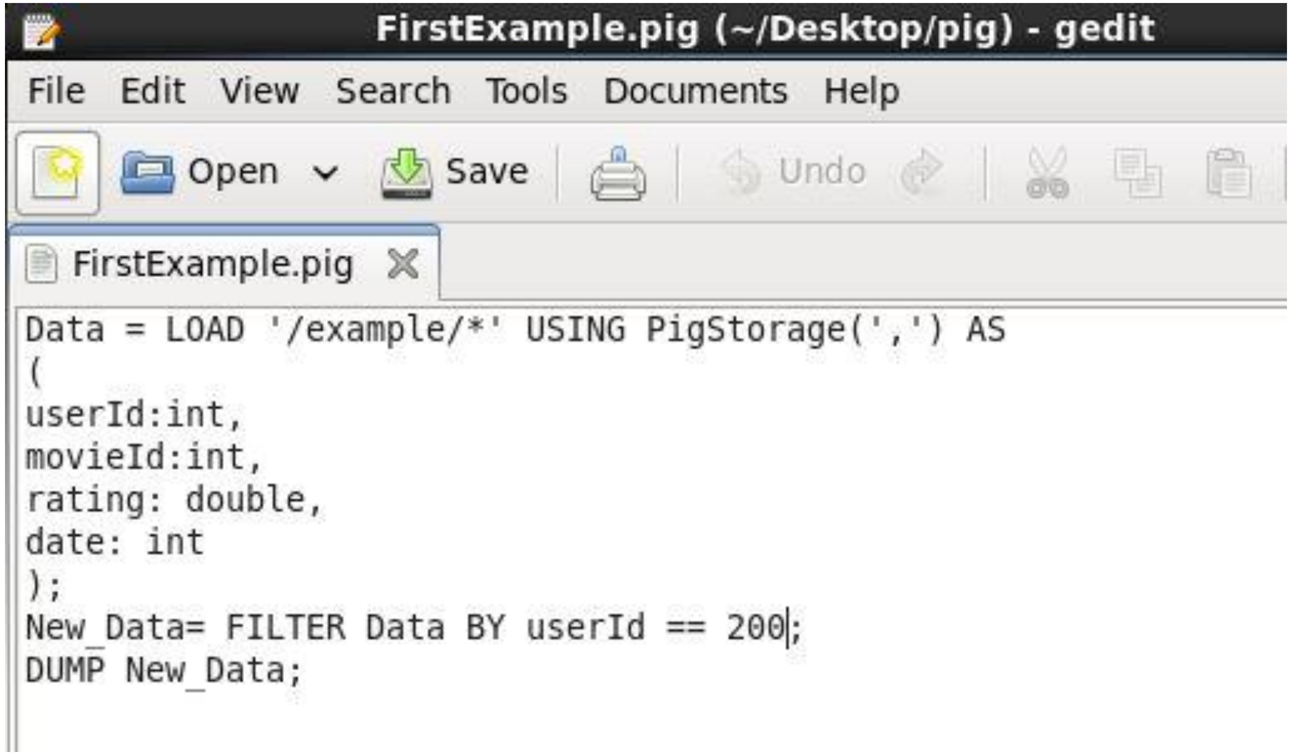


```
Data = LOAD '/example/*' USING PigStorage(',') AS
(
userId:int,
movieId:int,
rating: double,
date: int
);
New_Data= FILTER Data BY rating > 3.0;
DUMP New_Data;
```



```
[cloudera@quickstart ~]$ pig /home/cloudera/Desktop/pig/FirstExample.pig
```

EXAMPLES


FirstExample.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
FirstExample.pig

```
Data = LOAD '/example/*' USING PigStorage(',') AS
(
userId:int,
movieId:int,
rating: double,
date: int
);
New_Data= FILTER Data BY userId == 200;
DUMP New_Data;
```

FUNCTIONS AND OPERATORS

Id	Country	Duration Time	Search
253	US	9424	Bebek Bezi
234	TR	5462	Klavye
125	EN	3452	Deterjan
560	TR	1235	Süt
685	US	4564	Koltuk Takımı
456	EN	1249	Paspas
237	TR	8655	Halı

PIG ARITHMETIC OPERATORS

Operator	Symbol	Sample
Add	+	8+4=12
Subtraction	-	8-4=4
Multiplication	*	8*4=32
Division	/	8/4=2
Modulo	%	8%4=0
Bincond	?:	8==4?’eşit’.’eşit değil’

COMPARISON OPERATORS

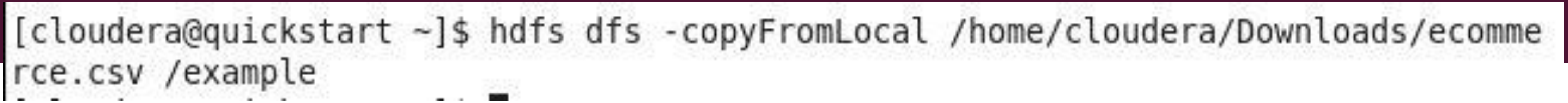
Operator	Symbol	Sample
Equal to	==	B= FILTER A BY(Id==560);
Not equal to	!=	B= FILTER A BY(Country!=‘TR’);
Less than	<	B= FILTER A BY(DurationTime<30000);
Greater than	>	B= FILTER A BY(DurationTime>1000);
Less than or equal to	<=	B= FILTER A BY(DurationTime<=2000);
Greater than or equal to	>=	B= FILTER A BY(DurationTime>=30000);
Regex	matches	B= FILTER A BY(Search **matches** ‘.*Koltuk.*’);

LOGICAL OPERATORS

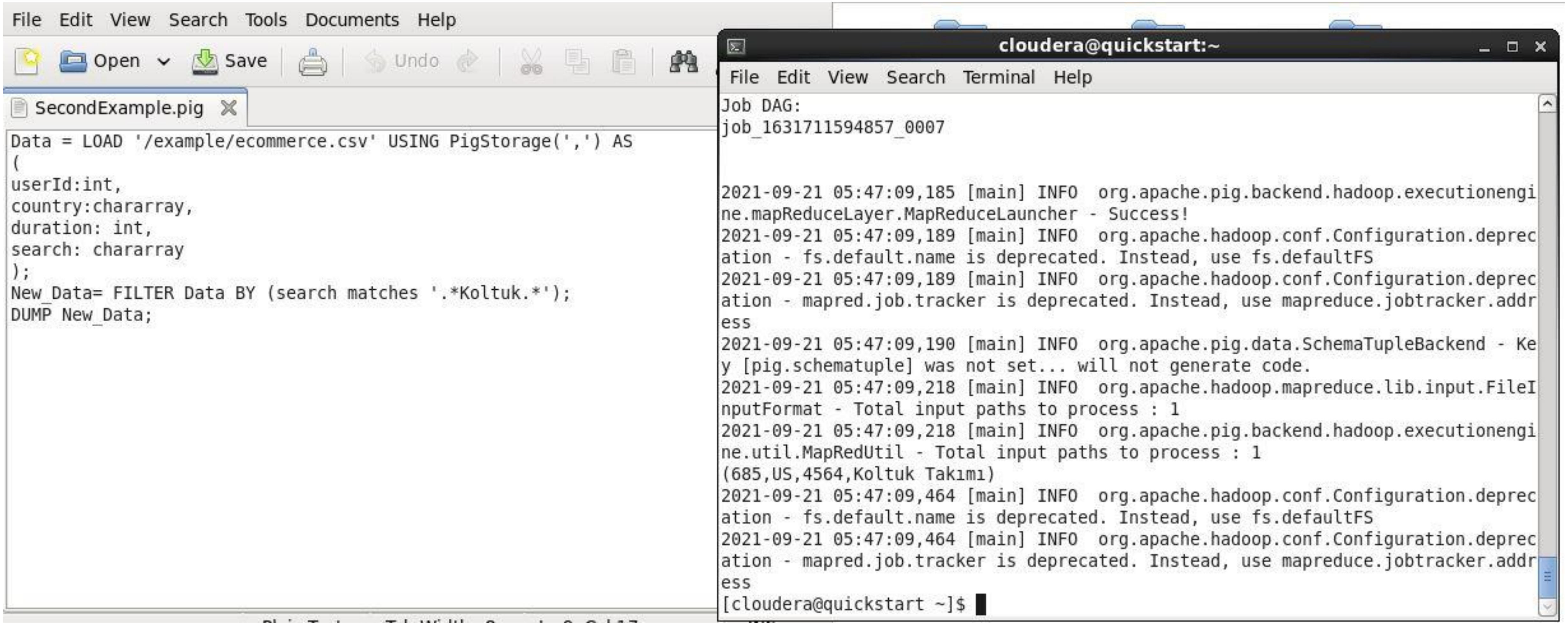
Operator	Symbol	Sample
AND	and	B= FILTER A BY(Country!='TR') **AND** (DurationTime > 3000);
OR	or	B= FILTER A BY(Country!='TR') **OR** (Country!='US');
NOT	not	B= FILTER A BY(**NOT** DurationTime < 30000);

Operator	Symbol	Sample
Is null	is null	B= FILTER A BY(Country **is null**);
Is not null	is not null	B= FILTER A BY(Country **is not null**);

EXAMPLE



```
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /home/cloudera/Downloads/ecomme
rce.csv /example
```



File Edit View Search Tools Documents Help
Open
Save
Undo
SecondExample.pig
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Data = LOAD '/example/ecommerce.csv' USING PigStorage(',') AS
(
userId:int,
country:chararray,
duration: int,
search: chararray
);
New_Data= FILTER Data BY (search matches '.*Koltuk.*');
DUMP New_Data;
```

```
Job DAG:
job_1631711594857_0007


2021-09-21 05:47:09,185 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-09-21 05:47:09,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 05:47:09,189 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-21 05:47:09,190 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-09-21 05:47:09,218 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-09-21 05:47:09,218 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(685,US,4564,Koltuk Takımı)
2021-09-21 05:47:09,464 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 05:47:09,464 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
[cloudera@quickstart ~]$ 
```

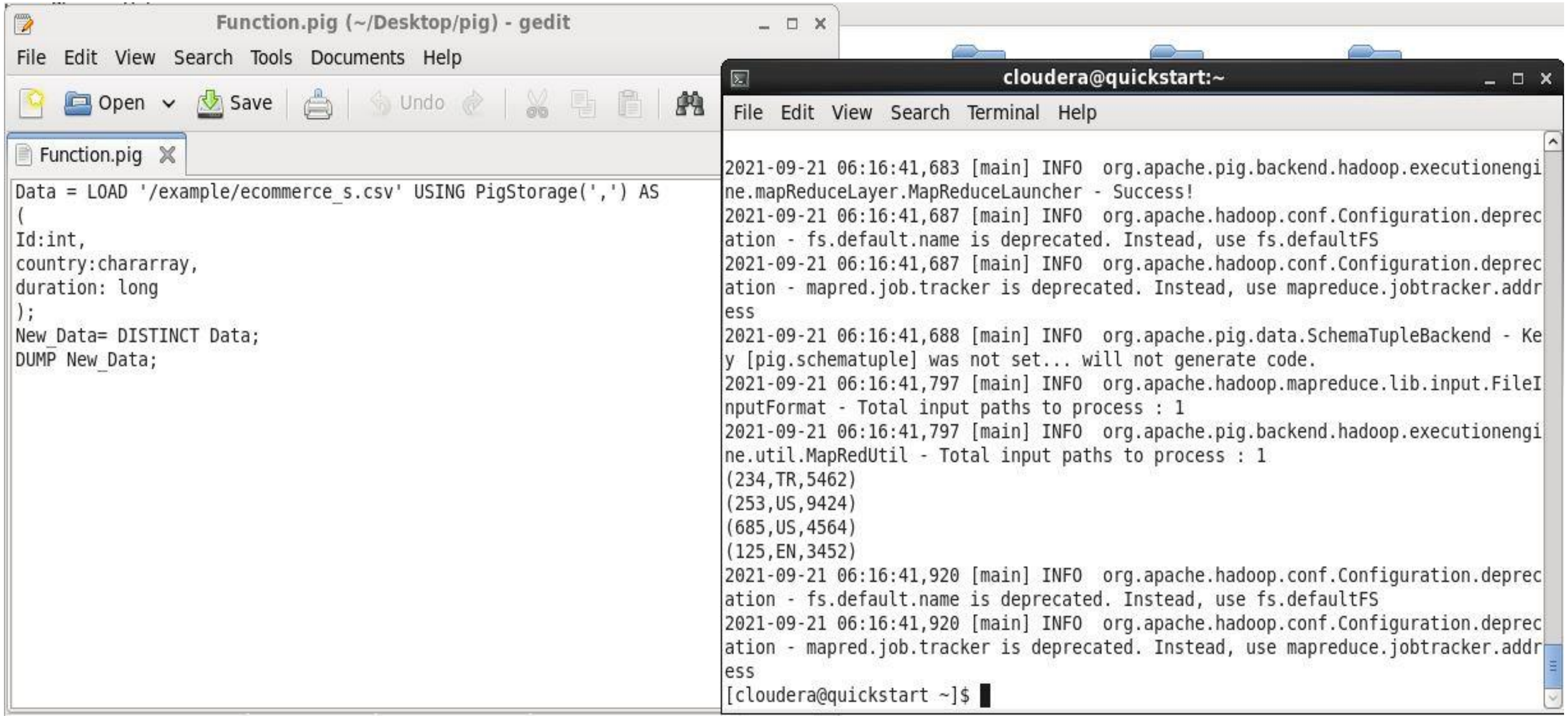
APACHE PING FUNCTIONS

- **DISTINCT FUNCTION : Deletes records with the same information.**

Id	Country	DurationTime
253	US	9424
234	TR	5462
125	EN	3452
234	TR	5462
685	US	4564

APACHE PING FUNCTIONS

- **DISTINCT FUNCTION : Deletes records with the same information.**

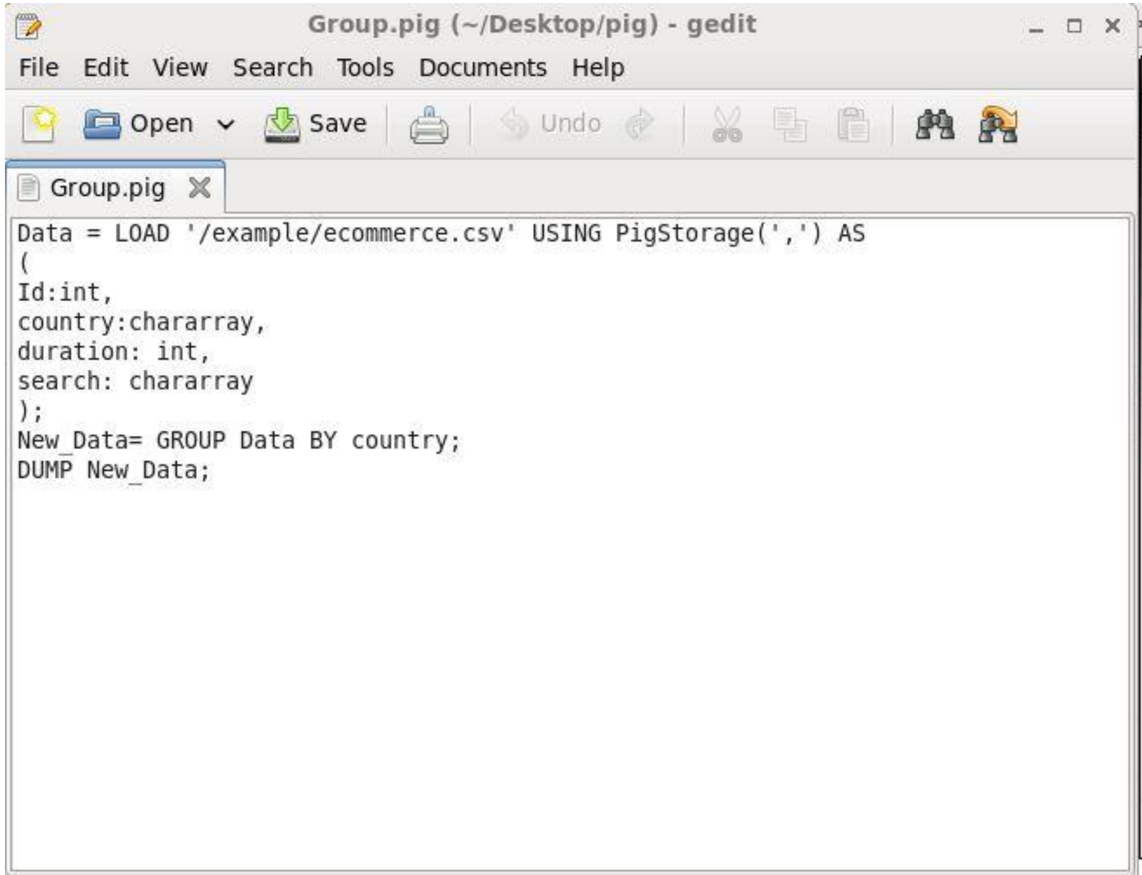

Function.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Function.pig
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Data = LOAD '/example/ecommerce_s.csv' USING PigStorage(',') AS
(
Id:int,
country:chararray,
duration: long
);
New_Data= DISTINCT Data;
DUMP New_Data;
```

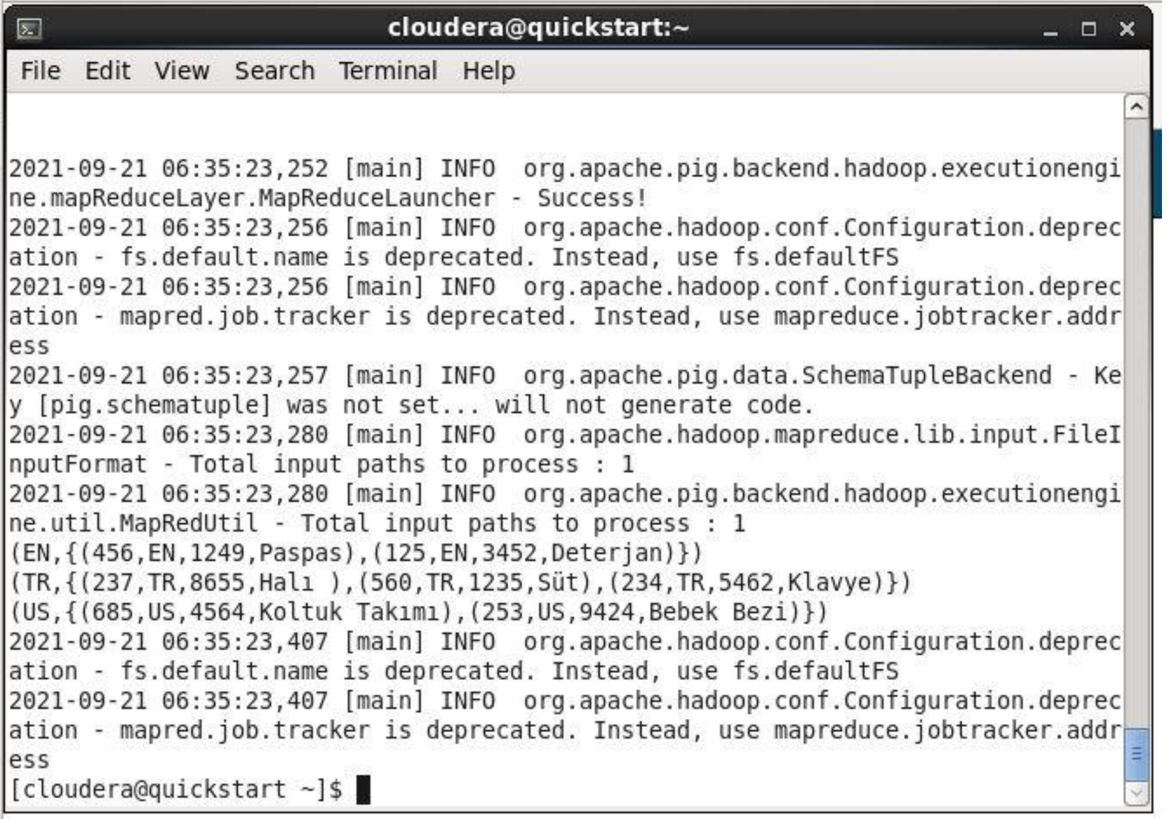
```
2021-09-21 06:16:41,683 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-09-21 06:16:41,687 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 06:16:41,687 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-21 06:16:41,688 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-09-21 06:16:41,797 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-09-21 06:16:41,797 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(234,TR,5462)
(253,US,9424)
(685,US,4564)
(125,EN,3452)
2021-09-21 06:16:41,920 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 06:16:41,920 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
[cloudera@quickstart ~]$ 
```

APACHE PING FUNCTIONS

- **GROUP FUNCTION : Grouping may be necessary for results such as the maximum, minimum, or average value within a field.**


Group.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Group.pig

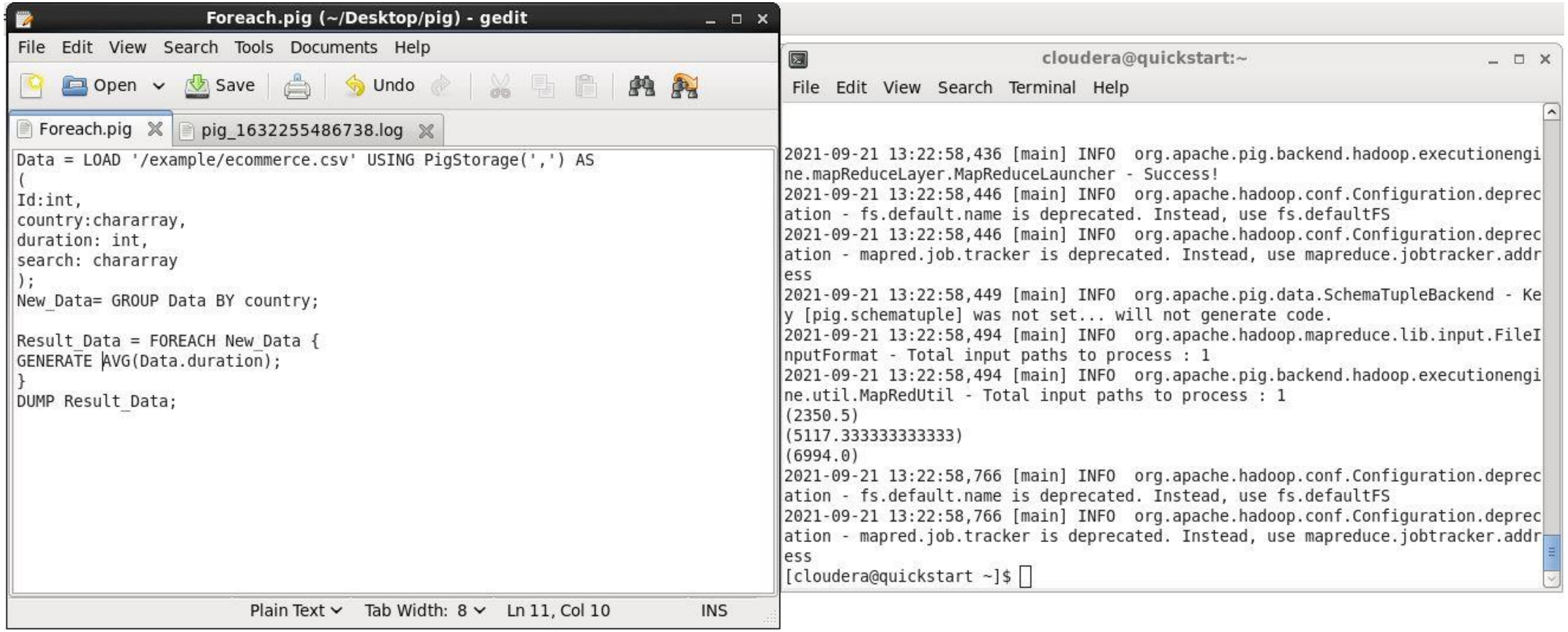
```
Data = LOAD '/example/ecommerce.csv' USING PigStorage(',') AS
(
Id:int,
country:chararray,
duration: int,
search: chararray
);
New_Data= GROUP Data BY country;
DUMP New_Data;
```



```
2021-09-21 06:35:23,252 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-09-21 06:35:23,256 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 06:35:23,256 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-21 06:35:23,257 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-09-21 06:35:23,280 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-09-21 06:35:23,280 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(EN,{(456,EN,1249,Paspas),(125,EN,3452,Deterjan)})
(TR,{(237,TR,8655,Halı ),(560,TR,1235,Süt),(234,TR,5462,Klavye)})
(US,{(685,US,4564,Koltuk Takımı),(253,US,9424,Bebek Bezi)})
2021-09-21 06:35:23,407 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 06:35:23,407 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
[cloudera@quickstart ~]$ 
```

APACHE PING FUNCTIONS

- **FOREACH FUNCTION : It allows us to navigate through the data with a loop.**

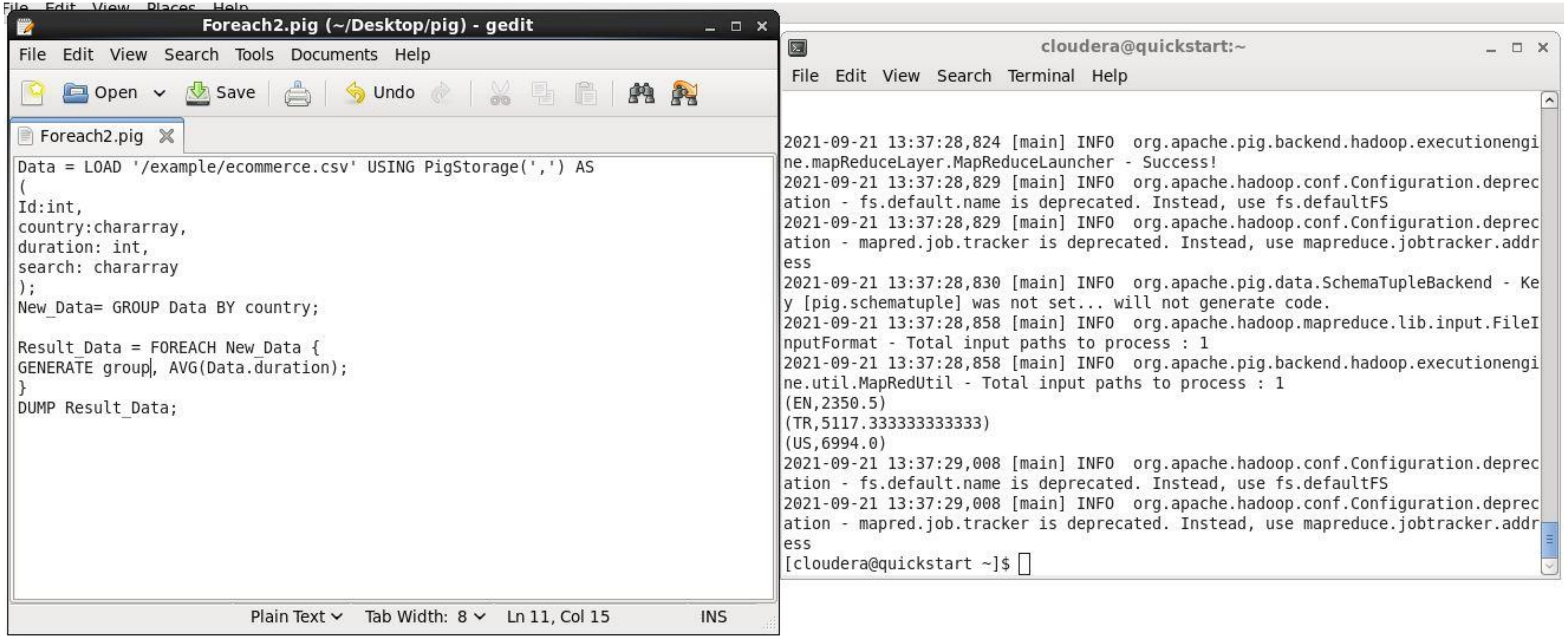

Foreach.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Foreach.pig
pig_1632255486738.log
Plain Text
Tab Width: 8
Ln 11, Col 10
INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Data = LOAD '/example/ecommerce.csv' USING PigStorage(',') AS
(
Id:int,
country:chararray,
duration: int,
search: chararray
);
New_Data= GROUP Data BY country;

Result_Data = FOREACH New_Data {
GENERATE AVG(Data.duration);
}
DUMP Result_Data;
```

```
2021-09-21 13:22:58,436 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-09-21 13:22:58,446 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 13:22:58,446 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-21 13:22:58,449 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-09-21 13:22:58,494 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-09-21 13:22:58,494 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(2350.5)
(5117.333333333333)
(6994.0)
2021-09-21 13:22:58,766 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 13:22:58,766 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
[cloudera@quickstart ~]$ 
```

APACHE PING FUNCTIONS

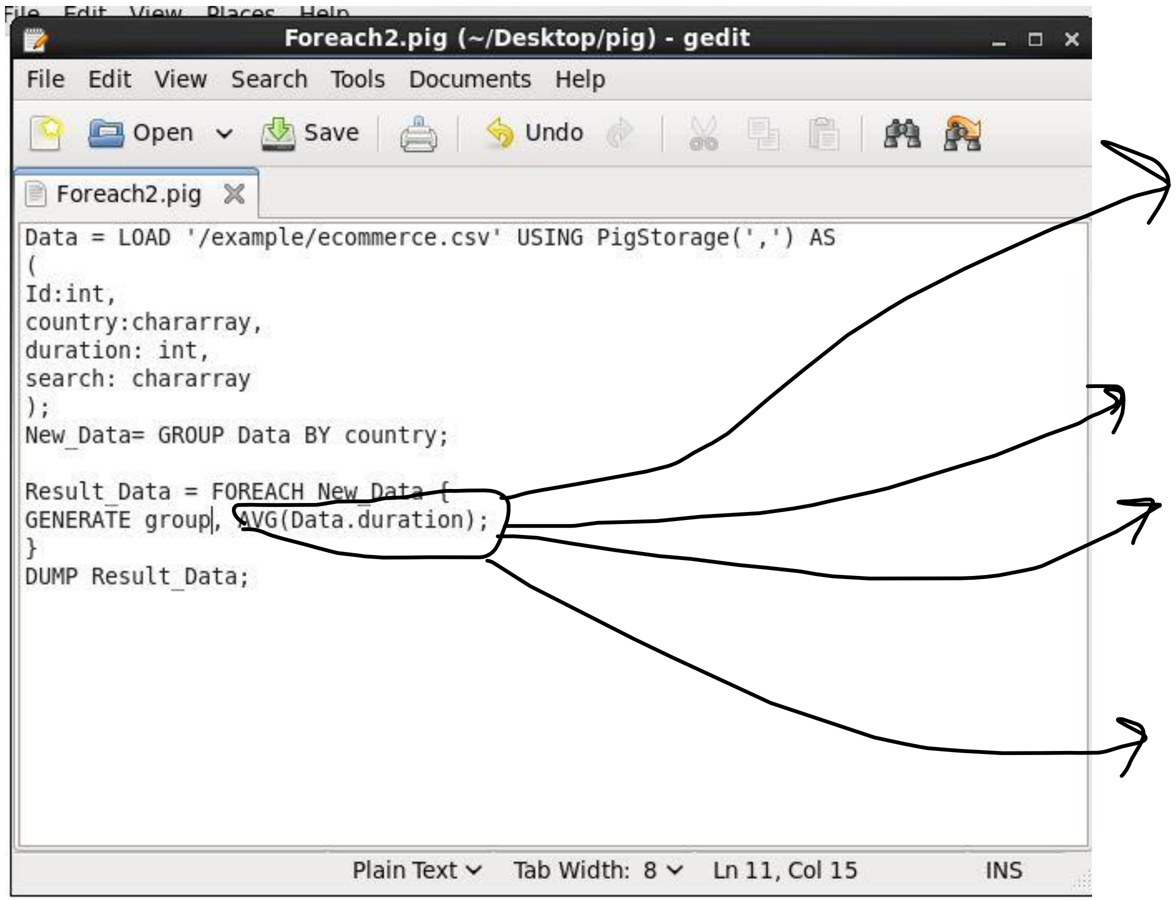
- **FOREACH FUNCTION : It allows us to navigate through the data with a loop.**


File Edit View Places Help
Foreach2.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Foreach2.pig
Plain Text Tab Width: 8 Ln 11, Col 15 INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Data = LOAD '/example/ecommerce.csv' USING PigStorage(',') AS
(
Id:int,
country:chararray,
duration: int,
search: chararray
);
New_Data= GROUP Data BY country;

Result_Data = FOREACH New_Data {
GENERATE group, AVG(Data.duration);
}
DUMP Result_Data;
```

```
2021-09-21 13:37:28,824 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.mapReduceLayer.MapReduceLauncher - Success!
2021-09-21 13:37:28,829 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 13:37:28,829 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2021-09-21 13:37:28,830 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Ke
y [pig.schematuple] was not set... will not generate code.
2021-09-21 13:37:28,858 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2021-09-21 13:37:28,858 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(EN,2350.5)
(TR,5117.333333333333)
(US,6994.0)
2021-09-21 13:37:29,008 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2021-09-21 13:37:29,008 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
[cloudera@quickstart ~]$ 
```

APACHE PING FUNCTIONS



File Edit View Places Help
Foreach2.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Foreach2.pig
Plain Text
Tab Width: 8
Ln 11, Col 15
INS

```
Data = LOAD '/example/ecommerce.csv' USING PigStorage(',') AS
(
Id:int,
country:chararray,
duration: int,
search: chararray
);
New_Data= GROUP Data BY country;

Result_Data = FOREACH New_Data {
GENERATE group, AVG(Data.duration);
}
DUMP Result_Data;
```

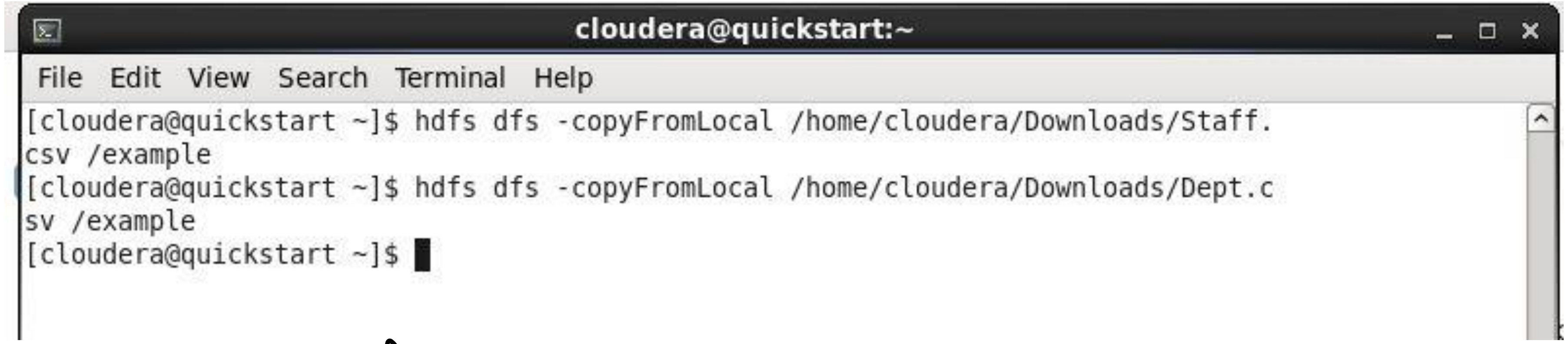
- COUNT(New_Data)
- MAX(Data.DurationTime) as maxDT
- MIN(Data.DurationTime) as minDT
- SUM(Data.DurationTime) as totalDT

APACHE PIG-JOIN APPLICATION

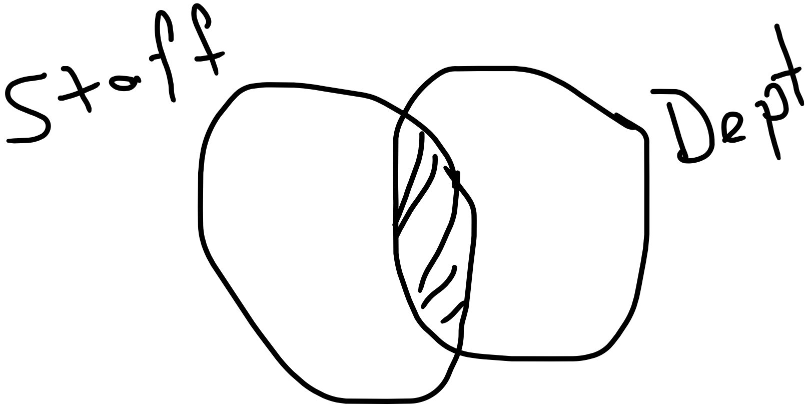
Name	Age	Dept_Id
Ahmet	27	1
Mehmet	35	2
Fatma	24	3
Seda	26	2
Cenk	34	3
Peter	30	1
Burak	29	2

Dept_Id	Dept_Name
1	Sales
2	Arge

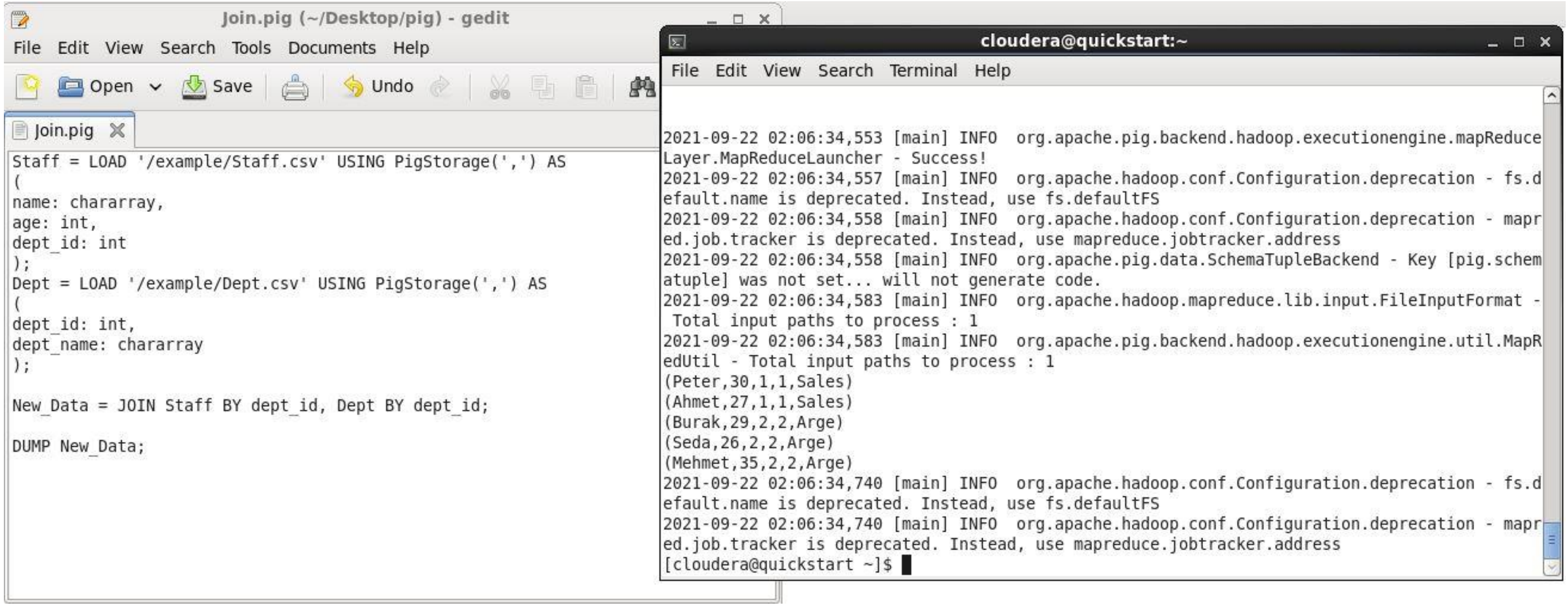
APACHE PIG-JOIN AND UNION



```
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /home/cloudera/Downloads/Staff.
csv /example
[cloudera@quickstart ~]$ hdfs dfs -copyFromLocal /home/cloudera/Downloads/Dept.c
sv /example
[cloudera@quickstart ~]$ 
```



APACHE PIG-JOIN AND UNION

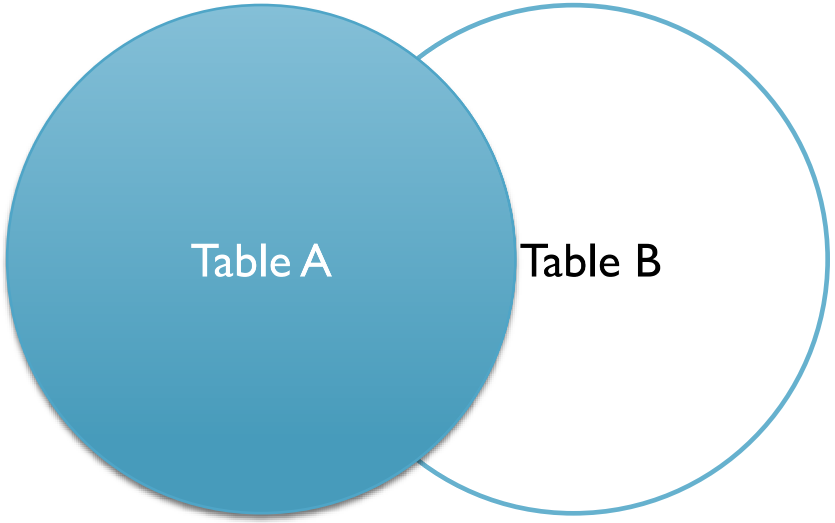

Join.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Join.pig
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Staff = LOAD '/example/Staff.csv' USING PigStorage(',') AS
(
name: chararray,
age: int,
dept_id: int
);
Dept = LOAD '/example/Dept.csv' USING PigStorage(',') AS
(
dept_id: int,
dept_name: chararray
);

New_Data = JOIN Staff BY dept_id, Dept BY dept_id;

DUMP New_Data;
```

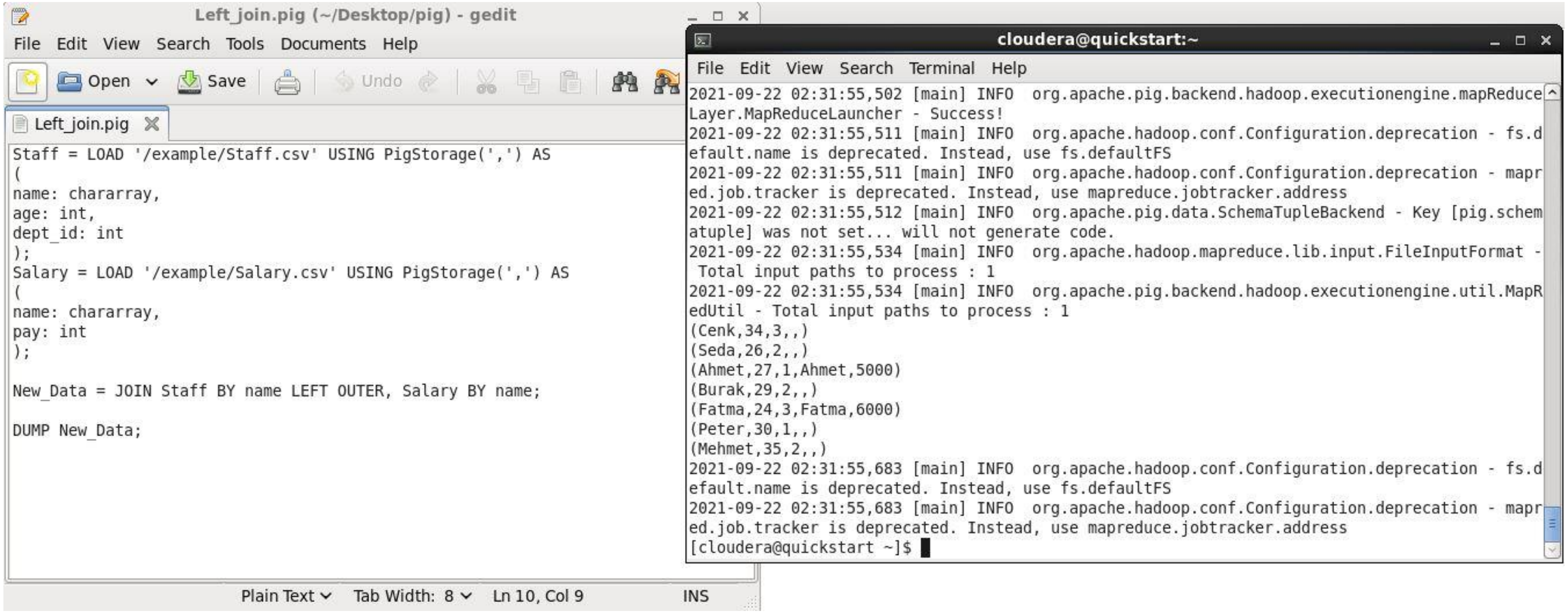
```
2021-09-22 02:06:34,553 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MapReduceLauncher - Success!
2021-09-22 02:06:34,557 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:06:34,558 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-22 02:06:34,558 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schem
atuple] was not set... will not generate code.
2021-09-22 02:06:34,583 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
 Total input paths to process : 1
2021-09-22 02:06:34,583 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapR
edUtil - Total input paths to process : 1
(Peter,30,1,1,Sales)
(Ahmet,27,1,1,Sales)
(Burak,29,2,2,Arge)
(Seda,26,2,2,Arge)
(Mehmet,35,2,2,Arge)
2021-09-22 02:06:34,740 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:06:34,740 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
[cloudera@quickstart ~]$ 
```

APACHE PIG-LEFT JOIN

- Left join: It is a join type. Returns all records from the left table. It returns only matching records from the table on the right.



APACHE PIG-LEFT JOIN APPLICATION

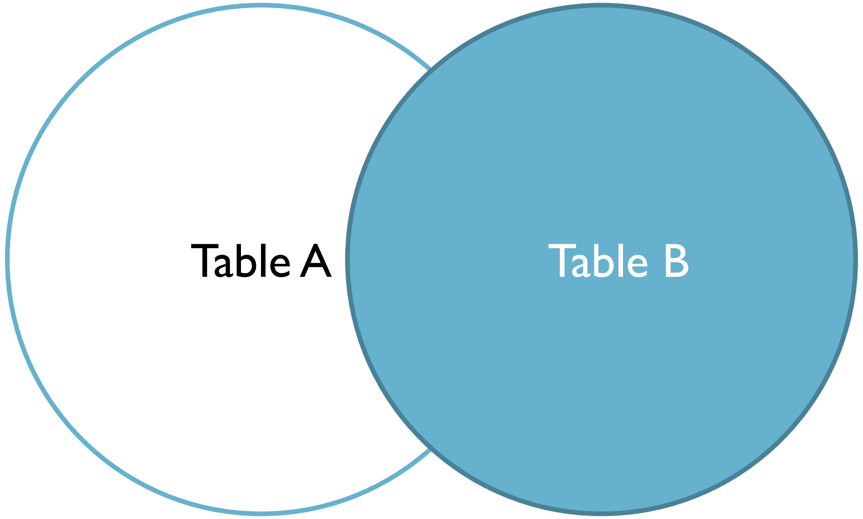

Left_join.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Left_join.pig
Plain Text
Tab Width: 8
Ln 10, Col 9
INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Staff = LOAD '/example/Staff.csv' USING PigStorage(',') AS
(
name: chararray,
age: int,
dept_id: int
);
Salary = LOAD '/example/Salary.csv' USING PigStorage(',') AS
(
name: chararray,
pay: int
);

New_Data = JOIN Staff BY name LEFT OUTER, Salary BY name;

DUMP New_Data;
```

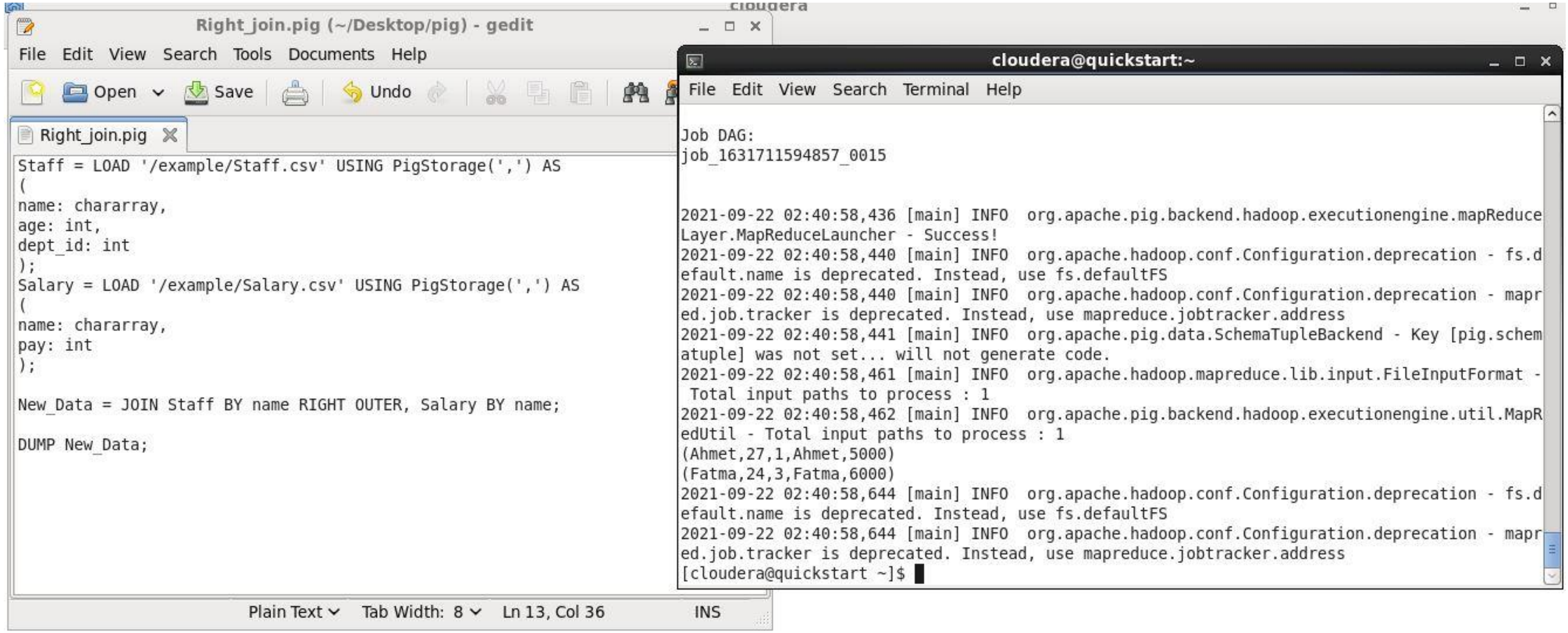
```
2021-09-22 02:31:55,502 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MapReduceLauncher - Success!
2021-09-22 02:31:55,511 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:31:55,511 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-22 02:31:55,512 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schem
atuple] was not set... will not generate code.
2021-09-22 02:31:55,534 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
 Total input paths to process : 1
2021-09-22 02:31:55,534 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapR
edUtil - Total input paths to process : 1
(Cenk,34,3,,)
(Seda,26,2,,)
(Ahmet,27,1,Ahmet,5000)
(Burak,29,2,,)
(Fatma,24,3,Fatma,6000)
(Peter,30,1,,)
(Mehmet,35,2,,)
2021-09-22 02:31:55,683 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:31:55,683 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
[cloudera@quickstart ~]$
```

APACHE PIG-RIGHT JOIN

- Right join: It is a join type. Returns all records from the right table. It returns only matching records from the table on the left.



APACHE PIG - RIGHT JOIN APPLICATION



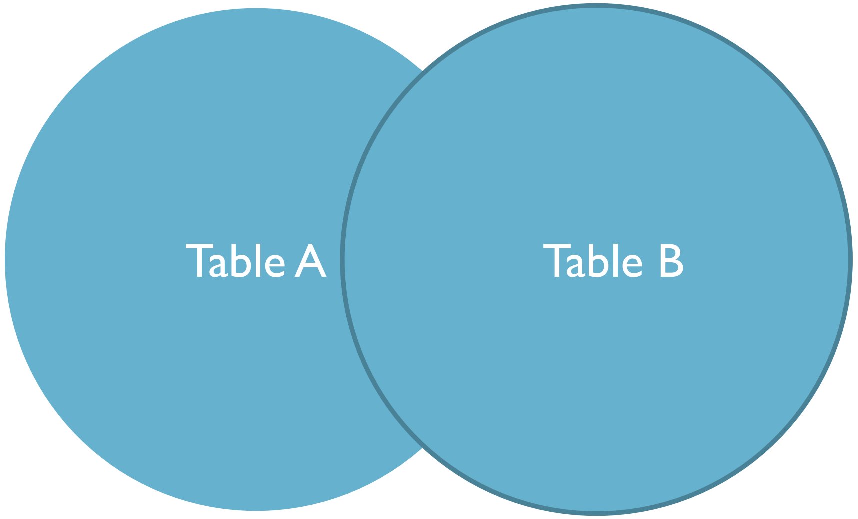
cloudera
Right_join.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
Right_join.pig
Plain Text Tab Width: 8 Ln 13, Col 36 INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Staff = LOAD '/example/Staff.csv' USING PigStorage(',') AS
(
name: chararray,
age: int,
dept_id: int
);
Salary = LOAD '/example/Salary.csv' USING PigStorage(',') AS
(
name: chararray,
pay: int
);

New_Data = JOIN Staff BY name RIGHT OUTER, Salary BY name;

DUMP New_Data;
```

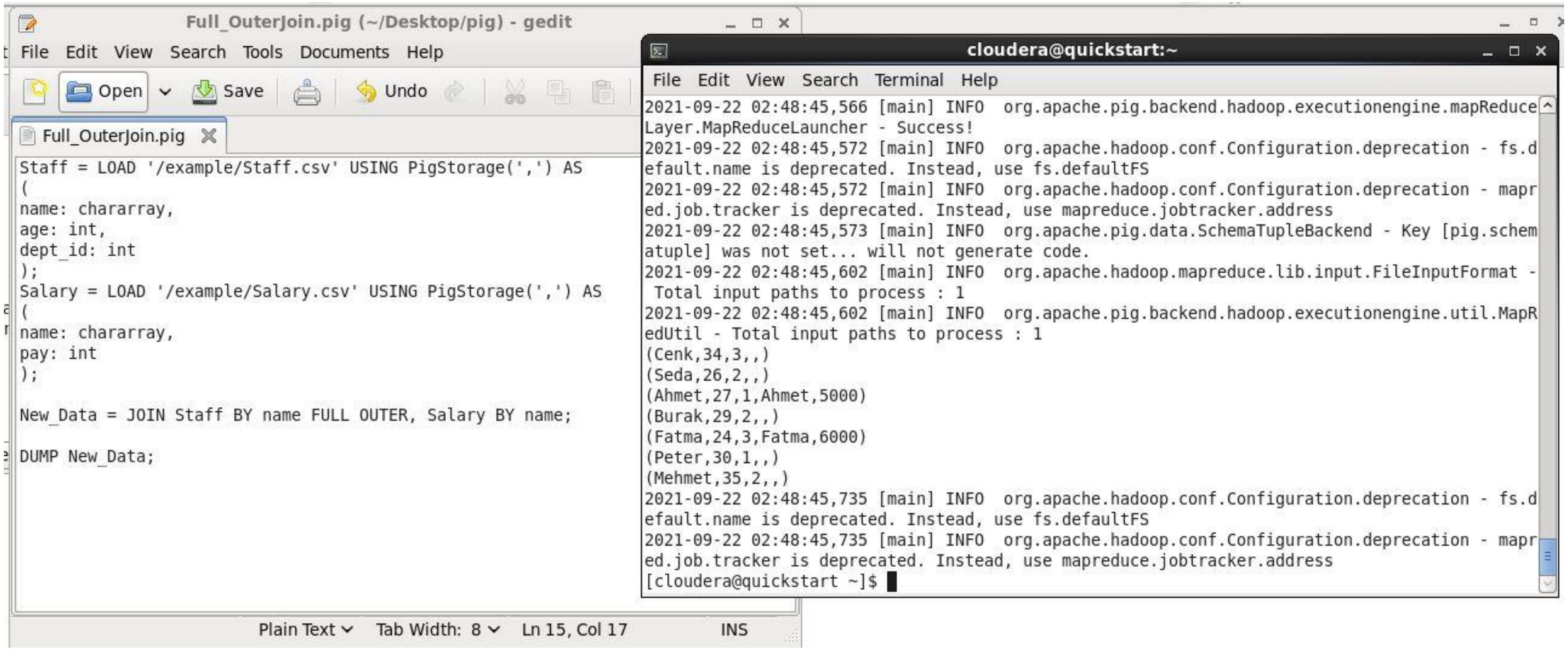
```
Job DAG:
job_1631711594857_0015


2021-09-22 02:40:58,436 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MapReduceLauncher - Success!
2021-09-22 02:40:58,440 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:40:58,440 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-22 02:40:58,441 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schem
atuple] was not set... will not generate code.
2021-09-22 02:40:58,461 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
 Total input paths to process : 1
2021-09-22 02:40:58,462 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapR
edUtil - Total input paths to process : 1
(Ahmet,27,1,Ahmet,5000)
(Fatma,24,3,Fatma,6000)
2021-09-22 02:40:58,644 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:40:58,644 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
[cloudera@quickstart ~]$
```

APACHE PIG-FULL OUTER JOIN

- Full outer join: It is a join type. It returns both matching records and all records to the right and left.



APACHE PIG – FULL OUTER JOIN APPLICATION

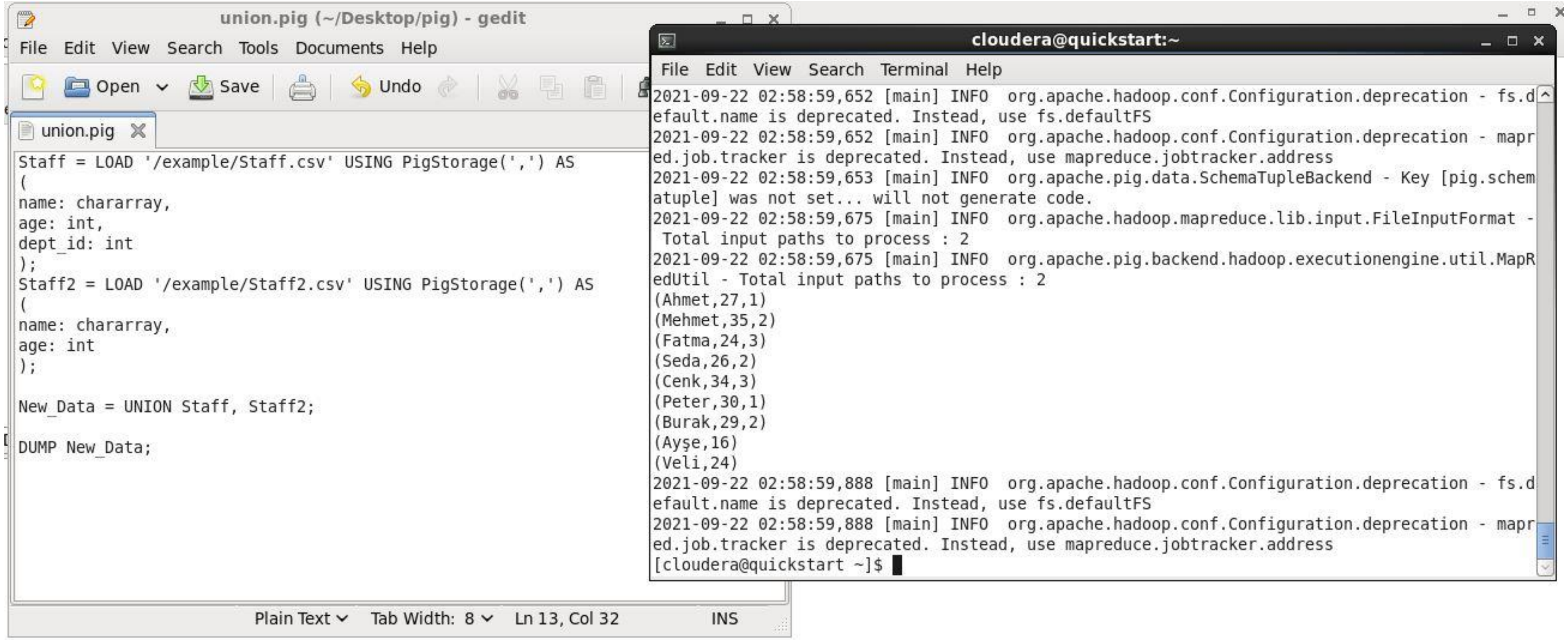

Full_OuterJoin.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Full_OuterJoin.pig
Plain Text
Tab Width: 8
Ln 15, Col 17
INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Staff = LOAD '/example/Staff.csv' USING PigStorage(',') AS
(
name: chararray,
age: int,
dept_id: int
);
Salary = LOAD '/example/Salary.csv' USING PigStorage(',') AS
(
name: chararray,
pay: int
);

New_Data = JOIN Staff BY name FULL OUTER, Salary BY name;

DUMP New_Data;
```

```
2021-09-22 02:48:45,566 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MapReduceLauncher - Success!
2021-09-22 02:48:45,572 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:48:45,572 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-22 02:48:45,573 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schem
atuple] was not set... will not generate code.
2021-09-22 02:48:45,602 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
 Total input paths to process : 1
2021-09-22 02:48:45,602 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapR
edUtil - Total input paths to process : 1
(Cenk,34,3,,)
(Seda,26,2,,)
(Ahmet,27,1,Ahmet,5000)
(Burak,29,2,,)
(Fatma,24,3,Fatma,6000)
(Peter,30,1,,)
(Mehmet,35,2,,)
2021-09-22 02:48:45,735 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:48:45,735 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
[cloudera@quickstart ~]$
```

APACHE PIG –UNION

- Combines two different datasets.

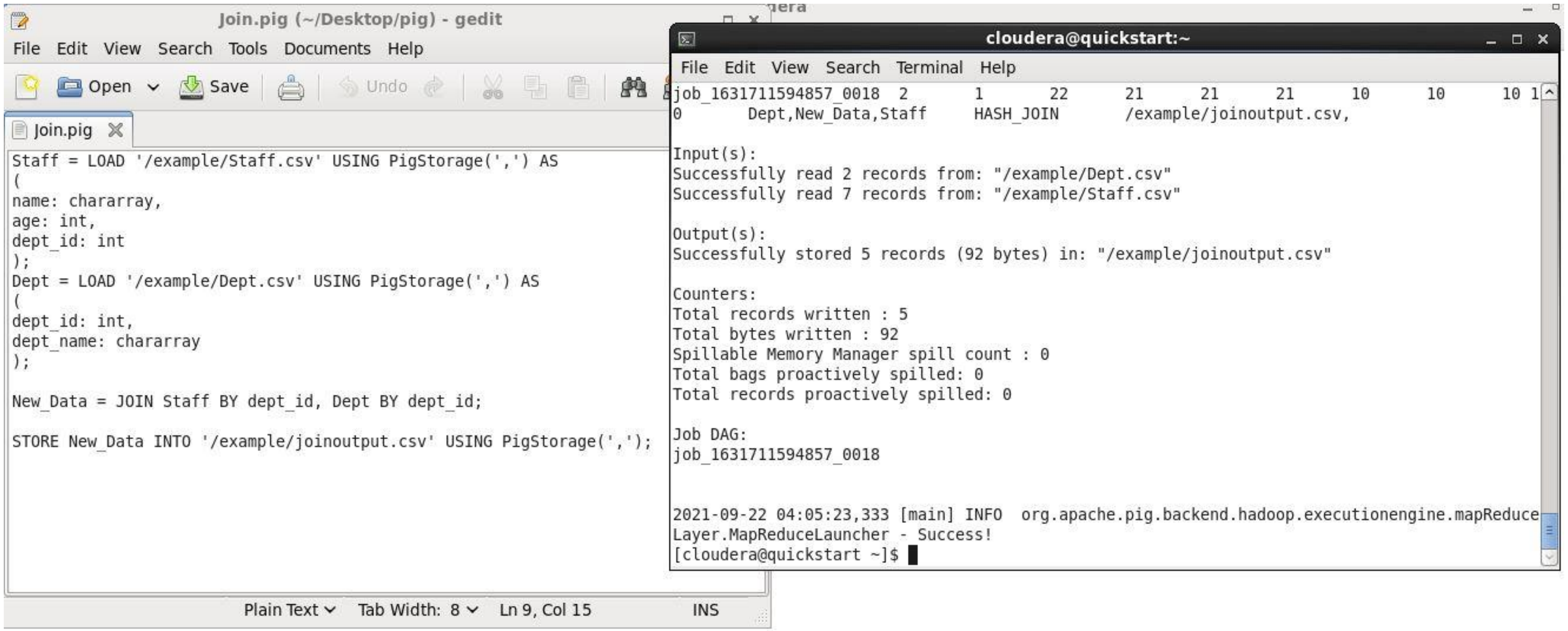


union.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
union.pig
Plain Text
Tab Width: 8
Ln 13, Col 32
INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Staff = LOAD '/example/Staff.csv' USING PigStorage(',') AS
(
name: chararray,
age: int,
dept_id: int
);
Staff2 = LOAD '/example/Staff2.csv' USING PigStorage(',') AS
(
name: chararray,
age: int
);

New_Data = UNION Staff, Staff2;

DUMP New_Data;
```

```
2021-09-22 02:58:59,652 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:58:59,652 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-09-22 02:58:59,653 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schem
atuple] was not set... will not generate code.
2021-09-22 02:58:59,675 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
 Total input paths to process : 2
2021-09-22 02:58:59,675 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapR
edUtil - Total input paths to process : 2
(Ahmet,27,1)
(Mehmet,35,2)
(Fatma,24,3)
(Seda,26,2)
(Cenk,34,3)
(Peter,30,1)
(Burak,29,2)
(Ayşe,16)
(Veli,24)
2021-09-22 02:58:59,888 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
2021-09-22 02:58:59,888 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapr
ed.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
[cloudera@quickstart ~]$
```

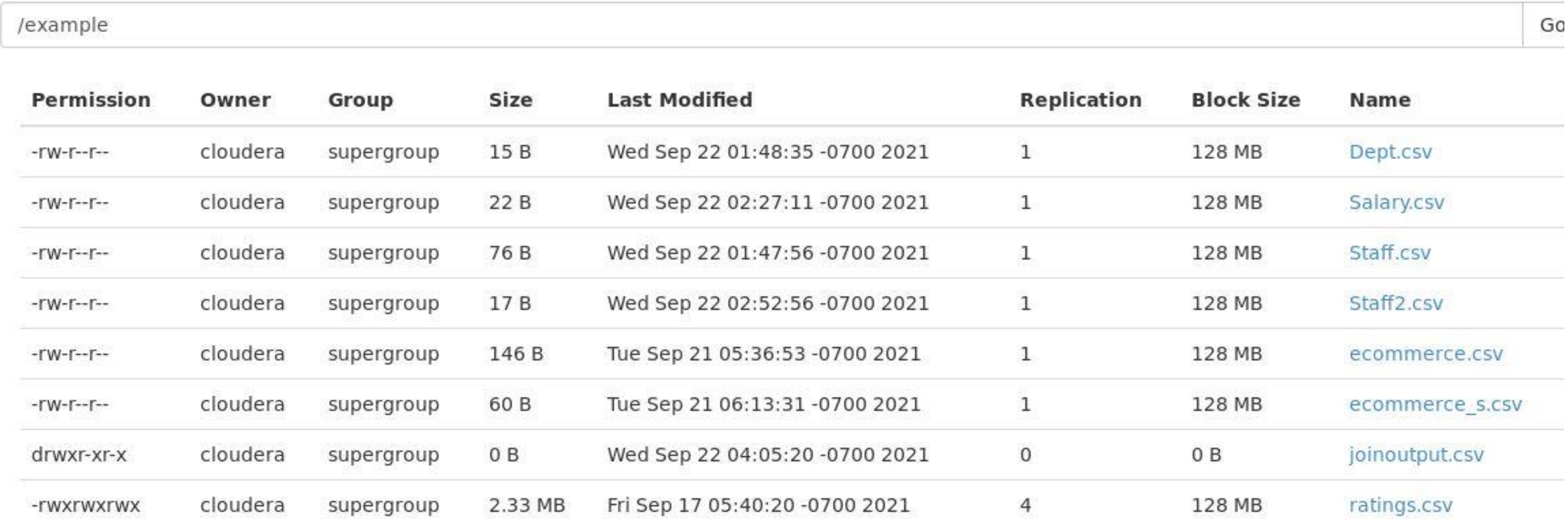
APACHE PIG- STORE OPERATOR


Join.pig (~/Desktop/pig) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
Join.pig
Plain Text
Tab Width: 8
Ln 9, Col 15
INS
cloudera@quickstart:~
File Edit View Search Terminal Help

```
Staff = LOAD '/example/Staff.csv' USING PigStorage(',') AS
(
name: chararray,
age: int,
dept_id: int
);
Dept = LOAD '/example/Dept.csv' USING PigStorage(',') AS
(
dept_id: int,
dept_name: chararray
);

New_Data = JOIN Staff BY dept_id, Dept BY dept_id;

STORE New_Data INTO '/example/joinoutput.csv' USING PigStorage(',');
```

```
job_1631711594857_0018  2       1       22      21      21      21      10      10      10 1
0       Dept,New_Data,Staff     HASH_JOIN       /example/joinoutput.csv,

Input(s):
Successfully read 2 records from: "/example/Dept.csv"
Successfully read 7 records from: "/example/Staff.csv"

Output(s):
Successfully stored 5 records (92 bytes) in: "/example/joinoutput.csv"

Counters:
Total records written : 5
Total bytes written : 92
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1631711594857_0018


2021-09-22 04:05:23,333 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduce
Layer.MapReduceLauncher - Success!
[cloudera@quickstart ~]$
```

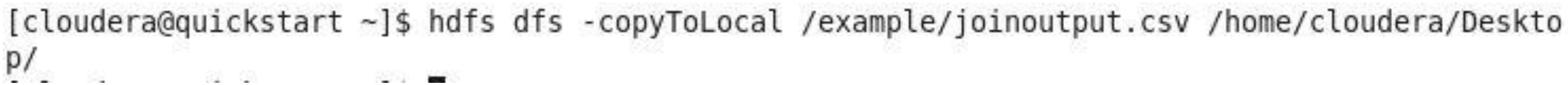
APACHE PIG- STORE OPERATOR



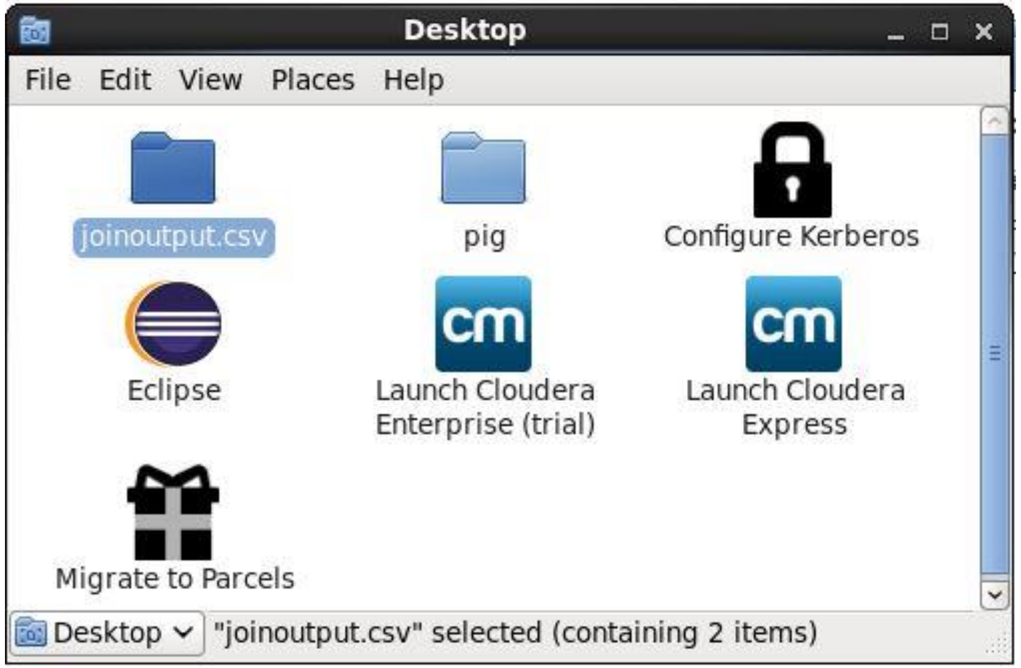
/example

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	supergroup	15 B	Wed Sep 22 01:48:35 -0700 2021	1	128 MB	Dept.csv
-rw-r--r--	cloudera	supergroup	22 B	Wed Sep 22 02:27:11 -0700 2021	1	128 MB	Salary.csv
-rw-r--r--	cloudera	supergroup	76 B	Wed Sep 22 01:47:56 -0700 2021	1	128 MB	Staff.csv
-rw-r--r--	cloudera	supergroup	17 B	Wed Sep 22 02:52:56 -0700 2021	1	128 MB	Staff2.csv
-rw-r--r--	cloudera	supergroup	146 B	Tue Sep 21 05:36:53 -0700 2021	1	128 MB	ecommerce.csv
-rw-r--r--	cloudera	supergroup	60 B	Tue Sep 21 06:13:31 -0700 2021	1	128 MB	ecommerce_s.csv
drwxr-xr-x	cloudera	supergroup	0 B	Wed Sep 22 04:05:20 -0700 2021	0	0 B	joinoutput.csv
-rwxrwxrwx	cloudera	supergroup	2.33 MB	Fri Sep 17 05:40:20 -0700 2021	4	128 MB	ratings.csv

APACHE PIG-STORE



```
[cloudera@quickstart ~]$ hdfs dfs -copyToLocal /example/joinoutput.csv /home/cloudera/Deskto
p/
```



"joinoutput.csv" selected (containing 2 items)

MONGODB

- MongoDB is a document-oriented Nosql database.
- In MongoDB, each record is a document.
- Documents are stored in Binary JSON(BSN) format.
- MongoDB supports real-time analytics with a wide variety of data.

MONGO DB & RELATIONAL DATABASE

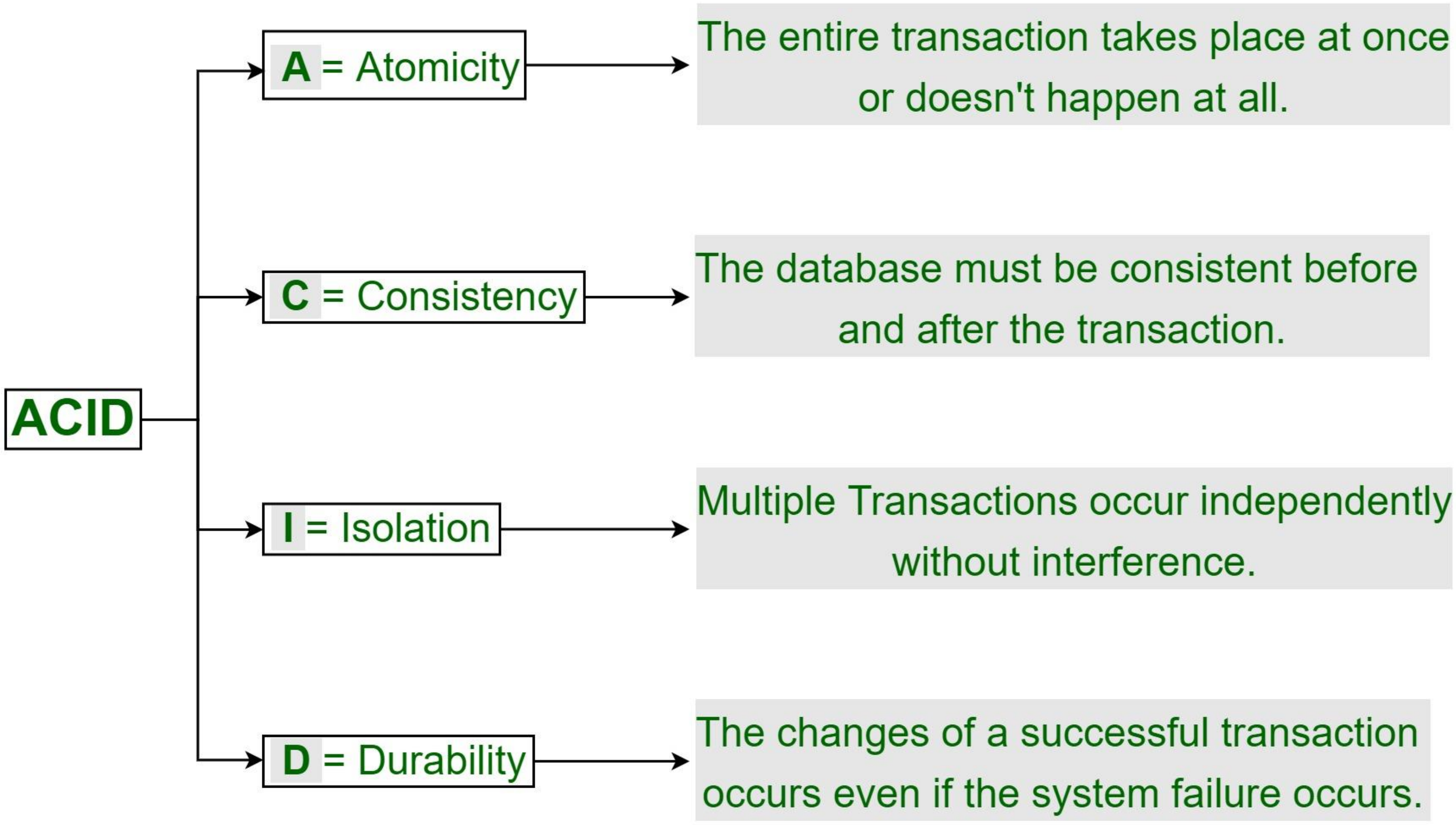
RDBMS

- It is a relational database.
- Not suitable for hierarchical data storage.
- It is vertically scalable i.e increasing RAM.
- It has a predefined schema.
- It is quite vulnerable to SQL injection.
- It centers around ACID properties (Atomicity, Consistency, Isolation, and Durability).

MongoDB

- It is a non-relational and document-oriented database.
- Suitable for hierarchical data storage.
- It is horizontally scalable i.e we can add more servers.
- It has a dynamic schema.
- It is not affected by SQL injection.
- It centers around the CAP theorem (Consistency, Availability, and Partition tolerance).

ACID Properties in DBMS



ACID

A = Atomicity

The entire transaction takes place at once or doesn't happen at all.

C = Consistency

The database must be consistent before and after the transaction.

I = Isolation

Multiple Transactions occur independently without interference.

D = Durability

The changes of a successful transaction occurs even if the system failure occurs.

CAP THEOREM

- Consistency
- Availability
- Partition tolerance

MONGO DB & RELATIONAL DATABASE

RDBMS

- It is row-based.
- It is slower in comparison with MongoDB.
- Supports complex joins.
- It is column-based.
- It does not provide JavaScript client for querying.
- It supports SQL query language only.

MongoDB

- It is document-based.
- It is almost 100 times faster than RDBMS
- No support for complex joins.
- It is field-based.
- It provides a JavaScript client for querying.
- It supports JSON query language along with SQL.

WHY MONGODB?

- Its flexible schema makes it easy to evolve and store data in a way that is easy for programmers to work with.
- MongoDB is also built to scale up quickly.
- MongoDB supports all the main features of modern databases such as transactions.
- MongoDB has a large community of users that can provide help, and enterprise-level support is available.

WHY MONGODB?

- Multiple copies of the data can be stored and no data loss.
- MongoDB allows cluster structure.

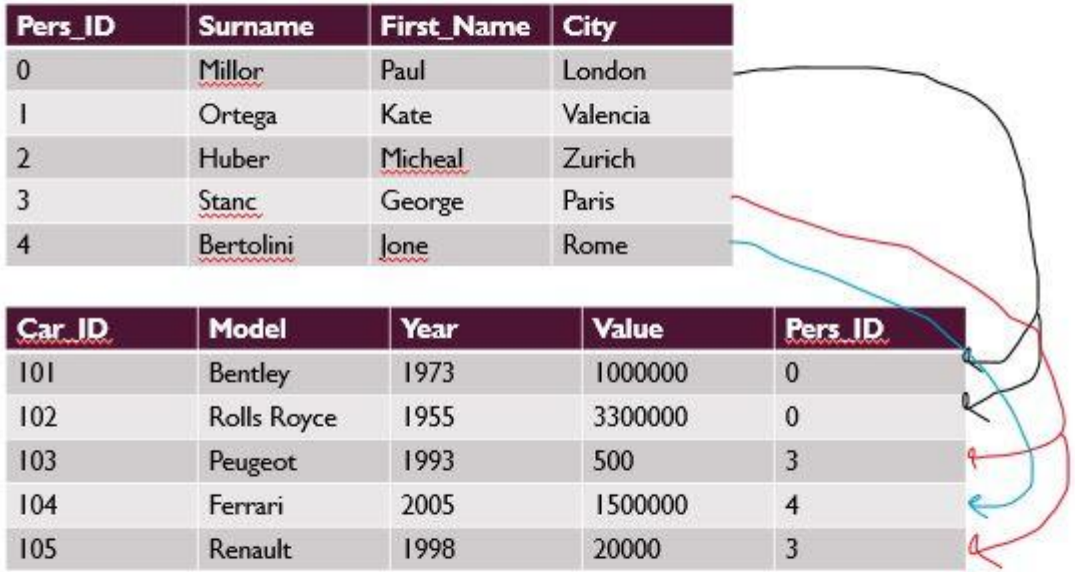
SAMPLE

Pers_ID	Surname	First_Name	City
0	Millor	Paul	London
1	Ortega	Kate	Valencia
2	Huber	Micheal	Zurich
3	Stanc	George	Paris
4	Bertolini	Jone	Rome

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	1000000	0
102	Rolls Royce	1955	3300000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	1500000	4
105	Renault	1998	20000	3

SAMPLE

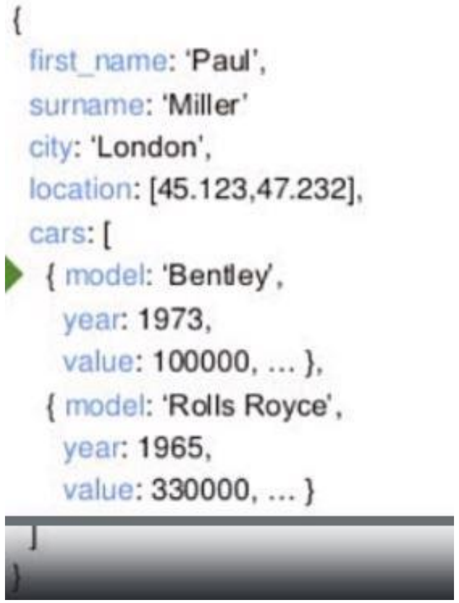
RDBMS



Pers_ID	Surname	First_Name	City
0	Millor	Paul	London
1	Ortega	Kate	Valencia
2	Huber	Micheal	Zurich
3	Stanc	George	Paris
4	Bertolini	Ione	Rome

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	1000000	0
102	Rolls Royce	1955	3300000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	1500000	4
105	Renault	1998	20000	3

MongoDB



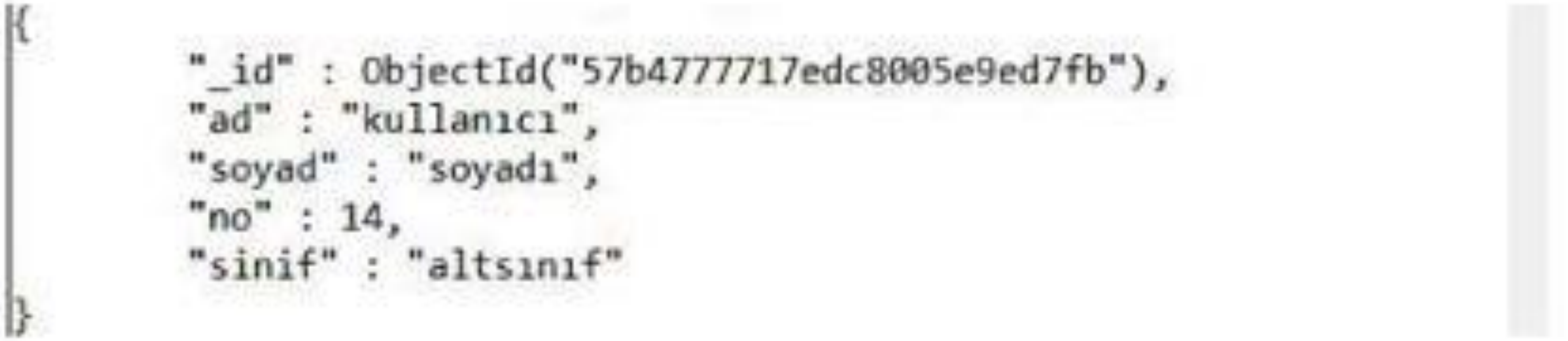
```
{
 first_name: 'Paul',
 surname: 'Miller'
 city: 'London',
 location: [45.123,47.232],
 cars: [
  { model: 'Bentley',
    year: 1973,
    value: 100000, ... },
  { model: 'Rolls Royce',
    year: 1965,
    value: 330000, ... }
 ]
}
```

MONGODB CONCEPTS

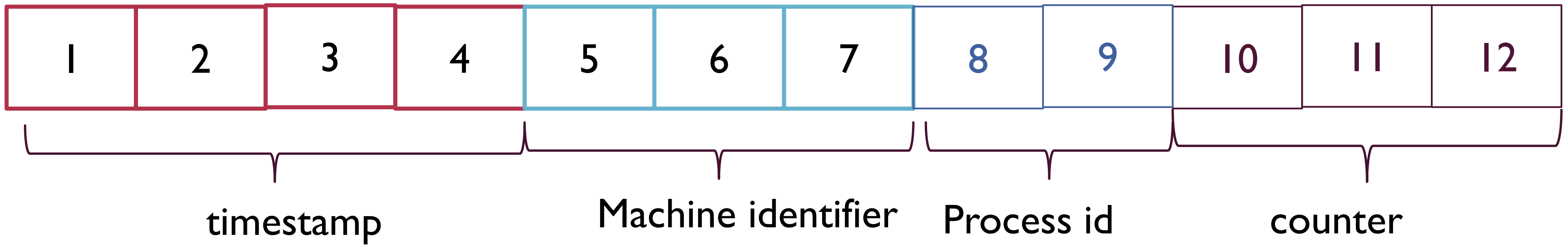
RDBMS	MongoDB
Table, View	Collection
Row	Document
Index	Index
Join	Embedded Document
Foreign Key	Reference
Partition	Shard

MONGODB ID INFORMATION

- When inserting a record on MongoDB, a field named _id is automatically added.
- This field can be entered by the user.
- If it is not entered by the user, it is saved with a unique value.



```
{
        "_id" : ObjectId("57b4777717edc8005e9ed7fb"),
        "ad" : "kullanıcı",
        "soyad" : "soyadı",
        "no" : 14,
        "sinif" : "altsınıf"
}
```



1 2 3 4 5 6 7 8 9 10 11 12

timestamp

Machine identifier

Process id

counter