BIG DATA ANALYTICS

ASSIGNMENT 6: Hands on over Apache PIG.

(Amitabh Saini, TY-E-4, 141541)

* **Introduction**

Apache Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyse larger sets of data representing them as data flows. Pig is generally used with Hadoop; we can perform all the data manipulation operations in Hadoop using Pig.

* **Architecture**



The language used to analyse data in Hadoop using Pig is known as Pig Latin.

**Parser**

Initially the Pig Scripts are handled by the Parser. It checks the syntax of the script, does type checking, and other miscellaneous checks. The output of the parser will be a DAG (directed acyclic graph), which represents the Pig Latin statements and logical operators.

In the DAG, the logical operators of the script are represented as the nodes and the data flows are represented as edges.

**Optimizer**

The logical plan (DAG) is passed to the logical optimizer, which carries out the logical optimizations such as projection and pushdown.

**Compiler**

The compiler compiles the optimized logical plan into a series of MapReduce jobs.

**Execution engine**

Finally the MapReduce jobs are submitted to Hadoop in a sorted order. Finally, these MapReduce jobs are executed on Hadoop producing the desired results.

* **PIG Data Model**

**Field**

It is stored as string and can be used as string and number. int, long, float, double, chararray, and bytearray are the atomic values of Pig. A piece of data or a simple atomic value is known as a field.

Eg. ‘Amitabh’,’4’

**Tuple**

A record that is formed by an ordered set of fields is known as a tuple, the fields can be of any type. A tuple is similar to a row in a table of RDBMS.

Eg. (Amitabh, 4)

**Bag**

A bag is an unordered set of tuples. In other words, a collection of tuples (non-unique) is known as a bag. Each tuple can have any number of fields (flexible schema). A bag is represented by ‘{}’. It is similar to a table in RDBMS, but unlike a table in RDBMS, it is not necessary that every tuple contain the same number of fields or that the fields in the same position (column) have the same type.

Eg. {(Amitabh, 4), (Rahul, 5)}

**Map**

A map (or data map) is a set of key-value pairs. The key needs to be of type chararray and should be unique. The value might be of any type. It is represented by ‘[]’.

Eg. [name#Amitabh,rollno#4]

* **Hands on PIG**
* Starting Pig

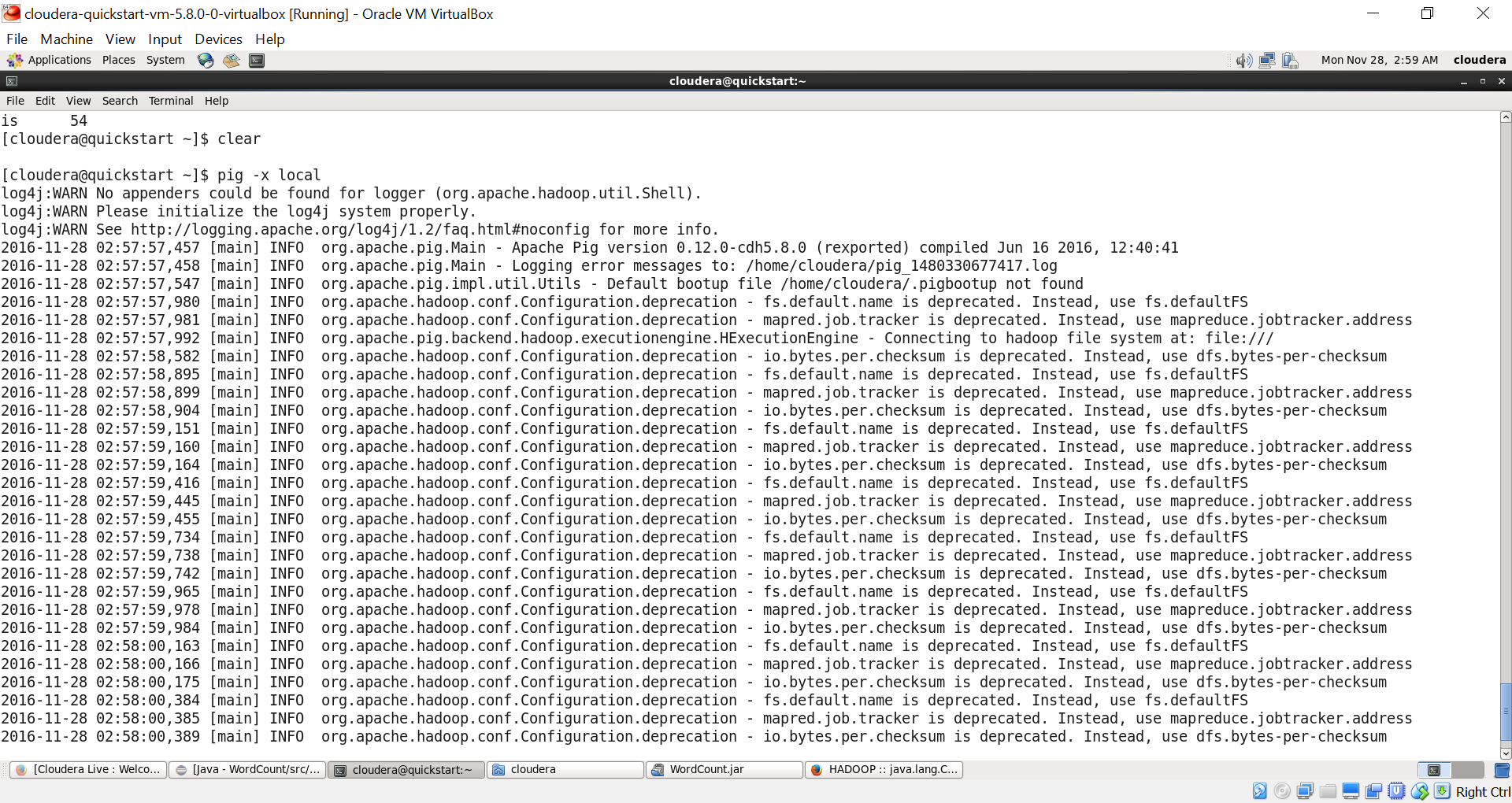
Pig can run in two modes local and map-reduce.

For Local – “pig –x local”

For MapReduce- “pig –x mapred”

This commands starts grunt shell used by pig.

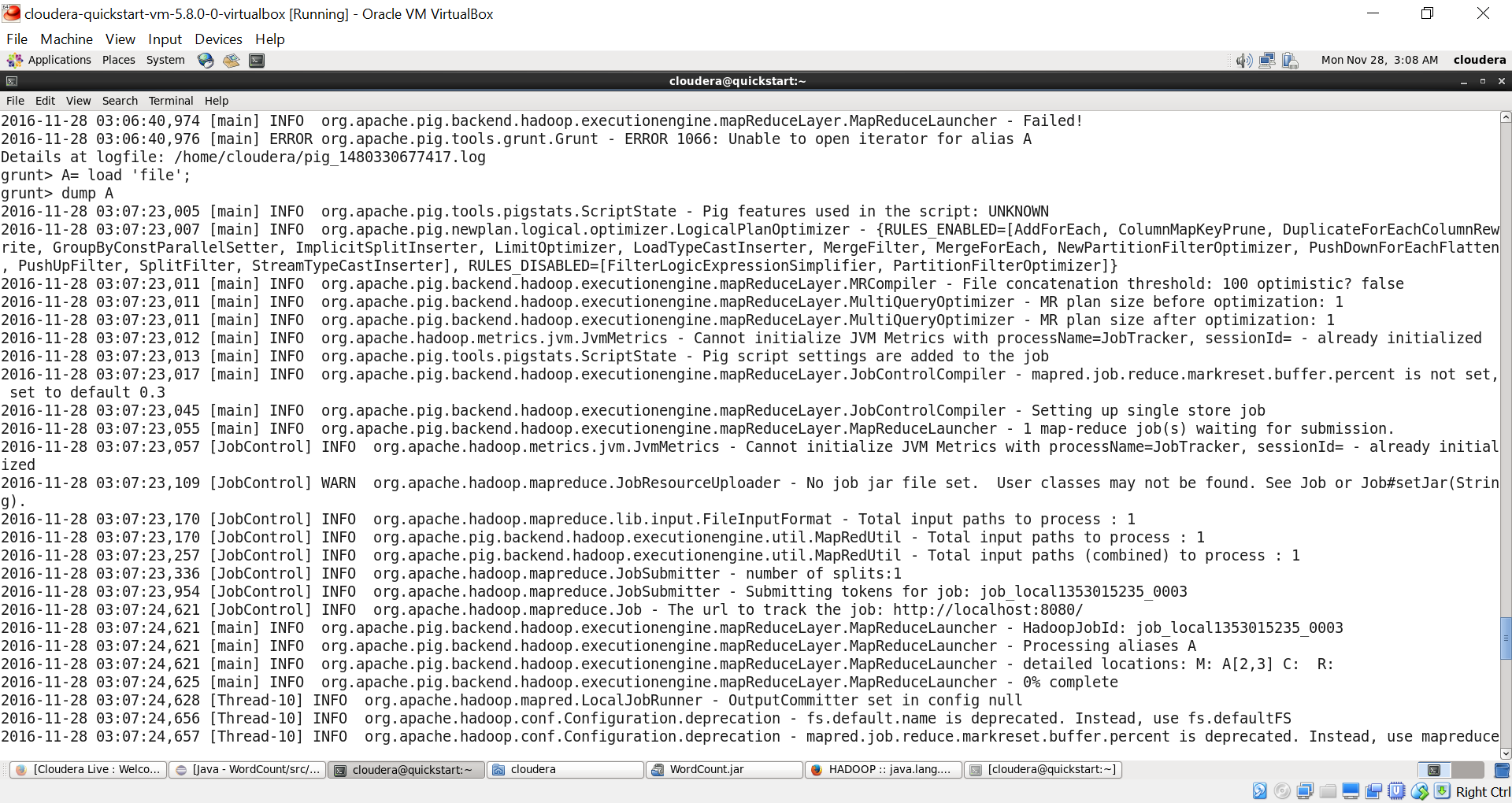
Further we will use local mode of pig.

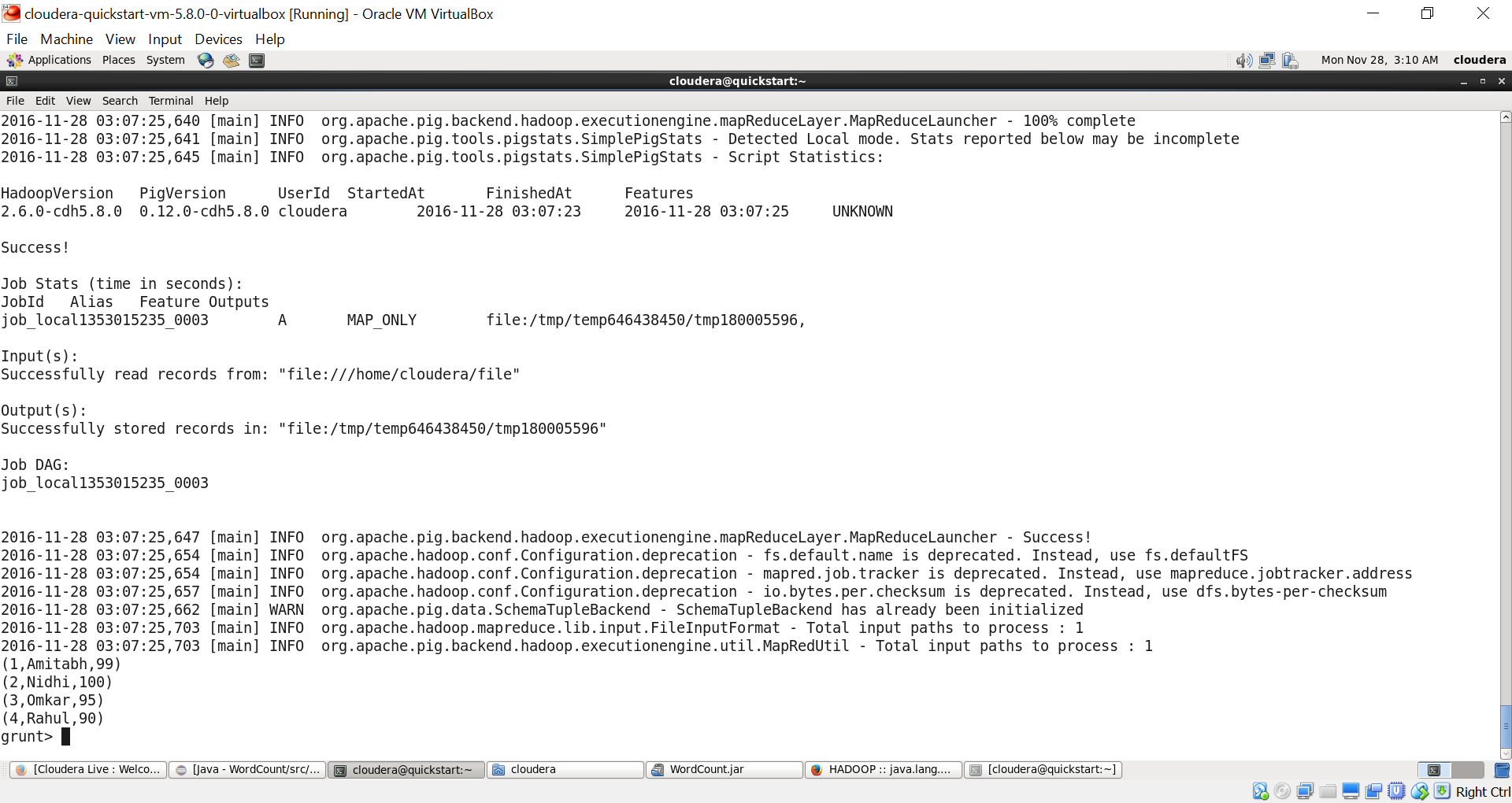


* Load and Dump

Loading file 🡪 A=load ‘<filepath>’;

Dump(See) 🡪 dump A;

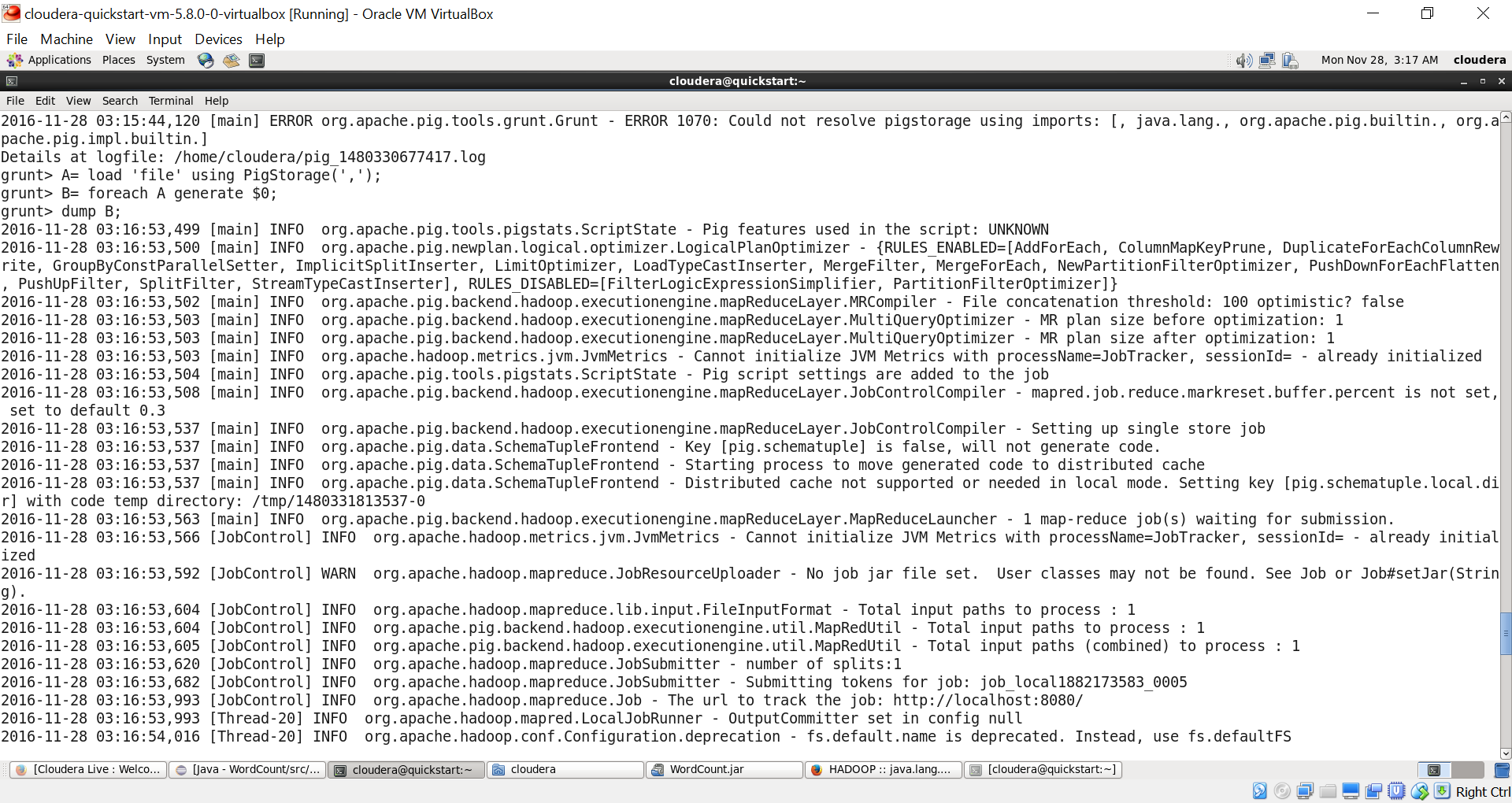


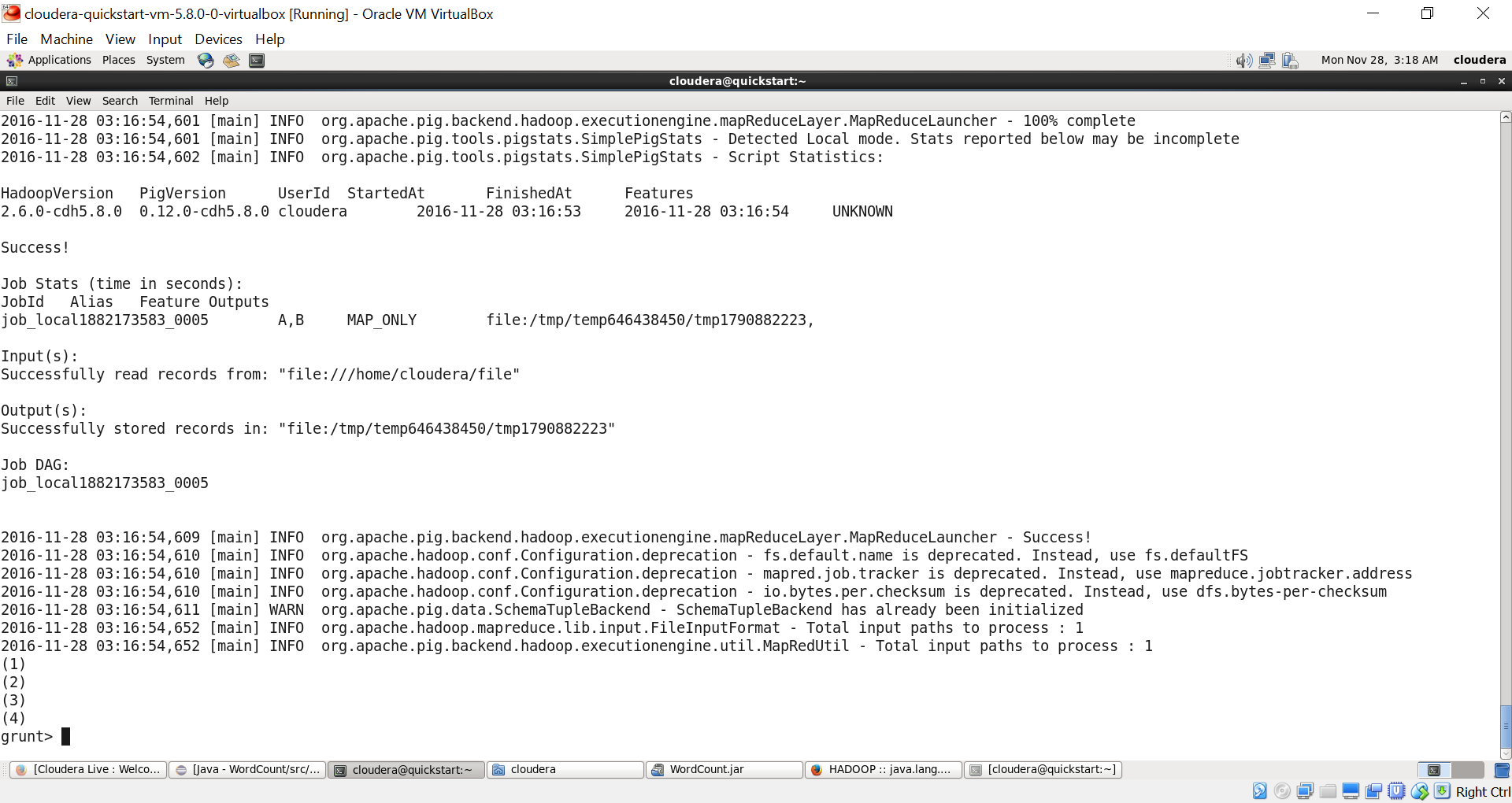


* Access fields

Load using PigStorage to separate fields by delimiter.

Eg. Accessing first fied in A 🡪 ‘FOREACH A GENERATE $0;’

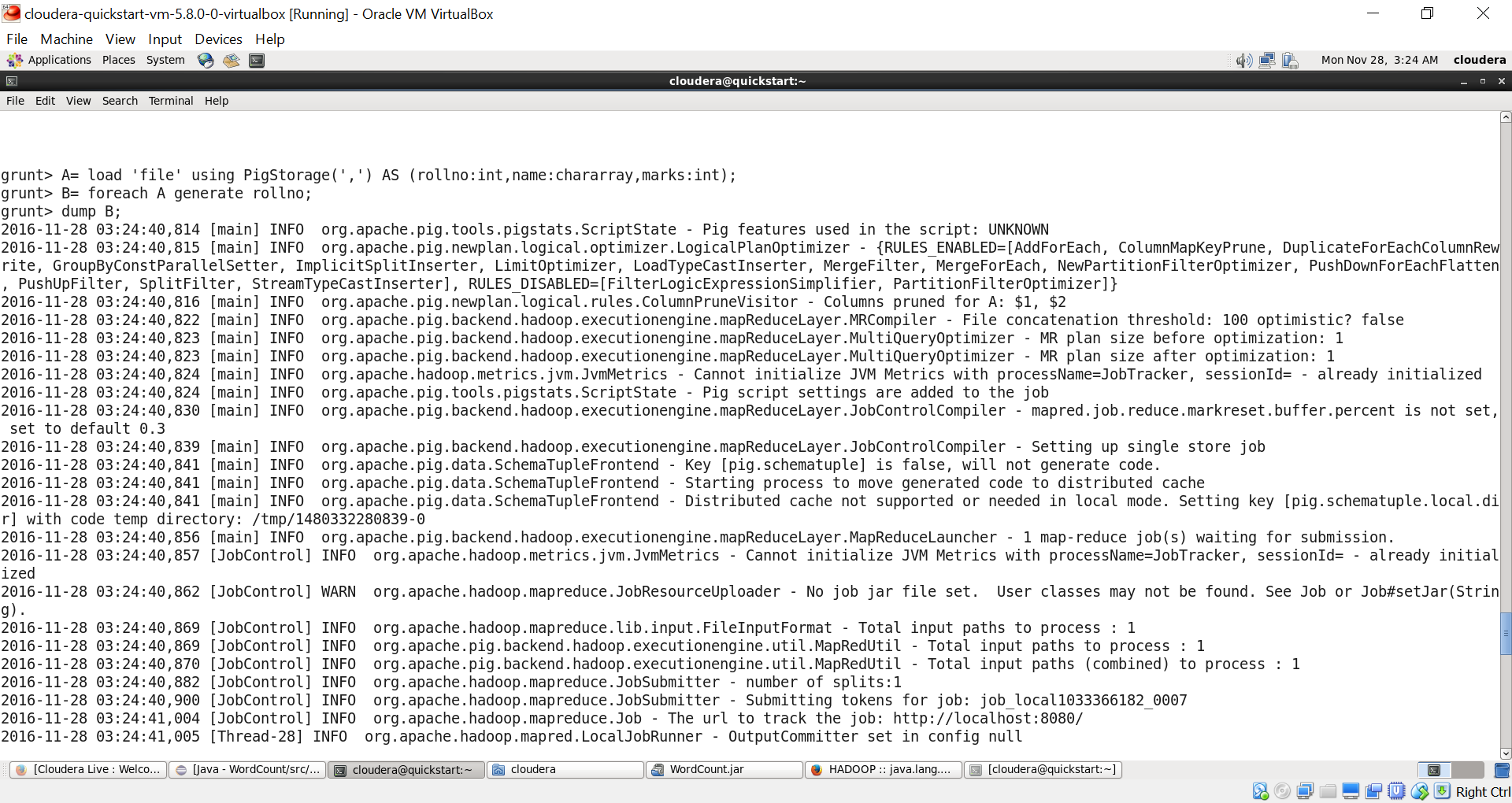


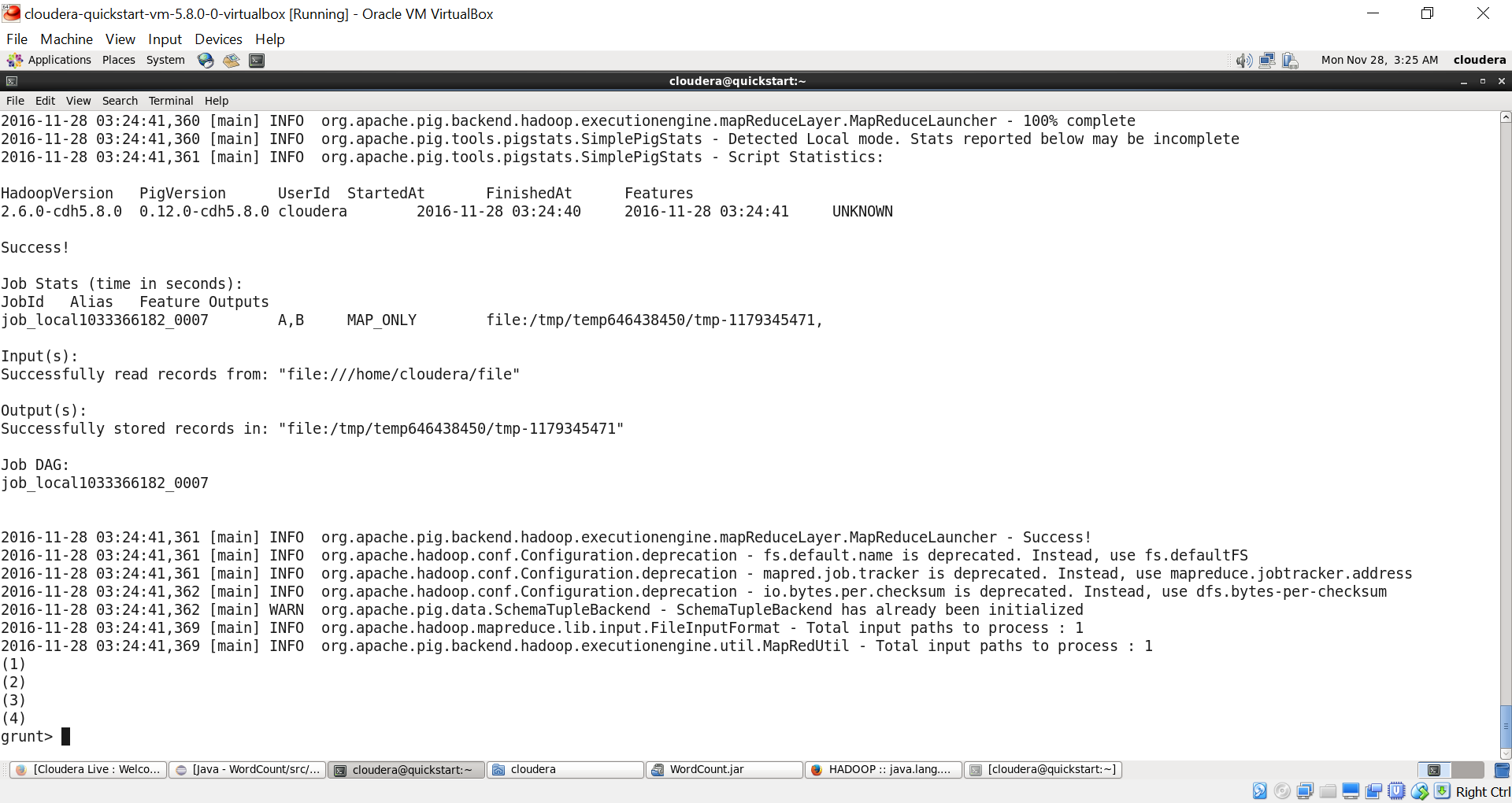


* Giving column names

A=load ‘file’ using PigStorage(‘,’) AS (rollno:int,name:CharArray,marks:int);

B=foreach A generate rollno;





* Order By and Limit

B=order A by marks;

C=Limit B 3;

Order by o/p:

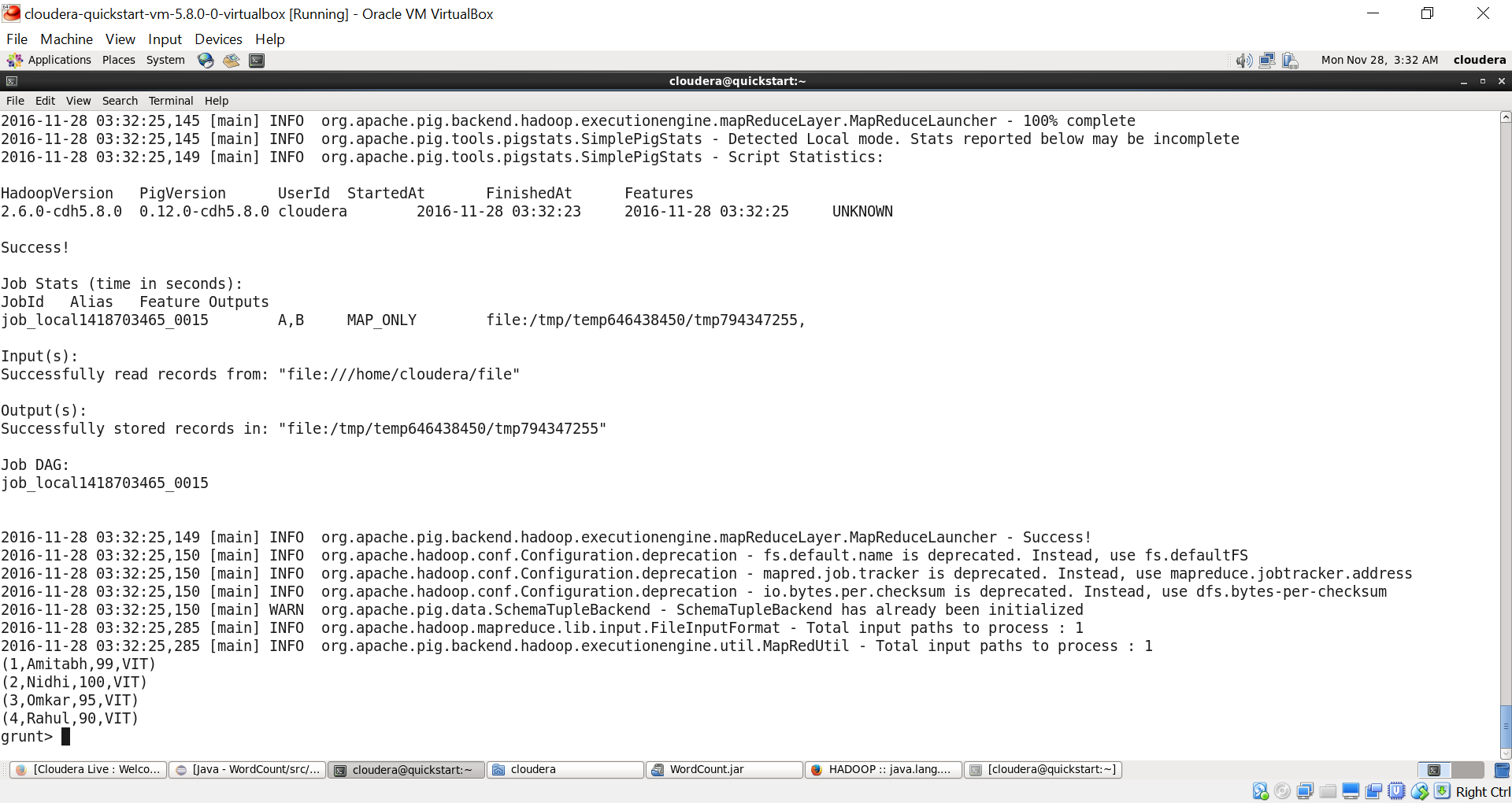


Limit o/p:



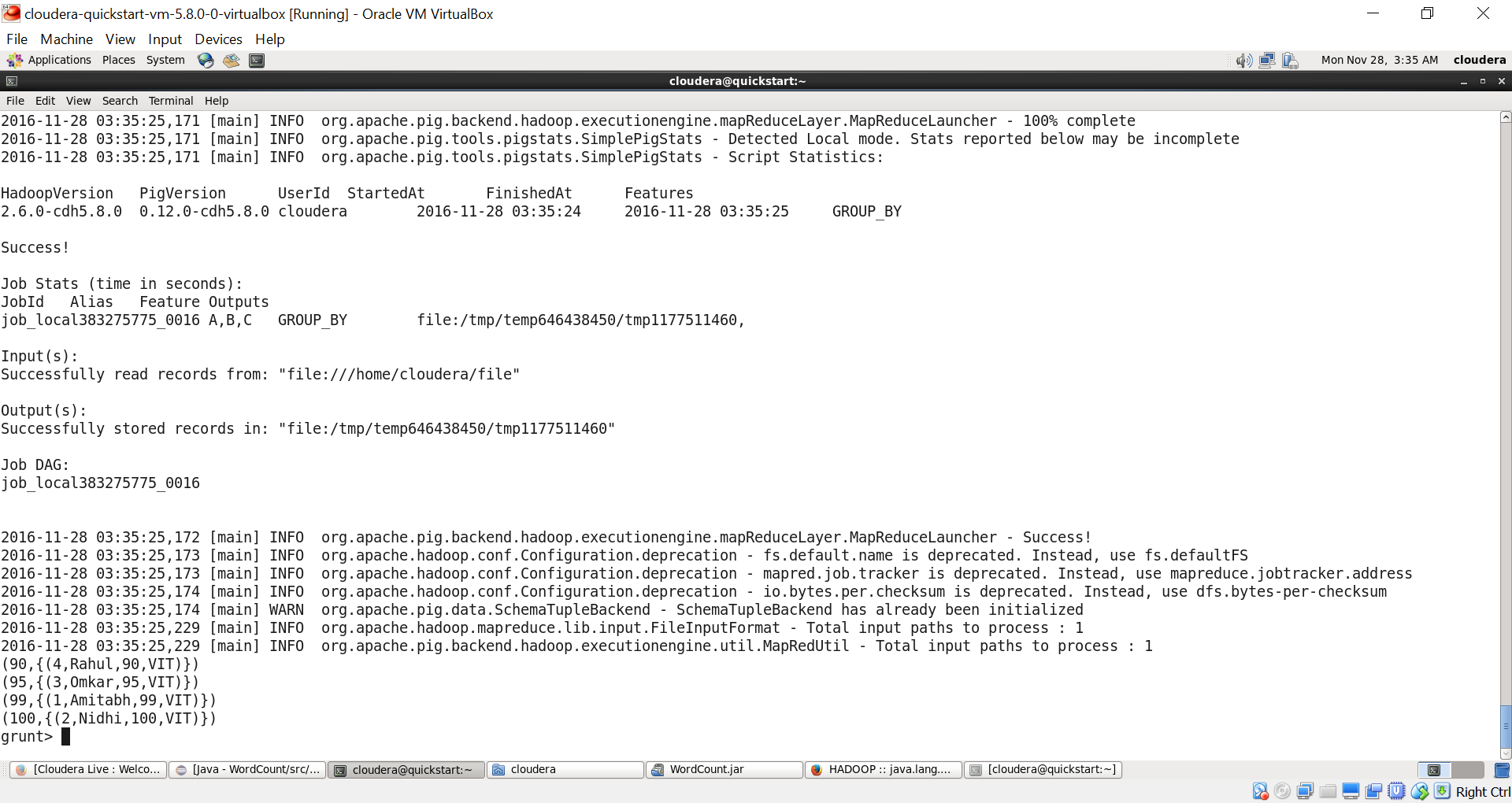
* Adding column

B=foreach A generate rollno AS rollno,name AS name,marks AS marks,’VIT’ AS college.

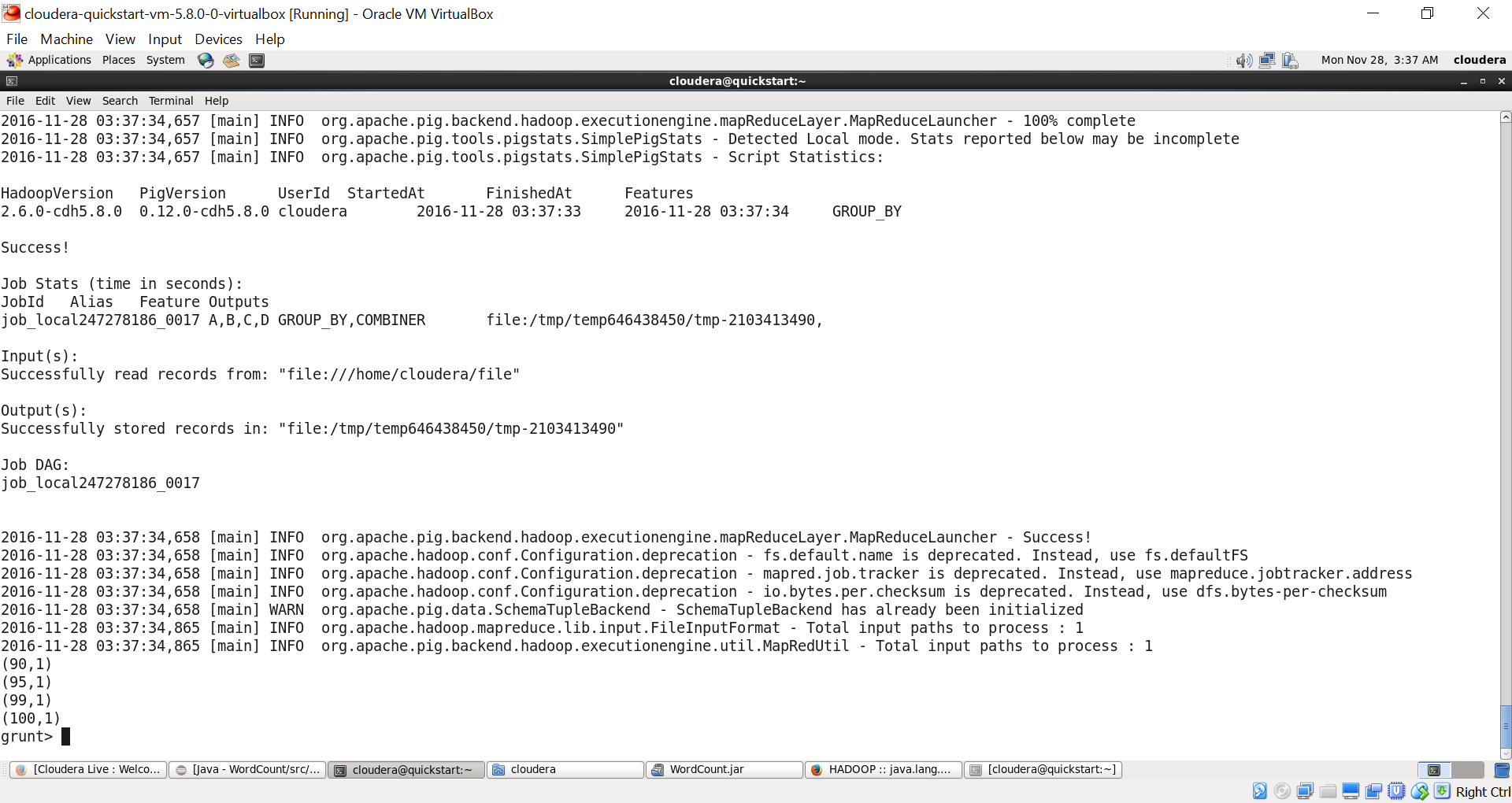


* Grouping

C= group B by marks;

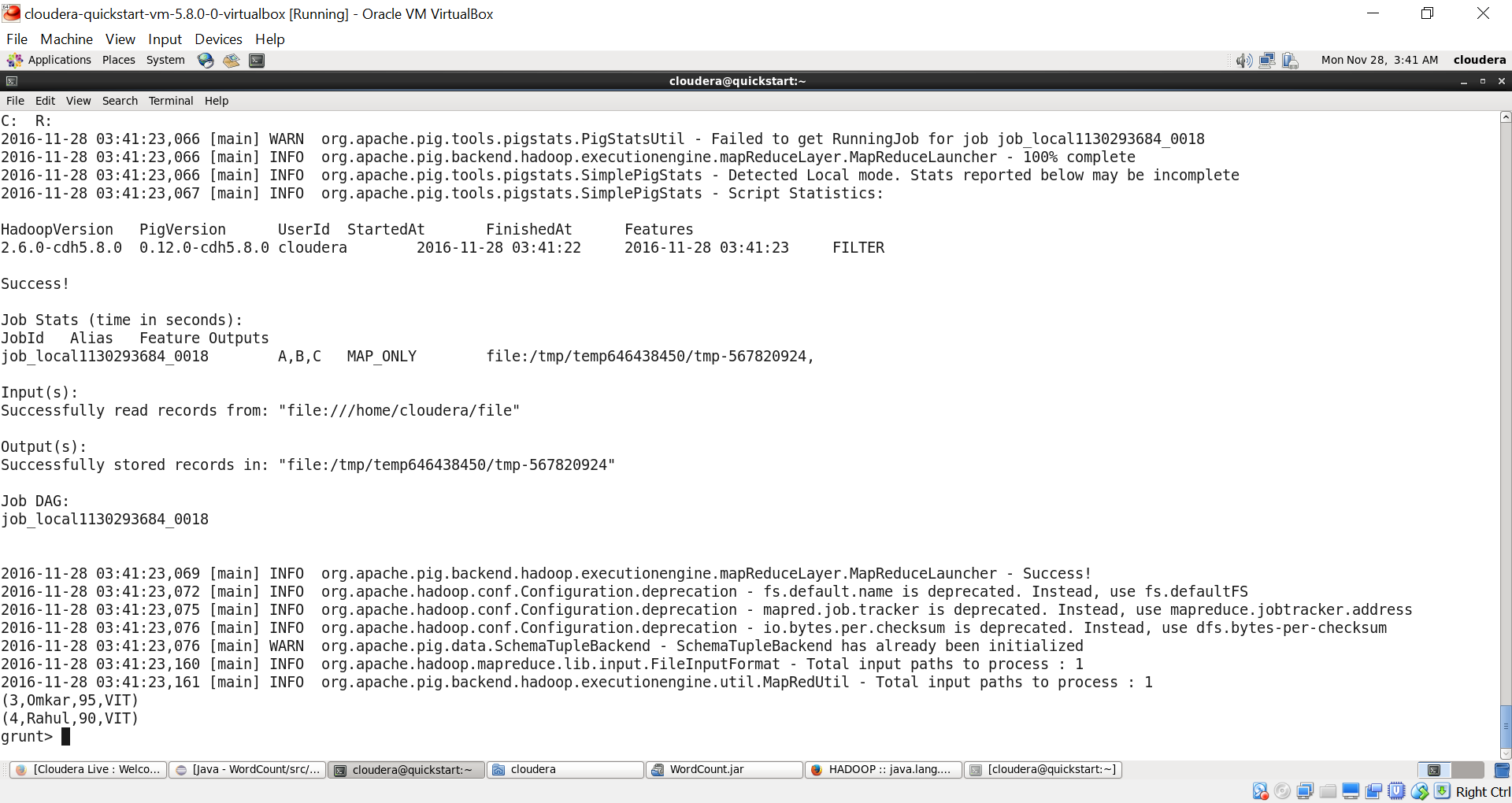


D= foreach C generate group,count(B.rollno);



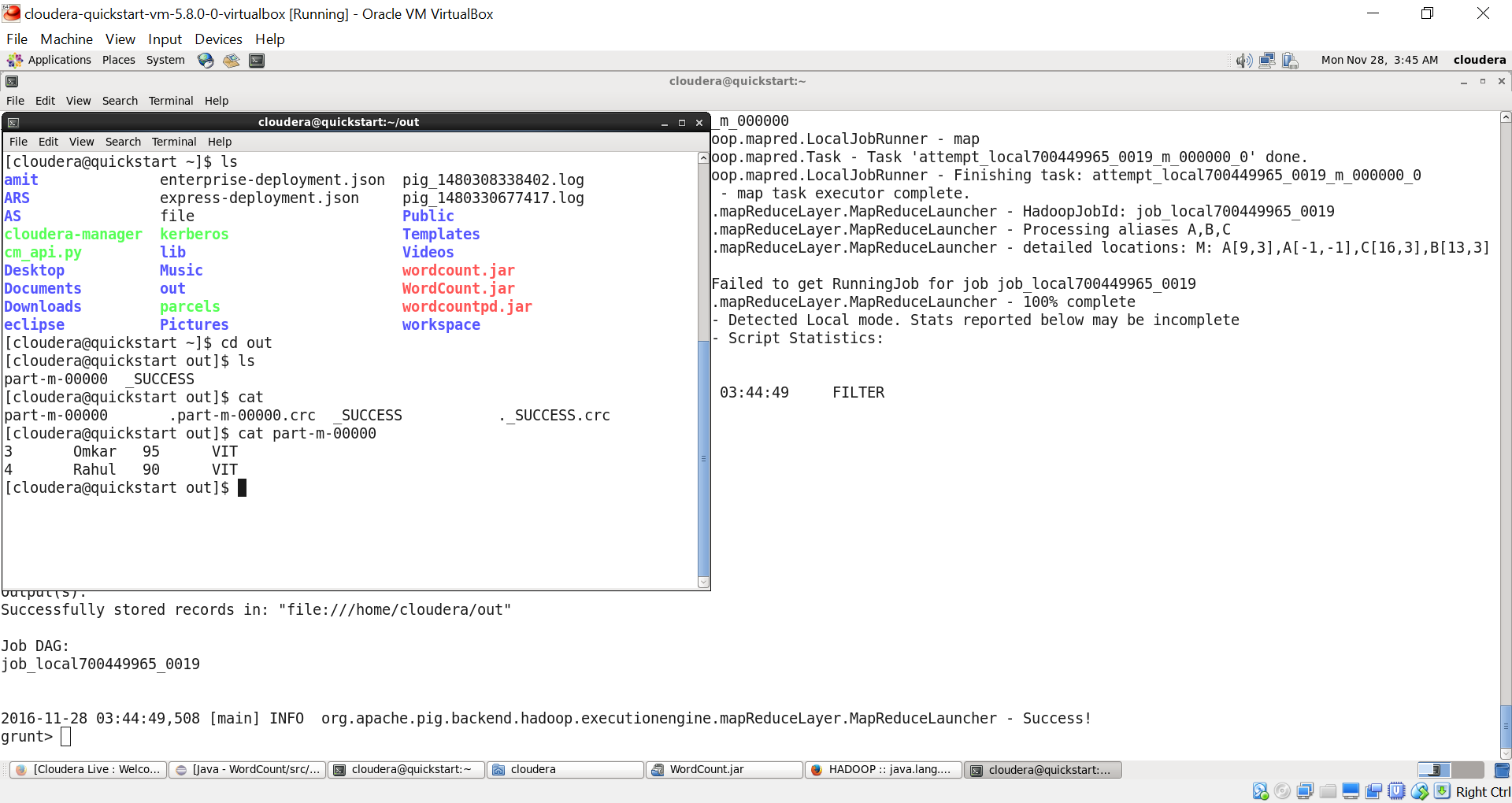
* Filter

C= Filter B by marks<96;



* Storing Dump

Store C into ‘out’;



* Problem: Get first and last.

B= order A by marks ASC;

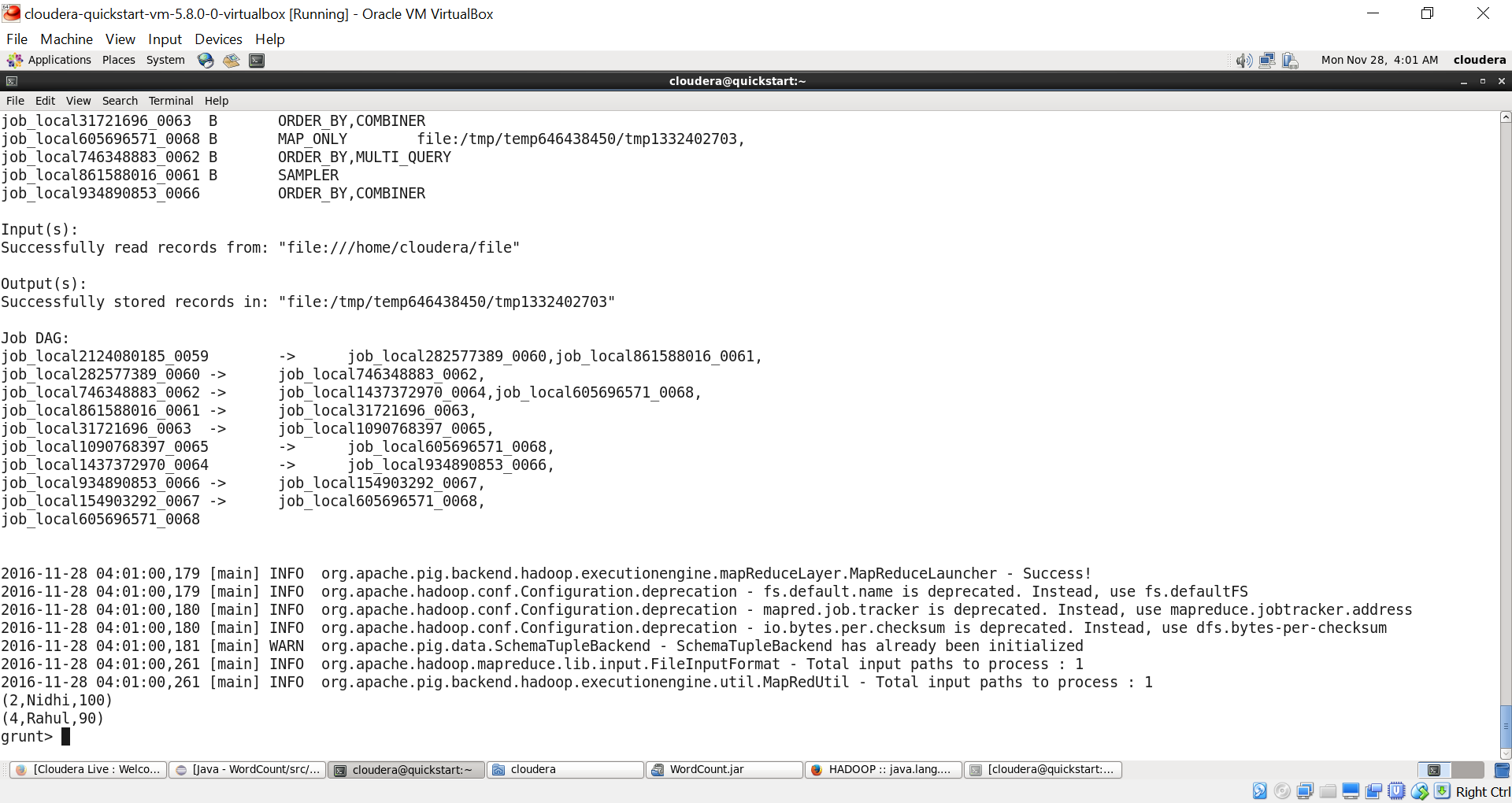
C= limit B 1;

B= order A by marks DESC;

D= limit B1;

B= filter B by marks==C.marks or marks==D.marks;

Dump B;

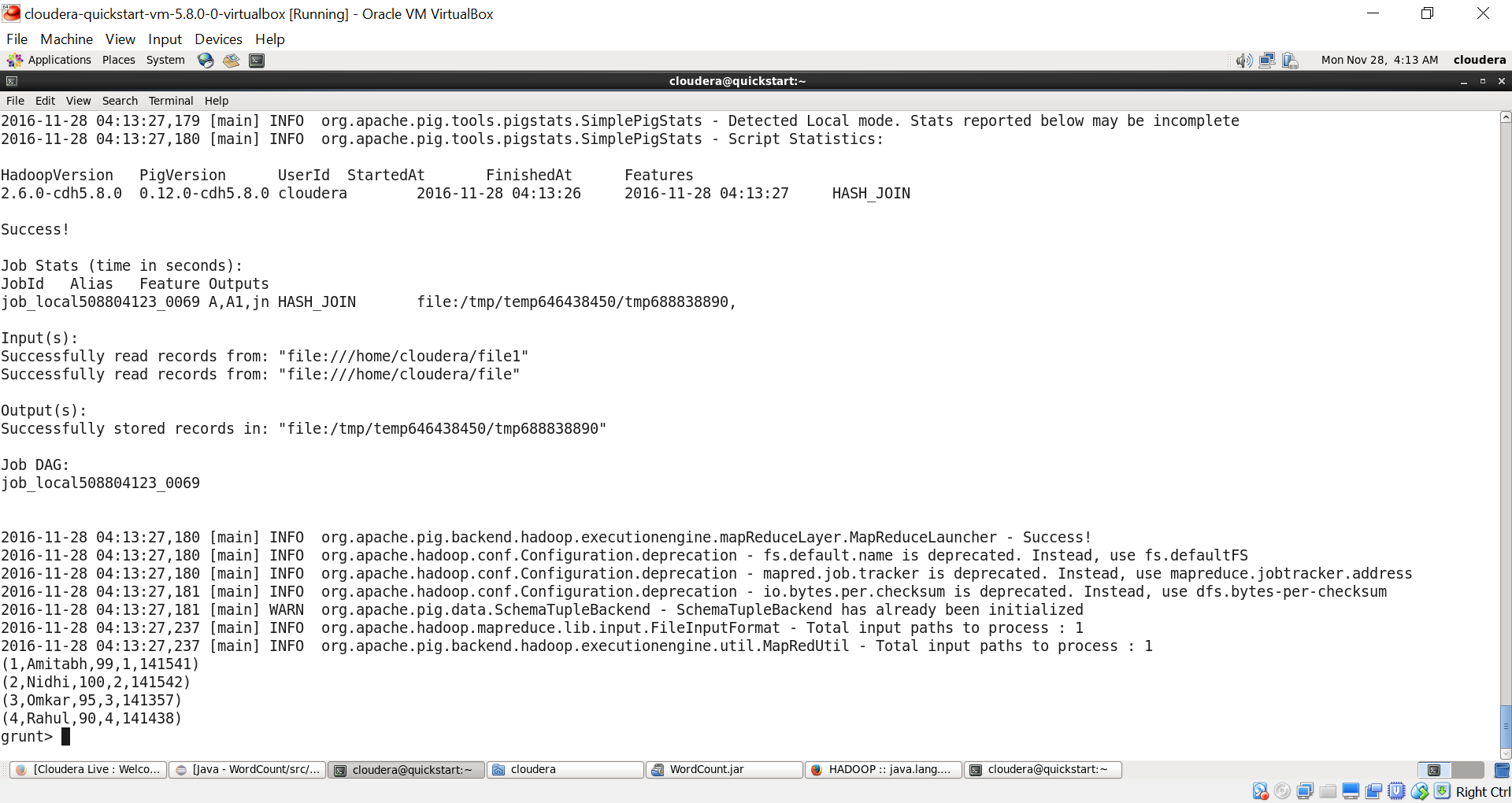


* Join

Types:

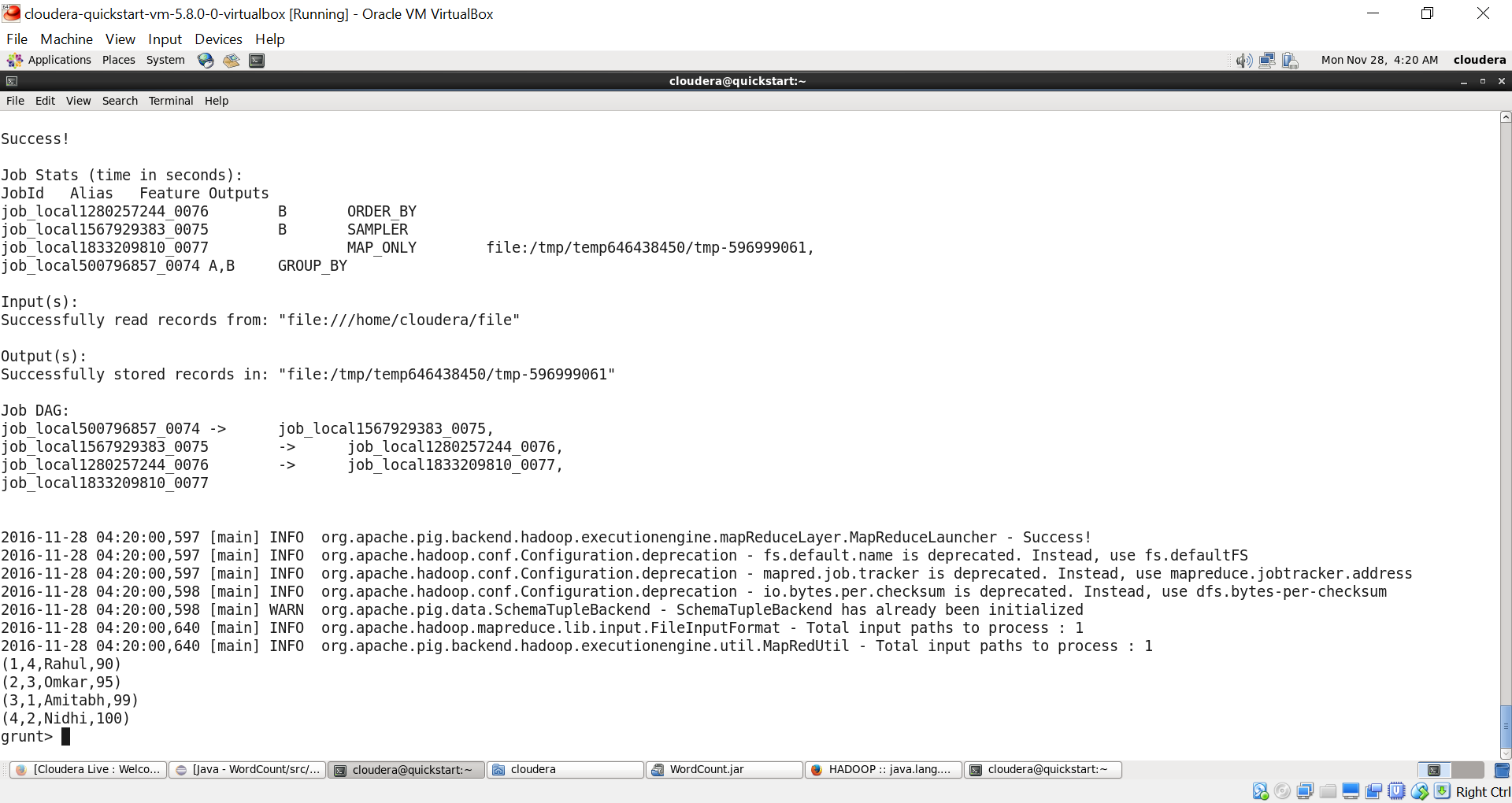
* Full – All tuples on both the files.
* Inner – The intersecting tuples of both files.
* Left – All the tuples in the left file joined with the common tuples in right file.
* Right – All the tuples in the right file joined with the common tuples in left file.

Syntax: jn=JOIN A by rollno [type], A1 by rollno. (Default: inner join, type: LEFT, RIGHT, FULL)



* Rank

B= rank A by name;



* You can write a .pig file and directly run as:

pig –x local file.pig