**SESSION 8: PIG – DEMO I**

**ASSIGNMENT 2**

**SAMPLE DATASET**

emp\_details.txt emp\_expenses.txt

101,Amitabh,Bachchan,20000,1

102,Shahrukh,Khan,10000,2

103,Akshay,Kumar,11000,3

104,Anubhav,Jain,5000,4

105,Pawan,Goel,2500,5

106,Aamir,Khan,25000,1

107,Salman,Khan,17500,2

108,Ranbir,Kapoor,14000,3

109,Katrina,Kaif,1000,4

110,Priyanka,Chopra,2000,5

111,Tushar,Kapoor,500,1

112,Ajay,Devgan,5000,2

113,Jubeen,Garg,1000,1

114,Madhuri,Dixit,2000,2

101 200

102 100

110 400

114 200

119 200

105 100

101 100

104 300

102 400

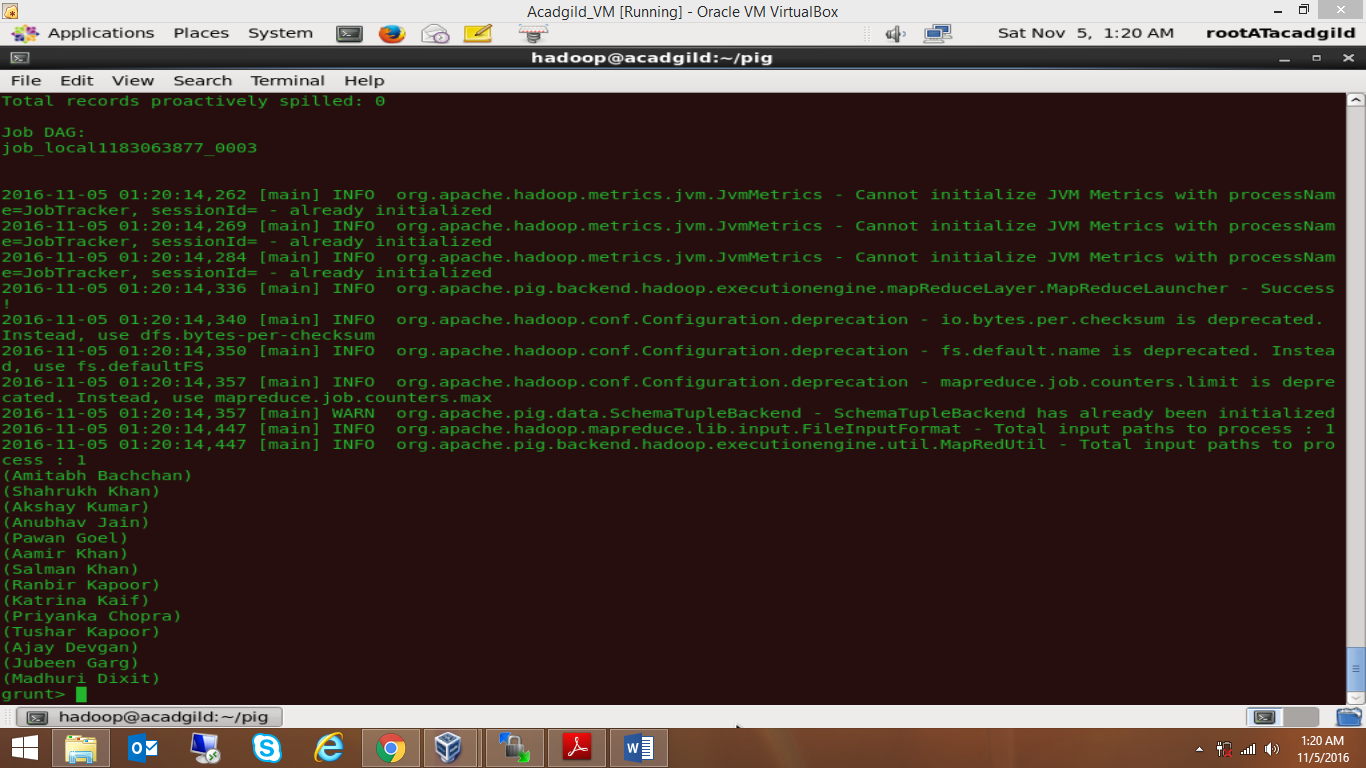
1. **CONCAT**

emp\_d = LOAD ‘emp\_details.txt’ using PigStorage(‘,’) AS (emp\_id:int, emp\_fname:chararray,

emp\_lname:chararray, emp\_salary:int, emp\_rating:int);

full\_name = FOREACH emp\_d GENERATE CONCAT(emp\_fname, CONCAT(‘ ‘, emp\_lname)) as fname;

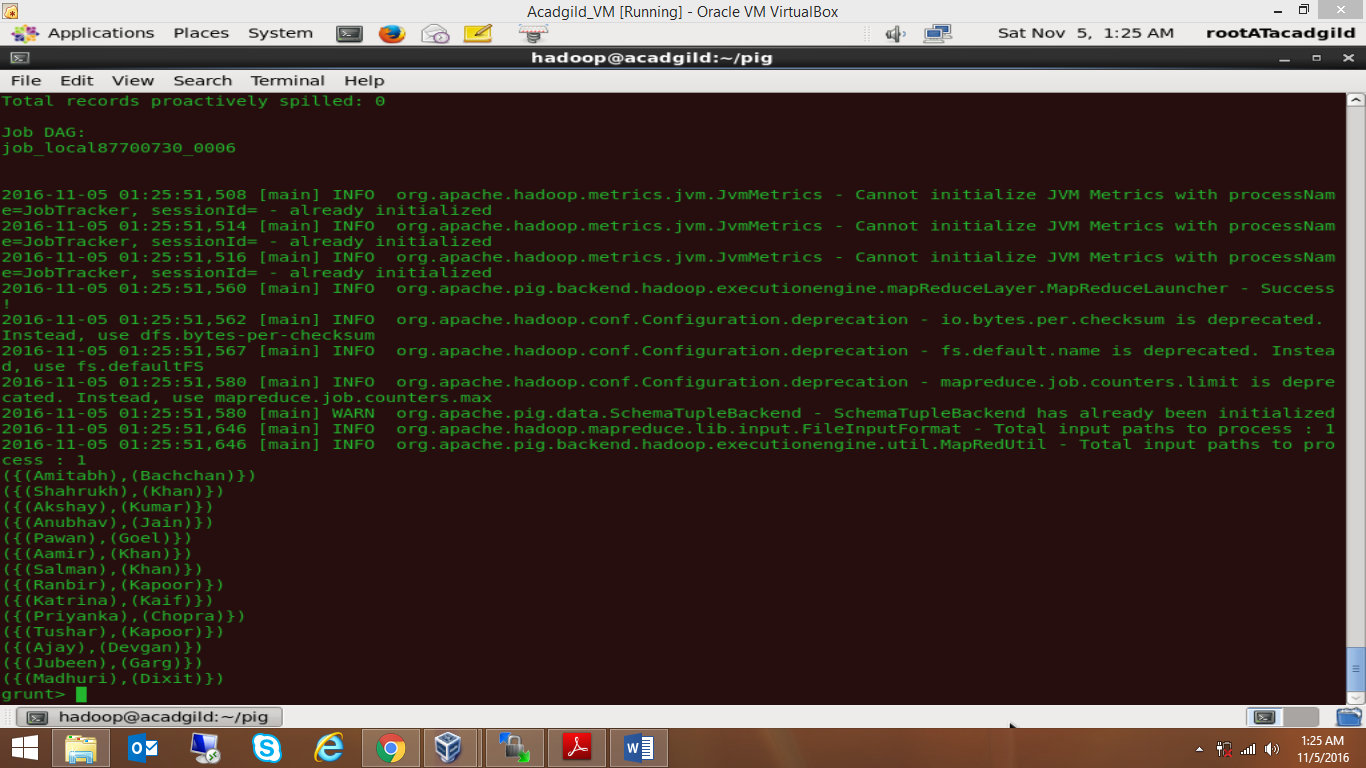
dump full\_name;



1. **TOKENIZE**

tokens = FOREACH full\_name GENERATE TOKENIZE(fname) as name\_token;

dump tokens;

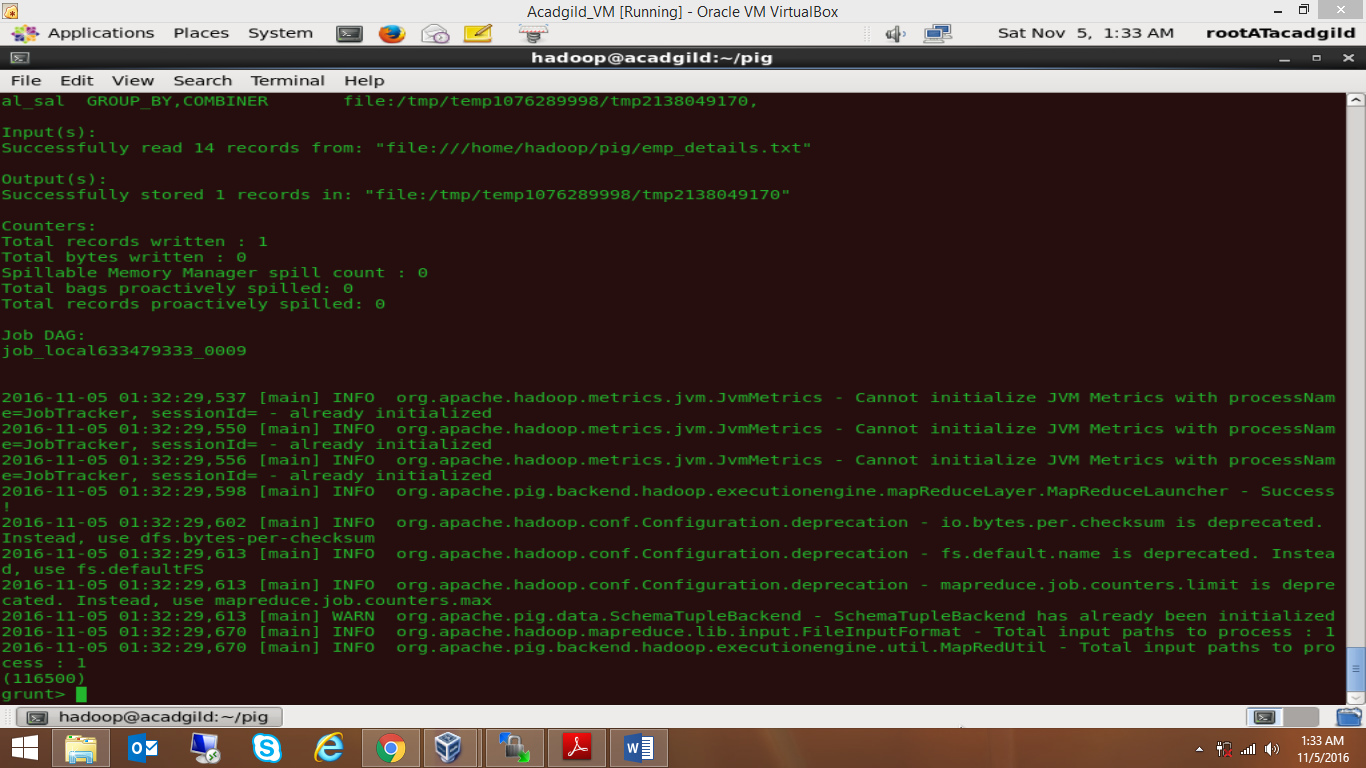


1. **SUM**

grouped\_emp = GROUP emp\_d all;

total\_sal = FOREACH grouped\_emp GENERATE SUM(emp\_d.emp\_salary);

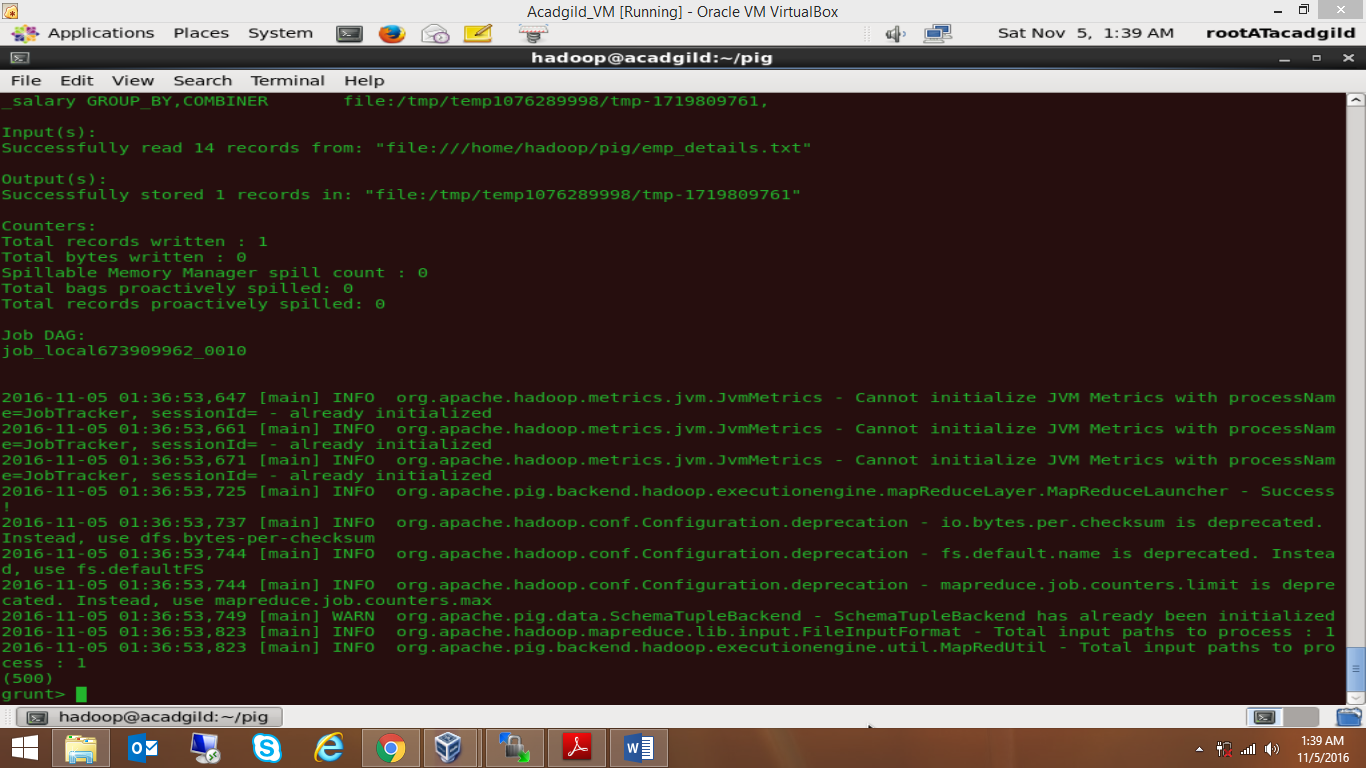
dump total\_sal;



1. **MIN**

min\_salary = FOREACH grouped\_emp GENERATE MIN(emp\_d.emp\_salary);

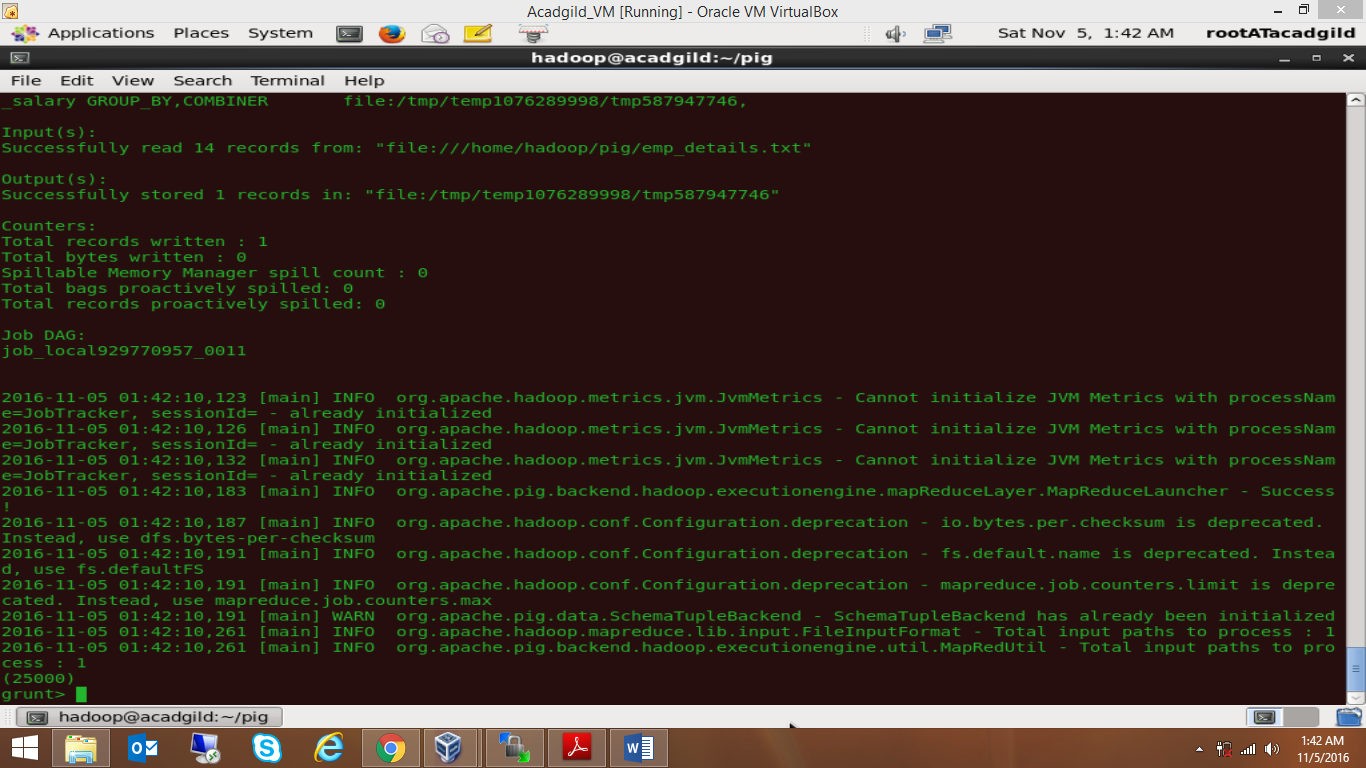
dump min\_salary;



1. **MAX**

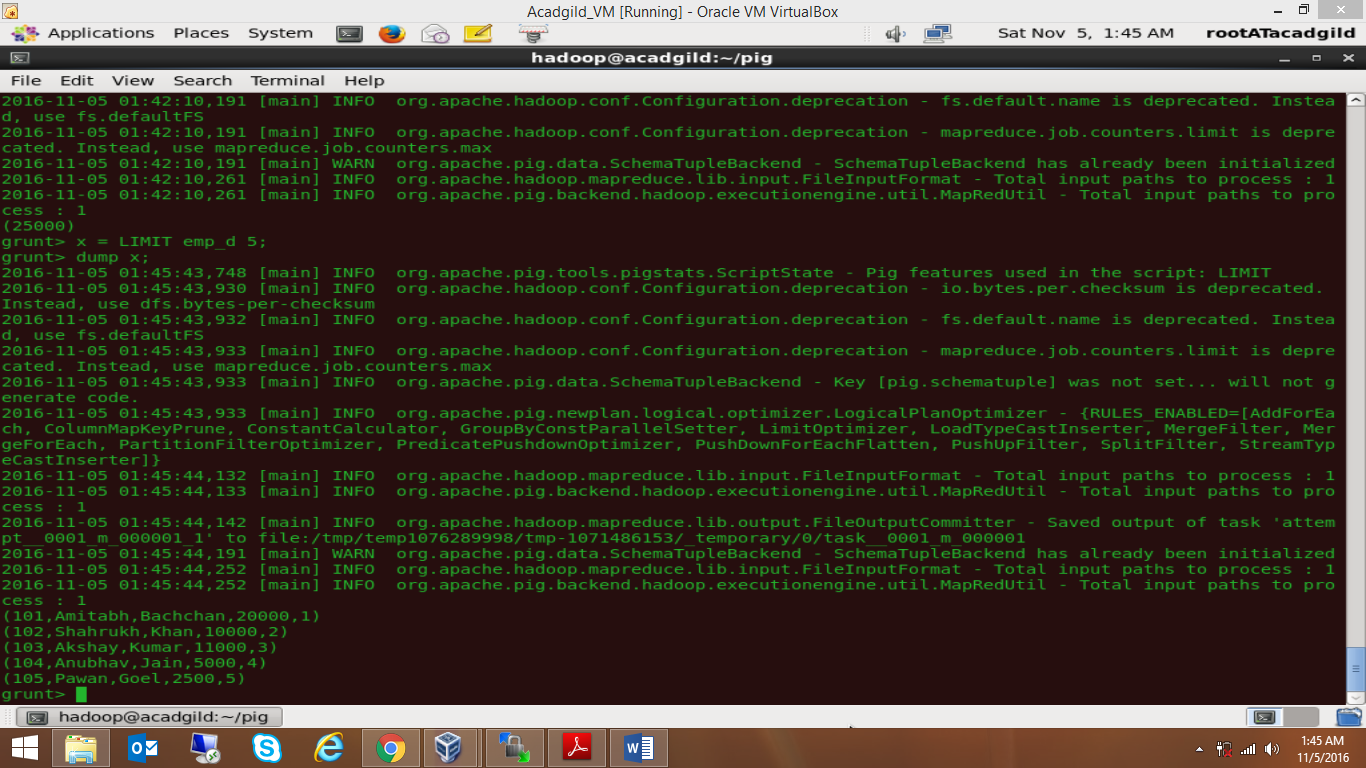
max\_salary = FOREACH grouped\_emp GENERATE MAX(emp\_d.emp\_salary);

dump max\_salary;



1. **LIMIT**

limit\_records = LIMIT emp\_d 5;



1. **STORE**

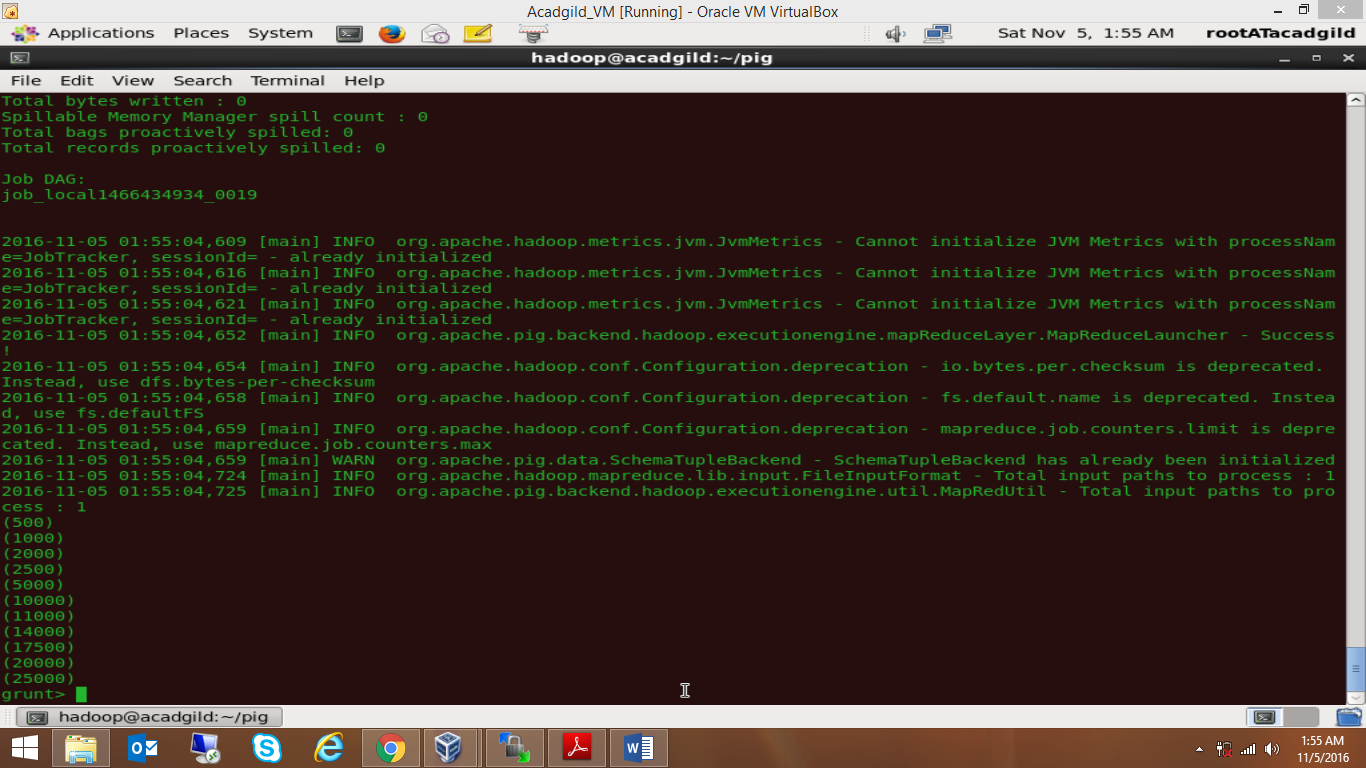
STORE limit\_records INTO ‘five\_emp\_records.txt’;

1. **DISTINCT**

sal = FOREACH emp\_d GENERATE emp\_salary;

distinct\_sal = DISTINCT sal;

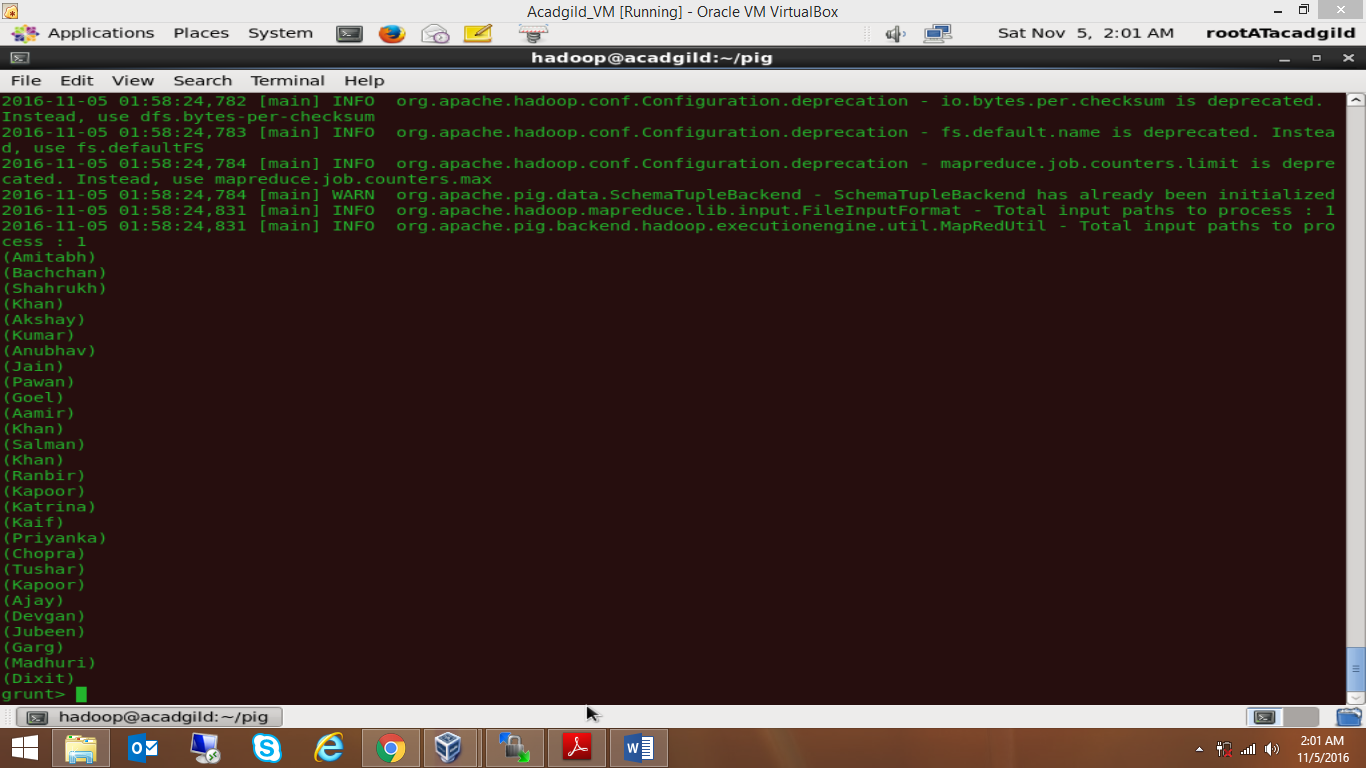
dump distinct\_sal;



1. **FLATTEN**

flat\_tokens = FOREACH tokens GENERATE FLATTEN(name\_token);

dump flat\_tokens;



1. **IsEmpty**

emp\_e = LOAD ‘emp\_expenses.txt’ AS (emp\_id:int, emp\_expense:int);

co\_emp = COGROUP emp\_d by emp\_id, emp\_e by emp\_id;

isempty\_data = FILTER co\_emp BY IsEmpty(emp\_e);

dump isempty\_data;

