A = load 'emp.csv' using PigStorage(',') as (eid: int, ename: chararray, epos: chararray, esal:int, ecom: int, edpno:int);

grunt> B = filter A by epos=='ANALYST';

grunt> dump B;

grunt> C = filter A by esal>=1500;

grunt> dump C;

[cloudera@quickstart ~]$ pwd

/home/cloudera

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/dept.csv Kaif/dept.csv

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/emp.csv Kaif/emp.csv

[cloudera@quickstart ~]$ hadoop fs -ls Kaif

hive> create table dept (edpno int, epos string, ecity string) row format delimited fields terminated by ',';

OK

Time taken: 0.706 seconds

hive> load data local inpath '/home/cloudera/Desktop/dept.csv' into table dept;

select \* from dept;

Hive

[cloudera@quickstart ~]$ hive

hive> show databases;

hive> create database kaiff;

hive> create table emp(eid int, ename string, epos string, esal int, ecom int, edpno int) row format delimited fields terminated by ',';

hive> load data local inpath '/home/cloudera/Desktop/emp.csv' into table emp;

hive> select \* from emp;

[cloudera@quickstart ~]$ hadoop fs -ls

[cloudera@quickstart ~]$ hadoop fs -mkdir kaif1

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/emp.csv/kaif1/emp.csv

PIG

[cloudera@quickstart ~]$ hadoop fs -mkdir kaif

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/emp.csv Kaif/emp.csv

[cloudera@quickstart ~]$ pig

grunt> cd kaif

grunt> A = load 'emp.csv' using PigStorage(',') as (eid: int, ename: chararray, epos:chararray, esal:int,ecom:int, edpno: int);

grunt> D = limit A 5;

grunt> dump D;

grunt> E = order A by esal;

grunt> dump E;

grunt> F = order A by esal desc;

grunt> dump F;

grunt> store F into '/user/cloudera/Kaif/pigout/' using PigStorage(',');

grunt> G = foreach A generate eid;

grunt> dump G;

grunt> H = foreach A generate \* , ecom \* 3 as bonus, esal \* 5 as incentive;

grunt> dump H;

grunt> I = foreach A generate SUBSTRING(ename,0,3);

grunt> dump I;

grunt> J = foreach A generate $0,$1;

grunt> dump J;

grunt> K = group A by edpno;

grunt> dump K;

grunt> M = group A by (edpno, epos);

grunt> dump M;

grunt> SPLIT A into B if edpno==10, C if edpno==20, D if epos=='Manager';

Joining

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/dept.csv kaif/dept.csv

again piggggg

grunt> A = load 'emp.csv' using PigStorage(',') as (eid: int, ename: chararray, epos:chararray, esal:int,ecom:int, edpno: int);

grunt> B = load 'dept.csv' using PigStorage(',') as (edpno: int, epos:chararray,ecity: chararray);

grunt> C = JOIN A by edpno, B by edpno;

grunt> dump C;

grunt> D = foreach C generate A::eid,B::epos;

grunt> dump D;

grunt> E = JOIN A by edpno RIGHT OUTER, B by edpno;

Word Count

(in another terminal)

[cloudera@quickstart ~]$ hadoop fs -mkdir kaif

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal /home/cloudera/Desktop/v.txt kaif/v.txt

(pig)

grunt> cat kaif/v.txt

hi

hi

i

grunt> lines = load '/user/cloudera/Kaif/v.txt' as (line:chararray);

grunt> dump lines;

grunt> token = foreach lines generate TOKENIZE(line);

grunt> dump token;

grunt> flats = foreach token generate FLATTEN($0);

grunt> dump flats;

grunt> group\_words = group flats by $0;

grunt> dump group\_words;

grunt> count\_words = foreach group\_words generate group as word, COUNT($1) as word\_occurence;

grunt> dump count\_words;