Nyse.zip contains lots of file, with huge number of records. It is taking a lot of time for executing the queries. Let's make it smaller so as to reduce the execution time for all our pocs.

Take NYSE\_daily\_prices\_A.csv

data = load '/home/hduser/nyse/NASDAQ\_daily\_prices\_A.csv' using PigStorage(',') as (stock\_exchange:chararray,stock\_symbol:chararray,date:chararray,stock\_price\_open:float,stock\_price\_high:float,stock\_price\_low:float,stock\_price\_close:float,stock\_volume:int,stock\_price\_adj\_close:float);

Converting date into datetime:

data1 = foreach data generate $0,$1,ToDate($2,'yyyy-MM-dd') as date,$3 .. $8;

Group them based on stock\_symbol(column 2)

grpd\_symbol = group data1 by stock\_symbol;

With in the bag, sort the data based on date descending order and get top 10 records from that group

sorted\_bag = foreach grpd\_symbol{

sort = order data1 by date desc;

lmt = limit sort 10;

generate FLATTEN(lmt);

};

Write this data into a file

store sorted\_bag into '/home/hduser/temp/sortnysesym' using PigStorage(',');