**Cleaning with PIG**

**Make New Dir and Put two files**

hadoop fs -ls /

hadoop fs -mkdir /user/data

hadoop fs -put athlete\_events.csv /user/data

hadoop fs -put noc\_regions.csv /user/data

**PIG: Load Files and Clean - (athlete\_events and noc\_regions)**

athletes = LOAD '/user/data/athlete\_events.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES\_MULTILINE','NOCHANGE','SKIP\_INPUT\_HEADER') AS (ID:int, Name:chararray, Sex:chararray, Age:int, Height:chararray, Weight:chararray, Team:chararray, NOC:chararray, Games:chararray, Year:int, Season:chararray, City:chararray, Sport:chararray, Event:chararray, Medal:chararray);

region\_noc = LOAD '/user/data/noc\_regions.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES\_MULTILINE','NOCHANGE','SKIP\_INPUT\_HEADER') as (NOC:chararray, region:chararray, notes:chararray);

athletes = FOREACH athletes GENERATE

ID as ID,

Name as Name,

Sex as Sex,

Age as Age,

Height as Height,

Weight as Weight,

Team as Team,

REPLACE(NOC,'SGP','SIN') as NOC,

Games as Games,

Year as Year,

Season as Season,

City as City,

Sport as Sport,

Event as Event,

Medal as Medal;

region\_noc = FOREACH region\_noc GENERATE

NOC AS NOC,

REPLACE(region,'Boliva','Bolivia') AS region,

notes AS notes;

region\_noc = FOREACH region\_noc GENERATE

NOC AS NOC,

REPLACE(region,',',' ') AS region,

notes AS notes;

region\_noc = FOREACH region\_noc GENERATE

NOC AS NOC,

REPLACE(region,'NA','Unknown') AS region,

notes AS notes;

athlete\_noc = JOIN athletes BY NOC LEFT, region\_noc BY NOC USING 'replicated';

athlete\_noc\_req = FOREACH athlete\_noc GENERATE athletes::ID AS ID, athletes::Sex AS Sex, athletes::Age AS Age, athletes::Height AS Height, athletes::Weight AS Weight, athletes::NOC AS NOC, athletes::Year AS Year, athletes::Season AS Season, athletes::Sport AS Sport, athletes::Medal AS Medal, region\_noc::region AS region;

STORE athlete\_noc\_req INTO '/user/data/athlete\_noc\_req' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',');

**HIVE – Create table and load the dataset**

create table athlete\_table(ID int, Sex string, Age int, Height string, Weight string, NOC string, Year int, Season string, Sport string, Medal string, region string) row format delimited FIELDS TERMINATED BY ',' location '/user/data/athlete\_noc\_req';

**Convert Table to csv file and export it**

hive -e 'set hive.cli.print.header=true;

select \* from athlete\_table' | sed 's/[\t]/,/g'  > /home/tenzin\_palbar2/atheletes\_noc\_final\_file.csv