PIG :
Pig is a high level scripting language that is used with Apache Hadoop. Pig excels at describing data analysis problems as data flows. Pig is complete in that you can do all the required data manipulations in Apache Hadoop with Pig. In addition through the User Defined Functions(UDF) facility in Pig you can have Pig invoke code in many languages like JRuby, Jython and Java. Conversely you can execute Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.l

Notes:
If permission denied:
sudo -u hdfs hadoop fs -mkdir /us

LINKS:
http://pig.apache.org/docs/r0.15.0/basic.htm
https://github.com/alanfgates/programmingpig/blob/master/examples/ch2/average_dividend.pig

https://hortonworks.com/tutorial/how-to-process-data-with-apache-pig/

https://mapr.com/blog/using-hive-and-pig-baseball-statistics/

https://github.com/alanfgates/programmingpig/blob/master/examples/ch4/no_schema.pig

Data Downloads:
http://www.retrosheet.org/game.html downloads

[cloudera@quickstart ~]$ hdfs dfs -mkdir /project/baseball

[cloudera@quickstart ~]$ hdfs dfs -put 2017eve /project/baseball

[cloudera@quickstart ~]$ hdfs dfs -ls /project/baseball/

Found 1 items
drwxr-xr-x - cloudera supergroup 0 2018-04-30 14:37 /project/baseball/2017eve
[cloudera@quickstart ~]$ hdfs dfs -ls /project/baseball/2017eve/*.EV{N,A} | wc -l

30

[cloudera@quickstart ~]$ hdfs dfs -ls /project/baseball/2017eve/*.EVA | wc -l15

[cloudera@quickstart ~]$ hdfs dfs -ls /project/baseball/2017eve/2017CHN.EVN | wc -l
1

……...

[cloudera@quickstart ~]$ pig -x mapreduce

grunt> raw = LOAD '/user/cloudera/projects/*.EVN' using PigStorage(',') as (type:chararray);

grunt> id = FILTER raw BY type MATCHES 'id';

grunt> g = GROUP id ALL;

grunt> result = FOREACH g Generate COUNT(id);

grunt> dump result;
STORE result INTO '/projects/name_to_id/' USING PigStorage(',');

……..

Local dir
grunt>records = LOAD 'sample.txt' AS (year:chararray, temperature:int, quality);

filtered_records = FILTER records BY temperature != 9999 ;

DUMP filtered_records;

Op
- Total input paths to process : 1
2018-05-03 14:04:01,781 [main] INFO
org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process
: 1
(1950,0,1)
(1950,22,1)
(1950,-11,1)
(1949,111,1)
(1949,78,1)
With hdfs :remember to start cloudera manager and services
records1 = LOAD '/user/cloudera/samples/sample.txt' AS (year:chararray, temperature:int,
quality);

……...