CISC 886
Cloud and Big Data

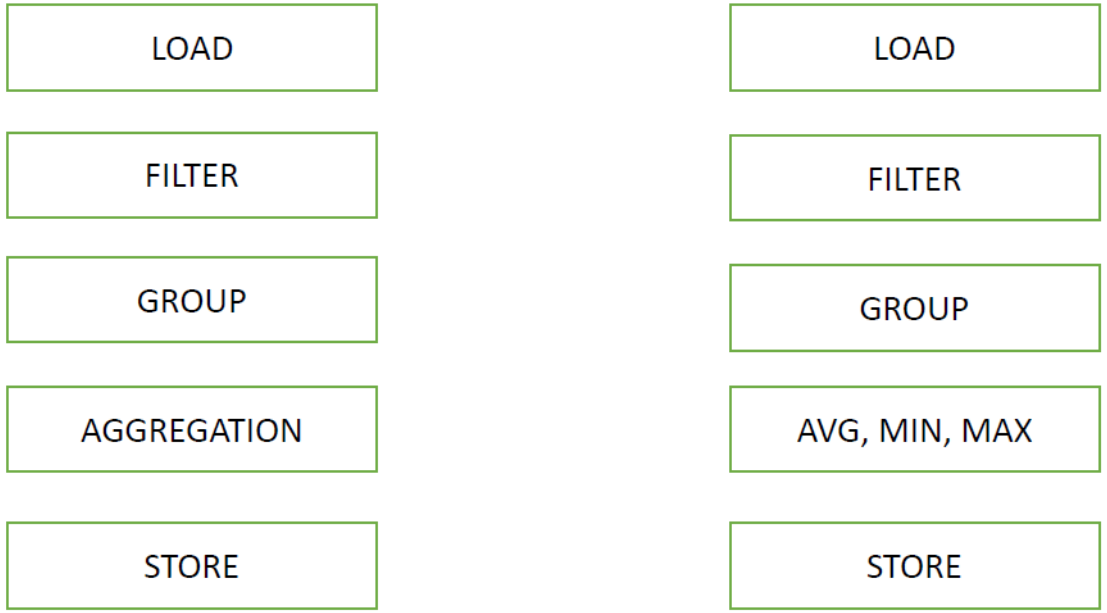
Lab 3

Shahenda Youssef

PIG

Data processing pipeline

- Most data problem can be broken down into list of operations and Pig provide instructions for each operations



PIG Command

- **Submit Pig Script** $pig /script_path.pig
- **View Result** $hdfs dfs -cat outputFilepath/part-r-00000
- **Load dataset** with column names and datatypes

  $stock_records = LOAD '/stocks' USING PigStorage(',') as
  (tickers:chararray, date:datetime, close:float,volume:int);
- **Describe dataset** $Describe stock_records;
- **Limit dataset** $Limit_data = limit stock_records 5;
- Dump limit_data;

PIG Command

- **Group by** $grp_by_tickers = GROUP stock_records BY tickers;
- Calculate maximum closing price

  $max_closing = FOREACH grp_by_tickers GENERATE group,
  MAX(stock_records.close) as maxclose;
- **Store output** $STORE max_closing INTO 'output/pig/stocks' USING PigStorage(',');
- **Select Columns** $foreach stock_records generate tickers, close;
- **Filter dataset** $filter stock_records by close > 100;
- **Order dataset** $order stock_records by $2 desc;

PIG – Word Count

- a = load ‘data/wordcount. as (x:chararray);
- b = foreach a generate tokenize(x) as x;
- c = foreach b generate flatten(x) as x;
- d= group c by x;
- e= foreach d generate group, count(c.x) as wordcount;

HIVE Command

- **Write Hive Command** $hive
- $show databases;
- $use database_name;
- $show tables;

- **Describe table** $Describe formatted table_name;

HIVE Command

- **CREATE EXTERNAL TABLE**

  $CREATE EXTERNAL TABLE IF NOT EXISTS stocks (

  ticker STRING,

  close FLOAT,

  volume INT)

  ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’

  LOCATION '/user/osboxes/input/stocks’;

- CALCULATE MAX CLOSING PRICE

  $SELECT ticker, max(close) max_close

  FROM stocks

  GROUP BY ticker;

HIVE Command

- **Select**

SELECT 100 RECORDS $SELECT * FROM stocks LIMIT 100;

$Select distinct ticker from stocks;

$Select * from stocks Where ticker =‘TSLA’;

$Select * from stocks Where ticker in (‘TSLA’, ‘ABC’);

$Select * from stocks Where ticker LIKE ‘TS%’;

Lab 3

This lab involves you to work with PIG and HIVE. Here is what you have to do:

1. Prepare a text file with WH questions by taking example sentences from the following link, name the file as 'whquestion.txt' and save it in your local drive: WH Question Words | Vocabulary | EnglishClub https://www.englishclub.com/vocabulary/wh-question-words.htm .

2. Copy the whquestion.txt file into a hdfs location

3. Write commands in PIG to count only the frequency of the WH question words (What, Which, Who, Whom, Why...) from the question file you put in hdfs location

4. Accumulate all the commands in a script file called pigwhquestion.pig

Lab 3

5. Run the script file to generate the final outcome at once.

6. Take a screen-shot of the script file content along with the final output the script file generates and put them in a word document named "lab3-yourname-ID.doc"

7. Use the same input data in whquestion.txt from within Hive and give the appropriate SQL commands to solve the same word frequency count problem above. Refer to any existing material you found on the web.

8. List all the Hive SQL commands used and **report your finding** in the same word document "lab3-yourname-ID.doc in a separate section.