Big Data bootcamp, Hive: hands on

# General tips

Use official documentation before googling: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DML>

<https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF>

In console: beeline -u jdbc:hive2://localhost:10000

# Report

A file in sensible format (.docx and .pdf are fine) with all the steps you have taken and answers to questions. Include exact commands you used (text is OK for most points, when you should take screenshot it’s said so). If you had some problems, describe what the problem was, what you tried and what the solution was. Push report to git.

# Task

1. Get data from https://drive.google.com/file/d/1AOytyTA2MYYQOs6x9Q5M3n24gkbU8Jet/view

*These data are LV wikipipedia TOP 1000 articles by month, each file should contain one month and have header*

1. Clean data if necessary.

Probably you will realize some cleaning is necessary few steps later. It’s ok. Come back and do some cleaning. It’s ok to do cleaning by hand. Extra points if you do it with script.

1. Put all data files on hdfs in directory /user/cloudera/wiki
2. Create database bootcamp
3. Create external table e\_wiki in database bootcamp, location: /user/cloudera/wiki(CREATE EXTERNAL TABLE IF NOT EXISTS e\_wiki(

project string, access string, year int, month int, day string, articles\_\_article string, articles\_\_views string articles\_\_rank string)

COMMENT 'external database wiki bootcamp' ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/user/cloudera/wiki';)

1. How can you be sure all the data from files are present in table? Think of some way how to do this check. Answer “it should be after my commands” is not enough, becasuse sometimes things break unexpectedly and sometimes people make a mihdfs stake. One of the simplest ways how to do it is check row count and number of items in new table.

(select count(\*) from e\_wiki) 5906

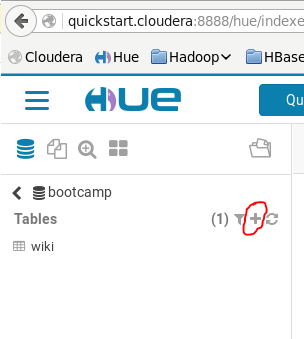
1. Create internal table t\_wiki and fill it with all the data from table e\_wiki and:
   1. Storage format ORC
   2. Schema:
      1. Year:int
      2. Month:int
      3. Article: string
      4. Views: int
      5. Rank: int

(CREATE TABLE bootcamp.t\_wiki (year INT, month int, article string, views int, rank int)

STORED AS ORC;)

1. Open one of the files in the new table. Are there any differences in directory structure? Why? How the file content differs (screenshot for this point with file content) from data you have on your laptop?

There are differences in structure, because we changed it from e\_wiki 

1. Try creating a table trough the HUE – easy way!
   1. Click on + sign in Hue  
      
   2. Work your way trough GUI
   3. What are advantages and disadvantages of this approach? Can you think about one use case for each?

Console works faster and gives faster results, writing in hue has better interface and windows like go trough, instead of terminals comands.

Console is definetly better for working trough larger amounts of data, since it works faster for example working with telecom data from last month. Hue has better more intuitive interface, you can look at the tables, for example if you need to see the contents of table to understand what data is there.

1. Write a queries to get to know following. If you notice any weird answers, point them out and fix them if it can be done in step 2
   1. How many records are in table t\_wiki ?

select count(\*) from t\_wiki; 5906

* 1. What’s the biggest view count in one month?

Select max(views)from t\_wiki GROUP by month;

Select sum(views) as Total

from t\_wiki

group by month

order by Total desc limit 1; =======  =

* 1. What is the top article in all months? What’s the top real article in all months?

Select max(views )from t\_wiki order by views desc limit 1;

"2018.\_gada\_Pasaules\_čempionāts\_hokejā" 92344

(Sākumlapa isn’t a real article. Special:<something> isn’t a real article)

* 1. What’s the second biggest view count?
  2. How many records there are for each month?

Select count(article)

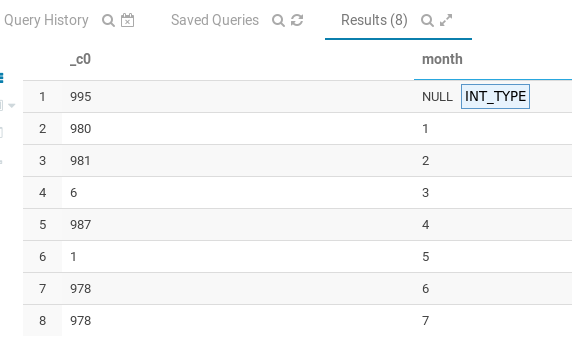
from t\_wiki

group by month;

* 1. How many real articles there are for each month?

Select count(article), month

from t\_wiki

group by month; 

some error here

* 1. Which article reached TOP 1000 in most months?
  2. What’s the average view count to reach TOP 1000?
  3. (for top achievers) Is there any correllation between page ranks between months? (hint: Hive udf exists for that, but you might rearrange data before. It’s OK to create a temporary table or a view for this)