НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО”
КАФЕДРА ІНФОРМАТИКИ ТА ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ІНСТРУКТИВНО-МЕТОДИЧНІ МАТЕРІАЛИ ДО ЛАБОРАТОРНИХ РОБІТ ТА КОМП'ЮТЕРНИХ ПРАКТИКУМІВ КРЕДИТНОГО МОДУЛЯ

“ Обробка надвеликих масивів даних”

для спеціальності

ЛАБОРАТОРНА РОБОТА

«Розподілена обробка даних в Apache Hadoop та Apache Hive»

Спеціальність 121“ Інженерія програмного забезпечення ”

Денної/заочної форми навчання

Київ - 2024

COMPUTER PRACTICE

Before the first computer workshop, the students should to split small team (2 member) for following activities:

- programming in Java / Phyton / Scala for downloading data, writing ETL procedures and data processing;
- configuring components hadoop, spark, hive, e.t.c
- designing the architecture of the database
- making reports.

Computer practice № 1

Preparation for a computer workshop

Для виконання практикуму попередньо необхідно завантажити дані про юридичні та фізичні особи України - To perform the workshop, it is necessary download the data on legal entities of Ukraine

CSV Data. FOP and UO located via URL:

Prepared data:

https://drive.google.com/drive/folders/1-7R7iK7J_lf2xfiMln0qRXxu36e_vgxM

- Uo - > Юридичні особи
- Fop - > Фізичні особи підриємці

Postgresql database dump loaded in the folder – https://drive.google.com/drive/u/1/folders/1QGydcDUeOOOdVbrmh_H1Aet_ILCPJK8R

You can use cloud solution cloud.google.com in educational porposes.

- For educational purposes. Create an account on the cloud.google.com cloud.

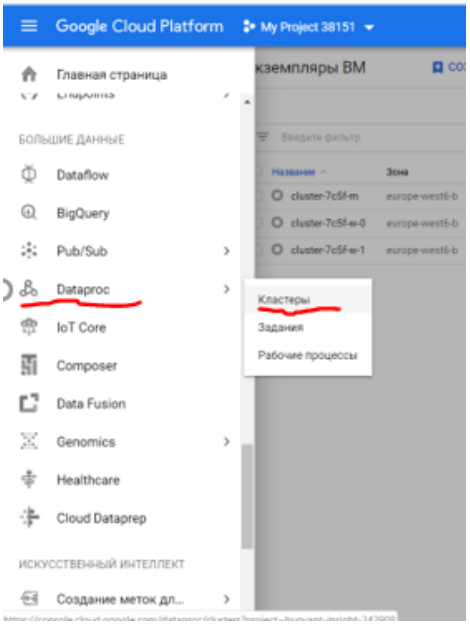
a) Create a Google Account.

b) sign up at http://cloud.google.com/

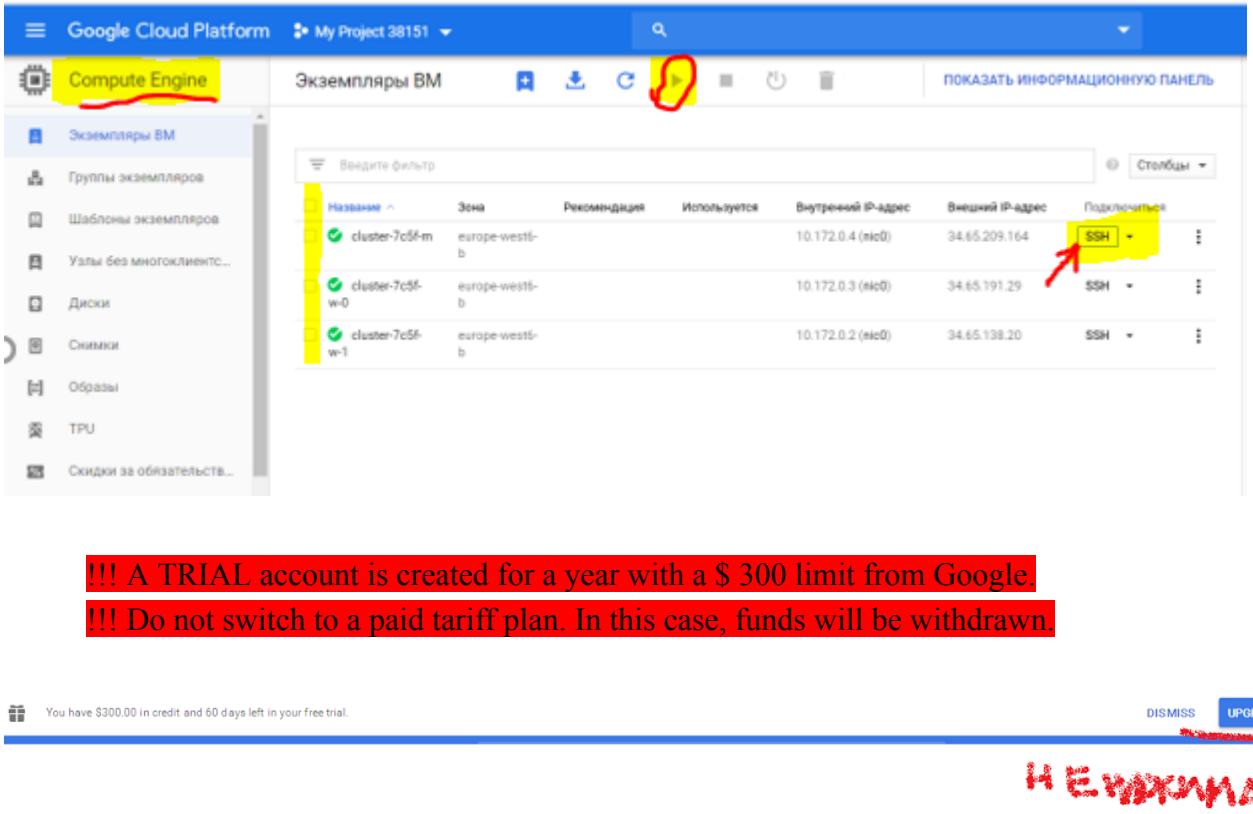
c) Create empty project.

d) Create cluster with at least 2 computing nodes and one main node.

(Menu -> DataProc ->Clusters)



To Open SHH client go to Computer Engine, select clusters, run, open SSH



!!! A TRIAL account is created for a year with a $ 300 limit from Google.

!!! Do not switch to a paid tariff plan. In this case, funds will be withdrawn.

Fig. 1. Do not click on the "Upgrade" fare.

!!! Actions after the end of the trial.

Google tries to take cats out of the trilogue without warning.

Watch out for billing. Exceed - exclude billing. Billing \ Disable billing.

Next to close the project on the google cloud, follow these steps:

In order to stop the charges from accruing, you can follow the steps below to shutdown your project and close the billing account.

To shut down a project:

1. Go to the Cloud Platform Console.

2. Open the console menu Gallery Menu on the top left and select IAM & Admin, then select All projects.

3. Find the name or project ID of the project you want to shut down, then click DELETE PROJECT. A confirmation screen describing what will happen appears.

4. To confirm, enter your project ID and click Shut down.

To close an account:

1. Go to the Cloud Platform Console.

2. Open the console menu and select Billing

3. If you have more than one billing account, select the billing account name.

4. Click Close billing account.

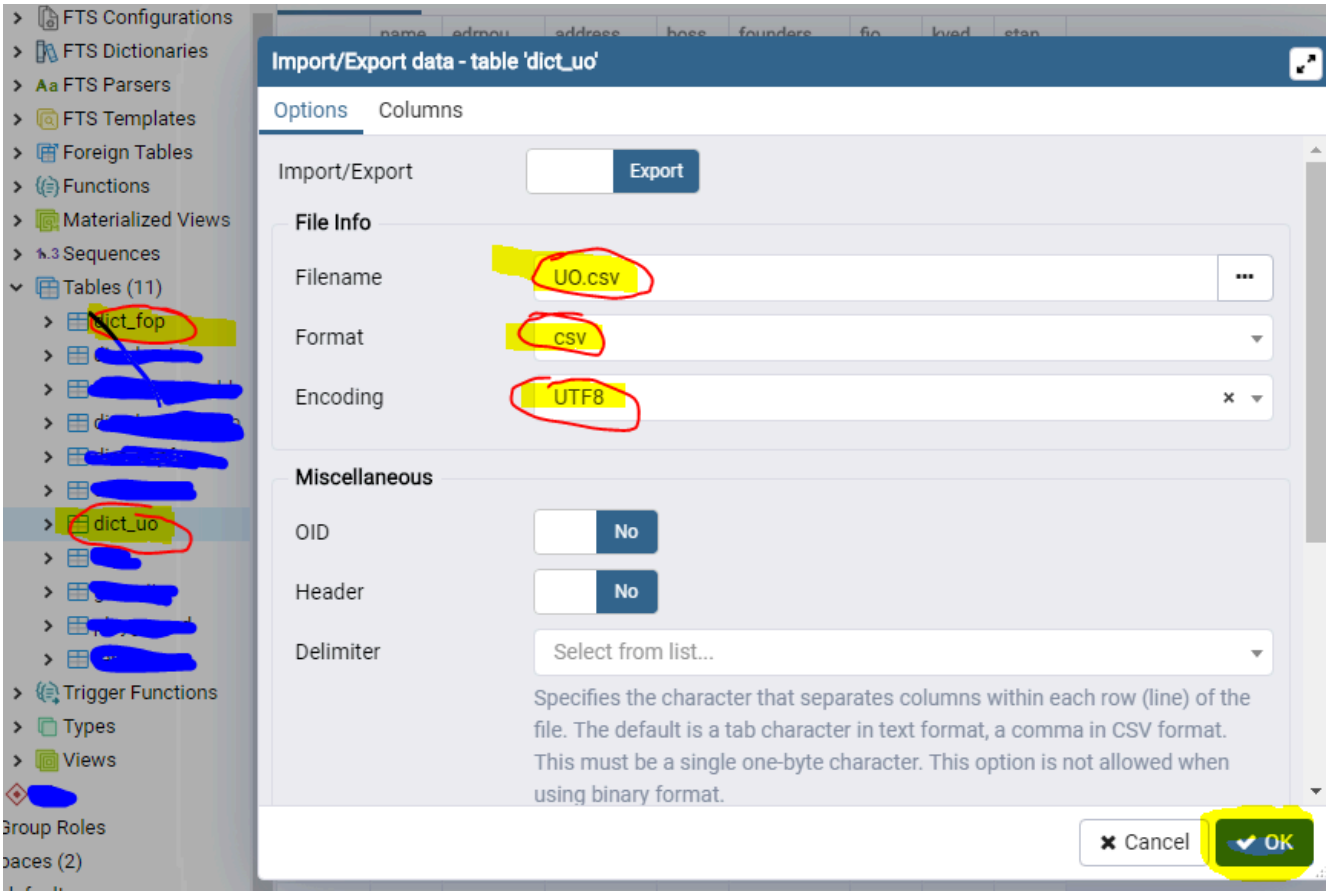
Task 1. Data loading and query execution

```
-- creating directories
hadoop fs -mkdir /tables_data
hadoop fs -mkdir /tables_data/UO
hadoop fs -mkdir /tables_data/FOP

-- loading data file
 in case using CSV files:
hadoop fs -put ./UO.csv /tables_data/UO/
hadoop fs -put ./FOP.csv /tables_data/FOP/
```

---

1. Create tables from PostgreSQL DUMP. //you may take it from google drive, url above




```
-- 2. Upload file to HDFS
-- 3. Create table based on CSV in Apache Hive
```

Use next script for table creation:

create external table UOtable(name string,EDRPOU string,ADDRESS string,BOSS string,founders string,fio string,KVED string,stan string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/tables_data/UO/';

```
create external table UO_table
(
name string,
EDRPOU string,
ADDRESS string,
BOSS string,
founders string,
fio string,
KVED string,
stan string
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE
LOCATION '/tables_data/UO/'
;
```

Make same actions for table FOP_table;

create external table FOP_table(fio string,address string,kved string,stan string) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS TEXTFILE LOCATION '/tables_data/FOP/';

```
create external table FOP_table
(
fio string,
address string,
kved string,
stan string

)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS TEXTFILE
LOCATION /tables_data/FOP/'
```

Laboratory workshop tasks

1.1 Determine the total records count

```
-- total number of entries
select count(*) from UO_table;
```

result: …

* Run this task on the RDBMS server (oracle, mysq, mssql, etc.). Compare the query execution

time.

1.2 Perform an analytical query

-- displays an additional serial number of the placement in the order of growth of the UO

* Run this task on the RDBMS server (oracle, mysq, mssql, etc.). Compare the query execution time.

```
select
name ,
edrpou ,
address ,
row_number() over (partition by address order by edrpou) as rn_by_place
from   UO_table
    limit 20;
```

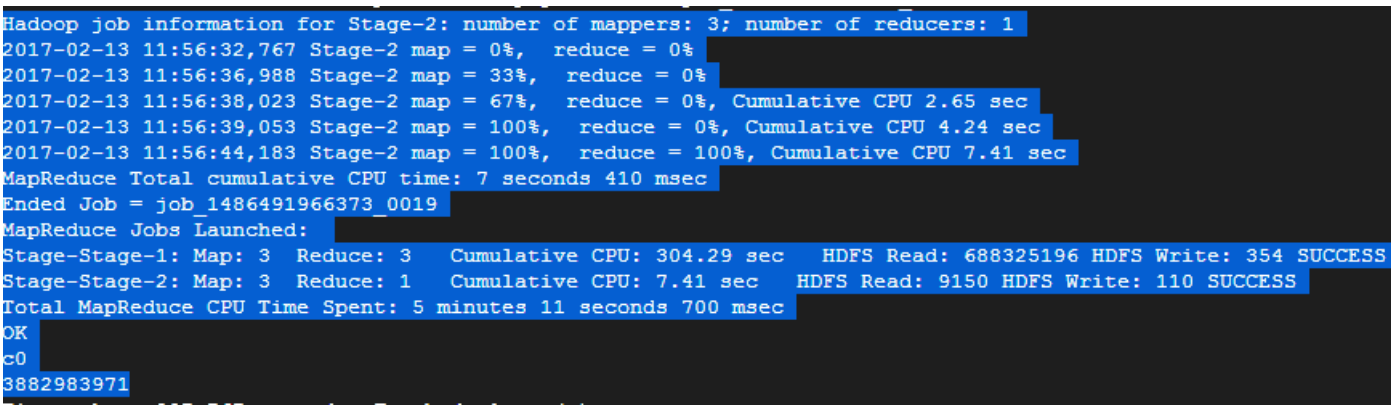
```
name    sname   edrpou  place   rn_by_place
ПОВНЕ ТОВАРИСТВО "ФІРМА "КЛИМЕНКО І.О., СПИРИДОНОВА М.П. І КОМПАНІЯ"    NULL    NULL    NULL    1
 [Назва]                32405541                1
,34604449,83114  Донецька обл.   місто Донецьк   Київський район         1
МАЛЕ ПІДПРИЄМСТВО "УНІВЕРСАЛ-ЛТД" ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ",МП "УНІВЕРСАЛ -ЛТД" /ТОВ/,19213810,65003      Одеська обл.    місто Одеса     Суворовсь
кий район       1
ФІРМА "ВОЛВА" ЛТД МАЛЕ ПІДПРИЄМСТВО У ВИГЛЯДІ ТОВАРИСТВА З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ",,13924094,65066   Одеська обл.    місто Одеса     Суворовський район      2
ПРИВАТНЕ ПІДПРИЄМСТВО "ЛФКС"    ПП "ЛФКС"       32887815        "ЛФКС, ВУЛИЦЯ ПАНІКАХІ, БУД.2, КВ.502, М.ДНІПРОПЕТРОВСЬК, ДНІПРОПЕТРОВСЬКА ОБЛ., УКРАЇНА        1
ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ " ВІКСАН "     ТОВ "ВІКСАН"    33053066        -, ВУЛ. КОЗАЧА, 29-А, МІСТО КИЇВ, ГОЛОСІЇВСЬКИЙ РАЙОН, УКРАЇНА  1
ПРИВАТНЕ ПІДПРИЄМСТВО ЦЕНТР ДУХОВНОГО ТА КУЛЬТУРНОГО РОЗВИТКУ ДІТЕЙ "НІРАНАНДА"       30184411        0, БРОВАРСЬКИЙ ПРОСПЕКТ, БУД.2, МІСТО КИЇВ, УКРАЇНА     1
ПРИВАТНЕ ПІДПРИЄМСТВО "ГЛОРІЯ - А"              24931588        0, БУЛЬВАР ВЕРХОВНОЇ РАДИ, БУД. 10 - А, КВ. 27, М.КИЇВ, УКРАЇНА 1
МАЛЕ ВИРОБНИЧЕ ПІДПРИЄМСТВО ФІРМА "БІОФАРМЕЛЕКТРО"              13689173        0, БУЛЬВАР ПЕРОВА, БУД.21-А, М.КИЇВ, УКРАЇНА    1
ВИРОБНИЧО-КОМЕРЦІЙНА ФІРМА "КРІОЛА"             21572812        0, БУЛЬВАР ПЕРОВА, БУД.48-А, КВ.100, МІСТО КИЇВ, УКРАЇНА       1
ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ "П.К.С."                25202118        0, ВУЛ. БЕРЕЗНЯКІВСЬКА, БУД.14-А, МІСТО КИЇВ, УКРАЇНА   1
МАЛЕ ПРИВАТНЕ ПІДПРИЄМСТВО ФІРМА "НОРД"         21461797        0, ВУЛ. БУДІВЕЛЬНИКІВ, БУД. 43/12, К. 10, М.КИЇВ, УКРАЇНА       1
МАЛЕ ПІДПРИЄМСТВО " ФІРМА "ВІКТОРІЯ - КОМПЛЕКС"        16476220        0, ВУЛ. ГАШЕКА, БУД. 4, М.КИЇВ, УКРАЇНА 1
ПРИВАТНЕ ПІДПРИЄМСТВО "УКРПРОМСПЛАВ"    ПП"УКРПРОМСПОАВ"        31170635        0, ВУЛ. ДЕГТЯРІВСЬКА, Б.31, М. КИЇВ, УКРАЇНА    1
ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ "МАРАФОН ЛТД"           21551075        0, ВУЛ. ЕНТУЗІАСТІВ, БУД. 9 - В, М.КИЇВ, УКРАЇНА       1
ПРИВАТНЕ ПІДПРИЄМСТВО "МАЖОР"           30252067        0, ВУЛ. КАУНАСЬКА, БУД. 10 - А, М.КИЇВ, ДНІПРОВСЬКИЙ, УКРАЇНА   1
ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ "БАЛТО"         30964726        0, ВУЛ. КАУНАСЬКА, БУД. 10 - А, ОФ. 604, М.КИЇВ, ДНІПРОВСЬКИЙ, УКРАЇНА  1
СПІЛЬНЕ ЗАКРИТЕ АКЦІОНЕРНЕ ТОВАРИСТВО УКРАЇНСЬКО-КІПРСЬКЕ "ДІМЕКС"             20037985        0, ВУЛ. КИБАЛЬЧИЧА, БУД.8, К.88, М.КИЇВ, УКРАЇНА       1
ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ "ГАРАНТ - СЕРВІС"               24743898        0, ВУЛ. КОСМІЧНА, БУД. 17, М.КИЇВ, УКРАЇНА      1
Time taken: 38.783 seconds, Fetched: 20 row(s)
hive>
```

1.3. Perform join.

```
SELECT
*
FROM
public.dict_uo uo
join dict_fop fop on uo.address = fop.address
join dict_fop fop1 on uo.address = fop1.address
```

* Run this task on the RDBMS server (oracle, mysq, mssql, etc.). Compare the query execution time.

-- join



!!! 1.4 - 1.6 might be executing by own source with > 1M rows.

1.4 Using PARTITIONED BY

Upload file Lending Club Loans_synthetic1.csv to the cluster.

Create managed table LENDING_CLUB using PARTITIONED BY constructions using "verification" fields.

Load data to LENDING_CLUB.

Show list of all table partitions.

Measure query time execution selection from different partitions.

1.5 Create table LENDING_CLUB_BUCKETS

Create managed table LENDING_CLUB_BUCKETS using BUCKETS for "working_years" field (10 BUCKETS).

Load data to LENDING_CLUB_BUCKETS.

Show bucket files for some partition.

Measure query time execution selection * from LENDING_CLUB_BUCKETS where working_years =??? (where ??? - is some values). .

1.6 Export data from LENDING_CLUB_BUCKETS

Export data from LENDING_CLUB_BUCKETSto different directories using distinct "working_years" values.

1.7 Counting the number of words in the files.

////////////////

```
hadoop jar hadoop-examples-1.2.1.jar wordcount
hadoop jar hadoop-examples-1.2.1.jar wordcount input output
```

-- counting the number of words, where "…" folder path to texy file, for example https://drive.google.com/drive/u/1/folders/1yPheTh7tLD27UNrUWCSUGzYyg-df5HZv

```
hadoop jar hadoop-examples-1.2.1.jar wordcount /…
```

9

```
17/02/21 18:36:54 INFO mapreduce.Job: Running job: job_1487698271776_0006
17/02/21 18:37:02 INFO mapreduce.Job: Job job_1487698271776_0006 running in uber mode : fal
17/02/21 18:37:02 INFO mapreduce.Job:  map 0% reduce 0%
17/02/21 18:37:14 INFO mapreduce.Job:  map 9% reduce 0%
17/02/21 18:37:17 INFO mapreduce.Job:  map 15% reduce 0%
17/02/21 18:37:18 INFO mapreduce.Job:  map 40% reduce 0%
17/02/21 18:37:19 INFO mapreduce.Job:  map 44% reduce 0%
17/02/21 18:37:20 INFO mapreduce.Job:  map 50% reduce 0%
17/02/21 18:37:23 INFO mapreduce.Job:  map 52% reduce 0%
17/02/21 18:37:25 INFO mapreduce.Job:  map 59% reduce 0%
17/02/21 18:37:26 INFO mapreduce.Job:  map 65% reduce 0%
17/02/21 18:37:27 INFO mapreduce.Job:  map 68% reduce 0%
17/02/21 18:37:28 INFO mapreduce.Job:  map 73% reduce 0%
17/02/21 18:37:32 INFO mapreduce.Job:  map 73% reduce 4%
17/02/21 18:37:33 INFO mapreduce.Job:  map 76% reduce 8%
17/02/21 18:37:34 INFO mapreduce.Job:  map 81% reduce 10%
17/02/21 18:37:36 INFO mapreduce.Job:  map 82% reduce 10%
17/02/21 18:37:37 INFO mapreduce.Job:  map 83% reduce 10%
17/02/21 18:37:39 INFO mapreduce.Job:  map 84% reduce 10%
17/02/21 18:37:40 INFO mapreduce.Job:  map 85% reduce 10%
17/02/21 18:37:42 INFO mapreduce.Job:  map 95% reduce 10%
17/02/21 18:37:43 INFO mapreduce.Job:  map 100% reduce 11%
17/02/21 18:37:44 INFO mapreduce.Job:  map 100% reduce 24%
17/02/21 18:37:45 INFO mapreduce.Job:  map 100% reduce 48%
17/02/21 18:37:46 INFO mapreduce.Job:  map 100% reduce 63%
17/02/21 18:37:49 INFO mapreduce.Job:  map 100% reduce 88%
17/02/21 18:37:51 INFO mapreduce.Job:  map 100% reduce 100%
17/02/21 18:37:51 INFO mapreduce.Job: Job job_1487698271776_0006 completed successfully
17/02/21 18:37:51 INFO mapreduce.Job: Counters: 52
```

```
	Job Counters
		Killed map tasks=2
		Killed reduce tasks=1
		Launched map tasks=7
		Launched reduce tasks=8
		Data-local map tasks=5
		Rack-local map tasks=2
		Total time spent by all maps in occupied slots (ms)=578229
		Total time spent by all reduces in occupied slots (ms)=789090
		Total time spent by all map tasks (ms)=192743
		Total time spent by all reduce tasks (ms)=131515
		Total vcore-milliseconds taken by all map tasks=192743
		Total vcore-milliseconds taken by all reduce tasks=263030
		Total megabyte-milliseconds taken by all map tasks=592106496
		Total megabyte-milliseconds taken by all reduce tasks=808028160
	Map-Reduce Framework
		Map input records=1479696
		Map output records=35975497
		Map output bytes=830441642
		Map output materialized bytes=153008126
		Input split bytes=726
		Combine input records=41065351
		Combine output records=9507113
		Reduce input groups=3548659
		Reduce shuffle bytes=153008126
		Reduce input records=4417259
		Reduce output records=3548659
		Spilled Records=13924372
		Shuffled Maps =48
		Failed Shuffles=0
		Merged Map outputs=48
		GC time elapsed (ms)=2159
		CPU time spent (ms)=171460
		Physical memory (bytes) snapshot=5879169024
		Virtual memory (bytes) snapshot=82795200512
		Total committed heap usage (bytes)=5885657088
	Shuffle Errors
		BAD_ID=0
		CONNECTION=0
		IO_ERROR=0
		WRONG_LENGTH=0
		WRONG_MAP=0
		WRONG_REDUCE=0
	File Input Format Counters
		Bytes Read=688296966
	File Output Format Counters
		Bytes Written=114102662
```

```
hadoop fs -tail /user/oliyura/output/part-r-00000
```

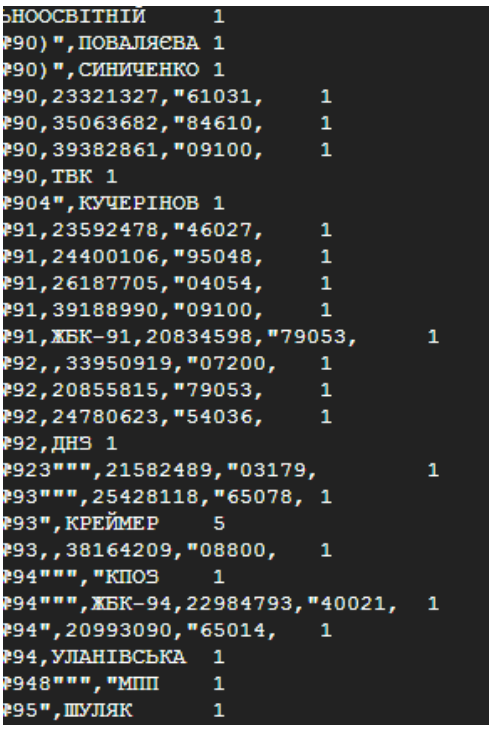


Figure. Counting words

Requirements for a computer laboratory workshop report

A computer worksheet report is executed in Microsoft Word and must contain the following sections:

1. Title.
2. Description of the task option.
3. SQL queries and performance results on Hadoop.
4. SQL queries and performance results on the RDBMS.
5. Comment on the Hadoop request log.

Links: [1,2].

Links

1. Harness the Power of Big Data. Paul C. Zikopoulos, Dirk deRoos, Krishnan Parasuraman, Thomas Deutsch, David Corrigan, James Giles. ISBN: 978-0-07180818-7
2. Big Data Beyond the Hype. A Guide to Conversations for Today's Data Center. Paul Zikopoulos Dirk deRoos, Christopher Bienko, Rick Buglio, Marc Andrews. ISBN: 978-0-07-184466-6
3. Чак Лэм. Hadoop в действии. - М.: ДМК Пресс, 2012. - 424с: ил. ISBN 978-5-94074-785-7.
4. Edward Capriolo, Dean Wampler, and Jason Rutherglen. Programming Hive. ISBN: 978-1-449-31933-5
5. https://drive.google.com/open?id=159vqraxrOrrZej2CqaOJzxe0wAuZyWcr