Урок 2

1. Поместите датасет ppkm_sentiment у себя в HDFS и дайте всем пользователям на них полные права
2. Определите расположение блоков
3. У вас 20 файлов, каждый размером 130 мб. Сколько блоков будет аллоцировано в NameNode, при условии, что размер блока по умолчанию у вас 128 мб, а фактор репликации равен 3?
4. У вас 1 файл, размером 1.56 Тб. Сколько блоков будет аллоцировано в NameNode, при условии, что размер блока по умолчанию у вас 128 мб, а фактор репликации равен 3?
5. В вашей компании развернут Hadoop кластер из 400 нод. Фактор репликации равен 3. Сколько единовременно может быть выведено машин из строя, чтобы не было потери данных?

Задание 1

Переносим данные в hdfs в директории ppkm

```
hduser@localhost:~$ ls
archive.zip  hadoop
hduser@localhost:~$ unzip archive.zip -d ppkm
Archive:  archive.zip
  inflating: ppkm/ppkm_dataset.csv
  inflating: ppkm/ppkm_test.csv
  inflating: ppkm/stopwordv1.txt
hduser@localhost:~$ ls
archive.zip  hadoop  ppkm
hduser@localhost:~$ rm archive.zip
rm: remove write-protected regular file 'archive.zip'? yes
hduser@localhost:~$ hdfs dfs -put ppkm/ /user/hduser/
hduser@localhost:~$ hadoop fs -ls /user/hduser/ppkm
Found 3 items
-rw-r--r--   1 hduser supergroup      43320 2022-05-15 14:07 /user/hduser/ppkm/ppkm_dataset.csv
-rw-r--r--   1 hduser supergroup        476 2022-05-15 14:07 /user/hduser/ppkm/ppkm_test.csv
-rw-r--r--   1 hduser supergroup       4015 2022-05-15 14:07 /user/hduser/ppkm/stopwordv1.txt
hduser@localhost:~$
```

Изменяем права для всех файлов в директории ppkm

```
hduser@localhost:~$ hadoop fs -ls ppkm/
Found 3 items
-rw-r--r--   1 hduser supergroup      43320 2022-05-15 14:07 ppkm/ppkm_dataset.csv
-rw-r--r--   1 hduser supergroup        476 2022-05-15 14:07 ppkm/ppkm_test.csv
-rw-r--r--   1 hduser supergroup       4015 2022-05-15 14:07 ppkm/stopwordv1.txt
hduser@localhost:~$ hadoop fs -chmod -R 747 ppkm/
hduser@localhost:~$ hadoop fs -ls ppkm/
Found 3 items
-rwxr--rwx   1 hduser supergroup      43320 2022-05-15 14:07 ppkm/ppkm_dataset.csv
-rwxr--rwx   1 hduser supergroup        476 2022-05-15 14:07 ppkm/ppkm_test.csv
-rwxr--rwx   1 hduser supergroup       4015 2022-05-15 14:07 ppkm/stopwordv1.txt
hduser@localhost:~$ hadoop fs -chmod -R 777 ppkm/
hduser@localhost:~$ hadoop fs -ls ppkm/
Found 3 items
-rwxrwxrwx   1 hduser supergroup      43320 2022-05-15 14:07 ppkm/ppkm_dataset.csv
-rwxrwxrwx   1 hduser supergroup        476 2022-05-15 14:07 ppkm/ppkm_test.csv
-rwxrwxrwx   1 hduser supergroup       4015 2022-05-15 14:07 ppkm/stopwordv1.txt
hduser@localhost:~$
```

Задание 2

С помощью команды fsck и флагов -files -blocks -locations выводим имена блоков для каждого файла. Далее вывол содержимое файла ppkm_test.csv в локакльной файловой системе.

```
hduser@localhost:~$ hdfs fsck ppkm/ -files -blocks -locations | grep repl=1 -B1
Connecting to namenode via http://localhost:50070/fsck?ugi=hduser&files=1&blocks=1&locations=1&path=%2Fuser%2Fhduser%2Fppkm
/user/hduser/ppkm/ppkm_dataset.csv 43320 bytes, 1 block(s):  OK
0. BP-1748043258-127.0.0.1-1651472891094:blk_1073741825_1001 len=43320 Live_repl=1 [DatanodeInfoWithStorage[127.0.0.1:50010,DS-9a15f614-5bbd-453c-a900-affcc9488697,DISK]]
--
/user/hduser/ppkm/ppkm_test.csv 476 bytes, 1 block(s):  OK
0. BP-1748043258-127.0.0.1-1651472891094:blk_1073741826_1002 len=476 Live_repl=1 [DatanodeInfoWithStorage[127.0.0.1:50010,DS-9a15f614-5bbd-453c-a900-affcc9488697,DISK]]
--
/user/hduser/ppkm/stopwordv1.txt 4015 bytes, 1 block(s):  OK
0. BP-1748043258-127.0.0.1-1651472891094:blk_1073741827_1003 len=4015 Live_repl=1 [DatanodeInfoWithStorage[127.0.0.1:50010,DS-9a15f614-5bbd-453c-a900-affcc9488697,DISK]]
hduser@localhost:~$ cat ppkm/ppkm_test.csv
label,comment
positif,Siap kawal dan amankan kebijakan kemendes palaksanaan PPKM untuk kemajuan dan kesejahteraan masyarakat desa
negatif,Aturan ini lah itu lah. Repot bener. Dia bilang kecewa sm ppkm tapi mobilitas tetap meningkat. Simple sebetulnya. Pake otak yg jernih. Stop atau tutup spbu. Jngn keluarkan atau jual bbm. Sudah. Tamat
netral,Sebenernya jika dari kasus Covid pertama kali ada seharusnya kita semua di rumah cuma 2 minggu pasti Covid berhenti total sekarang
hduser@localhost:~$ cat /tmp/hadoop-hduser/
dfs/           nm-local-dir/
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/
data/          name/          namesecondary/
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/data/
current/     in_use.lock
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/data/current/
BP-1748043258-127.0.0.1-1651472891094/ VERSION
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/data/current/BP-1748043258-127.0.0.1-1651472891094/
current/        scanner.cursor  tmp/
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/data/current/BP-1748043258-127.0.0.1-1651472891094/current/
VERSION    dfsUsed    finalized/ rbw/
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/data/current/BP-1748043258-127.0.0.1-1651472891094/current/finalized/subdir0/subdir0/blk_1073741826
blk_1073741825          blk_1073741825_1001.meta  blk_1073741826          blk_1073741826_1002.meta  blk_1073741827          blk_1073741827_1003.meta
hduser@localhost:~$ cat /tmp/hadoop-hduser/dfs/data/current/BP-1748043258-127.0.0.1-1651472891094/current/finalized/subdir0/subdir0/blk_1073741826
label,comment
positif,Siap kawal dan amankan kebijakan kemendes palaksanaan PPKM untuk kemajuan dan kesejahteraan masyarakat desa
negatif,Aturan ini lah itu lah. Repot bener. Dia bilang kecewa sm ppkm tapi mobilitas tetap meningkat. Simple sebetulnya. Pake otak yg jernih. Stop atau tutup spbu. Jngn keluarkan atau jual bbm. Sudah. Tamat
netral,Sebenernya jika dari kasus Covid pertama kali ada seharusnya kita semua di rumah cuma 2 minggu pasti Covid berhenti total sekarang
hduser@localhost:~$
```

Задание 3 и 4

Если честно не очень понятно, что значит "сколько блоков будет алоцировано"? Это означает сколько записей будет в Name Node или о скольких блоках будет храниться информация в каждой записи.

Name Node хранит мапинг о соотвесвии между файлами и блоками. На нейм ноде каждому файлу соотвесвует одна запись. Поэтому для 20 файлов размером 130мб будет 20 записей и для одного файла размером 1.56Тб будет 1 запись.

В каждой записи хранится данные о всех блоках. То есть в случае если у нас файл весит 130мб, а размер блока 128, то на каждый файл будет выделено по 2 блока, а с учетом репликации 6. то есть для 20 файлов 120. Файл размером 1.56Тб займет 12780 блоков, с репликацией 38340.

Задание 5

Тоже не очень понятный вопрос. Это же зависит от заполненности нод и если среди этих 400 нод, Нейм нода, а нет Secondary, то достаточно вывести из строя одну)

Если считать что ноды заполнены полностью, то чтобы не было потери данных можно допустить выход из тсроя не больше 2/3 нод.