ДЗ 3

Кондраев Дмитрий

1 Какие плюсы и недостатки у Merge Sort Join в отличии от Hash Join? (1 балл)

- недостаток: необходимость сортировки по ключу ($O(n \log n)$ vs $O(n)$). Но если данные *уже* отсортированы по ключу, то Merge Sort Join – лучший выбор.
- преимущество: использует меньше памяти (не нужно хранить хеши)
- преимущество: эффективнее при сильно отличающихся размерах сливаемых датасетов (но в этом случае будет лучше Broadcast Join)

2 Соберите WordCount приложение, запустите на датасете `ppkm_sentiment` (1 балл)

Исходники см. в репозитории https://github.com/mariohuq/2022-polis-ml. Команды выполняются из корневой директории репозитория:

```
$ git clone https://github.com/mariohuq/2022-polis-ml.git kondraev-ml
$ cd kondraev-ml
```

1. Скачать образ `img-hdp-zeppelin`.

   ```
   $ cd 03-spark/img-hdp-zeppelin
   ```

2. Сборка образа:

   ```
   $ systemctl start docker.service
   $ docker build -t img-hdp-zeppelin .
   ```

3. Первый старт образа:

   ```
   $ docker run -it --name zeppelin \
     -p 50090:50090 \
     -p 50075:50075 \
     -p 50070:50070 \
     -p 8042:8042 \
     -p 8088:8088 \
     -p 4040:4040 \
     -p 4044:4044 \
     -p 8888:8888 \
     --hostname localhost \
     img-hdp-zeppelin
   ```

   Последующие запуски контейнера `zeppelin`:

   ```
   $ docker start -i zeppelin
   ```

4. Сборка WordCount:

   ```
   $ git checkout 4934a2eb
   $ cd 03-spark/WordCount
   $ sbt package
   $ docker cp ./**/wordcount_2.11-0.1.0-SNAPSHOT.jar zeppelin:/home/hduser/
   ```

5. Перенос файлов `ppkm_sentiment`

   ```
   $ docker cp 01-hadoop-setup/ppkm.zip zeppelin:/home/hduser/
   ```

Внутри контейнера:

```
hduser@localhost:~$ unzip ppkm.zip -d ppkm && rm ppkm.zip
hduser@localhost:~$ hdfs dfs -put ppkm /user/hduser/
hduser@localhost:~$ hdfs dfs -chmod -R a+r /user/hduser/ppkm
hduser@localhost:~$ rm -r ppkm
```

Последняя команда удаляет ppkm с хоста в docker-контейнере. Содержимое hdfs сохраняется между запусками контейнера.

6. Запуск Spark

```
hduser@localhost:~$ spark-submit \
  --name WordCount \
  --class ok.ml.WordCount \
  --master yarn \
  --deploy-mode cluster \
  wordcount_2.11-0.1.0-SNAPSHOT.jar \
  /user/hduser/ppkm/ppkm_dataset.csv /user/hduser/scala-ppkm-rdd-out
```

```
<output trimmed for brevity>
22/10/16 19:16:29 INFO yarn.Client: Application report for
 application_1665947673451_0001 (state: FINISHED)
22/10/16 19:16:29 INFO yarn.Client:
   client token: N/A
   diagnostics: N/A
   ApplicationMaster host: localhost
   ApplicationMaster RPC port: 33147
   queue: default
   start time: 1665947739933
   final status: SUCCEEDED
   tracking URL: http://localhost:8088/proxy/application_1665947673451_0001/
   user: hduser
22/10/16 19:16:29 INFO util.ShutdownHookManager: Shutdown hook called
22/10/16 19:16:29 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
886c3466-bca3-4a81-bb91-8cf61d8108e9
22/10/16 19:16:29 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
b8f8c57c-5c26-405e-b6d2-6742d4875240
```

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-rdd-out/part-*
```

```
(butuh,1)
(keluarganya,1)
(positif,"Kebijakan,1)
(izin,1)
(Amati,1)
<output trimmed for brevity>
(Kegiatan,39)
(,41)
(yg,55)
(Mikro,56)
(di,74)
(PPKM,84)
(dan,84)
```

3 Измените `WordCount` так, чтобы он удалял знаки препинания и приводил все слова к единому регистру (1 балл)

```
$ git checkout e258e56e
$ cd 03-spark/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.2.0-SNAPSHOT.jar zeppelin:/home/hduser/
```

```
hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-rdd-out
hduser@localhost:~$ spark-submit \
      --name WordCount \
      --class ok.ml.WordCount \
      --master yarn \
      --deploy-mode cluster \
      wordcount_2.11-0.2.0-SNAPSHOT.jar \
      /user/hduser/ppkm/ppkm_dataset.csv /user/hduser/scala-ppkm-rdd-out

<output trimmed for brevity>
22/10/16 20:06:56 INFO yarn.Client: Application report for
  application_1665950652687_0001 (state: FINISHED)
22/10/16 20:06:56 INFO yarn.Client:
     client token: N/A
     diagnostics: N/A
     ApplicationMaster host: localhost
     ApplicationMaster RPC port: 46693
     queue: default
     start time: 1665950760527
     final status: SUCCEEDED
     tracking URL: http://localhost:8088/proxy/application_1665950652687_0001/
     user: hduser
22/10/16 20:06:56 INFO util.ShutdownHookManager: Shutdown hook called
22/10/16 20:06:56 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
c9bccb01-32a7-4463-b800-f5c864b92ad8
22/10/16 20:06:56 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
ed186df1-80ca-4491-98f7-cdac12f9c89b

hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-rdd-out/part-*

(butuh,1)
(keluarganya,1)
(rapih,1)
(dipikir,1)
(tuhanku,1)
(tmmd,1)
(swt,1)
(dpt,1)
<output trimmed for brevity>
(covid,93)
(negatif,100)
(netral,100)
(positif,105)
(mikro,111)
(ppkm,138)
```

4 Измените выход WordCount так, чтобы сортировка была по количеству повторений, а список слов был во втором столбце, а не в первом (1 балл)

```
$ git checkout 7eeb63cd
$ cd 03-spark/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.3.0-SNAPSHOT.jar zeppelin:/home/hduser/

hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-rdd-out
hduser@localhost:~$ spark-submit \
      --name WordCount \
      --class ok.ml.WordCount \
      --master yarn \
      --deploy-mode cluster \
```

```
      wordcount_2.11-0.3.0-SNAPSHOT.jar \
      /user/hduser/ppkm/ppkm_dataset.csv /user/hduser/scala-ppkm-rdd-out

<output trimmed for brevity>
22/10/16 20:40:32 INFO yarn.Client: Application report for
  application_1665952592009_0001 (state: FINISHED)
22/10/16 20:40:32 INFO yarn.Client:
     client token: N/A
     diagnostics: N/A
     ApplicationMaster host: localhost
     ApplicationMaster RPC port: 40327
     queue: default
     start time: 1665952785576
     final status: SUCCEEDED
     tracking URL: http://localhost:8088/proxy/application_1665952592009_0001/
     user: hduser
22/10/16 20:40:32 INFO util.ShutdownHookManager: Shutdown hook called
22/10/16 20:40:32 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
b9f514ea-73d3-4187-89d3-5e8ff3145561
22/10/16 20:40:32 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
da7547a9-0ee8-42ca-8885-f9645bd776f4

hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-rdd-out/part-*

(1,butuh)
(1,keluarganya)
(1,rapih)
(1,dipikir)
(1,tuhanku)
(1,tmmd)
<output trimmed for brevity>
(83,https)
(84,t)
(85,co)
(89,dan)
(93,covid)
(100,negatif)
(100,netral)
(105,positif)
(111,mikro)
(138,ppkm)
```

5 Измените выход WordCount, чтобы формат соответствовал TSV (1 балл)

```
$ git checkout 10462539
$ cd 03-spark/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.4.0-SNAPSHOT.jar zeppelin:/home/hduser/

hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-rdd-out
hduser@localhost:~$ spark-submit \
      --name WordCount \
      --class ok.ml.WordCount \
      --master yarn \
      --deploy-mode cluster \
      wordcount_2.11-0.4.0-SNAPSHOT.jar \
      /user/hduser/ppkm/ppkm_dataset.csv /user/hduser/scala-ppkm-rdd-out

<output trimmed for brevity>
22/10/16 21:00:59 INFO yarn.Client: Application report for
  application_1665953938530_0001 (state: FINISHED)
```

```
22/10/16 21:00:59 INFO yarn.Client:
     client token: N/A
     diagnostics: N/A
     ApplicationMaster host: localhost
     ApplicationMaster RPC port: 36525
     queue: default
     start time: 1665954011828
     final status: SUCCEEDED
     tracking URL: http://localhost:8088/proxy/application_1665953938530_0001/
     user: hduser
22/10/16 21:00:59 INFO util.ShutdownHookManager: Shutdown hook called
22/10/16 21:00:59 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
b26113dc-b32e-4a34-8214-0444a8cecec9
22/10/16 21:00:59 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-
fab90866-1b4c-4a19-bc5a-06461876f9db
```

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-rdd-out/part-*
```

```
1   butuh
1   keluarganya
1   rapih
1   dipikir
1   tuhanku
<output trimmed for brevity>
83  https
84  t
85  co
89  dan
93  covid
100 negatif
100 netral
105 positif
111 mikro
138 ppkm
```

6 Добавьте в `WordCount` возможность через конфигурацию задать список стоп-слов, которые будут отфильтрованы во время работы приложения (2 балла)

Использовался список стоп-слов, который шел вместе с датасетом.

```
$ git checkout 9197665e
$ cd 03-spark/WordCount
$ sbt package
$ docker cp ./**/wordcount_2.11-0.5.0-SNAPSHOT.jar zeppelin:/home/hduser/
```

```
hduser@localhost:~$ hdfs dfs -rm -r /user/hduser/scala-ppkm-rdd-out
hduser@localhost:~$ spark-submit \
      --name WordCount \
      --class ok.ml.WordCount \
      --master yarn \
      --deploy-mode cluster \
      wordcount_2.11-0.5.0-SNAPSHOT.jar \
      /user/hduser/ppkm/ppkm_dataset.csv /user/hduser/scala-ppkm-rdd-out
  /user/hduser/ppkm/stopwordv1.txt
```

```
<output trimmed for brevity>
22/10/16 21:49:30 INFO yarn.Client: Application report for
  application_1665956851047_0001 (state: FINISHED)
22/10/16 21:49:30 INFO yarn.Client:
     client token: N/A
```

```
     diagnostics: N/A
     ApplicationMaster host: localhost
     ApplicationMaster RPC port: 40375
     queue: default
     start time: 1665956923327
     final status: SUCCEEDED
     tracking URL: http://localhost:8088/proxy/application_1665956851047_0001/
     user: hduser
22/10/16 21:49:30 INFO util.ShutdownHookManager: Shutdown hook called
22/10/16 21:49:30 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-9861a94d-21b7-4af7-b2a1-4ad4c9f29d90
22/10/16 21:49:30 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-401f1382-16ce-4ba0-b224-7285220f9661
```

```
hduser@localhost:~$ hdfs dfs -cat /user/hduser/scala-ppkm-rdd-out/part-*
```

```
1   butuh
1   keluarganya
1   rapih
1   dipikir
1   tuhanku
<output trimmed for brevity>
83  https
84  t
85  co
93  covid
100 negatif
100 netral
105 positif
111 mikro
138 ppkm
```

7 Выполните broadcast join на двух датасетах, не используя метод `join()`. Можно использовать любые предварительные трансформации. В качестве исходных данных возьмите `Company.csv` и `Company_Tweet.csv` из датасета Tweets about the Top Companies from 2005 to 2020 (3 балла)

Загрузим датасет:

```
$ cd 03-spark/top-tweets
$ docker cp Company.csv zeppelin:/home/hduser/
$ docker cp Company_Tweet.csv.zip zeppelin:/home/hduser/
```

```
hduser@localhost:~$ unzip Company_Tweet.csv.zip -d . && rm Company_Tweet.csv.zip
hduser@localhost:~$ mkdir tweets && mv Company* tweets
hduser@localhost:~$ hdfs dfs -put tweets /user/hduser/
```

Код (см. `03-spark/Tweets.zpln`):

```
sc.textFile("tweets/Company.csv").first
```

```
val rdd = sc.textFile("tweets/Company.csv").mapPartitionsWithIndex {
    case (0, iter) ⇒ iter.drop(1)
    case (_, iter) ⇒ iter
}.map { row ⇒ row.split(",") match { case Array(symbol, name) ⇒ (symbol, name) } }
val companies = sc.broadcast(rdd.collect.toMap)
```

```
sc.textFile("tweets/Company_Tweet.csv").first
```

```
// broadcast join
val rdd = sc.textFile("tweets/Company_Tweet.csv").mapPartitionsWithIndex {
```

```
    case (0, iter) ⇒ iter.drop(1)
    case (_, iter) ⇒ iter
}
.flatMap { row ⇒ row.split(",") match {
    case Array(tweet_id, symbol) ⇒ companies.value.get(symbol).map(name
      ⇒(tweet_id, symbol, name)) }
}

print("%table\ntweet_id\tticker_symbol\tcompany_name\n"
  + rdd.takeSample(true, 25)
        .map { case (tweet_id, symbol, name) ⇒ s"$tweet_id\t$symbol\t$name" }
        .mkString("\n")
)
```

Результат работы:

tweet_id	ticker_symbol	company_name
1145068424655233026	GOOGL	Google Inc
1042088134014640128	TSLA	Tesla Inc
902693623908761600	AMZN	Amazon.com
756179657817485312	GOOG	Google Inc
1118566580445765636	AMZN	Amazon.com
1103770464172863515	TSLA	Tesla Inc
1103136574311710720	TSLA	Tesla Inc
1055255442472005632	TSLA	Tesla Inc
931937913021050886	AAPL	apple
1077229363102400512	GOOG	Google Inc
1108058122252369920	GOOGL	Google Inc
895714874986201088	TSLA	Tesla Inc
925030317441687552	AAPL	apple
919922864169672704	TSLA	Tesla Inc
720216498162114560	TSLA	Tesla Inc
800647227622432768	AAPL	apple
958325688393719809	AMZN	Amazon.com
655796160389451776	AAPL	apple
992126203522945024	MSFT	Microsoft
777908550282600449	AAPL	apple
1144255402999238656	TSLA	Tesla Inc
846740988811001856	TSLA	Tesla Inc
1016425371204386816	TSLA	Tesla Inc
740745977556934658	GOOG	Google Inc
1187011924065767424	MSFT	Microsoft