Task 2: how many more visitors are increased in weekends than week days

Step1: To do this job I have created hive table

create table ckin_pivot as select business_id ,COUNT(IF(date='mon',business_id,null)) as monday,COUNT(IF(date='tue',business_id,null)) as tuesday,COUNT(IF(date ='wed',business_id,null)) as wednesday,COUNT(IF(date='thu',business_id,null)) as thursday,COUNT(IF(date='fri',business_id,null))as friday,COUNT(IF(date='sat',business_id,null)) as saturday,COUNT(IF(date='sun',business_id,null))as sunday from ckin c where c.date IN('mon'|'tue'|'wed') GROUP BY business_id;

BusinessID	Monday.	Tuesday.	Wednesday	Thursday	Friday	Saturday	Sunday

```
hive> select * from ckin_pivot;
OK
-49WY_TEa9ZEcRk_GnuLog    4     0     3     0     2     1     0
05yzn45mHlTCbT25KmFVlQ    0     0     0     0     2     9     0
0dcNFXs5yMj1n2RtkgIoHg    4     4     3     1     1     5     1
0FL1-ZegpmlG8P9kZEoZjw    0     0     1     0     0     0     0
0Fln0xZBGfxvRd66pbZ3ag    1     0     2     2     6    12    10
0VJlqjDftGcrtgRhbKrTQA    0     0     0     1     0     1     0
0wLKvB8ui5SZdlj-r8EWtQ    1     0     0     0     0     0     0
0WsCtzG8xazAmFA5HhQE0A    2     1     0     1     1     3     4
1-rUjW65plaAjIMVMJjP_Q    1     0     0     1     3     1     0
1EH3UCsWkTHH4Wlwed72JA    0     0     0     0     8     0     0
1RV1FeZKJN7JJ1nZpEIVWw    0     0     1     0     0     1     0
1Xi4hGRzhXZleq37hdFLcg    0     0     2     0     0     6     0
243PQVgkx8eUzVaKCgC-gw    0     0     1     0     0     7    11
2pmOI_mrn1ZWB88IeiOMkA    0     0     0     0     0     0     4
2wfiUn6AZm2KO0FJYakVQ     0     0     0     0     0     9     0
2X9U7v-Avoib-ki0y85bA     5     5     4     2     2     3     1
2YIg-PAgCMXMLxw53T29Q     0     0     6     4     0     0     2
3XBN60tLWqr7WZuDZsyZkA    0     0     0     0     0    11     0
3ZQxw4vY0HBuS7KeiWmyew    7     3     5     3     5     6     0
49WY_TEa9ZEcRk_GnuLog     8     0     6     0    11    14    15
4iuFGkrZYnWZgOaxNantdw    6     5     0     0     0     0     0
4k2o2LNXm9ibZ1lXCV80Ug    4     6     7    10     6     2     8
4zOyZff4-FemUqRW3GXuCA    3     1     1     2     2     2     1
5A2BcDVbWGM7JEUgSfFleA    0     0     0     2     1     4     5
5YyLWps6xL372Lj2hRlPQw    9     5    11    19     7    13    12
6FkCrxJMq5KIgavHRqtlbQ    0     0     1     0     0     2     3
6gDRAYnqwHDirfA5gF1sYg    0     0     0     0     0     3     0
6iq7LAJRSstbaYcLzQ2EzA    8     8     5     6     8    12    13
6PgUz2MCsI8V-7utG_42LA    0     0     1     0     1     2     0
6pLZrlG99_Ig8Okbsqo28A    2     4     4     9     4     4     3
6RxtO_MxnLwZ7D7iYDE8Jg    9     5     7     4     7    11    11
7KPBkxAOEtb3QeIL9PEErg   10    10    12    11    12    13    12
7rHrh5QgqFBXj40jF0NDCA    2     4     0     1     2     1     6
7sakqP2av046jujj-ERHhg   11     4     6    12    18    23    18
8bY6M2yiWOF2ilfmGS34Fw    0     0     0     1     1     1     0
8ZL4ed3Y8_F6AgPAYxgi2g    4     0     0     0     0     1     0
a0H7W1poihF8S_OQzBs8Fw    2     3     3     0     1     1     1
Al-oKob7eDmt-_Te9FHuzw    0     0     0     0     0     0     1
aLnnIPzhKT7CBMACh95CQg    1     4     0     4     3     0     0
```

Step2:

create table difference as select business_id,ROUND(saturday/2+sunday/2)-ROUND(monday/4+tuesday/4+thursday/4+wednesday/4+friday/4) as weekendmorevisitors from ckin_pivot where ROUND(monday/4+tuesday/4+thursday/4+wednesday/4+friday/4) < ROUND(saturday/2+sunday/2);

business_id   more visitors on weekend

```
Time taken: 0.065 seconds, Fetched: 182 row(s)
hive> select * from difference;
OK
05yzn45mHlTCbT25KmFVlQ    4.0
0Fln0xZBGfxvRd66pbZ3ag    8.0
0VJlqjDftGcrtgRhbKrTQA    1.0
0WsCtzG8xazAmFA5HhQE0A    3.0
1RV1FeZKJN7JJ1nZpEIVWw    1.0
1Xi4hGRzhXZleq37hdFLcg    2.0
243PQVgkx8eUzVaKCgC-gw    9.0
2pmOI_mrn1ZWB88IeiOMkA    2.0
2wfiUn6AZm2KO0FJYakVQ     5.0
3XBN60tLWqr7WZuDZsyZkA    6.0
49WY_TEa9ZEcRk_GnuLog     9.0
5A2BcDVbWGM7JEUgSfFleA    4.0
6FkCrxJMq5KIgavHRqtlbQ    3.0
6gDRAYnqwHDirfA5gF1sYg    2.0
6iq7LAJRSstbaYcLzQ2EzA    4.0
6RxtO_MxnLwZ7D7iYDE8Jg    3.0
7rHrh5QgqFBXj40jF0NDCA    2.0
7sakqP2av046jujj-ERHhg    8.0
Al-oKob7eDmt-_Te9FHuzw    1.0
b2596IQibXyco49ctXEupA    2.0
ba2xyZXG8hRb3hh7uDZs_g    3.0
bANAMVwSMR-UAA54WdRsIA    1.0
beQdq_FSb-kL2e2ch7Bt6g    1.0
BGeJhokuC2wqsSKtFz47uQ    1.0
BH0K6SxPnY3wTRB7tA27hQ    2.0
bOJrPR_G0eDzpw8rGEi2mA    1.0
cB04zZj6HyWA7bmhA5Zkcg    1.0
d24LIT55-aAr7l1X99vpjw    2.0
D2RoOb4UkH2azf8DgvNfUA    3.0
dPxL6P_Gp74CqSPuL59Srg    2.0
FpW25osFTrQf2dCIzAPavQ    3.0
Fu-egVPTQZ3lNYM0PhpQiQ    1.0
FZ_edvdCN64D7JVWvm7fmw    3.0
gCfRcOwBLJEF1C6ilqhFuA    1.0
gHX6fK4O7NwP03aH9qT2AQ    2.0
glY_PSkxyM54izTHkd2mtw    1.0
hCxM2MH3cDaqozUDpydXZg    1.0
HRIUU2Z4EsZBo59P0I2ZYA    2.0
i4y0CrKYLh_Qiv6JGyABJw    1.0
```

STEP3:
sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/yelp_db --username root --password cloudera --split-by business_id --table ckin --target-dir /user/cloudera/ckin --fields-terminated-by "," --hive-import --create-hive-table --hive-table yelp_db.ckin

```
                          FILE: Number of large read operations=0
                          FILE: Number of write operations=0
                          HDFS: Number of bytes read=857
                          HDFS: Number of bytes written=109593
                          HDFS: Number of read operations=24
                          HDFS: Number of large read operations=0
                          HDFS: Number of write operations=12
                  Job Counters
                          Killed map tasks=1
                          Launched map tasks=6
                          Other local map tasks=6
                          Total time spent by all maps in occupied slots (ms)=127955
                          Total time spent by all reduces in occupied slots (ms)=0
                          Total time spent by all map tasks (ms)=127955
                          Total vcore-milliseconds taken by all map tasks=127955
                          Total megabyte-milliseconds taken by all map tasks=131025920
                  Map-Reduce Framework
                          Map input records=3257
                          Map output records=3257
                          Input split bytes=857
                          Spilled Records=0
                          Failed Shuffles=0
                          Merged Map outputs=0
                          GC time elapsed (ms)=2397
                          CPU time spent (ms)=4760
                          Physical memory (bytes) snapshot=958472192
                          Virtual memory (bytes) snapshot=9061216256
                          Total committed heap usage (bytes)=748683264
                  File Input Format Counters
                          Bytes Read=0
                  File Output Format Counters
                          Bytes Written=109593
18/05/23 08:28:54 INFO mapreduce.ImportJobBase: Transferred 107.0244 KB in 39.0609 seconds (2.7399 KB/sec)
18/05/23 08:28:54 INFO mapreduce.ImportJobBase: Retrieved 3257 records.
18/05/23 08:28:54 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ckin` AS t LIMIT 1
18/05/23 08:28:54 INFO hive.HiveImport: Loading uploaded data into Hive


Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties
OK
Time taken: 1.959 seconds
Loading data to table yelp_db.ckin
Table yelp_db.ckin stats: [numFiles=6, totalSize=109593]
OK
```

Step4:

sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/yelp_db --username root --password cloudera --split-by business_id --table ckin_pivot --target-dir /user/cloudera/ckin_pivot --fields-terminated-by "," --hive-import --create-hive-table --hive-table yelp_db.ckin_pivot

```
[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://quickstart.cloudera:3306/yelp_db --username root --password cloudera --split-by business_id --table ckin_p
get-dir /user/cloudera/ckin_pivot --fields-terminated-by "," --hive-import --create-hive-table --hive-table yelp_db.ckin_pivot
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
18/05/23 08:48:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
18/05/23 08:48:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
18/05/23 08:48:39 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
18/05/23 08:48:39 INFO tool.CodeGenTool: Beginning code generation
18/05/23 08:48:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ckin_pivot` AS t LIMIT 1
18/05/23 08:48:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ckin_pivot` AS t LIMIT 1
18/05/23 08:48:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/lib/hadoop-mapreduce
Note: /tmp/sqoop-cloudera/compile/3940a3258e622cce2d6ff6b87c9a6a4f/ckin_pivot.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
18/05/23 08:48:42 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/3940a3258e622cce2d6ff6b87c9a6a4f/ckin_pivot.jar
18/05/23 08:48:42 WARN manager.MySQLManager: It looks like you are importing from mysql.
18/05/23 08:48:42 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
18/05/23 08:48:42 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
18/05/23 08:48:42 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
18/05/23 08:48:42 INFO mapreduce.ImportJobBase: Beginning import of ckin_pivot
18/05/23 08:48:42 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
18/05/23 08:48:42 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
18/05/23 08:48:43 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
18/05/23 08:48:43 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
18/05/23 08:48:44 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1281)
        at java.lang.Thread.join(Thread.java:1355)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:952)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:690)
        at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:879)
18/05/23 08:48:45 INFO db.DBInputFormat: Using read commited transaction isolation
18/05/23 08:48:45 INFO db.DataDrivenDBInputFormat: BoundingValsQuery: SELECT MIN(`business_id`), MAX(`business_id`) FROM `ckin_pivot`
18/05/23 08:48:45 WARN db.TextSplitter: Generating splits for a textual index column.
18/05/23 08:48:45 WARN db.TextSplitter: If your database sorts in a case-insensitive order, this may result in a partial import or duplicate records.
18/05/23 08:48:45 WARN db.TextSplitter: You are strongly encouraged to choose an integral split column.
18/05/23 08:48:45 INFO mapreduce.JobSubmitter: number of splits:6
18/05/23 08:48:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1526588254685_0005
18/05/23 08:48:45 INFO impl.YarnClientImpl: Submitted application application_1526588254685_0005
18/05/23 08:48:45 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1526588254685_0005/
18/05/23 08:48:45 INFO mapreduce.Job: Running job: job_1526588254685_0005
18/05/23 08:48:53 INFO mapreduce.Job: Job job_1526588254685_0005 running in uber mode : false
18/05/23 08:48:53 INFO mapreduce.Job:  map 0% reduce 0%
```

File Edit View Search Terminal Help

```
                HDFS: Number of bytes read=857
                HDFS: Number of bytes written=6801
                HDFS: Number of read operations=24
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=12
        Job Counters
                Killed map tasks=1
                Launched map tasks=6
                Other local map tasks=6
                Total time spent by all maps in occupied slots (ms)=122785
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=122785
                Total vcore-milliseconds taken by all map tasks=122785
                Total megabyte-milliseconds taken by all map tasks=125731840
        Map-Reduce Framework
                Map input records=182
                Map output records=182
                Input split bytes=857
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2031
                CPU time spent (ms)=4470
                Physical memory (bytes) snapshot=949485568
                Virtual memory (bytes) snapshot=9060782080
                Total committed heap usage (bytes)=748683264
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=6801
18/05/23 08:49:21 INFO mapreduce.ImportJobBase: Transferred 6.6416 KB in 38.1557 seconds (178.2436 bytes/sec)
18/05/23 08:49:21 INFO mapreduce.ImportJobBase: Retrieved 182 records.
18/05/23 08:49:21 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `ckin_pivot` AS t LIMIT 1
18/05/23 08:49:21 INFO hive.HiveImport: Loading uploaded data into Hive


Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties
OK
Time taken: 1.76 seconds
Loading data to table yelp_db.ckin_pivot
Table yelp_db.ckin_pivot stats: [numFiles=6, totalSize=6801]
OK
Time taken: 0.386 seconds
[cloudera@quickstart ~]$
```

```
                                        cloudera@quickstart:~                                    _ □ X

File  Edit  View  Search  Terminal  Help
hive> SET hive.auto.convert.join=false;
hive> select b.name,b.city,b.state,d.weekendmorevisitors as visitors_on_weekend from difference d,business b where d.business_id = b.id;
Query ID = cloudera_20180523092525_82879134-8236-4a3e-8006-a8687180b827
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1526588254685_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1526588254685_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1526588254685_0008
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2018-05-23 09:25:08,614 Stage-1 map = 0%,  reduce = 0%
2018-05-23 09:25:19,572 Stage-1 map = 50%,  reduce = 0%, Cumulative CPU 1.22 sec
2018-05-23 09:25:20,605 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.33 sec
2018-05-23 09:25:28,084 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.84 sec
MapReduce Total cumulative CPU time: 3 seconds 840 msec
Ended Job = job_1526588254685_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 3.84 sec   HDFS Read: 70209 HDFS Write: 3334 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 840 msec
OK
Game Seven Grill        Phoenix AZ     4.0
Circle K        Phoenix AZ     8.0
Smile! Dental Center    Ste 30"      Gibsonia      1.0
Happy Moose Bar and Grill      Streetsboro    OH    3.0
Beaumont Tire & Auto Repair    Champaign      IL    1.0
Jody Maroni's Sausage Kingdom    3790 Las Vegas Blvd S" Las Vegas      2.0
Travelodge      Scottsdale    AZ     9.0
Tim Hortons    Markham ON     2.0
Men's Ultimate Grooming  Ste 103"      Gilbert 5.0
Epic Thrift     Las Vegas      NV     6.0
Politic Chester CHE    9.0
Double Play Sports Bar  Las Vegas      NV    4.0
The Independent  1 Floor"      Toronto 3.0
Knitting Tree   Madison WI     3.0
Rôtisserie Benny & Co   Sainte-Julie   QC    2.0
Wong's Jr      Peoria AZ     4.0
Rite Aid        Pittsburgh    PA     2.0
Travel Leaders  North Olmsted   OH     8.0
```

STEP5:
OUTPUT:
select b.name, b.city,b.state, d.weekendmorevisitors as visitors_on_weekend from difference
d,business b where d.business_id = b.id;
I have executed sqoop jobs to import data from my sql to hive database. The Map reduce jobs
are executed

name	city	state	visitors_on_weekend
Game Seven Grill	Phoenix	AZ	4
Circle K	Phoenix	AZ	8
Smile! Dental Center	Gibsonia	PA	1
Happy Moose Bar and Grill	Streetsboro	OH	3
Beaumont Tire & Auto Repair	Champaign	IL	1
Jody Maroni's Sausage Kingdom	Las Vegas	NV	2
Travelodge	Scottsdale	AZ	9
Tim Hortons	Markham	ON	2
Men's Ultimate Grooming	Gilbert	AZ	5
Epic Thrift	Las Vegas	NV	6
Politic	Chester	CHE	9
Double Play Sports Bar	Las Vegas	NV	4
The Independent	Toronto	ON	3
Rôtisserie Benny & Co	Sainte-Julie	QC	2
Wong's Jr	Peoria	AZ	4
Knitting Tree	Madison	WI	3
Rite Aid	Pittsburgh	PA	2
Travel Leaders	North Olmsted	OH	8
Bridal Fashion Debut	Phoenix	AZ	1
URBAN JUNGLE FUN PARK Mesa	Mesa	AZ	2
Mike Olbinski Photography	Phoenix	AZ	3
Advantage Rent-A-Car	Phoenix	AZ	1
Deli 194	Edinburgh	EDH	1
Boulangerie Arouch	Laval	QC	1
Sunnyside Grill	Toronto	ON	2
The Fry	North York	ON	1
Centra Food Market	Aurora	ON	1
Domino's Pizza	Tempe	AZ	2
New Wrap Order	Charlotte	NC	3
World Food Championships	Las Vegas	NV	2
Catholic Charities	Las Vegas	NV	3
MV Nail Spa	North Olmsted	OH	1
Jose' and Tony's Mexican Restaurant	Pittsburgh	PA	3
Mulberry Bar	Toronto	ON	1
Wendy's	Stow	OH	2
Matt's Big Breakfast	Phoenix	AZ	1
Zonie Cuts	Phoenix	AZ	1