Assignment 16.3

Before beginning with the transactions in Hive, let's look at the ACID properties, which are vital for any transaction.

What is ACID?

ACID stands for Atomicity, Consistency, Isolation, and Durability.

Atomicity means, a transaction should complete successfully or else it should fail completely i.e. it should not be left partially. Consistency ensures that any transaction will bring the database from one valid state to another state. Isolation states that every transaction should be independent of each other i.e. one transaction should not affect another. And Durability states that if a transaction is completed, it should be preserved in the database even if the machine state is lost or a system failure might occur.

These ACID properties are essential for a transaction and every transaction should ensure that these properties are met.

Transactions in Hive:

Below are the constraints before we start with Hive Transaction:

1. **Hive 0.14 must be installed since this version supports ACID properties of transaction.**
2. **File should be of ORC format.**
3. **Table needs to be bucketed.**

Enabling Transaction Feature in Hive:

Mysql service is running and hadoop daemons are running

transaction features present in Hive needs to be turned on, as by default they are turned off.

The below properties need to be set appropriately in **hive shell,** order-wise to work with transactions in Hive:

hive>set hive.support.concurrency = true;
hive>set hive.enforce.bucketing = true;
hive>set hive.exec.dynamic.partition.mode = nonstrict;
hive>set hive.txn.manager = org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;
hive>set hive.compactor.initiator.on = true;
hive>set hive.compactor.worker.threads =2;

```
hive> set hive.support.concurrency = true;
hive> set hive.enforce.bucketing = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive> set hive.txn.manager = org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;
hive> set hive.compactor.initiator.on = true;
hive> set hive.compactor.worker.threads =2;
```

If the above properties are not set properly, the 'Insert' operation will work but 'Update' and 'Delete' will not work

Creating a Table That Supports Hive Transactions

CREATE TABLE college
(
clg_id INT,
clg_name STRING,
clg_loc STRING
)
clustered by (clg_id) into 5 buckets
stored as orc
TBLPROPERTIES('transactional'='true');

```
hive> CREATE TABLE college(clg_id int,clg_name string,clg_loc string) clustered by (clg_id) into 5 buckets stored as orc TBLPROPERTIES('transactional'='tr
ue');
FAILED: LockException [Error 10280]: Error communicating with the metastore
```

Error received, upon further searching on web I have found a solution.

Changes made in hive-site.xml

```
<configuration>
   <property>
      <name>javax.jdo.option.ConnectionURL</name>
      <value>jdbc:mysql://localhost/metastore?createDatabaseIfNotExist=true</value>
      <description>metadata is stored in a MySQL server</description>
   </property>
   <property>
      <name>javax.jdo.option.ConnectionDriverName</name>
      <value>com.mysql.jdbc.Driver</value>
      <description>MySQL JDBC driver class</description>
   </property>
   <property>
      <name>javax.jdo.option.ConnectionUserName</name>
      <value>hiveuser</value>
      <description>user name for connecting to mysql server</description>
   </property>
   <property>
      <name>javax.jdo.option.ConnectionPassword</name>
      <value>password</value>
      <description>password for connecting to mysql server</description>
   </property>
   <property>
        <name>hive.in.test</name>
        <value>true</value>
   </property>
</configuration>
~
```

```
hive> CREATE TABLE college
    > (
    > clg_id INT,
    > clg_name STRING,
    > clg_loc STRING
    > )
    > clustered by (clg_id) into 5 buckets
    > stored as orc
    > TBLPROPERTIES('transactional'='true');
OK
Time taken: 0.365 seconds
hive> DESCRIBE college;
OK
clg_id                  int
clg_name                string
clg_loc                 string
Time taken: 0.185 seconds, Fetched: 3 row(s)
hive>
```

Inserting Data into a Hive Table

INSERT INTO table college values
(1,'nec','nlr'),
(2,'vit','vlr'),
(3,'srm','chen'),
(4,'lpu','del'),
(5,'stanford','uk'),
(6,'JNTUA','atp'),
(7,'cambridge','us');

```
hive> INSERT INTO table college values
    > (1,'nec','nlr'),
    > (2,'vit','vlr'),
    > (3,'srm','chen'),
    > (4,'lpu','del'),
    > (5,'stanford','uk'),
    > (6,'JNTUA','atp'),
    > (7,'cambridge','us');
Query ID = acadgild_20170829202222_dc618d00-017d-4ee2-b323-d38538a9a1ee
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 5
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_1504017329401_0001, Tracking URL = http://localhost:8088/proxy/application_1504017329401_0001/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1504017329401_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 5
2017-08-29 20:22:26,517 Stage-1 map = 0%,  reduce = 0%
2017-08-29 20:22:33,016 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.08 sec
2017-08-29 20:22:49,086 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 1.91 sec
2017-08-29 20:22:53,782 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 2.85 sec
2017-08-29 20:22:54,924 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 3.76 sec
2017-08-29 20:22:57,239 Stage-1 map = 100%,  reduce = 57%, Cumulative CPU 5.42 sec
2017-08-29 20:22:58,380 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 5.89 sec
2017-08-29 20:22:59,487 Stage-1 map = 100%,  reduce = 99%, Cumulative CPU 8.56 sec
2017-08-29 20:23:00,517 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.25 sec
MapReduce Total cumulative CPU time: 9 seconds 250 msec
Ended Job = job_1504017329401_0001
Loading data to table default.college
Table default.college stats: [numFiles=5, numRows=7, totalSize=3474, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 5   Cumulative CPU: 9.25 sec   HDFS Read: 358 HDFS Write: 3849 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 250 msec
OK
Time taken: 47.685 seconds
hive>
```

Check data added previously:

*select * from college;*

```
hive> select * from college;
OK
5       stanford        uk
6       JNTUA   atp
1       nec     nlr
7       cambridge       us
2       vit     vlr
3       srm     chen
4       lpu     del
Time taken: 0.159 seconds, Fetched: 7 row(s)
hive>
```

Insert again to check if data is appended in table:

INSERT INTO table college values
(8,'sit','ind');

```
hive> INSERT INTO table college values
    > (8,'sit','ind');
Query ID = acadgild_20170829202424_97e26ff1-2024-4260-8c66-4645d517be87
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 5
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_1504017329401_0002, Tracking URL = http://localhost:8088/proxy/application_1504017329401_0002/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1504017329401_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 5
2017-08-29 20:24:14,620 Stage-1 map = 0%,  reduce = 0%
2017-08-29 20:24:20,936 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.95 sec
2017-08-29 20:24:36,736 Stage-1 map = 100%,  reduce = 13%, Cumulative CPU 1.95 sec
2017-08-29 20:24:40,180 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 2.67 sec
2017-08-29 20:24:42,428 Stage-1 map = 100%,  reduce = 33%, Cumulative CPU 3.74 sec
2017-08-29 20:24:44,716 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 4.38 sec
2017-08-29 20:24:45,827 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 5.55 sec
2017-08-29 20:24:46,897 Stage-1 map = 100%,  reduce = 79%, Cumulative CPU 7.39 sec
2017-08-29 20:24:48,053 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.52 sec
MapReduce Total cumulative CPU time: 9 seconds 520 msec
Ended Job = job_1504017329401_0002
Loading data to table default.college
Table default.college stats: [numFiles=10, numRows=8, totalSize=4979, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 5   Cumulative CPU: 9.52 sec   HDFS Read: 286 HDFS Write: 1760 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 520 msec
OK
Time taken: 43.36 seconds
hive>
```

Check data added previously:

*select * from college;*

```
hive> select * from college;
OK
5       stanford        uk
6       JNTUA   atp
1       nec     nlr
7       cambridge       us
2       vit     vlr
3       srm     chen
8       sit     ind
4       lpu     del
Time taken: 0.065 seconds, Fetched: 8 row(s)
hive>
```

Updating the Data in Hive Table:

We cannot perfom update operation on the bucketed column in our case clg_id

UPDATE college set clg_name = 'iit' where clg_id = 8;

```
hive> UPDATE college set clg_name = 'iit' where clg_id = 8;
Query ID = acadgild_20170829202525_fe8dc4c5-38d5-42fd-ba69-dbee4e9674c6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 5
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_1504017329401_0003, Tracking URL = http://localhost:8088/proxy/application_1504017329401_0003/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1504017329401_0003
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 5
2017-08-29 20:25:51,364 Stage-1 map = 0%,  reduce = 0%
2017-08-29 20:26:09,198 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 1.49 sec
2017-08-29 20:26:12,555 Stage-1 map = 40%,  reduce = 0%, Cumulative CPU 2.83 sec
2017-08-29 20:26:14,856 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 4.21 sec
2017-08-29 20:26:18,308 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.05 sec
2017-08-29 20:26:31,935 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 8.07 sec
2017-08-29 20:26:33,055 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 9.16 sec
2017-08-29 20:26:34,118 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 10.29 sec
2017-08-29 20:26:35,201 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.61 sec
MapReduce Total cumulative CPU time: 12 seconds 610 msec
Ended Job = job_1504017329401_0003
Loading data to table default.college
Table default.college stats: [numFiles=11, numRows=8, totalSize=5658, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 5   Cumulative CPU: 12.61 sec   HDFS Read: 7111 HDFS Write: 918 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 610 msec
OK
Time taken: 53.521 seconds
hive>
```

Check table updated previously:

*select * from college;*

```
hive> select * from college;
OK
5       stanford        uk
6       JNTUA   atp
1       nec     nlr
7       cambridge       us
2       vit     vlr
3       srm     chen
8       iit     ind
4       lpu     del
Time taken: 0.051 seconds, Fetched: 8 row(s)
hive>
```

Deleting a Row from Hive Table:

DELETE from college WHERE clg_id=8;

```
hive> DELETE from college WHERE clg_id=8;
Query ID = acadgild_20170829202727_4cbeaaf3-294f-47e6-a4cf-190db5d905f2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 5
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_1504017329401_0004, Tracking URL = http://localhost:8088/proxy/application_1504017329401_0004/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1504017329401_0004
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 5
2017-08-29 20:27:23,754 Stage-1 map = 0%,  reduce = 0%
2017-08-29 20:27:41,510 Stage-1 map = 20%,  reduce = 0%, Cumulative CPU 1.19 sec
2017-08-29 20:27:44,923 Stage-1 map = 40%,  reduce = 0%, Cumulative CPU 2.61 sec
2017-08-29 20:27:47,170 Stage-1 map = 60%,  reduce = 0%, Cumulative CPU 4.03 sec
2017-08-29 20:27:49,497 Stage-1 map = 80%,  reduce = 0%, Cumulative CPU 5.55 sec
2017-08-29 20:27:50,667 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.84 sec
2017-08-29 20:28:04,425 Stage-1 map = 100%,  reduce = 20%, Cumulative CPU 7.94 sec
2017-08-29 20:28:05,536 Stage-1 map = 100%,  reduce = 40%, Cumulative CPU 9.01 sec
2017-08-29 20:28:07,727 Stage-1 map = 100%,  reduce = 60%, Cumulative CPU 10.18 sec
2017-08-29 20:28:08,860 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.57 sec
MapReduce Total cumulative CPU time: 12 seconds 570 msec
Ended Job = job_1504017329401_0004
Loading data to table default.college
Table default.college stats: [numFiles=12, numRows=7, totalSize=6151, rawDataSize=0]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 5   Cumulative CPU: 12.57 sec   HDFS Read: 8006 HDFS Write: 732 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 570 msec
OK
Time taken: 54.639 seconds
hive>
```

Check table where row with clg_id =8 has been deleted previously:

*select * from college;*

```
hive> select * from college;
OK
5       stanford        uk
6       JNTUA   atp
1       nec     nlr
7       cambridge       us
2       vit     vlr
3       srm     chen
4       lpu     del
Time taken: 0.052 seconds, Fetched: 7 row(s)
hive>
```