Carbondata Auto Cleanup

Background:

Openlookeng integration with Carbondata not supporting auto cleanup of vacuum/compaction tables. This becomes the limitation of openlookeng since other processing engine like spark provides the same.

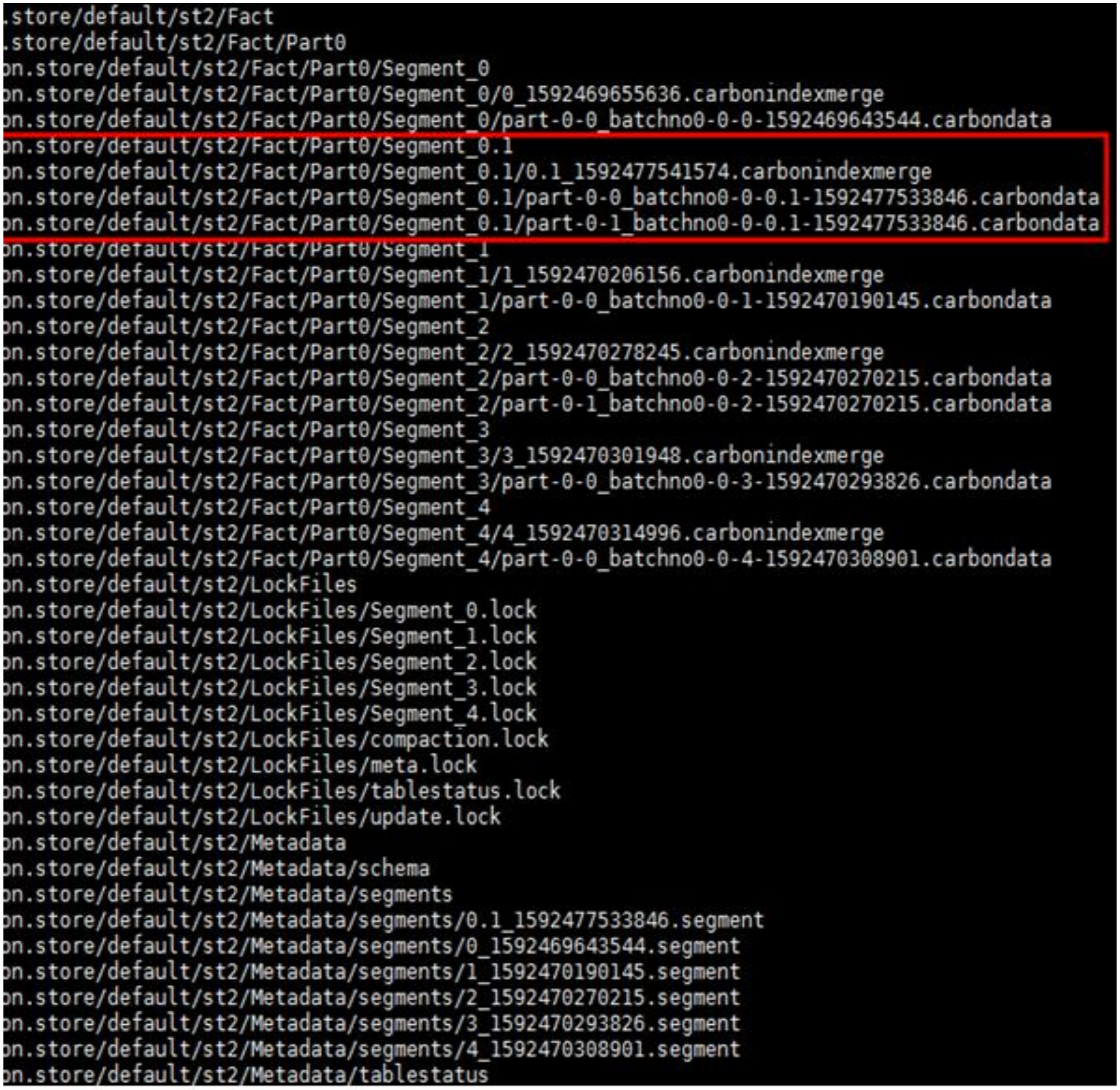
Objective:

After vacuum/compaction operation is completed on carbon tables, there will be unused base/ stale files which are left in HDFS. So Carbondata Auto cleanup is used to cleanup those files automatically.

It is will clean up the COMPACTED files which crossed more than 60 minutes.

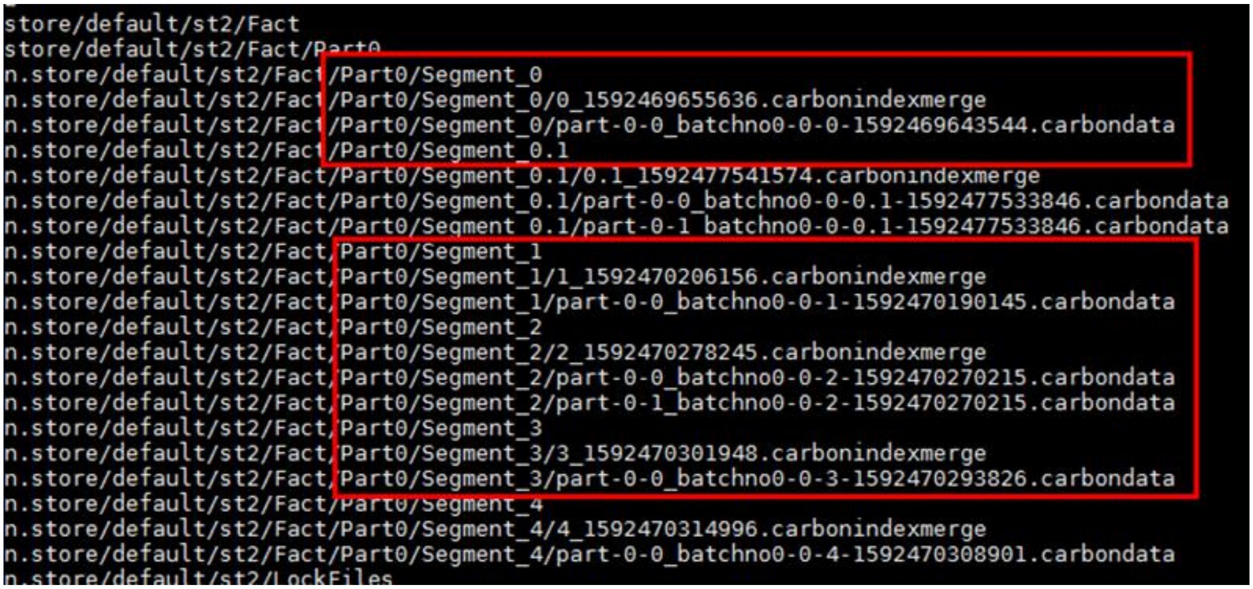
Auto cleanup will be triggered in case of insert, update, delete & vacuum.

After minor compaction



```
.store/default/st2/Fact
.store/default/st2/Fact/Part0
on.store/default/st2/Fact/Part0/Segment_0
on.store/default/st2/Fact/Part0/Segment_0/0_1592469655636.carbonindexmerge
on.store/default/st2/Fact/Part0/Segment_0/part-0-0_batchno0-0-0-1592469643544.carbondata
on.store/default/st2/Fact/Part0/Segment_0.1
on.store/default/st2/Fact/Part0/Segment_0.1/0.1_1592477541574.carbonindexmerge
on.store/default/st2/Fact/Part0/Segment_0.1/part-0-0_batchno0-0-0.1-1592477533846.carbondata
on.store/default/st2/Fact/Part0/Segment_0.1/part-0-1_batchno0-0-0.1-1592477533846.carbondata
on.store/default/st2/Fact/Part0/Segment_1
on.store/default/st2/Fact/Part0/Segment_1/1_1592470206156.carbonindexmerge
on.store/default/st2/Fact/Part0/Segment_1/part-0-0_batchno0-0-1-1592470190145.carbondata
on.store/default/st2/Fact/Part0/Segment_2
on.store/default/st2/Fact/Part0/Segment_2/2_1592470278245.carbonindexmerge
on.store/default/st2/Fact/Part0/Segment_2/part-0-0_batchno0-0-2-1592470270215.carbondata
on.store/default/st2/Fact/Part0/Segment_2/part-0-1_batchno0-0-2-1592470270215.carbondata
on.store/default/st2/Fact/Part0/Segment_3
on.store/default/st2/Fact/Part0/Segment_3/3_1592470301948.carbonindexmerge
on.store/default/st2/Fact/Part0/Segment_3/part-0-0_batchno0-0-3-1592470293826.carbondata
on.store/default/st2/Fact/Part0/Segment_4
on.store/default/st2/Fact/Part0/Segment_4/4_1592470314996.carbonindexmerge
on.store/default/st2/Fact/Part0/Segment_4/part-0-0_batchno0-0-4-1592470308901.carbondata
on.store/default/st2/LockFiles
on.store/default/st2/LockFiles/Segment_0.lock
on.store/default/st2/LockFiles/Segment_1.lock
on.store/default/st2/LockFiles/Segment_2.lock
on.store/default/st2/LockFiles/Segment_3.lock
on.store/default/st2/LockFiles/Segment_4.lock
on.store/default/st2/LockFiles/compaction.lock
on.store/default/st2/LockFiles/meta.lock
on.store/default/st2/LockFiles/tablestatus.lock
on.store/default/st2/LockFiles/update.lock
on.store/default/st2/Metadata
on.store/default/st2/Metadata/schema
on.store/default/st2/Metadata/segments
on.store/default/st2/Metadata/segments/0.1_1592477533846.segment
on.store/default/st2/Metadata/segments/0_1592469643544.segment
on.store/default/st2/Metadata/segments/1_1592470190145.segment
on.store/default/st2/Metadata/segments/2_1592470270215.segment
on.store/default/st2/Metadata/segments/3_1592470293826.segment
on.store/default/st2/Metadata/segments/4_1592470308901.segment
on.store/default/st2/Metadata/tablestatus
```

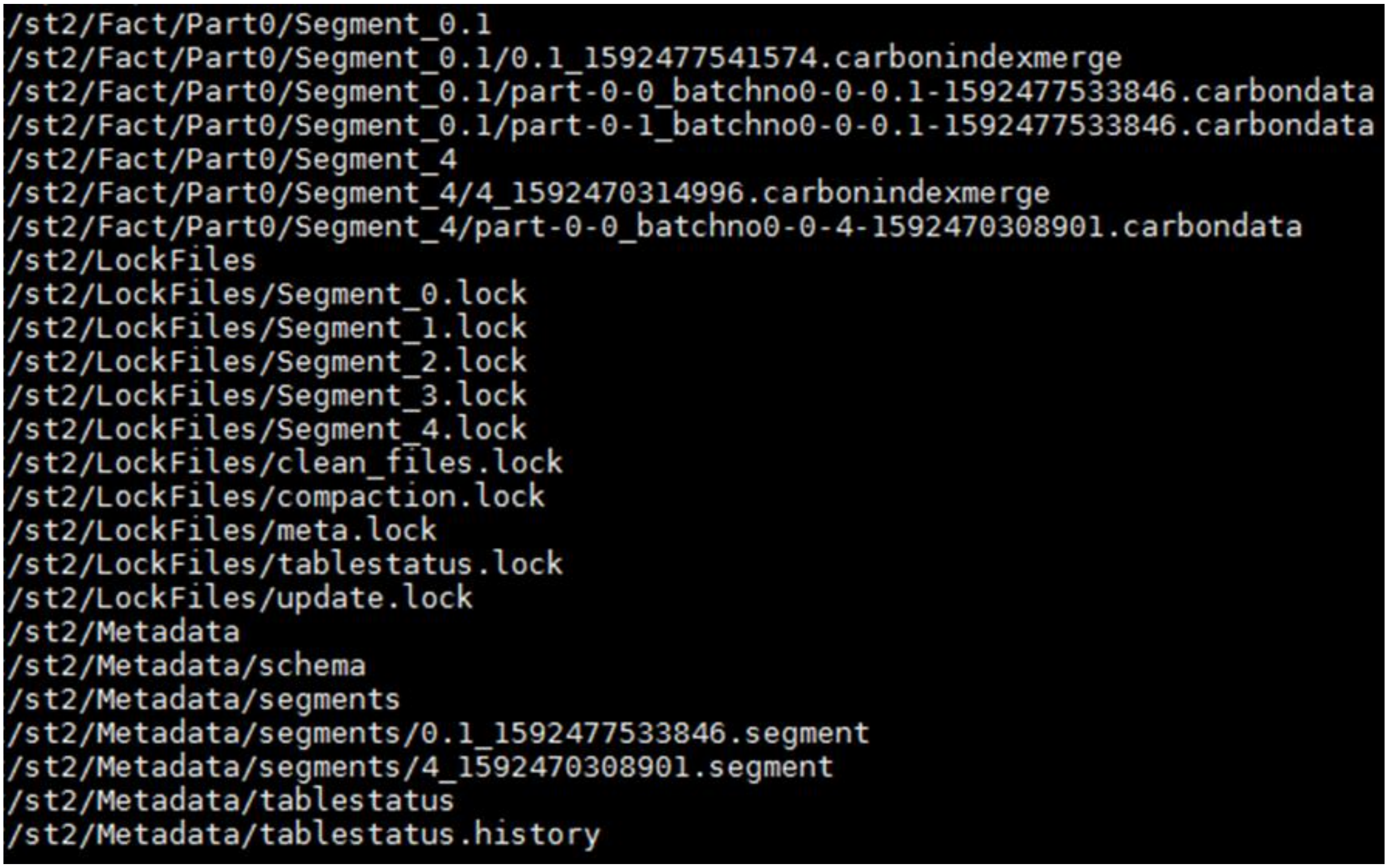
The file that will get cleanup in auto cleanup



```
store/default/st2/Fact
store/default/st2/Fact/Part0
n.store/default/st2/Fact/Part0/Segment_0
n.store/default/st2/Fact/Part0/Segment_0/0_1592469655636.carbonindexmerge
n.store/default/st2/Fact/Part0/Segment_0/part-0-0_batchno0-0-0-1592469643544.carbondata
n.store/default/st2/Fact/Part0/Segment_0.1
n.store/default/st2/Fact/Part0/Segment_0.1/0.1_1592477541574.carbonindexmerge
n.store/default/st2/Fact/Part0/Segment_0.1/part-0-0_batchno0-0-0.1-1592477533846.carbondata
n.store/default/st2/Fact/Part0/Segment_0.1/part-0-1_batchno0-0-0.1-1592477533846.carbondata
n.store/default/st2/Fact/Part0/Segment_1
n.store/default/st2/Fact/Part0/Segment_1/1_1592470206156.carbonindexmerge
n.store/default/st2/Fact/Part0/Segment_1/part-0-0_batchno0-0-1-1592470190145.carbondata
n.store/default/st2/Fact/Part0/Segment_2
n.store/default/st2/Fact/Part0/Segment_2/2_1592470278245.carbonindexmerge
n.store/default/st2/Fact/Part0/Segment_2/part-0-0_batchno0-0-2-1592470270215.carbondata
n.store/default/st2/Fact/Part0/Segment_2/part-0-1_batchno0-0-2-1592470270215.carbondata
n.store/default/st2/Fact/Part0/Segment_3
n.store/default/st2/Fact/Part0/Segment_3/3_1592470301948.carbonindexmerge
n.store/default/st2/Fact/Part0/Segment_3/part-0-0_batchno0-0-3-1592470293826.carbondata
n.store/default/st2/Fact/Part0/Segment_4
n.store/default/st2/Fact/Part0/Segment_4/4_1592470314996.carbonindexmerge
n.store/default/st2/Fact/Part0/Segment_4/part-0-0_batchno0-0-4-1592470308901.carbondata
n.store/default/st2/LockFiles
```

After auto cleanup

Cleanup the Fact/Part_x/segment folders , files .segment present in Metadata/segmets and also update tablestatus file

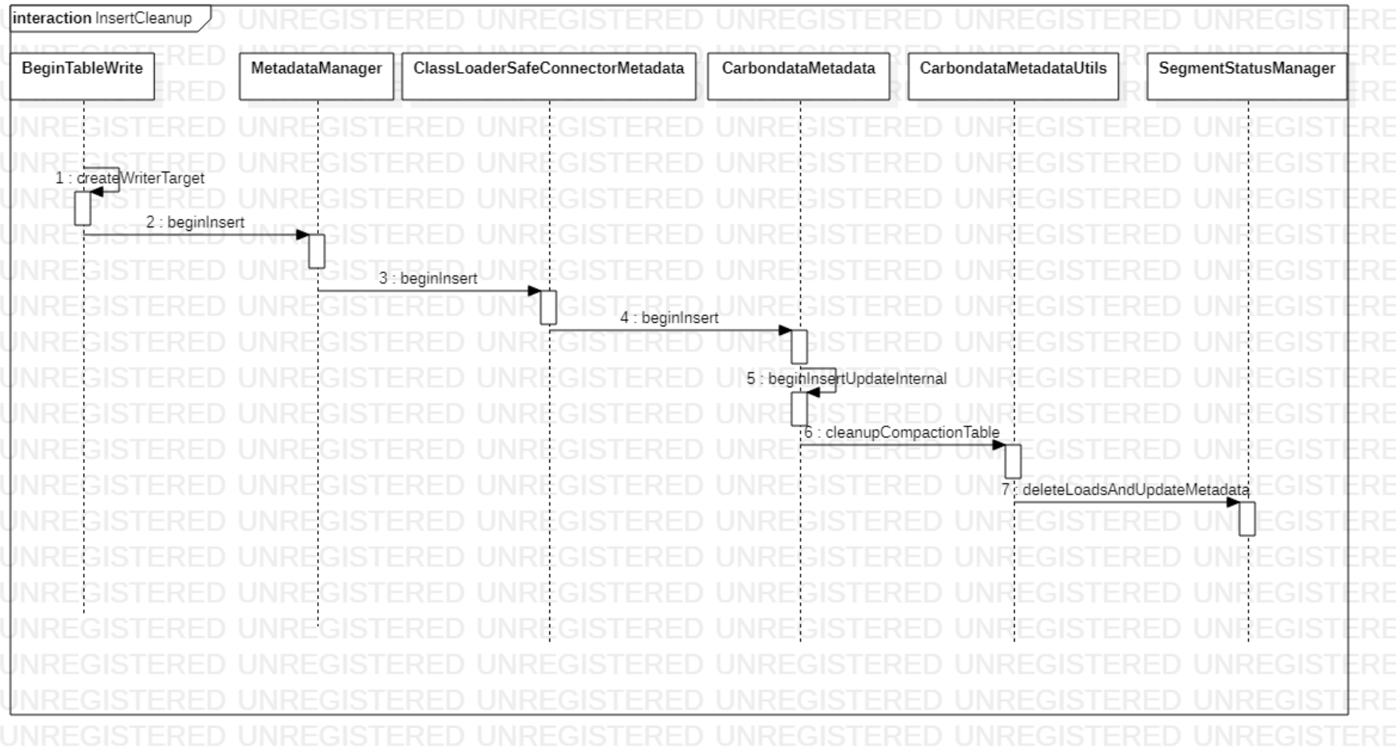


```
/st2/Fact/Part0/Segment_0.1
/st2/Fact/Part0/Segment_0.1/0.1_1592477541574.carbonindexmerge
/st2/Fact/Part0/Segment_0.1/part-0-0_batchno0-0-0.1-1592477533846.carbondata
/st2/Fact/Part0/Segment_0.1/part-0-1_batchno0-0-0.1-1592477533846.carbondata
/st2/Fact/Part0/Segment_4
/st2/Fact/Part0/Segment_4/4_1592470314996.carbonindexmerge
/st2/Fact/Part0/Segment_4/part-0-0_batchno0-0-4-1592470308901.carbondata
/st2/LockFiles
/st2/LockFiles/Segment_0.lock
/st2/LockFiles/Segment_1.lock
/st2/LockFiles/Segment_2.lock
/st2/LockFiles/Segment_3.lock
/st2/LockFiles/Segment_4.lock
/st2/LockFiles/clean_files.lock
/st2/LockFiles/compaction.lock
/st2/LockFiles/meta.lock
/st2/LockFiles/tablestatus.lock
/st2/LockFiles/update.lock
/st2/Metadata
/st2/Metadata/schema
/st2/Metadata/segments
/st2/Metadata/segments/0.1_1592477533846.segment
/st2/Metadata/segments/4_1592470308901.segment
/st2/Metadata/tablestatus
/st2/Metadata/tablestatus.history
```

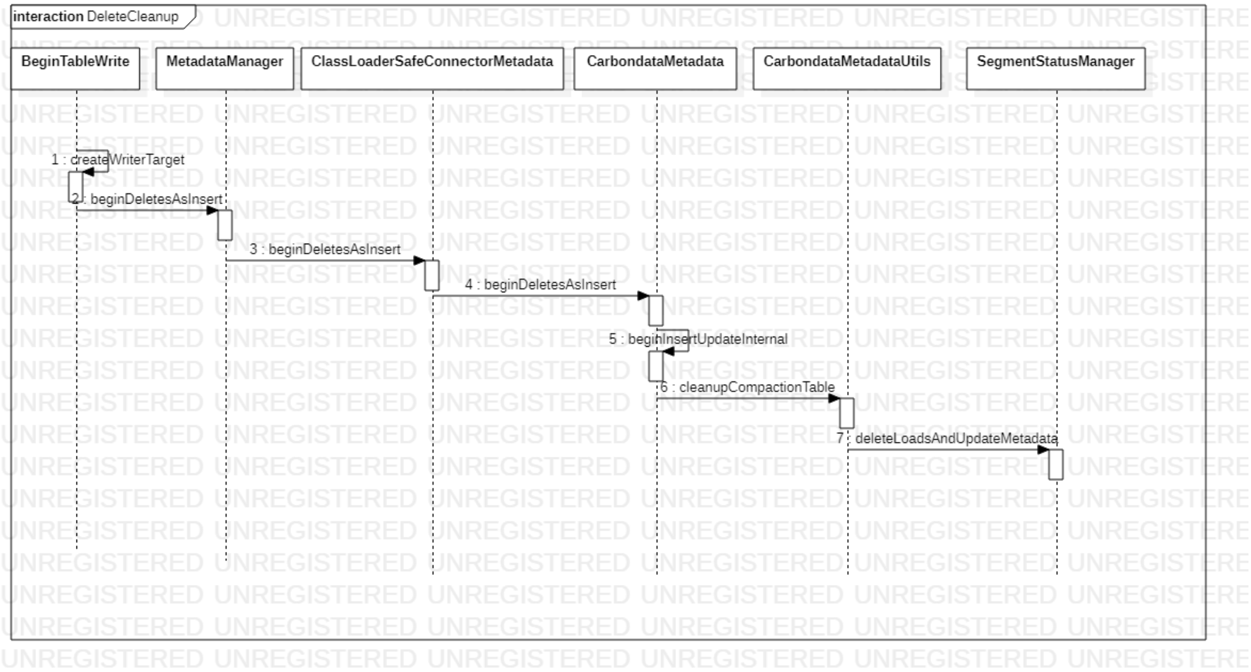
High level design:

deleteLoadsAndUpdateMetadata using tablestatus files , finds the which are COMPACTED, MARKED_FOR_DELETE, INSERT_IN_PROGRESS, INSERT_OVERWRITE_IN_PROGRESS files and which crossed more than 60 minutes and delete them.

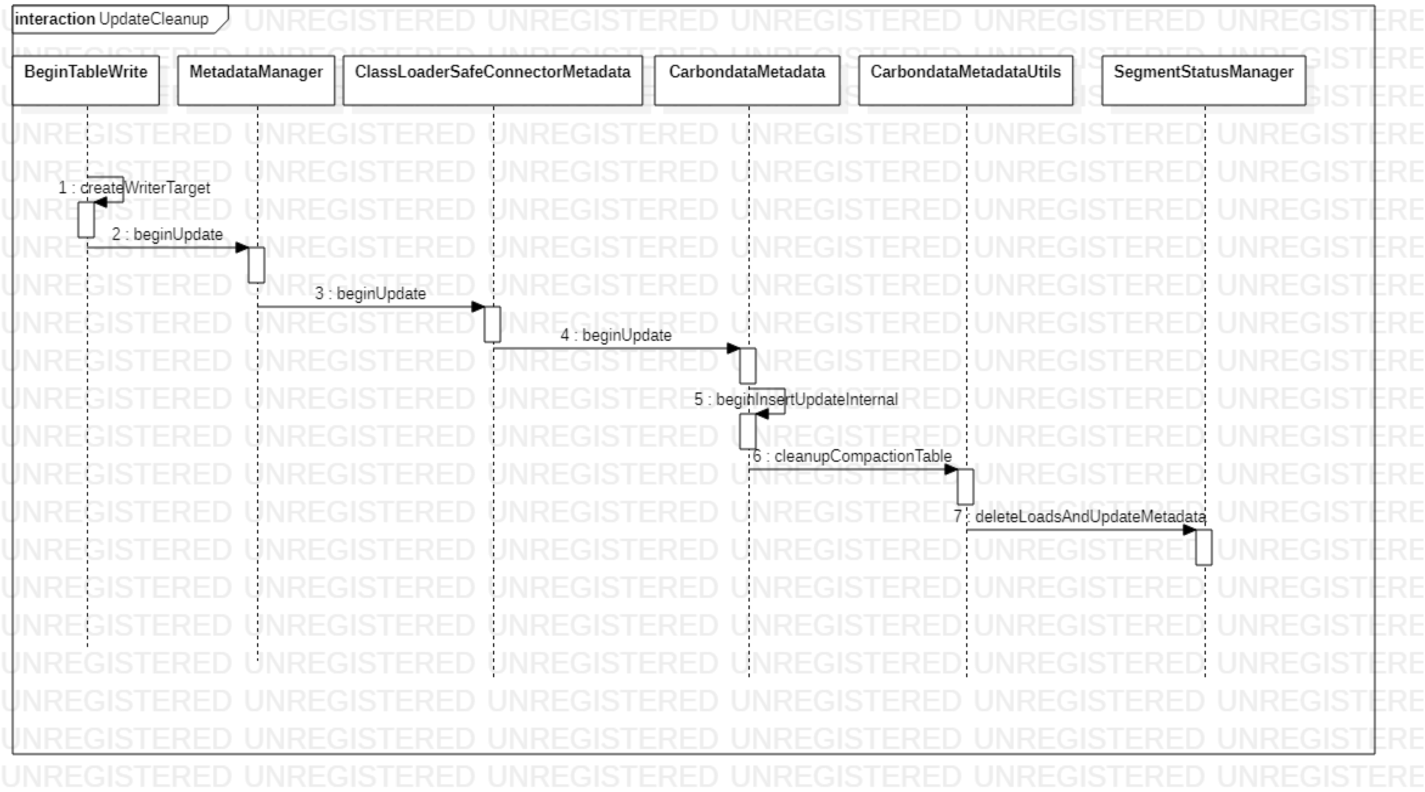
Auto cleanup getting triggered in case of insert



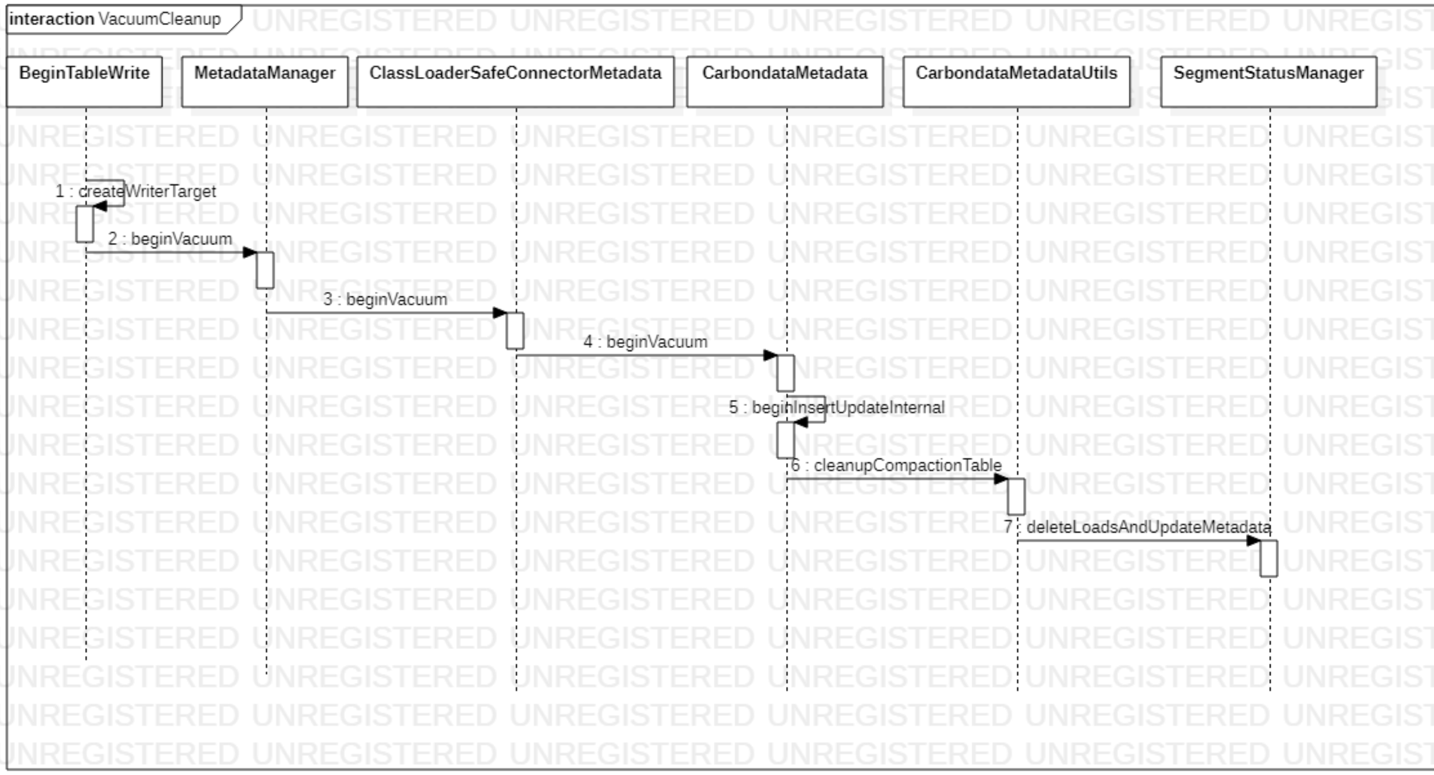
Auto cleanup getting triggered in case of delete



Auto cleanup getting triggered in case of update



Auto cleanup getting triggered in case of **vacuum**


interaction VacuumCleanup
BeginTableWrite
MetadataManager
ClassLoaderSafeConnectorMetadata
CarbondataMetadata
CarbondataMetadataUtils
SegmentStatusManager
1 : createWriterTarget
2 : beginVacuum
3 : beginVacuum
4 : beginVacuum
5 : beginInsertUpdateInternal
6 : cleanupCompactionTable
7 : deleteLoadsAndUpdateMetadata