How to clean up storage in Kylin 4

Background

Kylin will generate temporary files in HDFS during the cube building; Besides, when purge/drop/merge cubes, some parquet files may be left in HDFS and will no longer be queried; Although Kylin has started to do some automated garbage collection, it might not cover all cases; You can do an offline storage cleanup periodically.

Directory tree structure under Kylin 4.0 's working dir

Working Dir(ROOT)

- ***{PROJECT_NAME}*** ***[managed by tool]***
  - **parquet**
    - {CUBE_NAME} **[managed by tool]**
      - {SEGMENT_NAME} **[managed by tool]**
        - {CUBOID_ID}
          - parquet files
  - **spark_log**
    - driver
      - {JOB_ID}
        - drivers' log of cubing job
    - executor
      - {JOB_ID}
        - executors' log of cubing job
  - **dict/global_dict [managed by tool]**
    - {CUBE_NAME}
      - {COLUMN_NAME}
        - dict files
  - **table_snapshot [managed by tool]**
    - {SCHEMA_NAME.TABLE_NAME}
      - {JOB_ID}
        - parquet files
  - **job_tmp [managed by tool]**
    - {JOB_ID}
      - TBD
- **cube_statistics**
  - {CUBE_NAME}
    - {JOB_ID}
      - seq file of cuboid 's HLL
- **_sparder_log**
  - {DATE}
    - executors 's log of query job
- **resources-jdbc**
  - TBD

Summary

In above directory tree, the directory which end with "**managed by tool**" means **StorageCleanupJob** will try to check and delete useless files under these directory.

For directory **table_snapshot, dict/global_dict, parquet/{CUBE_NAME}, parquet/{CUBE_NAME}/{SEGMENT_NAME}** , Kylin will mark files which is unreferenced and stale(by checking last modified time) as garbage.

For directory **job_tmp,** Kylin will only check last modified time.

How to use

Option Table

Option	Data Type	Default Value	Comment

delete	Boolean	false	Boolean, whether or not to do real delete operation. Default value is false, means a dry run.
cleanupTableSnapshot	Boolean	true	Boolean, whether or not to delete unreferenced snapshot files. Default value is true .
cleanupGlobalDict	Boolean	true	Boolean, whether or not to delete unreferenced global dict files. Default value is true .
cleanupJobTmp	Boolean	false	Boolean, whether or not to delete job tmp files. Default value is false .
cleanupThreshold	Integer	168	Integer, used to specific delete unreferenced storage that have not been modified before how many hours (recent files are protected). Default value is 168 hours.

List help information

options

```
[root@cdh-master apache-kylin-4.0.0-SNAPSHOT-bin]# bin/kylin.sh org.apache.kylin.tool.StorageCleanupJob -help
Retrieving hive dependency...
Retrieving hadoop conf dir...
Retrieving Spark dependency...
...
Running org.apache.kylin.rest.job.StorageCleanupJob -help
usage: org.apache.kylin.rest.job.StorageCleanupJob
 -cleanupGlobalDict <cleanupGlobalDict>         Boolean, whether or not to
                                                delete unreferenced global
                                                dict files. Default value
                                                is true .
 -cleanupJobTmp <cleanupJobTmp>                 Boolean, whether or not to
                                                delete job tmp files.
                                                Default value is false .
 -cleanupTableSnapshot <cleanupTableSnapshot>   Boolean, whether or not to
                                                delete unreferenced
                                                snapshot files. Default
                                                value is true .
 -cleanupThreshold <cleanupThreshold>           Integer, used to specific
                                                delete unreferenced
                                                storage that have not been
                                                modified before how many
                                                hours (recent files are
                                                protected). Default value
                                                is 168 hours.
 -delete <delete>                               Boolean, whether or not to
                                                do real delete operation.
                                                Default value is false,
                                                means a dry run.
```

List directory which to be deleted

```
bin/kylin.sh org.apache.kylin.tool.StorageCleanupJob
```

Deleted them after confirm

```
bin/kylin.sh org.apache.kylin.tool.StorageCleanupJob --delete true
```

Only delete stale job_tmp and unreferenced cuboid files

```
bin/kylin.sh org.apache.kylin.tool.StorageCleanupJob --delete true \
 --cleanupJobTmp ture -cleanupTableSnapshot false \
 -cleanupGlobalDict false --cleanupThreshold 24
```