How to clean up storage in Kylin 4

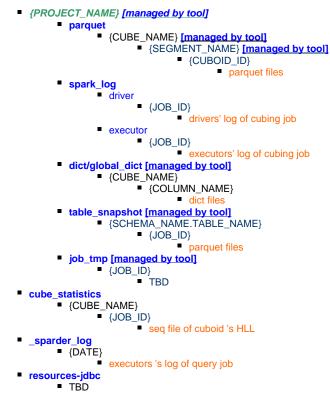
- Background
 - Directory tree structure under Kylin 4.0 's working dir
 - Summary
- How to use
 - Option Table
 - List help information
 - List directory which to be deleted
 - Deleted them after confirm
 - ° Only delete stale job_tmp and unreferenced cuboid files

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)



Summary

In above directory tree, the directory which end with "managed by tool" means <u>StorageCleanupJob</u> 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

a <i>i</i> :	.	B (1, 1, 1, 1)	
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	
Retrieving hive dependency	in]# bin/kylin.sh org.apache.kylin.tool.StorageCleanupJob -help
Retrieving hadoop conf dir	
Retrieving Spark dependency	
 Running org.apache.kylin.rest.job.StorageClean	un Tab balm
usage: org.apache.kylin.rest.job.StorageCleanu	
-cleanupGlobalDict <cleanupglobaldict></cleanupglobaldict>	Boolean, whether or not to
	delete unreferenced global
	dict files. Default value
	is true .
-cleanupJobTmp <cleanupjobtmp></cleanupjobtmp>	Boolean, whether or not to
	delete job tmp files.
	Default value is false .
-cleanupTableSnapshot <cleanuptablesnapshot></cleanuptablesnapshot>	Boolean, whether or not to
	delete unreferenced
	snapshot files. Default
	value is true .
-cleanupThreshold <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></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
```