The following DDL statements in hive result in creation of a directories on the DFS filesystem.

create database

create table

alter table <table> add partition

This is done to ensure that the user (be it be the service user or end user) has the privileges to access the underlying filesystem location. But it also sets up the subsequent queries from not having the burden of creating the directories “on demand”. For example, having the execution for “create table” statement also create the database root directories would be a little out of place.

If the Metastore does not create the directories for DDL statements, then replication manager snapshots are more consistent without any shell directories and easier to compare.

While the fix might seem small, it has implications for query processors like HS2 and Impala that have expectations of such directories existing. This may change the query planning. For example, insert into a table partition should not have to worry about the database directory being non-existent.

Expand the HMS API to accept additional parameters to skip the creation on a need-to basis. So essentially add APIs that we only anticipate the replication code to use.