Step 1. Data Preprocessing

- Data is assumed to be in Mysql initially.
- Copy the given csv files from local to hdfs
- Creation of Mysql Tables (Hint: Sno and Seq can be used as Primary Keys. Date can be varchar type)
- Sqoop export of data from HDFS to Mysql. (Hint : Use Staging table while exporting, use password encryption)
- Delete the data from hdfs post export

Step 2- Sqoop Import

- Bring the data from Mysql tables to HDFS
  Hint: Create sqoop jobs for importing both tables Use other necessary optimizations Use Password Encryption

Step 3- Create Hive External Tables on top of data in HDFS

Step 4- Create Optimized External tables in Hive

- Hint: File format -use TextFile for time being (though ORC is a better choice and can be used in future)
- Use Dynamic Partitioning on State Column, Bucketing on Date Column Note: There might be frequent queries on State and Date, so we choose these columns for Partitioning and Bucketing, both tables can be Partitioned on State and Bucketed on Date column.

Step 5: Load data to the optimized hive tables from normal hive tables.

- Hint: Use INSERT OVERWRITE clause
- Date has also to be formatted to proper hive format which is yyyy-mm-dd .Use from_unixtime , unix_timestamp

Step 6-Inner Join two tables in Hive and get a consolidated table.

- Hint: Perform Map-side join of the two tables. Join columns can be ‘date’ and ‘state’ for better optimization. Here it is assumed that the State_Testing table is small enough to fit in memory.

Step 7: Analysis

For Example: Ideally, the number of samples that tested positive and the number of covid cases confirmed must be the same. See which state/states have more consistent data collection like The number of positive samples (table1) match mostly with the number of confirmed cases(table2), for which state. For every state, find the total number of confirmed cases reported and also the total number of positive samples tested, in the entire duration of 2 months, starting with the state with the highest cases.