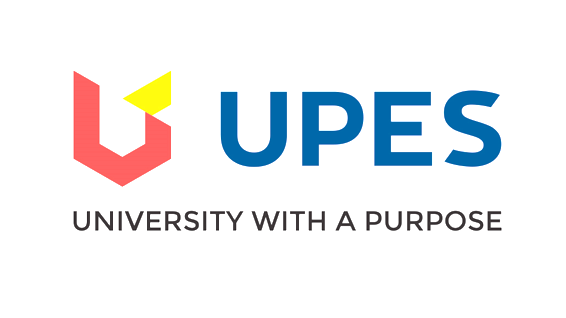
****

**Big Data Analysis**

**Project based Assignment**

Corona Model

By: Rishi Pandey

CSE-Big Data-B2

Roll no: 97

Project based Assignment: -

*Choose a domain where you can apply big data analysis. In the selected domain:*

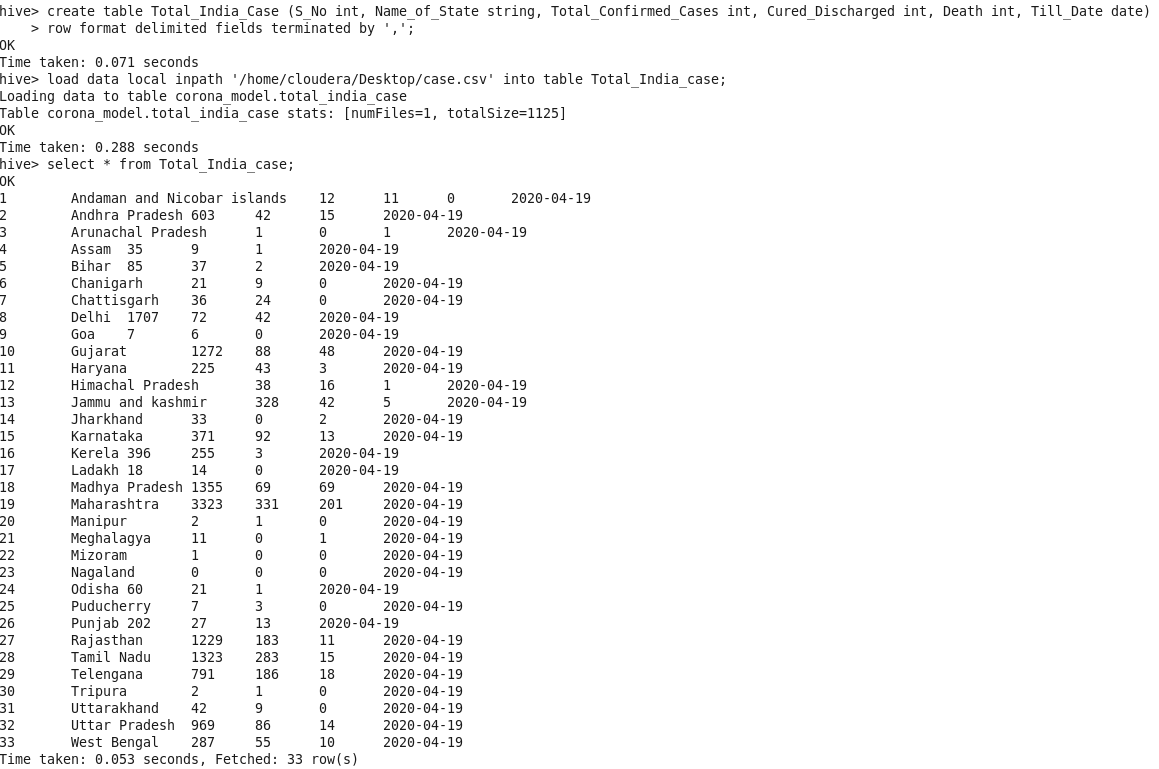
1. *Write a schema of some important tables.*
2. *Identify 10 problems that an individual might face.*
3. *Find a solution to these challenges using big data analysis. Explain your solution with the help of HiveQL****, if possible*.**

My chosen domain is **Government and Health Sector.**

There are 5 tables of different data of specifically current COVID-19 pandemic, they are named as: -

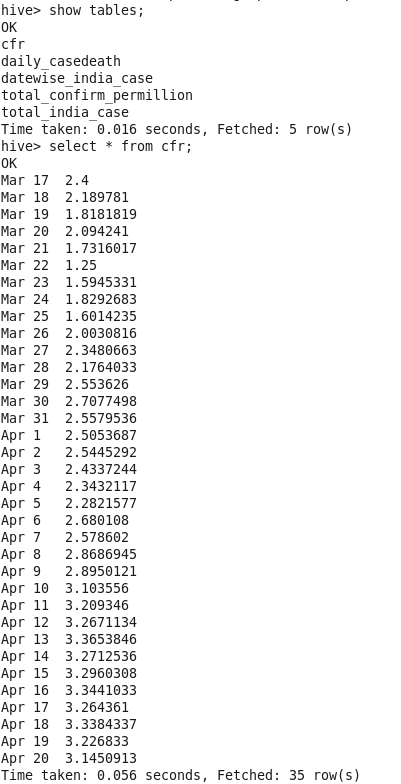
1. Total\_India\_case
2. Datewise\_India\_case
3. Daily\_casedeath
4. CFR
5. Total\_confirm\_permillion

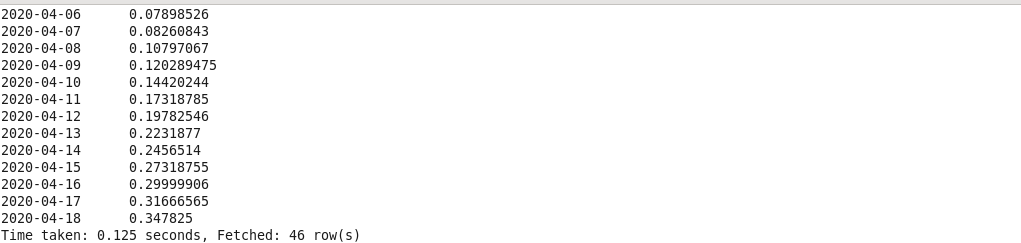
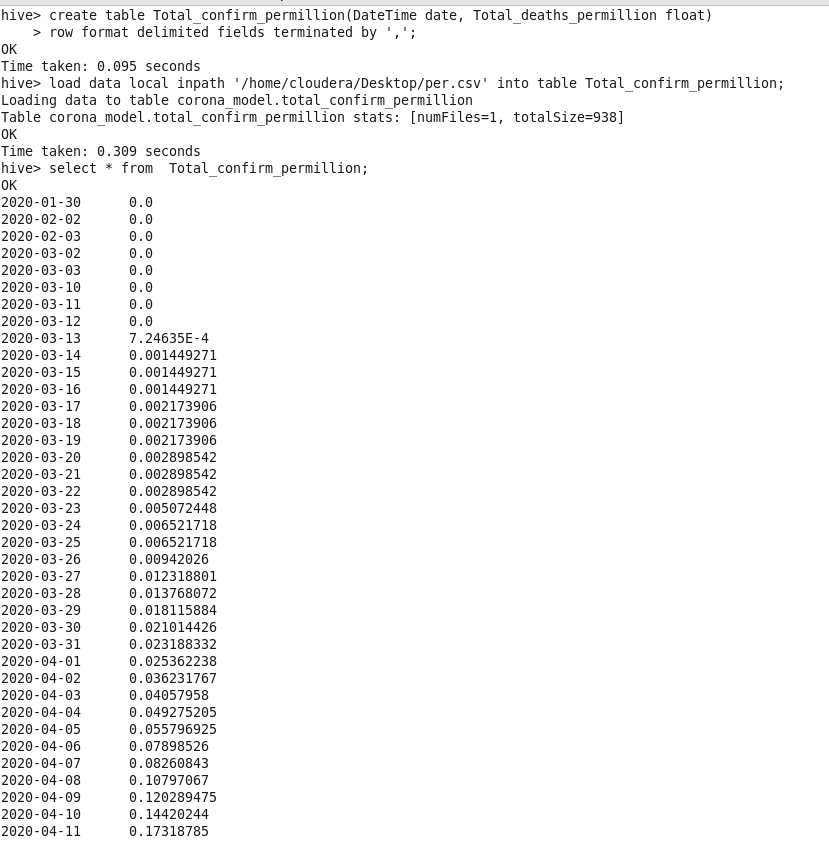
The snapshot of tables:

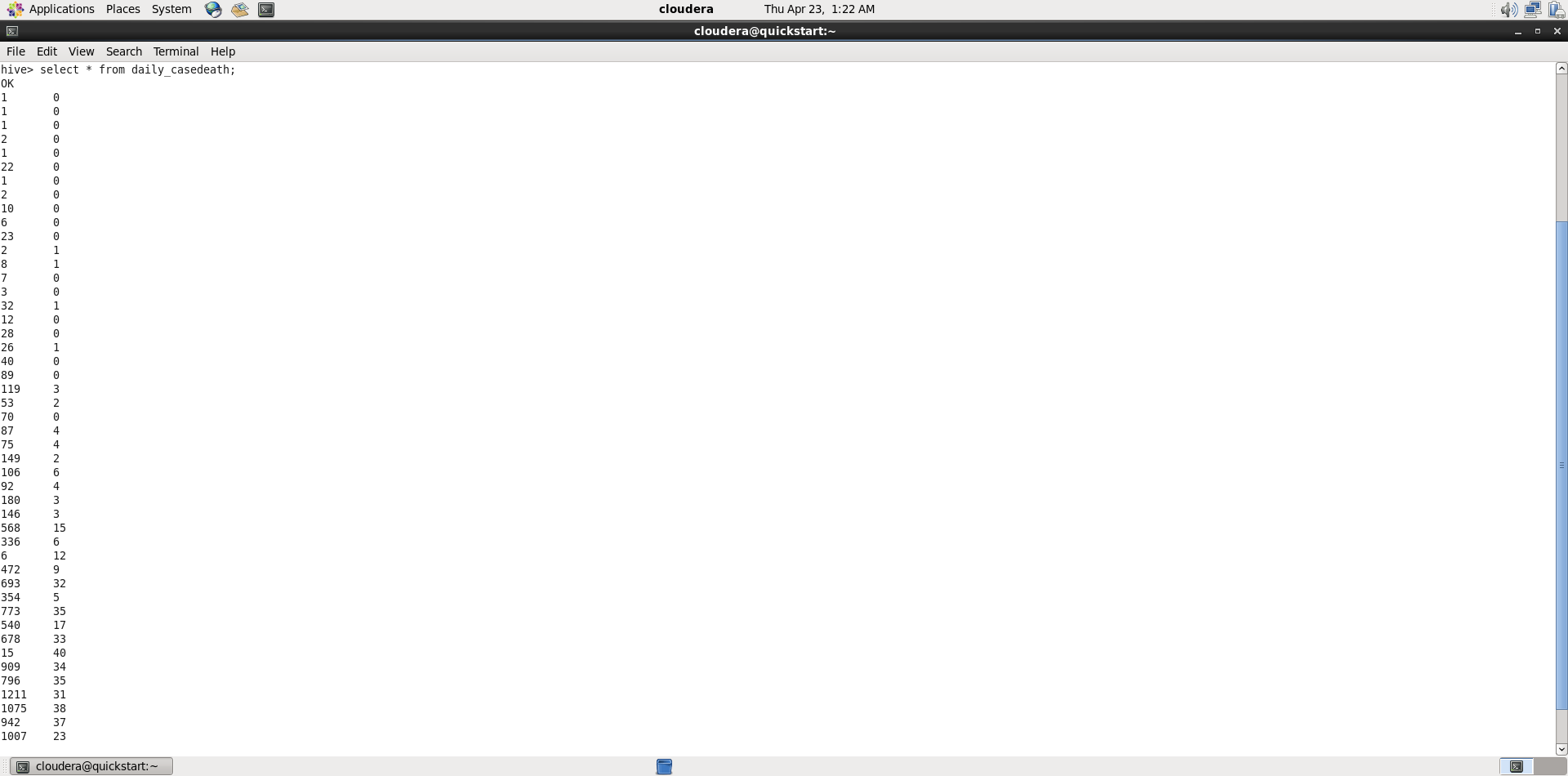




……….







The 10 problems which I executed/tried on hive are:

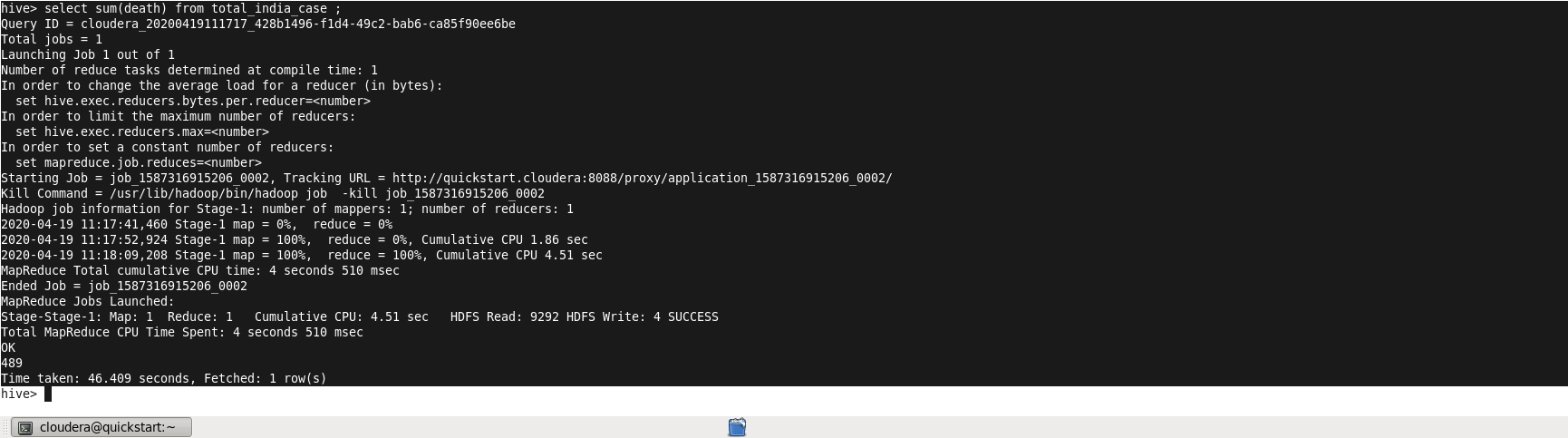
1. What is the total number of confirmed deaths due to current pandemic in India?
2. What is the daily number of confirmed cases in our country?
3. In India, which state is adversely affected due to Covid-19 mass-spread?
4. Which state is in good condition and is predicted to recover soon from this mass-spread?
5. What is the case fatality rate (CFR) of each state in India?
6. What is the current average CFR of India? What does it tell us?
7. Maharashtra has been declared severely affected. How much has the Maharashtra and all other affected states contributed to the Total Indian cases of Covid-19?
8. Which state is considered to be untouched/not-affected with the Covid-19 virus?
9. What does the current data on occurring of confirmed cases and deaths due to it tell us about the mortality risk of Covid-19?
10. What is the estimated IFR (Infection Fatality Rate) for the whole India?

There were few more problems which I tried but could not take them to the end result…

The solutions to these problems in terms of HQL are:

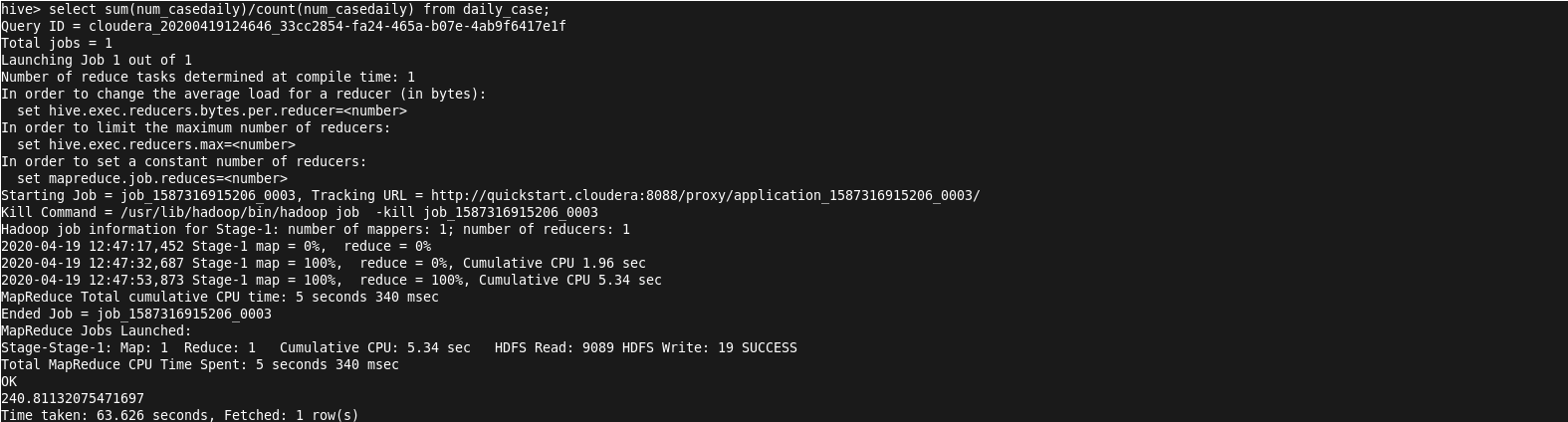
1. What is the total number of confirmed deaths due to current pandemic in India?

This is calculated using function sum () on table total\_india\_case in hive whose result came out to be **489** till 2020-04-18.



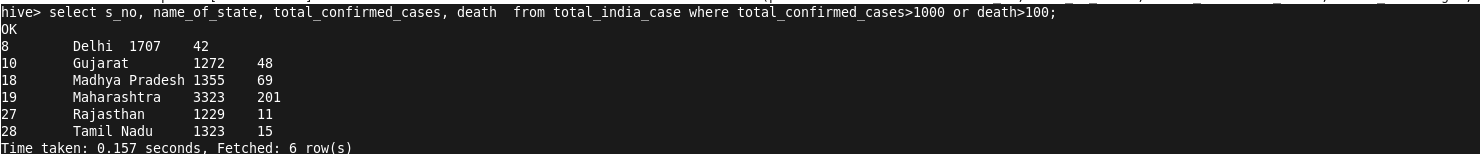
1. What is the daily number of confirmed cases in our country?

This was calculated using different function and operation on queries i.e. sum (), count () whose result came out to be **240.8113** till 2020-04-18.



1. In India, which state is adversely affected due to Covid-19 mass-spread?

This problem was calculated using simple SQL concepts of where clause and multiple conditions on required table. The result is clear these **6 states** are conditioned in worse state of pandemic.



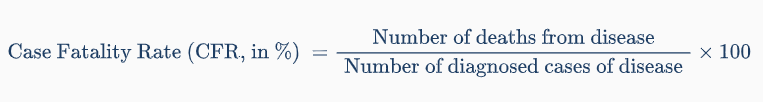
1. Which state is in good condition and is predicted to recover soon from this mass-spread?

This problem is to figure out the recovering state or the state in good condition, here multiple conditions were used on the table to get result, here the % of recoverins where lesser the % better is the result.



1. What is the case fatality rate (CFR) of each state in India?

CFR is an important factor in determining the extend of mass-spread which awakens us and is easy to calculate

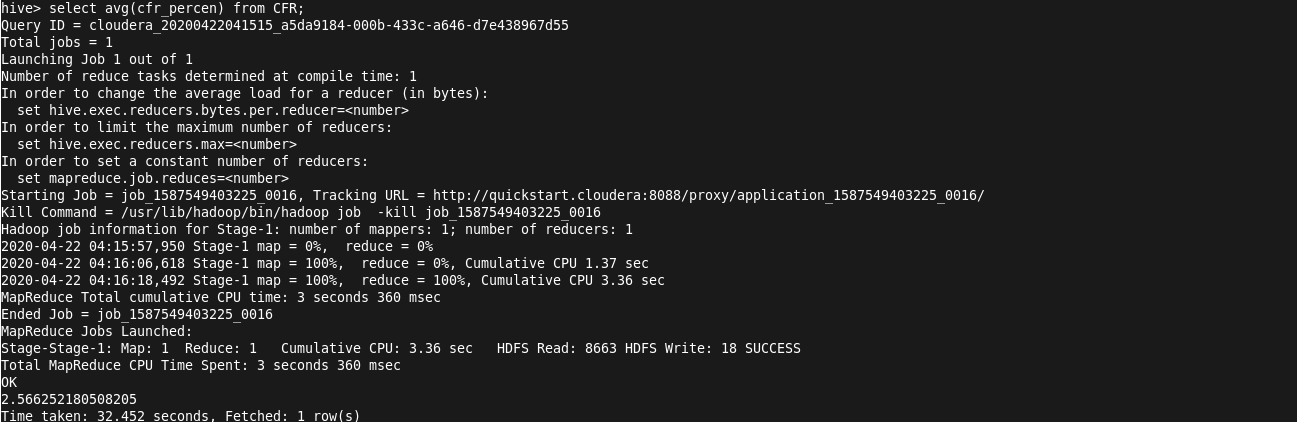


CFR is not constant it varies with many parameters but as per acquired and available data, this is the current CFR of each state. When there are people who have the disease but are not diagnosed, the CFR will overestimate the true risk of death. There might be many undiagnosed people. So testing is also major factor on which CFR depends.



1. What is the current average CFR of India? What does it tell us?

This is the problem where we used SQL function avg () to calculate the average CFR in whole India, which turned out to be **2.566**



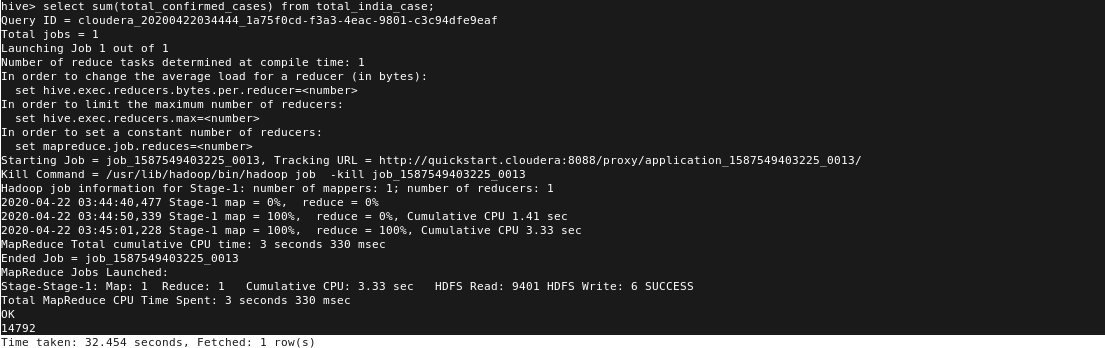
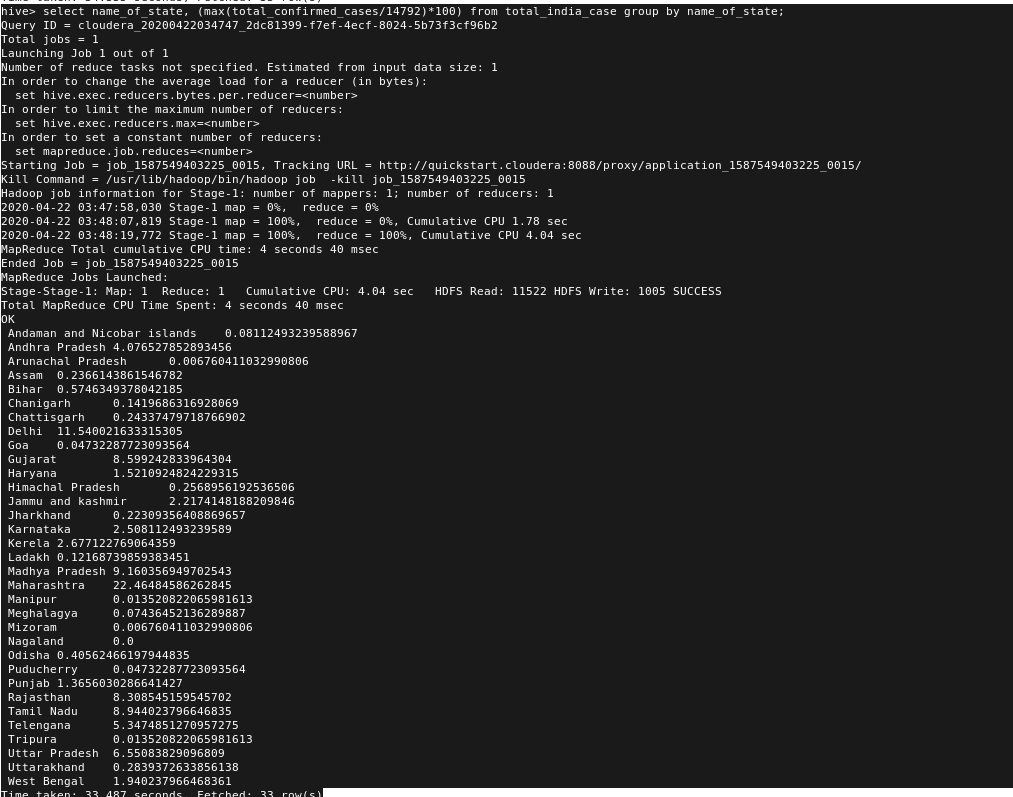
1. Maharashtra has been declared severely affected and so many others. How much has the Maharashtra and all other affected states contributed to the Total Indian cases of Covid-19?

This problem took two queries to complete as together they got very complex. Here Contribution in total confirmed cases was measured firstly taking the total confirmed cases in India and then Using it in the formula

In the query using max (), group by etc…

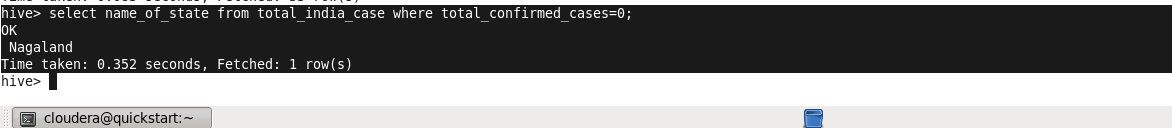
*Contribution per state = (Total confirmed in state/*

*Total in India) \*100*

1. Which state is considered to be untouched/not-affected with the Covid-19 virus?

This problem is simple SQL query to find a state where the effect of pandemic and mass-spread is almost null. The query result came out to be **Nagaland**.



1. What does the occurring of confirmed cases and deaths due to it tell us about the mortality risk of Covid-19?

This is the only problem which cannot be solved by hive,

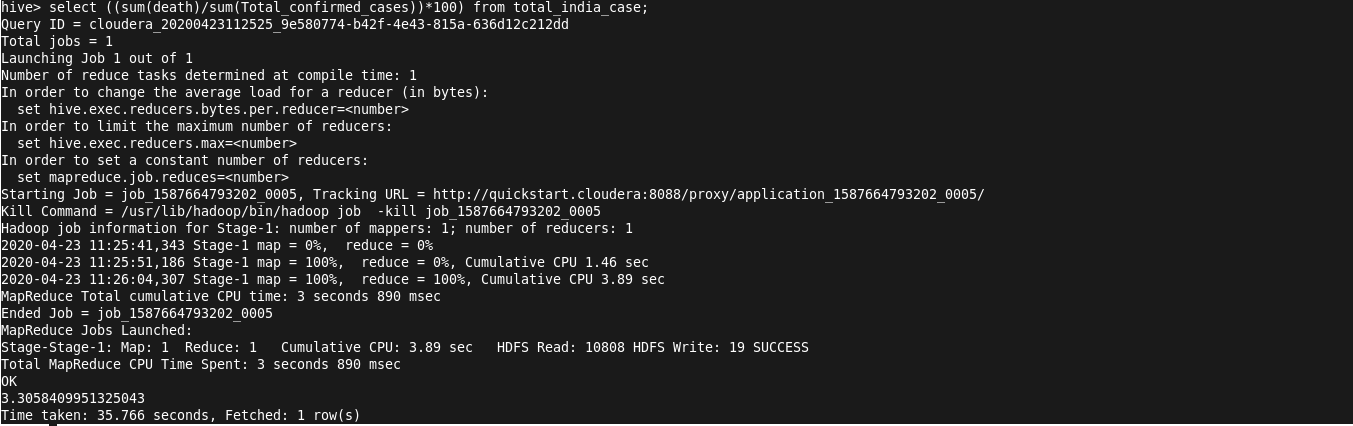
Maybe very few points with low information query can give minuet result. This kind of problem requires large datasets of continuous data input like spark, and can also done using RStudio, visualizations, GIS tools for Hadoop, Orange etc…

Mortality risk is the measure having there very common yet different aspects. CFR (case fatality rate), CMR (crude mortality rate) and IFR (infection fatality rate). We may be able to estimate the total number of cases and use it to calculate the IFR and researchers do this. But the total number of cases is not known, so the IFR cannot be accurately calculated. The CFR of COVID-19 differs by location, and has changed during the early period of the outbreak. At the end, these aspects are themselves calculated from the various parameters and various data sources which was not possible in hive.

1. What is the estimated IFR (Infection Fatality Rate) for the whole India?

This problem is solved using various function and formula where the data was limited and not updated regularly so IFR for India till date 2020-04-18 is **3.305**

The IFR = number of deaths from a disease divided by the total number of cases.



Thank You