Covid-19 Data Analysis
(Increment 2)

Team Members (Team#5):

1. Neha Navgale
2. Andrew Poitras
3. Krishnapriya Akula
4. Christian Barlow
5. Jayadeep Kumar Reddy Singareddi

Introduction:

Covid-19 has impacted each and every person across the globe and every one is taking necessary measures to reduce its impact. Our project is to study and analyze its impact on different sectors and geographic locations and provide users with interactive visualizations.

Background:

There are some projects out there which shows analysis and visualization related to covid data. The scope of analysis is limited to identifying the number of active, recovered cases and deaths per country or region.

https://www.tableau.com/covid-19-coronavirus-data-resources
https://www.worldometers.info/coronavirus/#countries
https://coronavirus.jhu.edu/map.html

Goals and Objectives:

Motivation:

As the coronavirus is spreading rapidly, everyone looks for a place where they can find meaningful insights on its impact on different geographical locations, health care facilities and on the economy. In order to provide information, we are motivated to generate new insights in support of the ongoing fight against this infectious disease using big data techniques.

Significance/Uniqueness:

Every single person on the planet has been affected by COVID-19 and have been doing everything they can to avoid it. This project could help people avoid hot zones by analyzing the correlation between geographical location and the number of confirmed cases of the virus. The research gathered will also allow us to identify locations that may not be following proper social distancing and mask regulations as they are going to have a higher population of people with the virus. Insights gathered will also allow us to see what age groups the disease has been most commonly spreading. We will also be able to identify the impact on different business sectors and by looking into trends we can identify its impact in future. This project will serve as a single place to understand its overall impact.

Objectives:

Our objective is to understand the datasets and utilize latest big data techniques to extract meaningful insights which can help to understand the trend of coronavirus and the correlation of number of cases with locations, age group, business and economy.

Features:

Our project will include following features:

1. Collect real time data of covid-19 for each state of the USA
2. Identify answers to below questions based on data analysis:
    a. Identify the total number of cases and total death in each state
    b. Identify the highest single day rise in a state.
    c. Identify the death rate by each state.
    d. Identify the most affected county/region.
    e. Identify its impact on different age groups.
    f. How healthcare is responding and providing facilities
    g. How it has impacted economy and to what extent
    h. What businesses are badly impacted.
    i. How the job market looks like after the pandemic started.
3. Develop interactive front end application for users to understand the analysis.
4. Create graphical visualization of all the analysis completed in feature 2.

Features Developed:

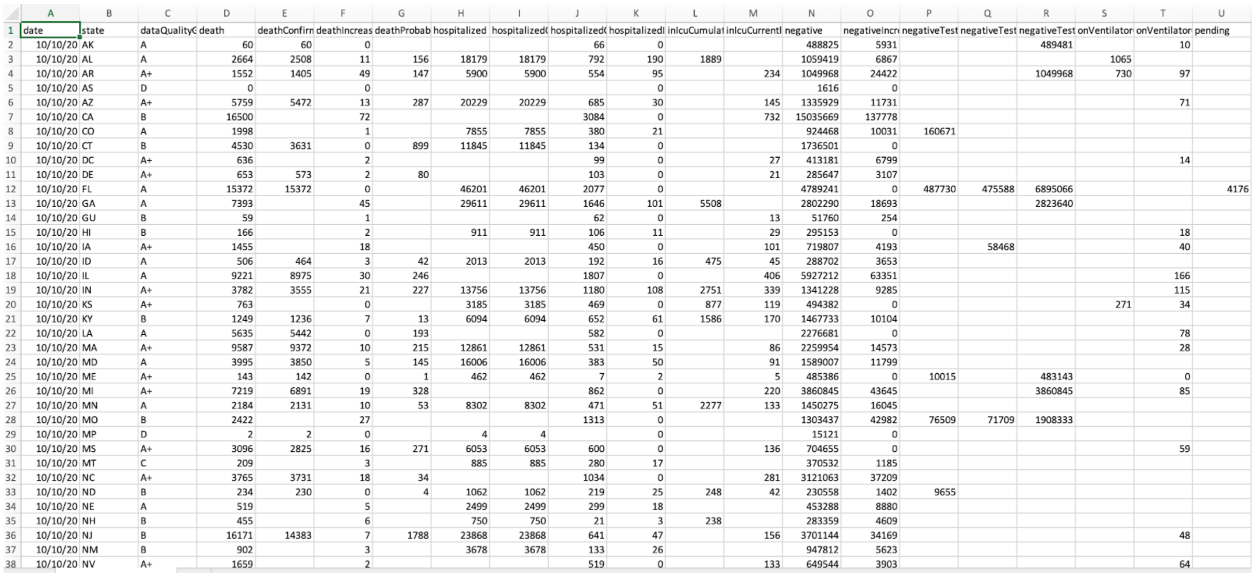
Dataset

USA Data:

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

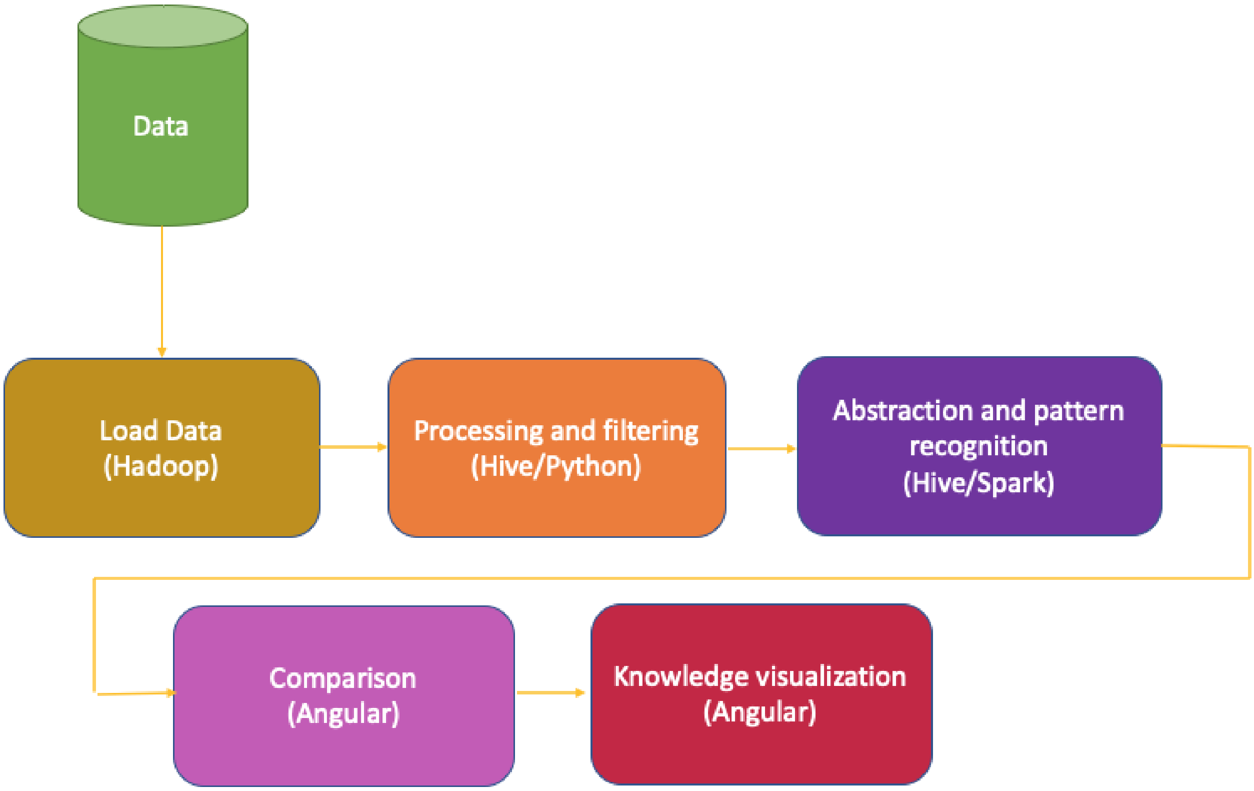
Hospitalization Data of USA - https://covidtracking.com/data

World Data - https://ourworldindata.org/coronavirus-source-data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	date	state	dataQualityG	death	deathConfirn	deathIncreas	deathProbab	hospitalized	hospitalized(	hospitalized(	hospitalizedI	inIcuCumulat	inIcuCurrentl	negative	negativeIncr	negativeTest	negativeTest	negativeTest	onVentilator	onVentilator	pending
2	10/10/20	AK	A	60	60	0				66	0			488825	5931			489481		10	
3	10/10/20	AL	A	2664	2508	11	156	18179	18179	792	190	1889		1059419	6867				1065		
4	10/10/20	AR	A+	1552	1405	49	147	5900	5900	554	95		234	1049968	24422			1049968	730	97	
5	10/10/20	AS	D	0		0					0			1616	0						
6	10/10/20	AZ	A+	5759	5472	13	287	20229	20229	685	30		145	1335929	11731					71	
7	10/10/20	CA	B	16500		72				3084	0		732	15035669	137778						
8	10/10/20	CO	A	1998		1		7855	7855	380	21			924468	10031	160671					
9	10/10/20	CT	B	4530	3631	0	899	11845	11845	134	0			1736501	0						
10	10/10/20	DC	A+	636		2				99	0		27	413181	6799					14	
11	10/10/20	DE	A+	653	573	2	80			103	0		21	285647	3107						
12	10/10/20	FL	A	15372	15372	0		46201	46201	2077	0			4789241	0	487730	475588	6895066			4176
13	10/10/20	GA	A	7393		45		29611	29611	1646	101	5508		2802290	18693			2823640			
14	10/10/20	GU	B	59		1				62	0		13	51760	254						
15	10/10/20	HI	B	166		2		911	911	106	11		29	295153	0					18	
16	10/10/20	IA	A+	1455		18				450	0		101	719807	4193		58468			40	
17	10/10/20	ID	A	506	464	3	42	2013	2013	192	16	475	45	288702	3653						
18	10/10/20	IL	A	9221	8975	30	246			1807	0		406	5927212	63351					166	
19	10/10/20	IN	A+	3782	3555	21	227	13756	13756	1180	108	2751	339	1341228	9285					115	
20	10/10/20	KS	A+	763		0		3185	3185	469	0	877	119	494382	0				271	34	
21	10/10/20	KY	B	1249	1236	7	13	6094	6094	652	61	1586	170	1467733	10104						
22	10/10/20	LA	A	5635	5442	0	193			582	0			2276681	0					78	
23	10/10/20	MA	A+	9587	9372	10	215	12861	12861	531	15		86	2259954	14573					28	
24	10/10/20	MD	A	3995	3850	5	145	16006	16006	383	50		91	1589007	11799						
25	10/10/20	ME	A+	143	142	0	1	462	462	7	2		5	485386	0	10015		483143		0	
26	10/10/20	MI	A+	7219	6891	19	328			862	0		220	3860845	43645			3860845		85	
27	10/10/20	MN	A	2184	2131	10	53	8302	8302	471	51	2277	133	1450275	16045						
28	10/10/20	MO	B	2422		27				1313	0			1303437	42982	76509	71709	1908333			
29	10/10/20	MP	D	2	2	0		4	4		0			15121	0						
30	10/10/20	MS	A+	3096	2825	16	271	6053	6053	600	0		136	704655	0					59	
31	10/10/20	MT	C	209		3		885	885	280	17			370532	1185						
32	10/10/20	NC	A+	3765	3731	18	34			1034	0		281	3121063	37209						
33	10/10/20	ND	B	234	230	0	4	1062	1062	219	25	248	42	230558	1402	9655					
34	10/10/20	NE	A	519		5		2499	2499	299	18			453288	8880						
35	10/10/20	NH	B	455		6		750	750	21	3	238		283359	4609						
36	10/10/20	NJ	B	16171	14383	7	1788	23868	23868	641	47		156	3701144	34169					48	
37	10/10/20	NM	B	902		3		3678	3678	133	26			947812	5623						
38	10/10/20	NV	A+	1659		2				519	0		133	649544	3903					64	

Project Workflow:



Detail design of Features

Feature 1
Collect real time data of covid-19 for each state of the USA

- Analyzed different datasets and finalized 3 datasets out of all.
- For increment 2, we collected the data in csv format. For the next increment we will write python code to consume API to get real time data.

Feature 2
Identify answers to below questions based on data analysis:

1. **Identify the total number of cases in state of USA**
    - Loaded the data into hdfs
    - Wrote mapreduce program to count the cases
    - Stored output on hdfs
    - Displayed the output on Hue.

2. **Identify the total number of deaths in state of USA**
    - Loaded the data into hdfs
    - Wrote mapreduce program to count the deaths
    - Stored output on hdfs
    - Displayed the output on Hue.

3. **Highest single day rise.**
    - Created the hive table.
    - Loaded data in hive table
    - Wrote HQL query using aggregate and group by function to find the single day rise
    - Executed query on hive editor of Hue and displayed the output.

4. **October cases by state**
    - Created the hive table.
    - Loaded data in hive table
    - Wrote HQL query to sort the states with decreasing order of cases in october
    - Executed query on hive editor of Hue and displayed the output.

5. **Death rate as compared to total cases.**
    - Created the hive table.
    - Loaded data in hive table
    - Wrote HQL query to find the percentage of death as compared to total cases registered.

- Executed query on hive editor of Hue and displayed the output.

Analysis (Details about data):

We collected the data in csv format. The dataset is very huge and splitted by state. There are some fields which we have not considered in scope of this project. We may consider that as we proceed further. Below are the definition of fields that are in scope.

Data Definitions:

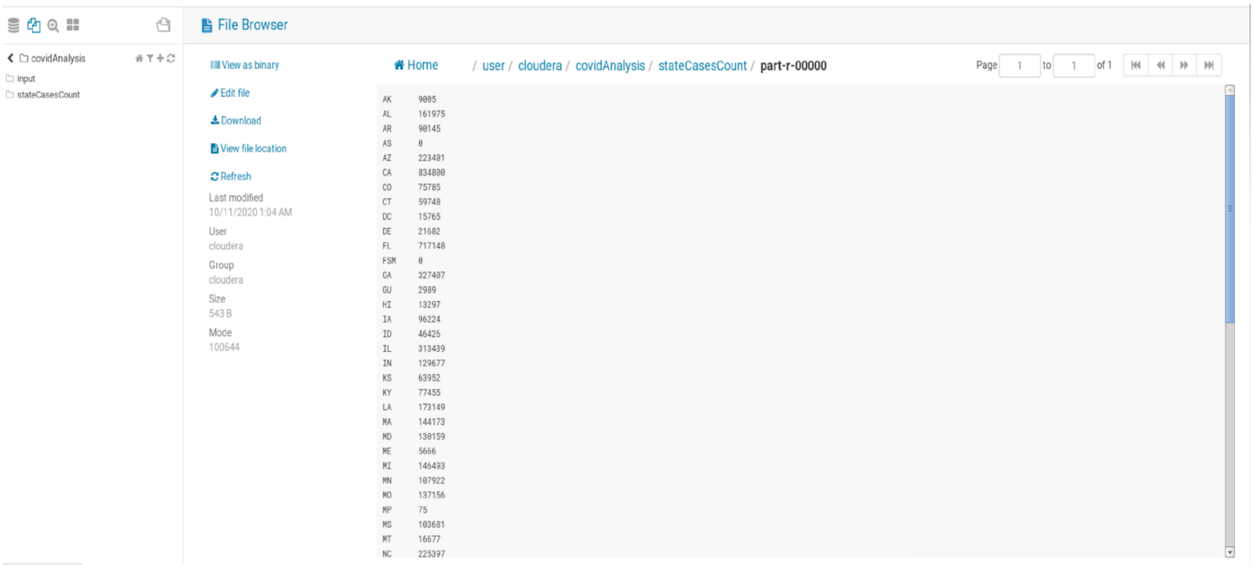
1. New Cases: This will indicate the daily increase in the number of cases in each state.
2. Recovered: Daily increase in total number of recovered patients
3. Death: Daily increase in total number of deaths.
4. Total Test Results: Field will have a total number of tests performed so far.
5. Currently Hospitalized: Count of individuals currently hospitalized because of covid.
6. Cumulative hospitalized: Count of individuals who have ever been hospitalized because of covid.
7. onVentilatorCumulative: Count of individuals who have ever been hospitalized under ventilation because of covid.
8. onVentilatorCurrently: Count of individuals who are currently hospitalized under ventilation because of covid.

Implementation:

Technologies/Platform used: Hadoop, Mapreduce, Hive, Hue.

Preliminary Results (Visualization of Results):

Total number of cases in each state:

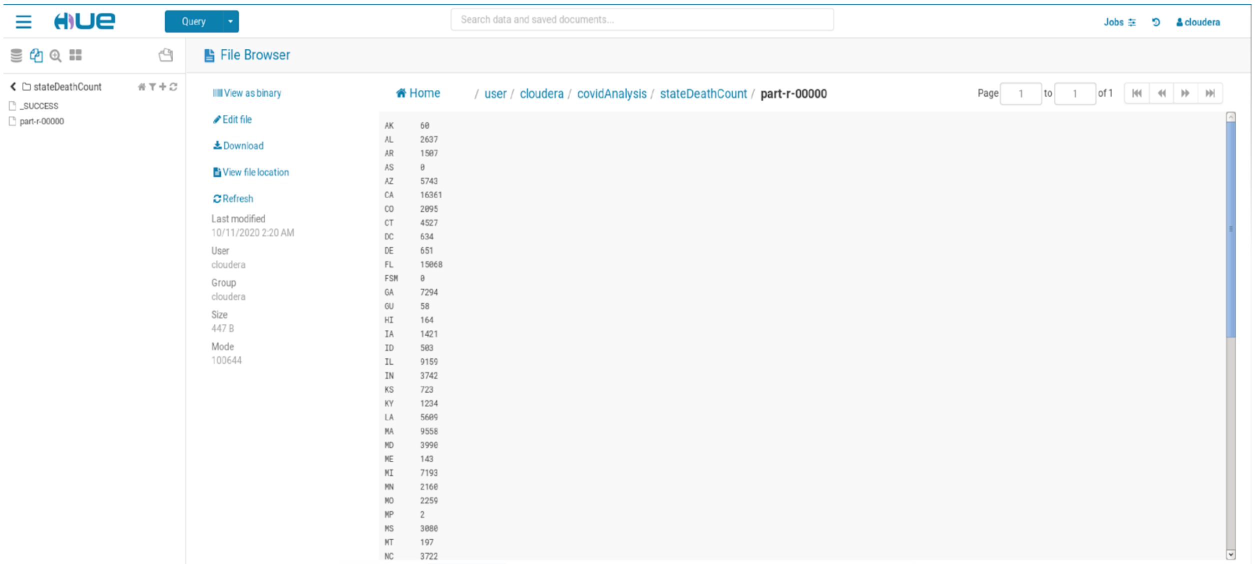


/ user / cloudera / covidAnalysis / stateCasesCount / part-r-00000

Last modified
10/11/2020 1:04 AM
User
cloudera
Group
cloudera
Size
543 B
Mode
100644

```
AK    9005
AL    161975
AR    90145
AS    0
AZ    223401
CA    834800
CO    75785
CT    59748
DC    15765
DE    21682
FL    717148
FSM   0
GA    327407
GU    2989
HI    13297
IA    96224
ID    46426
IL    313439
IN    129677
KS    63952
KY    77455
LA    173149
MA    144173
MD    130159
ME    5666
MI    146493
MN    107922
MO    137156
MP    75
MS    103681
MT    16677
NC    225397
```

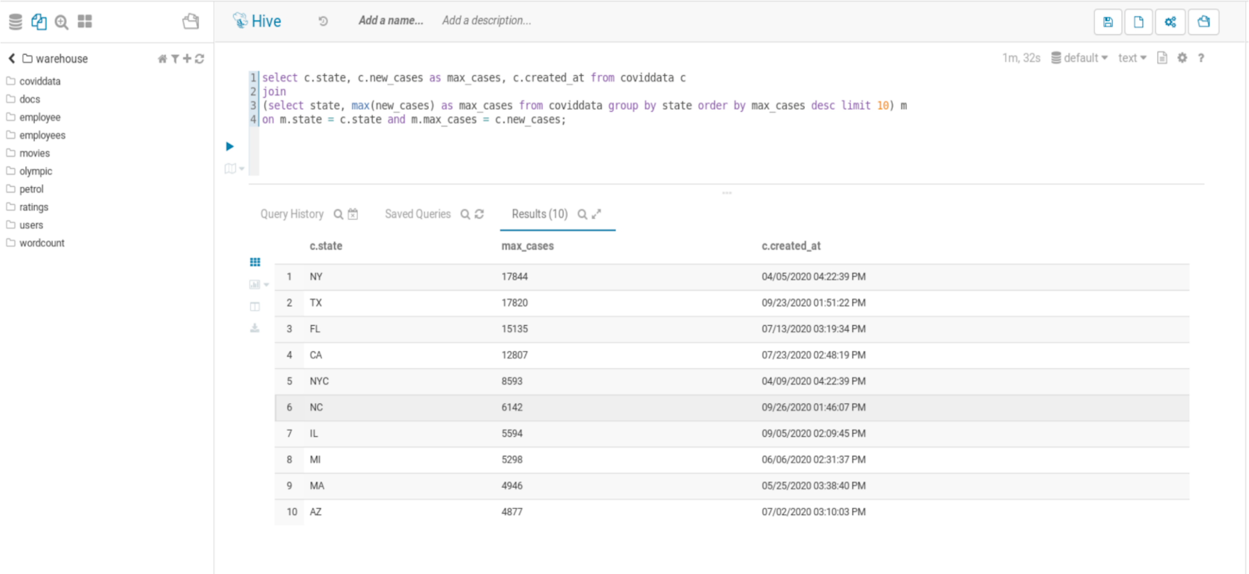
Total number of deaths in each state:



Last modified
10/11/2020 2:20 AM
User
cloudera
Group
cloudera
Size
447 B
Mode
100644

```
AK     60
AL     2637
AR     1507
AS     0
AZ     5743
CA     16361
CO     2095
CT     4527
DC     634
DE     651
FL     15068
FSM    0
GA     7294
GU     58
HI     164
IA     1421
ID     503
IL     9159
IN     3742
KS     723
KY     1234
LA     5609
MA     9558
MD     3990
ME     143
MI     7193
MN     2160
MO     2259
MP     2
MS     3080
MT     197
NC     3722
```

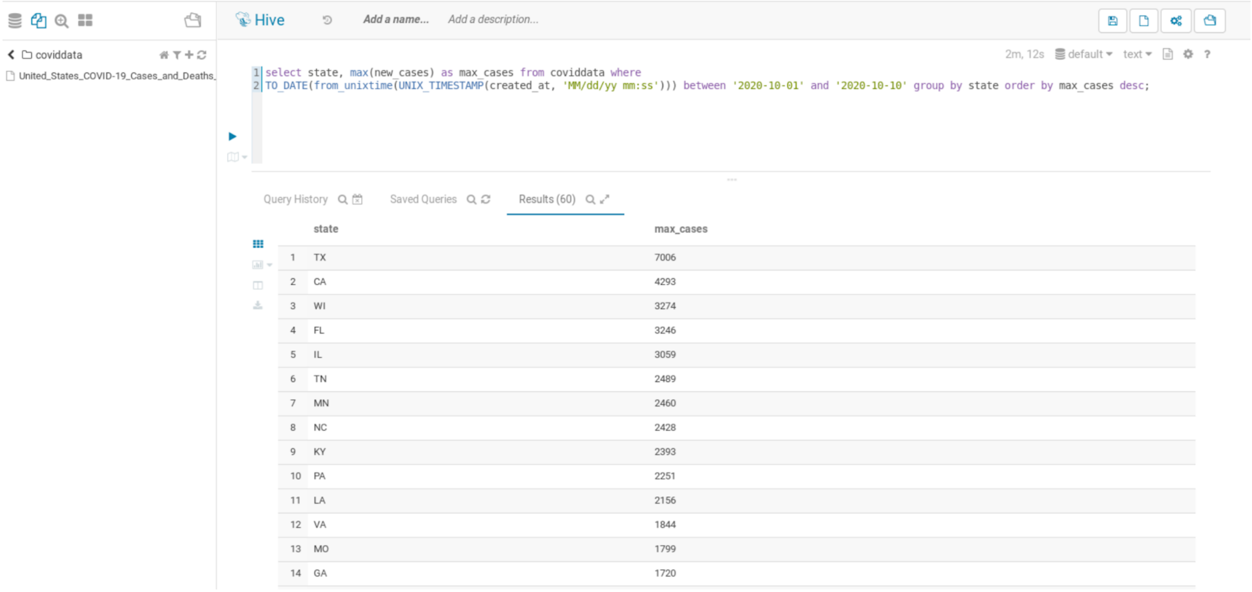
Highest single day rise:



```
select c.state, c.new_cases as max_cases, c.created_at from coviddata c
join
(select state, max(new_cases) as max_cases from coviddata group by state order by max_cases desc limit 10) m
on m.state = c.state and m.max_cases = c.new_cases;
```

	c.state	max_cases	c.created_at
1	NY	17844	04/05/2020 04:22:39 PM
2	TX	17820	09/23/2020 01:51:22 PM
3	FL	15135	07/13/2020 03:19:34 PM
4	CA	12807	07/23/2020 02:48:19 PM
5	NYC	8593	04/09/2020 04:22:39 PM
6	NC	6142	09/26/2020 01:46:07 PM
7	IL	5594	09/05/2020 02:09:45 PM
8	MI	5298	06/06/2020 02:31:37 PM
9	MA	4946	05/25/2020 03:38:40 PM
10	AZ	4877	07/02/2020 03:10:03 PM

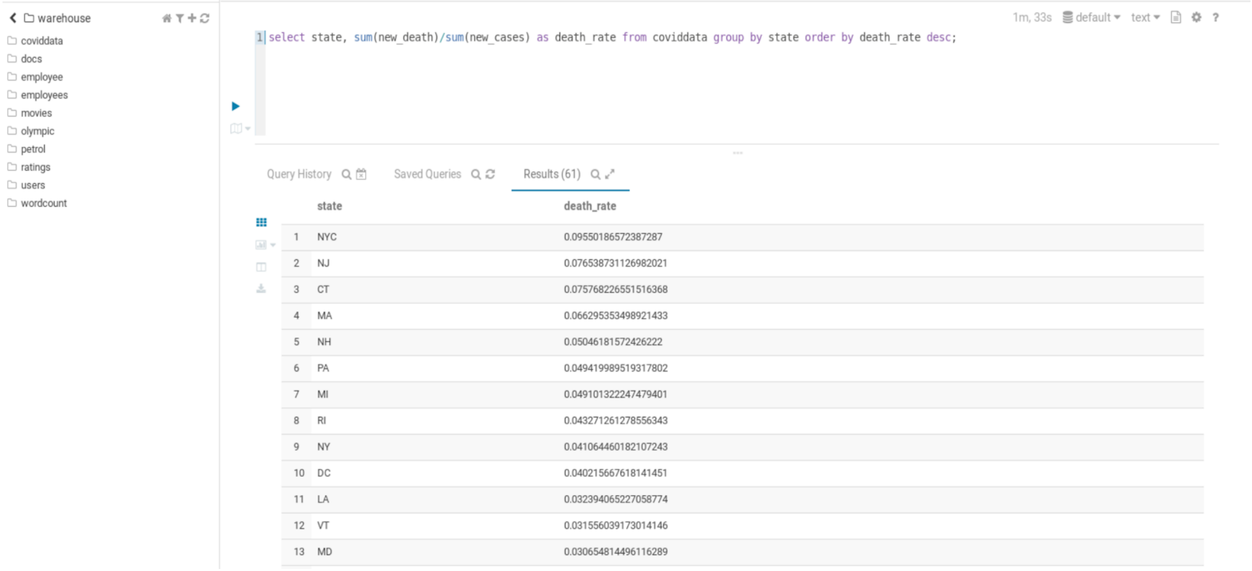
October highest cases by state:



```
select state, max(new_cases) as max_cases from coviddata where
TO_DATE(from_unixtime(UNIX_TIMESTAMP(created_at, 'MM/dd/yy mm:ss'))) between '2020-10-01' and '2020-10-10' group by state order by max_cases desc;
```

	state	max_cases
1	TX	7006
2	CA	4293
3	WI	3274
4	FL	3246
5	IL	3059
6	TN	2489
7	MN	2460
8	NC	2428
9	KY	2393
10	PA	2251
11	LA	2156
12	VA	1844
13	MO	1799
14	GA	1720

Death rate as compared to total case:



```
select state, sum(new_death)/sum(new_cases) as death_rate from coviddata group by state order by death_rate desc;
```

	state	death_rate
1	NYC	0.09550186572387287
2	NJ	0.076538731126982021
3	CT	0.075768226551516368
4	MA	0.066295353498921433
5	NH	0.05046181572426222
6	PA	0.049419989519317802
7	MI	0.049101322247479401
8	RI	0.043271261278556343
9	NY	0.041064460182107243
10	DC	0.040215667618141451
11	LA	0.032394065227058774
12	VT	0.031556039173014146
13	MD	0.030654814496116289

Project Management:

Work completed:

We did analysis on different datasets available and identified the relevant dataset as per our requirement. We also did analysis on big data technology learned in class to finalize the technology for our project. Each team member selected one technology and worked on that. We have completed all the features mentioned above.

Responsibility (Task, Person)

Neha Navgale:

1. Identified the correct dataset.
2. Loaded data in hadoop for mapreduce operation.
3. Loaded data on Hive table to perform analysis to get answers to some of the questions mentioned in feature 2
4. Contributed to the project report.

Christian Barlow:

1. Wrote Sqoop code to transfer data to MySQL database
2. Research front end technologies
3. Contributed to project presentation
4. Contributed to project report

Krishnapriya Akula:

1. Worked along with Jayadeep for importing the dataset into Cassandra and compiled some basic CQL queries.
2. Contributed to project presentation
3. Contributed to the project report.

Jayadeep Kumar:

1. Importing the dataset to Cassandra.
2. Compiled some basic CQL queries to answer some of the questions.
3. Contributed to project presentation
4. Contributed to the project report.

Andrew Poitras

1. Researched and analyzed use of Apache Solr and Apache Lucene for project application use
2. Contributed to the project presentation
3. Contributed to the project report

Contributions (members/percentage)

We equally contributed to the project. We divided the tasks and completed our tasks on time. We gathered on a timely basis on zoom call and updated each other with the latest status.

Work to be completed

1. Write a python program to consume API to get real time data.
2. Write spark queries on real time data to solve given questions.
3. Develop front end applications for visualization.

Responsibility (Task, Person):

Neha Navgale:

1. Research on using spark in python.
2. Research on how to output spark output on the angular front end.
3. Analysis on different visualization libraries.
4. Development work of python backend and angular front end application.
5. Write some of the spark queries.

Christian Barlow:

1. Development work of python backend and angular front end application.
2. Write some of the spark queries.

Krishnapriya Akula:

1. Development work of python backend and angular front end application.
2. Write some of the spark queries.

Jayadeep Kumar:

1. Writing a python program to consume API to get real time data
2. Write some of the spark queries.

Andrew Poitras

1. Development work of python backend and angular front end application.
2. Write some of the spark queries.

We will further split the development work into tasks and divide the work.

Issues/Concerns

No issues so far.

Story Telling:

Who:

The dataset is about the people who were impacted due to covid 19 in different states and territories of the USA. They are the representative of the main characters of the story as this will help to understand the trend of covid -19 impact on different locations of the USA. The data does not contain any identifying information nor does it have risks of disclosing identifiable information, it is mostly anonymous geographical and medical information.

What:

The dataset records the number of people tested positive, recovered, deceased, hospitalized in each and every state on each day.

When:
The data is collected everyday from official sites of each state and placed at covidtracking.com. As part of the first and second increment, we collected the data in csv format but for next increment we will be consuming API to get the real time data. Covid-19 started spreading in the USA from March 2020, hence the data is available from March and gets updated everyday.

Where:
Global data is available but we narrowed down the scope of our project to the USA hence we have collected data for all the states and territories of the USA. The state variable in the dataset geographically separates out the data.

Why:
The data is very crucial to understand the ongoing pandemic and its effect on every sector. The data is collected to analyze, understand and identify the gaps in preventive measures taken.

References:

https://covidtracking.com/
https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36
https://ourworldindata.org/mortality-risk-covid
http://ocel.ai/