Project-3: Big Data Engineering and Architecture

A walk in the spark: Covid County and State analysis of USA

Shashank Magdi

Project Plan:

Goals: Analysis of COVID-19 can unearth a rich array of possibilities, predictions and insights due to a plethora of available data from different domains which sometimes not only helps us develop some intuition from the past, but also make us understand how might our future stand. My current goal in this data movement against COVOD-19 is to understand how the disease has affected us at an even more microscopic level such as counties and states within the United States before diving deep into the global level where even diverse fields of data is available to use. Once the underlying patterns at this level are identified, the same can be aimed to be executed at a global level where a lot more parameters are at hand and a lot less clean data available. This project intends to understand how the virus has spread across different counties, which ones were most affected, how did each of the states fare whilst battling the pandemic. A hypothesis one can think of, when are the surge in the cases the highest, festive times? Interesting questions like these can be answered. This information at microscopic level, especially while understanding the trends like during what period there were a surge of cases, in what counties surge continued for a prolonged period of time and what could be the possible predictions based on the cleaned data at hand can be provided to Government authorities to take precautionary measures and prevent another wave of pandemic starting at a local level.

Data Description:

The dataset I have chosen, after a lot of deliberation is the New York times Covid 19. The New York Times has been releasing a series of data files, each having cumulative counts of the coronavirus in the United State, at both state and county level, over time.
I would be using the US-counties, dataset for primary analysis. The US-daily state wise vaccination data and the US-daily State wise case data from the John Hopkins(JHU) repository have been chosen to try and combine those datasets and gain insights.
The link to the dataset can be found here : https://github.com/nytimes/covid-19-data.
The link to the secondary dataset can be found here:
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

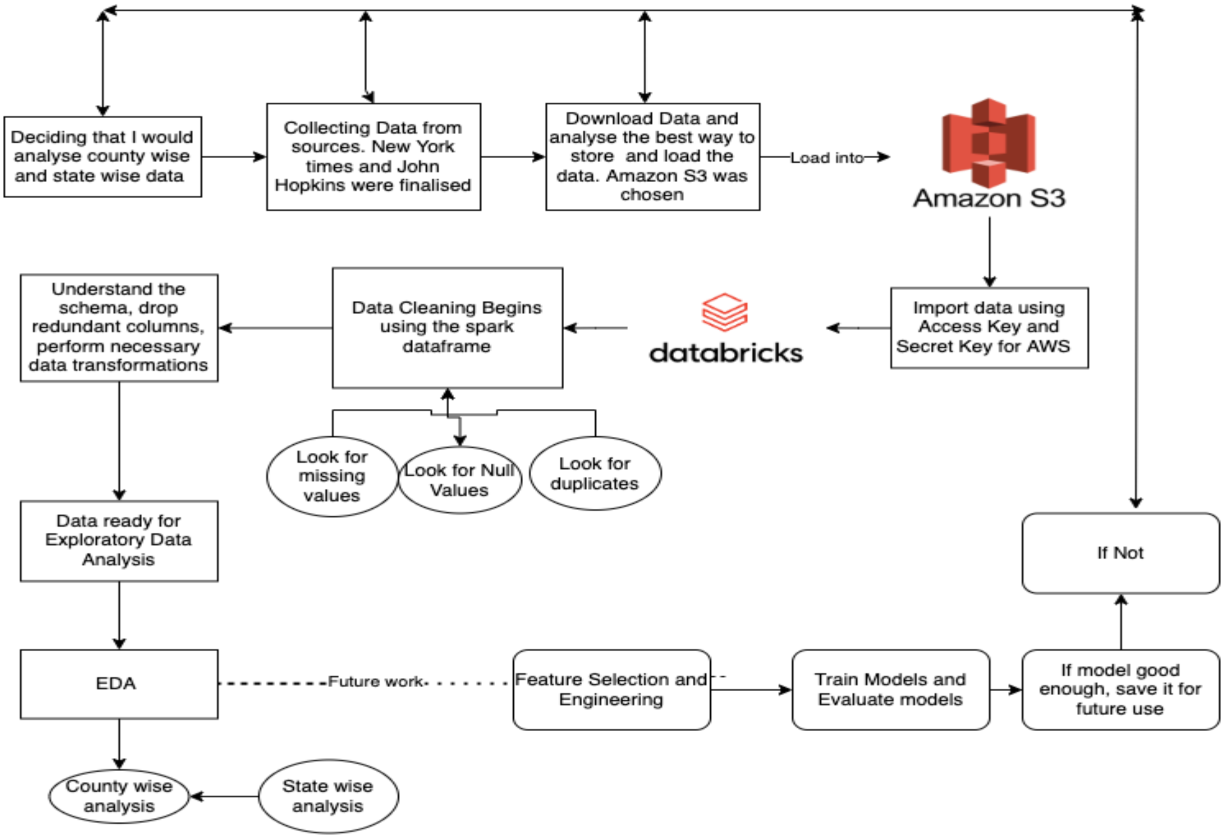
About the data :

The data is the end product of dozens of journalists working across various time zones, monitoring news conferences, readily analyzing the releases of data and seeking clarification from officials on how they categorize the cases. The primary data that has been published is the daily cumulative number of cases and deaths reported across each county and states,. NYT have additionally published cases about cases in Prison, county wise mask usage and college wise usage as well. I had taken the decision to use both NYT and JHU datasets as I felt the quality of the data for counties was very well showcased in NYT and JHU dataset for state wise daily cases and vaccinations was a succinct yet powerful format. These sources also have been validated several times, which is always necessary.

Data Dictionary for the data:

Data Field	Data Type	Data Description	Analyzed	Nullable
Date	Timestamp	Date of Covid case record	Yes	True
County	String	Name of the county	Yes	True
State	String	Name of the state	Yes	True
Fips	Integer	Federal Information Processing Standard code	Yes	True
Cases	Integer	Cases on the particular day	Yes	True
Deaths	Integer	Deaths on the particular day	Yes	True
Province State	Integer	Name of the state	No	True
Lat	Double	Latitude	No	True
Long	Double	Longitude	No	True
Combined key	String	Combined State, county	No	True
People fully vaccinated	Integer	Number of people fully vaccinated	Yes	True
People partially vaccinated	Integer	Number of people partially vaccinated	Yes	True

Architecture and Analytical lifecycle of the project:



Deciding that I would analyse county wise and state wise data
Collecting Data from sources. New York times and John Hopkins were finalised
Download Data and analyse the best way to store and load the data. Amazon S3 was chosen
Load into
Amazon S3
Import data using Access Key and Secret Key for AWS
databricks
Data Cleaning Begins using the spark dataframe
Understand the schema, drop redundant columns, perform necessary data transformations
Look for missing values
Look for Null Values
Look for duplicates
Data ready for Exploratory Data Analysis
EDA
Future work
Feature Selection and Engineering
Train Models and Evaluate models
If model good enough, save it for future use
If Not
County wise analysis
State wise analysis

Methodology:

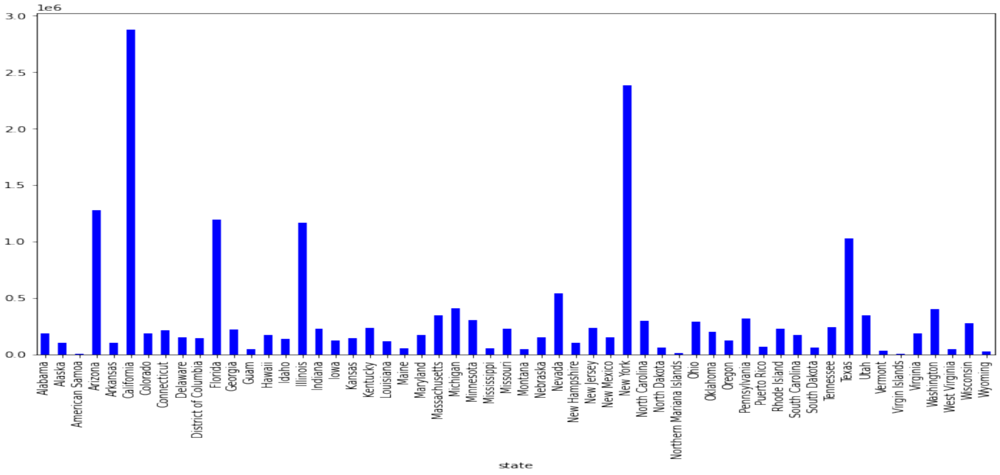
a) Data Loading:
The data has been loaded into the Databricks environment with the help of Amazon S3 bucket. Once a mount has been established between Amazon S3 bucket and data bricks environment, one can simply read files as they wish. Spark Data frames were created for each of the files and using spark functionality these were displayed to get a better understanding of what the data meant.

b) Data Cleaning:
Summary statistics of the dataset conveyed the underlying numerical trend using count, average values. Before trying to impute any uncleaned values, knowing the schema of data is vital. The first dataset, US counties did not seem to have any duplicate values, missing values, although there were a few null values that had to be cleaned. Null values in the death column constituted to 2.3 % of all values, to 0.95% of all FIPS values. Another observed problem whilst trying to understand the data was that there were plenty of unknown county values, constituting to 0.85 % of all counties. These were the first things to be taken care of. The deaths column null values were set to zero, while drastic measures like imputing the null values in other columns weren't taken as this could be done if these values drastically skew the results. Next, the vaccine and the state daily records data frames were to be cleaned. Similar methodology and steps have been implemented here as well. One important note, plenty of cleaning on the columns had to be performed on the Vaccine and state data frames, like setting null values to zero for the days when vaccinations weren't available yet. Also unnecessary columns like Latitude, longitude, combined key were omitted amongst other cleaning tasks.

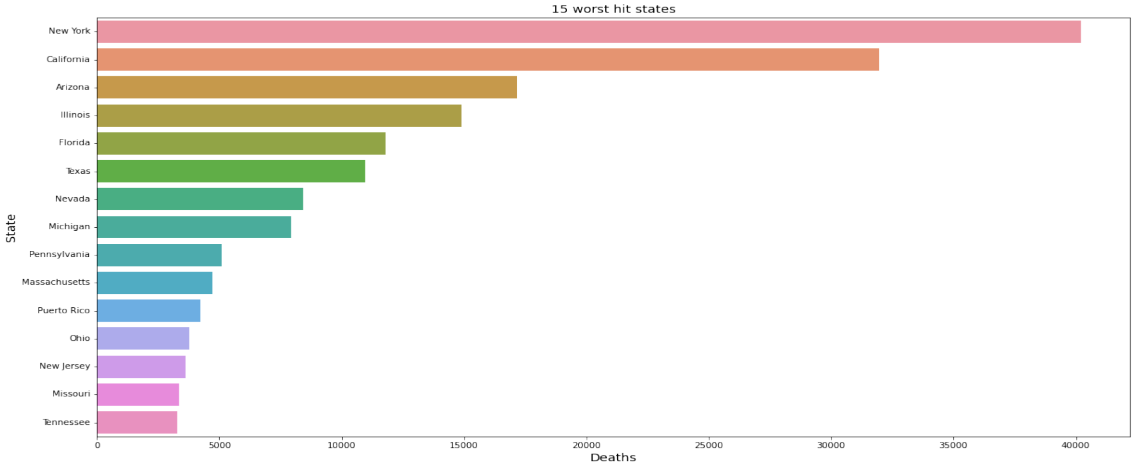
c) Data frame Join:
The states, vaccine data frames were converted from spark data frame to Pandas data frame and an outer join was used to join the tables pivoted on FIPS code, Province name columns. The columns deemed unnecessary were imputed to make the data look cleaner.

d) Exploratory Data Analysis:

Q) How did the states fare in terms of cases recorded?
Ans) We can clearly see how badly the states of California and New York were hit, with Texas and Florida and Arizona also badly affected.



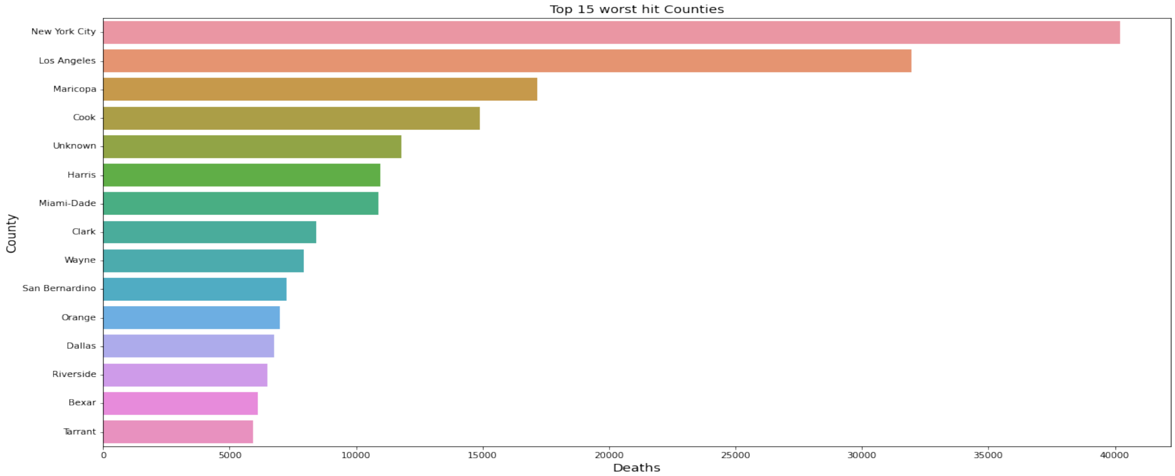
Q) What were the top 15 worst hit states?



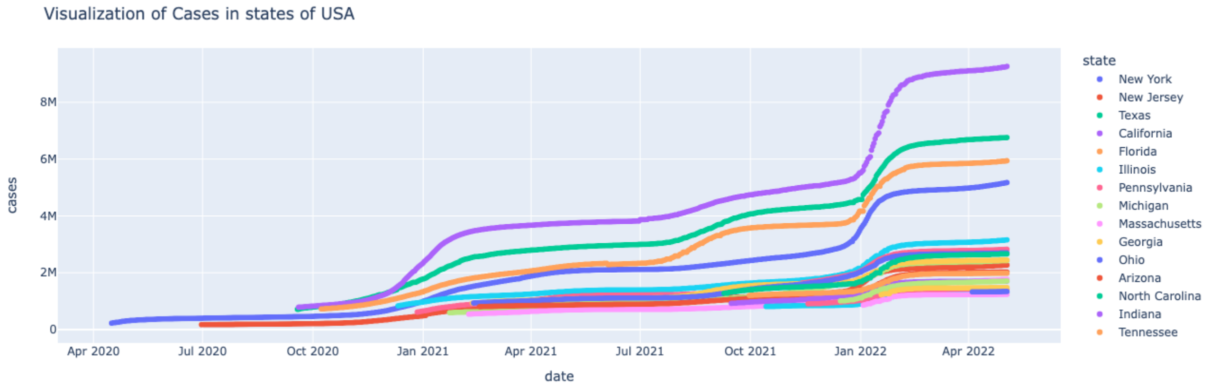
This visual clearly depicts the magnitude of cases that were recorded in states like New York and California over the pandemic, which is certainly the case in reality as these states suffered heavy causalities and recorded humongous numbers in terms of cases recorded.

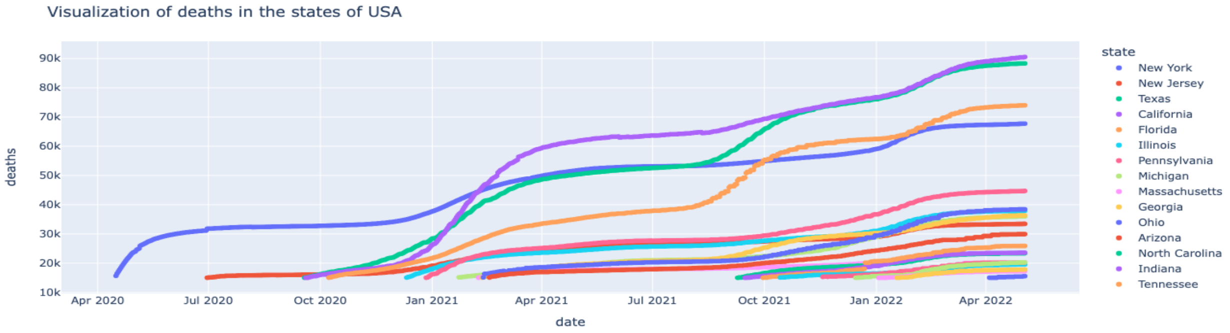
Q) What were the top 15 worst hit counties?

New York City, Los Angeles were the usual suspects given how widespread coverage the surge in cases of these cities received. However Maricopa, Cook counties also were drastically effected.



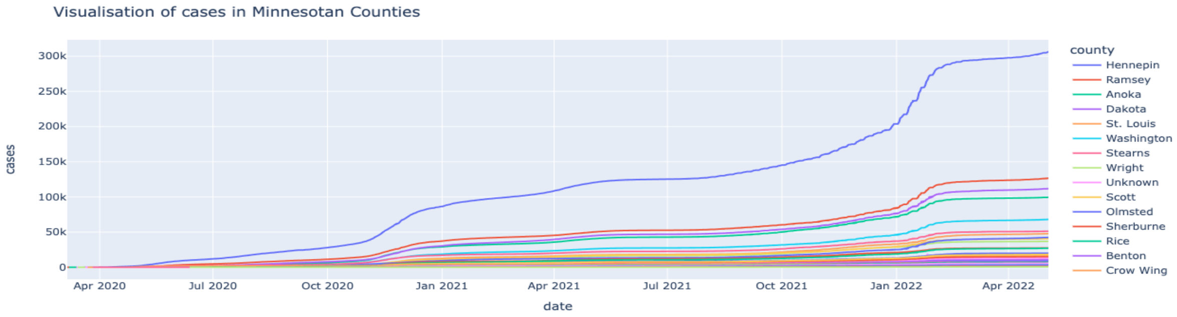
Q) How was the cases and deaths trend in the states of USA?


Visualization of Cases in states of USA

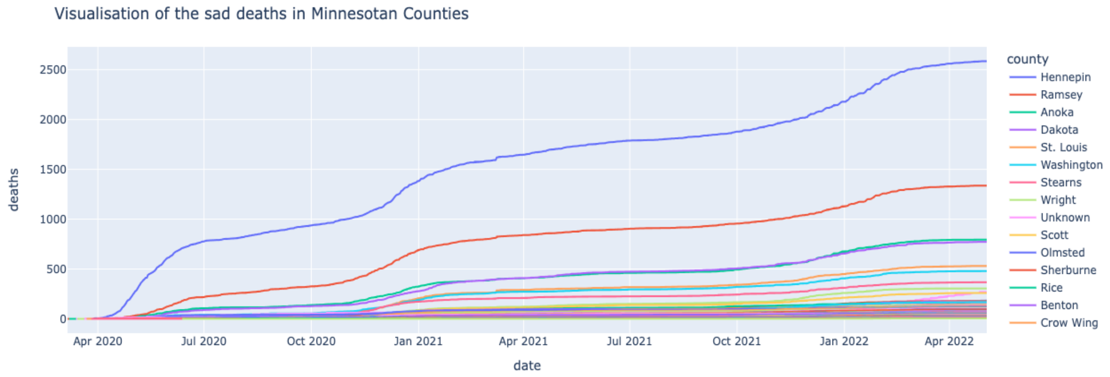

Visualization of deaths in the states of USA

There is a sharp spike in New York between January and April of 2021 and the cases were on a constant increase since then. Almost 1.5 million case increase happened in the state of California within 2 weeks between 31st December and 15th January 2022, which is quite peculiar.

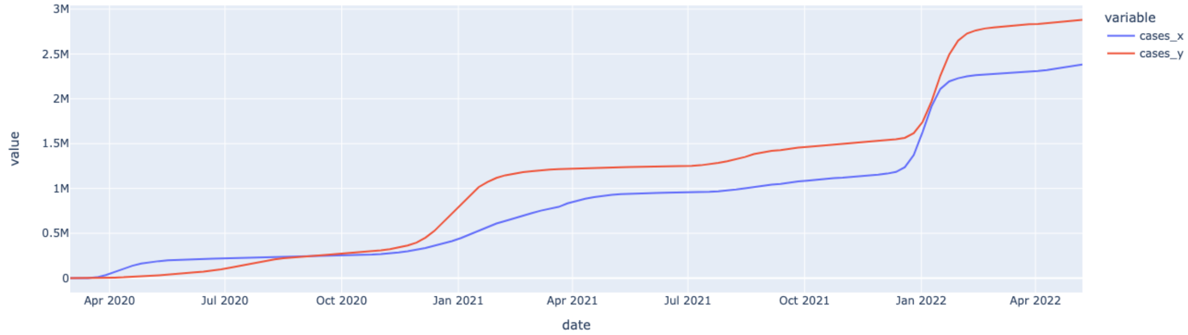
Q) How was the local counties of Minnesota effected during the pandemic?


Visualisation of cases in Minnesotan Counties

It can be clearly seen that the Hennepin County was the worst hit county of all of Minnesota with more than 300k cases since the advent of Pandemic. The same trend follows with the scenario of deaths too. One can clearly see the rise in cases around the time of New Year during both 2021 and 2022 during which people usually tend to socialize and gather around, therefore increasing the chances of the spread of virus!



Q) Since New York and California were the worst hit states, How did they fare against each other?



Ans) In terms of cases both New York and California were equally hit in the early phase of Pandemic till December 2021. However, there has been a constant spike of cases(again particularly during Jan 2021 in California cases as the cases were almost double at 1.0M on Jan 17 in California, whilst New York reported cases tad higher than 500k.

Lessons Learnt :

1) Data is messy!, very messy! Especially when one is dealing with time series data in correlation with real life, there are a lot of inconsistencies such as counties not being assigned, left unknown. Not to mention the number of null values that exist, and a sharp call needs to be taken on which ones are to be dropped and which are to be imputed.
2) A thorough understanding of a spark cluster anatomy, with the driver at the apex and worker – executor architecture helps in distributing the compute power across multiple nodes. Spark job being broken into multiple stages and each stage having multiple tasks certainly helps achieve distributed computing in the way Spark does..
3) Some functionality in spark data frame is not in line with Pandas data frame. One such instance I had encountered is while filling the null values was one could replace null values with values of any data type in Pandas, but spark data frame has a restriction of filling with the same data type.
4) Data Cleaning is one of the most important aspects in an analytical project lifecycle. Most of EDA and ML modelling depends on how clean the data is. If one figures out the underlying pattern and trends of the data with insightful EDA after proper Data Cleaning, the modelling part always becomes easier to implement as one understands the features that can be engineered correctly.

Future Improvements:

With the data cleaning knowledge and combining datasets experience that has developed throughout the course of the project, the next step will be to implement visually succinct heat maps which describe how the cases vary across different time periods. Powerful libraries like Geopandas and comparative population dataset sources like Vega datasets can be used to construct a choropleth is something that I would look forward to do. Such graphs truly underline the saying that one picture speaks a thousand words and purely from a data point of view, understanding the presence or strength of a statistical trend spread across time certainly got me curious. Also the next step would be trying to implement robust ML models to carefully try and predict future scenarios with caution as this is not a trivial aspect to be dealt with and would certainly need intensive mathematical modelling to mimic the scenario and even predict a chunk of the future!

Link to the notebook: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1441892554850136/2454928379343047/256159671845597/latest.html

References :

1) https://github.com/nytimes/covid-19-data.
2) https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.
3) https://medium.com/analytics-vidhya/cleaning-and-understanding-multivariate-time-series-data-6554eefbda9c
4) https://towardsdatascience.com/python-data-preprocessing-using-pandas-dataframe-spark-dataframe-and-koalas-dataframe-e44c42258a8f
5) https://app.diagrams.net/
6) https://www.districtdatalabs.com/how-to-start-your-first-data-science-project
7) https://aws.amazon.com/s3/getting-started/
8) https://www.kaggle.com/code/tanujdhiman/us-countries-covid-19-report
9) https://towardsdatascience.com/making-heat-maps-with-literal-maps-how-to-use-python-to-construct-a-chloropleth-6b65e4e33905
10) https://www.kaggle.com/code/fireballbyedimyrnmom/a-basic-covid-19-comparison
11) https://towardsdatascience.com/how-to-replace-null-values-in-spark-dataframes-ab183945b57d
12) https://github.com/hritikbhandari/Exploratory-Data-Analysis-of-COVID19-with-python/blob/master/COVID_19_EDA.ipynb
13) https://stackoverflow.com/questions/52287553/how-to-create-a-copy-of-a-dataframe-in-pyspark