Project B - Big Data Concepts
Capital Bikeshare*

Deepak Khirey
Indiana University, Bloomington
(Dated: December 3, 2019)

I. INTRODUCTION

Artificial Intelligence and Machine Learning has become the buzzword in the field of Data Analytics and Business Intelligence. Although Machine Learning algorithm is a critical piece of the puzzle, there needs to be an entire data centered eco-system built to realize and support reliable predictions.
It has to start with a definite Business goal in mind and a carefully crafted Business Strategy to collect and leverage real time data. We have to implement end-to-end Data Pipeline consisting of **Data Acquisition, Data Exploration, Data Profiling and Cleanup, Data Transformations, Data Ingestion and Data Visualization**. All these activities are vital and ensure that the Data Lifecycle is robust to support tangible outcomes.
For this project, I am going to use Capital Bikeshare dataset and will implement Data Lifecycle and Data Pipeline to optimize bike utilization at each Bike stations. I find this dataset particularly interesting because it is actual data from the field and it provides an opportunity to think of data management from business perspective.

II. BACKGROUND

Business Model

Capital Bike share launched in 2010 is a bicycle share company which operates in and around Washington DC. [1] It has 4300 bikes and more than 500 bike stations around the city. All of its renting actions are done through a mobile app. so User has to download App from Playstore on Mobile and after registration, user book the bike from any location in the city. Once booked, user has to just unlock the bike from station, they can ride anywhere in the city and can return the bike to any station when they are done. Rides have 3 types of booking ways - Single trip - It is 30 min ride and charged $2 flat. 24-Hour pass - Users can keep the bike with them for a day at charge of $8. Annual Membership - It is $85 per year and comes with unlimited 30 min rides.[2]

Business Goal

This is a very interesting business model and profit can be made by adding more and more Annual Memberships and 24-Hour passes. Where that would need extensive Marketing campaign, another way to improve operational efficiency is by predicting time taken by each user, frequency and popularity of station and making sure that bikes are available as per demand. This is the exact Business Goal I'm trying to deal with in this project.

BigData Strategy

As we saw in "Big Data Dreams", Mazzei and Noble(2017) [3], Big Data and Strategy are complimentary to each other. Rather than defining strategy first, it benefits more to leverage data and implement strategy around it. To improve predictions about usage of bike, it is very important to gather the data on every ride. This is enabled by locking-unlocking of the bikes for every ride. This provides exact time of pickup and drop off for bike along with location.

Insights to uncover

Once the data is available, business strategy can be formulated to maximize profit by studying the patterns of user behavior. What is most popular bike station? What is peak time for demand? Who are the most frequent users? Based on answers of these questions, it is the possible to ensure that users are given incentives, bikes are present when there is demand at particular station etc.

III. DATA DETAILS

Capital Bikeshare provides data from 2010 to 2017 in year-wise folders and 2017,2018 data in month-wise folder. It also provides real time data for creating streaming applications. However, considering scope of this project I used 2017 data only. It is complete set for all four quarters and provides necessary use cases to be covered in this project and establish a full data cycle. [4]

Data Model

Capital Bike share has taken efforts to make their bike rides data available publicly for general studies. This data is stored in ZIP format in year-wise manner. Each zip file contains 4 CSV files corresponding to each quarter of the year. CSV file contains following fields [TABLE:I]-

Data Quality

The company has ensured that the data does not contain

* I535 Management, Access and Use of Big and Complex Data, Fall 2019

Column Name	Column Description
Duration	Duration of trip
Start Date	Includes start date and time
End Date	Includes end date and time
Start Station	Includes starting station name and number
End Station	Includes ending station name and number
Bike Number	Includes ID number of bike used for the trip
Member Type	Indicates whether user was a "registered" member (Annual Member, 30-Day Member or Day Key Member) or a "casual" rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

TABLE I. Columns in the Dataset

data related to test stations, false starts (less than 1 minute ride). So it is clear that the data is from actual customers and there is no bias or noise due to bad data. The dataset does not have any NULL, NaN values. So there is no need to cleanup data in that aspect.
All the data is in string format in CSV. We will have to apply respected Data Types before putting it to use.

Tidy Data
As we have studied in "Tidy Data", Hadley Wickham,[5] it is greatly advised that data should be arranged in proper format so that it can be processed efficiently. Data has to follow certain guidelines which are proven best practices.
Capital Bikes has thankfully provided this dataset in a Tidy Data format. We can see prominent feature of a well formatted Tidy dataset as follows-

- Each variable in this dataset forms a column and it provides information about that specific feature. For example, Duration tells us how much time the ride was taken.
- Each row tells us information about one particular ride.
- All records containing rows and columns together forms an observational unit of bike rides in the given time period.

IV. METHODOLOGY

For a successful execution of any data analytics project, it is very important to identify suitable methods of Data Lifecycle. This is generally driven by nature of business and nature of data being collected. Once we finalize lifecycle for data, then we can decide on Data Pipeline ie actions to be performed to realize the lifecycle. Of course, we need to setup our environment with all applications that are necessary to perform these actions may it be storage, transformations, analysis and visualization etc. We also need to devise a strategy of how we are going to store this data and for how long, what kind of access will be provided to consumers of the data throughout the lifecycle. I have followed this thought process for Capital Bike share project as below-

Data Lifecycle
Data Lifecycle tells us about maturity of data at that stage. What to expect from data and what needs to be done further. It is kind of high level way which leads to design data flow.
For bike rides dataset, I think **USGS Data Lifecycle** is most suitable to follow. It has lifecycle states as **Plan - Acquire - Process - Analyze - Preserve - Publish**. [6] At a high level this flow caters to our needs of getting new data frequently, manipulate it as we need for analytics input, store in NoSQL database so that large volume and schema changes can be accommodated and then build visualizations to monitor health of the process and gain insights that are actionable by business.

Data Pipeline
Data Pipeline, on the other hand, explains how to achieve a stage of data. It is more about action level details.
This data lifecycle fits perfectly well with Data Pipeline where Capital Bike plans to get data from every bike ride and makes it available over the internet. We can acquire data by manually downloading it at frequent interval. We can then transform this data to suit our needs with tools like OpenRefine and then analyze for predictions using python libraries such as sklearn. We can preserve this data into NoSQL database, MongoDB in this case and then visualize in a dashboard format with Tableau.

Technological Setup
There are various tools and applications play a very vital role to execute data pipeline actions. This is infrastructure layer which enables the business strategy based on BigData. For current project, I did following technological setup [TABLE: II]

V. EXECUTION AND RESULTS

ETL Approach
As we discussed above in Data Pipeline section, we are following a typical ETL (Extract-Transform-Load) approach for this project. This is most proven approach and it is used in the industry for long time. [7]. Benefits of this approach are as below-

- This is simplistic process generic for most of datasets
- we can sequence the steps which are repeatable for similar data increments

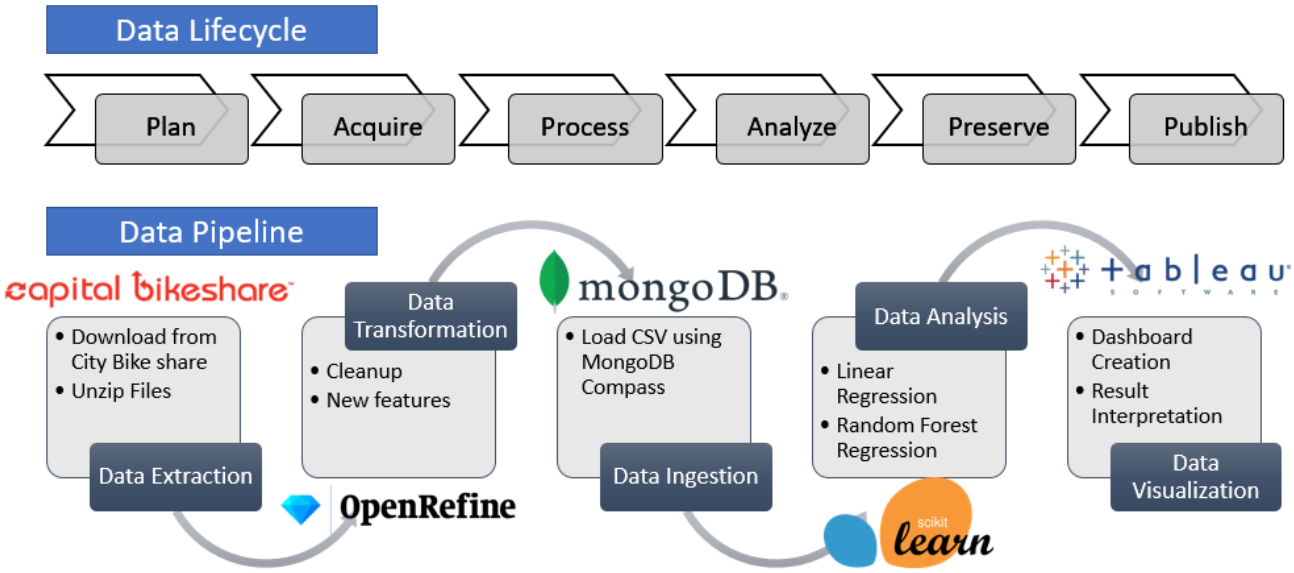

FIG. 1. Data Management Plan - Data Lifecycle and Data Pipeline

Application	Purpose
MongoDB	NoSQL Document Database to store CSV files
Tableau Desktop	Visualization Tool for creating Dashboard
Jupyter Notebook	IDE for writing and executing Python code
OpenRefine	For Data Cleanup and Transformations
python libraries	pandas,pymongo,scipy,scikitlearn,numpy
MongoDB Compass	User Interface for MongoDB
Anaconda	Distributive Environment for Python libraries and Jupyter Notebook
7-Zip	For unzipping acquired data from website

TABLE II. Technological Setup

- we get an opportunity to explore and enrich data before it gets loaded into database
- We can ensure loaded data confirms data quality requirements

Data Extraction

Capital Bike data is available for extraction on company website over the internet. It is publicly free and it requires no authentication. So the data need to be downloaded on local environment. Once the data is downloaded, it is in ZIP format. It needs to be unzipped so that we get CSV file corresponding to each quarter in the year.

Data Exploration

Next step after extraction of data is to get familiar with it. Here we have to take help metadata information provided by Capital Bike. Below are key points observed about the dataset-

- We have to look for any inconsistencies in the dataset. Fortunately, this dataset is clean and there are no major inconsistencies found.
- Start Date and End date are consistent in the format yyyy-MM-dd hh:mm:ss eg.2017-01-01 00:00:41
- Duration is the time difference between End Date and Start Date. It is expressed in terms of seconds.
- Start Station and End Station are string fields and they are expressed as crossroads with road names separated by ""
- Member Type has only two type "Member" and "Casual". However, it does not indicate what kind of pass is used.
- Bike number starts with "W" followed by 5 numbers.

Data Transformation

Information available in the dataset is correct and clean but it is not enough to develop a predictive model. We have to perform additional Feature Engineering to derive derive new features which we think can help to get specific insights. We need to transform data to generate these features and I have chosen OpenRefine as tool to accomplish this task.

As we have seen in the OpenRefine assignment, OpenRefine is an opensource tool developed by Google. It is

great tool for Data Manipulation and it has many key functionalities to experiment with data.[8]
For Capital Bike dataset, I performed below transformations on the dataset in OpenRefine [FIG: 2]-

1. Every value in each field is surrounded by double quotes, and it needs to removed because it will treat every field as string. We need to apply suitable Data Types to each field so while importing the CSV I selected the option "Use character " to enclose cells containing column separators"

2. Member Type - This field is in String format. However, for predictive algorithm it is suitable if this field is in Boolean format. So assigned "Casual"=0 and "Member"=1 using Text facet option.

3. Start Station, End Station - This field has camel case words, so applied toLowercase() function for uniform lower case letters using Common Transforms option.

4. Start Date, End Date - This field contains pickup and dropoff timestamp, hence applied Datetime Data Type to this field by todate() function using Common Transforms option.

5. Start Month - Derived this feature as month value from Start date with value.datePart("month") function using "Add column based on this column..." option.

6. Start Day - Extracted Day of the ride from Start Date column with value.datePart("weekday") function using "Add column based on this column..." option.

7. Start Hour - Derived this feature as hour value from Start date with value.datePart("hour") function using "Add column based on this column..." option.

8. Start weekend - This column is to indicate whether the ride was on weekend or not. This is accomplished using Text Facet by setting Saturday/Sunday = 1 and others values as 0.

OpenRefine comes with a very nice feature to export actions in a JSON format which works as a template for future use. It helps to standardize sequence of actions on raw data to get uniformity in data Transformation over the time.
All above actions need to be repeated on every quarterly CSV input file. I used this feature to process all CSV files before proceeding further to Data Ingestion.

Critical Insights

1. We are dealing with on Start date because in a Predictive model for Duration we will not know the end date. We will have to predict End Time based on Start time

2. Start month will help us considering that Bike rentals will be on peak in Spring and Summer seasons. Month number will show co-relation with number of rides.

3. Start hour will help illustrate morning, afternoon, evening and night hours rentals. Logically, morning and evening would be high demand time periods.

4. Although Bike Number is given, I don't think it would have any co-relation with its ride frequency. So we are not leveraging this field.

Data Ingestion

It is very important to store collected data in a suitable Database. For the scope of this project, I have chosen MongoDB as my NoSQL Document database.[9]

Why MongoDB?
As we have seen in IFRI case study, we have to choose NoSQL database which is most suitable to the structure of data.

- In case Capital Bike dataset, data is available in CSV format. So it is more or less suitable for RDBMS database, however given the volume of data, it is advisable to use NoSQL database which can leverage Distributive storage capabilities.

- Here, MongoDB is chosen because each record of this dataset is complete in itself and can be stored as JSON file. So it can be imagined as storage of large number of JSON files which may have dynamic schema if business model evolves.

- Also MongoDB has seamless connectivity with Python, Pyspark through pymongo library. This makes it best choice from analytics point of view. We can directly connect with database and perform predictive algorithms on it.

- MongoDB also connects with Tableau visualization tool seamlessly, so we can utilize MongoDB BI Connector to create dashboards directly from database.

Once the database is selected, it is relatively easy to import CSV files into the database with OOTB capabilities. Below steps are performed to ingest data into MongoDB [FIG: 3] -

1. Create database schema and collection.

2. Open MongoDB Compass and select Import Dataset option.

3. Select CSV as dataset type and hit import.

4. Dataset is automatically converted into JSON format. Each row of CSV becomes one JSON file and it gets unique "ID" assigned to it.

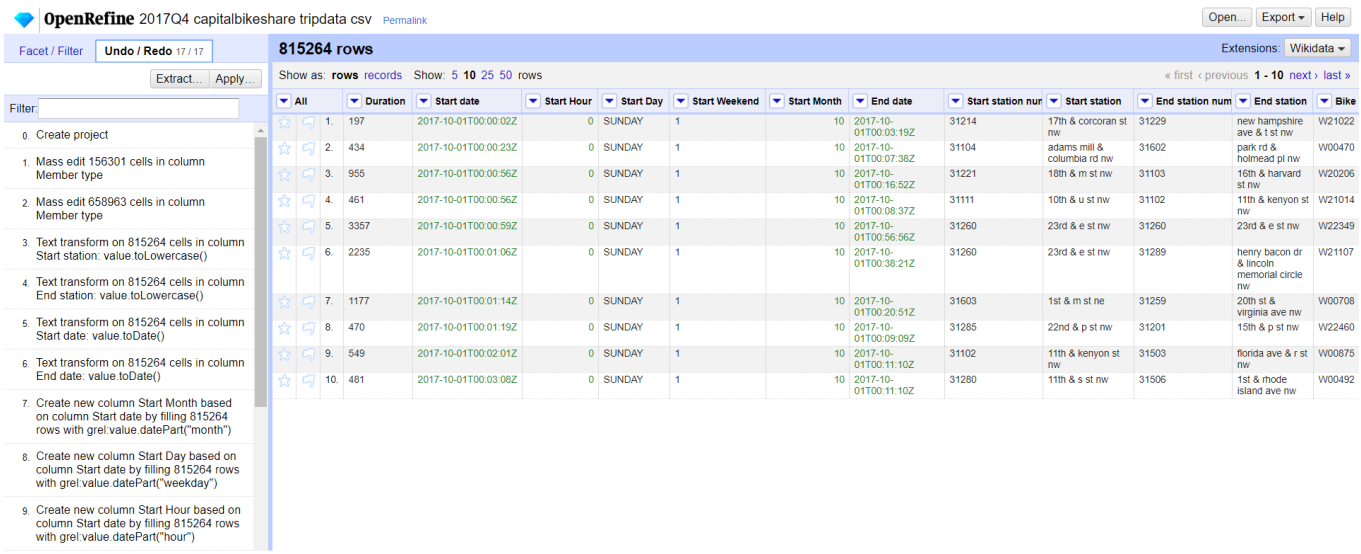

FIG. 2. Data Transformation using OpenRefine

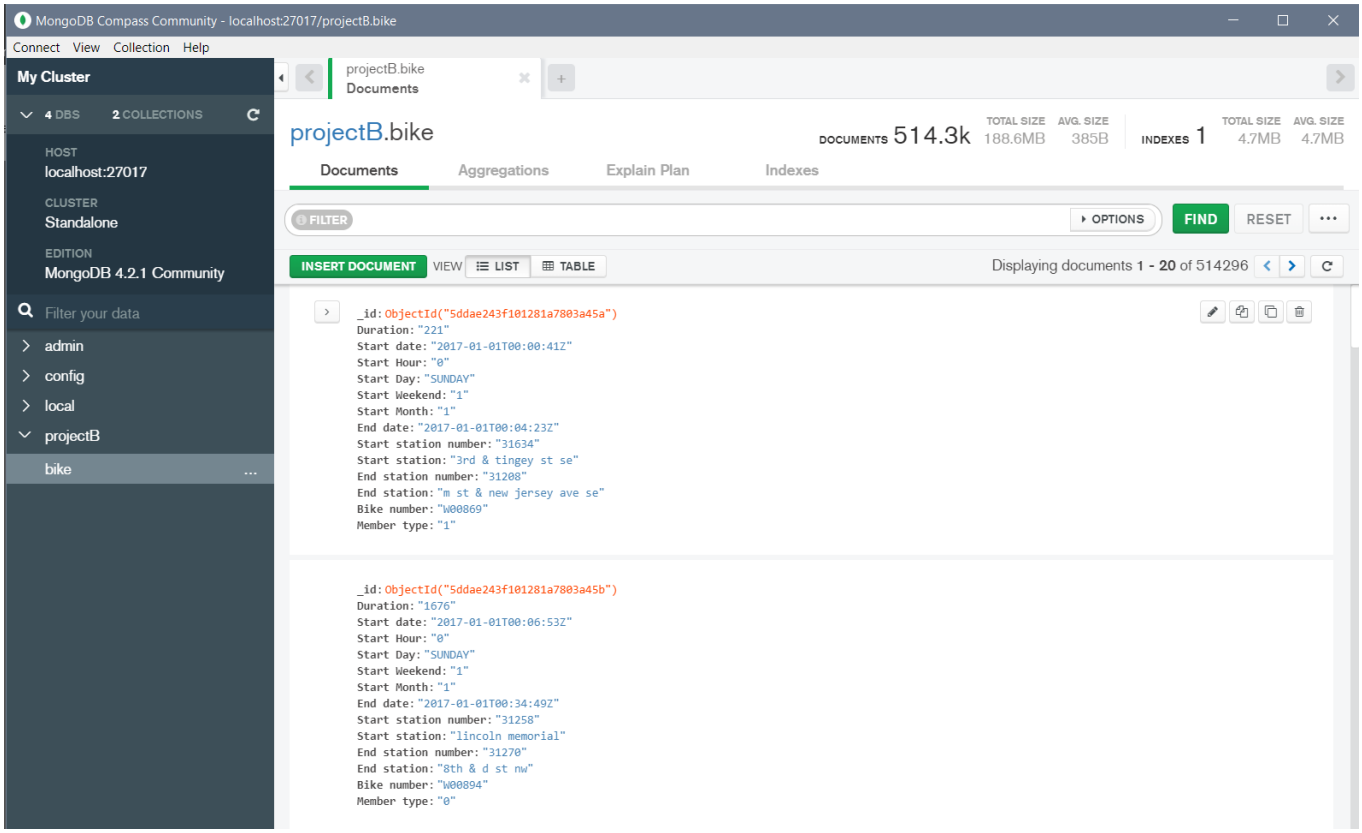

FIG. 3. Data Ingestion using MongoDB Compass

Predictive Analytics

After ETL process execution, data is now ready for performing Analytics operations on it. This is most important step to derive insights from the data. Considering this dataset, there can be different types of analysis that can be applied such as-

- KMeans clustering to identify most popular Bike Stations
- Geosptial Analysis to figure out peak demand areas of the city

- Classification of usage patterns to figure out frequent customers
- Baysian Analysis to predict End Station based on Start station
- Regression Analysis to predict how long bike will be rented

Each of this analysis could be a separate topic of study and it will cater to different business decisions. For the scope of project, I will focus on last use case which is Predictive Analysis by Regression.

Use case statement
Can we predict what time customer will return the bike if we know the Start Time, Start Station and Member Type?

Connecting to Database To execute any analysis on data, we have to first connect to the dataset. MongoDB makes this task very simple. There is opensource python library called pymongo which allows us to connect to desired collection in MongoDB programmatically.
Once connection is established with database, then we can leverage python libraries like pandas to get data into Dataframe and then perform required statistical algorithms on it.

Applying Data Types Although we have data in Dataframe format, it is all string type values because we have not applied any schema in MongoDB. We have not defined any data types on database because it give us flexibility to store all incoming values without any data loss. However, while performing analysis, we have to make sure that they are in correct data type. So we apply numeric data type to columns like Start Month, Start Hour, Member Type etc.

Histogram To gauge the distribution of the data, histogram is a very popular technique. It tells us about frequency distribution of data. So that we can understand what are most likely values in any field of the dataset. Histogram for Capital Bikes dataset shows us that,

- Duration field values are right skewed distribution.
- Members are using the service more than casual users.
- Some station numbers are more popular than others.
- Rental service is used most in daytime from 8 am to 10 pm. There is spike in the morning hours which could be business start time.

Outliers Duration is the most critical field in this analysis because it tells us how long the bike was rented. If we look at Box-Plot of this field, we can see that it is highly skewed. It has too many outliers. [FIG: 4]

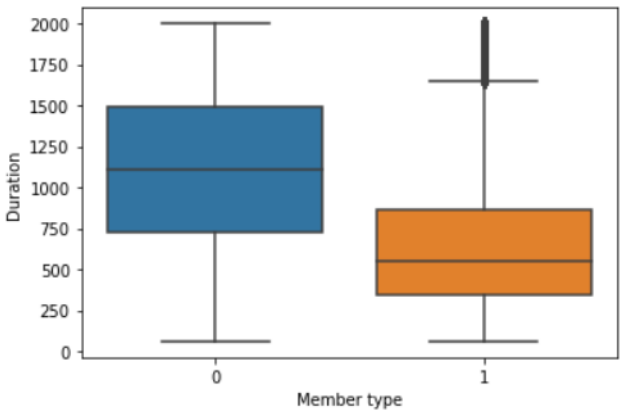

FIG. 4. Outliers in the Dataset

Critical Insight
This can be explained by the business model of Capital Bikes. Their pricing suggests they have $2 for 30 min ride and the $8 for day long ride. So there must be user behavior to return the bike either within 30 min or to keep it for day. That's why we see duration values ranging between 300-2000 sec and then suddenly 20000+ sec.
It really makes sense to predict when short time users will return the bike, because that will affect the availability of bikes at every station. So we will consider only duration less than 2000 seconds. We will remove all outlier values with high Duration.

Correlation Matrix After removing outliers, we will build the Correlation Matrix to check which fields are most related with the Duration. From [FIG: 5], We can see that-

- Member Type is most co-related field with Duration
- Start Weekend is second most related field with Duration. Its logical because Bike renting trend can be seen more on weekends.
- Start Month is third co-related field followed by Start Hour. Its again logical because season and hour of day affects renting behavior.

From the correlation matrix we come to know that we have derived some useful feature during Data Transformation process.

Test-Train Split To test our Regression model, it is good practice to split data corpus into Test Data and Train data.Before splitting test train, I defined target variable Y as "Duration" in a separate vector and features X in separate matrix. Then I split our data in 70:30

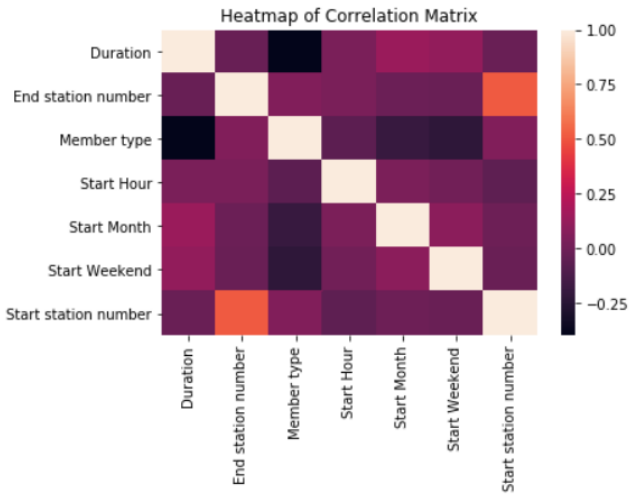

FIG. 5. Correlation Matrix wrt Duration

proportion. This will help us to establish our model parameters and check if it is overfitting/underfitting.

Normalization It is also important to normalize the data. This is a necessary step because field values are ranging differently. Duration has range of 300 to 80000 where as Start Month ha range of 1 to 12. So we have to bring them on same ground statistically so that they are not biased while building the model. StandardScaler() action makes sure that all value are normalized between the the scale of 0 to 1.

Linear Regression Now that all data preparation is completed, I applied Linear Regression model to the Capital Bike dataset. Linear Regression is most basic algorithm and it is often used as benchmark for performance of regression accuracy.
Here, I used scikit-learn package to execute Linear-Regression() algorithm. I fitted training data so that algorithm learns model parameters and then I used predict() to predict the Duration for test data.
I used RMSE (Root Mean Squared Error) as accuracy matrix. Using this matrix, we can check how our predictions are closer to actual values. We can see that RMSE is 383.41.

Random Forest Regression It is always good practice to deploy multiple algorithms on same datasets and the compare their accuracy to come up with best suitable method of prediction. So I tried Random Forest Regression algorithm as challenger to Linear Regression. This is a tree based method and it is generally considered as good alternative for regression analysis.
For this algorithm also, I use scikit-learn package [10] to execute RandomForestRegressor() method. I fitted the training data to this model and then applied the trained model to predict Duration for test dataset. I again used RMSE as accuracy matrix and this time we can see that accuracy of the model is slightly improved to 365.18.

Comparison of models Hence based on above exercise of applying regression techniques, we see that **Random Forest Regression is performing better than Linear Regression** in terms of accuracy. So we can deploy this model for all future data of Capital Bike rides.

Data Visualization
A good visualization helps u to get meaningful insights from the data. I decided to use Tableau Desktop [11] as my tool to generate visualizations. My main aim was to see bike ride patterns based on Start Date and Member Type. By generic common sense we can guess some patterns like bikes rides will be in demand on weekends, near downtown, at office time etc. It was necessary to check those assumptions with data and match it with the facts. It would also make sense to see if there are any surprises or anti-patterns which are counter intuitive.
I created Following visualizations for Capital Bike rides dataset-

Bike Rides Timeline

- This is the first graph to get generic feel of the data. Here I have plotted daily number of rides. We can see a lot of fluctuations over the bike demand which is justified by dynamic nature of the Bike share business.
- We cannot see any specific pattern about usage of bike hare service. I would call it a 50000 feet view of the dataset. It indicates that we need to have a closer and deeper look at the data.
- There could be sudden high demand on some days and similarly there could be slowdown. This makes perfect business sense to develop a prediction model which can optimize company's resources and thereby maximizing profits.

Bike Rides Month-wise

- If we look at month-wise distribution of bike rides, we can see that bike sharing is on high demand in summer and fall seasons. Winter has the lowest demand. It is logical considering weather conditions of Washington DC.
- We can see that high demand is sustained throughout the months of July, August, September and October. This is a good thing and it indicates that Capital Bike share is a popular service and people love it.
- This graph tells us that if company wants to maximize its profits, it has to focus on nice weather time of the year. Company might consider for a promotion campaign based on this input.

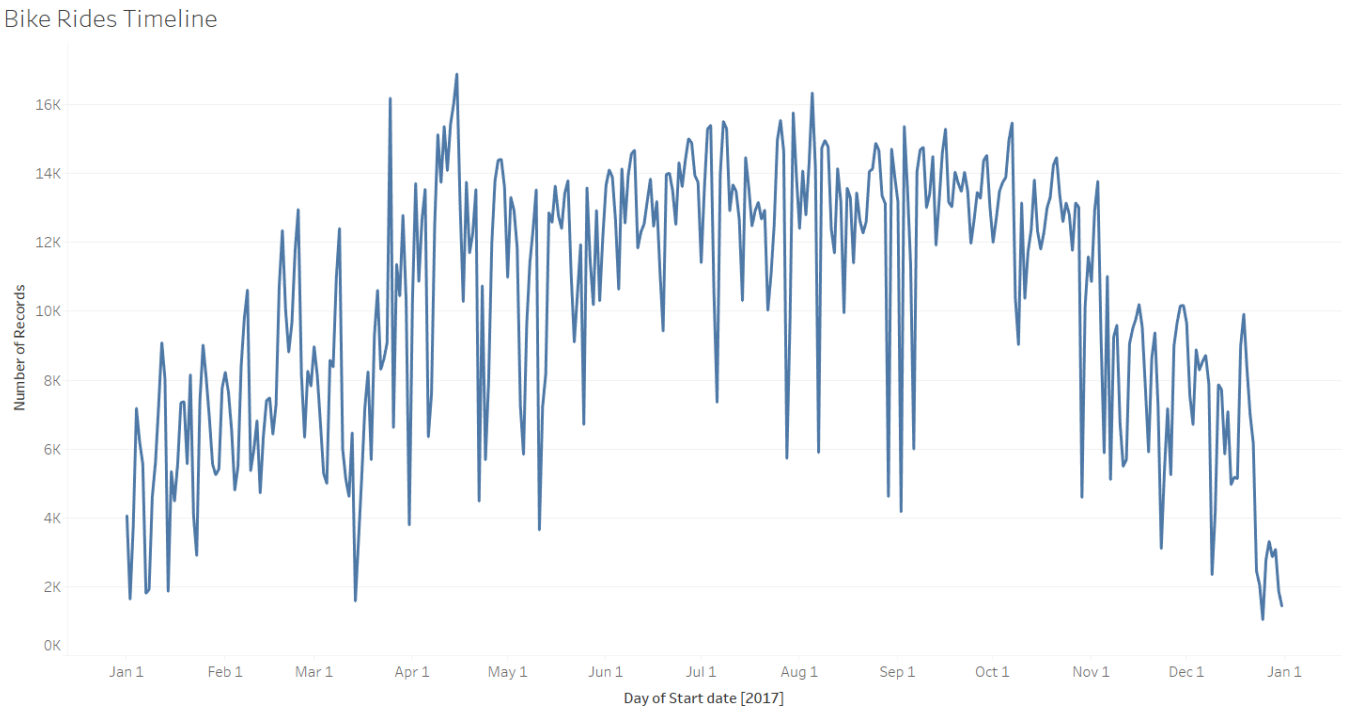

FIG. 6. Bike Rides Timeline

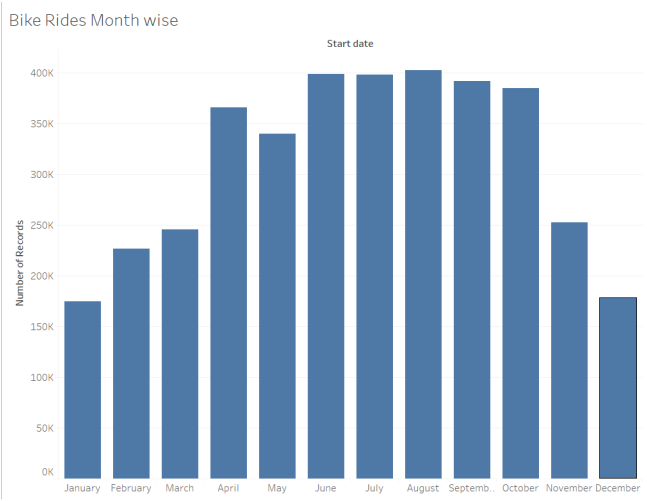

FIG. 7. Bike Rides Month-wise

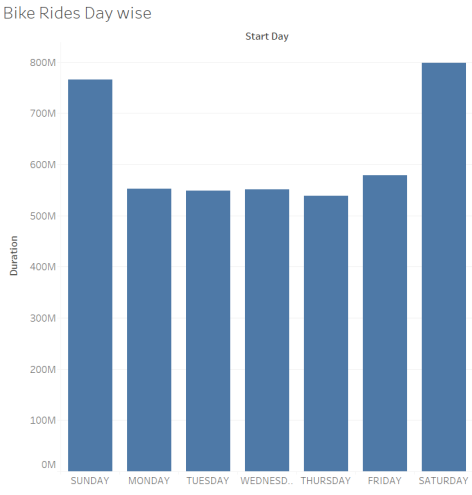

FIG. 8. Bike Rides Day-wise

Bike Rides Day-wise

- Day-wise graph is on the lines of expectations that Saturday and Sunday are most busy days for Bike sharing stations. People like to get on bikes and roam around the city on a leisure day.
- We also notice that demand over the weekdays is table and people routinely use the bike share service. This is a good sign of loyal customers necessary for a stable business model.
- Friday seems to be picking up in demand. Probably people get early from office and go for outdoors. This can be a valuable customer behavior input for marketing analysis.

Bike Rides Hour-wise

- Hour-wise rides count i no surprise at all. We can see that bike shares are at peak on 5 pm which is office time. People take bikes to go home and avoid traffic hassles. In this case company has to ensure

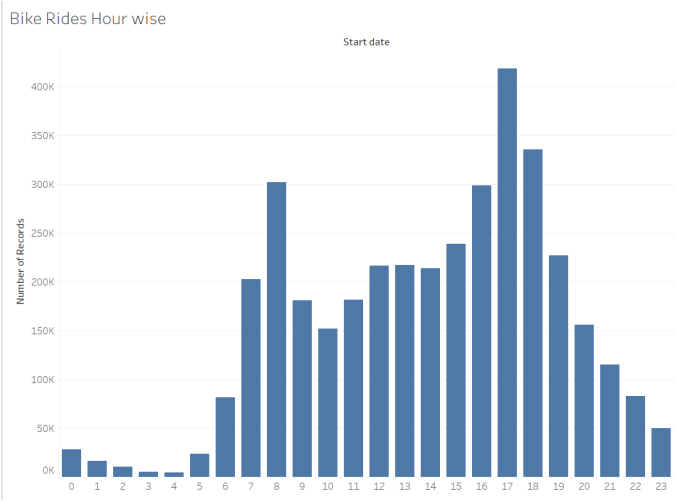

FIG. 9. Bike Rides Hour-wise

enough bikes at this time in Business districts.

- Night and late night hours are not much for business of bike sharing. This is idle time due to lack of customer activity. However company can make some profit through late night charges on customers.

- This distribution curve looks more or less similar to Poisson's distribution. It is an important input for Predictive Analysis.

Top Start Station Member Type wise

- This graph is the most important of all visualizations for this dataset. All above graphs are very much as expected that there not much insight other than common knowledge.

- This graph show a peculiar Power law distribution which means very few start stations are heavily used by users and there are many stations towards the tail which are rarely used or not even used. This reminds us of 80:20 principal ie. 20% stations are making 80% business activity.

- Columbus Circle/ Union stations seems to be most popular Bike sharing start station followed by Lincoln Memorial.

- We can also see a great divide on usage patterns of casual user and members. While members are making most number of rides, casual user are taking more duration for rides.

Critical Insight

Most probably, bike sharing service is popular among tourists which are casual users and they use it more on weekends.
Members, on the other hand, are office going people who use it on weekdays for daily commute.

VI. DISCUSSION

In this project, we saw one of the way to implement data-centric business strategy for Capital Bikes dataset. It is necessary to evaluate the outcome of this process-whether this strategy worked, how far it is effective, can it be extended further, what went right and what went wrong. This kind of holistic analysis gives us an opportunity to appreciate current process and improve it further.

Interpretation of Results

Capital Bikes has done a great job to make their data publicly available which enables us to deploy this project. The dataset is quite huge and it is certainly to the tune of BigData hence it may become the central strategic piece for business growth by optimizing resources.
The Data Lifecycle and Data Pipeline devised worked well in tandem and it allowed us to perform all operations in a streamlined manner. Here, each previous step worked as base for building next block on top of it. All these steps are progressing towards common goal which was drive by business needs. The business goal of optimizing bike availability for peak hours percolated to design thinking and all analysis done with that focus.
This was quite an iterative process throughout the pipeline. First time execution of all steps revealed shortcomings and necessities which fuelled next iteration. For example, first time data exploration showed that Start time is a critical field and useful features can be derived from it. So in next iteration data transformation step, those features were generated. Similarly first time regression analysis revealed that Duration field has lot of outliers which are result of Capital Bikes business model. Now we can not change the business model, but we can certainly omit those outliers for better prediction accuracy. So that change was incorporated in next iteration.

What went well

- Technological setup was a basic step of enabling this project. Installations of various applications like MongoDB, Anaconda, Tableau Desktop went fine and also got an opportunity to fine tune their parameters to operate with large datasets.

- I liked the fact that the dataset and the predictions were pretty much as anticipated. We arrived at logical outcome and visualizations. There were no surprises in terms of behavior of bike ride users. The data was not much messy, thanks to Capital Bikes.

- Another thing to notice here is that we are using many applications developed by different vendors like Google OpenRefine, MongoDB, Tableau, Python. However they all work cohesively on the dataset and we do not see any inconsistencies or

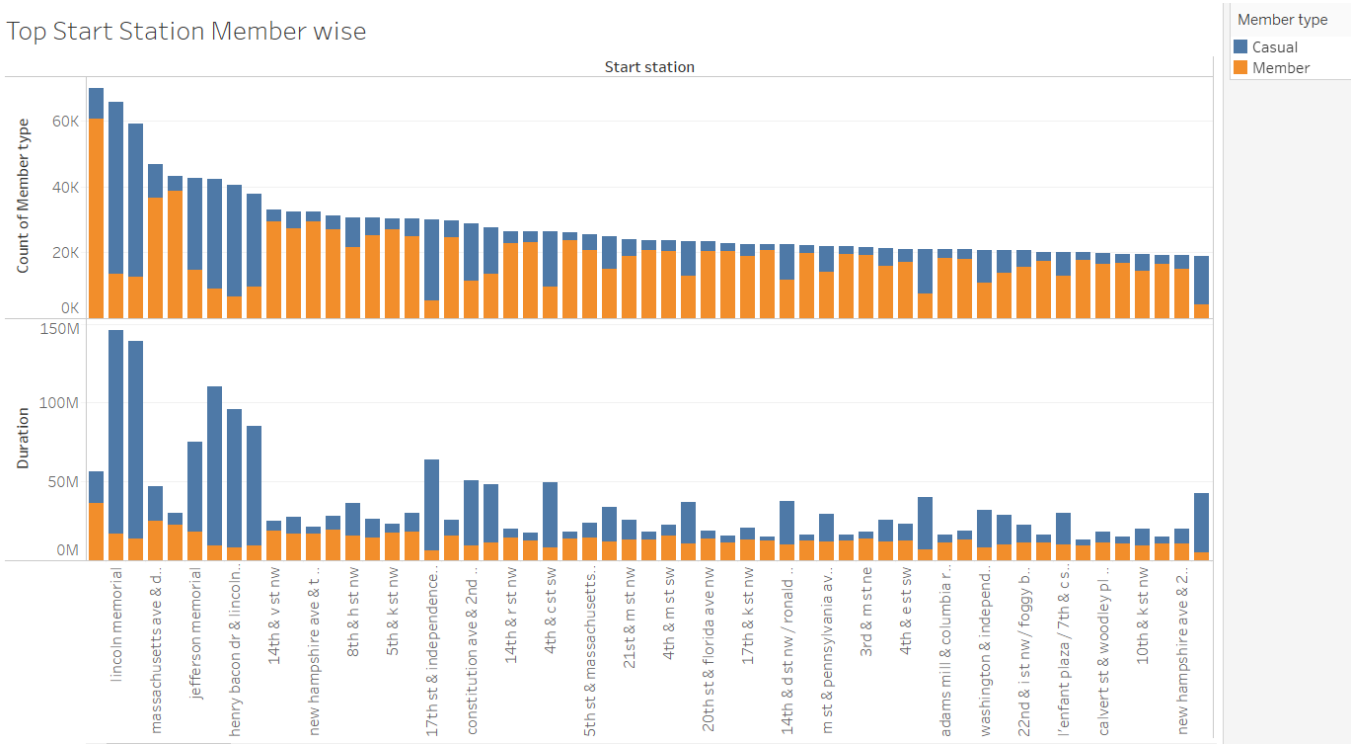

FIG. 10. Top Start Station Member Type wise

limitations due to different tools we are working with.

- We can see that with this structured top-down approach, all pieces fall in place and we could arrive at final result of prediction. We deployed two different regression models and compared their outputs. Linear Regression has RMSE of 383,41 as against Random Forest RMSE of 365.18. **In Functional terms, it means that we can predict bike ride time with +/- 5 min accuracy.** This will definite help business to ensure that bike stations have bike availability all the time with respect to demand.

Challenges/Failures

I come across some limitations while performing some of the actions.

- I could not leverage Jetstream VM provided by Indiana University because it did not have all python libraries required eg. sklearn. It did not have applications required like Tableau, OpenRefine. This put limitation on my capacity to process data as I had to rely on my laptop. I could process only 2017 data for Bike rides. I could have considered all the data available if I could leverage distributed environment.

- There is limitation while using opensource tools like OpenRefine, Tableau Desktop, MongoDB Compass. They are not built for scale and some of the features are available only in Enterprise edition. However, It is of great learning to get experience of these tools and see how they work in sync.

Connection with Course

This project was a reflection of what we have learnt so far in the semester for I535 Big Data and Access. I got to apply all the concepts that we have learnt theoretically. It helped to see how technology works on the ground. I was able to apply concepts of Big Data driven Business Strategies, Data Lifecycle and Data Pipeline, Tidy Data, Selection of NoSQL Database and Predictive Analytics etc. in this project.

VII. NEXT STEPS

This project work can be further extended in many ways.

- **More Overlapping data** We can use weather data, public holiday data and local events data to predict bike sharing period more accurately because these things are directly affecting user demand and it will certainly help to leverage this additional data.

- **Data Pipeline Automation** Since the prediction is a repetitive process for business, we have to keep running this pipeline periodically and diligently. It

makes a great sense to automate this pipeline using DevOps techniques. We can definitely deploy a tool like Apache Jenkins which will ensure timely execution when there is new data or when there is code change.

- **Real Time Data Processing** We have downloaded this data from Capital Bike hare in zip format which is a data dump of historical events. The company also provides real time data which can be ingested and we can derive real time predictions. We have to deploy Apache Spark Streaming for such real time processing of data.

VIII. CONCLUSION

Every business has to draw its own unique data management plan to suit its need and goals. In this project we saw the implementation of complete end-to-end Big Data strategy catering to Business Goals. We studied the Business Model and challenges faced by Business which forces to look out for feasible solutions satisfying Business goal. In this case, challenge was to ensure bike availability at all stations based on dynamic customer demand. Then we saw how Big Data comes into play to realize potential solutions. Real Time data collection from each station for each ride enables us to analyze customer behavior and respond to it appropriately. Sine this is large amount of continuous data, it is crucial to adopt an efficient Data Management Plan to collect, store and consume data seamlessly. Selection of suitable NoSQL database is key aspect of this strategy. Similarly, Data Transformation steps are foundational to enrich data and making it ready for analytics. All this preparation leads us to actual predictive analysis which has business value. Bike rent time predictions in this project make business capable of forward planning. Finally all the steps are monitored via a dashboard visualization.

ACKNOWLEDGMENTS

I would like to thank Prof. Inna Kouper (Indiana University, Bloomington) for providing her valuable Suggestions and Guidance for this project and throughout the course of I535 Management, Access and Use of Big and Complex Data.

[1] https://www.capitalbikeshare.com/ (2019).
[2] https://www.capitalbikeshare.com/how-it-works (2019).
[3] D. N. Matthew J. Mazzei, Big data dreams: A framework for corporate strategy, in *ScienceDirect*, Vol. 60 (Business Horizons, 2017) p. 405—414.
[4] https://www.capitalbikeshare.com/system-data (2019).
[5] H. Wickham, Tidy data, in *Journal of Statistical Software*, II, Vol. VV (American Statistical Association).
[6] Usgs data management,u.s. geological survey (2018).
[7] N. U. Katharina Ebner, Thilo Bühnen, Think big with big data: Identifying suitable big data strategies in corporate environments, in *Hawaii International Conference on System Science*, Vol. 47 (2014).
[8] I. David Huynh, Metaweb Technologies, https://github.com/openrefine/openrefine/wiki (2012).
[9] https://docs.mongodb.com/manual/introduction/ (2019).
[10] https://scikitlearn.org/stable/supervisedlearning.html (2019).
[11] https://tableau.com/current/pro/desktop/enus/mongodb.html (2019).
[12] J. H. Saltzerand M. F. Kaashoek, *Principles of Computer System Design: An Introduction* (Morgan Kaufmann, 2009).