MSBD5003 Project Proposal

Group 16

20787575 FENG Geqin, 20413514 OU Huiyi, 20784183 Zhu Zhenyi, 20783842 YANG Yuke

I. Task

Rental of bikes is ubiquitous in most modern cities and we would like to analyze factors related to demands for bikes, specifically in Seoul. We would apply different big data technologies and try to predict the bike count needed at each hour and ensure enough supply of rental bikes.

II. Dataset

Our dataset is available at UCI website and the original source is from South Korea public holidays website. The dataset contains count of public bikes rented per hour each day during the time period from December 2017 to November 2018 in Seoul Bike sharing System, the corresponding weather data and holidays information. In detail, weather information includes eight numerical attributes, which are Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Rainfall and Snowfall. In addition, the dataset also contains several nominal attributes, such as the season, holiday and functional day information.

III. Technologies Used

We plan to use several technologies in data storage and connector, data preprocessing, visualization and modeling parts.

- **Data Storage**

For data storage, we decide to store the data in Microsoft Azure cloud database for its convenience, flexibility and affordability. On one hand, cloud databases offer nearly unlimited space compared to physical hard disks on our own computers. On the other hand, there are existing connectors to connect to cloud databases, which paves the way in extracting data from the cloud.

Microsoft Azure offers different choices of databases, including SQL, NoSQL and data warehouse. Since our data is in a structured format, we will choose the SQL-like databases. Specifically, Microsoft provides three kinds of Azure SQL databases, namely SQL server, PostgreSQL and MySQL. PostgreSQL is a feature-rich and object-relational database that supports array, json, graph-like structure and can handle complex queries and massive databases, while MySQL is a simpler relational database that is fast, reliable and easy to manage. Given the fact that the bike demand data we use is relatively simple and there are

no complicated data types involved, except for integer, float and string types, we decide to use the MySQL database.

- **Data Connector**

Connectors are used when we need to read data from the database and load into the Spark environment in localhost as Dataframe. Both JDBC driver and Apache Spark Connector could enable SQL databases on cloud as input data sources. Compared to the built-in JDBC connector, the Apache Spark connector provides the ability to bulk insert data into SQL databases with faster performance. But it is an open-source project without any Microsoft support, which may result in some unknown issues. Therefore, we decide to try both connectors first and see which one works better.

- **Data Processor**

For the data processor, we decided to use spark RDDs or Dataframe, depending on the exact data type after loading the data from cloud databases using data connectors. Similar to RDD, DataFrame is also a distributed data container. However, DataFrame is more like a two-dimensional table of a traditional database. In addition to the data, it also records the schema. Since our dataset has the requirement of schema, spark Dataframe is better in this case. During the data processing, we may need to handle issues such as missing values, type conversion and reformatting after querying the data. These can be solved by the transformation and action operations in the RDDs and DataFrame. The future decision of the choice between these two techniques will be made when we go through this process.

- **Visualization**

For the visualization of the output, Matplotlib, Plotly and Seaborn will be used to graph our observation during the exploratory data analysis process. For example, plotting the relationship between the weather and the demand for bikes could help us perceive the predictive power of the weather factor.

- **Algorithm**

The Spark's MLLib library will be used to model the relationship between environmental factors and the demand for bikes in Seoul. The library provides various available algorithms including generalized linear regression, decision trees, etc. Since our task is a regression problem, we will use the regression and tree-based algorithms in MLLib to build the prediction model.