Group 8: Xiaoxiang Zhang (xz2631), Minghao Li (ml4025), Ruimin Zhao (rz2390)

a. Title: Delay Prediction based on the data of weather and New York Subway

b. Data to be used: OpenWeatherMap (weather dataset), MTA (transportation dataset)

c. What will you do? Application/Techniques/Systems you want to use

Generally speaking, based on historical weather and traffic information, we want to predict how long the delay will be for some future weather and time(9am,10am,etc) inputs. Then, if time permits, we will recommend optimized route based on the predicted time delay(use GOOGLE map API to get the several routes and our models to predict which route takes least time).

1. Firstly, we are going to parse real time weather and traffic data from Internet. The result can be tuples which will be transformed to spark-streaming system for preprocessing and storage.
2. Then we will apply Kernel Regression algorithm to our data to get a model evaluating the relationship between parameters. Another potential model LSTM is to predict current delay using several previous delays. Next we could combine the 2 results using weights.
3. Finally, we are going to predict the traffic delayed time based on the real-time weather plus previous delayed time and offline trained model, as well as the optimized route. All the results will be presented through a web application.

Techniques: Sparking Streaming, Kernel Regression, LSTM, Flask, Google Map APIs, MTA API

d. How will you show any results?

1. Hopefully, a web application will be presented in the end. Users can see the predicted delay time of subway stations. Besides, user can obtain our recommended public transportation route based on the predicted delay.
2. Data visualization using packages such as Pandas based on the geographic information obtained from MTA data source.

e. What do you think is novel about your approach?

1. Rather than only focusing on one data source, we would utilize two different streaming data sources (MTA public transportation data and weather data) and examine the correlation between them. In this way, hopefully, we are able to obtain more comprehensive and interesting data analysis results such as transportation delay prediction, transportation route recommendation etc.
2. The main purpose of our project is to predict the delay situation of New York subway system. We noticed that we can only get some vague information from MTA website, such as which line is delayed or which line is in good service. However, people have no idea about the exact delay time. This is the point we will focus on.

f. Any references: [1] http://alert.mta.info/ [2] https://www.dot.ny.gov/wta

Following are some technical details to clarify our proposal.

1. Can you please include one concrete example of each:
MTA Data that you plan to use
NOAA Weather data that you plan to use

We plan to use weather data from openWeather instead of NOAA because we can get full weather history in 5 years easily. This dataset is used for training based on some extracted attributes, which we will discuss below. Plus real-time weather information as for online streaming prediction.

Besides, we decided to use MTA dataset to predict the delay time of subway system in New York.

2. How will you collect data for training the model? Will this be historical data for some period of time?

As mentioned above, we will collect weather data through openWeather API for both historical data during the past 5 years. These data are for the training purpose. The label is the delayed time under each weather condition.

3. Do you know what types of data cleaning you need to do, for each data stream, and how you will put it all together?

Data parsing and extraction, filtering for some obviously wrong data, quantization for some character-like feathers, feature scaling using MinMaxScaler.

For the MTA data cleaning, we are stilling working on it and trying to utilize the available data resource as much as possible so as to provide reasonable datasets for later training.

For the problem about how to put it together, we plan to predict the delayed time respectively based on the weather and MTA data, then combine the two times using some merge weight, if the two times have same timestamps we created when preprocessing the streams.

4. What attributes of the weather data will you use? Temperature, precipitation? Will you use the weather forecast?

Temperature, pressure, humidity, month, date, hour, wind speed, wind direction, cloud, weather code. These are all possible features we plan to use but not necessary. We will have some validation and feature selection when training the model for final confirm.

We will use the weather forecast for delayed time prediction. Because our goal is to combine weather information and previous delay for current delay on one specific route.

5. For the MTA data will you look at trains/buses/both?

Subways only.

6. How will you use the Google Maps API? Can you be more specific? How will you get suggested routes?
(If time allows) By setting google.maps.TravelMode as TRANSIT, we can get several feasible routes from our starting point to destination. In this way, we can get the transit point in these routes as well. Based on this information and our predictions on delay time, we can derive a path which will spend the minimum time.

7. For the prediction, will you use also the current delay as reported by MTA, or just weather data?

For prediction purpose, we won't use the current delay, but also some previous delays and current weather.

8. How will you evaluate the performance of Kernel regression?

MSE: we plan to seperate the dataset into training set and testing set. Our goal is to make the two MSE as low and as close to each other, in order to make the model neither overfitting nor underfitting.

9. Can you say something about the rate of the streaming data you will process?

Weather: at least 1 message/2 hr
Transportation: 1 message/30 sec

Schedule Update:
As discussed in the phone call, the data processing part is challenging since there is no direct resource of historical delay. We have to calculate the historical delay by ourselves using the schedule timetable and the historical data downloaded via MTA API. A lot of assumptions have to be made during this process. Also, the weather data require processing. We assume this data processing stage will take quite a lot time (We will try to limit it within 1.5 weeks).
After that, we would do the data analysis (model training and performance Testing). Since we might also have to learn how to use Kafka and Spark Streaming for the data streaming step, we assume this data analysis stage would also take a lot of time (We hope to complete it within another 1.5 weeks).
As a result, we might not be able to actually provide a web application in the end if the data processing and data analysis do not go very smoothly. However, we would try our best to eventually provide some visualized results to demonstrate our project.