**NAYA College**

**Data Engineering Course Project**

**Israel Public Transportation**

**Detailed Design**

Guy Levinhr

Meirav Mazala

December 2020

Contents:

Scope and Background …………………….3-4

Architecture Requirements Overview……...5

Technological Components Overview……..6-12

Architecture second option………………...13

Sizing and Costing…………………………..14

Results and conclusions…………………….15

Appendix A- Data Files………………………16-17

Scope and Background

Lately the Israeli public transportation, opened Service Interface for Real time Information,

Since it’s a new platform and since the Israeli public transportation doesn’t have the analysis from the data and really hope to get it from developers, we decided to take their data As our data platform in this project.

The following SIRI <https://www.gov.il/he/departments/general/real_time_information_siri>

Allow Developers to receive the following data:

* Plan for the next week transportation arrival- Date, time and location- this data saved every night as CSV files and updated for the next week.
* Actual Real time Data- Current location of the vehicles and their attributes (ID, Line, code..)

Based on this data, we can compare the plan vrs the actual and get material conclusions about the delays in the transportation agencies.

The outcome of this project will help to improve the transportation services given by the Israeli public transportation unit and will be a trigger to the suppliers to work according to the agreements.

Today there are basic measures in the agreements but no enforcement. For example, page 3 "רמת שירות מינימאלית":

<https://www.gov.il/BlobFolder/guide/agreements_for_operating_service_lines_in_public_transportation/he/PetahTikvaRoshHaayinagreement.pdf> .

Lately, the Israeli public transportation started to install sensors on the bus doors in order to collect the quantity of travelers on the bus. This part just started- the installations already done 30% in the center and its not open to use (as they allowed with the real time location).

Having the results of this project with the new layer of data in the future will be even more material for the conclusions.

The Project technical requirements are:

1. To create a multi-purpose data pipeline system which will acquire data from the real API
2. To Create a platform for performing data filtering, data cleansing, data verification and data enrichment to perform analytical operations for the organizational and business purposes and to transform raw data into knowledgeable data.

The API and the additional files that are given include all the agencies and lines.

In this project, we will focus only on two Bus lines.

Line number 171 and line number 88 in Gush Dan.

Architecture Requirements Overview

Night task- once a day- forecast files:

1. Workflow management platform:

* Download the csv files
* Delete the previous files (saved on the night before) and save the new ones.
* Unzip the files
* save the relevant files and delete the ones are not relevant
* bring all the CSV files to DataFrame
* filter the data by document ZONE in order to get:

Dan and Eged- only

From Tel Aviv

All this data will go to DB- planed time schedules and path

* create one DataFrame to consolidate all data on each trip

1. save the data as Json on a database

Real Time Data:

1. Connect the Israeli public transportation API, collect and do the ETL processes real time:
   1. Connect the API and bring the plans for the chosen lines
   2. Send the data to a pipeline
2. Pipe line to collect, process and save the data on real time to database after filtering, formatting and cleansing

In this part, need to consolidate the data from the SIRI with the stream data we receive- this part will enrich the stream data with the plans and the additional data that is saved once a day.

In this part, the data will be organized- data, hour.

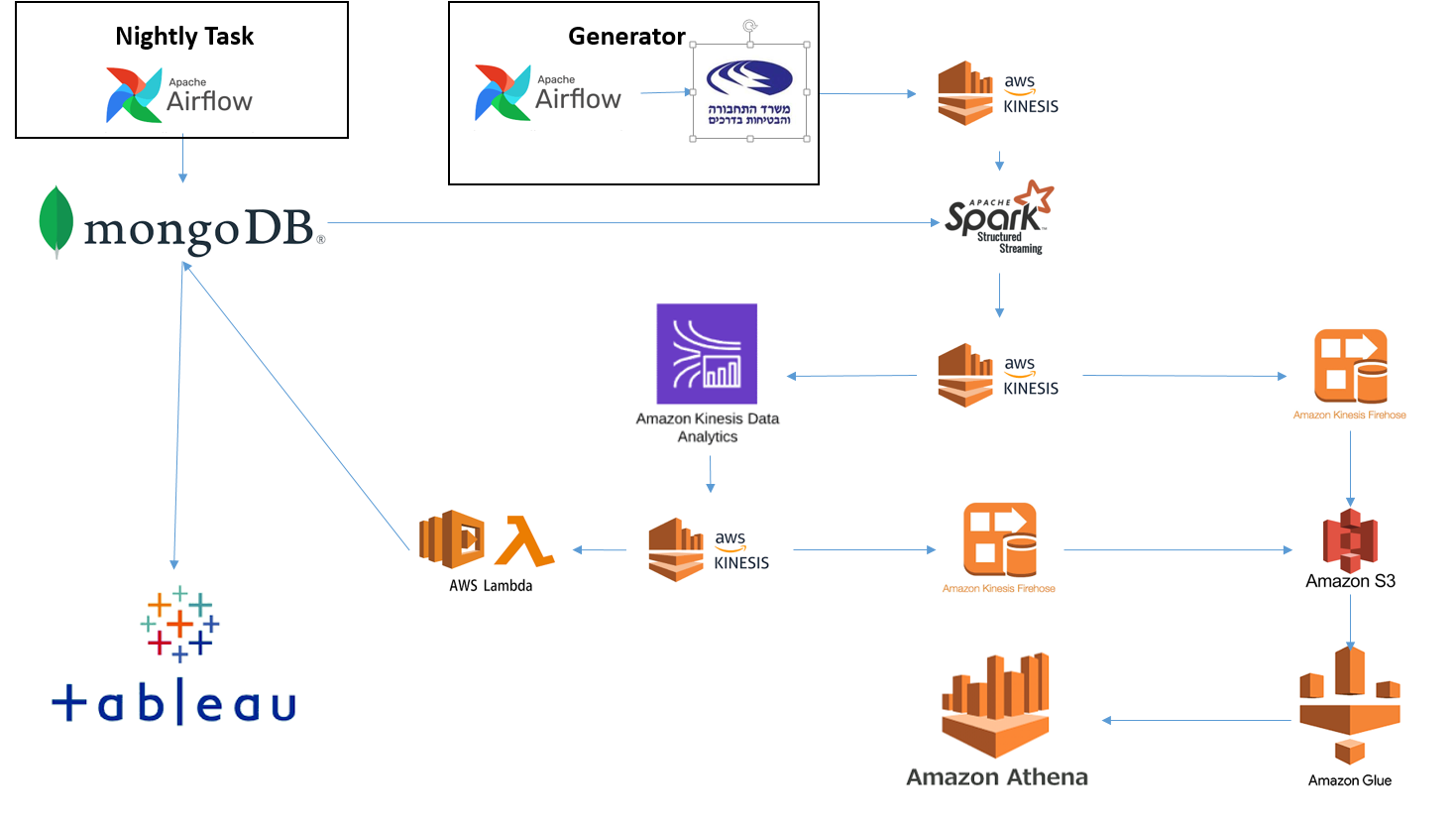
Very important role in this part, to find the closest bus stop station. This will allow knowing what is the distance from the closest bus stop station.

The enriched data will be sent for analysis and will be saved also in the archive for history analysis.

1. Data reporting, analysis and visualization

Technological Components Overview

Considering all the requirements, we chose the following components:



Night Task- forecast files:

In the night task, the Airflow will run the process to download the csv files and process them. Using Pandas to create the csv files into DataFrame.

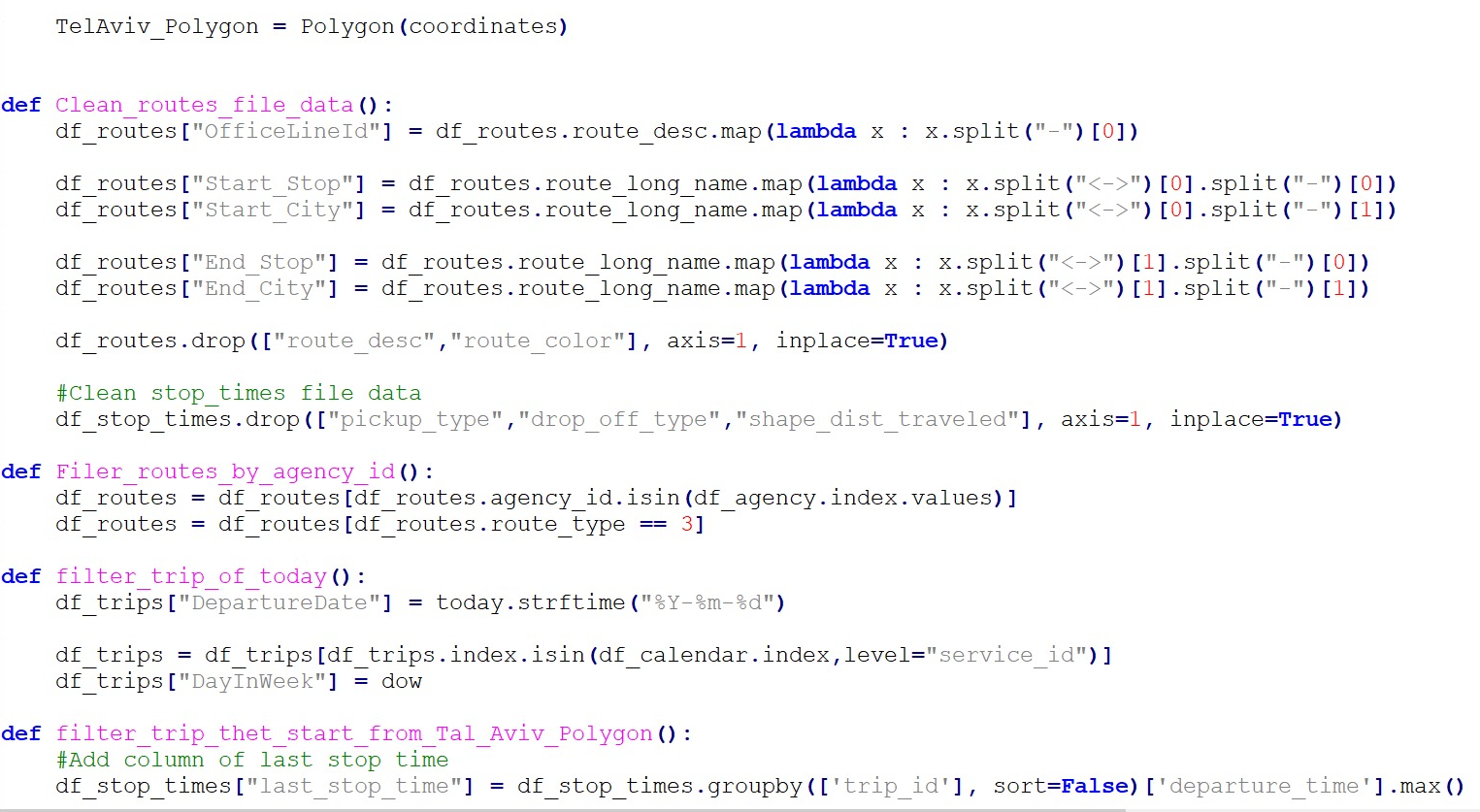
In the end of this part, the data saved as Json in MongoDB.

In this part, we filter the data to agencies Dan and Egged and Tel Aviv area.

Airflow- schedule and monitor workflows as Directed Acyclic Graphs (DAGs) of tasks.

Pandas- is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool.

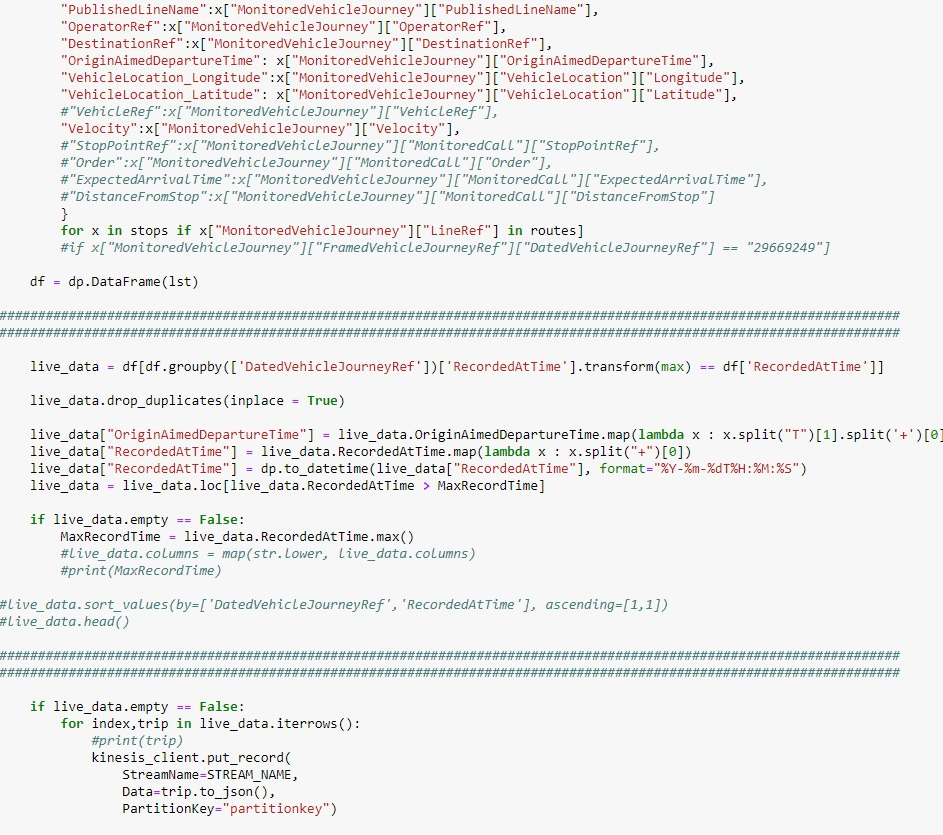
MongoDB- a cross-platform document-oriented database program.



Real time – Actual data process:

We used the Airflow in the following process for the API data:

Generator connect to the API (every 15 seconds), search for the relevant data on buses locations and the lines we choose (88 and 171), bring real time data, take from MongoDB the existing lines and their Trip number and send to Kinesis.



Spark- enrich the data from Mongo with kinesis data.

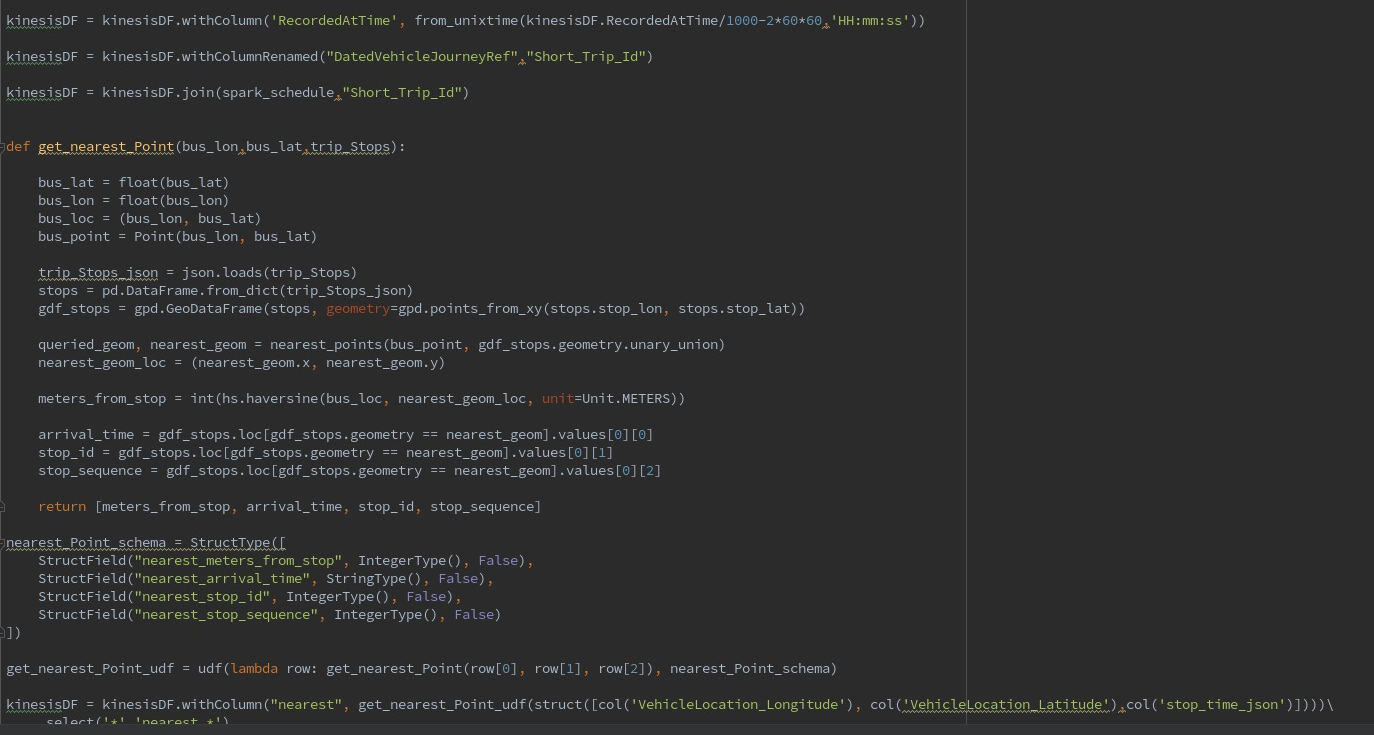
Connecting both and find the closest station near the bus and what is the distance (meter). This data goes to kinesis after it was adjusted to fit kinesis.

Kinesis answer the need in this part to Capture, process, and store data streams.

Also used Kinesis data firehose to save the data in partitions every hour.

Kinesis data analytics filter the stream data in S3. If the bus distance from the station is less than 50 meter, than the bus is in the station.

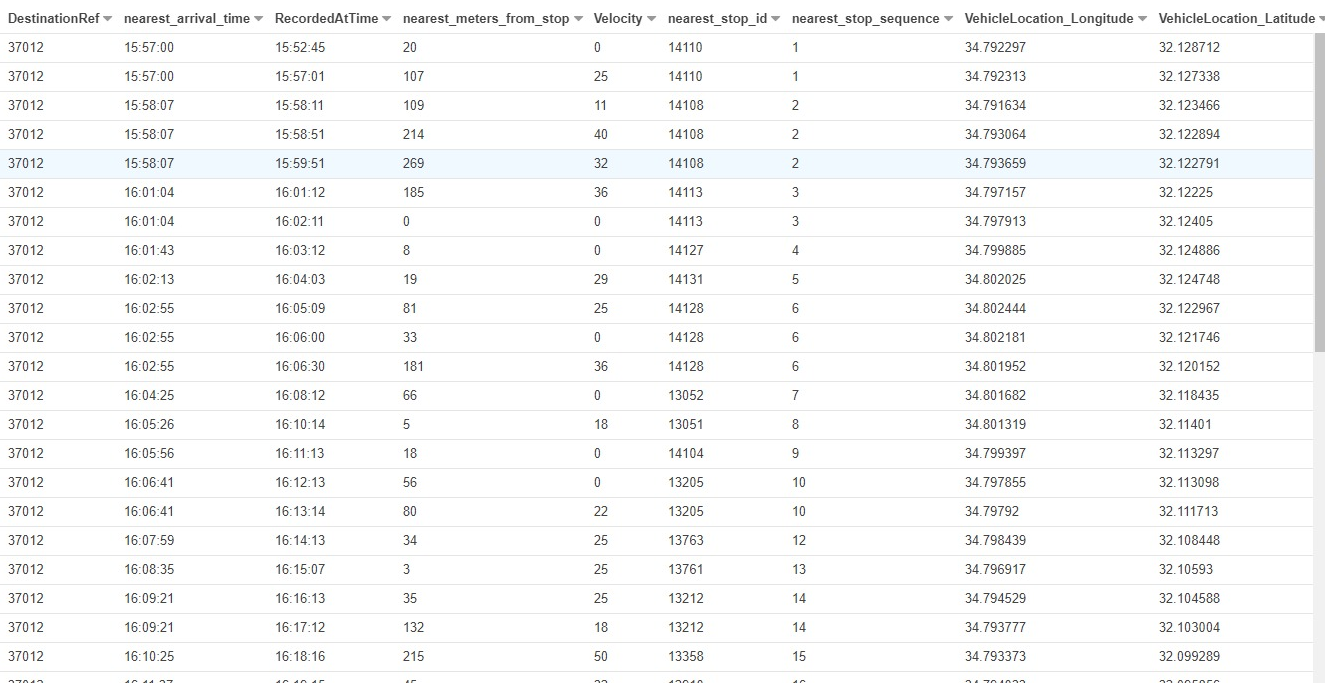
Used Lambda to take the data from kinesis to the Mongo



Aws glue

Crawler every hour to update the new partitions that were created. This part is done to allow Athena read from S3 when a table exist.

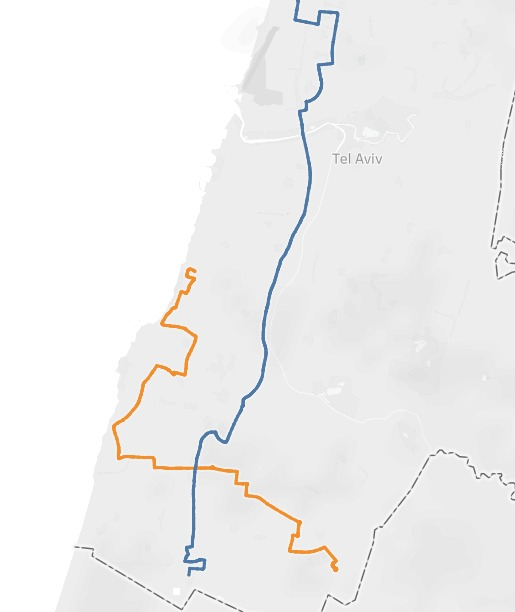
Athena- Execute reports and statistics to compare actual arrival time for each line and agency. Analyze the Data in S3.



In this table exported from Athena, we see:

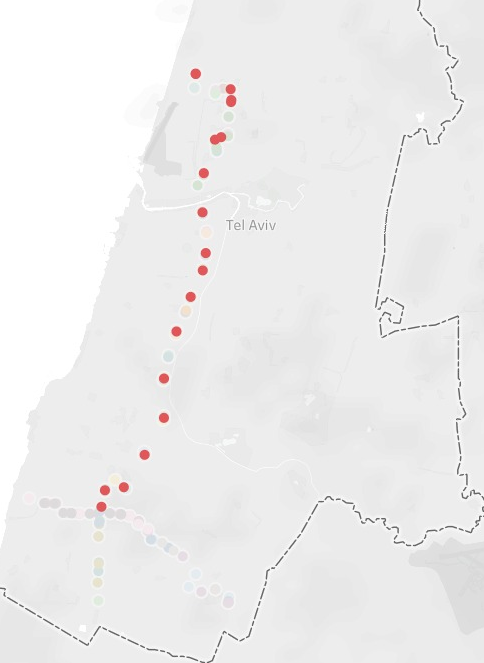
* Planed arrival time
* Actual arrival time to the station
* Distance from station
* Velocity
* Stop id
* Location on path

Tableau- Dashboards to present the data and maps of the paths and their data.



Blue- line 171

Orange- Line 88



Why these tools?

We chose SAAS- AWS from comfortable reasons, good performance, no need to maintain, not expensive.

We could have use Hadoop to get similar results (but in this case, it require maintenance).

Challenges and complication in this Architecture:

1. Connect the stream live data (real time data) with the static data.
2. In the stream data- find the closest station – used the Gio Location.

Architecture- Alternative solution

Alternative solution for these needs from the data, can be with Batch.

Take the data every minute or 10 minutes, instead of stream- real time data.

In this case, we can use Spark and Hadoop. Hive and HDFS for storage. Also Oozie for workflow scheduling.

The queries will be on a table in Hive.

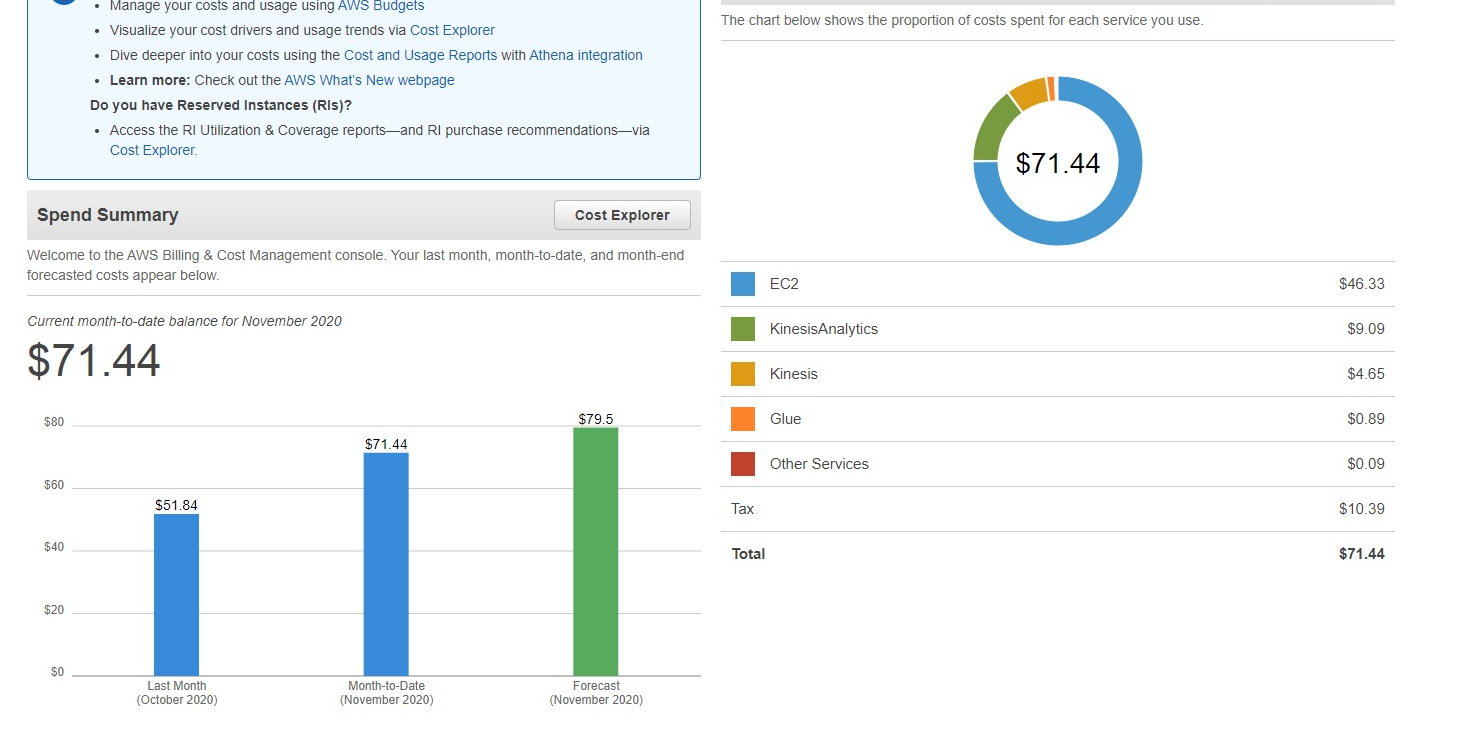
Sizing and Costing

The cost will be according to the storage we have on S3 and Mongo

Machine cost 1 Dollar per day 1 Giga, including BI connector its **6$** per day for two lines If it will run 24 hours.

The total cost will be around 180$ Dollar a Month for the two lines.

The cost will be depend on how many paths- lines we want to collect and save the data.



Results and conclusions:

The data received in Athena answer the following questions:

Timeline- Plan vrs Actual

How long it take to get to each station

Does the bus start its path on time.

What is the time between arrivals to the last stop and starting it as first stop (Do they have enough break time?)

Specific areas with many delays.

How many delays each day

How many time it was forecasted to take 3 minutes and it took 4.

Next step can be Alerts for delays or long time that no bus arrived to the station or even more than one bus in the same station.

Getting these answers will bring better forecasting- plans and better services as a result of better enforce for the agencies agreements.

“Turn your Data into Knowledge”

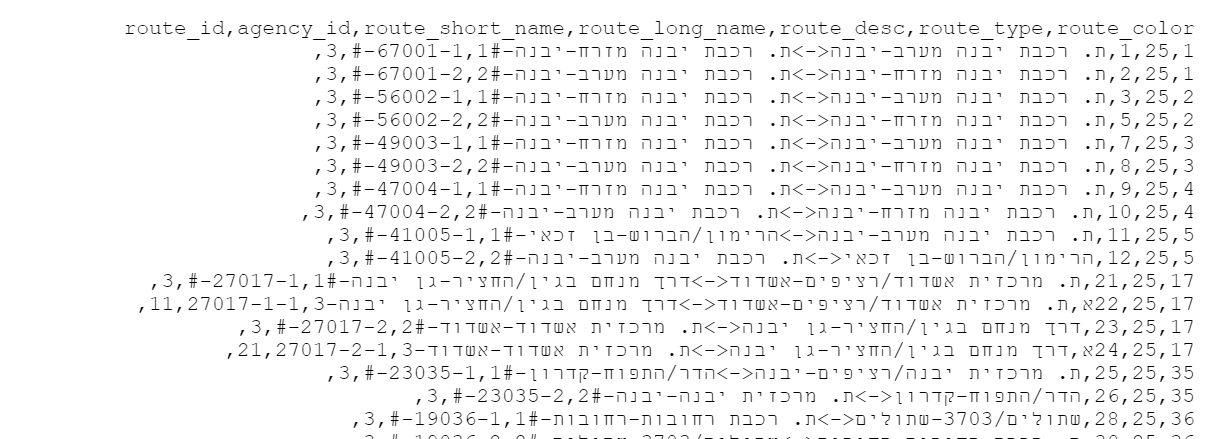
Now the data is ready to be analyzed and be the base for better conclusions and action items.

Appendix A- Data Files- Examples

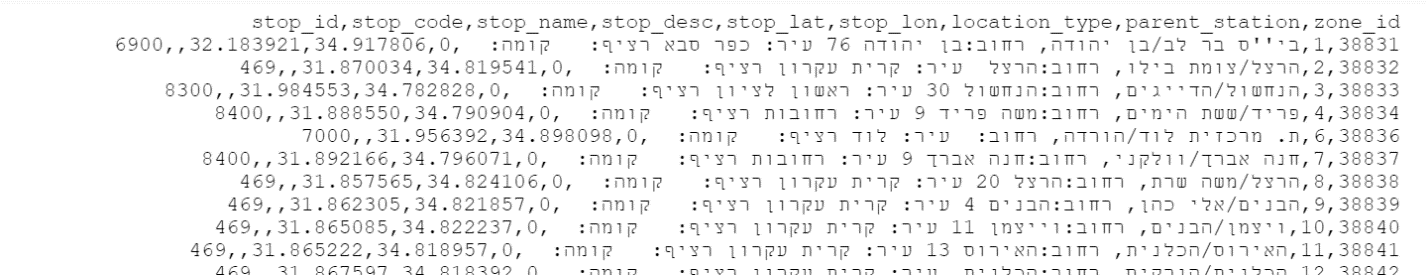
Agency-



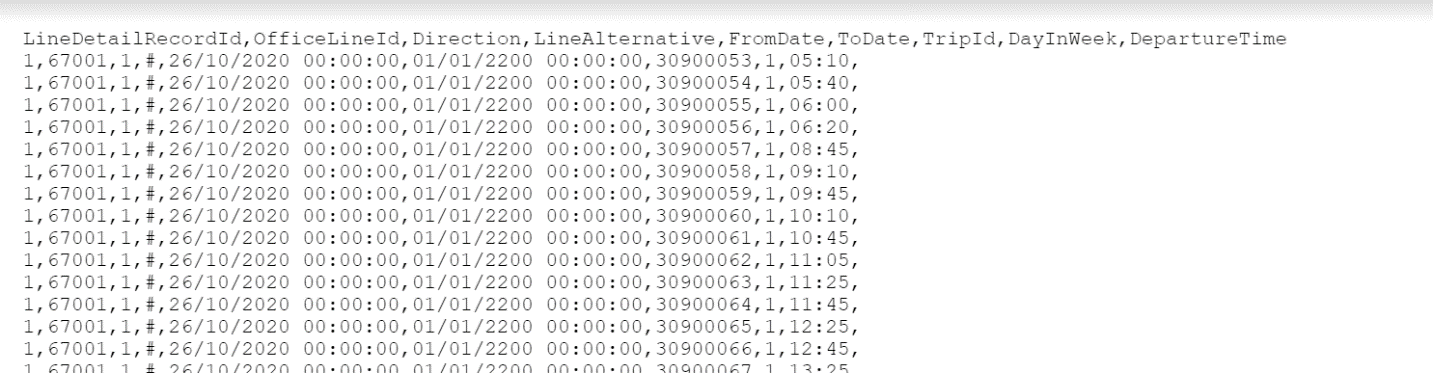
Routes-



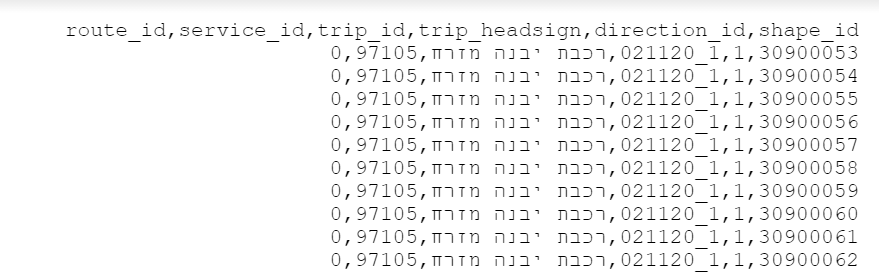
Stops-



Trip id to date



Trips-



Zones-

