Assignment 2

Charlie Mei cm3947

Scenario 1

I would use a data lake to store and manage the data. This is because:

- *Variety of data:* the data collected are of different formats and from different sources e.g. spatial data for car location, external data on current driving conditions, monitoring data from car sensors.
- *Volume of data:* there is a large volume of different formats of data, which require different ETL processes to integrate into a schema.
- *Velocity of data:* the data that is to be collected is at 15-minute intervals, as required to make regular adjustments to the business model or to the type of routine maintenance of the cars. This high velocity would be too cumbersome and slow in a data warehouse.
- *Applications needed*: a variety of applications would be needed to support this data lake. Data extraction, extraction and storage would require applications such as SQL, Hive and/or Spark. Meanwhile, data analysis would require additional visualization applications such as Microsoft BI, Tableau, R or Python.

Scenario 2

I would use a data warehouse to store and manage the data. This is because:

- *Variety of data:* the data collected is largely of the same format which can also be easily summarized at different levels of granularity through ETL processes.
- *Volume of data:* there is a large volume of data, however is of the same, unified and stable format.
- *Velocity of data:* the data is required only on an ad hoc basis, therefore does not require frequent modification of the schema.
- *Applications needed:* SQL and similar database management systems (mySQL, postgresSQL etc.) is sufficient to extract the necessary information for customers. To visualize the data, applications such as Microsoft BI, Tableau, R or Python could be used.