**Data Visualization with ELK**

**Problem Statement:-**

We as IT people or any business people may need to handle huge amount of raw data, obtained from different sources. But these raw data is useless, unless we have some mechanism to visualize what that dataset is all about and what kind of information we can obtain from it and based on the insight we extract the data for our business use.

**Solution:-**

This application that uses open ELk stack, handles any dataset provided in the form of CSV, which is be loaded into **Elasticsearch** with the help of **Logstash** and can be visualized with **Kibana,** by creating customized dashboard.

This is just a MVP, which can be enriched with more features in future.

**Demo Steps (Standalone mode):-**

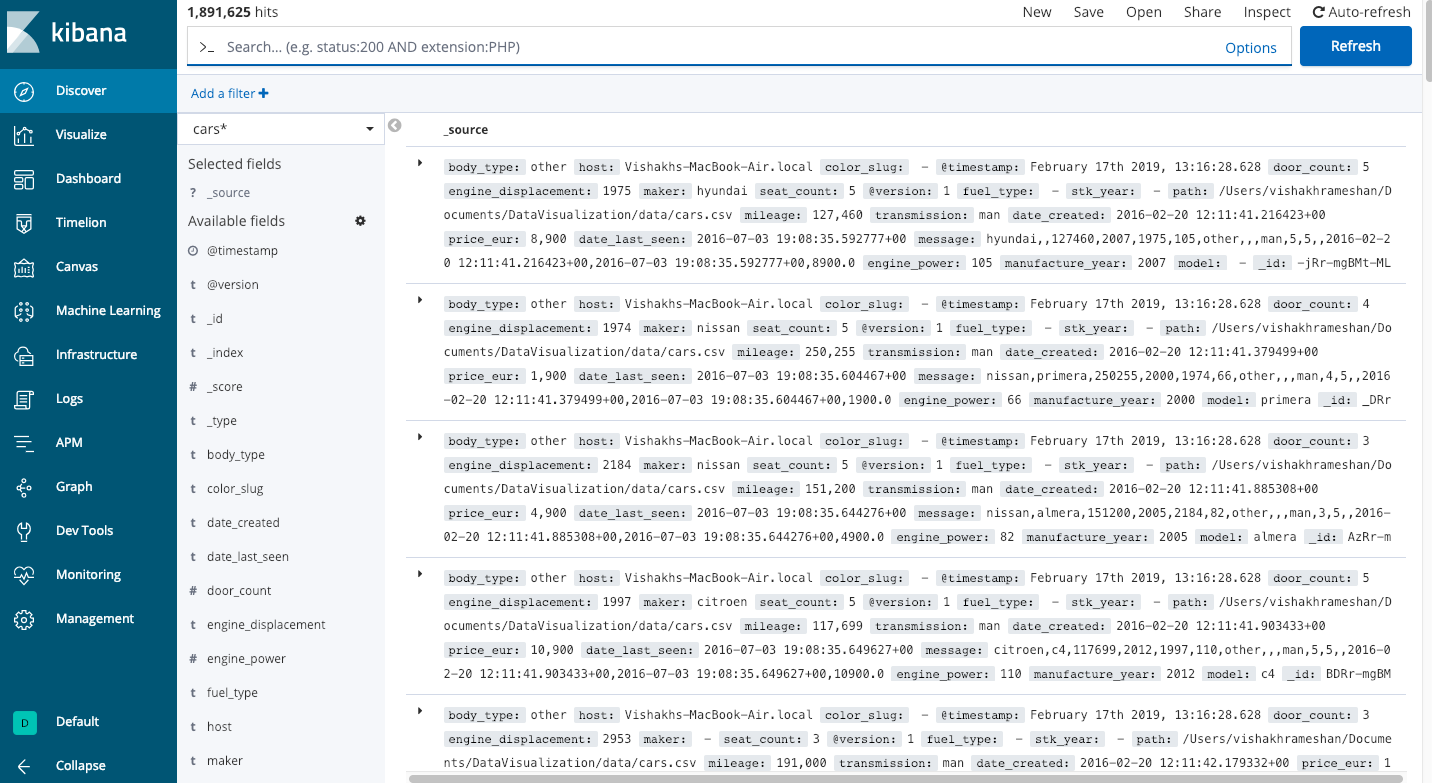
For demo purpose we have collected open dataset from <https://www.kaggle.com>. The dataset is the details of used cars for sale in Germany and Czech Republic since 2015. The details include:-

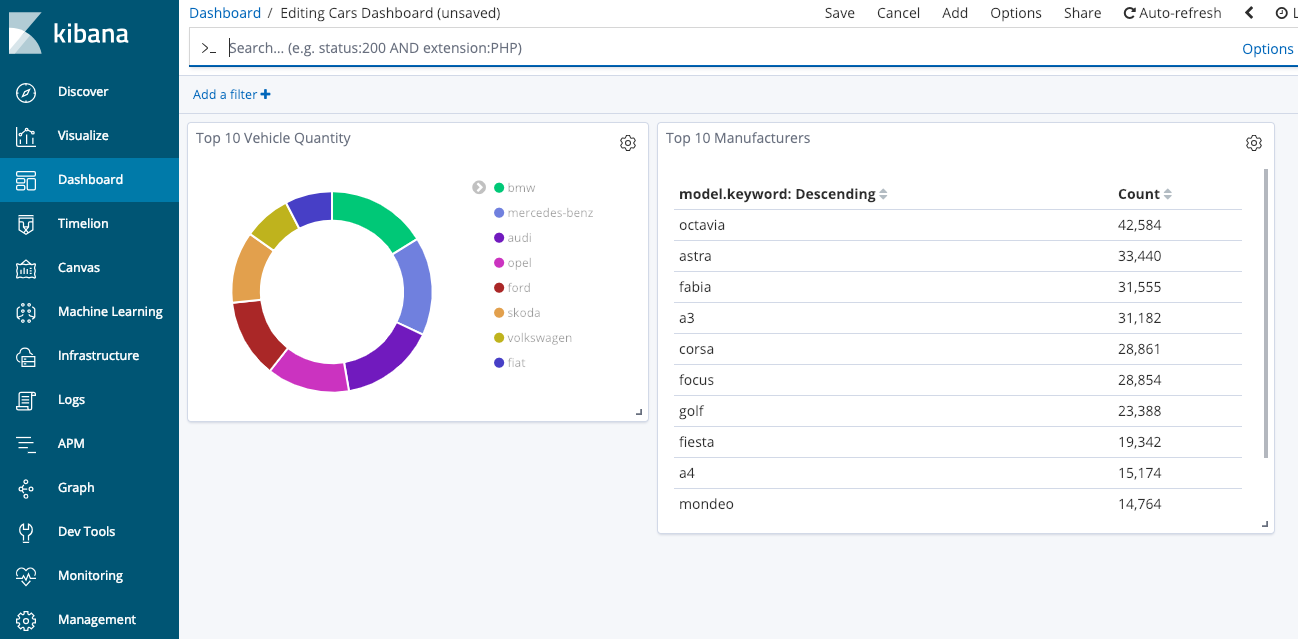
There are roughly 3.5 Million rows and the following columns:

1. maker - normalized all lowercase
2. model - normalized all lowercase
3. mileage - in KM
4. manufacture\_year
5. engine\_displacement - in ccm
6. engine\_power - in kW
7. body\_type - almost never present, but I scraped only personal cars, no motorcycles or utility vehicles
8. color\_slug - also almost never present
9. stk\_year - year of the last emission control
10. transmission - automatic or manual
11. door\_count
12. seat\_count
13. fuel\_type - gasoline, diesel, cng, lpg, electric
14. date\_created - when the ad was scraped
15. date\_last\_seen - when the ad was last seen. Our policy was to remove all ads older than 60 days
16. price\_eur - list price converted to EUR

**To Run ELK and View the dataset:-**

1. Run ElasticSearch - **./bin/elasticsearch**
2. Verify it by clicking <http://localhost:9200/>
3. Run Kibana - **./bin/kibana**
4. Verify it by clicking <http://localhost:5601>
5. Create logstash\_cars.conf file
6. Run logstash - **./bin/logstash -f location\_to\_config\_file**
7. We can see logstash running and importing each document (row) into ElasticSearch
8. Open Kibana and specify the index as specified in logstash\_cars.config
9. Open Visualize and create visualization graphs and charts and save it
10. Import that to Dashboard in Kibana
11. Also visualize the data trends with Machine Learning feature available in Kibana. Create a job with dataset having a time based attribute in it and predict the anomalies in it in the next few weeks or months





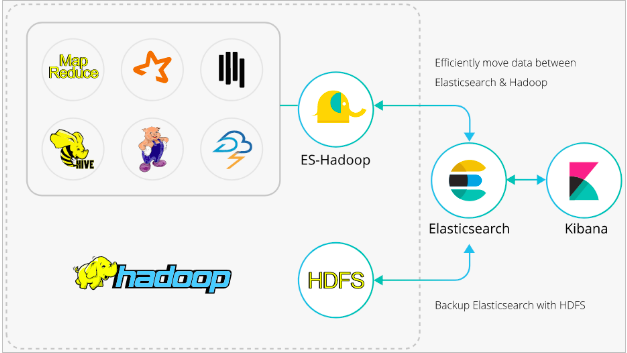
**Advantage**

1. **Visualize any document (csv, avro, parquet, txt, json etc.)**

With a proper spark/hadoop job to convert the big data into csv format and then use logstash to move and index the data into ElasticSearch, we can visualize the data in Kibana by creating custom dashboards

1. **Integrate Hadoop ecosystem with Elasticsearch**

Hadoop ecosystem and Elasticsearch can be integrated with ES-Hadoop. This will enable us the feature to feed live streaming as well as batch data into Elasticsearch and visualize data in Kibana



1. **Modularity**

As each module (Hadoop ecosystem, Logstash, Elasticsearch, Kibana) are independent, we can replace any module and have a more flexible and customizable module. For example, we can replace Kibana with our own front end with a more features to upload input file to ElasticSearch and visualize based on some parameter.

1. **Helpful to anyone**

Any person (testers, developers, architects, PMs, AMs and even the customers) who needs to visualize any raw big data to get an insight to it is possible.

**Future Scope**

We can create our own UI to visualize multiple dataset at the same time and merge the data together with different parameters to get a good insight.

For example:- we can merge 2 datasets (income of people in an area, and the household property sold and brought), we can merge these two datasets to see if there are any chances of people likely to sell and buy new homes in coming years.