The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

Here I have used the Pyspark technology to explore the data set. The data set is obtained from the ArcGIS website.

### Justification

Since our data set contains more than 1.9 lakh rows of data i.e. it’s a big data. So to handle this much amount of data we needed a good and efficient way, hence we have used the pyspark.

## Enterprise Data

### Technology Choice

Here we’ve defined the pyspark technology. Now, if the question is, “when should we use pyspark?” so the simple answer to that is, when we need to handle the big data.

### Justification

In the modern world problems data sets are very large, so to handle this data with only python becomes a tedious and lengthy task which is also computationally very costly. That’s why we have used the pyspark technology.

## Streaming analytics

### Technology Choice

For doing the data analytics we have used the SQL querries which are in the form of pyspark are executed in our model.

### Justification

Query languages are often offer an ease to handle the data sets of higher order. In pyspark it is very easy to implement basic SQL querries on to the data set for analytics. Also various data visualizations are done to understand the analysis much better using the matplotlib technology.

## Data Integration

### Technology Choice

Pyspark and SQL queries

### Justification

Query languages are often offer an ease to handle the data sets of higher order. In pyspark it is very easy to implement basic SQL querries on to the data set for integrating the whole data set.

## Data Repository

### Technology Choice

Google colab (drive) / Github

### Justification

I made this notebook on google colab hence to store the data set I used the drive. Also one can also use the Git.

## Discovery and Exploration

### Technology Choice

The data visualization has helped me to explore the various aspect of the data set.

### Justification

Matplotlib and seaborn are the best libraries for exploring the data using visualizations. As they give various plots which help to explore the data thoroughly.

## Actionable Insights

### Technology Choice

The finding of the project is that, the severe cases can be reduced if we provide the better light at the intersections of the road during winter and rainy seasons especially.

### Justification

The data analysis proves that many severe collisions had happened due to bad light.

## Applications / Data Products

### Technology Choice

For the betterment of transport system

### Justification

To save the lives of the citizens on the roads, projects of these can help a lot.

## Security, Information Governance and Systems Management

### Technology Choice

Various Government organizations can use projects of these kind to improve their quality of transport system.

### Justification

To save the lives of the citizens on the roads, projects of these can help a lot.