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

<https://developer.ibm.com/articles/data-science-architectural-decisions-guidelines/>

## Data Source

### Technology Choice

Kaggle - database about startups in kickstart in three years 2015 to 2018

CSV , no real time delivery

### Justification

Please justify your technology choices here.

Not to big .. about 400K lines only

## Enterprise Data

No other data sources. So location is in my notebook

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Nothing else other than the data source

### Justification

Please justify your technology choices here.

All scope is in the collected data from data source

## Streaming analytics

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Nothing of Apache Spark, Apache Kafka, Samza, Apache Flink, Apache Storm, Total.js Flow, Eclipse Kura, and Flogo

### Justification

Please justify your technology choices here.

No need

## Data Integration

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Nothing of ApacheSpark , ETL, or SQl

### Justification

Please justify your technology choices here.

No need

## Data Repository

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

My Notebook disk as the data is about 20MB

### Justification

Please justify your technology choices here.

## Discovery and Exploration

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Pandas, Matplotlip, Jupyter, Python, sklearn no thing of : pyspark, PixieDust, we use histograms for specific features and scatters

### Justification

Please justify your technology choices here.

Easy to use and enough for the volume of data

## Actionable Insights

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Pandas, PCA in sklearn, Keras and TensorFlow

### Justification

Please justify your technology choices here

correlation is in some features is high, so used PCA to merge columns.

Used Standardization instead of Normalization for better performance. We see that outliers are not there significantly, but the distribution is along all values. Histograms show good guassian values.

## Applications / Data Products

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

No need

### Justification

Please justify your technology choices here.

No need

Easy and data is not too big

## Security, Information Governance and Systems Management

My notebook secuirty

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

My notebook security

### Justification

Please justify your technology choices here.

I am using my notebook