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

We do not use any technology such as IBM DB2 and we do not do any real time data analysis. The data is just downloaded into jupyterlab and it is available in csv format.

### Justification

The data is a simple household survey data and not an enterprise data. Hence, we do not need any technology such as IBM DB2

## Enterprise Data

### Technology Choice

No technology was used.

### Justification

The data is not enterprise data and hence we do not need to transfer subsets of it to a cloud.

## Streaming analytics

### Technology Choice

No technology was used

### Justification

Streaming analytics is not needed for this project. All the data is used at the same time and hence we do need processes like batching process.

## Data Integration

### Technology Choice

Jupyterlab notebook.

### Justification

The data is simple and already available in a csv format and can be easily cleaned and transformed using simple python libraries. We do not need advanced technologies like Apache Spark and others.

## Data Repository

### Technology Choice

No technology used here.

### Justification

The data is not an enterprise data and need not to be stored. It is already available at the Ghana Statistical Service website.

## Discovery and Exploration

### Technology Choice

Python libraries pandas and matplotlib available in jupyterlab were used for data visualization and exploration.

### Justification

Matplotlib and pandas libraries supports the widest range of possible visualizations.

## Actionable Insights

### Technology Choice

Python, pandas and scikit-learn, keras and a tensorflow backend.

### Justification

Python is easier to use than R, and moreover, scikit-learn nicely groups all necessary machine learning algorithms together. Pandas also makes data manipulations relatively easier. More importantly, we do not need to scale our data and hence we do not need ecosystems such systemML and Apache Spark. Tensorflow is one of the most widely used deep learning frameworks and provides a backend for keras. Keras is simpler to use compared to other deep learning frameworks.

## Applications / Data Products

### Technology Choice

PDF report, IBM machine learning.

### Justification

The report will be shared among various stakeholders involved in policy making. The model has also been deployed on the IBM machine learning environment for classifying new data by the various stakeholders interested in the findings of this project.

## Security, Information Governance and Systems Management

### Technology Choice

No technology is used here.

### Justification

The data is freely available on the internet, and the results of this project is available for use by any entity, whether an individual, business or the government.