The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

The data source for this project is a CSV file obtained from Kaggle. This data file contains demographics and health information that was collected by trained personnel for JPAC Center for Health Diagnosis and Control using direct interviews, examinations, and blood samples.

The CSV data file is loaded to the IBM Cloud Object storage and then Python is being used for extracting and processing the data from the Object storage. We will also use Spark SQL to connect to the Object Store and handle it like a database.

### Justification

IBM Cloud Object storage is used as the Data storage solution for the following reasons:

1. Most of the data is being expected in the form of Data files. Hence storing the data files becomes easier in an object storage.
2. For this Capstone project, we currently have not yet explored the need for a SQL database due to the high cost of such databases. However this is something that can be looked into in the future if the need arises.

## Enterprise Data

### Technology Choice

Not applicable as of now.

### Justification

At present as we do not see any Enterprise data needs for this project. As of now this is being envisioned as a stand-alone model / application / product. However this is something that can be looked into in the future if the need arises.

## Streaming analytics

### Technology Choice

Not applicable as of now.

### Justification

Our present use-case does not have the need for streaming analytics. However as we move forward and work towards refining and enhancing our use-case, there is definitely scope for using streaming analytics to be able to stream real-time data for use in our model. This will need to be revisited in the future based on the need for streaming data.

## Data Integration

### Technology Choice

At present Data Integration is pretty straight forward and is mostly manual. The data is made available to us as csv or Excel files containing the data. This is downloaded and kept in IBM Object store from where this is loaded into our Jupyter Notebook using native pandas or Apache Spark libraries.

### Justification

Our present model is a very simple one with no requirement for frequent and automatic data updates via data integration frameworks. However this is something that will need to be considered in the future as we enhance our model and try and integrate it with real-time data capturing and monitoring systems.

## Data Repository

### Technology Choice

We are currently using IBM Object store as our data repository for storing the source data files and also for storing the data used by our models.

### Justification

IBM Cloud Object storage is used as the Data storage solution for the following reasons:

1. Most of the data is being expected in the form of Data files. Hence storing the data files becomes easier in an object storage.
2. For this Capstone project, we currently have not yet explored the need for a SQL database due to the high cost of such databases. However this is something that can be looked into in the future if the need arises.

## Discovery and Exploration

### Technology Choice

For the Discovery and exploration of the data and our model, we are currently using the following.

* Apache Spark Framework
* Jupyter Notebooks with Python 3.5 and Spark 2.3
* IBM Watson Studio as our cloud based platform for building and deploying our solution
* IBM Object Storage as our data repository

### Justification

Most of the tools and technology used by us in the project are Open source and IBM Watson studio is one of the best a cloud based platform which allows us to build and deploy AI solutions built using these Open Source tools. This works best for our project based on our current needs. However in the future this can be revisited based on the need.

## Actionable Insights

### Technology Choice

None as of now.

### Justification

There are no actionable insights at present. However this can be relevant in the future as we enhance our model / data product.

## Applications / Data Products

### Technology Choice

None as of now.

### Justification

There are no applications / Data Products in scope of our project at present. This is something that will be addressed as part of the enhancement scope in the future.

## Security, Information Governance and Systems Management

### Technology Choice

Not applicable as of now. There is however some Systems Management that is required which is being performed manually on a need basis.

### Justification

At present our model is stand-alone model which does not interact or integrate with any external application or systems. Hence there is no real need or scope for implementing Security, Information Governance and Systems Management. This is also something which will need to be considered in the future as we enhance our model and start streaming data and also start interacting with external systems / data products.