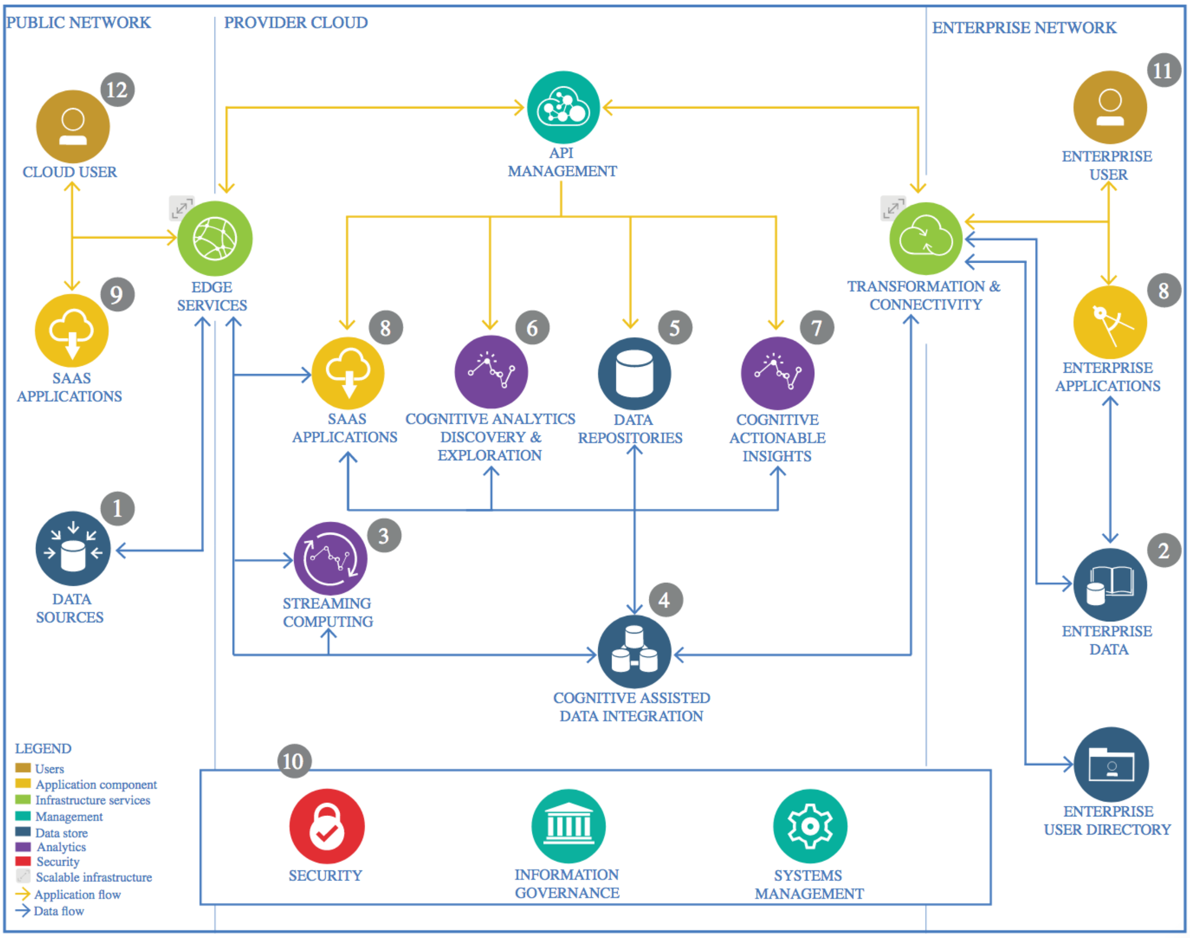
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

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

The data were obtained from https://opendata.cityofnewyork.us/. The Jupyter Notebook ad Pandas were used.

1.1.2 Justification

Please justify your technology choices here.
The data are easily accessible. The Jupyter notebook is open source and supports many frameworks and/or APIs. Pandas was used to allow creation of a data frame. The data frame was created for the data with great efficiency in the Jupyter notebook.

1.2 Enterprise Data

1.2.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.
The data were defined, integrated, modelled and retrieved using the Jupyter notebook. The IBM cloud object storage was used for storage. The notebook was shared using GIST, onto GitHub.

1.2.2 Justification

Please justify your technology choices here.
Whereas the data used are open source and no particular enterprise or organization is directly mentioned in line with the data, probably the biggest 'organization' indirectly mentioned are the people including New Yorkers and the general public. The data were sourced from 'Open Data for All New Yorkers'. This is a bigger organization than any other. My involvement in looking at these data adds value to the data by extracting a meaning out of them. The way this is done must be well organized for the benefit of anyone that might be interested (which makes them a stakeholder) including my Coursera learning colleagues. On the platform mentioned above, there is easy and/or efficient data access and retrieval from storage and, good data security handled by IBM. If anyone in the future would like to retrieve the stored data, it will be possible for them to do so from a secure platform. The platform provides confidence and trust the data set.

1.3 Streaming analytics

1.3.1 Technology Choice

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

1.3.2 Justification

Please justify your technology choices here.
This was not used since the data were collected cross-sectionally for the different years 2007-2014.

1.4 Data Integration

1.4.1 Technology Choice

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

This was not done.

1.4.2 Justification

Please justify your technology choices here.

Data integration was not done since the data are provided already in a relational database.

1.5 Data Repository

1.5.1 Technology Choice

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

For these data, the data were stored in the IBM cloud storage object for data analysis, sharing and also reporting as shown in the power point presentation. The data were stored as Jupyter notebooks.

1.5.2 Justification

Please justify your technology choices here.

The IBM cloud storage object is safe and secure.

1.6 Discovery and Exploration

1.6.1 Technology Choice

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

Python 3.7 - Jupyter notebook, Pandas, Scikit-Learn and Matplotlib, TensorFlow and Keras.

1.6.2 Justification

Please justify your technology choices here.

The Jupyter notebook is open source and easily accessible. Pandas allowed the creation of data frames, which were explored using correlation statistics, visualization (using Matplotlib) and descriptive analytical methods. Scikit-Learn was used for modelling – analysis of variance (ANOVA) and multiple linear regression in order to determine the association of race or sex with mortality in New York City. There are several efficient modules available in Scikit-Learn and Matplotlib that were called for the various analyses and visualization. TensorFlow (TF) was used as a deep learning framework as was Keras and, both are open source. They were both used for greater comparison of cost or loss with traditional regression methods. TF would require a lot more data and no convergence was seen with the gradient descent algorithm used. On the other hand, convergence was achieved with Keras.

1.7 Actionable Insights

1.7.1 Technology Choice

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

Jupyter notebook, Scikit-Learn, Pandas, TensorFlow and Keras.

1.7.2 Justification

Please justify your technology choices here.

These methods were used to model the data for the analytics that helped discover the relationship between race or sex with mortality in New York City.

1.8 Applications / Data Products

1.8.1 Technology Choice

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

Jupyter notebook, Pandas, Scikit-Learn, TensorFlow and Keras, GIST.

1.8.2 Justification

Please justify your technology choices here.

Jupyter notebook, Pandas, Scikit-Learn, TensorFlow and Keras were used to model the data in order to produce an output. The file was then shared using GIST. All these are readily available for use. The Python script is therefore available on GitHub.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

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

IBM cloud storage object. Jupyter notebook.

1.9.2 Justification

Please justify your technology choices here.

The IBM cloud storage object is safe and secure. The storage platform is professionally managed and the scripts are quickly and/or efficiently retrieved from storage whenever required. Large sets of data can be stored effectively.