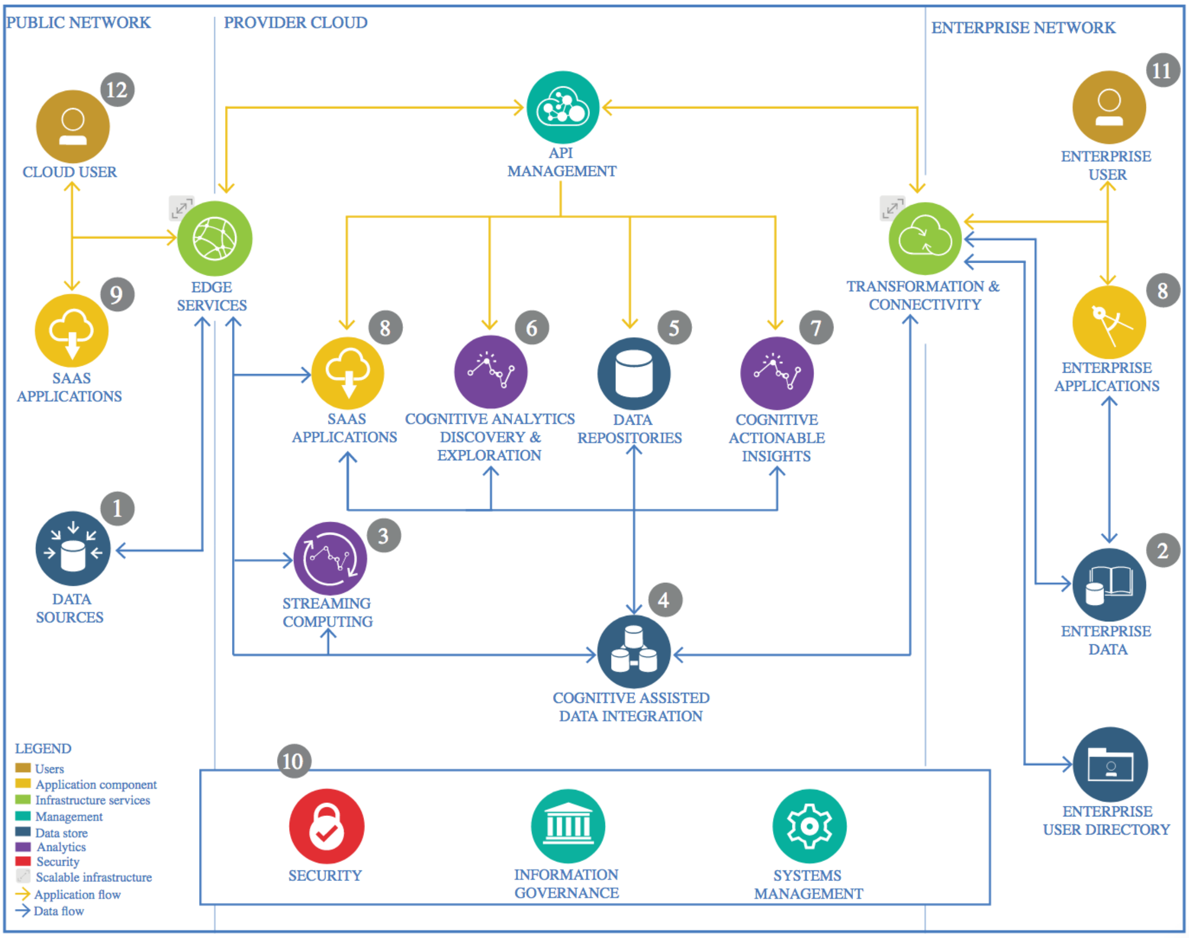
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

This is a description of the technical solutions used in the IBM advanced data science capstone project.

The dataset contains information on the income of households collected by the national statistics agency (NSA) during a 2016 nation-wide survey in a southern African country known as Namibia. It is anonymized and hence there is no way to identify any particular household.

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

The Namibia Statistics Agency maintains a central data catalogue at https://nsa.org.na/page/central-data-catalogue/, where public datasets can be downloaded after proper login.

1.1.2 Justification

The data was freely available since meant for use by the researchers, data scientists, and the public at large.

1.2 Enterprise Data

1.2.1 Technology Choice

No enterprise data sources was used.

1.2.2 Justification

The best available data for the project was available from the NSA web portal.

1.3 Streaming analytics

1.3.1 Technology Choice

Streaming was found not necessary; however Apache Spark in the IBM Cloud environment could have been used.

1.3.2 Justification

Users can easily open an account on the IBM Cloud.

1.4 Data Integration

1.4.1 Technology Choice

Apache Spark, IBM Developer Skills Network Labs (CCLabs), Jupiter Notebooks with Python 3 (pandas, numpy). In this project, all data came from one single source used by the notebooks. Data encoding and other forms of cleaning done for ulterior processing.

1.4.2 Justification

These technologies are easily accessible and perform quite well.

1.5 Data Repository

1.5.1 Technology Choice

Cloud Object Storage in IBM Cloud, CSV files in IBM Developer Skills Network Labs.

1.5.2 Justification

The dataset have thousands of rows and thousands of columns of data. This load is easily handled by these technologies.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupiter Notebooks on Apache Spark and CCLabs, Python 3 (pandas, numpy, seaborn, Matplotlib).

1.6.2 Justification

They are cheaply available and do the job.

1.7 Actionable Insights

1.7.1 Technology Choice

Jupiter Notebooks on Apache Spark and CCLabs, Python 3 (pandas, scikit-learn, Keras, TensorFlow using deeplearning (sequential model and non deeplearning algorithms (kNN, Logistics Regression, GBT, AdaBoost, RandomForest).
The processing can be scaled to multiple processors if necessary with Apache Spark and SystemML.

1.7.2 Justification

These powerful tools for getting useful insights quickly are available on the IBM Cloud

1.8 Applications / Data Products

1.8.1 Technology Choice

At this there will be a PDF document to show the results since there is no need for now to use the model in production. This can be easily done however in IBM Watson on the Cloud or using Node-RED.

1.8.2 Justification

The focus as of now is to know whether classifying the households in income percentiles can be automated using machine learning algorithms. No need for production.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

The system could be run from the web servers of the agency or from the cloud. There is no plan as yet to go that route

1.9.2 Justification

As already stated there is no plan for production yet.