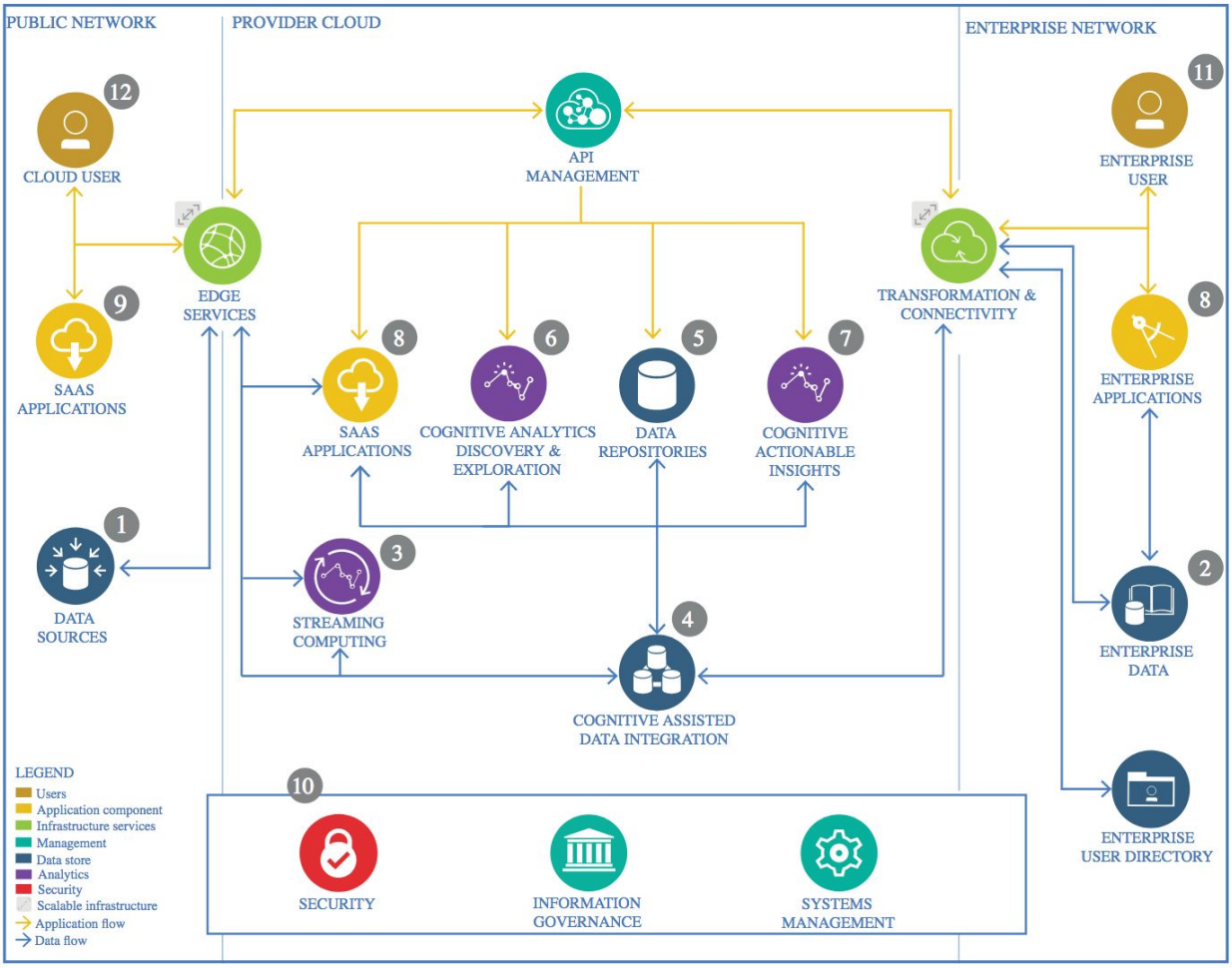
Architectural Decisions Document (ADD)

The Lightweight IBM Cloud Garage Method for Data Science

Project: Student Recruitment Analysis: Classification and Regression

1. Architectural Components Overview



1.1. Data Source

1.1.1. Technology Choice
The dataset has been taken from Kaggle, which is online platform where lots of data scientists meet and compete. They work on different datasets and bring a solution to real-world problems. The dataset is a recruitment dataset.

1.1.2. Justification
The dataset is publicly available and it points to an interesting problem. Very few people have worked with it.

1.2. Enterprise Data

1.2.1. Technology Choice
Jupyter Notebooks on IBM Watson will be used to analyze the data and run models.

1.2.2. Justification
Jupyter Notebooks are very interactive when we need to explore and at the same time work on it. Since our dataset is small, thus it will be sufficient to work with the Jupyter Notebooks. Also, it can be shared with others for feedback and help.

1.3. Streaming Analysis

1.3.1. Technology Choice
The dataset doesn't involve streaming.

1.3.2. Justification
It is not real-time data being obtained from the sensor, instead, it is data for students applying for recruitment, with small size.

1.4. Data Integration

1.4.1. Technology Choice
Not available

1.4.2. Justification
The dataset already comes in single CSV file and doesn't need any integration.

1.5. Data Repository

1.5.1. Technology Choice
The dataset is available on Kaggle (link mentioned in initial sections). To work with it on the IBM Watson cloud, it is uploaded in the object storage and was accessed.

1.5.2. Justification
Since size of data is small, thus it would be better to upload and access, rather than streamlining it from other sources.

1.6. Discovery and Exploration

1.6.1. Technology Choice
Python numpy, pandas, matplotlib, and seaborn was used for plotting and exploring data. It was initially loaded as a data frame and further work is carried on for the data.

1.6.2. Justification
Since the dataset had multiple rows and columns, thus working with the data frame seems justifiable as it gives overall structure of the data.

1.7. Actionable Insights

1.7.1. Technology Choice
Some preprocessing is done and other ML models were loaded using scikit learn library of Python.

1.7.2. Justification
Preprocessing and feature selection are important steps, which are easily available from scikit learn library as it brings very effective solutions for the task.

1.8. Application / Data Products

1.8.1. Technology Choice

- SytemML for Apache Spark (distributed Machine Learning)
- SciKit-Learn (in-memory Machine Learning)
- IBM Cloud Machine Learning for the deployment of a pre-trained model.

1.8.2. Justification

SciKit-Learn provides access to maximum ML models for working on own machine, deployment to SystemML and Apache Spark servers would be a good choice for larger data sets. The Cloud platform Provides ease of access to already trained models. SystemML is used for deployment off deeplearning models on IBM Watson cloud platform

1.9. Security, Information Governance, and Systems Management

As of now, no action is taken, however, any form of redistribution of this work needs prior permission. If you want to download and further improve it, do contact me through the e-mail: srivastava_pravesh@iitgn.ac.in