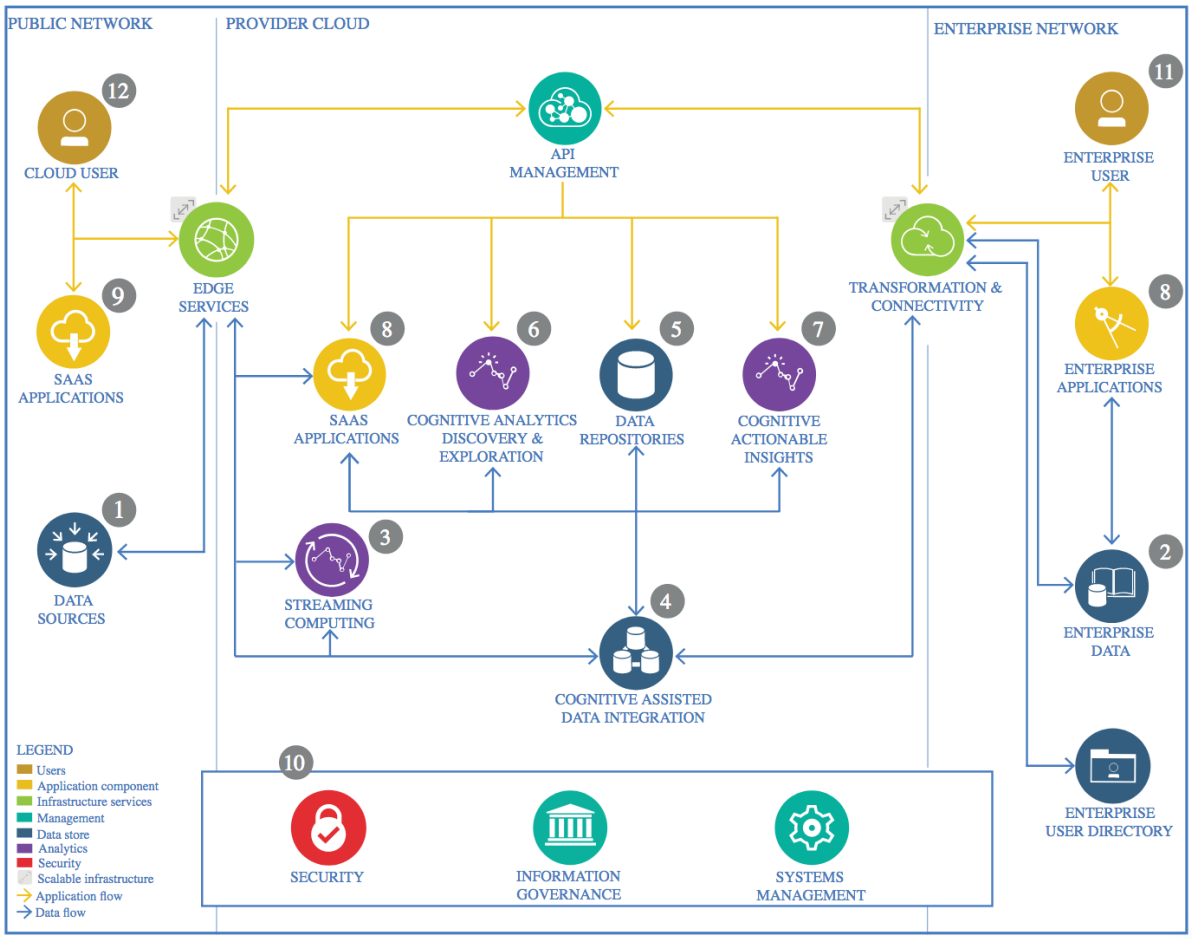
The Lightweight IBM Cloud Garage Method for Data Science

Telecom Churn Case Study

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

Understanding data is very important, as a first step, before starting any machine learning process. This data is available in Kaggle and part of a competition.

Data available:

https://www.kaggle.com/competitions/telecom-churn-case-study-hackathon-c43/data

1.1.2 Justification
It was data for a Kaggle competition.

1.2 Enterprise Data

1.2.1 Technology Choice
GitHub Repository

1.2.2 Justification
The Notebooks and Data available everywhere and up to date.

1.3 Streaming analytics

1.3.1 Technology Choice
N/A

1.3.2 Justification
N/A

1.4 Data Integration

1.4.1 Technology Choice
N/A

1.4.2 Justification
N/A

1.5 Data Repository

1.5.1 Technology Choice
Local Hard Drive and Jupyter Notebooks

1.5.2 Justification
The amount of data didn’t justify using IBM storage or other.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Pandas library: For loading data and exploring the data
Seaborn and Matplotlib: for metrics and data visualization.
Numpy: array processing

1.6.2 Justification

For the kind of that being used, these libraries can do the job.

1.7 Actionable Insights

1.7.1 Technology Choice

Data Quality assessment:

- Descriptive and Exploratory Analysis;
  - check duplicated data;
  - check overall null/nan values;
  - check overall binary class balance

Feature Engineering and data transformation:

- Column removal and/or filling with data (mean and mode);
- Standard Scaling;
- Over Sampling (SMOTE), RandomUnderSampling – to balance dataset;

Algorithms used in the project:

- Decision Tree Classifier
- Random Forrest Classifier
- Keras Deep Learning Model

Frameworks used in the project:

- Sklearn
- Keras
- ImbLearn

Metrics considered to evaluate the models:

- Confusion Matrix
- Accuracy
- Precision
- Recall

1.7.2 Justification

Descriptive and Exploratory analysis enables us to understand what kind of that we are dealing with, the problems the it might have and some general statistics.
Feature Engineering and data transformation to prepare the data for further usage in the models, solve the issues identified in the previous step and in some cases reduce the overall size of data.
Algorithms, usually Tree bases algorithms perform better at imbalanced datasets.
Frameworks used in the project, open source with already pre-built models and a lot of documentation.
Metrics considered to evaluate the models, accuracy considered but not the best option, since we are dealing with an imbalanced dataset, so the most important should be Recall.

1.8 Applications / Data Products

1.8.1 Technology Choice
Jupyter Notebooks

1.8.2 Justification
Jupyter Notebooks, it is a tool that enables us to run just a few cells of code at a time and at the same time, produce a report containing the code used, visuals and text.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice
N/A

1.9.2 Justification
N/A