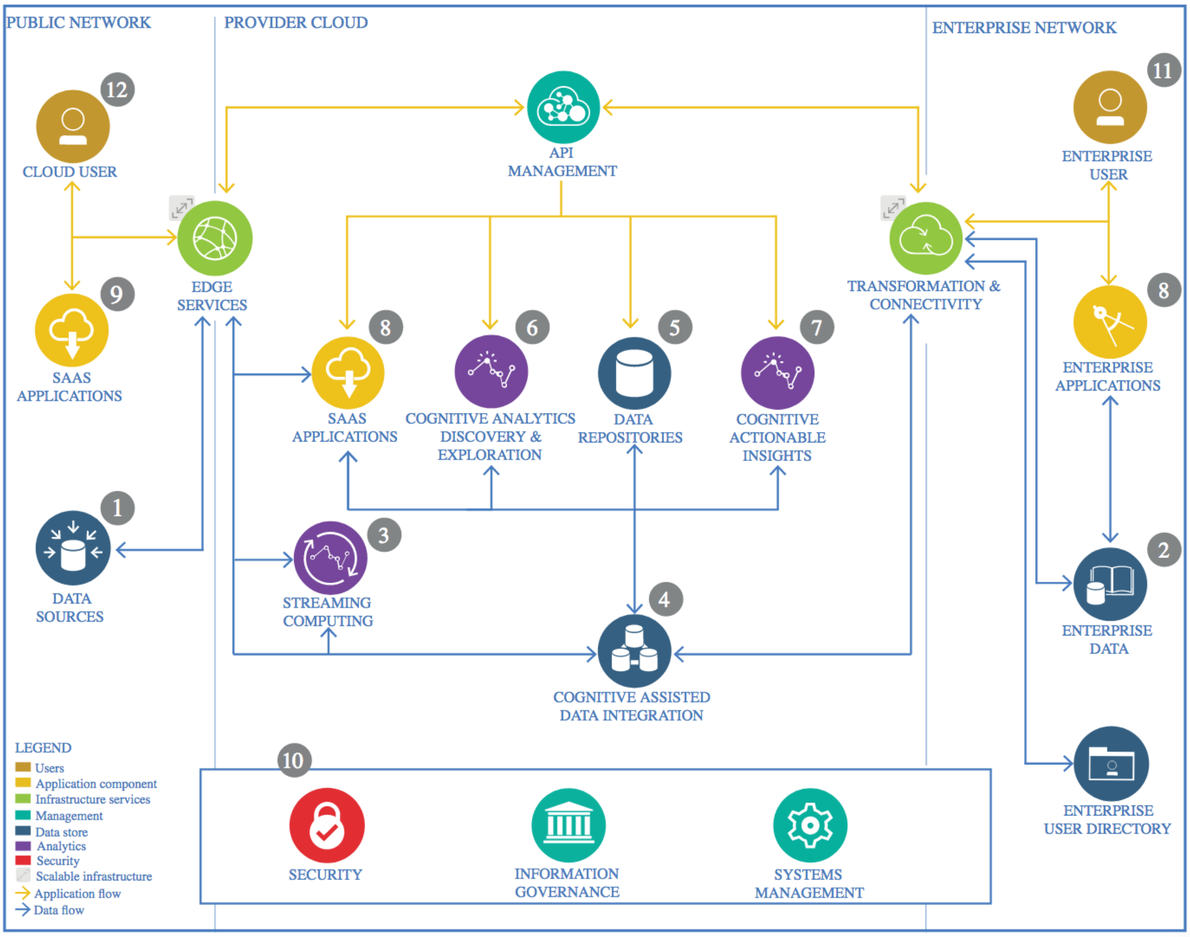
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

Data source is an external data set provided by Kaggle and made available in .csv format. Data is on auto insurance claim with details about customer, claim, auto, incident and a flag if the claim is recognized as a fraud.

1000 records of auto-insurance claims with customer and accident data. Records are labeled (fraud_reported) when fraud is reported making this a binary classification dataset.

1.1.1 Technology Choice

Excel format is convenient for export and transfer of this dataset.

1.1.2 Justification

No limitations.

1.2 Enterprise Data

When new claims are recorded it is necessary to make them available for use in machine learning algorithm. For that purpose, a connection to enterprise data system or ETL pipeline for exporting the data is required.

1.2.1 Technology Choice

Direct connection or ETL pipeline are standard solutions for in enterprise environments.

1.2.2 Justification

NA

1.3 Streaming analytics

Not applicable. We are not working with streaming data. Data is obtained once from a database. Final product can take data in batches as well.

1.3.1 Technology Choice

This component is not needed.

1.3.2 Justification

This project doesn’t involve working with real time stream data.

1.4 Data Integration

In the data integration stage, data is cleansed, transformed, and if possible, downstream features are added

1.4.1 Technology Choice

It is a single data source, already merged with data from different databases. All the pre-processing is done with python and pandas dataframes.

1.4.2 Justification

Cleaning the data, transforming it into a format suitable for analytics and extracting important features is crucial for achieving high performance with machine learning models.

1.5 Data Repository

Persistent storage for your data.

1.5.1 Technology Choice

CSV file with one record per auto-insurance claim. Dataset is small enough to fit into the excel file.

1.5.2 Justification

No need for anything more complex that a excel.

1.6 Discovery and Exploration

Statistical analysis and visualization of the dataset features. When dealing with high-dimensional and large datasets we use statistical measures to learn more about the data.

Correlations are very useful for uncovering relationships in the dataset. With correlation analysis we take bird-level view and look for features that have high positive or negative correlation. Correlated features are not useful for modeling and can even negatively impact model performance so we remove them from the dataset.

We also look at numerical and categorical distributions to inspect value ranges, outliers and clusters as shown on the right.

1.6.1 Technology Choice

Numpy, Pandas and Seaborn library for exploration and static visualizations during data exploration phase.

1.6.2 Justification

All these components are widely used by data science community. They are open-source, with extensive documentation and community support.

1.7 Actionable Insights

Python with data science libraries (numpy, pandas, seaborn, scikit-learn and keras) is used. IBM Watson Studio with **jupyter notebooks** is selected as platform.

- Dedicated jupyter-notebooks for each phase of the Lightweight IBM Method for Data Science. Named according to conventions.
- Pre-run jupyter-notebook detailing each step with explanations in text.
- Chosen graphs from EDA and model evaluation results are made available in PowerPoint presentations for stakeholder and data science peers.
- Video presentation detailing the whole project is uploaded to youtube.

We are dealing with binary classification and an unbalanced dataset, so the we chose **f1-score** as evaluation metric.

In our case F1score is more appropriate than for example accuracy, which tells us proportion of true results among the total number of cases. On a imbalanced dataset, as ours is, we can have high accuracy just by predicting everything with majority label resulting in a useless classifier.

Non linear models perform best on this dataset.

- LightGBM is best performing model.
- Fully connected single hidden layer neural network is 2nd best model.

1.7.1 Technology Choice

All these components are widely used by data science community. They are open-source, with extensive documentation and community support. Using scikit-learn and keras, all state-of-the-art models and metrics are supported.

1.7.2 Justification

Technologies are widely used, accepted by community and de-facto standard in open-source domain.

The F_1 score is the harmonic mean of the precision and recall, where an F_1 score reaches its best value at 1 (which means perfect precision and recall) and worst at 0. This makes it a very robust measure in which false positives and false negatives are penalized.

1.8 Applications / Data Products

1.8.1 Technology Choice

- Jupyter notebook with extensive textual explanation detailing the whole process
- Trained model and weights are export in a .zip file ready for deployment.

1.8.2 Justification

- Delivering the jupyter notebook with machine learning model makes it easy to start testing the application in local environment.
- Exported model can easily become data product when deployed in e.g. docker container and exposed with an API.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Notebooks are stored on IBM Watson Cloud. Git is used for code storage. Cloud filesystem for document storage.

For final product deployment we need to consider local resources, hybrid cloud or full cloud soluion.

1.9.2 Justification

All environments are managed and exposed as a Service. Security is managed by the vendor so we can focus on the work. Standard practice in rapid prototyping.