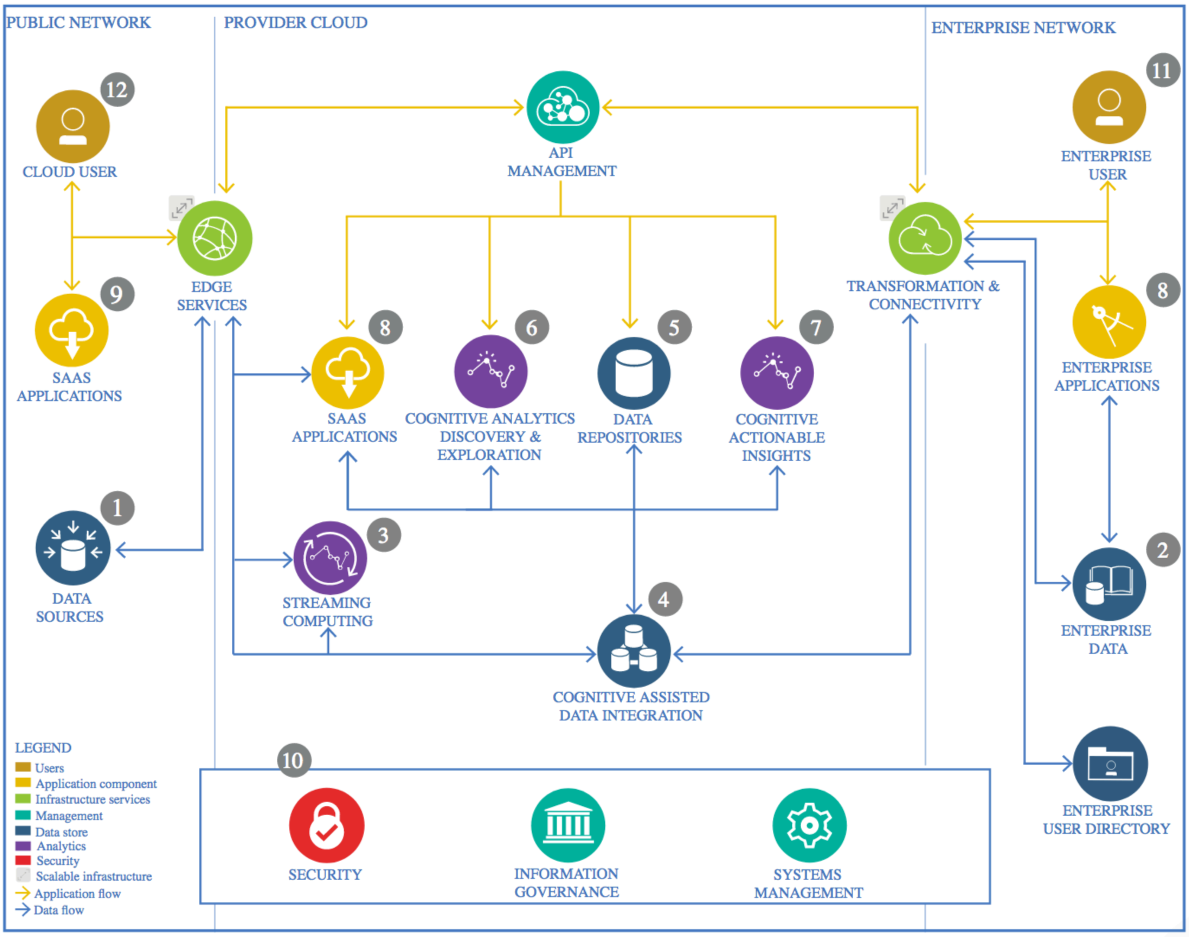
The Lightweight IBM Cloud Garage Method for Data Science

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

For the data source I utilized the Credit Card Fraud Detection from Kaggle. https://www.kaggle.com/mlg-ulb/creditcardfraud

1.1.2 Justification

I chose to do this data set in order to utilize machine learning and deep learning algorithms in a real-world example. I wanted to focus simply on the curation of data and then to be able to apply my model while worrying less about the wrangling and cleansing of data. I also wanted to create a type of anomaly detector on a static dataset.

1.2 Enterprise Data

1.2.1 Technology Choice

For the enterprise data I chose to store all of my data as a parquet file on IBM Cloud Storage.

1.2.2 Justification

In order to stay within the infrastructure provided by Coursera with the IBM Cloud framework I chose to stay within the IBM ecosystem. The dataset is not that large so it was easy for me to move it around and perform machine learning on it without having to worry about storage solutions.

1.3 Streaming analytics

1.3.1 Technology Choice

The dataset is static. N/A

1.3.2 Justification

The data set is static so no streaming analytics was required.

1.4 Data Integration

1.4.1 Technology Choice

I chose Apache Spark with Python 3.7, pyspark, to integrate my data.

1.4.2 Justification

The entire size of the data set was only 60 MB which made it extremely easy to run spark jobs with a single worker on a single worker. The data types where simply float64 types and the source system was just a CSV file. In order to code all of the model's extensive knowledge in Python 3.7 specifically the Pandas, Pyspark, Scala, Keras, Seaborn, and Matplotlib libraries was required. I turned some of the csv data tables into Spark Data frames which enabled SQL queries to be run on the data set in the Spark Session.

1.5 Data Repository

1.5.1 Technology Choice

For my persistent storage I utilized IBM's Cloud Object Store. which is

1.5.2 Justification

Cloud Object Store is implemented within the IBM Watson development environment.

- How does this impact cost of storage
  - For this Coursera course and for the data storage amoung and computational power I used the free account was sufficient
- Which data types are supported?
  - Resembles a file system, any datatype is supported.
- How good must point queries be supported?
  - RDBMS are king at point queiries becasu an index can be created on each column
- What skills are required
  - Apache SparkSQL Python, Scala, and SQL

- How good must full table scans be supported
  - Full table scans are just bound by I/OP bandwith of the OS
- What is the amount of storage needed
  - Will use only 80MB of storage
- Growth and scaling?
  - Fully elastic on Cloud Object Storage

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupyter notebooks, python, pyspark, scikit-learn, pandas,matplotlib, seaborn.

1.6.2 Justification

- What type of visualizations are needed
  - Matplotlibe and Seaborn supports the widest possible visualiztions including runc charts, histograms, box-plots, and scatter plots
- Are interactive visualizations needed?
  - NO
- Are coding skills available/required?
  - Yes. For both matplotlib and seaborn
- Do metrics and visualizations need to be shared with business stakeholders?
  - The notebook can be shared through jupyter notebooks.

1.7 Actionable Insights

1.7.1 Technology Choice

In order to identify fraudulent credit cards we created three models each with different selections in technology. Scikit learn for a single node Logisitic Regression model, pyspark's logistic regression model, and Keras Tensorflow backend sequential feed forward neural model made of Dense layers. All of which were implemented with python, pandas, apache spark.

1.7.2 Justification

- What are the available skills regarding programming language?
    - Apache spark supports python and python skills are widely available.
- What are the cost of skills regarding the programming language?
    - Cost are usually low as python is open source.
- What are available skills regarding frameworks?
    - Pandas and scikit-learn are both clean and easy to learn, skills are widely available
- What are the costs regarding frameworks?
    - Python is open source, so cost is low. Keras and Tensorflow skills however are could be much more expensive.
- Is model interchange required
    - Scikit-learn, keras, tensorflow can be serialized and we can save Keras model and export/load them as need be
- Is parallel or GPU based training or scoring required?
    - Not applicable in this case but in Apache Spark this can very quickly be integreated by loading a trained keras model into Apache Spark.
-

1.8 Applications / Data Products

1.8.1 Technology Choice

N/A

1.8.2 Justification

N/A

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

N/A

1.9.2 Justification

N/A