The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

The data used in this study is export from Enterprise dashboard.

The other part of data (Interest rate ) is exported from Reserve Bank of NZ website

### Technology Choice

The data source is in CSV format and with raw format and mix of different data from dashboard. Since this data is obtain from enterprise, it is only for private use. Data also might be adjusted to keep its confidentiality and only use for the purpose of this report.

Due to this is raw data, ETF processing is needed and described later in this document in Data Integration section.

Since the form of data is downloadable files, there is no need for connection, streaming or database connection.

However, in the future, if this model predicts well, we might scrape data relevant features data from dashboard to make a real time prediction.

* Jupyter
* Python
* Pandas

### Justification

Due to data is in CSV format and small dataset with monthly record from 1995 (<300 rows), it is easy to load with Pandas in Python Jupyter

## Enterprise Data

### Technology Choice

Files will be stored in the Object Store belonging to the enterprise cloud environment

**Tools:**

* • IBM Watson Studio environment.
* • Python 3 programming language with Pandas library
* • Spark 2.3 provided from the IBM Watson Studio.
* • Spark ML library for classification, clustering, and machine learning algorithms
* • Keras API for AI neural networks configuration and training.
* • Jupyter Notebook as a Python editor.

### Justification

I have used IBM Watson and it is very easy to use

## Streaming analytics

### Technology Choice

N/A

### Justification

N/A

## Data Integration

Data will be downloaded from company dashboard

The downloaded CSV will be processed:

* Remove duplicate record if any
* Check for NaN values
* Formatting

### Technology Choice

* Python
* Pandas
* Seaborn
* Tensorflow
* Keras
* Jupiter notebook

### Justification

Pandas is quick and simple solution to read CSV

**Why I have chosen a specific framework?**

* **Pandas** for data manipulate for easy to use
* **Seaborn** for data visualization due to its popularity and easy to use but powerful with 1 line of code
* **Tensorflow** for ML as it is open source and support both regression and neural networks
* **Keras** for deep learning due to its simplicity interface. Also it provide rapid prototyping and also work with TensorFlow
* **Python** is very powerful language and easy to learn. It is also very popular so most of problem we face can be found on internet.
* **Jupiter notebook** easy to use

## Data Repository

Folder include:

* CSV files
* Output of ETL
* Output after preprocessed data and feature creation

### Technology Choice

* Object-store Watson in this study
* Firebase object on Google Cloud for deployment

### Justification

* Since the report will be run and publish for marking on Watson Studio

## Discovery and Exploration

### Technology Choice

* **Pandas** for data manipulate for easy to use
* **Seaborn** for data visualization due to its popularity and easy to use but powerful with 1 line of code

### Justification

**Why I have chosen a specific method for feature engineering?**

* Compose datetime column for our LTSM model
* Remove missing values in account management activities (explained above)
* **Feature creation:** We will create 3 new features for our analysis
  + % service= total account activities/number of total accounts

The reason we created this feature due to we want to reduce the impact of account growth during this period. For example: from 2000-2014 total number of accounts grow from 2000-11000 accounts so total account manager activities also increase but % account get service might not follow.

* + Revenue per account – same explain above
  + Velocity of trading= total opening balance/ total trading volume

This feature will help to explain how many times money flow between the exchanges. The more they circulate the better with same money supply

## Actionable Insights

### Technology Choice

* Python for programming language
* Pandas for data exploration
* Seaborn for data visualation
* Sklearn Keras for deep learning
* LSTM for alternative model as we are dealing with time series data

### Justification

* All above tools are easy to use and popular
* Can find solution to any problem by google/ stackoverflow

**Why I have chosen a specific method for data quality assessment?**

* Calculate missing value NaN from 1995-2000 (25% of total rows)

However company doesn’t have any system back then so data is unavailable

Also first 5 years was a startup company so most number are outliers so we decide to exclude from our dataset. Only take data from 2000-2019

* Duplicated number of total rows-> need to remove
* Datetime is in wrong format. We will have to rebuild datetime feature.
* In general, due to datasource come from enterprise dashboard it is accurate and reliable.
* Interest rate download from different source so it will need to reformat datatime in order to merge to our main dataframe,

**Why I have chosen a specific algorithm?**

* We will try with a basic linear regression first due to our quick plot show a good linear relationship.
* Then we will conduct a deep learning model with Keras
* Lastly we will try with LSTM due to the nature of time series data and also sometimes a lot of variable are not controllable except for account manager’s activities – but in our analysis it shows a weak relationship with revenue (coeff=<0.5)

## Applications / Data Products

### Technology Choice

* Google Cloud Platform
* Firebase storage

### Justification

* To deploy the model as API service and easier to access

## Security, Information Governance and Systems Management

### Technology Choice

* As data source comes belongs to enterprise. Some data will be adjusted/ renamed
* All will set to private sharing mode

### Justification

* It is confidential

**Why I have chosen a specific method for data quality assessment?**

* Calculate missing value NaN from 1995-2000 (25% of total rows)

However company doesn’t have any system back then so data is unavailable

Also first 5 years was a startup company so most number are outliers so we decide to exclude from our dataset. Only take data from 2000-2019

* Duplicated number of total rows-> need to remove
* Datetime is in wrong format. We will have to rebuild datetime feature.
* In general, due to datasource come from enterprise dashboard it is accurate and reliable.
* Interest rate download from different source so it will need to reformat datatime in order to merge to our main dataframe,

**Why I have chosen a specific method for feature engineering?**

* Compose datetime column for our LTSM model
* Remove missing values in account management activities (explained above)
* **Feature creation:** We will create 3 new features for our analysis
  + % service= total account activities/number of total accounts

The reason we created this feature due to we want to reduce the impact of account growth during this period. For example: from 2000-2014 total number of accounts grow from 2000-11000 accounts so total account manager activities also increase but % account get service might not follow.

* + Revenue per account – same explain above
  + Velocity of trading= total opening balance/ total trading volume

This feature will help to explain how many times money flow between the exchanges. The more they circulate the better with same money supply

**Why I have chosen a specific algorithm?**

* We will try with a basic linear regression first due to our quick plot show a good linear relationship.
* Then we will conduct a deep learning model with Keras
* Lastly we will try with LSTM due to the nature of time series data and also sometimes a lot of variable are not controllable except for account manager’s activities – but in our analysis it shows a weak relationship with revenue (coeff=<0.5)

**Why I have chosen a specific framework?**

* **Pandas** for data manipulate for easy to use
* **Seaborn** for data visualization due to its popularity and easy to use but powerful with 1 line of code
* **Tensorflow** for ML as it is open source and support both regression and neural networks
* **Keras** for deep learning due to its simplicity interface. Also it provide rapid prototyping and also work with TensorFlow

**Why I have chosen a specific model performance indicator?**

* We will use MAE and MSE to measure our model as they are the most common metrics used to measure accuracy for continuous variable

**USE CASE:**

Company B is a B2B exchange found in 1995 where clients can trade with each other without the need of cash. Company B will act as a third party ledger recorder and also a marketing channel to promote members. Company B will charge a % of each transaction as success fee and this will be Company B’s revenue.

Company B employs a number of account managers to look after these accounts and facilitate trading as the mode trading -> the more profit.

Account managers KPI are based on service call, service visit or annual review of each account.

Recently our director is interested to find out if we could predict our revenue based on account managers’ input as some of his questions below:

1. Are account manager’s activities a good indicator for revenue?

2. Is floating interest a good indicator for revenue?

3. Does transaction fee impact revenue?

4. Can we find any good predictor for number of Closures using account manager activities?

Data of most variables: total accounts, transactions, volume, account manager activities … are available from company portal as CSV and can be download to process. However data are formatted in presentation ready form so it will need to be processed and converted to proper format for each value.