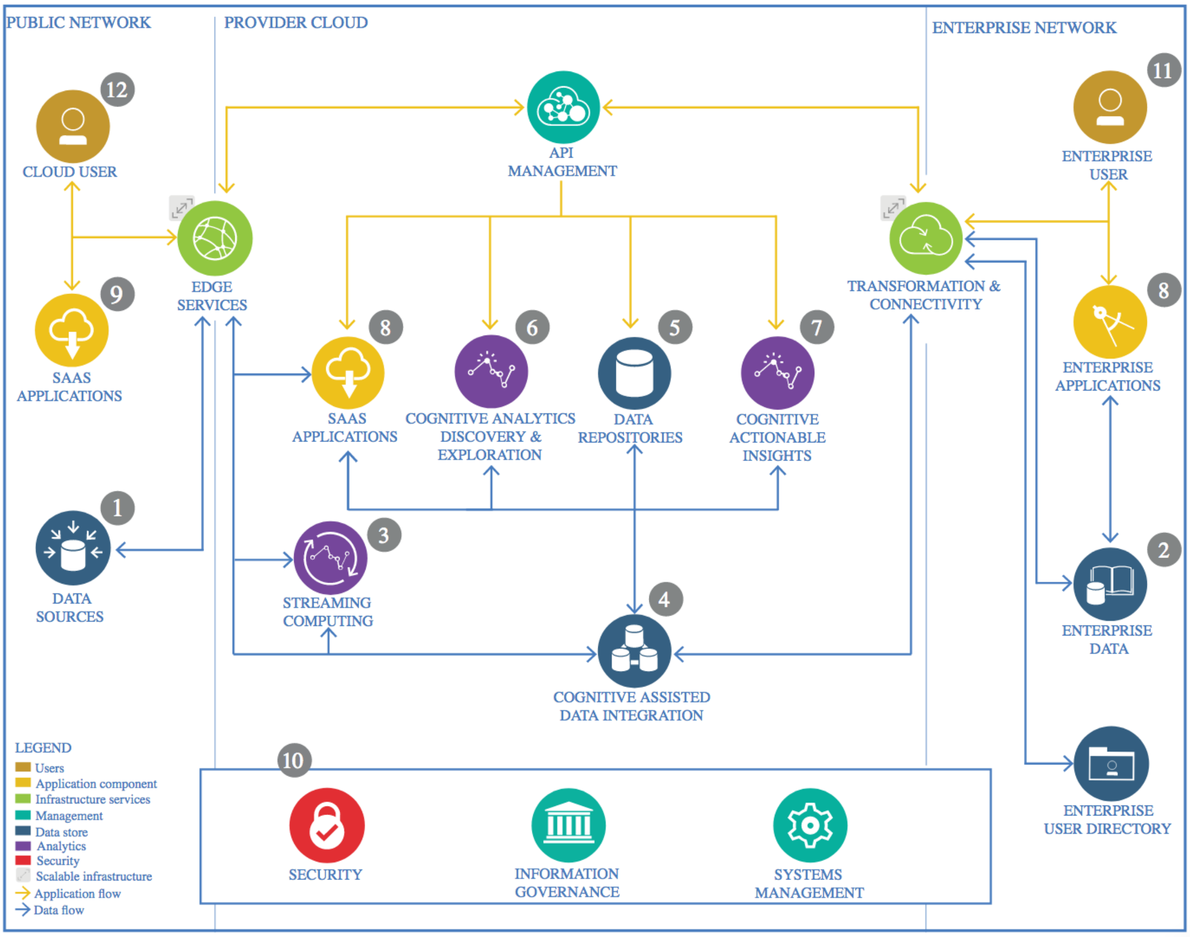
Architectural Decisions Document
SARS-CoV-2

1 Architectural Components Overview



1.1 Data Source

1.1.1 Technology Choice

The data source is a repository from github : https://github.com/CSSEGISandData/COVID-19. It gets updated on a daily basis which makes it an ideal choice

1.1.2 Justification

The data repository is owned and operated by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) who are constantly working towards the cause i.e. analysis of COVID-19 and is deemed as a trustable source

1.2 Enterprise Data

1.2.1 Technology Choice

This technology is not currently included in this data science project.

1.2.2 Justification

As of now the data is collected from the GitHub repository operated by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE).

1.3 Streaming analytics

1.3.1 Technology Choice

This technology is not currently included in this data science project.

1.3.2 Justification

As the data update frequency is every day so right now there is no real-time analysis done. The models can be run once per day to get the updated results.

1.4 Data Integration

1.4.1 Technology Choice

This technology is not currently included in this data science project.

1.4.2 Justification

Adding new data to the original dataset is done by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) daily.

1.5 Data Repository

1.5.1 Technology Choice

Data repository used is on GitHub - https://github.com/CSSEGISandData/COVID-19.

1.5.2 Justification

The data repository, owned and operated by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), has a good level of verbosity which is good for providing flexibility in the analysis.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Correlation between the features is used as data exploration. Also, many other modifications like – renaming , pruning, filling empty records is done after being discovered

1.6.2 Justification

Correlation is strong approach to draw out a relation between the various features which helps us to identify the features which are significant and contains the most information

1.7 Actionable Insights

1.7.1 Technology Choice

Data preprocessing – Scaling , Normalization, Data Visualizations and Performance of the model is enough to get some insights and perform action accordingly

1.7.2 Justification

After reviewing the dataset we must performing appropriate steps to get the most out of the data.

1.8 Applications / Data Products

1.8.1 Technology Choice

Produce visually interactive plots as per the available libraries. Use the trained machine learning / deep learning model for performing tasks (e.g. Prediction, Projection etc.) as per the requirements. No products are produced

1.8.2 Justification

The project runs on a notebook which cannot be considered as a product. Also, the application has its limits

1.9 Security, Information Governance and Systems Management

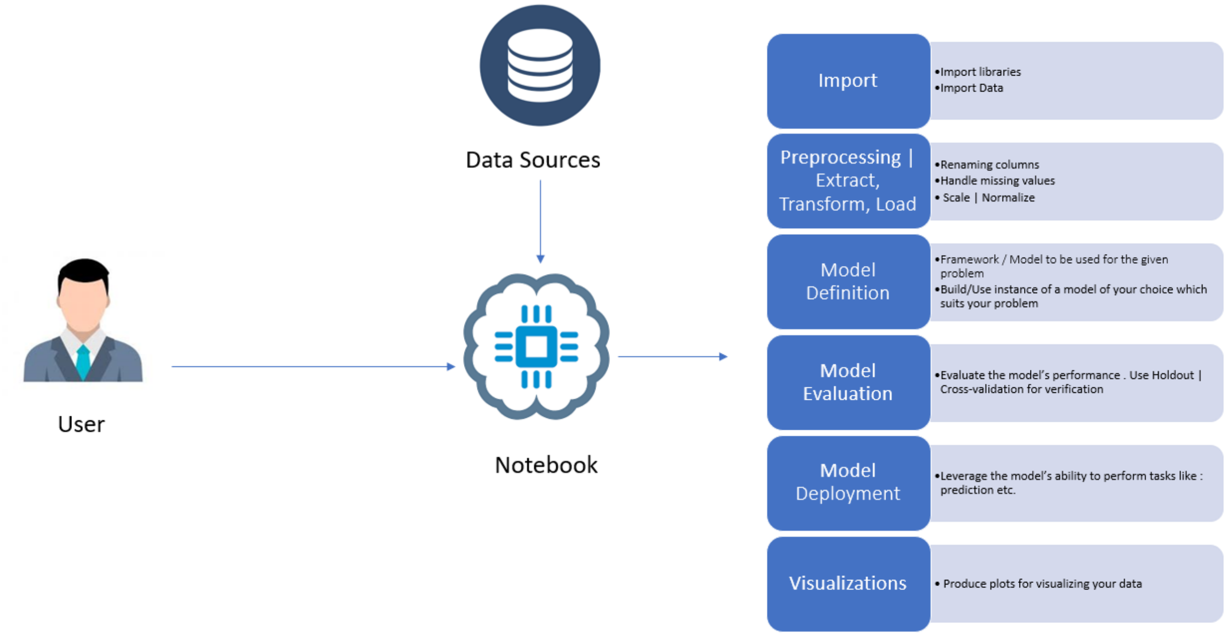
1.9.1 Technology Choice

This technology is not currently included in this data science project.

1.9.2 Justification

Currently the project runs on a jupyter notebook so no security, Information Governance and Systems Management is present.

2 Architecture components specific to my current project



Import
- Import libraries
- Import Data

Preprocessing | Extract, Transform, Load
- Renaming columns
- Handle missing values
- Scale | Normalize

Model Definition
- Framework / Model to be used for the given problem
- Build/Use instance of a model of your choice which suits your problem

Model Evaluation
- Evaluate the model's performance . Use Holdout | Cross-validation for verification

Model Deployment
- Leverage the model's ability to perform tasks like : prediction etc.

Visualizations
- Produce plots for visualizing your data

2.1 Data Import

2.1.1 Technology Choice

The data source is a repository from github : https://github.com/CSSEGISandData/COVID-19. It gets updated on a daily basis which makes it an ideal choice

Import the required libraries and packages for analyzing the data

2.1.2 Justification

The data repository is owned and operated by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) who are constantly working towards the cause i.e. analysis of COVID-19 and is deemed as a trustable source

Data analysis is a resource intensive task and should be done properly to gather information from the data

2.2 Pre-processing

2.2.1 Technology Choice

Data preprocessing – Scaling , Normalization, Data Visualizations and Performance of the model is enough to get some insights and perform action accordingly

Correlation between the features is used as data exploration. Also, many other modifications like – renaming , pruning, filling empty records is done after being discovered

2.2.2 Justification

Correlation is strong approach to draw out a relation between the various features which helps us to identify the features which are significant and contains the most information

After reviewing the dataset we must performing appropriate steps to get the most out of the data

2.3 Model Definition

2.3.1 Technology Choice

Framework Used : Keras
Machine Learning Model used : Support Vector Regressor and polynomial regression

2.3.2 Justification

The dataset I am using is highly dimensional which is suitable for models like Support Vector Machines & Polynomial Regression.

SVM is famous for efficiently performing a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Also, the dataset which I am using grows daily making SVM and polynomial regression an ideal choice as they scale relatively well in these cases.

Keras a high-level deep learning framework that sits on top of backend frameworks like TensorFlow.

Keras is excellent because it allows you to experiment with different neural-nets with great speed! It sits atop other excellent frameworks like TensorFlow and lends well to the experienced as well as to novice data scientists! It doesn’t require nearly as much code to get up and running!

Keras provides you with the flexibility to build all types of architectures; that could be recurrent neural networks, convolutional neural networks, simple neural networks, deep neural networks, etc.

2.4 Model Evaluation

2.4.1 Technology Choice

Holdout and Cross Validation

2.4.2 Justification

Simple holdout with stratification is performed for the regression model as the data is highly dimensional and Cross validation takes much time

Cross validation is performed while determining the best parameters for the machine learning model as only a sample of features are considered

2.5 Model Deployment

2.5.1 Technology Choice

The deep learning model is exported which can be loaded and used whenever required. Predictions are performed using the deep learning models after reaching a satisfying level of accuracy

2.5.2 Justification

Training a model takes a lot of time so its better to export and reuse again provided the data does not change too much in the future. Leveraging the model to predict and then test against the real (observed) data is always a good way to keep the model in check

2.6 Visualizations

2.6.1 Technology Choice

In my case, various visualizations for data , features, models, predictions etc. is drawn using a no. of packages (Plotly, folium etc.). These involve figures like – Line charts, Pie charts , World Map, Heatmaps etc.

2.6.2 Justification

Visualizations should be strong and as much detailed as possible for grasping the most knowledge of the data. So, its better to use well-defined packages instead of the conventional packages for getting the best visuals.