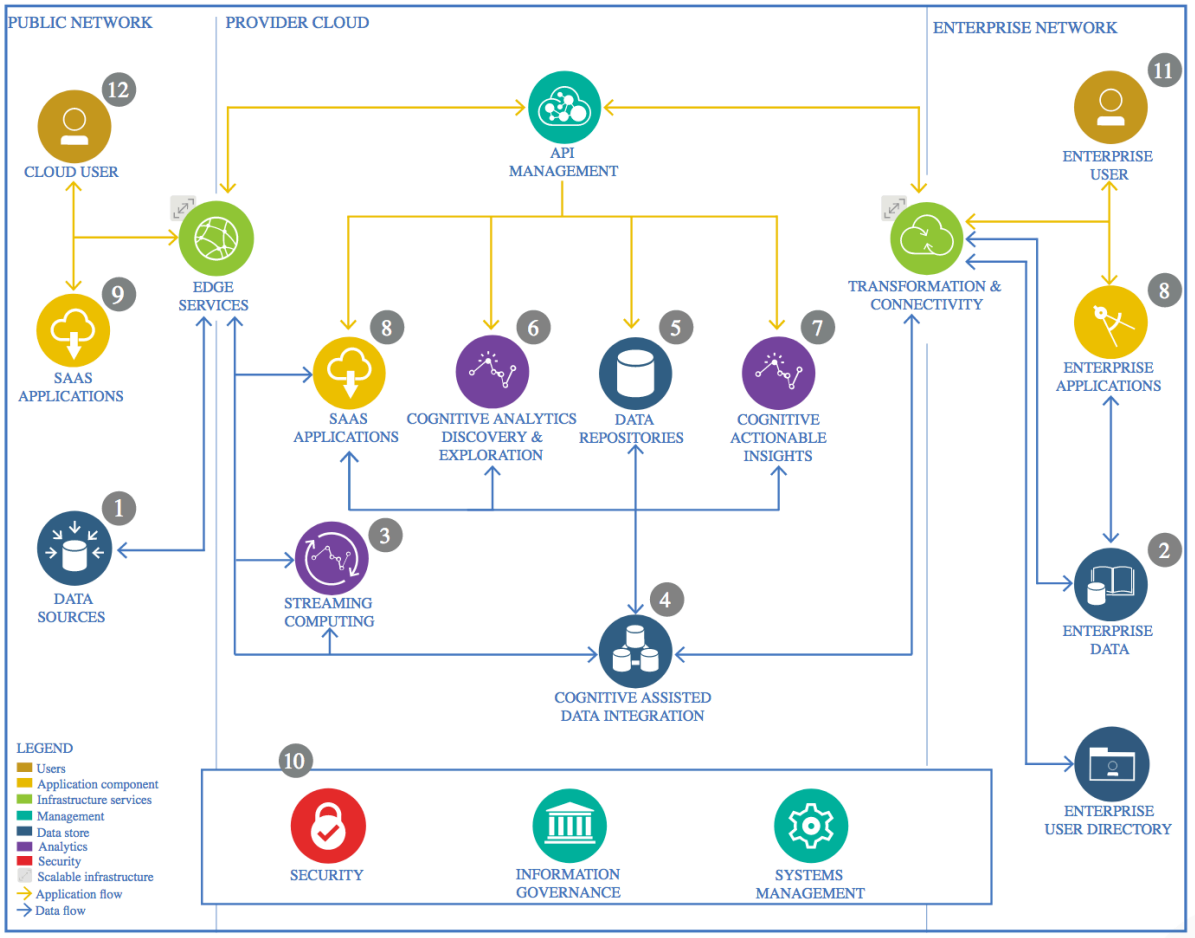
The Lightweight IBM Cloud Garage Method for Data Science

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

My Dataset is Novel Corona Virus 2019 Dataset collected from kaggle.com .This dataset has daily level information on the number of affected cases, deaths and recovery from 2019

novel coronavirus.Dataset is provided by Sudalai Rajkumar (Data Scientist | Kaggle Grandmaster) https://www.kaggle.com/sudalairajkumar.

1.1.2. Justification

Dataset is initially managed byJohns Hopkins Github repository . Sudalai Rajkumar was able to provide an organized and structured dataset for my need .Accurate data is available from 22 Jan, 2020 to 04 June ,2020(updated every other day..) which helped for amore accurate prediction .

1.2. Enterprise Data

1.2.1. Technology Choice

NAN
Data collected was already Enterprises from various sources by Johns Hopkins Github repository and later reformatted and arranged by Sudalai Rajkumar to CSV files.

1.2.2. Justification

The data required for this project was already collected from various sources .

1.3. Streaming analytics

1.3.1. Technology Choice

Once the Stream data is available in the root drive the model can be re performed for further prediction. Technologies uses for forecasting Stream data

- Holt's Linear Model
- AR Model prediction
- Facebook Prophet's predicton

1.3.2. Justification

Of the models I tested these forecasting models performed better for small-moderate Pandemic datasets.

1.4. Data Integration

1.4.1. Technology Choice

Input Dataset is contained in CSV file which is integrated into panadas Dataframe using pandas read_csv method.Temporary data storage is also performed using CSV files stored using pandas to_csv methord

- Pandas dataFrame (CSV)

1.4.2. Justification

As the Dataset size is small CSV was a better option than other forms such as Parquet, binary, etc..
Pandas CSV methods provides a better integration with Numpy arrays for better data processing and conversion.

1.5. Data Repository

1.5.1. Technology Choice

For storing training data and temporary data I used

- IBM Cloud Storage

1.5.2. Justification

The storage options on IBM Cloud™ offer inherent scalability and flexibility.
The storage options on IBM Cloud are easier to provision, deploy and access. This flexibility helps relieve the headaches caused when managing large amounts of disparate types of storage locally. Select from object, file, and block storage services, as well as cloud data migration options.
It is more easy to access data while working with IBM Watson Studio.

1.6. Discovery and Exploration

1.6.1. Technology Choice

Data Discovery and Exploration is Done using following technologies

- Pandas Dataframes
- Numpy arrays
- Sklearn libraries
- Datetime Libraries
- Plotly
- Matplotlib
- Statsmodels
- Pmdarima
- Facebook Prophet

1.6.2. Justification

- Pandas data frame provide methods to access CSV files manage and format them based on need.It includes native functions to provide summary of data (.Info() , .describe()) and methods to manage corrupted or missing data for evaluation.
- Numpy array provides native methods for fast and flexible multi-dimensional data processing for different type of data.

- sklearn is a Python module integrating classical machine learning algorithms in the tightly-knit world of scientific Python packages (numpy, scipy, matplotlib).
  It provide simple and efficient solutions to learning problems that are accessible to everybody and reusable in various contexts: machine-learning as a versatile tool for science and engineering.

- The datetime module includes functions and classes for doing date and time parsing, formatting, and arithmetic

- The plotly Python library (plotly.py) is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

- Pmdarima is a Python & Cython wrapper of several different statistical and machine learning libraries (statsmodels and scikit-learn), and operates by generalizing all ARIMA models into a single class.

- Prophet is an open source library published by Facebook that is based on decomposable (trend+seasonality+holidays) models. It provides us with the ability to make time series predictions with good accuracy using simple intuitive parameters .

1.7. Actionable Insights

1.7.1. Technology Choices

Insights were obtained with the help of graphs and correlation matrix
- Mathplotlib
- Plotly
- Numpy
- Statsmodels
- Pmdarima
- Facebook Prophet

1.7.2. Justification

By visualizing the output we were able to determine the parameters which are more closely correlated and contributed to the spread of the virus.
I was able to determine high correlation between :

- Confirmed cases and recovery suggesting lower mortality rate
- Tourists were affected in regions of higher population densities
- People with median age were more prone to affect
- From prediction graphs confirmed an exponential growth if not acted efficienty
- Cross 3 lakh as early as 12th of June
- The doubling rate suggests that the number of cases is expected to be doubled in 10-15 days

1.8. Applications / Data Products

1.8.1. Technology Choice

The output of the notebooks are graphical representation of increase or decrease in the overall total number of cases in India for the conning days.

- Pandas data frame prediction for coming days .
- Plots of increase or decrease in the overall affected people.
- Provides a margin in which data is expected to fall using graphs.

Note: The number of days can be adjusted but the accuracy decreases as the forecasting moves much farther off than trained data.

1.8.2. Justification

Healthcare organizations are in an urgent need for decision-making technologies to handle this virus and help them in getting proper suggestions in real-time to avoid its spread. AI works in a proficient way to mimic like human intelligence. It may also play a vital role in understanding and suggesting the development of a vaccine for COVID-19. This result-driven technology is used for proper screening, analyzing, prediction and tracking of current patients and likely future patients. The significant applications are applied to tracks data of confirmed, recovered and death cases.