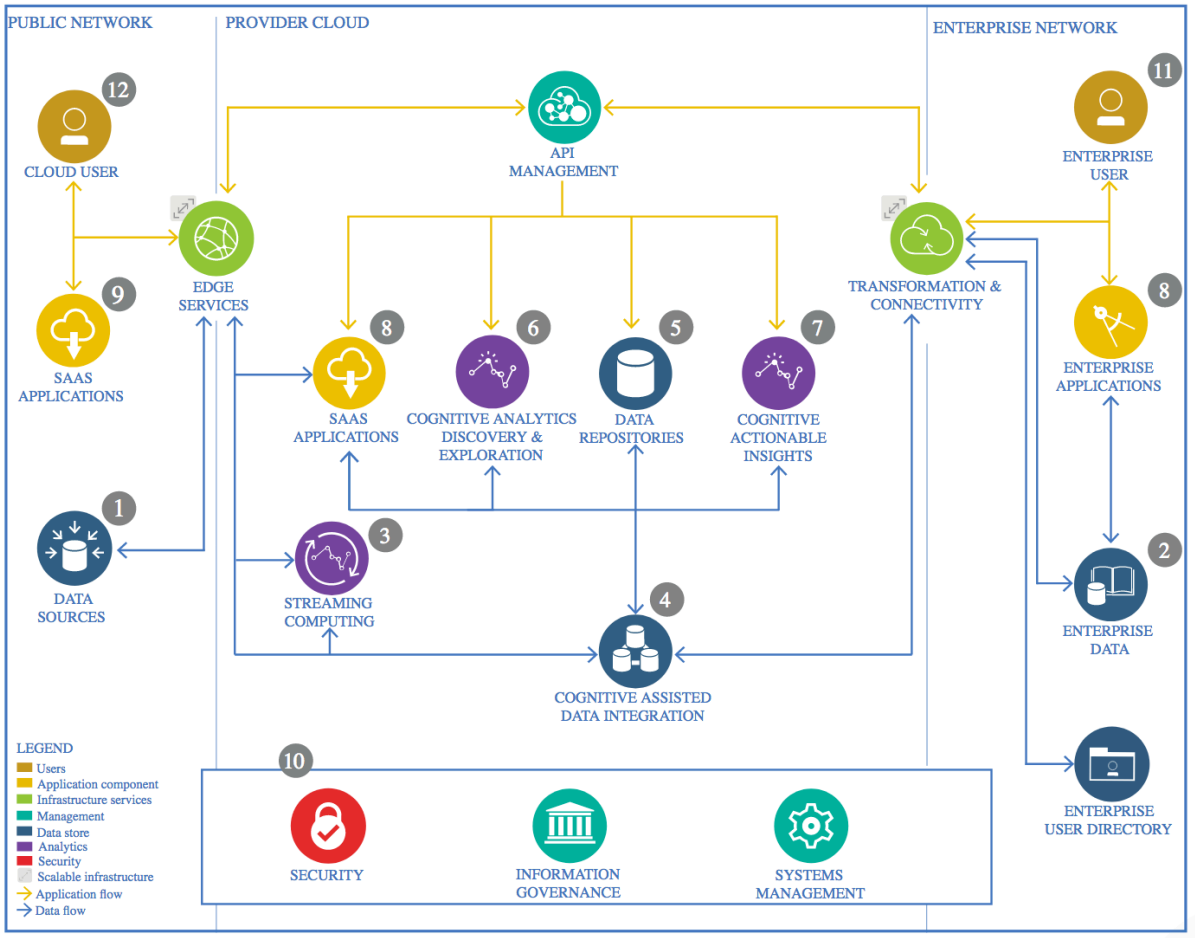
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1. Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Covid19 India EDA and Forecasting

1.1. Data Source

Ministry of Health & Family Welfare (https://www.mohfw.gov.in/)

covid19india API (https://api.covid19india.org/raw_data.json)

Acknowledgements

Thanks to Indian [Ministry of Health & Family Welfare] (https://www.mohfw.gov.in/) for making the data available to general public.

Thanks to [covid19india.org] (https://covid19india.org) for making the individual level details and testing details available to general public.

1.1.1. Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

1. Python Request, JSON modules
2. Python EDA and time series analysis libraries - Pandas, Numpy, Stats, Folium, Plotly, SARIMAX, Prophet

1.1.2. Justification

The data gets updated daily with cases, tests details. Hence API based approach to extract data. Also the data size is in limit to be handled with python. Hence Python (instead of Spark)

1.2. Enterprise Data

1.2.1. Technology Choice

Python based Data Extraction Utility

1.2.2. Justification

An independent python utility to extract the data from trusted sources is built. This idea is to have data extraction layer scalable enough to be configured with various data sources

1.3. Streaming analytics - NA

1.3.1. Technology Choice

Currently the data gets updates twice a data at the sources and data extraction is capable of extracting data in batch mode.

1.3.2. Justification

Streaming is not the idle case as source is not realtime

1.4. Data Integration

1.4.1. Technology Choice

Batch-wise data ingestion at separate directory. This is used to read the data for analysis.

1.4.2. Justification

Data Extraction updates the data in separate repository. This can be moved to Cloud based Object Stores in future. The analysis refers the data from this directory

1.5. Data Repository

1.5.1. Technology Choice

Python API to fetch and push data in data repository

1.5.2. Justification

Data gets updated daily basis hence an api based mechanism is used to pull data and push it in data repository for further analysis.

1.6. Discovery and Exploration

1.6.1. Technology Choice

Python libraries like Panda, Numpy, Seaborn, Plotly, Folium are used

1.6.2. Justification

The size of the dataset was the key factor in deciding data exploration tools. The current data small enough to be processed on a single computer ruling out the need for distributed processing (Spark, Pyspark)

1.7. Actionable Insights

1.7.1. Technology Choice

TimeSeries libraries and Prophet based forecasting

1.7.2. Justification

To understand the pattern and growth of Covid 19, a through analysis using TimeSeries and Prophet based models was conducted along with various visualisations to understand the pandemic spread. Forecast models were built to project the COVID-19 cases

1.8. Applications / Data Products

1.8.1. Technology Choice

Data visualisation libraries , TimeSeries libraries and Prophet based forecasting

1.8.2. Justification

A Jupiter book and Forecast models were built to project the COVID-19 cases. This can be deployed as an API solution in future.

1.9. Security, Information Governance and Systems Management

1.9.1. Technology Choice

None.

1.9.2. Justification

Data is captured from trusted source and is available for public use. Also model is but for educational purpose.