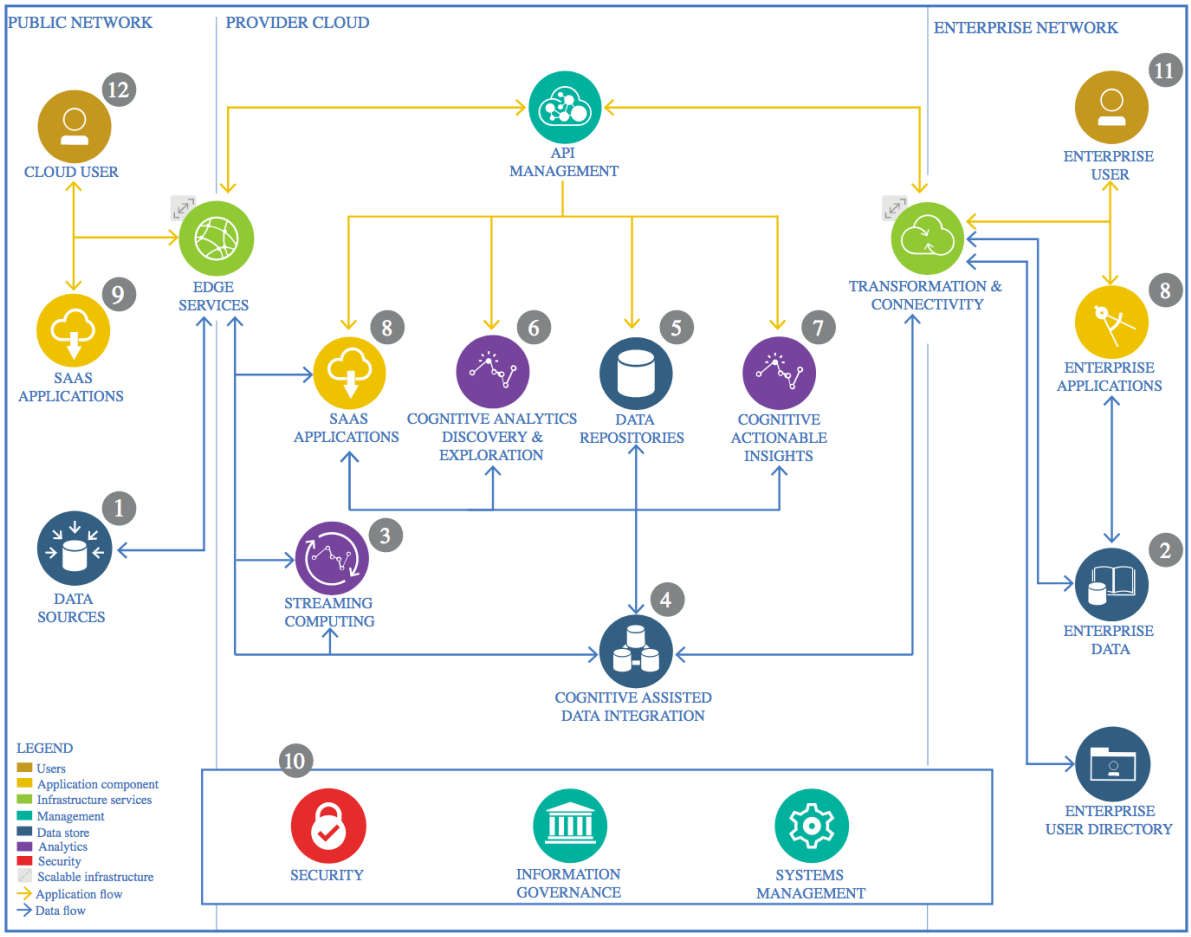
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

Mexican government – Ministry of Health
https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

1.1.1 Technology Choice

The database used has a daily update and needs to download continuously. For this case, the information is collected using “wget” command in Python. This command helps to retrieve information directly from the internet and stored into specific location.

1.1.2 Justification

As a matter of fact, the CoVid information is updated daily and for that reason it is necessary to create a specific process to collect the information and update continuously.
This information can be used to report in “real-time” the contagious patter of this hazardous disease.

1.2 Enterprise Data

1.2.1 Technology Choice

Not used or apply

1.2.2 Justification

We are using python and wget command to extract the information. Also, there are some python functions used to verify the data as well as to erase the old information in the system to avoid volume space consumption.

1.3 Streaming analytics

1.3.1 Technology Choice

Not used or apply

1.3.2 Justification

The information is updated daily. Furthermore, this project does not used any “real-time” information to work. The information is collected daily morning.

1.4 Data Integration

1.4.1 Technology Choice

The database is stored into a separate folder. This folder is checked continuously to verify the last information available.

1.4.2 Justification

There are two different sources for the data. The first is the daily database with all the information about the CoVID cases. The other file is a descriptor table with the different fields description to enrich the database.
The Mexican government decide to split the information in this form to reduce the volume size information.

1.5 Data Repository

1.5.1 Technology Choice

The CoVID information is in special open source platform to be downloaded by anyone. This project will stored into GitHub.

1.5.2 Justification

GitHub is very useful to control the versions and track the changes in the project. For that reason, the project is clearly keep into github and check continuously to analyze the information. On the other hand, it helps to create a user control to verify who changes or modify the project.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Python libraries such as Pandas, numpy, seaborn and matplotlib will be used in the project

1.6.2 Justification

The database size was the key factor to decide use data exploration tools. On the other hand, the process can perform into a computer, but it will consume lot of time and resources. However, there are new cloud technologies that help us with this problem such as spark.

1.7 Actionable Insights

1.7.1 Technology Choice

Timeseries and forecasting

1.7.2 Justification

The growth pattern can study and analyze using time series. In this brief study, the model will try to determine the principal factors to get sick. Nevertheless, there are many different variables that this study will not include such as the exposure in crowded events. On the other hand, this study should try to focus in several elements that can increase the probability to sick.

1.8 Applications / Data Products

1.8.1 Technology Choice

The information can easily display in many different formats. Nevertheless, the most used is power point slides.

1.8.2 Justification

There are many different visualization tools such as Qlik Sense or Tableau, but these tools are not free and most of the time needs to connect to a server to display the dashboard. For that reason, we can use matplotlib and seaborn in python to construct the graphics and displayed in power point. Maybe, the graphics will not be dynamic, but the information can show to anyone and also in the news.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

The CoVID database is uploaded by the Mexican government daily. Furthermore, this information needs to be downloaded by a python process once a day. It is preferably to do that in the morning and analyze the information.

1.9.2 Justification

Every day, the Mexican government collect the information from all the hospitals and keep into a database. This database is uploaded to an open source repository. The information can check by anyone to verify the numbers that are daily shown in the news.