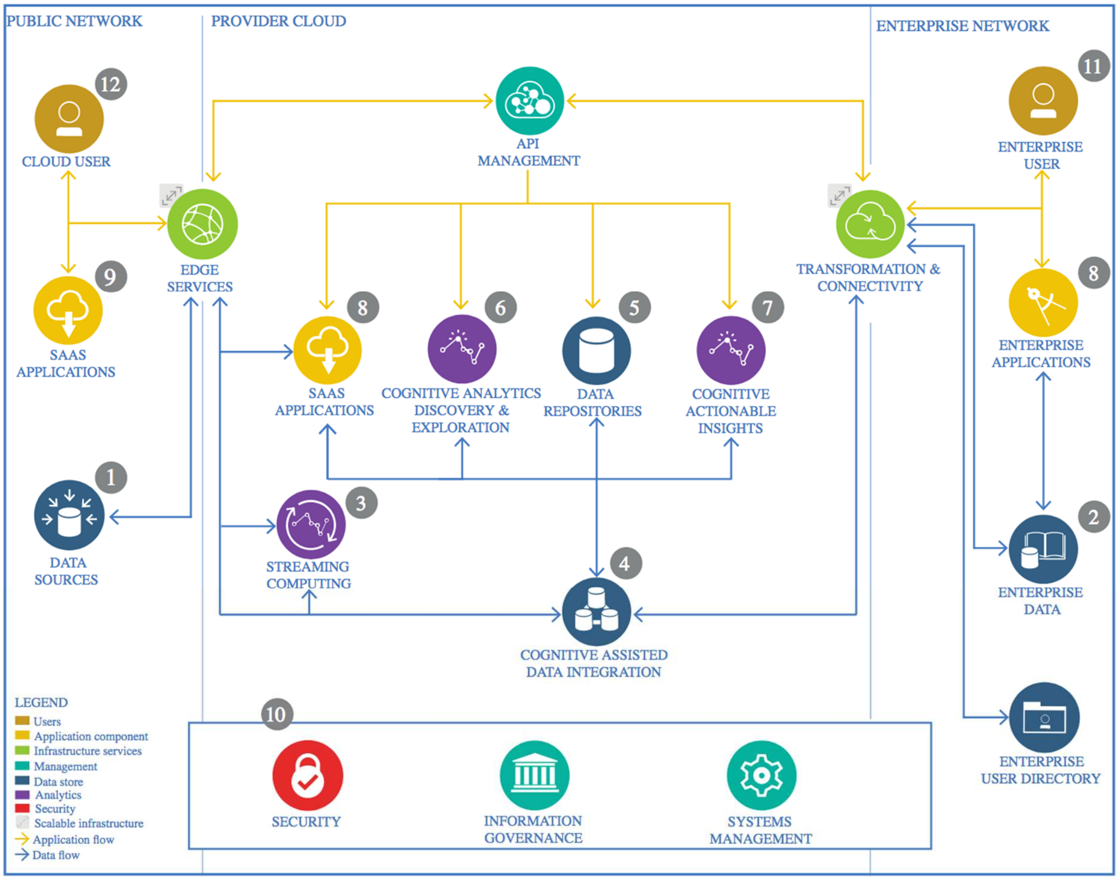
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Problem Statement:
Segretation/Classification of the images received on WhatsApp. It is really becoming tedious to manage several unwanted images received and which are getting backed up frequently. To resolve this problem, this solution will help in segregating the images into multiple types such as

1. Images with single person
2. Images with two persons (Couple)
3. Images with more than two people (Group)
4. Wishes/quotes

1.1 Data Source

Data source are the images received on whatsapp. All these are getting backed up onto the Onedrive/Google drive. For the project purpose, shared 1000+ images (training) and 200+ images for validation through github. Both the databases are labeled and useful for supervised learning.

1.2 Enterprise Data

Backup frequency of my whatsapp data is weekly. Every week this data will be backed up into the Google Drive.

1.3 Streaming analytics

Not needed. As this application is off line, it won’t require streaming analytics.

1.4 Data Integration

Multiple technologies were used in this project. As this is a image classification project used Keras predominantly to develop a deep learning model.

Following were used:
Python
Apache Spark
Keras
Pandas dataframe
OpenCV

1.5 Data Repository

As the total training data size is small copied the training/validation data from github and performed the training by using Data Flow generator in Keras.

1.6 Discovery and Exploration

Images were viewed using OpenCV, matplotlib etc. to understand how the images were preprocessed for deep learning. Initially identified training data set was not balanced i.e. some categories were having more images. To rectify this problem duplicated and created multiple variants to build a balanced training dataset. This is critical for deep learning training. Otherwise model will be biased to few classes.

Deep learning model, explored multiple activation functions, learning rates and optimizers. Specifically, for this problem, RMS optimizer is not stable enough. Whereas SGD (Stochastic Gradient Descent) is relatively stable in terms of loss and accuracy.

1.7 Actionable Insights

Learning rate, batch size has the impact on model training speed. LeakyRelU provided better accuracies compared to other activation functions. Increasing the model size requires more training data. With the current data available for training model size is looking optimal and provided classification accuracy up to 85%.
To improve accuracy further, aspect ratio of the images needs to be managed better. At this point of time used the image pre-processing steps of the Keras ImageDataGenerator, but these needs to be better understood and modify such that the input images aspect ratios are kept intact.

Once the required accuracies achieved using this model, would like to delete all the images belongs to wishes category. Only keep few photos from the group category. As in WhatsApp for each group/party several photos are shared and all repetitive. Keep only few photos so that my storage is optimized. Single/Double category identify the people in those photos and do the actions based on the people in those photos. Lot more work to do to achieve the above end goal. Nevertheless, first step of broad categorization is achieved through this step.

Note: Somehow each iteration is providing very different results interms of model accuracy and loss.

1.8 Applications / Data Products

To deploy this application Watson Cloud storage as a repository. But at this point of time not deployed this model as it requires some more work. As such I understood how to deploy using IBM data cloud using REST API.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

None required. All the data is personal.