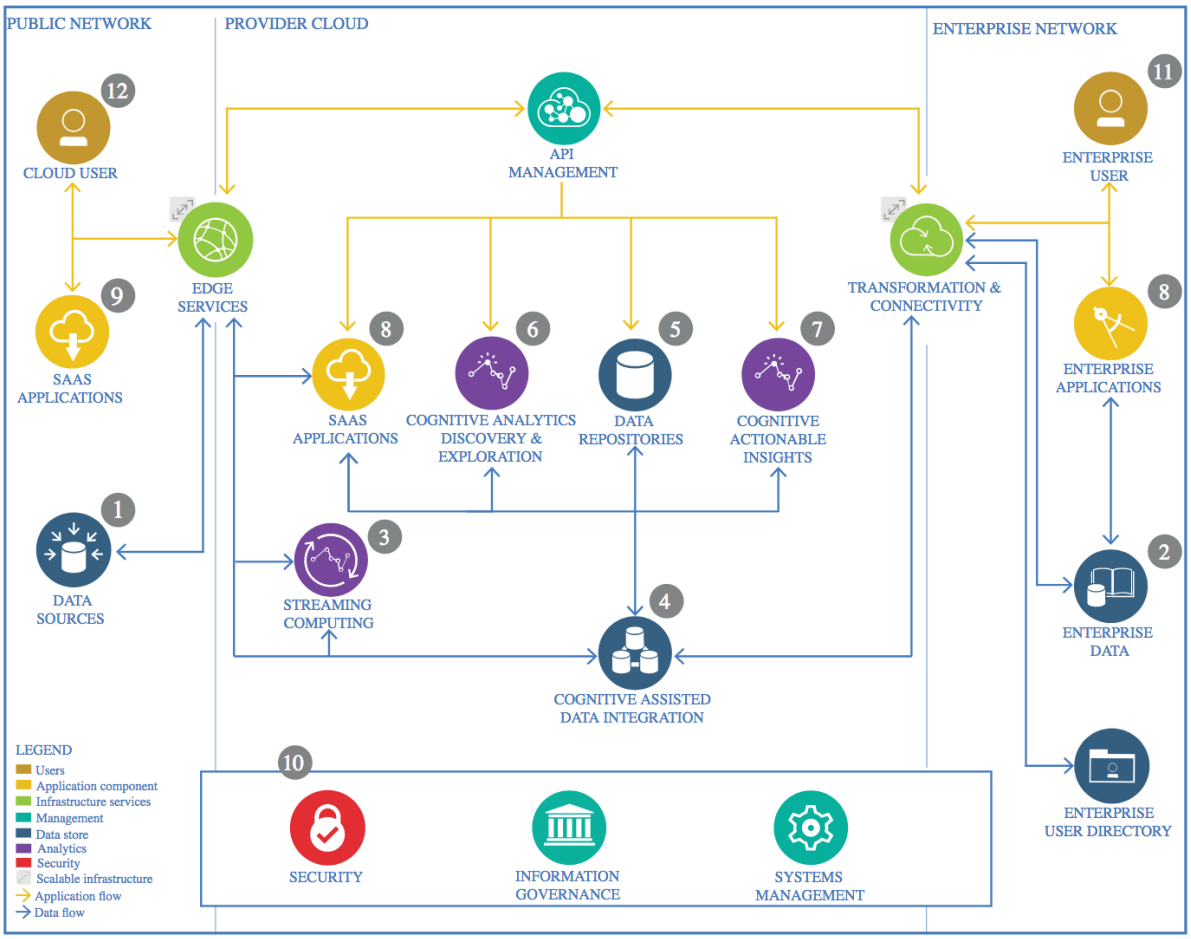
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

The source of this data comes from the following Keras dataset, it consists of images stored in two different folders “Parasitized” and “Uninfected”, these folders indicate the label of the images that are contained within.

1.1.1 Technology Choice

The path to each image is extracted by glob and stored in a pandas DataFrame.

1.1.2 Justification

The nature of the dataset, made it so that dynamically exploring the folders and getting the path to each image while storing their label (from which folder were they extracted) was a fairly simple way on preparing the images to be processed.

1.2 Enterprise Data

1.2.1 Technology Choice

No technology was employed for working with Enterprise Data.

1.2.2 Justification

Since the dataset is stored as local assets there is no need for this component.

1.3 Streaming analytics

1.3.1 Technology Choice

No technology was employed for Streaming data.

1.3.2 Justification

Training, validation and testing are all made by batch processing images from a Keras.ImageDataGenerator, rather than stream data in real time, it would make sense to classify batches of data, after they are obtained from the blood smears that were captured by microscopes.

1.4 Data Integration

1.4.1 Technology Choice

The pandas DataFrame that contains the path to each labeled image is processed by reading and resizing the images (to 125x125px) with the opencv-python library, they are stored in a numpy array.

1.4.2 Justification

The OpenCV API is a fairly popular Computer Vision Library, employing this tools allows for reading the images to memory and resizing the differently sized images to a standard.

1.5 Data Repository

1.5.1 Technology Choice

Aside from storing the local dataset assets in the hard drive or the IBM Cloud, no technology was used to store the processed Dataset (after resizing the images).

1.5.2 Justification

Since the size of the dataset isn't massive (353MB), storing it directly in the Assets of the IBM Cloud was a reasonable decision, instead of using a more complex form of storage.

1.6 Discovery and Exploration

1.6.1 Technology Choice

By using matplotlib we were able to show samples of the dataset after reading the images, show how data was augmented by applying transformations to the images, and plotting the training results for the model. Jupyter also facilitates displaying data neatly through printing pandas dataframes (such as the evaluation metrics of the dataset).

1.6.2 Justification

All these technologies are part of the standard toolkit of the python data scientist, they more than suffice for displaying the relevant information for this dataset and the model that classifies it.

1.7 Actionable Insights

1.7.1 Technology Choice

The model was built with Keras Sequential API, one model is a Convolutional Neural Network built from scratch, whereas the other models are variations on a Neural Network that applies transfer learning by employing the Inception V3 pre-trained network with the imagenet dataset as backbone network.

1.7.2 Justification

Keras very simple API allows for the creation of complex models with a really low amount of lines of code, and most of the complexity of the model is handled by Keras, although, this doesn't normally provides the best and most performant models, they are more than sufficient for a Proof of Concept.

1.8 Applications / Data Products

1.8.1 Technology Choice

No technology was employed for working with Data Products.

1.8.2 Justification

Predicting newly extracted images was out of the scope of this project, since the logic of extracting images from the blood smears captured with microscopes is of higher complexity and has to be tailored to the business logic for the laboratory that will employ the service.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

No technology was employed for working with System Management and securing access to the information.

1.9.2 Justification

Since the dataset is public and contains no information whatsoever on data from the patient on which the blood smear was taken, access to this information is not compromised.