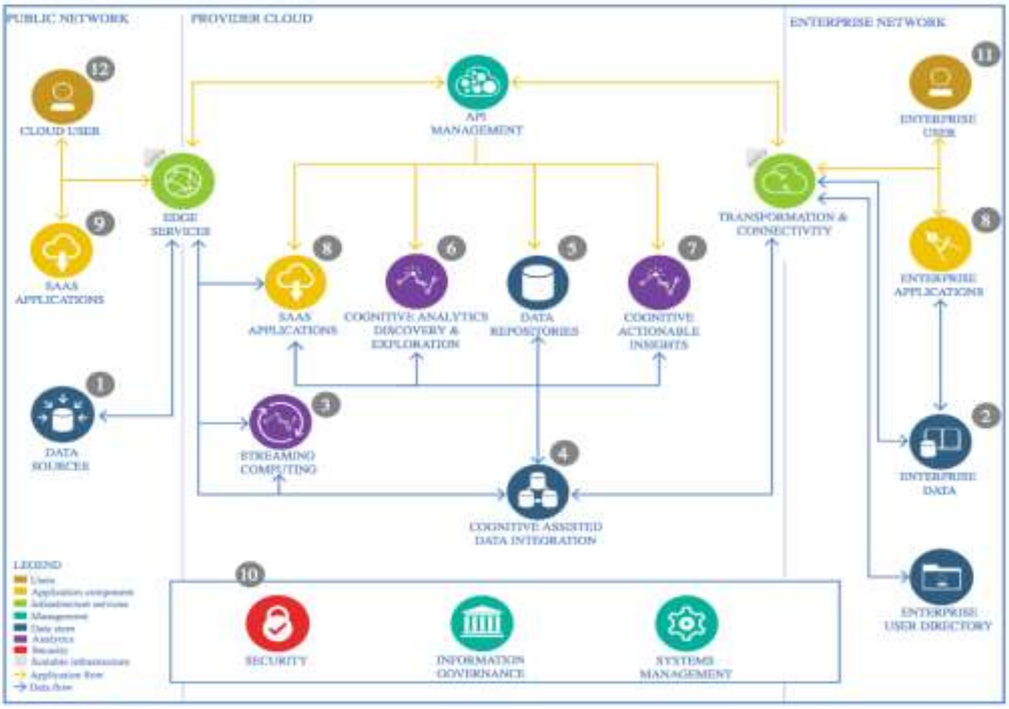
SPECIALIST
IBM
Advanced Data Science
Powered By
coursera

Architectural Decisions Document

Project Title	Predicting breast cancer patients living status after 5 years
Author	Yudy Yunardy
Date	December 2021

Architectural Components Overview

The project “Predicting breast cancer patients living status after 5 years” uses the lightweight IBM Cloud Garage Method process model. The lightweight IBM Cloud Garage Method for data science includes a process model to map individual technology components to the reference architecture. This method does not include any requirement engineering or design thinking tasks. Because it can be hard to initially define the architecture of a project, this method supports architectural changes during the process model.



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1. Data Source

Data of Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)

Targeted sequencing of 2509 primary breast tumors with 548 matched normals

Is available at http://www.cbioportal.org/study/summary?id=brca_metabric

1.1. Technology Choice

Pandas in Jupyter notebook to download the CSV data into the system.

1.2. Justification

CSV file with the results of science research. It is a typical format for the stable research data.

2. Enterprise Data

2.1. Technology Choice

N/A

2.2. Justification

N/A

3. Streaming analytics

3.1. Technology Choice

cBioPortal websites provides data on cancer genomics nearly in real-time. That means, the prediction model can be fed in almost real-time and we can have a streaming analytics and forecasting. In this project, streaming analytics is not used for simplicity. But it can be implemented at any time. For example, using IBM Streaming Analytics service.

3.2. Justification

IBM Streaming Analytics provides fast streaming application delivery using Python. Data Scientists and Developers can use existing Python code for building Streams applications without starting from scratch.

4. Data Integration

4.1. Technology Choice

Pandas and Scikit-learn. All the datasets have been downloaded to a local machine. In the ETL jupyter notebook can be seen that data is cleaned, merged and get ready for building a model. (In case of a real project or huge data, they can be loaded into a data warehouse for example IBM Object Storage)

4.2. Justification

Pandas were selected as it had a simple interface and a big community where you can find the help if you need. It is flexible and easy to use data analysis and manipulation tool. Also, the dataset size is not big, so we don't need parallelization.

Sklearn provides easy tools to input features and do further analytics.Jupyter notebooks and python are now mostly used by data scientists and it they are easy technologies to work with. That's why everything is done using python.

5. Data Repository

5.1. Technology Choice

Part of the job is done locally. So, there's a directory with all the data on local machine. Moreover, they are pushed regularly to a GitHub repository as backup. The other part of the job, which includes training, is done on cloud, specifically on IBM Watson studio. The models are then stored to IBM Object Storage, and finally downloaded to local machine.

5.2. Justification

It's easy to integrate and call from the notebook. It ensures you can save and load data on every step of your pipeline and between the notebooks, So you don't need to repeat the same steps in every notebook. It allows you to divide the process of development into structured steps and save the data on each of them.

6. Discovery and Exploration

There is a Jupyter notebook especially for EDA. In these notebooks, data is explored. The breast cancer diagnosis data is visualized using matplotlib.

6.1. Technology Choice

Pandas, Matplotlib, Seaborn

6.2. Justification

Matplotlib and Seaborn are common, and handy tools for data visualization. They got a lot of built-in functions to plot correlation matrices, histograms, scatter plots, and other useful things.

7. Actionable Insights

7.1. Technology Choice

PySpark.

7.2. Justification

In-Memory cluster computing in Spark, parallelization, speed.

8. Applications / Data Products

8.1. Technology Choice

N/A

8.2. Justification

The realization of our research as data product lies beyond the project, but the possible application, that can be based on it is the website or analytical system, that could provide a doctor help in deciding the chances of a patient, looking for similar cases, deciding type of the therapy. All of the above can be done on the base of the primary analysis, as we proved that even the result after the 5 years could be predicted based on the initial tests. Possible technology is any standard web stack, for example, MySQL + Django application.

9. Security, Information Governance and Systems Management

9.1. Technology Choice

N/A

9.2. Justification

Security should be maintained on the analytics side to avoid un-anonymizing of patients against their will. In our project, we are using already anonymized data, where each patient goes with just an ID, and we don’t need to perform any actions on the security side.