The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document : **Breast Cancer Cells**

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Identifying the problem and **Use external data sources for data enrichment**

## Data Source

Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a results of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed.

### Technology Choice

Breast Cancer dataset is available In **Kaggle.com** it is a csv file that contain 569 rows with 32 features.

### Justification

Understanding data is one of the most important part when designing any machine learning algorithm. In the notebook, I will use a data set which is available at kaggle site which is the most important community in data science

## Enterprise Data

### Technology Choice

* IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas, matplotlib

### Justification

**scikit-learn** is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.

**The Jupyter** Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

**IBM Watson Studio** provides tools for data scientists, application developers and subject matter experts to collaboratively and easily work with data to build and train models at scale. It gives you the flexibility to build models where your data resides and deploy anywhere in a hybrid environment so you can operationalize data science faster.

## Streaming analytics

### Technology Choice

This component allows for visualization and creation of metrics of data. In various process models, data visualization and exploration are one of the first steps. Similar tasks are also applied in traditional data warehousing and business intelligence.

IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas, matplotlib, **Seaborn**

### Justification

* **Seaborn** for statistical data visualization. Seaborn is a library for making attractive and informative statistical graphics in Python. It is based on matplotlib. Seaborn aims to make visualization a central part of exploring and understanding data.
* **Matplotlib** supports the widest range of possible visualizations including bar charts, run charts, histograms, box-plots and scatter plots, Whereas matplotlib needs coding skills,. For computing metrics, some code is necessary in Python
* **Watson Studio** supports sharing of jupyter notebooks, also using a fine-grained user and access management system

## Data Integration

### Technology Choice

* IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas, matplotlib

### Justification

IBM Watson Studio provides tools for data scientists, application developers and subject matter experts to collaboratively and easily work with data to build and train models at scale. It gives you the flexibility to build models where your data resides and deploy anywhere in a hybrid environment so you can operationalize data science faster..

## Data Repository

### Technology Choice

There exists an extremely huge set of technologies for persisting data. IBM Cloud Object Store is used for the data storage in the project

* IBM Watson Studio jupyter notebooks, scikit-learn, pandas, matplotlib
* IBM Cloud object storage

### Justification

There exists an extremely huge set of technologies for persisting data**. IBM Cloud Object Store** is used for the data storage in the project. It’s stable and no specific skills is required.

## Discovery and Exploration

### Technology Choice

* IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas, matplotlib

### Justification

**matplotlib** creates static plots, pixiedust supports interactive ones and matplotlib needs coding skills (I have). For computing metrics, some code is necessary in Python.

Using **scikit-learn** and pandas, all stateof-the-art metrics are supported

**Watson Studio** supports sharing of jupyter notebooks, also using a fine-grained user and access management system

## Actionable Insights

### Technology Choice

### Python, pandas and scikit-learn

### Justification

Python is a much cleaner programming language than R and easier to learn therefore. Pandas is the python equivalent to R dataframes supporting relational access to data. Finally, scikit-learn nicely groups all necessary machine learning algorithms together. It’s supported in the IBM Cloud via IBM Watson Studio as well.

Python skills are very widely available since python is a clean and easy to learn programming language.

## Applications / Data Products

### Technology Choice

* Node-RED

### Justification

Although Node-RED is a No-Code/Low-Code data flow/data integration environment, due to its modular nature it supports various extensions including the dash boarding extension. This extension allows for fast creation of user interfaces including advanced visualizations which are updated in real-time.

Due to the completely graphical user interface based software development approach, only basic skills are required to build data products with Node-RED

## Security, Information Governance and Systems Management

### Technology Choice

* IBM Cloud PaaS/SaaS

### Justification

BM Cloud PaaS/SaaS eliminates operational aspects from data science project since all components involved are managed by IBM