The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document
Advanced Data Science Capstone Project
by Kostyantyn Kravchenko

1 Architectural Components Overview

1.1 Data Source

1.1.1 Technology Choice
External data source.
CSV (coma separated values format) file provided by a client.
92.254 MB of data.

1.1.2 Justification
The CSV file provided is a common format for table data.

1.2 Enterprise Data

1.2.1 Technology Choice
Not applicable.

1.2.2 Justification
Not applicable.

1.3 Streaming analytics

1.3.1 Technology Choice
Not applicable.

1.3.2 Justification
Not applicable.

1.4 Data Integration

1.4.1 Technology Choice
Watsons Studio, IBM Cloud Object Storage.

1.4.2 Justification
IBM Cloud Object Storage provides a free plan and is easy to integrate into Watson Studio projects:

- 1 COS Service Instance
- Storage up to 25 GB/mo.
- Up to 20,000 GET requests/mo.
- Up to 2,000 PUT requests/mo.
- Up to Data Retrieval 10 GB/mo.
- Up to 5GB Public Outbound

1.5 Data Repository

1.5.1 Technology Choice
IBM Cloud Object Storage.

1.5.2 Justification
IBM Cloud Object Storage provides a free plan and is easy to integrate into Watson Studio projects.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Watson Studio, Jupyter Notebooks.

1.6.2 Justification
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 modeling, data visualization, machine learning, and much more.

1.7 Actionable Insights

1.7.1 Technology Choice
Watson Studio, Jupyter Notebooks, Python (pandas, sklearn).

1.7.2 Justification
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 modeling, data visualization, machine learning, and much more.

1.8 Applications / Data Products

1.8.1 Technology Choice

Jupyter Notebook with full pipeline: takes in raw data in CSV and outputs the fitted pipeline with trained sklearn model saved to pickle format and ready for production.
Model deployed as service with Dash application running on Flask REST API. Access through web-interface.

1.8.2 Justification

The model is easy to access through web-interface and the service can be easily adjusted for the batch predictions or web-application response.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Not applicable.

1.9.2 Justification

Not applicable.