Architectural Decision Document

The Lightweight IBM Cloud Garage Method for Data Science

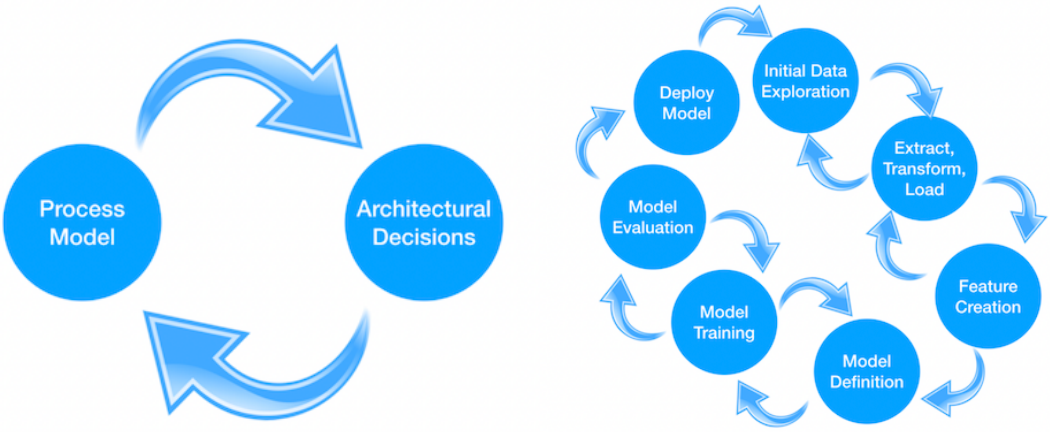
Kumar Hemant[†]

[†]*attendee @ IBM Advanced Data Science via Coursera*

'A PROCESS MODEL TO MAP INDIVIDUAL TECHNOLOGY COMPONENTS TO THE REFERENCE ARCHITECTURE'

April 1, 2020

The IBM Garage Method for Cloud is IBM's approach to enable business, development, and operations to continuously design, deliver, and validate new solutions. The practices and workflows cover the entire product lifecycle from inception through capturing and responding to customer feedback and market changes.

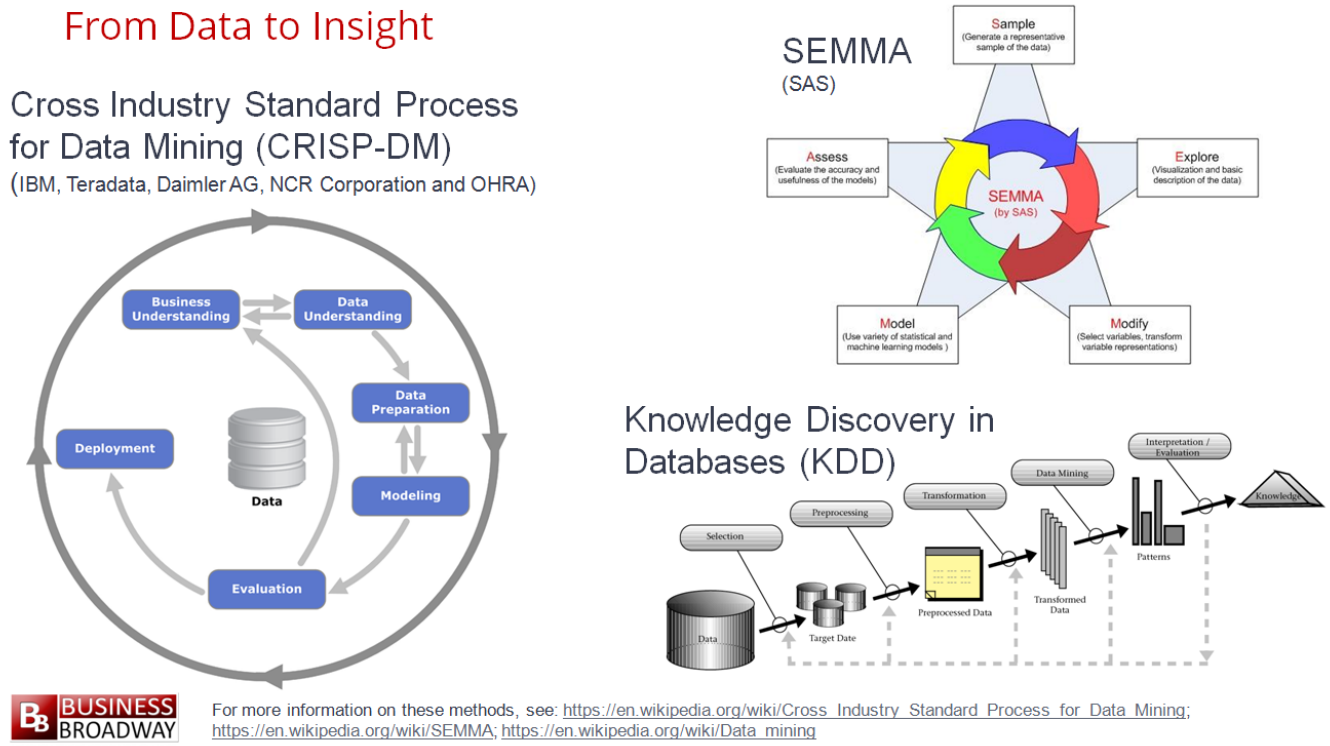


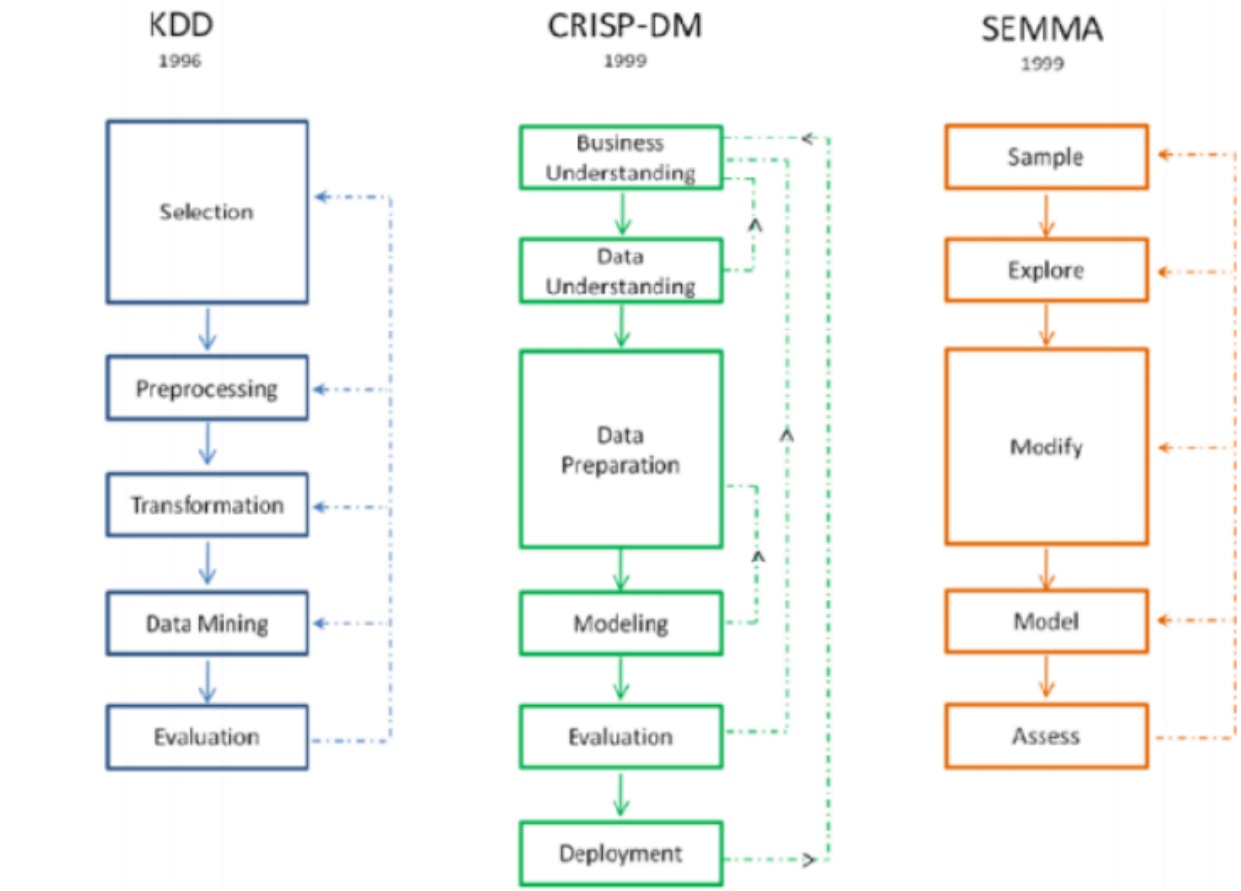
The lightweight IBM Cloud Garage Method for data science includes a process model to map individual technology components to the reference architecture. This method does not include any requirement engineering or design thinking tasks. Because it can be hard to initially define the architecture of a project, this method supports architectural changes during the process model. A separate companion article discusses the architectural decision guidelines

1 The Standard Architecture and Process Models

- KDD
- CRISP-DM
- SEMMA

There are three major standardized process model in data science.
The following pictorial illustration shows who have championed and what are the steps involved.



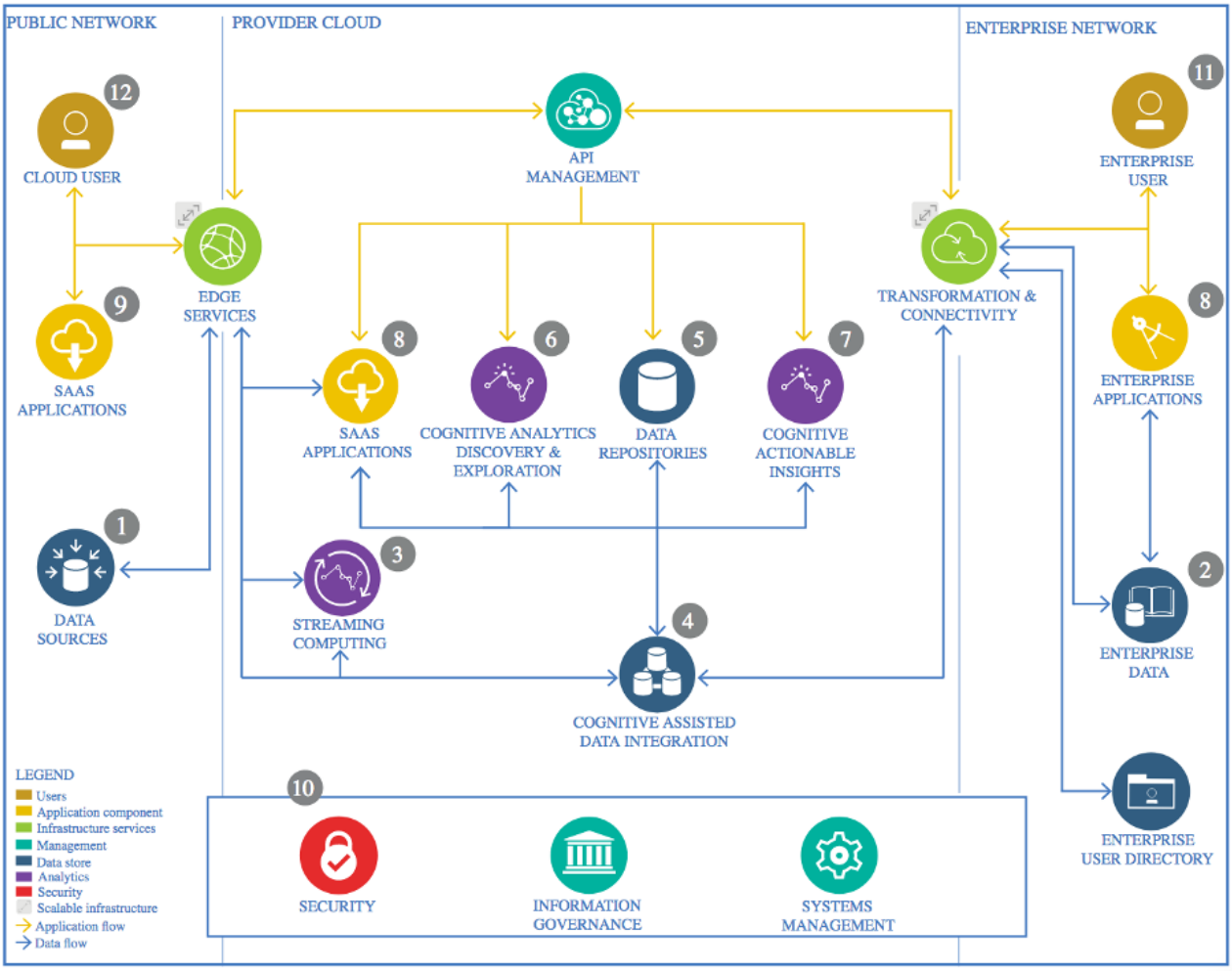


Since among all, KDD, CRISP-DM and SEMMA, the CRIP-DM has been widely accepted and used championed by IBM. We can breakdown the steps as re-explained and re-imagined by IBM in particular by the famous IBM Garage Model practices. The steps of IBM garage model can be well be sectionalized at individual level and clearly be broken-down.
The following sections explain the individual tasks as per IBM Garage (Lightweight) Method for Data Science.

2 The lightweight IBM Cloud Garage Method for data science process model

This section introduces this lightweight process model. There is IBM Garage Model (which is vast and encompass large suit of tools and technologies), but we focus of lighter version of it.

The Reference Architecture of IBM Garage (Lightweight) Method



3 Initial data exploration

This task is crucial for understanding your data.

Initial Data Exploration & Data Dictionary: Data Dictionary to understand the data and its attributes qualitative what they mean and what they stands for. Exploratory Data Analysis, Visual Data Analysis, Data Distribution, etc for further understanding of the data.

3.a Tools guidance

The tools used

- IBM Watson Studio
- Jupyter Notebooks,
- scikit-learn,
- pandas,
- Matplotlib, seaborn,
- git/github

4 Extract, transform, load (ETL)

This task is an important step in transforming the data from the source system into data suitable for analytics.

4.a Tools guidance

Tools that you can use to perform this task include:

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

5 Feature creation

This task transforms input columns of various relations into additional columns to improve model performance. A subset of those features can be created in an initial task (for example, one-hot encoding of categorical variables or normalization of numerical variables). Some others require business understanding or multiple iterations to be considered. This task is one of those benefiting the most from the highly iterative nature of this method.

5.a Tools guidance

Tools that you can use to perform this task include:

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

6 Model definition

This task defines the machine learning or deep learning model. Because this is a highly iterative method, various iterations within this task or including up- and downstream tasks are possible. I recommend starting with simple models first for baseline creation after those models are evaluated.

6.a Tools guidance

Tools that you can use to perform this task include:

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

7 Model training

This task trains the model. The task is set apart from model definition and evaluation for various reasons.

7.a Tools guidance

Tools that you can use to perform this task include:

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

8 Model evaluation

This task evaluates the model's performance.

8.a Tools guidance

Tools that you can use to perform this task include: RMSE a metric was computed using actual and predicted values from model. This was the criteria.

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

9 Model deployment

This task deploys the model. The task depends heavily on the use case, especially, on the stakeholder's expectation on consuming the data product. So, valid ways of deployment include:

- An interactive Jupyter Notebook
- An export of an already run, static Jupyter Notebook or some type of report

9.a Tools guidance

Tools that you can use to perform this task include:

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

10 Naming convention

Need a structure to name your assets? Here's our recommended convention. Note that we recommend to always use `project_name`, while the others are optional.

```
[project_name].data_exp.<technology>.<version>.<extension>
[project_name].model_deployment.<technology>.<version>.<extension>
```

On same guideline, the naming convention is used like:

```
Capital-Bikeshare-1-Random-Forest.ipynb
Capital-Bikeshare-2-Neural-Network.ipynb
```

11 Summary

This article information on the lightweight IBM Cloud Garage Method for data science. It included a process model to map individual technology components to the reference architecture. A separate companion article discusses the architectural decision guidelines.

Architecture Decision Document for the Capital Bikeshare Capstone Project

The followings sub-sections will discuss on the guidelines of IBM Lightweight Garage Model what are the factors have been taken into account for this Capital Bikeshare Rental Forecast project.

Project: Capital Bikeshare - Rental Demand Forecast (Hourly/Daily)

1 Data Source

External data source was chosen. The URL for the original data of the Capital Bikeshare of Washington DC during 2011-2012 was taken into consideration.

Trip History Data URL: https://www.capitalbikeshare.com/system-data

- Duration – Duration of trip
- Start Date – Includes start date and time
- End Date – Includes end date and time
- Start Station – Includes starting station name and number
- End Station – Includes ending station name and number
- Bike Number – Includes ID number of bike used for the trip
- Member Type – Indicates whether user was a ”registered” member (Annual Member, 30-Day Member or Day Key Member) or a ”casual” rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

1.a. Technology choice

The CSV file provided is a common format for table data. Used python ‘pandas’ and ‘urllib2’ to read the data.

1.b. Justification

The data is flat files despite large number of records, the size of the file is few MB. However, the data to hold in memory was sufficient on 16GB machine and 8GB on IBM Cloud.

2 Enterprise Data

Not applicable. No enterprise data considered.

2.a. Technology choice

Not applicable. However, the Jupyter Notebooks was used on IBM Watson data platform to analyze the data and run models.

2.b. Justification

Jupyter Notebooks are sufficient for this small project that does not require deployment. The Jupyter Notebooks can be shared with other Data Scientists in the department.

3 Streaming Analytics

3.a. Technology choice

The data set is a finished product past historical record that does not require real-time streaming for processing in this scope of exercise.

3.b. Justification

The data is not in motion but stationary. The data set is historical download-able or readable using any standard stream or object reader. Python csv reader is needed not any real-time streaming.

4 Data Integration

4.a. Technology choice

No data integration applied.

4.b. Justification

The draw data set used for this project comes from a single csv file and was already itntegrated.

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

5 Data Repository

5.a. Technology choice

The data is stored as csv file and can be loaded from the GitHub repository. It was uploaded to IBM Cloud Object Storage as well and connected to the Watson Studio project. But due to security of access through key prefer to read directly as url using urllib. From there it can be loaded as streaming body into the notebook.

5.b. Justification

The data is light enough to be stored and loaded either way described above.

6 Discovery and Exploration

6.a. Technology choice

An ETL notebook section was created analyzing the data with Pandas data frames and Matplotlib. A more detailed exploration was started in the first model main notebook with Pandas data frames and seaborn plots.

6.b. Justification

The data are given as continuous values in a csv file perfectly suited for Pandas data frames.

7 Actionable Insights

7.a. Technology choice

The data set required scaling due to strongly different value ranges. The SciKit-Learn functions for preprocessing were used. Watson Studio, Jupyter Notebooks. The data quality is assessed with EDA performed with pandas, pandas-profiler and matplotlib libraries. No missing values.

7.b. Justification

The SciKit-Learn library for Python is the gold standard for preprocessing and machine learning. Whereas 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.

8 Applications / Data Products

8.a. Technology choice

- SciKit-Learn (in-memory Machine Learning)
- Watson Studio, Jupyter Notebooks, Python (pandas, sklearn).
- Feature Engineering is based on date-based features. Other feature engineering was season, weather and logarithmic scaling for large values of features.
- The variables are pre-processed to work with linear models.

8.b. Justification

SciKit-Learn was run in the Watson Studio 1-CPU environment and 2-CPU for random forest model for cross validation. Scaling the data increased in memory especially the Random Forest while many fold (10) cross validations and feature engineering, validation and computing mean error (RMSE) significantly.

The Cloud deployment makes it easy to predict values from anywhere with the API and access credentials.

9 Security, Information Governance and Systems Management

9.a. Technology choice

Not applicable.

9.b. Justification

Not applicable.

References

- IBM Garage Method
  https://www.ibm.com/garage/method/
- The lightweight IBM Cloud Garage Method for Data Science
  https://developer.ibm.com/technologies/artificial-intelligence/articles/the-lightweight-ibm-cloud-garage-method-for-data-science/
- Architectural Decisions Guidelines: An Architectural Decisions Guide for Data Science
  https://developer.ibm.com/articles/data-science-architectural-decisions-guidelines/
- IBM Data Science Reference Architecture
  https://www.ibm.com/cloud/architecture/architectures/dataArchitecture