The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

Using building permit data that is provided from United States Census Bureau that can be found at <https://www.census.gov/construction/bps/>.

The data that is being pulled for analysis is located within the ASCII files by State, MSA, Count or Place and from there we have selected State (<https://www2.census.gov/econ/bps/State/>).

The files are stored as .txt files but are structured as CSV so we can utilize the read\_csv function within Pandas.

### Justification

This data has shown to have the relevant information that is required for the analysis requirements.

## Enterprise Data

### Technology Choice

No enterprise data is required for this project.

### Justification

The data used is on the United States Census Bureau website and is easily obtainable and is not large enough to warrant the use of Enterprise Data.

## Streaming analytics

### Technology Choice

Watson Studio will be utilized for the Streaming Analytics and will implement an instance of pyspark to do the machine learning aspect of the project.

### Justification

Watson Studio and pyspark will be able to handle the load of the data.

## Data Integration

### Technology Choice

Data integration, such as SQL, will not be required for this project.

### Justification

The dataset that is being used is not large enough to warrant being stored within the cloud and using ETL.

## Data Repository

### Technology Choice

The data being used will remain in it’s original state and will be retrieved whenever it is required. The application will be able to retrieve current data and transform it whenever needed.

### Justification

Keeps the data current and not large enough to warrant developing another repository of the information.

## Discovery and Exploration

### Technology Choice

We will be creating a Jupyter notebook and utilizing Python, pandas, pyspark, scikit-learn and Matplotlib.

### Justification

The data, once retrieved, will create a visual chart of the historical flow of permits for new homes.

## Actionable Insights

### Technology Choice

Will be using pyspark and developing a machine learning model to predict the upcoming permits of new homes in each of the divisions, regions and states.

### Justification

Pyspark will be sufficient given the data size.

## Applications / Data Products

### Technology Choice

No additional applications or data products are required for this project

### Justification

Watson studio and the other libraries are sufficient.

## Security, Information Governance and Systems Management

### Technology Choice

This project will be limited to myself and will be shared with others on a case-to-case basis.

### Justification

Not for public view.