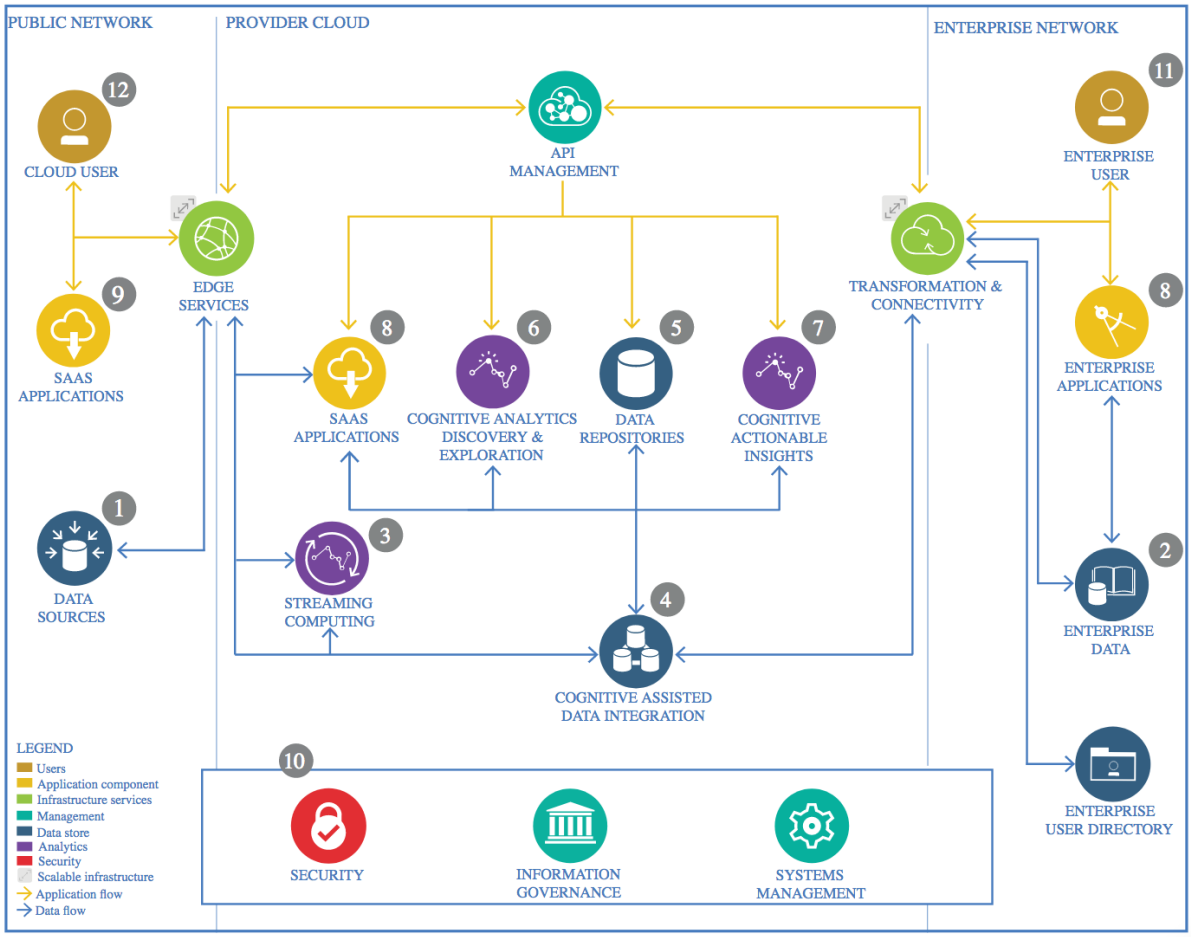
Architectural Decisions Document (ADD)

The Lightweight IBM Cloud Garage Method for Data Science

Project:

Power Output Prediction of Combined Cycle Power Plant (CCPP)

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

The data was retrieved from UCI machine learning repository and loaded on IBM cloud object storage. The UCI repository link is given below.

https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

1.1.2 Justification

The data for environmental conditions was required to estimate CCPP power out. Therefore, online repository was searched to retrieve the data.

1.2 Enterprise Data

1.2.1 Technology Choice

IBM cloud object storage will be used to analyze the data and run models.

1.2.2 Justification

Free credits have been provided to use IBM cloud object storage therefore no economic analysis was conducted to make decision if cloud storage is a viable option.

1.3 Streaming analytics

1.3.1 Technology Choice

No streaming technology for data will be used.

1.3.2 Justification

The historical data will be used for modelling and forecast. No streaming technology is required.

1.4 Data Integration

1.4.1 Technology Choice

Pandas was chosen as a technology for data exploration, cleansing and transforming. Box plots (method) were chosen to assess data quality. After data exploration and wrangling, it was loaded to apache spark. IQR-Score Method was used to remove the outliers. For feature engineering, correlation matrix method was used to determine correlation and decide about number of input features as model input parameters. Spark SQL and PySpark were used for data visualization and feature engineering puproses.

1.4.2 Justification

The data are given as continuous values in an excel file perfectly suited for Pandas data frames. Also, Pandas is a great tool for exploratory data analysis, quality assessment and transformation. Box plots method was chosen for data quality assessment because it tells

about outliers present in the data. IQR-Score Method was used to remove the outliers because it performed better than Z-score method. The draw data set used for this project comes from a single excel file and was already integrated. Apache spark works best with structured data. Correlation matrix displays strength of correlation of each independent variable to output variable (to forecast) in a very comprehensive manner. Pipelines were built with different number of input features for modeling and evaluation by using PySpark and Spark SQL.

1.5 Data Repository

1.5.1 Technology Choice

The data uploaded to IBM Cloud Object Storage and connected to the Watson Studio project. From there it can be loaded into the Jupyter notebook.

1.5.2 Justification

The data is not huge therefore it can be stored and loaded to notebook easily.

1.6 Discovery and Exploration

1.6.1 Technology Choice

A detailed data exploration notebook was created with Pandas data frames, seaborn plots and matplotlib. Box plots (method) were chosen to assess data quality.

1.6.2 Justification

The data are given as continuous values in an excel file perfectly suited for Pandas data frames. Box plots method was chosen for data quality assessment because it tells about outliers present in the data

1.7 Actionable Insights

1.7.1 Technology Choice

Regression was chosen as the algorithm of choice for modeling. And R-squared was chosen as the model performance indicator. It is also called the coefficient of determination.
Following technologies were chosen for actionable insights.
PySpark(distributed Machine Learning)
SciKit-Learn (in-memory Machine Learning)
Keras (in-memory Deep Learning)
IBM Watson Studio
Jupyter Notebooks

1.7.2 Justification

Since the main objective of the project is to forecast pant power output from ambient conditions (as input features), regression was chosen as the algorithm of choice for modeling. In order to solve regression problems, a fitted line is drawn or calculated by using

algorithms. The better the data scatter around the fitted line, the better the regression model is. R-squared calculates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination
PySpark machine learing was used for regression and hyperparameter tuning by using grid search and cross validation. Linear regression, radom forest and decision tree regressors were chosen as the models.
SciKit-Learn was run in the Watson Studio 1-CPU environment.
Keras was used to run and tune neural networks.

1.8 Applications / Data Products

1.8.1 Technology Choice

The final product is a correlation to predict plant power output by using environmental conditions. Therefore a PDF report is selected.

1.8.2 Justification

A PDF report is enough to explain correlation for predicting power output of power plant and will be handed over to the client with details about business understanding, data exploration and transformation, feature engineering, modeling, and evaluation.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Git was used for code storage. Notebooks were stored on IBM Watson cloud. There is no security protocol over the data and the models.

1.9.2 Justification

Security is managed by the service providers (IBM Watson and git).