School of Computing & Information Sciences

Spring 2019 Senior Design Project

FIU Engineering & Computing
School of Computing & Information Sciences

Data Science Workflow Manager

Students: Serge Metellus
Mentor: ***Steven Luis, Dr. Miguel Alonso,*** Florida International University
Instructor: *Dr. Masoud Sadjadi*, Florida International University

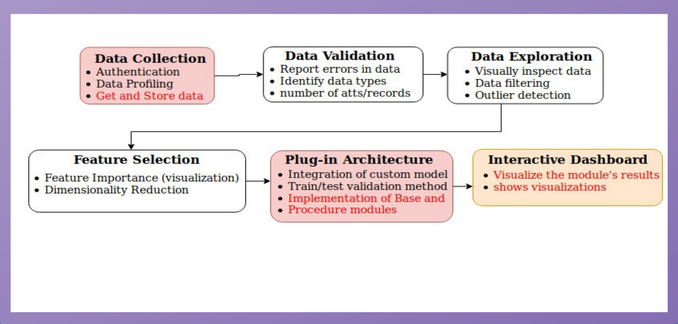
Problem

The airline industry currently does not have the capability to fully utilize the large Amounts of data generated every day to enhance customer experience, increase the efficiency of operations or make better data-driven decisions. Following the latest data science tools and techniques, we can leverage valuable insights from airline data.

Current System

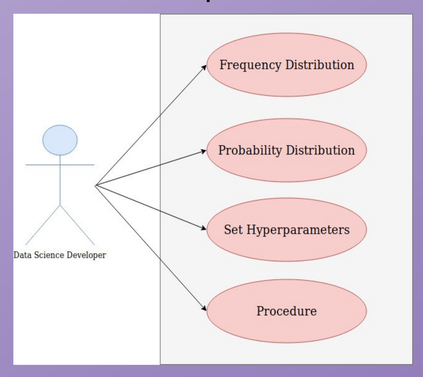
Version 1.0 System:

- Users are able to see the visualizations of the
- data collected and processed from CPI and Oil.
- The different distributions can be selected from a drop down and the responding graph will be displayed.
- The graphs display data returned by the procedure method of each distribution class



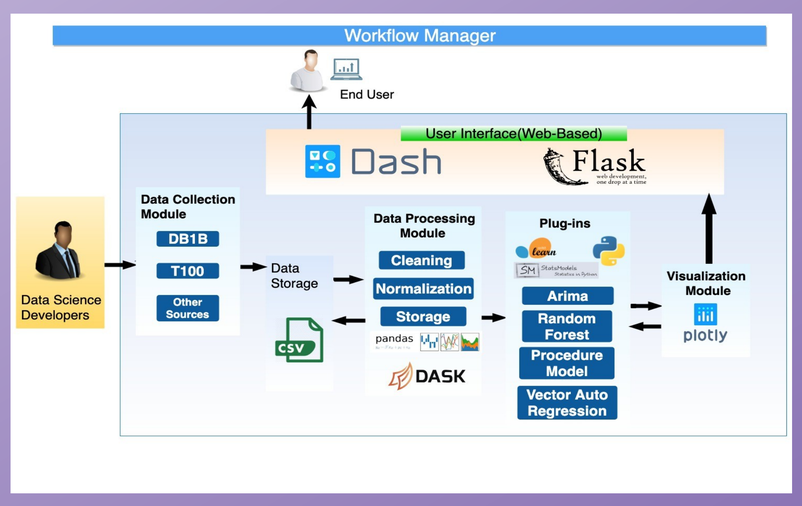
Requirements

The WFMG project requires the development of



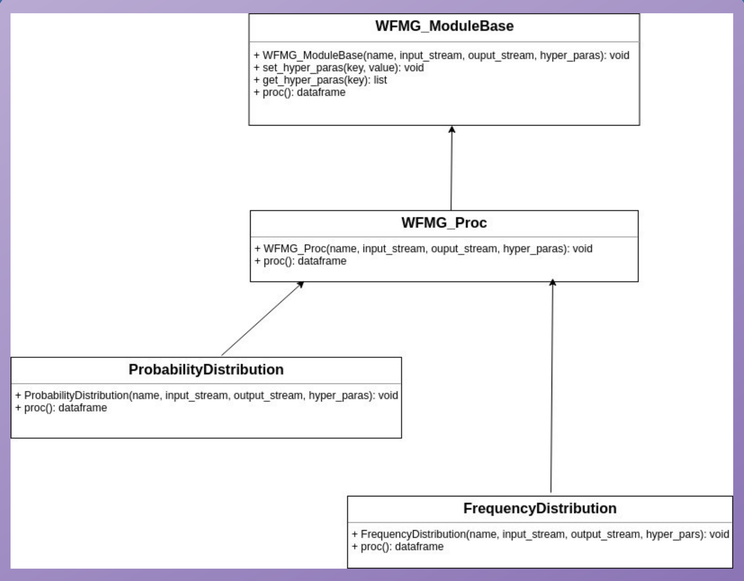
- A front-end that displays visualizations of data to the users and allows the users to filter and compare data from different sources
- A WFMG Module, shown in the diagram, that returns the probability and frequency distributions of the various data sources
- Data processing scripts, one for each source, automated by Make that processes and cleans the data before visualizing
- A back end web application that interfaces the WFMG module, data processing scripts, and the user interface.

System Design



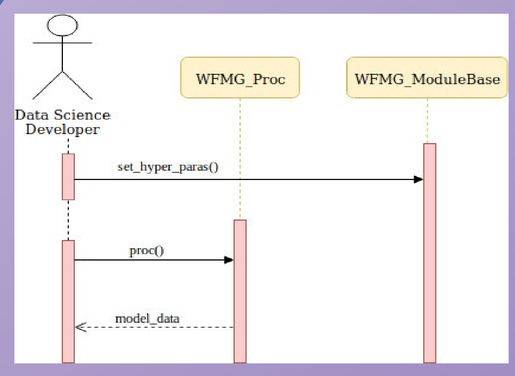
An overview of the entire system architecture highlighting the different modules in the system.

Object Design



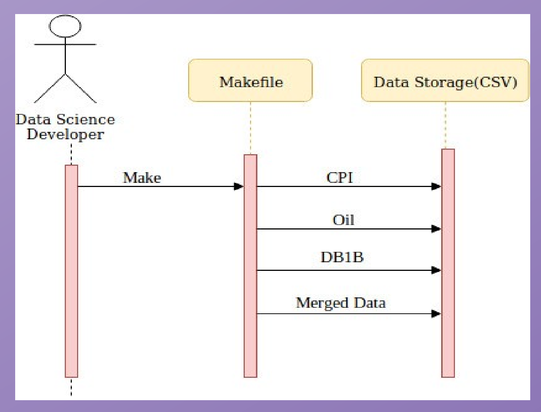
Class Diagram depicting the WFMG Module base class and the Procedure class. The distribution sub-classes of WFMG_Proc returns the specified distribution of the input Stream.

Implementation



WFMG Module Base

Virtualization module that allows data science developers To find the probability or frequency distribution of a data set.



Data Processing Module

- Make file that allows developers to process multiple sources of data concurrently.
- The processed data is then Imported by the controller for visualization.

Verification

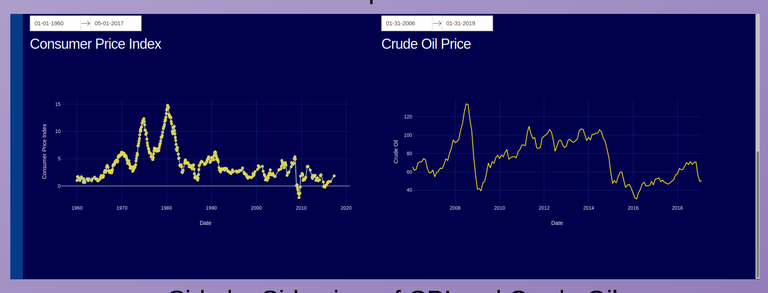
Integration and Unit Testing were used to verify the System.

Test case ID: WFMG-24-UT	
Purpose	Validate that the developer is able to instantiate the distribution classes and use the specified methods.
Precondition	The developer must have the WFMG module in the present working directory or imported from somewhere else.
Expected Results	The developer should be able to instantiate the Probability and Frequency sub classes.
Actual Results	The results are as expected, the models used to instantiate the distribution sub classes are able to use the methods from those classes and the inherited methods from the super-class
Status	Pass

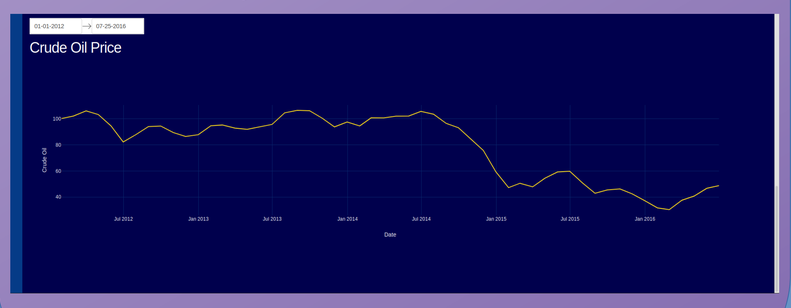
Test case ID: WFMG-24-IT	
Purpose	Validate that the visualizations in the graphs are able to use the methods from the WFMG module.
Precondition	The application (browser) must be installed and running.
Expected Results	The graph should use data from a data fame return by a call to the proc method.
Actual Results	The graph is visible from the dashboard with the expected style and arrangement and the data points come from a dataframe returned by the proc method.
Status	Passed

Screenshots

Visualization of the Coupon Distribution from DB1B



Side by Side view of CPI and Crude Oil



Crude Oil Price for a specific date range

Summary

The goal of this project is to create a workflow manager that allows developers working on a data science project to introduce new models as plugins into data processing to allow end users to visualize and explore processed data and various data distributions.

In this first iteration, we focused on the plugin architecture and the interactive dashboard for users interested in analyzing the different data distributions.

Acknowledgement

The material presented in this poster is based upon the work supported by FIU AIRLab and FareLogix . I am thankful for the help that I received from my group member Tomas Ortega