Project Requirements

Requirements

The basic need to be satisfied by this project is to automate the process of extracting the quarterly information from Empirical model sheets and Regression Model sheets of each client data excel into the consolidated excel document in the specified format. This activity is currently a monthly activity.

The project requirements for infrastructure, hardware, software is explained on business requirements. The solution is expected to be 100% open source and automated.

We have asked a solution from each participating team to develop and present pragmatically. The winning team will receive a reward.

Feature Highlight

- Extract files from the zip file
- Read required information from Empirical model and Regression model screens
- Write the required information into Dataout excel file
- Delete the extracted files
- Archive the zip file
- Persist the data on a Sql DB
- The code should be deployed on the server through a CI/CD mechanism with appropriate CI and CD tools
- The solution should have well cloud architecture best practices.
- The teams should work end to end provide infrastructure and operation support to all applications requirements moving from one environment to production platform environment
- The teams should be able to troubleshooting application issues as well as platform issues running in Cloud Infrastructure, co-ordinate issues resolutions with operations, functional and technical teams.
- The developers should work independently on while dependencies are yet to be solved and the code should be demonstrated during any phase of the project irrespective of the simple to medium complex projects.

Business Requirements

Req No.	Description
BR1	Automate the creation of a CentOS FTP server on any cloud
BR2	The ‘/opt’ folder should be mounted with an encrypted drive to ensure that the data is secure
BR3	Create a cronjob to trigger a Linux script that will perform to trigger the following actions: 1. Extract files from the Zip file in the same folder 2. Trigger a Python script
BR4	Create an Excel file in datadev ‘/opt/dataout’ folder as output file
BR5	Name of the output excel file shall be “model_YYYYMMDD”. “YYYYMMDD” refers to the current date or date of execution. For example: model_20190220
BR6	Once the extraction of the data from all files is complete, delete extracted excel files
BR7	After deleting excel files, archive/move the zip file to ‘/opt/dataout’
BR8	Each execution of the script should generate a new Excel file in the dataout folder with the filename as mentioned in BR4
BR9	Read Estimated Total Sold, Minimum Sold and Maximum sold values from the Empirical Model from each excel
BR10	Read Forecast w/o SA Actual, Minimum and Maximum values from the regression Model from each excel file.
BR11	The source excel sheet may have more than 1 Empirical model and 1 regression Model sheets with type included in the sheet name.
BR12	The information should be read from each of Empirical model and regression model screens

BR13	Output excel should have the following columns: • Date (current date) • Ticker (Company Code, picked from source excel file name) • Type • Quarter • Year • Estimated Total Sold • Estimated Sold Max • Estimated Sold Min • Forecast w/o SA Actual • Forecast w/o SA Max • Forecast w/o SA Min
BR14	Data from each source file should be entered as a record in the output file
BR15	If the source file has multiple sheets, then create a separate record for each type of sheet. Example: Total Americas, knees, hips etc.
BR16	Type – refers to the type of Empirical and Regression model sheets. When Type is provided in Empirical and Regression sheet name, the same should be mentioned in Type on output file record.
BR17	When no type is mentioned with Empirical and regression sheet names, then Type should be NULL in the output file.
BR18	Quarter – refers to quarter for which data is collected (mentioned Col C of Regression Model sheet of the source file)
BR19	Year – refers to the year for which data is collected (mentioned Col C of Regression Model sheet of the source file)
BR20	Estimated Total sold – Estimated quarterly value mentioned in Col F of Empirical Model sheet of the source file
BR21	Estimated Max sold – Estimated quarterly value mentioned in Col F of the empirical Model sheet of the source file
BR22	Estimated Min sold – Estimated quarterly value mentioned in Col F of the empirical Model sheet of the source file
BR23	Forecast w/o SA Actual – Quarterly Forecast value mentioned in Col R of Regression Model sheet of the source file
BR24	Forecast w/o SA Max – Quarterly Forecast value mentioned in Col R of Regression Model sheet of the source file
BR25	Forecast w/o SA Min – Quarterly Forecast value mentioned in Col R of Regression Model sheet of the source file
BR26	Empirical Model data and Regression model data of the same type should be part of the same record. Example: for knees, empirical Model data and Regression Model data should be part of the same record in the output file.
BR27	Post the output data in an MySQL Database named AnalystData which is in a Database server
BR28	If the program is run twice on the same day, then delete the data that was previously posted on that day before posting the data
BR29	The latest month's Data that is stored in the Datasheet of the excel files also has to be posted in the Database in a database called Data in the following format ['Date', 'FacilityType', 'BedSize', 'Region', 'Manufacturer', 'Ticker', 'Group', 'Therapy', 'Anatomy','SubAnatomy', 'ProductCategory', 'Quantity', 'AvgPrice', 'TotalSpend']

Sample output format of the xlsx file

Date	Ticker	Type	Quarter	Year	Estimated Te	Estimated Se	Estimated Te	Forecast w/o SA	Forecast w/o SA Max	Forecast w/o SA Min
03/26/19	ABMD	Null	3Q	2019	175149385.	177776626.	172522145.	172293029.6	174877425.0	169708634.1
03/26/19	ATRC	Null	4Q	2018	43165589.6	43813073.5	42518105.8	42157411.3	42789772.4	41525050.1
03/26/19	CRY	Null	4Q	2018	36273795.8	36817902.8	35729688.9	37271577.5	37830651.1	36712503.8
03/26/19	CSII	Null	2Q	2019	60788655.0	61700484.8	59876825.1	60719959.9	61630759.3	59809160.5

Non-Functional Requirements

Reliability

Tool must be reliable and perform defined actions maintaining the integrity and atomicity of the data

Robustness

Tool should be robust and must be able to handle any number of files and sheets within, without affecting the user experience

Scalability

Script should be scalable and able to integrate with any other platform in the future and do the desired action.

Challenges:

1. The entire end-to-end flow takes not more than 30 mins to complete.
2. The proper error-handling mechanisms are set in place to ensure that the process does not get interrupted abruptly