PaperID **AU226**

Author **VIKRAM KUSHWAHA , ONGC , India**

Co-Authors **Ms. Anna Tamuly, Sh. Sanjay Chakraborty**

Data Analysis and Visualisation in EPINET (Exploration and Production Information Network) data repository using Python

Abstract

In the world of computers, internet and increasing technology has led to the increase in rapid growth of data. Data which was considered to be too huge are considered to be very small sets of data in present time. There is no stop to the data and with increasing technology and users, data have increased to an unimaginable level. The oil industry is one industry which has been dealing with huge datasets since a long time. All the data produced in this industry is considered to be big data, as it is impossible to find small sets of data in this industry.

With such huge data which is at present approximately 75 - 80 TB, comes the task of analysing these data. There are usage of various tools and techniques for the analysis, cleaning of this data to gather meaningful Information and knowledge from the data. Python as a programming language is used quite a lot for data science for analysis of big data and visualising it. The main goal was to use Python programming language to various trends and patterns hidden in different types of data through various analytical techniques and visualisation.

Introduction

EPINET centre at Jorhat facilitates in collection, storage, upkeep and maintenance of ONGC's E & P data pertaining to entire North East area. It is one of the five regional centres of EPINET. The different data classes handled under this category are Well, Log, Seismic, Laboratory, Reservoir, Drilling, Production and Well Stimulation. In addition to the standard E & P data it stores, it also serves as a repository of other important information like reports from various groups and hosts s/w applications of other groups like GTO, Data Centre etc.

Case 1: Analysis of user activity log to find out data availability status from the database

All users of EPINET have unique username and passwords to log into the database web portal and then search for relevant data as per their project requirement. All the activity at the portal is archived. Log in / Log out times, IP address, data searched for and the success and failure of data search is all archived. This information can later be used to check the data access success and failure rate and subsequent improvement required in maximising the data availability for the user.

A large chunk of this data was taken to perform data analysis to visualize the various trends of pattern for the data usage. The main idea was to produce some inference of this data on the following basis:

a) Analysis was done to check the success and failure rate of the users query according to well names. A dataframe was created, grouping the description and result together from existing dataframe. Successful data search from the result was ignored and only the failed attempts was then plotted in form of a graph to produce the output showing which description has the highest fail rate.

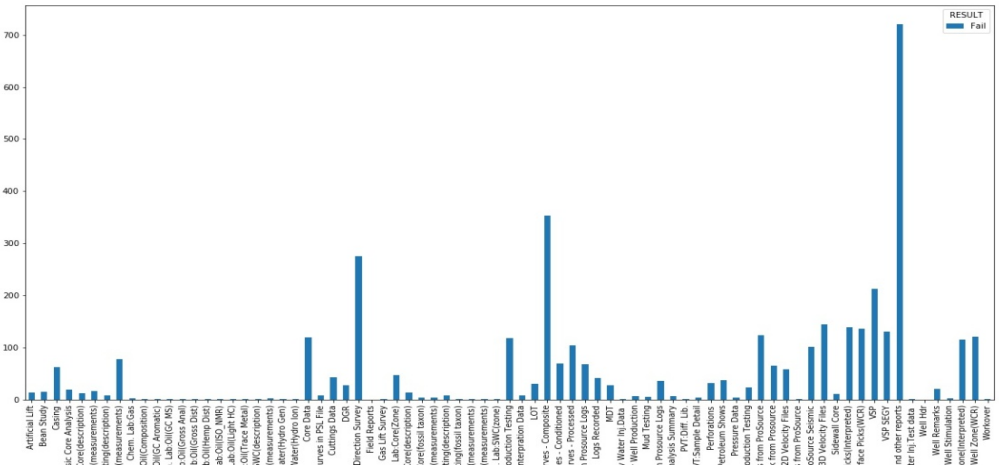

Fig. 1.1: Analysis of the dataset to find maximum no. of fails according to the description.

b) Analysis was done to check the success and failure rate of the users according to their groups. A dataframe was created grouping the department and result together from existing dataframe. Successful data search from the result was ignored and only the failed attempts was then plotted in form of a graph to produce the output showing which description has the highest fail rate grouped according to groups to find out which area has maximum data gaps .

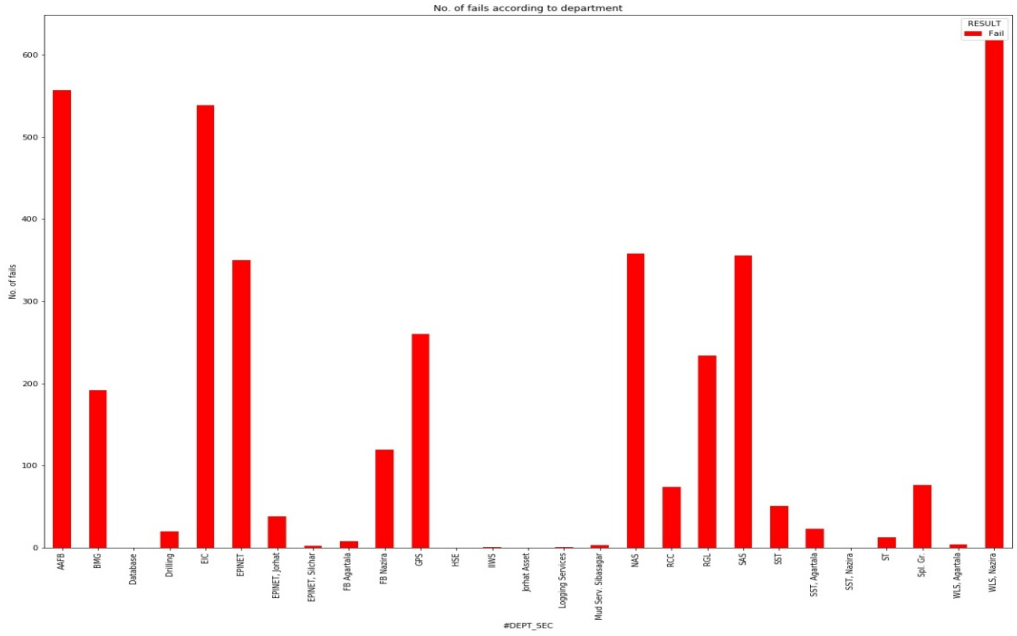

Fig 1.2: Analysis to find out the total no. of fail rate on the basis of groups of the users.

c) Analysis was done to check the success and failure rate of the users according to username. A dataframe was created grouping the department, username and result together from existing dataframe. Successful data search from the result was ignored and only the failed attempts was then plotted in form of a graph to produce the output showing which user has the highest data access fail rate.

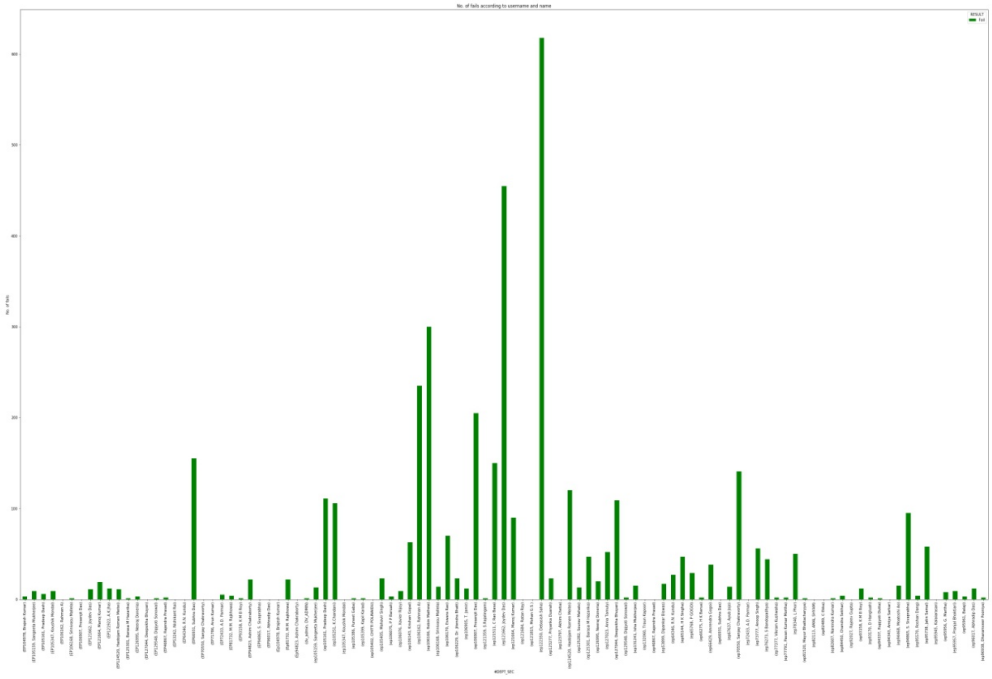

Fig 1.3: Analysis to find out the maximum no of fails according to the username.

Case 2: Analysis of production data to find data gaps

In EPINET production allocation data is stored with respect to individuals wells for each month. A large chunk of data was collected for a particular field with well names, month and year of production. Data availability was marked with one and non-availability as zero. The main idea was the use of Python and produce inference on month-wise data gaps. The dataset was divided on the basis of months from 2011 to 2016. The pie chart with value 0 shows the wells on which data is not available on that particular month. Similarly, the pie diagram with value 1 shows the wells on which data is available on that particular month of a particular year.

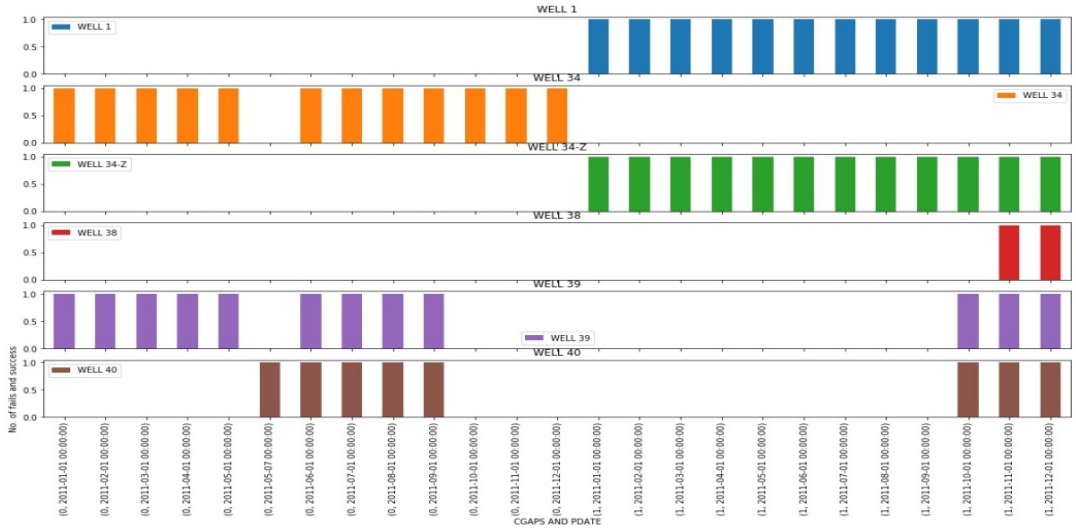

Fig 2.1: Showing data availability and gaps for 6 wells over a period of 2 years.

Conclusion

This analysis uses the Python programming language and different packages involved within it like pandas and matplotlib to perform data analysis of the given datasets.

In the first case “Analysis of user activity log to find out data availability status”, it is about using pandas to group the various activity log data together, and then deleting the least important columns in the dataframe so as to get the failure rate of the various data search outcomes like, a) data search failure with respect to wells, b) groups /areas, and c) by name/username of employees.

In the second case “Analysis of production data to find data gaps”, it is about using pandas to extract the dataframes to excel worksheet to find the data gaps between the start of production of a particular well to the current date of working of the well. It also uses pandas and matplotlib to find the data gaps and work done on the wells, month-wise since January 2011.

EPINET is a big database covering entire range of exploration and production data. The data present have both diverse varieties as well as they are big in numbers. There are large number of users from various sections of the organisation who use these data for their project activities as well as day to day data requirement. It is very essential that proactive monitoring of both user activities and availability of data is been done, so that the database serves the purpose for which it is been set up.

The present case studies were very helpful as it helped in identifying the number of failures the users have while searching for data in the EPINET database. By categorizing data search failures under various category like data types, groups, and areas, it became easier in identifying the areas of data gaps. Similarly it also helped in finding pockets of areas where there exists data gaps w.r.t. production data and subsequent tallying with original data source gave the exact extent of data gaps. Subsequent corrective measures ensured filling of those data gaps resulting in enhanced data availability for users.