Approach towards Data Mining

As the businesses are becoming complex day by day, everyone can see the amount/quantum of data being generated and recorded in higher volumes. Technology has added value to our personal as well as our commercial status, thereby increasing the recording of data reached at wider level. With respect to the personal status the data may include security through the surveillance system, communication by social media and various reports of personal activity. The commercial status witnesses a lot of compliance which in today's time requires usage of the electrical mode mainly. Every regulatory process, even something as simple as day to day payment, depends on technology which may range from recordable data to non-recordable data, for instance, it can be said that a bank account transaction necessitates the data to be recorded for future purposes. However, the transmission of an informal message unrelated to the professional relationship between two business holders , such as exchanging wishes on a special occasion, does not require recording the data as it is not mandatory. To summaries, data is generated for a multitude of activities, day to day activities as well as important data necessary for future records. The increasing capacity of hard disc drives in laptops and computers, from a mere 32 GB to even 500 GB or 1 TB, showcases the need to store large volume of data as every year proceeds. In the last two decades, the world has witnessed a 100x increase in the volume of data being saved, which makes the procurement of useful data from this huge ocean of data a big challenge. Handling of this data is very complicated, but it becomes easier with many cloud-based servers providing the facility of storing valuable data in secured servers. If an MIS system provides strong value added information in a timely manner, handling business transactions becomes an easier job. Hence, for proper running of enterprises, the MIS must be prepared such that it can process almost all the data belonging to day-to-day transactions. Processing of this data is easier when the size of data is smaller, i.e., has smaller volume. As was mentioned above, the processing of bigger data is a huge challenge. However, if for some reason, the data is huge and cannot be pruned, Data Mining and Data Analysis can be used for the same. Despite it being a challenge, many applications are available for one's help in such as case. These applications are usually used based on the person's knowledge and experience using them. Many of us use basic applications such as MS Excel however; each application has some limitation with regards to the configuration of one's personal computer and processor. The limitation of application software hence extends to system/server configuration too. In case of BIG DATA lot of application we can use, however in my view machine learning language to mostly uses for Artificial Intelligance and BIG DATA mining. Biggest advantage of machine learning i.e R and Python is that processing time as compare to normal application drastically shorter, hence durability of our system configuration wider.

In case of any data mining, our approach is such that the outcome is either helpful to management or the end user to arrive at any decision or to detect a fraud. In case of fraud detection, we categories it in two types, viz; Red Flag and Green Flag. Red Flag is raised for frauds identified through data trail where any abnormality in transaction flow is observed and respective Green Flag is identified through data trail where every trend is found in right direction. Data mining plays an important role in detection of frauds in various departments thereby strengthening the enterprises overall by helping plug all the loopholes. Data Mining approach can be defined in our mind through approach of various data file structure and in that case there is an important role of understanding of DBMS i.e. Data Base Management System, when we start in our mind about data mining we need to understand data structure of various files. In short, data mining can be done with the combination of various files of any enterprises or relevant data files to make a usable file which provides understanding of various data and results in abnormal trend identification.

Data mining can be used most effectively through best data sample identification. Samples are drawn using our experience as well as available data analysis trend,. For an in-depth data analysis or for best Data mining, we need to understand the data structure and functionality of business structure. In case of a small file containing whole data of the enterprises, which can be understood or assessed easily, the full file shall be taken for data mining. However, in case of voluminous files with large number of data, we shall need to deploy a higher and more advance level of data analysis. In such cases, we shall use the following points in our approach for effective data mining:

1. Understanding whole data with different types of files
2. Filter all data files to usable file
3. Complete understanding of usable file data structure i.e. field and records with type of field i.e. number, string, Memo, date etc.
4. Co-relation of these files based on our analysis approach.
5. Creation of data files for analysis
6. Different type of algorithm/approach applied on data file to draw most suitable or best sample for audit.

Normally without data mining, we start our audit through review of financial details comparing current year figure with previous year's figure and pick sample with high value of deviation. In simple way we can say that our review process starts from front view i.e balance sheet, profit and loss account or General Leger balance to analysis, compare and detect high deviations. Apart from the above, can we draw correct samples on the basis of vouchers punched in system and thereby analysis through voucher level deviation? We have done Data mining using this approach though we did have thorough knowledge of the specific industry.

Case Study – Bank Audit

In case of bank audit DATA MINING, the bank audit is indeed a challenging audit because most of the financial frauds are linked with banks. In Past many such frauds have been declared and further most of such frauds are under investigation. The quantum of frauds in a bank is not small be it in value or in volume. Hence the audit of financial transaction is always a challenge. The bank are obviously having BIG DATA because of basic role of bank taking deposits of various customers and alternatively also providing loan facility to borrowers, resulting in two way transactions with very high volume of transactions. In India normally a bank branch itself is an entity and prepares financials as a standalone Balance Sheet. Hence lot of inter office adjustment entries are also passed in accounting system. A branch audit is thus having more challenge because the system which is used by the bank as such are different, hence hand holding in banking system is a challenging task.

1. In case of a bank audit, the first requirement is to understand the business module and process used by the bank, as well as the different types of policies issued by the bank, which may be directly or indirectly linked with the RBI. Following this basic study, the audit with Data Analysis and Data Mining can begin. Understand the various accounting systems issued by the bank, along with its different versions.
2. Identify a list of different types of menu/ transaction codes used in the system, which are provided access through our Auditor Login ID
3. Identification of a report which provides application software
4. Generate a report from the application software as well as various MIS details that are available in the core banking solution
5. Identify the files which can be used and understand their data structure
6. Identify various fields in a file, as well as the number of records in one file
7. There is also a need to know the file size
8. Correlate the field levels among the file for further analysis
9. Usage of different type of Tools for creation of data files, be it MS Excel or any other advance level tools to achieve desired output in time bound manner

10. Apply algorithm for different queries on data files, to identify best sample which might arise a red flag or green flag in data
11. Verification of sample along with physical verification of the supporting documents and voucher
12. Respective query mark in LFAR and Audit report.

Hence the above process defines a lot of sample with various combinations of files which also support us during audit documentation. When we start processing with the approach of DATA Mining, our role might be wide and we may have to work on a lot of data In my view, starting the audit with all the reports at once may not be an easy task, because when we started it, it required at least 12 to 13 months for analysis, with the data ranging from 100 to 200 MB in each branch.