e-PGPathshala

Subject : Computer Science

Paper: Data Analytics

Module: Analytics – How it must be?

Module No: CS/DA/5

Quadrant 1 – e-text

1.1 Introduction

The aim of the data analysis phase is to transform the data collected into valid or acceptable proof or evidence about the development of the intervention and its performance. This process usually includes rearranging and organizing the data for analysis, describing the data and interpreting the data. Here we discuss the criterion for good analysis, reporting, problem framing and inferences based on analysis. We will familiarize some open source tools also.

1.2 Learning Outcomes

- To understand the criteria for a Good Analysis
- To learn how to frame the problem correctly
- To know making good inferences
- To know about open source supports for analytics

1.3 .Data Analysis

Data analysis is not a passive process, it is an interactive and active process. It is a process of transforming data for discovering useful information, conclusion, and for decision-making. Analysis refers to dividing or breaking the whole into its small modules for individual examination. Otherwise, it is a process of obtaining raw data and converting it into useful information for decision-making. Data is collected and analysed to answer questions, test hypotheses etc. The main aim of the data analysis is to convert the data collected into useful data for improving the performance of anaylsis.

1.3.1 **Difference between Analysis and Reporting**

There is a marginal difference between reporting and analysis.

Reporting is the process to arrange or organize data into informative summaries, which helps to monitor how different areas of a business or an organization are performing. It is an aggregate of three actions: extract, transform and load. Reports extract meaningful, actionable insights, which can be transformed into useful and understandable summaries and that can be load or feed into any management system to improve business performance.

Analysis is the process of transforming data into useful developments. It is an interactive and active process which finally interpret the result.

Reporting and analysis are also different in their terms of its purpose, tasks, results etc.

Purpose - Translating raw data into information is called reporting. The scope and style of reports may vary depends on topic. Transforming data and information into insights is called analysis. Reporting helps companies to monitor the status of their online business. Reporting will raise questions about the business from its end users. The purpose of analysis is to answer questions by clarifying the data and giving directions.

Tasks - By examining the basic tasks performed by the analytic team we can distinguish between reporting and analysis. The main tasks in analysis are data collection, evaluation, interpretation and decision making. The main tasks in reporting are data collection, build and configure the data, consolidate and summarize the content in a systematic manner or in a suitable format.

Outputs

Deliverables of reporting and analysis may look similar with lots of charts, tables, graphs etc. Look closer, and we can see some differences.

Reporting usually follows a *push* approach, where reports are *passively* pushed to users to extract meaningful insights and take appropriate actions for themselves. But analysis follows a *pull* approach, where the analyst *actively* pulls some data to answer specific business questions and provide recommended next steps with possible outcomes.

Context is another important difference between reporting and analysis . Reporting provides little or limited context about what is happening in the data. But for a good analysis context is a critical factor.

Reporting	**Analysis**
Provides data in a suitable format	Provides answers or conclusion
Provides only what is asked for	Provides what is required for
Is standardized	Is customized
Personal involvement is necessary	Personal involvement not necessary
Moderately inflexible	Extremely flexible

1.3.2 Analysis

The analysis have to be done effectively. There are many factors that has to come together for an analysis, then only it will have some impact and value. Essential factors that drives analysis are "G.R.E.A.T" criteria: Guided, Relevant, Explainable, Actionable and Timely. (**Source:** Bill Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics", John Wiley & sons, 2012.)

Guided

There should be a business need and a great analysis will be directed by that need. It will not do because of any fun or special interest. With big data in particular, it is easy to get drawn into a lot of interesting but irrelevant work. A great analysis is one that starts through the identification of a specific business problem. Once identified, the analysis is guided by what is required to solve that problem. Every step of the analysis should be guided by the needs of the problem.

Relevant

Any great analysis has to be relevant to the business. This means that not just choose an arbitrary business problem. The problem needs to be one that the business feels needs a solution, and it has to be a problem that the business has an ability to address.

Explainable

A great analysis will need to be explained effectively to those tasked with acting on it. It is possible to get carried away with formulas, algorithms and statistics. Technical details may be the proof required behind the problem that an analysis is valid, the results need to be explained in terms that decision makers can understand. A great analysis will be explainable and easy for the decision to make a decision.

Actionable

A great analysis will be actionable. It will point to specific steps that can be taken to improve a business. There is no point to an analysis showing that moving a few stores a mile down the road will increase sales if there is no way the company would ever actually move the stores .Analysis become useless if it is not providing the ability to be acted upon it.

Timely

A great analysis will be delivered on time. Time is critical because the data should be needed for decision making. Having the answer to a question next month doesn't do any good if the decision needs to be made next week. It is possible for an analysis to be great in every aspect, but it just can't be completed in time for the decision it supports.. A late analysis is not great.

1.3.3 Core Analytics versus Advanced Analytics

1.3.3.1 Core Analytics

It is referred that non advanced analytics as core analytics, to keep it simple. Usually in core analytics asking simple questions and providing simple answers are natural process. A core analytics process is going to investigate what happened, when it happened, and what the impact was etc. Note that all the data for the core analysis can be provided by standardized reports. The analysis itself is the process of examining those reports, making inferences, and suggesting action. In this case, the analysis will consist of looking at the numbers and determining if the goals were met or not.

1.3.3.2 Advanced Analytics

Advanced analytics includes everything from complex ad hoc SQL, to forecasting, to data mining, to predictive modeling. One question that often arises is how advanced analytics is different from data mining, forecasting, or predictive modeling. The answer is that everything you would think about when you think of those activities is encompassed within advanced analytics. However, advanced analytics also includes other processes that aren't necessarily algorithm-intensive, such as ad hoc SQL—not basic everyday SQL queries, but highly complex SQL queries that involve combining data sources in complex ways. The reason activities like advanced SQL are included in the definition is that the main goal of advanced analytics is to quantify the cause of events, predict when they might happen again, and identify how to influence those events in the future. Sometimes it doesn't require a model to get the insights you require to answer those questions

1.4 Problem Framing

One of the most important tasks in project definition or research is to frame the problem correctly. Organisations frame problems in rigid and narrow manner that may lead to a poor outcome or the wrong tactics being employed. If we spend more time time seriously in defining a problem can improve the success of the organization in dimensions. Our frame should be sharp and precise. The solution depends on the problem frame, so good framing leads to good solutions.

1.4.1 Framing the Problem Correctly

Framing the problem means ensuring that important questions have been asked and critical assumptions have been laid out up front.Great analysis starts with framing the problem correctly. This includes assessing the data correctly, developing a solid analysis plan, and taking into account the various technical and practical considerations that are in play. Arguably, framing the problem is the most critical step of an analysis, because if it isn't done right, neither will be anything that follows.

1.4.2 Statistical Significance versus Business Importance

Analytic professionals are interested in statistical significance. The key is that statistical significance is only part of delivering a great analysis. Testing for statistical significance takes a set of assumptions and determines the probability that the results seen would happen if the assumptions are correct. We can take an example a coin, then it will land heads and tails each 50% of the time. With a fair coin, the odds of getting 10 tails in a row are very small. If 10 tails in a row are seen, there are only two possibilities. Landing a head is only one in 1,024 attempts and the second is that the coin isn't really fair after all. A significance test related to a run of 10 tails would say that you can be 99.9 percent or so confident that the coin isn't fair. This is because a fair coin will only yield such a result 0.1 percent of the time. Statistical significance is all about such computations. It is necessary to differentiate between statistical significance and business importance.

1.4.3 Statistical Significance

Statistical significance is used frequently for averages and percentages. It is also used to evaluate the parameter estimates that come out of statistical models. Testing for statistical significance can be valuable to make sure the importance of data. There are times when differences that appear to be significant will not be and times that differences that appear small will be found significant. A statistical test will make sure the right conclusions are reached. There is an entire discipline built around testing. A common term in the business world for this discipline is test and learn. In a test and learn environment, an experiment is designed so that it is possible to

specifically measure the effects of one or more options and identify which of the options is going to work best.

1.4.3 Business Importance

Statistics has a very important place in business decision analysis and decision making. In current ruthless marketing environment, a business can not outlast by simply making decisions on guesswork or approximations. Dedicated and accurate analysis on the collected scientific data and information can help to make profitable decisions for the business and organizations.

1.4.4 Samples Versus Populations

Today the systems are available to wok with the whole population, Sampling is also a useful common practice. The main difference between a population and sample depends on the observations assigned to the data set.

- A **population** consists of all items or elements from a set of data which has same characteristics.

- A **sample** includes of one or more observations or data from the entire population.

Depending on the sampling method, a sample can have lesser number of observations than the population. But the population can have the same or more number of observations. From the same population one or more samples can be derived. Mean, standard deviation, variance are some measurable characteristic of a population, they are known as **parameters.** A measurable characteristic of a sample is called a **statistic**.

1.5 Inference

Inference is not physical process, it is a mental process by which people reach specific conclusion based on certain evidence. Inferences are used in each and every field of human society. Diagnosing diseases by doctors, repairing the engine problem of machines by mechanic, estimating the loss and profit of business by analysts everywhere we are using inferences. It is essential for a human being to live in a society. Actually we engage in different types of inference in our day to day life. We interpret actions to be examples of behavior characteristics, intents, or expressions of particular feelings. We infer it will rain when we see someone with an umbrella or cloudy atmosphere. We infer people are thirsty if they ask for water.. Inferences are not random and ,they may be guesses, but they are educated and thoughtful guesses based on supporting evidence and experience. The evidence

give an impression that we reach a conclusion. Inferences are not achieved by mathematical calculation.Inferences tend to reflect prior knowledge.,experience and personal beliefs and assumptions.

1.5.1 Role of Data Visualization in Inference

Data visualisation has high impact on users,because it communicate with users clearly and efficiently via the statistical and analytical graphics, plots, tables, and charts. An effective and precise visualization helps users to analyse the data and evidence clearly. It makes any complex and huge data more simple, understandable and usable. When presenting an idea that references to data, it is more effective when we use graphs, charts and table. These visual methods can make our point much stronger than simply describing the data.ie,a figure equals more than hundreds of words . Tables are used for compare the values of variable and random access of specific part.Charts are used to show relationships in the data for one or more variables.

Data visualization is both an art and a science. Because we have to present in a impressive manner. The rate at which data is generated has increased ,nowadays. The main challenges in data visualization lies in bigdata processing, analyzing and communicating . Visualization has mainly two important roles in data analysis:
a)Visual comparisons are more effective
b)We can easily summarize estimates and uncertainties

1.6 Open Source Softwares

Open Source softwares are publically accessible they offers more benefits when compared to some commercial products.They are shared and modifiable. Commercial products typically have some features such as reliability, security and similar less glamorous attributes, but they are not modifiable.

Some notable benefits from the use of Open Source Software are following:

- Reliability
- Stability
- Auditability
- Cost
- Flexibility and freedom
- Support and accountability

1.6.1 Big Data Analysis Platforms and Tools

Many free and proprietary tools, platforms, etc are available for analysing big data. Few of the most commonly used technologies and tools are explained in this section.

Hadoop is an open-source framework, which is used for big data processing and storage in a distributed environment.It uses simple programming models. It scalable,so it can accommodate thousands of machines. Operating System: Windows, Linux, OS X.

MapReduce is originally developed by google. It is a programming model and capable to process big data.. It is OS Independent.
GridGrain is an alternative to Hadoop's MapReduce .It is compatible with the Hadoop DFS. It provides fast analysis of real-time data. From GitHub we can download the open source version . O S: Windows, Linux, OS X.

HPCC LexisNexis Risk Solutions developed HPCC. The developers offer higher performance than Hadoop. Both free versions and paid versions are available. O S: Linux.

Storm Twitter is the current owner of Storm. It provides testing life time activities and distributed real-time computation capabilities, more over it is scalable, robust and fault tolerant . It is often known as the "Hadoop of realtime." It goes up with all most everyl programming languages. O S used: Linux.

Databases/Data Warehouses

Cassandra It is originally developed by Facebook, but Apache Foundation is now managing NoSQL database. Many organizations like Netflix, Twitter, Urban Airship, Reddit, Cisco and Digg etc are using it with large dataset. Commercial support and services are provided by the vendors(third party).O S used: OS Independent.

HBase is another Apache project. The non-relational data store for Hadoop. It has so many features lke scalability, consistency in reads and writes, automatic failover support and so. O S: OS independent.

MongoDB is simple and more acceptable open source document oriented databases. It NoSQL schema less database.The main features are high availability, full index support, replication and more. O S used : Windows, Linux, OS X, Solaris.

CouchDB It is designed for the Web applications. CouchDB stores data in JSON documents that can be accessed through browser via HTTP using JavaScript. It offers distributed scaling and fault-tolerant storage. Operating system used: Windows, Linux, OS X, Android.

FlockDB It is kown as Twitter's database and stores graph data (i.e., who is following whom and who is blocking whom). It provides l scaling and fast data input and output. Operating System used: OS Independent.

Business Intelligence

Talend makes a number of business intelligence , data mart and data warehouse products.It supports HDFS, Hive, Hbase ,Pig etc. Operating System used: Windows, Linux, OS X.

Palo is framework of memory resident multidimensional database server. The open source Palo Suite support Palo for Exel, OLAP Server, Palo ETL,Palo Web tc. Jedox also offers the similar tools . Operating System used: OS Independent.

Pentaho offers business and big data analytics tools capable with data mining, reporting and dashboard processing. It is used by more than 10,000 companies,Operating System used: Windows, Linux, OS X.

BIRT is reporting tool for business intelligence. .It is an Eclipse-based tool that adds reporting features to Java applications. Actuate is the supporting company of BIRT,which offers different types of software, based on the open source . Operating System used: OS Independent.

Data Mining

Mahout This Apache project offers algorithms for filtring, clustering and classification which can run on top of Hadoop. The project's goal is to build scalable machine learning libraries.OS used -: OS Independent.

Orange It is a very popular and useful open source tool. It offers a wide variety of visualizations and a toolbox of more than 100 widgets. OS used: Windows, Linux, OS X.

Weka (Waikato Environment for Knowledge Analysis. Weka provides a set of algorithms for data mining . It is part of a larger machine learning project, and it is also sponsored by Pentaho. OS used: Windows, Linux, OS X.

KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source tool(java software tool).It is used for clustering, classification and regression,mainly for knowledge discovery tasks .It helps to evaluate algorithms for data mining problems. It includes many existing algorithms that it uses to compare and with new algorithms. OS used : OS Independent.

Rattle is data mining GUI for R tool,that can be directly available with R.It makes easier for non-programmers to use the R language . It is used to make statistical and visual data summaries , draw graphs, score datasets, build models and more. OSused: Windows, Linux, OS X.

File systems

Gluster FS: It is a free open source scale-out network attached storage file system, which uses more off the shelf hardware. It has a client server component. It gives a shared storage place so it is suitable for cloud computing services. Gluster FS was developed originally by Gluster.
Hadoop DFS: The Hadoop Distributed File System (HDFS) is a distributed file system .Basic features of HDFS are high throughput, highly fault tolarent,suitable for huge data etc.It is designed to run on commodity hardware. It is similar with many existing distributed file systems.

Programming Languages

Pig is a high-level platform for developing MapReduce programs.It is usualyl used with Hadoop. It was originally developed by Yahoo. **Pig Latin** is the language used for this platform. Pig Latin abstracts the programming from the Java.

ECL (Enterprise Control Language) is a declarative, data centric programming language designed and used with HPCC.It is specially designed for data management and query processing.

Summary

- G.R.E.A.T criteria will add value to analysis
- Framing the problem correctly is the seed for good analysis
- Validation using statistical methods are essential for conclusion
- Various open source supports are available for analytics