A REVIEW ON DATA ANALYSIS TECHNOLOGIES AND PLATFORMS

[1] Manoj Kumar, [2]Hemant Rathore
[1]Assistant professor, [2]Assistant professor
[1]Department of Computer Science Engineering
[1]Rajasthan Insitute Of Engineering and Technology,Bhankrota,Jaipur,India

Abstract - Once data have been collected, the next step is to look at and to identify what is going on – in other words, to analyse the data. Here, we refer to "data analysis" in a more narrow sense: as a set of procedures or methods that can be applied to data that has been collected in order to earn one or more sets of results. A list of specific analytical procedures and methods is provided below.
Because there are different types of data, the analysis of data can proceed on different levels. The wording of the questions, in fusion with the actual data collected, have an influence on which procedure(s) can be used – and to what effects.
The task of matching one or more analytical procedures or methods with the collected data often involves considerable thought and reflection. As a GAO report puts it, "Balancing the analytic alternatives calls for the exercise of considerable judgment." This is a rather elegant way of saying that there are no simple answers on many occasions.

Index Terms – Data Analysis, Exploratory Analysis,Trend Analysis, Data Analysis as a Linear Process, Data Analysis as a Cycle Process.

INTRODUCTION

Data is short hand for "information," and whether you are collecting, reviewing, and/ or analysing data this process has always been part of Head Start program operations.
Children's enrolment into the program requires many pieces of information. The provision of health and dental services includes information from screening and any follow-up services that are provided. All areas of a Head Start program – content and management – associate the collection and use of substantial amounts of information. For doing Data Analysis different kind of Analysis performs on data which represented data in different dimension[1].

I. DATA STRATEGIES

Spark is an open source cluster computing (distributed computing) framework. Distributed computing is a field of computer science that studies distributed systems. A distributed system is a model in which components located on networked computers communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal. Spark provides an interface for programing for cluster with implicit data parallelism and fault tolerance. Data parallelism is a form of parallelization of commuting across multiple process it basically focus on distributing the data across different parallel nodes. Data parallelism can be achieved when in multiprocessor system each processor performs the same task on different pieces of distributed data.
Strategy: Visualizing the Data
Involves: Creating a visual "picture" or graphic display of the data.
Reason(s): a way to begin the analysis process; or as an aid to the reporting/ presentation of findings.
Strategy: Exploratory Analyses
Involves: Looking at data to identify or describe "what's going on"? – creating an initial starting point (baseline) for future analysis.
Reason(s): Like you have a choice?
Strategy: Trend Analysis Involves: Looking at data collected at different periods of time.
Reason(s): to identify and interpret (and, potentially, estimate) change.
Strategy: Estimation Involves: Using actual data values to predict a future value.
Reason(s): to combat boredom after you have mastered all the previous strategies. Also to answer PIR and Community Assessment items and tasks[2].

II. VISUALIZING DATA

Visualizing data is to literally design and then consider a visual display of data. Technically, it is not analysis, nor is it a substitute for analysis. However, visualizing data can be a useful starting point prior to the analysis of data.
Consider, for example, someone who is interested in understanding Migrant and Seasonal[3]..

III. EXPLORATORY ANALYSIS

Exploratory analysis entails behold at data when there is a low level of knowledge about a particular indicator (teacher qualifications, first and second language acquisition,etc.)It could also include the relationship between indicators and/or what is the cause of a particular indicator[4].

IV. TREND ANALYSIS

The most general goal of trend analysis is to look at data over time. For example, to discern in case a given indicator such as the number of children with disabilities has increased or decreased over time, and if it has, how quickly or slowly the increase or decrease has occurred.. This form of trend analysis is carried out in order to assess the level of an indicator before and after an event.

V. DATA ANALYSIS AS A LINEAR PROCESS

A closely linear approach to data analysis is to work through the components in order, from beginning to end. A possible advantage of this approach is that it is structured and organized, as the steps of the process are arranged in a fixed order. In addition, this linear conceptualization of the procedure may make it easier to learn. A possible disadvantage is that the step-by-step nature of the decision making may obscure or limit the power of the analyses – in other words, the structured nature of the process limits its effectiveness [5].

VI. DATA ANALYSIS AS A CYCLE PROCESS

A cyclical approach to data analysis provides much more flexibility to the nature of the decision making and also includes more and different kinds of decisions to be made.
In this approach, different components of the process can be worked on at different times and in different sequences – as long as everything comes “together” at the end. A possible advantage of this approach is that program staff are not “bound” to work on each step in order. The potential exists for program staff to “learn by doing” and to make improvements to the process before it is completed.
IT DEPENDS. Rather than chose to present ‘data analysis’ as either linear or cyclical, we have pick to present both approaches. Hopefully, this choice will give MSHS program staff the options and flexibility to make informed decisions, to utilize skills that they already have, and to grow and develop the ability to use data and its analysis to support program/agency purposes and goals[6].

VII. MANAGING THE DATA COLLECTION PROCESS

In order to successfully manage the data collection process, programs need a plan that addresses the following:
- What types of data are most appropriate to answer the questions?
- How much data are necessary?
- Who will do the collection?
- When and where will the data be collected?
- How will the data be compiled and later stored?

By creating a data collection plan, programs can proceed to the next step of the overall process. In addition, once a particular round of data analysis is completed, a program can then step back and reflect upon the contents of the data collection plan and identify “lessons learned” to inform the next round.

VIII. EVALUATION

The final step of the data analysis procedure is evaluation. Here, we do not refer to conducting a program evaluation, but rather, an evaluation of the preceding steps of the data analysis process.
Here, program staff can review and reflect upon:
Purpose: was the data analysis process consistent with federal standards and other, relevant regulations?

Questions: were the questions worded in a way that was consistent with federal standards, other regulations, and organizational purposes? Were the questions effective in guiding the collection and analysis of data?
Data Collection: How well did the data collection plan work? Was there enough time allotted to obtain the necessary information? Were data sources used that were not effective? Do additional data sources exist that were not utilized? Did the team collect? too little data or too much?
Data Analysis Procedures or Methods: Which procedures or methods were chosen?
Did these conform to the purposes and questions? Were there additional procedures or methods that could be used in the future?
Interpretation/Identification of Findings: How well did the interpretation process work? What information was used to provide a context for the interpretation of the results? Was additional compatible data not utilized for interpretation? Did team members disagree over the interpretation of the data or was there consensus? Writing, Reporting, and Dissemination. How well did the writing tell the story of the data? Did the intended audience find the presentation of information effective? [7].

CONCLUSION

Once a set of results has been gain from the data, we can then turn to the explanation of the results.
In some cases, the results of the data analysis speak for themselves. For example, if a program's teaching staff all have bachelor's degrees; the program can report that 100% of their teachers are credentialed. In this case, the results and the interpretation of the data are (almost) identical.
However, there are lots of other cases in which the results of the data analysis and the interpretation of those results are not identical. For example, if a program reports that
30% of its teaching staff has an AA degree, the interpretation of this result is not so clear-cut.
In this case, interpretation of the data associates two parts:
1) presenting the result(s) of the analysis; and 2) providing additional information that will allow others to understand the meaning of the results. In other words, we are placing the results in a context of relevant information. Obviously, interpretation involves both decision making and the use of good judgments! We use the term results to refer to any data obtained from using analysis procedures. We use the term findings to refer to results which will be agreed upon by the data analysis team as best representing their work. In other words, the team may generate a large number of results, but a smaller number of findings will be written up, reported, and disseminated.
On a final note, it is important to state that two observers may legitimately make different interpretations of the same set of data and its results. While there is no easy answer to this issue, the best approach seems to be to anticipate that disagreements can and do occur in the data analysis process. As programs develop their skills in data analysis, they are encouraged to create a process that can accomplish dual goals: 1) to obtain a variety of perspectives on how to interpret a given set of results; and 2) to develop procedures or methods to resolve disputes or disagreements over interpretation.

REFERENCES

[1] Basic statistical tools in research and data analysis
Zulfiqar Ali and S Bala Bhaskar[1] Indian J Anaesth. 2016 October; 60(10): 790.
[2] C. W. Olofson and D. Vesset. Worldwide Hadoop-MapReduce ecosystem software 2012–2016 forecast. May 2012.
[3] D. Fisher, I. Popov, S. Drucker, and m. schraefel. Trust me, I'm partiallyright: Incremental visualization lets analysts explore large datasets faster.
[4] B. Kwon, B. Fisher, and J. S. Yi.Visual analytic roadblocks for novice investigators. In
[5] González-Vidal, Aurora; Moreno-Cano, Victoria (2016). "Towards energy efficiency smart buildings models based on intelligent data analytics". Procedia Computer Science. 83 (Elsevier): 994–999. doi:10.1016/j.procs.2016.04.213
[6] Chekanov, S. (2016) *Numeric Computation and Statistical Data Analysis on the Java Platform*, Springer. ISBN 978-3-319-28531-3
[7] *Grandjean, Martin (2014). "La connaissance est un réseau" (PDF). Les Cahiers du Numérique.* ***10*** *(3): 37–54. doi:10.3166/lcn.10.3.37-54*