Systematic Approaches and Tools for Big Data Analysis in Cloud-based Environment for Managing Healthcare Data

K. Roslin Dayana[1] and M. Vigilson Prem[2]
[1]Assistant Professor, RMD Engineering College, Kavaraipettai
dayana.moncy@gmail.com
[2]Professor, RMD Engineering College, Kavaraipettai
vigiprem@gmail.com

Abstract: There are several data analysis tools invented recently which are powerful to extract the meaningful information from a large set of regularly growing, heterogeneous industrial and other data. Cloud computing technology is used in large scale data analysis to provide scalability, fault tolerance and minimum cost. In this paper, we have done a survey by comparing most widely used big data analysis tools and see that the tools like OpenRefine, Unified Data Analytics Suite (UDAS) and TIBCO Clarity satisfy most of the data analysis functionalities and features like data collection, refinement, delivery, reproducibility and reusability.

Keywords: Data analysis tools, Big Data Analytics, Cloud-based Healthcare data, predictive analytics, machine learning.

Introduction

The different data analysis tools widely used to collect, refine and deliver Big data in Cloud are compared and discussed below. Before this, we will see briefly

about analytics, predictive analytics, machine learning, big data analytics in healthcare, cloud-based big data in healthcare, in the following subsections.

Analytics

Analytics is the systematic use of data and related business insights developed through applied analytical disciplines, for example, statistical, contextual, quantitative, predictive, cognitive, other models.

Evolution of Analytics

The various stages of analytics from 1980's to till date are:

- Descriptive Analytics – It is a preliminary stage of data processing that creates a summary of historical data.
- Diagnostic analytics – It is examining data or content to answer the question "Why did it happen?".
- Predictive analytics – It is the advanced analytics that uses both new and historical data forecast activity.
- Prescriptive analytics – It is the area of business analytics dedicated to find the best course of action for a given situation.

Importance of Analytics in Healthcare

In Healthcare, analytics is used for the following purposes: 1. Collection of large volume of data on a daily basis. 2. Absolute necessity to extract valuable information from them owing to the importance. 3. Time-sensitiveness of the industry. 4. Need for efficient medical diagnosis, prognosis and treatment.

Predictive Analytics

It is developing mathematical models and algorithms that make predictions by applying a large variety of mathematical techniques to historical data. Steps to be followed in predictive analysis are: Understand data, Prepare data, Model, Evaluate, Deploy, Monitor.

Predictive analytics is used in Healthcare industry for the following purposes: 1. Disease management 2. Enhancing patient care 3. Optimizing resource utilization 4. Improving clinical outcomes 5. Increasing income and revenue.

Machine Learning

It is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" with data, without being explicitly programmed.

Types of Machine Learning

The main types of machine learning are: Supervised learning, Unsupervised learning, Reinforcement learning

Big Data Analytics in Healthcare

It is the process of examining large and varied data sets i.e., big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. For example: Real-time Alerting, Telemedicine.

Cloud-based Big Data in Healthcare

Lot of data is produced on a routine basis by hospitals, laboratories, retail, and non-retail medical operations and promotional activities. Cloud computing is used to store and compute medical imaging like radiology, genomic data offloading and collecting Electronic Health Records and increases collaboration and security in a cost-effective manner.

Data Analysis

Data is the prerequisite of any statistical analysis. The computed analysis for the given data will be better for the higher samples. It leads to better decision

making. Efficient decision making improves the operational efficiency and productivity of the organization. Fig 1. shows the steps in data analysis.

Several tools have been developed for doing data collection, data refinement and data delivery in data analysis process. For example, R [1], SAS [2], and SPSS [3]. Many tools will not support all the three major functions of data analysis with the exception of Deducer [4] and SAS. Table 1 shows the classification of existing data analysis tools. They are classified based on functionality, feature and environment.

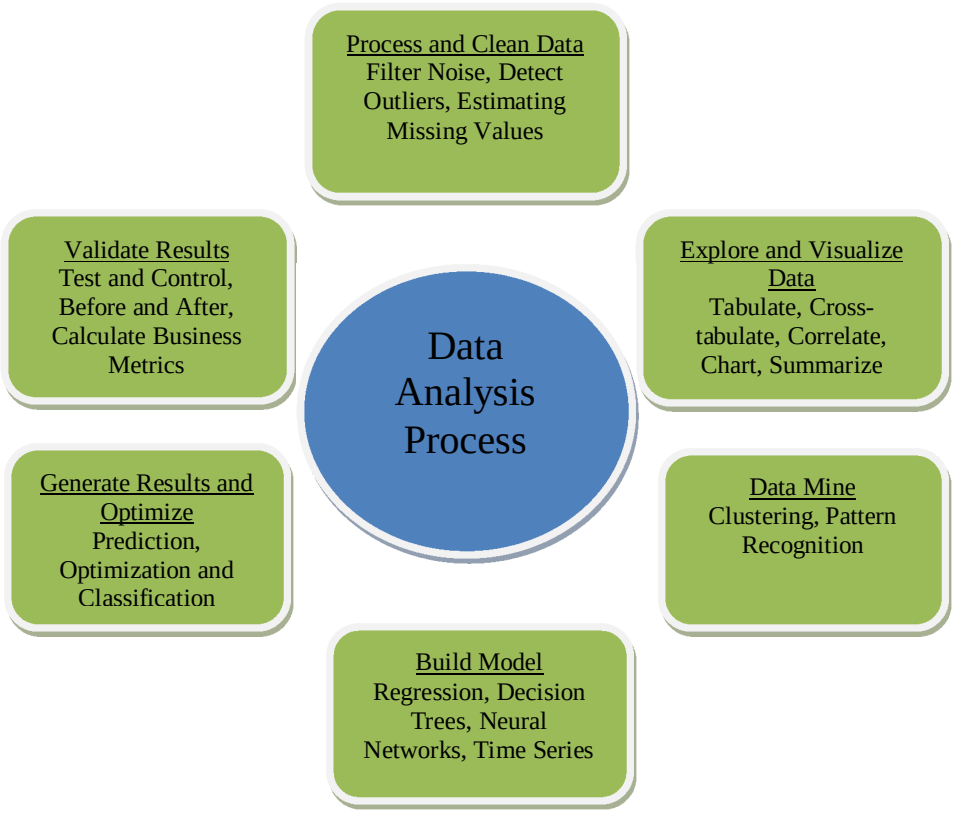

Figure 1. Steps in Data Analysis

Tools and interfaces such as RStudio [5], RKWard [6], JGR [15], RCommander [7]. Rattle [8], and Deducer [4] are developed to be used with R. Apache Spark has released SparkR [9] for interoperability with R. Microsoft has released the Microsoft Machine Learning Server, an open source analytics platform. SRC-

STAT [10] was developed for the users who are unfamiliar with statistical processing and analysis. Stat! [11] refines appropriate queries for data scientists to explore data and provides tables and graphs of data analysis results through multiple queries.

ManyEyes is a web-based system that allows users to collaborate on data analysis and visualization [12]. OpenRefine is the only data analysis tool that provides data cleaning and data transformation, as well as automation for refinement [13]. UDAS provides all the functions necessary for data analysis, such as data collection, refinement, and delivery in distributed cloud-based system [14].

Table 1. Classification of existing Data Analysis Tools

CLASSIFICATION OF EXISTING DATA ANALYSIS TOOLS							
Tools	**Functionality**			**Feature**		**Environment**	
	Collection	**Refinement**	**Delivery**	**Reproducibility**	**Reusability**	**Cloud Service**	**Open Source**
RStudio	WeaklySupports	N/A	Supports	N/A	Weakly Supports	N/A	Supports
RKWard	WeaklySupports	N/A	Supports	N/A	Weakly Supports	N/A	Supports
JGR	WeaklySupports	N/A	Supports	N/A	Weakly Supports	N/A	Supports
RCommander	WeaklySupports	N/A	Supports	N/A	Weakly Supports	N/A	Weakly Supports
Rattle	N/A	N/A	Supports	N/A	Weakly Supports	N/A	Supports
Deducer	Supports	WeaklySupports	Supports	N/A	Weakly Supports	N/A	N/A
SAS		WeaklySupports	Supports	N/A	Weakly Supports	N/A	N/A
SPSS	WeaklySupports	N/A	Supports	N/A	Weakly Supports	N/A	N/A
SRC-STAT	N/A	WeaklySupports	Supports	N/A	N/A	N/A	Supports
STAT	N/A	WeaklySupports	WeaklySupports	N/A	N/A	N/A	N/A
ManyEyes	Supports	N/A	Supports	N/A	WeaklySupports	N/A	Supports
OpenRefine	N/A	Supports	Supports	N/A	WeaklySupports	N/A	Supports
UDAS	Supports	Supports	Supports	Supports	Supports	Supports	Supports
TIBCO Clarity	Supports	Supports	Supports	Supports	Supports	Supports	Supports

TIBCO Clarity provides on-demand software services as Software-as-a-Service. It is used to find, profile, cleanse, and standardize raw data gathered from many sources to supply quality data for accurate analysis and decision-making.

Conclusion

We have compared different data analysis tools. Based on this survey, we see that the tools OpenRefine, UDAS and TIBCO Clarity are very rich in analytic features and we can use them to handle Cloud-based Healthcare system.

References

[1] R Development Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org.

[2] SAS Institute, "SAS/STAT 9.1 User's Guide the Reg Procedure: (Book Excerpt)", SAS Institute Inc., Cary, NC, USA, 2008. "The R Commander: A Basic-Statistics Graphical User Interface to R", *Journal of Statistical Software*, vol. 14, no. 9, pp. 1-42, 2005.

[3] J.Pallant, and S.S.Manuel, A step by step guide to data analysis using SPSS, Berkshire, UK: McGraw-Hill Education, 2010.

[4] I.Fellows, "Deducer:: a data analysis GUI for R", Journal of Statistical Software, vol 49, no.8, pp.1-15, 2012.

[5] RStudio Team, "RSTUDIO: integrated development for R", RStudio Inc., Boston, MA, USA, 2015. Available: http://www.rstudio.com.

[6] S. Rodiger, T.Friedrichsmeier, P.Kapat, and M.Michalke, "RKWard: A Comprehensive Graphical User Interface and Integrated Development Environment for Statistical Analysis with R", Journal of Statistical Software, vol. 49, no.9, pp.1-34, 2012.

[7] J.Fox, "The R Commander: A Basic-Statistics Graphical User Interface to R," Journal of Statistical Software, vol. 14, no. 9, pp. 1-42, 2005.

[8] G.J.Williams, "Rattle: A Data Mining GUI for R", R Journal, vol. 1, no. 2, pp. 45-55, 2009.

[9] S. Venkataraman, Z. Yang, D. Liu, E. Liang, H. Falaki, X. Meng, ... and M. Zaharia, "Sparkr: Scaling r programs with spark," in Proc. of the 2016 International Conference on Management of Data, 2016, pp. 1099-1104.

[10] G.Han, Y.-H. Kim, D. Shin, Y. Lee, and J. Seo, "Design and Development of Visual Analytic Tools in SRC-STAT," *HCI Korea*, pp. 229-231, 2014.

[11] M.Barnett, B. Chandramouli, R. DeLine, S. Drucker, D. Fisher, J. Goldstein, P. Morrison, and J. Platt, "Stat!: An interactive analytics environment for big data," in *Proc. ACM SIGMOD*, New York, NY, USA, 2013, pp. 1013-1016.

[12] F.B.Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, "Manyeyes: a site for visualization at internet scale," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1121-1128, 2007.

[13] www.openrefine.org

[14] Hyunjin Choi, Jangwon Gim, Young-Duk Seo, and Doo-Kwon Baik, " VPL-based Big Data Analysis System:UDAS", IEEE Transactions, Open access, 10.1109/ACCESS.2018.2857845, July 2018.

[15] M.Helbig, M. Theus, and S. Urbanek, "JGR: A Java GUI for R," *The Computing and Graphics Newsletter*, vol. 16, no. 2, pp. 9-12, 2005.