Exploring Big Data Analysis Pipeline and Comparative Summarization of Mining Tools

Ms. Deepali Bajaj[1] Ms. Asha Yadav[2]
[1,2]Shaheed Rajguru College of Applied Sciences for women, University of Delhi

Abstract— From past some years Big data has become a big hype and buzz word in IT industry. It is a major concern of research for data driven industries where massive data needs to be processed and analysed to acquire in-depth knowledge of useful information. It has a huge potential for unlocking the emerging trends, projecting the upcoming growth techniques, increasing productivity, and competitiveness for entire sectors and economies. In this paper, we have reviewed the elements of big data and discussed the phases of processing pipeline of Data Analysis. Further the paper elaborates the most popular and specialized Big Data mining tools as per the survey conducted by KDNuggets in 2015 on analytics and data mining community and vendors. This paper also elucidates the contrasts between two top-ranking tools R and Rapid-I Rapidminer.

Keywords: Rapid-I Rapidminer, KDNuggets, Processing Pipeline Phases

I. INTRODUCTION

We are living in the age of big data, the term which has gained a significant hype from past few years in data science industry. In a wide range of application areas, data is being generated and collected at unmatched scale. The urge to implement a data-driven decision making to gain insights of the current market and industry trend has become an alluring offer for the industries to invest in big data.

"Big Data" is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies" [1].

Some alarming artifacts about big data are:

Wal-Mart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data [2].

Decoding the human genome involves analysing 3 billion base pairs—which took ten years the first time it was done, in 2003, but can now be achieved in one week [3].

Radio frequency identification (RFID) systems used by retailers and others can generate 100 to 1,000 times the data of conventional bar code systems [4].

Facebook handles more than 250 million photo uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc.) – each day [5]. More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide [6].

Big shot of IT industry deem that proper investment in Big Data will definitely lead to a new beckon of essential technological advances that will be made flesh in the next generations of Big Data management and analysis platforms, products, and systems. Most of the industries have recognized role and potential benefit of applying big data techniques. It now drives nearly every aspect of our modern society, including mobile services, manufacturing, retail, finance and many more.

There have been credible cases made for the value of Big Data for urban planning using blend of high-fidelity geographical data, intelligent transportation by analysis of live and detailed road network data, environmental modelling via sensor networks all over collecting data [7], financial systemic risk analysis done by analysing integration of web of contracts to find dependencies between financial entities) [8], and so on and so forth.

Thus people are concerned for effective data analysis to rapidly extract important information from huge data set so as to incorporate value addition to market decision for enterprises and individuals.

The major issues that obstruct the progress of big data implementation in all its phases are heterogeneity, timeliness, and complexity, scale and privacy problems. The problems creeps right away during data acquisition phase, when we need to decide what to keep from the bulk of acquired data although in an ado manner, and what to discard, and format to store key data reliably with the appropriate metadata information. Since the data is collected over from varied sources, it is not natively in structured format; for example, blogs are pieces of text, while images and video are well thought-out for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge.

These 3 V's of big data provide both an opportunity as well challenge for effective implementation and steadfast result derivation from available dataset. These V's are discussed as below:

A. Volume

Many factors contribute to the increase in data volume like transaction-based data stored throughout the years, unstructured data streaming from electronic media or enormous amounts of sensor data being collected. With ever decreasing costs of storage devices, data companies are constantly dumping large data volumes of streaming data from various sources. This leads to another problem that how to use analytics on this plonk of data to extract relevant information.

B. Velocity

Data is generated in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors data and smart metering are driving the need to deal with torrents of data in almost real time.

C. Variety

Data today comes in varied formats. From structured, numeric data in traditional databases to amorphous blogs, text documents, tweets, email, video, audio and financial transactions. Collecting, unifying and managing varieties of data are a throttle task that many organizations still struggle with.

The analysis phase of Big Data involves many distinct stages illustrated below. Each phase is challenging in its entirety. Mostly people focus just on the analysis/modelling phase as this is the most critical phase of analysis pipeline. Further there are poorly understood complexities in analysis phase. Since multiple user program are running simultaneously to dig out relevant data from multi-tenanted clusters. The lack of expertise to seek proper question from data set is yet another hurdle. We can overcome this by supporting many levels of engagement with the data, which may not demand an exhaustive database expertise. It requires a vital reconsideration on how we manage the Data Analysis.

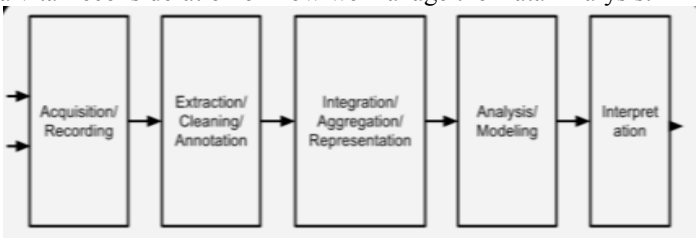

Fig. 1: Major steps of big data analytics pipeline [9].

II. PROCESSING PIPELINE PHASES

A. Data Acquisition

1) Data Acquisition and Recording

Big data is nothing but the normal day to day data generated from various data sources at a volume which exceeds the capacity of traditional analysis. For example, data recorded from geographical sensors, satellites, measuring toxin level of air, patient's diagnostic history, all this and many more accumulate up to approx. 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations have capacity to produce petabytes of data.

It is estimated that the business data volume of all companies in the world may double every 1.2 years [10], in which, the business turnover through the Internet, enterprises to enterprises, and enterprises to consumers per day will reach USD 450 billion [11].

One major confront is to filter data in such a way that it do not discard useful and relevant information. We need research in the area of data reduction that can smartly process this raw data to an extent that its users can handle while not missing the needle in the haystack. Further, "on-line" analysis techniques are required that can process streaming data on the fly, since in such cases store and reduce is not affordable.

The other challenge is to generate the right metadata automatically to describe how and what data is recorded and measured. Such metadata acquisition systems can reduce the human load in recording metadata.

One more issue is data provenance. Recording all the information about the data at its origin is not of use unless this information can be interoperated, tapped and carried next to the data analysis pipeline. Thus we need research in both areas generating apt metadata and data systems that take the provenance of data and its metadata throughout data analysis pipelines.

2) Extraction and Cleaning

The data collected so far is not in a format that can be directly analysed. For example, consider the health records from a hospital, comprising of transcribed dictations from physicians, controlled data from sensors and other machines, image from x-rays etc.

Another example is related to computational biology. GenBank is a nucleotide sequence database maintained by the U.S. National Bio-Technology Innovation Center. Data in this database may double every 10 months. By August 2009, GenBank had more than 250 billion bases from 150,000 different organisms [12].

Above listed examples of unstructured data can't be effectively analyzed. Thus an information extraction process is well required that pulls out the requisite information from the basic sources and depict them in a structured form apposite for analysis. Doing this accurately and entirely is a regular technical confront. Also the analysis dug out from Big Data may not always be correct as it depends on the validity of data (like all other data based analysis). Any missing or mishandled data may lead to accuracy issues in results. Existing methods of data cleaning incorporates well-recognized constraints on valid data and clearly stated error models.

3) Data Integration, Aggregation, and Representation

Data integration is the cornerstone of modern commercial informatics, which involves the combination of data from different sources and provides users with a uniform view of data [13].

Given the heterogeneity of the torrent of data, it doesn't suffice merely to record and throw data into a warehouse. Analysis is significantly more exigent than just locating, identifying, understanding, and citing data. For large-scale analysis, all of these mentioned steps have to ensue in a totally automated way. This requires expressing of data structurally and semantically in such forms that are computer logical, and can be resolved robotically. Data integration can provide some of

the substantial answers. However, additional work is to be done in order to achieve automated error-free difference resolution.

Even for simpler analyses involving only one data set, the question of suitable database design persists. Same information can be stored in various ways and certainly some designs may prove superior over others for certain requirements, and possibly drawbacks for other purposes.

Thus Database design is an art, and is cautiously executed in the enterprise framework by highly-paid professionals, such as domain scientists. They create effective designs, through devising tools to help them in the design progression or they may forgo the design progression entirely and developing techniques to use database effectively in the absence of an intelligent database design.

4) Query Processing, Data Modelling, and Analysis

Big data querying and mining is different from traditional data analysis, as big data is often dynamic, diverse, noisy and unreliable. But the observation of such data obtained via frequent patterns and association analysis often discloses more dependable hidden patterns and facts. Unified Big Data forms big heterogeneous information networks, which can be used to explore information redundancy to compensate for missing values, to find conflicting cases, to disclose intrinsic clusters, and to expose hidden relationships and models.

Mining requires the data to be integrated, cleaned, efficiently accessible and trustworthy. Also it needs declarative query and mining interfaces, big-data computation environment, and scalable mining algorithms.

Big data is growing to leaps and bounds towards the future generation of interactive data analysis in getting real-time answers. Nevertheless the next step could be an auto generated queries for Big Data towards content formation on websites, to populate recommendations, and to supply an adhoc analysis of the value of a data set.

A major challenge with current Big Data analysis is the lack of management between host database systems, which provide SQL querying, and analytics packages which carry out various forms of non-SQL tasks, such as mining and statistical analysis. Nowadays analysts are impeded by a monotonous process of exporting data, performing a non-SQL process and bringing the data back. This is a hindrance to the increasing demand of carrying over the interactive style of the SQL-driven OLAP systems for the data mining. Thus coupling between declarative query and the functions of such packages will advantage both clarity and performance of the analysis.

5) Interpretation

As rightly said any analysis stands incomplete and will lead to futile work unless it could be explained to its end users. Since the ultimate aim is to interpret the result and calculate the decision. In words of Carly Fiorina, Former CEO of HP "The goal is to turn data into information, and information into insight."[14]. through analysis we convert information to insight and interpretation helps to convert this insight to action.

This interpretation cannot occur in a vacuum. It involves investigating all the assumptions made and then retracing the analysis. Nevertheless sources of error in data need to be counted too. For this reasons, no end user will surrender authority to the computer. Rather he will try to comprehend, and authenticate, the results produced by the computer. The system must make it easy for him to do so. This is one of the posed challenges with Big Data due to its complexity.

It is hardly ever enough to provide just the results. Instead supplementary information is also required that can explain how each consequence was derived, and based on what inputs. By studying how effectively we can capture, store, and query supplementary information, in combination with capturing right metadata, we can offer users with the ability to understand analytical outcomes obtained as well as to reiterate the analysis with distinct assumptions, parameters, or data sets.

III. BIG DATA ANALYSIS TOOLS

Big data is distributed over hundreds and thousands of commercial data servers. Many tools are available for big data mining and analysis. Like specialized and amateur software, high-priced commercial as well as open source software. In this section, we here briefly review the top three most widely used software, according to a survey report "R leads RapidMiner, Python catches up, Big Data tools grow, Spark ignites", The survey was conducted by KDNuggets in 2015, on analytics and data mining community and vendors, with about 2,800 voters, who chose from a record number of 93 different tools [15].

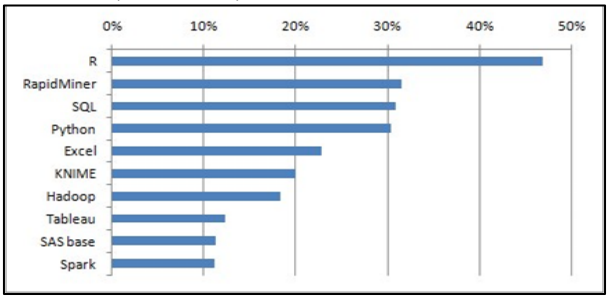

Fig. 2: Top analytics, data mining, data science software used 2015, Top Analytics Tools and Trends [15]

A. R (approx. 46.9%)

R, an open source programming language cum software environment which is designed for data mining and analysis. While computing with R rigorous tasks are executed, code snippets programmed in C, C++ or FORTRAN may be called within the R working environment. R is based on S language (developed by AT&T Bell Labs), it is an interpreted language and is used for data searching, statistical analysis, and drawing graphical plots. R ranks top in the KDNuggets 2015 survey as a big data

analysis tool. With the popularity and usage of R, database manufacturing giants such as Oracle and Teradata; have released their products supporting R.

B. Rapid-I Rapidminer (approx. 31.5%)

Rapidminer is again open source software for data mining and predictive analysis. In a survey of 2011 done by KDNuggets, it was more popularly used than R. Machine learning and Data Extraction programs provided by RapidMiner include a list such as

1) ETL-Extract, Transform and Load
2) Data pre-processing and visualization,
3) Data modelling and result evaluation,
4) Deployment programs

Rapid-Miner is coded in Java. The data mining process is developed in XML and depicted via GUI (Graphical User Interface). The learner and evaluation technique is integrated from Weka (a well know data mining tool). The entire process flow can be represented as a manufacture line of an industrial unit, with original raw data as input and modelling results outcome. The operators used in between can be considered as some functions with different set of input and output characteristics.

Characteristic	RapidMiner	R
Developer:	RapidMiner, Germany	worldwide development
Programming language:	Java	C, Fortran, R
License:	open s. (v.5 or lower); closed s., free Starter ed. (v.6)	free software, GNU GPL 2+
Current version:	6	3.02
GUI / command line:	GUI	both; (GUI for DM = Rattle)
Main purpose:	general data mining	sci. computation and statistics
Community support (est.):	large (~200 000 users)	very large (~ 2 M users)

Fig. 3: Contrast between RapidMiner and R [16]

IV. CONCLUSION

In this paper, we have reviewed the concept of big data thoroughly. Firstly we have explained the general characteristics of Big data along with emphasis on how they pose the technical challenges for the analysis phase of its analysis. Further paper describes the processing steps of data analysis pipelines and challenges that pertain to all the phases. In the remainder of this section, we summarized comparative analysis of top two data mining tools R and RapidMiner.

REFERENCES

[1] Big Data A New World of Opportunities Jun Hou, Lei Xu,"A Testing Tool for Composite Web Services based on data flow", Sixth web information systems and applications conference, 2009.
[2] wikibon.org/blog/big-data-statistics/
[3] www.economist.com/node/15557443
[4] Mayer-Schönberger, Viktor, and Kenneth Cukier. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013.
[5] https://www.ibm.com/developerworks/community/blogs/2251858c-fb45-49e5-9684033c60f33771/entry/what_is_big_data2?lang=en
[6] http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf
[7] A Sustainable Future. Computing Community Consortium. Summer 2011.
[8] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.
[9] http://www.slideshare.net/yuhuang/large-scale-machine-learning-for-big-data
[10] Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute
[11] Gantz J, Reinsel D (2010) the digital universe decade-are you ready. External publication of IDC (Analyse the Future) information and data, pp 1–16.
[12] Bryant RE (2011) Data-intensive scalable computing for scientific applications. Computer Sci Eng 13(6):25–33
[13] Lenzerini M (2002) Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACTSIGART symposium on principles of database systems. ACM, pp 233–246
[14] http://www.analyticshero.com/2012/10/25/31-essential-quotes-on-analytics-and-data
[15] http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html.
[16] http://www.infoivy.com/2014/06/not-all-data-mining-packages-are.html