International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

An Hybrid Model Based Technique in Large Data Analysis Used In Project Management

Shanker Chandre

Research Scholar, Computer Science & Engineering
Sri Sthya Sai University Of Technology & Medical Sciences
Sehore, Bhopal, M.P, (India)

ABSTRACT

Information technology (IT) projects are vulnerable to changes in the business environment, and the increasing rate of change in universal business is challenging the management of enterprise systems such as Data Analysis of project management. At the same time, business success depends on the rigor of the business management processes. Extent creep, poor risk management, inadequate allotment of human resources above time as well as vendor management has a few frequent problems related through the implementation of an enterprise system. These issues pose threats to the winning of a wide-ranging software project, this is Data Analytics. This research adopts a case study approach to observe how poor data management is able to imperil the implementation of a Data Analytics method. Having learned the lessons from the breakdown of its initial Data Analytics implementation, in this case the business is reengineered its data management practices are successively carried out its next Data Analytics implementation. Several critical project management factors are contributed to the collapse and achievement of this company's data Analytics method. The existing models have mostly relied on soft computing techniques and weighting methods. Although they have reduced the complexity and vagueness of Business project attributes, attempts are ongoing to develop more accurate and reliable estimation models. In the research for the development of effort estimation process the analogy-based estimation (ABE) is used. Still it could not conquer data analytics based business data management as there is lot of variance in the weights of attributes depending on data and transactional data used. The process of attribute weighting should be customized based on the nature of project being estimated. To overcome this ABE is enhanced with Genetic algorithm and the data driven attribute selection which uses Bayesian approach. A comparison between the estimates achieved by the proposed model and those obtained from previous are compared.

Keywords: Big data analytics, Datamining,Data warehouse,Hadoop, NOSQL, (APACHE)SPARK

1.INTRODUCTION

1.1 .What is Data Analytics:

Data analytics refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is extracted and categorized to identify and analyze behavioral data and patterns, and techniques vary according to organizational requirements.

Data analytics is primarily conducted in business-to-consumer (B2C) applications. Global organizations collect and analyze data associated with customers, business processes, market economics or practical experience. Data is categorized, stored and analyzed to study purchasing trends.

Evolving data facilitates thorough decision-making. For example, a social networking website collects data related to user preferences and community interests and segment according to specified criteria, such as demographics, age or gender. Proper analysis reveals key user and customer trends and facilitates the social network's alignment of content, layout and overall strategy.

1.2. Benefits of data analytics tools:

The data analytics tools offer many benefits. The main business advantages of big data generally fall into one of three categories: cost savings, competitive advantage, or new business opportunities.

Cost Savings

The data analytics tools like Hadoop allow businesses to store massive volumes of data at a much cheaper price tag than a traditional database. Companies utilizing big data tools for this benefit typically use Hadoop clusters to augment their current data warehouse, storing long-term data in Hadoop rather than expanding the data warehouse. Data is then moved from Hadoop to the traditional database for production and analysis as needed. Versatile big data tools can also function as multiple tools at once, saving organizations on the cost of needing to purchase more tools for the same tasks.

Competitive Advantage

According to a survey of 540 enterprise decision makers involved in big data purchases by Webopedia's parent company QuinStreet, about half of all respondents said they were applying big data and analytics to improve customer retention, help with product development, and gain a competitive advantage. One of the major advantages of big data analytics is that it gives businesses access to data that was previously unavailable or difficult to access. With increased access to data sources such as social media streams and clickstream data, businesses can better target their marketing efforts to customers, better predict demand for a certain product, and adapt marketing and advertising messaging in real-time. With these advantages, businesses are able to gain an edge on their competitors and act more quickly and decisively when compared to what rival organizations do. Needless to say, a business that effectively utilizes big data analytics tools will be much better prepared for the future than one that doesn't understand how important those tools are.

New Business Opportunities

The final benefit of big data analytics tools is the possibility of exploring new business opportunities. Entrepreneurs have taken advantage of big data technology to offer new services in AdTech and MarketingTech. Mature companies can also take advantage of the data they collect to offer add-on services or to create new product segments that offer additional value to their current customers. In addition to those benefits, big data analytics can pinpoint new or potential audiences that have yet to be tapped by the enterprise. Finding whole new customer segments can lead to tremendous new value.

These are just a few of the actionable insights made possible by available big data analytics tools. Whether an organization is looking to boost sales and marketing results, uncover new revenue opportunities, improve customer service, optimize operational efficiency, reduce risk, improve security, or drive other business results, big data insights can help.

1.3.Various Data Analytics Tools:

The following are the various popular data analytics tools used in general for different business applications.

1.3.1.Tableau Public:

Tableau democratizes visualization in an elegantly simple and intuitive tool. It is exceptionally powerful in business because it communicates insights through data visualization. Although great alternatives exist, Tableau Public's million row limit provides a great playground for personal use and the free trial is more than long enough to get you hooked. In the analytics process, Tableau's visuals allow you to quickly investigate a hypothesis, sanity check your gut, and just go explore the data before embarking on a treacherous statistical journey.

1.3.2.OpenRefine:

Formerly GoogleRefine, OpenRefine is a data cleaning software that allows you to get everything ready for analysis. What do I mean by that? Well, let's look at an example. Recently, I was cleaning up a database that included chemical names and noticed that rows had different spellings, capitalization, spaces, etc that made it very difficult for a computer to process. Fortunately, OpenRefine contains a number of clustering algorithms (groups together similar entries) and makes quick work of an otherwise messy problem.

1.KNIME: KNIME allows you to manipulate, analyze, and modeling data in an incredibly intuitive way through visual programming. Essentially, rather than writing blocks of code, you drop nodes onto a canvas and drag connection points between activities. More importantly, KNIME can be extended to run R, python, text mining, chemistry data, etc, which gives you the option to dabble in the more advanced code driven analysis.

2.RapidMiner: Much like KNIME, RapidMiner operates through visual programming and is capable of manipulating, analyzing and modeling data. Most recently, RapidMiner won KDnuggets software poll, demonstrating that data science does not need to be a counter-intuitive coding endeavor.

3.Google Fusion Tables: Meet Google Spreadsheets cooler, larger, and much nerdier cousin. Google Fusion tables is an incredible tool for data analysis, large data-set visualization, and mapping. Not surprisingly, Google's incredible mapping software plays a big role in pushing this tool onto the list. Take for instance this

map, which I made to look at oil production platforms in the Gulf of Mexico. With just a quick upload, Google Fusion tables recognized the latitude and longitude data and got to work.

4.NodeXL takes that a step further by providing exact calculations. If you're looking for something a little less advanced, check out the node graph on Google Fusion Tables, or for a little more visualization try out Gephi.

5.Import.io: Web scraping and pulling information off of websites used to be something reserved for the nerds. Now with Import.io, everyone can harvest data from websites and forums. Simply highlight what you want and in a matter of minutes Import.io walks you through and "learns" what you are looking for. From there, Import.io will dig, scrape, and pull data for you to analyze or export.

6.Google Search Operators: Google is an undeniably powerful resource and search operators just take it a step up. Operators essentially allow you to quickly filter Google results to get to the most useful and relevant information. For instance, say you're looking for a Data science report published this year from ABC Consulting. If we presume that the report will be in PDF we can search it in other form then underneath the search bar, use the "Search Tools" to **NodeXL**: NodeXL is a visualization and analysis software of networks and relationships. Think of the giant friendship maps you see that represent linkedin or Facebook connections. limit the results to the past year. The operators can be even more useful for discovering new information or market research.

7.Solver: Solver is an optimization and linear programming tool in excel that allows you to set constraints (Don't spend more than this many dollars, be completed in that many days, etc). Although advanced optimization may be better suited for another program (such as R's optim package), Solver will make quick work of a wide range of problems.

8.WolframAlpha: Wolfram Alpha's search engine is one of the web's hidden gems and helps to power Apple's Siri. Beyond snarky remarks, Wolfram Alpha is the nerdy Google, provides detailed responses to technical searches and makes quick work of calculus homework. For business users, it presents information charts and graphs, and is excellent for high level pricing history, commodity information, and topic overviews.

II. BIG DATA ANALYTICS:

Big data analytics 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.

The definition of big data holds the key to understanding big data analysis. According to the Gartner IT Glossary, Big Data is high-volume, high-velocity, and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.

Volume refers to the total amount of data. Many factors can contribute to high volume: sensor and machine-generated data, networks, social media, and much more. Enterprises are awash with terabytes and, increasingly,

petabytes of big data. As infrastructure improves along with storage technology, it has become easier for enterprises to store more data than ever before.

Variety refers to the number of types of data. Big data extends beyond structured data such as numbers, dates, and strings to include unstructured data such as text, video, audio, click streams, 3D data, and log files. The more sources that data is collected from, the more variety will be found within data assets.

Velocity refers to the speed of data processing. The pace at which data streams in from sources such as mobile devices, clickstreams, high-frequency stock trading, and machine-to-machine processes is massive and continuously fast moving. The faster that pace becomes, the more data can be analyzed for discovering new insights.

Like conventional analytics and business intelligence solutions, big data mining and analytics helps uncover hidden patterns, unknown correlations, and other useful business information. However, big data tools can analyze high-volume, high-velocity, and high-variety information assets far better than conventional tools and relational databases that struggle to capture, manage, and process big data within a tolerable elapsed time and at an acceptable total cost of ownership.
Organizations are using new big data technologies and solutions such as Hadoop, MapReduce, Hadoop Hive, Spark, Presto, Yarn, Pig, NoSQL databases, and more to support their big data requirements.

2.1.Various Tools of Big data analytics

Apache Hadoop

Hadoop is an open source software framework originally developed by Doug Cutting and Mike Cafarella in 2006. It was specifically built to handle very large data sets. Hadoop is made up of two main parts: the Hadoop Distributed File System (HDFS) and MapReduce. HDFS is the storage component of Hadoop. Hadoop stores data by splitting files into large blocks and distributing it across nodes. MapReduce is the processing engine of Hadoop. Hadoop processes data by delivering code to nodes to process in parallel.

Apache Spark

Apache Spark is quickly growing as a data analytics tool. It is an open source framework for cluster computing. Spark is frequently used as an alternate to Hadoop's MapReduce because it it is able to analyze data up to 100 times faster for certain applications. Common use cases for Apache Spark include streaming data, machine learning and interactive analysis.

Apache Hive

Apache Hive is a SQL-on-Hadoop data processing engine. Apache Hive excels at batch processing of ETL jobs and SQL queries. Hive utilizes a query language called HiveQL. HiveQL is based on SQL, but does not strictly follow the SQL-92 standard.

NoSQL Databases

NoSQL databases have grown in popularity. These Not Only SQL databases are not bound by traditional schema models allowing them to collect unstructured datasets. The flexibility of NoSQL databases like MongoDB, Cassandra, and HBase make them a popular option for big data analytics.

The existing models have mostly relied on soft computing techniques and weighting methods. Although they have reduced the complexity and vagueness of Business project attributes, attempts are ongoing to develop more accurate and reliable estimation models. In the research for the development of effort estimation process the analogy-based estimation (ABE) is used. Still it could not conquer data analytics based business data management as there is lot of variance in the weights of attributes depending on data and transactional data used.

III. EXISTING SYSTEM

Generally it is a challenging task to the I.T. companies w.r.to. managing the enterprise systems such as Data Analysis of Project Management, due to the changes in the business environment. At the same time, business success depends on the rigor of the business management processes. Extent creep, poor risk management, inadequate allotment of human resources w.r.to. time as well as vendor management has a few frequent problems related through the implementation of an enterprise system.

IV. PROPOSED SYSTEM

In this research proposal I observe ,how poor data management is able to imperil (put at risk) the implementation of a Data Analytics method.

Having learned the lessons from the breakdown of its initial Data Analytics implementation, in this case the business is reengineered its data management practices are successively carried out its next Data Analytics implementation. Several critical project management factors are contributed to efficient the company's Data Analytics method.

The existing models have mostly relied on soft computing techniques and weighting methods. Although they have reduced the complexity and vagueness of Business project attributes, attempts are ongoing to develop more accurate and reliable estimation models.

In the research, for the development of effort estimation process the analogy-based estimation (ABE) is used. Still it could not conquer data analytics based business data management as there is lot of variance in the weights of attributes depending on data and transactional data used.

To overcome this, ABE is enhanced with Genetic algorithm and the data driven attribute selection that uses Bayesian approach.

Analogy Based Estimation (ABE):

Analogy Based Software Estimation is based on the principle that actual values achieved within the organization in an earlier and similar project are better indicators and predict the future project

performance much better than an estimate developed afresh from scratch. It also facilitates bringing the organizational experience to bear on the new projects.

However, to use this technique, it is necessary for the organization to put in place certain pre-requisites, such as

1. The organization ought to have executed a number of projects
2. The organization should be keeping meticulous records of past projects
3. The organization must be conducting project post mortem for every project and causes for variances must be identified using meticulous methods and the actual values validated depending on the causes. Care must be taken to prevent erroneous data to influence future projects.
4. The organization should have a well organized and maintained Knowledge Repository from which it is feasible to locate similar past projects and extract the validated project data
5. The estimators should be trained in drawing analogies accurately and in accessing the Knowledge Repository and extracting validated data and extrapolate the same for the current project Once these pre-requisites are in place, this technique can very profitably be used in the organization.

1. Application Domain- This is perhaps the single most important feature to be considered. It would not make sense to draw analogy between two different domains. For example – would it make sense to select a Marketing Information project to draw analogy for a Material management Information project? Therefore, draw analogy from similar application domain.

2. Organization size of the prospective client – The extent of functionality would differ between different sizes of organizations even if the domain is the same. The functionality of Material management, for example, for medium-sized organizations would significantly differ from a large sized organization. Select a past project that is comparable in size with the current project.

3. Number of Locations of the prospective Client– The functionality for a single location would be vastly different for a multi-location organization. Therefore, select a past project that is similar in number of locations of the client organization with the current project.

4. Nature of modules in the application – the past project selected needs to include majority of the modules that the current project has. We can adjust and extrapolate for extra modules for one or two modules but not for a majority.

The following parameters from the development platform need to be considered –

1. Number of application tiers – A two-tier application would significantly differ from a three-tier project.

2. Backend– In present day, almost all applications are built with an RDBMS. As long as the backend is an RDBMS in both the cases it can be considered equivalent. However if one of them is flat files and the other is RDBMS, then they would be different

3. Web Server– Different web servers cause different amount of work. We may need to extrapolate based on the web servers used. This would be applicable in web based application development.

4. Middleware – Different middleware have different impacts on the amount of effort required.

5. Rules Engines– if the proposed project uses a Rules Engine, it would be desirable to select a past project that also used a Rules Engine. Also significant is the fact that different Rules Engines would have different impacts on the a mount of effort required for software development.

6. Programming language – The amount of work is influenced to a large extent by the language in which programs are developed. If the past project used a different programming language than the present project, we may need to adjust the estimate for difference in programming language.

7. Development environment– The type of tools used for editing the programs, debugging, compiling etc have a large impact on the productivity of programmers. Hence it is important to select past projects that have similar software development environments.

8. Software Development Process used – It is also important to select projects that are similar in the manner of developing software conforming to the process that is likely to be used in the current project.

9. Location of Development– Development locations can be either in-house or a t client location. It is better to select a past project that used similar location.

Merits of Analogy based Estimation

1. It is based on actual values achieved within the organization in an earlier project and hence are more reliable than other methods of estimation

2. Easy to learn and very quick to come out with a good estimate

3. For new organizations, they can purchase estimation data from organizations such as ISBSG (International Software Benchmarking Standards Group) and make use of this technique. While these estimates are not from within the organization, they provide variety of estimates to choose from and provide a starting point.

4. This technique facilitates use of organizational expertise and experience to be brought forth for the current project like no other technique of software estimation.

Demerits of Analogy based Estimation

1. The short-listing of past projects and selection of the final project do require meticulous record keeping and software tool support. Any laxity in this step would have serious consequences for the estimate

2. Organization needs to maintain well designed Knowledge Repository and maintain it conforming to meticulous process

3. The current project may not have any relevant past projects at all in the organization .

4. This can not be implemented in a new organization .

V.CONCLUSION

In the research, for the development of effort estimation process the analogy-based estimation (ABE) is used. Still it could not conquer data analytics based business data management as there is lot of variance in the weights of attributes depending on data and transactional data used.To overcome this, ABE is enhanced with Genetic algorithm and the data driven attribute selection that uses Bayesian approach. In this research collecting

the huge amount of data from multiple data sources and applying BIG data analytics to analyze the data easily to develop the business transactions.

ABOUT AUTHOR

SHANKER CHANDRE, Pursuing **Ph.D** in **Computer Science & Engineering** at **Sri Sathya Sai University Of Technology & Medical Sciences**, Worked as a Assistant Professor in CSE Department in various Engineering colleges, Having 10 Years of Teaching Experience. Studied M. Tech. at J.B.Institute Of Engg. & Technology, Hyderabad. B.Tech at Vijay Rural Engg. College, Nizamabad & Interested areas Are: Data Mining, Software Engineering, Network Security, Cloud Computing, and Big Data.

REFERENCES

[1.] Wu X, et al. Data mining with big data. Knowl Data Eng IEEE Trans.2014;26(1):97–107.

[2.] Hashem IAT, et al. The rise of "big data" on cloud computing: review and open research issues. Info Syst. 2015;47:98–115.

[3.] Learning Spark: Lightning-Fast Big Datanalysis" by Holden Karamu.

[4.] Shafer J, Rixner S, Cox AL. The hadoop distributed filesystem: balancing portability and performance. In: IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS); 2010. p. 122–3

[5.] Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. Commun ACM. 2008;51(1):107–13.

[6.] Grossman M, Breternitz M, Sarkar V. Hadoopcl: Mapreduce on distributed heterogeneous platforms through seamless integration of hadoop and opencl. In: IEEE 27th International on Parallel and Distributed Processing Symposium Workshops & Ph.D. Forum (IPDPSW); 2013. p. 1918–27.