Volume 6, Issue 5, May 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

www.ijarcsse.com

Research Paper

Available online at: www.ijarcsse.com

Enchanced Approach for Attainment of BigData Agility

Sheikh Ikhlaq
Research Scholar
Department of Computer Sciences
Suresh Gyan Vihar University, India

Bright Keswani
Associate Professor
Department of Computer Sciences
Suresh Gyan Vihar University, India

Abstract— In early days of BigData what was important, was how fast we could capture the growing data and the Process it .This ability was termed as Data Agility. Now the Scenario has changed what we need are the values from this data that too really quick otherwise it will be nothing but a junk. It means that there is a greater need to bring Agility in BigData Processing. How present Databases can't be used as they can't adapt the changing nature of data. In this Paper we will see the problems associated in bringing Agility to Big Data, Technologies and ways/steps to obtain agility, and how other modes like Cloud computing that can be viable.

Keywords— Data Agility, BigData, Cloud Computing, Data Warehouse, Schemas

I. INTRODUCTION

Ever since the rise of digitization, organizations from various fields have combined huge amounts of digital data, capturing trillions of bytes of information, which ranges from their customers, to suppliers and operations. Volume by which data growing is exponential, due to data generated by machine (data records, web-log files, and sensor data) and from various growing human engagement within the multiple social networks. Variety of data has also increased from text to audio, images and videos. The velocity at with this data is growing is tremendous. All these three V's i.e. Volume, Variety and Velocity have led to collection of mammoth amount of data termed as Big Data. The growth of data can never stop or for that matter it can't be restricted. According to IDC Digital Universe Study published in 2011, 130 Exabyte's of data was produced and stored in 2005. The amount grew exponentially to 1,227 Exabyte's in 2010 and was projected to grow at 45.2% to 7,910 Exabyte's in 2015 [1]·This data can be used to do wonders extracting the hidden wealth of information inside it. On the other side if this data is left as such it will be nothing but garbage. Big Data is considered as a data analysis methodology that is enabled by a new generation of various technologies and architecture that can supports high-velocity data capture, storage, and then analyze it. Data sources have now gone beyond our traditional corporate database, which makes them to include e-mail, mobile output, data generated by sensors, and the output by social media [2]. Data here is no more restricted to structured database items but unstructured data is also included [3].We now know that Big Data needs massive amount of storage space. As the cost of storage continues to reduce, the resources required to deal with big data can still create various financial difficulties for most of the small to medium sized organizations. According to McKinsey, Big Data is considered to be as datasets whose size is outside the capability of typical database software tools. As present there is no plain definition of how large a dataset needs to be in order to be termed as Big Data[4]. "IDC explained that Big Data technologies as a combination of new generation technologies and architectures which are designed so as to extract value economically from very huge volumes heterogeneous data" .This is done by making high velocity detain, discovery and then analysis [5]. "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it"[6].Analysis of big data is done using a called Map Reduce[7].In this Map Reduce, a query is made and then data are mapped to find key values which are considered to relate to the query; the results after that reduced to a dataset which provides answer to the query. Most enterprise data management tools present today (database management systems) is designed to execute basic queries run quickly. Data is indexed so that only very small portions of the data need to be examined in order to answer a query. This solution is not applicable to data that cannot be indexed, i.e., in semi-structured form (text files) or unstructured form (media files). To reply and execute a query successfully in this case, all the data has to be examined [8]·

To explain Map Reduce Further, Map Reduce is a data processing algorithm which uses a parallel programming implementation. In basic terms, Map Reduce is a programming paradigm which involves distributing a work across multiple nodes executing a "map" function. The map function accepts the problems, splits it into sub-parts and then sends them to different machines so that the entire sub-parts can executed concurrently. The results from these parallel map functions are collected and distributed to a set of servers that are running "reduce" functions, which then collects the results from the sub-parts and then re-combines them to get the single answer.

Many of the technologies within the Big Data environment are of an open source origin, due to participation, innovation, development and sharing between the commercial providers that are in open source development projects.

The Hadoop framework based on Map Reduce, in conjunction with additional software components like R language and a wide range of open source Not Only Structured Query Language (NoSQL) tools which include Cassandra and Apache HBase, have become the basics of many Big Data discussions today. Vendors have now launched their own versions of such tools (e.g., Oracle's version of the NoSQL database) [9] or tointegrate these tools with their own products (e.g., EMC's Greenplum Community edition which includes the open source Apache Hive, HBase and ZooKeeper) [9].

The HDFS (Hadoop data file system) is a considered to be fault-tolerant storage system that can store large volumes of information, scale up incrementally and survive on storage failure without losing data. Hadoop clusters are generally built with inexpensive computers. If one node (computer) fails, the cluster can still continue to operate without losing data or for that matter interrupting work by simply re-distributing the task to the remaining machines in the cluster. HDFS manages the storage on cluster actually by breaking files into small blocks and then storing duplicated copies of them over the pool of nodes. Comparing it with other redundancy techniques, including the strategies that are employed by Redundant Array of Independent Disks (RAID) machines found, HDFS has extra two key advantages. Firstly, HDFS does not require any special hardware as it can be built from hardware that is common. Secondly, it enables an efficient method of data processing in the form of MapReduce [10].

In addition to Map Reduce and HDFS, Hadoop also refers to a collection of other software projects that uses the MapReduce and HDFS framework. Some of the tools include: HBase, Hive, Pig, Mahout, Zookeeper, and Sqoop. NoSQL database management systems (DBMSs) are available as open source software and are designed for use in high data volume applications in the clustered environments. They usually do not have any fixed schema and are non-relational, unlike the traditional SQL database management system (also known as RDMS) present in many data warehouses today. Because they don't adhere to a fixed schema, NoSQL DBMS permit us with more flexible usage, allowing high-speed access to both semi-structured and unstructured data. However, SQL interfaces are also being used alongside the MapReduce programming paradigm.

There is a known fact that mid-sized companies and governments can't afford Hadoop as it costs are two high to be beard. Also Hadoop has a limitation that data is still scattered and we can't go for multidimensional view or for that matter we can go for drill down approach .The solution to it to have an approach like warehousing and mining. Since now no new tool for this purpose with respect to Hadoop has not been developed. It was in 1993, much before Big Data Problem, the father of data warehousing, Bill Inmon penned down the definition of data warehouse as: "A data warehouse is a subject oriented, integrated, time-variant, non-volatile collection of data in support of management decisions."[11] Recent advances in computer and networking technology have led to the development of hardware and software platforms that can help to collect, manage and distribute large amount of pertinent data. Data Warehousing is most interesting and dynamic among the new technological transitions available. It works as a repository of subjectively selected and adapted operational data, which can be successfully used to answer any ad hoc, complex, statistical or analytical queries. Also, it provides a mechanism for implementing effective decision support system by utilizing data which scattered all over the organization. Data warehousing provides us with the information that is useful to us from summarized data. Now there was a need of something that could provide us with patterns or trends (Knowledge extraction) in the data that was not immediately apparent by just summarizing the data. This led to us a technology named Data Mining. Data mining is said to be originated from three branches of artificial intelligence i.e. a) neural networks, b) machine-learning and c) genetic algorithms that have brought us to great analytical advancement [12][13][14][15]. It is method which we use to predict the future (predictive analytics) by providing the answer to "How" or "why". One problem with traditional warehouse is that Agile organizations do not have the luxury to take weeks or months to gather and normalize data into a data warehouse, as was historically done. Data is scattered and there is no proper data management [].This approach worked well in the past because data was mostly structured and the concept of a "single source of the truth" was actually attainable (if rarely achieved).With the passage of time agility in warehouses was attainted and these warehouses were termed as agile warehouses. These Warehouses can be used for the cause.

Cloud Computing is a word used to describe a that new class of network based computing which takes place over the Internet or a model that depends on a large, centralized data centre to store and process a wealth of information. Cloud computing has fundamentally been derived by the need to process a huge quantity of data [16] .Cisco has observed that Data today is no longer measured in gigabytes but in Exabyte's as we are "Approaching the Zetta Byte Era" Cloud computing is all intended to access huge amounts of computing power by combining resources and offering a single system view [17]. Cloud computing has become a powerful architecture to perform extensive and complex computing, and has transformed the way that computing setup is abstracted and used. In addition to this, an important goal of these technologies is to deliver computing as a solution for tackling big data, such as large scale, multi-media and high dimensional data sets. Cloud computing is associated with new model for the provision of computing infrastructure and method for processing big data with all kinds of resources. Moreover, some innovative cloud-based technologies have to be implemented because dealing with big data for parallel processing is difficult. Cloud deployment solutions provide services that businesses would otherwise not be able to afford under the traditional hardware and software acquisition method. Cloud computing transforms the way information is handled, the typical organization models for cloud computing includes: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS) and hardware as a service (HaaS).With All these things it's clear that Cloud Computing has inborn feature of being agile.

II. DATA AGILITY

Bringing agility into big data (and small data) analytics, though, has been a challenge for many talented data scientists and engineers around the globe. The reasons are very similar to the struggle in adopting agile application software development: mainly governmental culture and team dynamics. In the space of agile analytics, the problem is further enlarged as the sponsors often go beyond IT to include marketing and other officials.

For many enterprises, their ability to collect data has exceeded their capability to organize it quickly enough for analysis and act. Managers, IS staff, and analysts similar have been irritated with traditional inflexible processes for data processing that require a sequence of steps before data is actually ready for analysis. Relational databases and data warehouses have helped businesses well for collecting and normalizing other relational data from data sources where the data format and schema is known and doesn't change often. Though, the relational model and procedure for defining schema in advance can't keep speed with the rapidly changing variety and format of data.

Sometimes an analyst just wants to start working with data just to understand what it holds and what new understandings it can reveal before the data is actually modeled and then added to the concerned data warehouse schema. Sometimes one is not even sure what questions to ask. This process drives up the budgets for using traditional relational databases and data warehouses due to the fact that DBA resources are mandatory to compress, summarize, and then fully construct the data, and these DBA costs can cause delay access to new data sources. Legacy databases can't be agile enough to meet the ever increasing needs of most organizations today. Hadoop has become a conventional technology for storing and processing large amounts of data, but now the discussion has changed. Presently, it's not about how much amount of data you can store and process. Rather, it's about *data agility*, which means how fast one can extract value from your peaks of data and how quickly one can translate that information into achievement? After all, there is still a need for someone to apply schema to the data after which it can be analysed. Getting data *into* Hadoop easily is never a guarantee an analyst can easily get it *out*.

Managers need their teams to focus on business effect, not on how they should store, process, and analyse their data. Now the some questions arise, how does the ability to process and analyse data impact their operations? How swiftly can they adjust and respond to changes in customer favourites, market situations, economical actions, and operations? These questions are supposed to direct the investment and also the scope of big data projects in 2015 as enterprises change their focus from just capturing and managing data to vigorously using it.

This concept can be not just applied to big data infrastructure; it can also be applied across all business activities, ranging from risk management to marketing movements to supply sequence optimization.

In the beginning when the concept of data agility was first talked about, the discussion focused on an organization's ability of how quickly they could gather business intelligence. However, the concept of data agility can also apply to data warehouse architecture. With traditional i.e., data warehouse architectures based on relational database systems, the data schema needs to be carefully designed and maintained. If the schema is changed, it can sometimes take up to a full year to make the changes to an RDBMS. After all this , the process of extracting data from a data store and then loading it into a data warehouse can take up an entire day before it's made available to be analyzed.

Hadoop makes storing a new kind of data easy as it doesn't mean having to redesign the schema. It is as simple as creating a new folder and then moving the new type of files to that newly formed folder. Hadoop can therefore be used for storing and processing data, further teams can develop products in a much smaller timeframe.

III. REAL HINDRANCE TO DATA AGILITY

Traditional databases need a predefined schema before writing data. Pairing that with the time which is needed to get the data into the database and then the process can no longer be considered as agile. There are worse times when those DBAs need to perform complicated processes which require dropping of foreign keys or exporting data, altering table designs, and then even reloading data in a specific order to that it can satisfy the table design. There are some of big data technologies like Apache Hive which are able to get about the schema-on-write but still they require defining a schema much before the users can ask for the very first question.

IV. KNOWING AGILITY BY KNOWING IT

New technologies for data discovery and data exploration are being developed so they provide greater flexibility. Apache Drill is a great that can bring data agility. Inspired primarily by Google's Dremel , Dremel is a scalable, interactive ad-hoc query system for analysis of read-only nested data. By combining multi-level execution trees and columnar data layout, it is capable of running aggregation queries over trillion-row tables in seconds. The system scales to thousands of CPUs and petabytes of data, and has thousands of users at Google[18]. Apache Drill is an open source, low-latency SQL query engine developed for Hadoop and NoSQL that can be used to query across data sources. It can handle both flat fixed schemas and semi-structured/nested data. Drill is an Apache open-source SQL query engine for Big Data exploration. Drill is designed from the ground up to support high-performance analysis on the semi-structured and rapidly evolving data coming from modern Big Data applications, while still providing the familiarity and ecosystem of ANSI SQL, the industry-standard query language[19]. Drill provides plug-and-play integration with existing Apache Hive and Apache HBase deployments.

Drill is like opening the door to this inevitable future of short timed life cycles for data processing so that it can support faster responses to various opportunities and threats. What it matters I the end is, the faster one can ask a question and get the right answer, the better for business.

Drill has the ability to implement schema-on-the-fly, means that whenever a new data format arrives, nothing needs to be done to process the data with Drill. Now we don't need DBAs to build and maintain schema designs. Generally used Commercial off-the-shelf business intelligence tools can connect easily with Drill as Drill implements standards. It is ANSI SQL: 2003-compliant and comes with JDBC and ODBC drivers. This means that business doesn't need to adopt any new tools to work with all types of data from all data sources.

There are certain opposing views on this new technology which need to be considered. The question that may arise here is: Why do we need these new technologies? The main change in the industry gets on the deployment of data interchange formats such as JSON (JavaScript Object Notation) which is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate. Data which comes from various applications that issue data in JSON do not need a DBA so as to structure the incoming data because it displays up already structured, thus eliminating the process bottleneck[20].

Drill boosts data agility by allowing various users to perform self-service data ingestion and data source management, whether if it's due adding a new data source or adapting for a change in the incoming data structure.

V. AGILITY IN ENTERPRISE

Data agility would be an important feature of all your big data initiatives in the future. Individuals would be able to analyse and explore data directly. Self-service data exploration removes the dependency on IT to create data definitions and structures, and also frees up IT staff to perform more valuable activities.

By bringing agile technologies such as Hadoop and Apache Drill on board, into your enterprise and existing data management and analytics capabilities, guides organization's agility towards real-time business impact.

Even though we have above mentioned technologies for bringing agility but we need to have a proper ways to utilize them. We are exactly in the middle of a big data technology race that will decide the winners and the losers across many areas of private enterprise and public life. The companies, candidates, and organizations who express the most agility in adapting to the new big data environment have the great chances of surviving and getting success in the future.

VI. WAYS TO ATTAIN AGILITY

In present chaotic digital atmosphere, keeping up with the exponential growth or for that matter explosion of new data is a huge task. There is no surprise that organizations are in need and want to make more and better use of this new data and data sources, but getting started can be difficult. To remain agile, businesses need to change the views they ponder about data and analytics. Here are six ways/steps organizations can start filling themselves with big data agility.

Step One: Having Big Data Team

Generally, we would hire a data scientist who has the right amount of math, statistics, business, and also programming skills to lead big data efforts. The only problem here is that data scientists are tough to find and the price they come for. Universities are moving to rise up production of data scientists, but that takes years which organizations can't afford don't have.

While there is a need of Data scientists for some types of big data works, agile organizations are know that they can still drive success with that thing: a big data team with a mature blend of technical, analytical, and of course business skills. It is sure that in the future, business analysts and data analysts are going to become the main force of big data projects at almost every organization.

Step Two: Transition from Old Data Warehousing to Agile Data Warehousing

The second step needed in achieving big data agility to move ahead from old concepts of traditional data warehousing (even though not necessarily the technology; column-oriented data stores, in particular, will have a bright future in big data analytics). Agile organizations do not have time to gather and normalize data into a data warehouse, as was traditionally done. This approach was a success in the past as data was mostly structured.

Today data is heterogeneous in nature i.e. the semi-structured and unstructured data which is being generated by the Web, smartphones, the Internet of Things and other data producing sensors. Thrashing these huge data sets into a structured format can neither feasible nor desirable. A more flexible and agile approach is absolutely required. Hadoop is better suited for storing this type of data but agility still remains a concern therefore agile data warehousing is an optimal option.

Step Three: Scheme for Flexible Schemas

We need to adopt more agile schema-on-read methodologies used in agile Data warehouses, as opposed to schema-on-write or schema-on-load which were used in traditional Data warehouses. When we talk about schema on write or schema on load, we mean that, for a given set of datasets or dataset, the schemas–and also the relationships between the datasets–are all well understood and defined at the time the data is loaded.

The concept of schema on read or schema on use, there's the knowledge where we will load data into the database–or, more probable, Hadoop–so that data is kept in its raw format onto the file system or in the data store, when an analyst wants to use that data, at the time they use the data, that's only when they'll start structuring the data and defining its schema and then potentially defining relationships between this data set with other data set in universe of data sets.

Step Four: Concentrating On Data That Is Quick

As explained above how various open source technologies are enabling us to store heterogeneous data. NoSQL databases have brought new capabilities to store and crunch huge amounts of both semi-structured and unstructured data.

But in many cases, agile organizations are separated from slow-footed ones by the ability to process and react to rapidly moving data in a real time. There's a great opportunity for organizations to find more and valuable data from where patterns can be found and on which to react in real time

Step Five: Management of Flexible Workflows

Once big data sources are identified who are going to drive new analytic decision-making machine (perhaps residing in Hadoop, perhaps cloud), everything needs to be put into action. Big challenges in the agile world are not exactly around modelling, predicting, and forecasting rather there are some challenges, especially when to get to large-scale data and in this early phase of analysis, i.e. data transformation, data attack, or data wrangling, data cleaning.

There are many data transformation tool, including Trifacta's, which aims at helping organizations do the data preparation(early phase) work necessary for analytics without bringing people with advanced data science degrees on board. As data becomes quicker and more versatile, agile organizations will always look to this particular class of tools so that they can keep the data science characteristics of big data in check, while they take full advantage of automation and repeatability

Step Six: Taking Advantage of Cloud Computing for its Agility and Cost

Cloud Computing as we know is the new phenomenon that has revolutionized the world of computing .Now we are in an era where we don't have to worry about the space or computing power or the changing trends of applications or there development framework. Cloud Computing is proving to be a boon for data scientists .One major advantage of cloud computing is that there is no limit of adding infrastructure or bringing changes in the existing environment to fit whatever we want to. This means that Big Data and the problems associated with its computation can easily be dealt with this technology. Also Cloud computing is much more economical that implementing Hadoop. Implementing various open source technologies on cloud environment will unearthin the gold for us which is hidden in the mountains of data.

VII. CONCLUSIONS

Since much has been done to solve the problem of heterogeneity of data and the Big Data agility by bringing the technologies like Apache Drill, JSON, Dremel and Cloud Computing who perform under certain standards, are playing their role in achieving agility but still not much work has been to remove the randomness in data which hamper the agility .Also we can use agile warehouse to perform data mining without modifying the existing algorithms, methods or tools. Presence of Randomness can hamper gold mining too, this means that we still don't have a perfect data management solution. Work needs to be done in pre analysis stage so that the results which will be achieved are more agile and perfect.

REFERENCES

[1] International Data Corporation, (2011), "*The 2011 Digital Universe Study: Extracting Value from Chaos*", Accessed at: http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

[2] R. L. Villars, C. W. Olofson, and M. Eastwood, "*Big Data: What It Is and Why You Should Care*," White Paper, IDC, 2011.

[3] C. Coronel, S. Morris, & Rob,"*Database Systems: Design, Implementation, and Management*", (10th Ed.). Boston: Cengage Learning P, 2013.

[4] J.Manyika, etal., "*Big data: The next frontier for innovation, competition,andproductivity*".Accessedfrom:http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[5] IDC Market Analysis, (2012), "*Worldwide Big Data Technology and Services 2012 – 2015 Forecast*", Accessed at: http://www.idc.com/research/viewtoc.jsp?containerId=233485

[6] E. Dumbill, "*What is big data?*", Accessed from: http://radar.oreilly.com/2012/01/what-is-big-data.html

[7] Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, "*Understanding big data: Analytics for enterprise class Hadoop and streaming data*", New York: McGraw-Hill,2012.

[8] J. Dean, S. Ghemawat, "*MapReduce: Simplified Data Processing on Large Clusters*", Accessed from: http://static.usenix.org/event/osdi04/tech/dean.html

[9] D. Henschen, (2012), "*Oracle Releases NoSQL Database, Advances Big Data Plans*", Accessed at: http://www.informationweek.com/software/informationmanagement/oracle-releases-nosql-database-advances/231901480

[10] C. W. Olofson, D. Vesset (2012), "*Worldwide Hadoop – MapReduce Ecosystem Software 2012-2016*",Accessed at : http://www.idc.com/getdoc.jsp?containerId=234294

[11] W.H. Inmon, "*Building the Data Warehouse",* John Wiley. pp. 33, 1996

[12] M.S. Chen, J. Han, and P.S. Yu, " *Data Mining: An Overview from a Database Perspective*", IEEE Transactions on Knowledge and Data Engineering, 8(6), pp. 866-883, 1996.

[13] R, Agrawal , R. Srikant, "*Mining Sequential Patterns*", In: Yu P, Chen A (Eds).Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan, pp.3–14,1995.

[14] R.P, Schumaker, et al., "*Sports Data Mining Methodology, Sports Data Mining, Integrated*", In: Information Systems 26, Springer Science+ Business Media, LLC,2010.

[15] S. Shabia, M.A.Peer, "*Expedition for the exploration of Apposite Knowledge*", International Journal of Computer Science and Information Technologies, 3 (5),pp.5164 – 5168,2012.

[16] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, M. Zaharia, "*A view of cloud computing*", Communications of the ACM, 53(4),2010.

[17] http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481374_ns827_Networking_Solutions_White_Paper.html

[18] http://research.google.com/pubs/pub36632.html

[19] https://drill.apache.org/docs/drill-introduction/

[20] http://www.json.org/

[21] S. Ikhlaq and B. Keswani.*"Computation of Big Data in Hadoop and Cloud Environment*",IOSR Journal of Engineering (IOSRJEN),6(1),31-39,2016