IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. V (May-Jun. 2016), PP 120-123
www.iosrjournals.org

Big Data Implementation Challenges and Solutions

G.Narendra[1], Dr.B.V. Ramana Reddy[2] , M.Jeevan Kumar[3]
[1,2,3]*(Cse, Nec Nellore/ Jntu Ananthapuramu, India)*

Abstract: *Big Data – Adoption challenges and solutions Enterprises are encumbered by certain challenges in their Big Data adoption journey. Some of these challenges are: Inhibition in making the first move for a particular use case Reluctance towards making Big Data strategy investment in the current financial year Absence of any single Big Data vendor and Integration of already available traditional data with Big Data q and Scarcity of combined Big Data and domain skills Our research indicates that the biggest challenges while deriving business value from Big Data are as much cultural as technological. It is imperative for organizations experimenting with Big Data to have a thought-through enterprise-wide data strategy, which caters to the integration of Big Data with already existing traditional data. Entering the Big Data arena – Is it Easy? From big vendors to small boutique firms, organizations are identifying lists of Big Data capabilities and offerings that they provide. Technology giants have made significant investments in software, infrastructure and R&D, foreseeing the tremendous opportunities that the Big Data future holds. Numerous lean startups are offering niche and customized Big Data solutions.*

Keywords: *Big data, velocity, volume, secondary node, data node*

I. Introduction

Big Data, an 'in the news' technology, is gaining attention from organizations looking to seize early applicant opportunities. Namely Volume, Variety, and Velocity, are the key characteristics that are fundamental to its evolution. Current definitions qualify any data that is difficult to manage using traditional systems as Big Data. Big Data has evolved from a bid to derive value out of huge volumes of available unstructured data - overlooked until now because of the existing systems' inability to process them. Big Data has application opportunities across all value industries. Adoption of Big Data in sales and marketing is gradually gaining power primarily due to the underlying characteristics of the function that necessitates interfacing through the web to listen to the 'voice of people'. Big Data has huge potential in Research and Development because of its intrinsic ability to process data from multiple sources, such as millions of documents, protocols, study records, images, and applications; and provide a unified view. Like every nascent technology, Big Data adoption entails some challenges. In addition to planning a significant investment in this evolving technology, collecting data elements lying in organizational silos introduces a cultural challenge in Big Data initiatives.

Recruiting and retaining the right workforce with a balance of domain knowledge and Big Data skills is another challenge organizations face today. It is imperative for any organization to formulate a well-defined strategy for its Big Data implementation to ensure alignment with its business objectives. Given that Big Data is still in its early days, organizations looking to evaluate it can begin small – by taking up proofs of technology and factoring in multiple iterations. The recommended approach to adopting Big Data is to start slow, realize benefits, pause, think and take the next step, and reap subsequent benefits. Big Data Implementation – Does it require a specialized skill set? It is imperative for organizations implementing Big Data projects, to understand the importance of the special skills required. Since Big Data is not a single technology, skills relevant to it cannot be acquired in silos or through traditional training methods. Organizations implementing Big Data initiatives will require the expertise of data scientists, system analysts, infrastructure analysts, domain experts, technology implementers, solution architects, data integrators, reporting and analytics experts and software developers among others.

An organization taking up a Big Data implementation can adopt a step-wise approach for acquiring Big Data skills. As a first step, organizations should establish enterprise-wide awareness about Big Data and its capabilities. Data scientists, system analysts and domain experts can then work towards defining the problem statement and corresponding solution. Technology implementers, data integrators, reporting and analytics experts can then implement custom Big Data solutions. The Big Data solution architect presents the integrated view of the problem statement.Big Data Technologies – Making Choices Although we have presented a variety of use cases, their relevance, impact and potential business value change based on each organization's context and maturity. Factors that influence the Big Data adoption decision to determine its success include the current environment of the organization, relevant data sources to be accessed and the insights needed. Organizations must opt for Big Data tools/technology based on the type of data sources – a high-performance message delivery data system, or data in motion, or static data as well as the threshold on infrastructure costs and performance

benchmarks. Exhaustive due diligence for vendor and tools selection should be carried out. It is extremely important to verify the fitment of Big Data in the existing enterprise IT landscape.

Is cost the only consideration for adoption? In a turbulent economy, businesses need to justify the cost involved in embracing any new technology platform vis-à-vis its ROI. Innovation can be sustained if it harmonizes capability and cost. The cost consideration to adopt new technology depends on various factors such as implementation cost, maintenance cost, skills availability, and upgradability, all of which businesses need to evaluate within the current system. If the cost factors of a technology changeover are identical to those of the existing technology, then questions arise regarding its cultural acceptance. While on the subject of Big Data solutions and their cost effectiveness, the market is still evaluating both Hadoop solutions and non Hadoop solutions. A Hadoop ecosystem utilizes massive parallel processing computing techniques using just commodity grade desktop machines. This capability of the Hadoop platform overcomes the challenge of optimum performance in a cost-effective manner and has been proven to be highly scalable. NonHadoop solutions are largely customized solutions for addressing specific and specialized requirements. Cost consideration has also taken a step back for use cases where the requirement is to process near real-time data to be able to generate more advanced analytics.

II. Frameworks in Big Data Image processing

In the last decade, Hadoop has become a standard framework for big data processing in the industry. Although Hadoop today is primarily applied to textual data, it can be also used to process binary data including images. A number of frameworks have been developed to increase productivity of developing Hadoop based solutions. The amount of images being uploaded to the internet is rapidly increasing, with Facebook users uploading over 2.5 billion new photos every month, however, applications that make use of this data are severely lacking. Current computer vision applications use a small number of input images because of the difficulty is in acquiring computational resources and storage options for large amounts of data. As such, development of vision applications that use a large set of images has been. Many image processing and computer vision algorithms are applicable to large-scale data tasks. It is often desirable to run these algorithms on large data sets (e.g. larger than 1 TB) that are currently limited by the computational power of one computer. These tasks are typically performed on a distributed system by dividing the task across one or more of the following features: algorithm parameters, images, or pixels. Performing tasks across a particular parameter is incredibly parallel and can often be perfectly parallel. Face detection and landmark classification are examples of such algorithms. The ability to parallelize such tasks allows for scalable, efficient execution of resource-intensive applications. The Map Reduce framework provides a platform for such applications.

The Hadoop Map reduce platform provides a system for large and computationally intensive distributed processing, though use of Hadoop system is severely limited by the technical complexities of developing useful applications. With the proliferation of online photo storage and social medias from websites such as Face book, Flickr, and Picasa, the amount of image data available is larger than ever before and growing more rapidly every day. This alone provides an incredible database of images that can scale up to billions of images. Incredible statistical and probabilistic models can be built from such a large sample source. For instance, a database of all the textures found in a large collection of images can be built and used by researchers or artists. The information can be incredibly helpful for understanding relationships in the world1 . If a picture is worth a thousand words, we could write an encyclopedia with the billions of images available to us on the internet. These images are enhanced, however, by the fact that users are supplying tags (of objects, faces, etc.), comments, titles, and descriptions of this data for us. This information supplies us with an amazing amount of unprecedented context for images. Problems such as OCR that remain largely unsolved can make bigger strides with this available context guiding them.

It is these reasons that motivate the need for research with vision applications that take advantage of large sets of images. Map Reduce provides an extremely powerful framework that works well on data-intensive applications where the model for data processing is similar or the same. It is often the case with image-based operations that we perform similar operations throughout an input set, making Map Reduce ideal for image-based applications. However, many researchers find it impractical to be able to collect a meaningful set of images relevant to their studies. Additionally, many researchers do not have efficient ways to store and access such a set of images. As a result, little research has been performed on extremely large image-sets. Hadoop uses a distributed file system to store files on various machines throughout the cluster. Hadoop allows files be accessed, however, without knowledge of where it is stored in the cluster, so that users can reference files the same way they would on a local machine and Hadoop will present the file accordingly. When performing Map Reduce jobs, Hadoop attempts to run Map and Reduce tasks at the machines were the data being processed is located so that data does not have to be copied between machines.

As such, Map Reduce tasks run more efficiently when the input is one large file as opposed to many small files . Large files are significantly more likely to be stored on one machine whereas many small files will

likely be spread out among many different machines, which requires significant overhead to copy all the data to the machine where the Map task is. This overhead can slow the runtime ten to one hundred times. Simply put, the Map Reduce framework operates more efficiently when the data being processed is local to the machines performing the processing.

III. Challenges of Big Data Analytics in the industry

Big data is now a reality: The volume, variety and velocity of data coming into your organization continue to reach unprecedented levels. This phenomenal growth means that not only must you understand big data in order to decipher the information that truly counts, but you also must understand the possibilities of big data analytics. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. With big data analytics, data scientists and others can analyze huge volumes of data that conventional analytics and business intelligence solutions can't touch. Consider that your organization could accumulate (if it hasn't already) billions of rows of data with hundreds of millions of data combinations in multiple data stores and abundant formats. High-performance analytics is necessary to process that much data in order to figure out what's important and what isn't. Enter big data analytics. Why collect and store terabytes of data if you can't analyze it in full context? Or if you have to wait hours or days to get results? With new advances in computing technology, there's no need to avoid tackling even the most challenging business problems. For simpler and faster processing of only relevant data, you can use high-performance analytics. Using high-performance data mining, predictive analytics, text mining, forecasting and optimization on big data enables you to continuously drive innovation and make the best possible decisions. In addition, organizations are discovering that the unique properties of machine learning are ideally suited to addressing their fast-paced big data needs in new ways.

Big Data Analytics Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools. The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing. These advances have created several differences between traditional analytics and Big Data analytics. Big Data technologies can be divided into two groups: batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion. Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies like Drill and Dremel provide new paradigms for data analysis.

Hadoop is one of the most popular technologies for batch processing. The Hadoop framework provides developers with the Hadoop Distributed File System for storing large files and the MapReduce programming model which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized. Several tools can help analysts create complex queries and run machine learning algorithms on top of Hadoop. These tools include Pig (a platform and a scripting language for complex queries), Hive (an SQL-friendly query language), and Mahout and RHadoop (data mining and machine learning algorithms for Hadoop).

New frameworks such as Spark 4 were designed to improve the efficiency of data mining and machine learning algorithms that repeatedly reuse a working set of data, thus improving the efficiency of advanced data analytics algorithms. There are also several databases designed specifically for efficient storage and query of Big Data, including Cassandra, CouchDB, Greenplum Database, HBase, MongoDB, and Vertica. Stream processing does not have a single dominant technology like Hadoop, but is a growing area of research and development. One of the models for stream processing is Complex Event Processing, which considers information flow as notifications of events (patterns) that need to be aggregated and combined to produce high-level events. Other particular implementations of stream technologies include InfoSphere Streams5 , Jubatus6 , and Storm7 .

Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view. Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses. Off-the-shelf Big Data tools and techniques are now bringing attention to analytics for fraud detection in healthcare, insurance, and other fields.

IV. Conclusion

Hadoop analyze big data but faces some problem of storage and computation power, storage will be overcome with help of HDFS module and processing will run under the control of Map Reduce, The technological advances in storage, processing, and analysis of Big Data include , the rapidly decreasing cost of storage and CPU power in recent years; the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing.

References

[1]. Garlasu,D.Sandulescu,V.; Halcu,I.;Neculoiu,G.; Grigoriu,O.; Marinescu,M.; Marinescu, V., " A big dat a implement at ion based on Grid comput ing" , Roedunet Internat ional Conference (RoEduNet), 2013
[2]. C.Chandhini, MeganaL.P , "Grid Comput ing-A Next Level Challenge wit h Big Dat a" , Internat ional Journal of Scient ific & Engineering Research , Issue3, Mar-2013
[3]. A big data implementation using gridcomputing",www.pantechproed.com/download_project .php
[4]. An Oracle White Paper March 2013 , " Big Data analytics : Advanced analytics in oracle dat abase"
[5]. "Ideas Economy: Finding Value In Big Dat A", www.oracle.com/us/technologies/big-data/index.html
[6]. Ann Chervenak; Ian Foster; Carl Kesselman; Charles Salisbury; Steven Tuecke, "The Dat a Grid: Towards an Archit ect ure for the Dist ribut ed Management and Analysis of Large Scient ific Dataset s"
[7]. Apachen Hadoop" http://en.wikipedia.org/wiki/Apache_Hadoop
[8]. GridComputing"http://en.wikipedia.org/wiki/Grid_computing