International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 5, Issue 3 (March 2016), PP.32-38

Survey on Real Time Data Analytical Structural design

[1]Mr.Jayashakthivelmurugan, [2]Sharon Sushantha.J
[1]Associate Professor, Jeppiaar Engineering College [2]PG Student Jeppiaar Engineering College

Abstract:- *"Big Data" is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value. It requires new data architectures, analytic sandboxes, new tools, new analytical methods, integrating multiple skills into new role of data scientist. Organizations are deriving business benefit from analyzing ever larger and more complex data sets that increasingly require real-time or near-real time capabilities .Big data can come in multiple forms. Everything from highly structured financial data, to text files, to multi-media files and genetic mappings. The high volume of the data is a consistent characteristic of big data. Remote sensing is the acquisition of information about an object or phenomenon without making physical contact with the object and thus in contrast to on site observation. Remote sensing is a sub-field of geography. In modern usage, the term generally refers to the use of aerial sensor technologies to detect and classify objects on Earth (both on the surface, and in the atmosphere and oceans) by means of propagated signals (e.g. electromagnetic radiation). It may be split into active remote sensing (when a signal is first emitted from aircraft or satellites) or passive (e.g. sunlight) when information is merely recorded. Since the volume of data recorded by remote sensing application is massive we use big data techniques to manage the data.*

Keywords:- *Bigdata, Hadoop, Mapreduction, data streaming, parallel computing, Remote Sensing*.

I. INTRODUCTION

The emergence of big data application has had a major impact on the current internet world.

More and more services are emerging on the internet and huge number of service relevant elements are generated and distributed across the network which cannot be effectively addressed by the traditional databases. In remote access systems, the information source, for example sensors can deliver a staggering measure of crude information. We refer it to the first step, i.e., data acquisition, in which much of the data are of no interest so they can be filtered or compressed by orders of magnitude. But with a view to using such filters, they do not discard useful information. For instance, in consideration of new reports, is it adequate to keep that information that is mentioned with the company name? Alternatively, is it necessary that we may need the entire report, or simply a small piece around the mentioned name? The second challenge is by default generation of accurate metadata that describe the composition of data and the way it was collected and analyzed. Such kind of metadata is hard to analyze since we may need to know the source for each data in remote access.Timely and cost-effective analytics over "Big Data" is now a key ingredient for success in many businesses, scientific and engineering disciplines, and government endeavors. The Hadoop software stack—which consists of an extensible MapReduce execution engine, pluggable distributed storage engines, and a range of procedural to declarative interfaces—is a popular choice for big data analytics. Most practitioners of big data analytics—like computational scientists, systems researchers, and business analysts—lack the expertise to tune the system to get good performance. Unfortunately, Hadoop's performance out of the box leaves much to be desired, leading to suboptimal use of resources, time, and money (in payas-you-go clouds). We introduce Starfish, a self-tuning system for big data analytics. Starfish builds on Hadoop while adapting to user needs and system workloads to provide good performance automatically, without any need for users to understand and manipulate the many tuning knobs in Hadoop. While Starfish's system architecture is guided by work on self-tuning database systems, we discuss how new analysis practices over big data pose new challenges; leading us to different design choices in Starfish.

Big Data concerns massive, heterogeneous, autonomous sources with distributed and decentralized control. These characteristics make it an extreme challenge for organizations using traditional data management mechanism to store and process these huge datasets. It is required to define a new paradigm and re-evaluate current system to manage and process Big Data. In this paper, the important characteristics, issues and challenges related to Big Data management has been explored. Various open source Big Data analytics frameworks that deal with Big Data analytics workloads have been discussed. Comparative study between the given frameworks and suitability of the same has been proposed.

Digital universe is flooded with huge amount of data generated by number of users worldwide. These data are of diverse in nature, come from various sources and in many forms. To keep with the desire to store and analyze ever larger volumes of complex data, relational databases vendors have delivered specialized analytical platforms that come in many shapes and sizes from software only to analytical services that run in third party hosted environments. In addition new technologies have emerged to address exploding volumes of complex data, including web traffic, social media content and machine generated data including sensor data, global positioning system data. New non-relational database vendors combine text indexing and natural language processing techniques with traditional database technologies to optimize ad-hoc queries against semi-structured data. A number of analytical platform are available in the market for analysis of complex structured and unstructured data, each of which is designed to handle specific type of data/workload. In this paper we will discuss three open source Big Data Analytics frameworks suitable for different types of workload.

Big Data can be characterized by different aspects. The commonly used aspects are Volume, Velocity and Variety. Veracity and Value are also used to characterize Big Data. They are helpful lens through which we can understand the nature of Big Data and the platform available to exploit them.

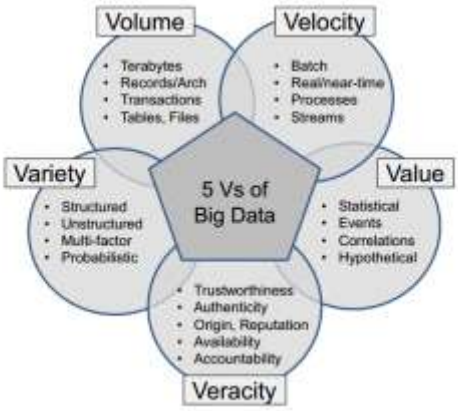
Volume - As infrastructure becomes increasingly available and affordable, data generated by different sources is very huge in size; petabytes or zettabytes. This huge amount of data is called Big Data.

Velocity - The sheer velocity at which we are creating data is huge cause of Big Data. Digital universe expands from 130 million to 40 trillion in 8 years (2005-2013). The data generated from various sources range from Batch to Real time. So this high velocity data defines new term called "Big Data".

Variety - The representation of data generated by various sources are diverse in nature; for example ecommerce web sites deal with structured data, web server logs deal with semi structured data and social websites deal with unstructured data like audio, video, images etc… Hence big data can be categorized into structured, unstructured and semi structured types and digital universe deals with combination of all.

Veracity - Duo to sheer velocity of some data we cannot spend time in cleans the data before using it. Compiling multisource data and use it for decision making for business requiresmechanisms that deal with imprecise data. Hence combination of precise, imprecise, accurate, data can be called big data.

Value - By processing huge volume, high velocity, variety and veracity of data, presents a new dimension for analyzing big data called "value". Collaborating different types of data, putting them all together in order to extract hidden knowledge for business and getting competitive advantage from it represents value of big data.



II. RELATED WORK

Salim Raza Qureshi et.al [1] has concluded that Big Data will be information whose scale, conveyance, assorted qualities, and/or convenience obliges the utilization of new specialized architectures and examination to empower bits of knowledge that open new wellsprings of business worth. The information, because of its size or level of structure, can't be effectively broke down utilizing just customary databases or techniques. It requires new information architectures, systematic sandboxes, new apparatuses, new expository techniques, coordinating various aptitudes into new part of information researcher. Associations are getting business advantage from breaking down ever bigger and more intricate information sets that undeniably oblige constant or close ongoing capacities.

Lakshmish Ramaswamy et.al [2] has envisioned a cloud-based eco-framework in which great information from extensive quantities of autonomously managed sensors is shared or even exchanged in recent times. Such an eco-framework will fundamentally have numerous partners, for example, sensor information suppliers, space applications that use sensor information, and cloud foundation suppliers who may collaborate and compete. This work proposed a cloud based huge information structural planning for supporting sensor

administrations. A key part of our structural planning is that DQ is a first class outline antiquity that is pervasive all through the system. This work exhibited a remarkable DQ-empowered XML-based markup dialect for expounding sensor sustains as well as for do-principle applications to determine their sensor bolster prerequisites & itemized investigation of the advantages and impediments of surely understood huge information strategies.

Parth Chandarana et.al [3] Computerized universe is overflowed with enormous measure of information produced by number of clients around the world. These information are of differing in nature, originate from different sources and in numerous structures. To keep with the craving to store and investigate ever bigger volumes of complex information, social databases merchants have conveyed particular logical stages that come in numerous shapes and sizes from programming just to expository administrations that keep running in outsider facilitated situations. Apache Hadoop is suited for workload where time is not separating variable however Project storm is proper for data stream examination in which examination performed is constant and Apache drill is best for natural and uncommonly selected examination.

Zhi-Hua Zhou et.al [4] has identified machine learning among the core techniques for data analytics. He has clarified three common but unfortunately misleading arguments about learning systems in the big data era. It is difficult to identify totally new issues brought about by big data. Nonetheless, there are always important aspects to which one hopes to see greater attention and efforts channeled. First, although we have always been trying to handle (increasingly) big data, we have usually assumed that the core computation can be held in memory seamlessly. Whereas the current data size reaches to such a scale that the data becomes hard to store and even hard for multiple scans. However, many important learning objectives or performance measures are non-linear, non-smooth, on-convex and non-decomposable over samples. For example, AUC (Area Under the ROC Curve), and their optimizations, inherently require repeated scans of the entire dataset. Is it learnable by scanning the data only once, and if it needs to store something, the storage requirement is small and independent to data size? We call this "one-pass learning" and it is important because in many big data applications, the data is not only big but also accumulated over time, hence it is impossible to know the eventual size of the dataset. Fortunately, there are some recent efforts towards this direction, including.

Xiaoquan Li et.al [5] has summarized that in this present data age, the fast advancement and promotion of the Internet, so that the remarkable increment in data limits, data accumulation, safeguarding, support, and sharing undertakings confronting new difficulties. This work contributes; "Big data" have started to exist. Enormous Data in volume, assortment, and speed, genuine and precise qualities such as lead era innovation wave. It truly means is that the user can dissect and use information, through the trading of information joining and investigation, the disclosure of new learning, make new esteem, bringing the "huge learning", "enormous science and innovation to the" huge benefits "and" huge advancement ".Not just the information from the substantial organizations can "dig commercial gold mine" for the knowledge offices who need to screen the "bad guys," big data is "invaluable."

Divyakant Agarwal et.al [6] has said that despite the fact that scalable data management has been a dream for over three decades and much research has concentrated on large scale data management in customary venture setting, distributed computing brings its own particular arrangement of novel difficulties that must be tended to guarantee the achievement of information administration arrangements in the cloud environment. This work study envelops both classes of frameworks: (i) for supporting redesign overwhelming applications, and (ii) for ad-hoc analytics and choice backing & it concentrate on giving a top to bottom investigation of frameworks for supporting upgrade serious web-applications and give a review of the cutting edge in this area**.**

AvitaKatal et.al [7**]** has mentioned that as data is developing at an enormous pace making it hard to handle such substantial measure of information .The fundamental trouble in taking care of such vast measure of data is on account of that the volume is expanding quickly in correlation to the Computing assets. To accept and adapt to this new technology many challenges and issues exist which need to be brought up right in the beginning before it is too late. All those issues and challenges have been described. These challenges and issues will help the business organizations which are moving towards this technology for increasing the value of the business to consider them right in the beginning and to find the ways to counter them. This work portrayed the new idea of big data, its significance and the current undertakings. Hadoop device for Big information is portrayed in point of interest concentrating on the territories where it should be enhanced so that in future Big information can have innovation and in addition abilities to work with.

Xiaomeng Yi et.al [8] in his work has examined the unique challenges when big data meet networks, and when networks meet big data. & check the state of the art for a series of critical questions: What do big data ask for from networks? Is today's network infrastructure ready to embrace the big data era? If not, where are the bottlenecks? And how could the bottlenecks is lifted to better serve big data applications? We take a close look at building an express network infrastructure for big data. Our study covers each and every segment in this network highway: the access networks that connect data sources, the Internet backbone that bridges them to remote data centers, and the dedicated network among data centers and within a data center. We also present

two case studies of real-world big data applications that are empowered by networking, highlighting interesting and promising future research directions.

Deepa Gupta et.al [9] has proposed about few visualization techniques like Visualization based data discovery toolsand nano cube technology. We also discussed Hadoop. For future we can work upon the optimization ofpseudocode to build nanocubes.Nano cubes still take more memory than we would like. The envisiondynamic control over the cardinality of dimensions, buleave that for future work & would also like to explore hybrid solutions that utilize both on-disk and in memory data structure to enable more complexnano cubes. The proposed work demonstrates the effectiveness ofour technique on a variety of real-world datasets, and present memory, timing, and network bandwidthmeasurements. Nano cubes offer efficient storage andquerying of large, multidimensional, spatiotemporal datasets, but are not without limits. So in future here is a need to get solution to the limitations like disk storage and allowing multiple spatial dimensions.

Emmanuel Christophe et.al [10] has proposed to address one of the main shortcomings, which are the relatively slow computation in double precision using the new Fermi architecture. This work definitely witness an increasing number of GPU implementations for remote sensing processing algorithms in the near future.Still, benefiting from this massive speed-up requires one tocarefully select those algorithms which fit well in the GPU com-putting architecture, identify the critical sections to optimize, andlook closely at how things are implemented.All this complexity should remain hidden to the end-user,which is exactly what the high level of abstraction providedby the Orfeo Toolbox framework allows. Further improvementscan be made in that direction by proposing a mechanism toswitch seamlessly from CPU to GPU versions of algorithmsdepending on available hardware.

III. SYSTEM DESCRIPTION

1. Predictive Model Representation and Comparison: Towards Data and Predictive Models Governance

The procedures of creating prescient models include information planning, checking of data quality, lessening, displaying, expectation, and investigation of results. Creating superb prescient models is a time consuming activitybecause of the tuning process in finding optimum modelparameters. Extraction of information mining models is an essential issue. Delivering the most helpful information mining models is inadequate without translating and handling learning from the models. This proposed work has an adaptable information and model representation in a more broad system towards information and model administration. Separating the model parameters from XML representation outlines that the models are profitable as far as understanding, and sensible for further utilization. PTML representation gives a helpful answer for separating information structure data mining.

2. Predicting and Mitigating Jobs Failures in Big Data Clusters

Motivated by the significant amount of resource waste, in terms of computational time, CPU, RAM and DISK, caused by job failures at big-data clusters, the aim to capture failed jobs upon their arrival and minimize the resulting resource waste. To incorporate transient and complex system dynamics, we consider extensive static and system features that capture the disparities of jobs' multiple tasks and system load across priorities. The first explore four supervised classification techniques, namely LDA, ELDA, QDA, and LR, in a sliding window fashion, when developing an on-line prediction model for job failures. Based on the prediction results, the developed a delay-based mitigation policy that proactively terminates predicted-to-fail jobs after a certain grace period. The optimal choice of the classification technique, size of the sliding window, and length of grace period are determined during the training phase, so as to achieve low misclassification rates and mitigated false negative rates.

3. Big Data Processing for Prediction of Traffic Time based on Vertical Data Arrangement

This work discovered new problems of predicting various traffic conditions according to time and location with historical traffic data for long-term prediction, and the problems indicate historical data aggregation and a variety of spatiotemporal traffic conditions. To solve the data aggregation issue, the proposed novel method called vertical data arrangement, which aggregates matching items of historical data into the same time slot. For a variety of spatio temporal traffic conditions & suggests constructing a spatiotemporal prediction map for each road and each day. By using the prediction map, the work can select suitable time-series forecasting methods for specific traffic conditions according to the location and time by analyzing the characteristics of historical data for each road and each day of the week. Moreover, both methods involve big data processing & constructed a big data processing Framework to handle the complete series of processes.

IV. PREDICTING I-PHONE SALES FROM I-PHONE TWEETS

Illustration from the hypothetical structure of AIDA and Hierarchy of Effects models in advertising combined with an assumption that social media actions such as tweeting, liking, commenting and rating are proxies for user/consumer's attention to a particular object/product, the proposed work how social media data from twitter can be used to predict the sales of i-Phones. The developed and evaluated work possess a linear regression model that transforms i-Phone tweets into a prediction of the quarterly i-Phone sales with an average error close to the established prediction models from investment banks. This strong correlation between i-Phone tweets and i-Phone sales becomes marginally stronger after incorporating sentiments of tweets.

4. Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients

This work proposed the big data solution for predicting the 30-day risk of read mission for the CHF patients. The proposed solution leverages big data infrastructure for both information extraction and predictive modeling. We study the effectiveness of our proposed solution with a comprehensive set of experiment, considering quality and scalability. As ongoing work, we aim at leveraging big data infrastructure for our designed risk calculation tool, for designing more sophisticated predictive modeling and feature extraction techniques, and extending our proposed solution to predict other clinical risks.

5. An Initial Study of Predictive Machine Learning Analytics on Large Volumes of Historical Data for Power System Applications

We are in a new era of industrial growth that combines computers, sensors, data repositories, high bandwidth networks, mobile devices, autonomous machines, and data analytics that drive industrial innovation and growth. More and more industrial data are being collected and stored by these industrial systems. For this reason, Industrial Analytics requires more powerful and intelligent machine learning tools, strategies, and environments to appropriately extract knowledge from the large volumes of industrial data to unleash its great potential value. We started our research on predictive machine learning analytics for Big Data by conducting a comprehensive literature survey of machine learning libraries and tools for Big Data analytics, and initial studies on how to forecast substation faults and power loading. Our results indicated that it is feasible to forecast substations fault events and power load using Naïve Bayes algorithm in MapReduce paradigm or machine learning tools specific for Big Data. We will collect more industrial data on these two cases and more industrial analytics domains in ABB. More statistical and machine learning algorithms will be developed, utilized, and verified to mine more values from our industrial data.

6. Mining the Situation: Spatiotemporal Traffic Prediction with Big Data

In this paper, we proposed a framework for online traffic prediction, which discovers online the contextual specialization of predictors to create strong hybrid predictor fromseveral weak predictors. The proposed framework matches the real-time traffic situation to the most effective predictor constructed using historical data, thereby self-adapting to the dynamically changing traffic situations. We systematically proved both short-term and long-term performance guarantees for our algorithm, which provide not only the assurance that our algorithm will converge over time to the optimal hybrid predictor for each possible traffic situation but also provide a bound for the speed of convergence to the optimal predictor. Our experiments on real-world dataset verified the efficacy of the proposed scheme and showed that it significantly outperforms existing online learning approaches for traffic prediction. For future work, we plan to extend the current framework to distributed scenarios where traffic data is gathered by distributed entities and thus, coordination among distributed entities are required to achieve a global traffic prediction goal.

7. Predicting Days in Hospital Using Health Insurance Claims

Healthcare administrators worldwide are striving to lower the cost of care whilst improving the quality of care given. Hospitalization is the largest component of health expenditure. Therefore, earlier identification of those at higher risk of being hospitalized would help healthcare administrators and health insurers to develop better plans and strategies. In this paper, a method was developed, using large-scale health insurance claims data, to predict the number of hospitalization days in a population. We utilized a regression decision tree algorithm, alongwith insurance claim data from 242,075 individuals over three years, to provide predictions of number of days in hospital in the third year based on hospital admissions and procedure claimsdata from the first two years. The proposed method performs well in the general population as well as in sub-populations. Results indicate that the proposed model significantly improvespredictions over two established baseline methods (predicting a constant number of days for each customer and using the number of days in hospital of the previous year as the forecast for the following year). A reasonable predictive accuracy (AUC= 0:843) was achieved for the whole population. Analysis of two sub-populations - namely elderly persons aged over 63 years

or older in 2011 and patients hospitalized for at least one day in the previous year - revealed that the medical information made more contribution to predictions of these two sub-populations, incomparison to the population as a whole.

8. Towards Efficient Big Data and Data Analytics: A Review

Big Data" will be information whose scale, conveyance, assorted qualities, and/or convenience obliges the utilization of new specialized architectures and examination to empower bits of knowledge that open new wellsprings of business worth. The information, because of its size or level of structure, can't be effectively broke down utilizing just customary databases or techniques. It requires new information architectures, systematic sandboxes, new apparatuses, new expository techniques, coordinating various aptitudes into new part of information researcher. Associations are getting business advantage from breaking down ever bigger and more intricate information sets that undeniably oblige constant or close ongoing capacities.

9. Towards A Quality-Centric Big Data Architecture for Federated Sensor Services:

Envision a cloud-based eco-framework in which great information from extensive quantities of autonomously managed sensors is shared or even exchanged in recent times. Such an eco-framework will fundamentally have numerous partners, for example, sensor information suppliers, space applications that use sensor information, and cloud foundation suppliers who may collaborate and compete. This work proposed a cloud based huge information structural planning for supporting sensor administrations. A key part of our structural planning is that DQ is a first class outline antiquity that is pervasive all through the system. This work exhibited a remarkable DQ-empowered XML-based markup dialect for expounding sensor sustains as well as for do-principle applications to determine their sensor bolster prerequisites & itemized investigation of the advantages and impediments of surely understood huge information strategies.

IV. RESULTS AND DISCUSSIONS

Big data usually includes data sets with sizes beyond the ability of commonly used software tools that is used to capture, manage, and process the data within a tolerable elapsed time. Big data "size" is a target that moves constantly, as the previous year’s ranging from dozen of terabytes to more petabytes of data. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.

The challenges include analyzing the data, capturing, curating, searching, sharing, storing, transferring, visualization, and other privacy violations. A drastic move to larger data sets is due to additional information that are derived from the analysis of a large set of data that are related to each other, and as compared to smaller sets along with the similar total amount of data, by allowing correlations that are already found to spot latest trends in business, diseases prevention, crime detection and so on. Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on several thousands of servers.

Data age, the fast advancement and promotion of the Internet, so that the remarkable increment in data limits, data accumulation, safeguarding, support, and sharing undertakings confronting new difficulties. This work contributes; "Big data" have started to exist. Enormous Data in volume, assortment, and speed, genuine and precise qualities such as lead era innovation wave. It truly means is that the user can dissect and use information, through the trading of information joining and investigation, the disclosure of new learning, make new esteem, bringing the "huge learning", "enormous science and innovation to the" huge benefits "and" huge advancement ".Not just the information from the substantial organizations can "dig commercial gold mine" for the knowledge offices who need to screen the "bad guys," big data is "invaluable."

V. CONCLUSION

Data is developing at an enormous pace making it hard to handle such substantial measure of information .The fundamental trouble in taking care of such vast measure of data is on account of that the volume is expanding quickly in correlation to the Computing assets. To accept and adapt to this new technology many challenges and issues exist which need to be brought up right in the beginning before it is too late. All those issues and challenges have been described. These challenges and issues will help the business organizations which are moving towards this technology for increasing the value of the business to consider them right in the beginning and to find the ways to counter them. This work portrayed the new idea of big data, its significance and the current undertakings. Hadoop device for Big information is portrayed in point of interest concentrating on the territories where it should be enhanced so that in future Big information can have innovation and in addition abilities to work with.

REFERENCES

[1]. Salim Raza Qureshi and Ankur Gupta,"Towards efficient Big data and data analytics: A Review", IEEE 2004.

[2]. LakshmishRamaswamy, Victor Lawson and Siva VenkatGogieni,"Towards a Quality Centric Big data architecture for federated sensor services", 2013 IEEE International Congress on Big data.

[3]. ParthChandarana and M.Vijayalakshmi,"Big data Analytics Frameworks", 2014 International conference CSCITA.

[4]. Zhi-Hua Zhou et.al,"Big data opportunities and challenges: discussions from big data perspective",IEEE Nov 2014.

[5]. Xiaoquan Li et.al,"Research on big data architecture, Key technologies and its measures",IEEE 2013.

[6]. Divyakant Agarwal et.al,"Big data and cloud computing: Current state and future opportunities",EDBT 2011,March.

[7]. AvitaKatalet.al,"Big data: Issues, Challenges, Tools and good practices", IEEE 2014.

[8]. Xiaomeng Yi et.al,"Building a network highway for big data: Architecture and Challenges", IEEE July 2014.

[9]. Deepa Gupta and SameeraSiddiqui,"Big data implementation and visualization", IEEE International Conference ICAETR-2014.

[10]. Emmanuel Christophe et.al,"Remote sensing Processing: from Multicore to GPU", IEEE Journal of earth science, Sep 2011.