ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Big Data: Performance Analysis of Vendor and Value Creation through Big Data Analytics

R.Kanagalakshmi

* Department of Computer Science, Dhanalakshmi Srinivasan, Institute Of Research and Technology, Perambalur, India

Abstract

There has been a dramatic increase in the amount of data generated which is currently measured in zettabytes and still increasing. Managing the fast paced increase in data volume is going to be challenging as ever before as the datasets is getting even more diverse. Business analytics is the focal point of computer science, statistical science and management science. Business analytics entails providing strategy and process optimization which in turn steps up the competitive advantage. The world of business is undergoing a revolution through data and analytics acts as an aid to decision making process. Many of the leading organizations have started recognizing data as a strategic asset. Hence we focus on how the vendor processes Big Data and the architecture of the vendor service. It also emphasizes on the agent-based computational model, effective analytics and statistical techniques that can be applied on Big Data. Data is usually analyzed to streamline the enterprise with means of analytical insights with which the efficiency, profit and competitive position can be stepped up. As the concept of Big Data is loosely defined and kind of fuzzy, uncertainty pertaining to cost and advantages of undertaking a novel Big Data project prevails. Also it is not evident earlier that Big Data is transformed into valuable information and insights that justifies the investment made.

Keywords: Agent-based computational model, Big Data analytics method, EARL, GLADE..

Introduction

Data is exploding at an astounding rate. The enormous amount of data increases in an incredible manner and has already surpassed the capability of available tools and applications to capture, process, curate, store, analyze and understand the collection of data. In 1998 the web pages indexed by Google was around one million which was 1 billion in 2000 and exceeded 1 trillion in 2008. One such reason for this dramatic increase is due to social networking applications that let users to add to the already enormous volume of web data. Moreover the real time data that could be fetched from mobile also alarmingly raise the data volume. The data generated in two days is estimated[1] to be 5 Exabyte. This enormous amount of data poses new challenges. Applications such as sensor networks, network traffic analysis, web click streams, email, blogging, RFID readers, traffic cameras etc. generate massive datasets which may be either structured or unstructured. Unstructured data is heterogeneous in nature which may include text, image, audio, video, etc. and is rapidly increasing than the structured data. As per a 2011 IDC study 90 percent of all data created in the next decade would be unstructured.

In the high speed business world acquiring quick results is a key factor for "Big Data"[2]. While analyzing large amount of data well formed theory and substantial experiments from statistics carrying out thorough sampling on data and then computing early results from those samples provide a fast and successful way to acquire approximate results within the stipulated level of accuracy for considerable number of applications. The data and analytics is driving the realm of business to make better decisions and hold the competitive advantage.

BIG data

Big Data[1] refers to datasets that cannot be managed with the available database and software tools due to its enormous volume. It can also be given as the information asset that is of high volume, velocity and variety which claims for innovative and cost effective methods of processing in order to create value through insight obtained. This ever increasing data can be handled only through new and innovative algorithms and tools. The 3 V's in Big Data management include:

Volume: There is more data than ever before; its size continues increasing, but not the percent of data that our tools can process.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Variety: There are many different types of data, as text, sensor data, audio, video, graph, and more.
Velocity: Data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time.

Fig.1 Big-Data Represent diagram

Nowadays, there are two more V's:
Variability: There are changes in the structure [2]of the data and how users want to interpret that data
Value: Business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach.

BIG data analytics

Big Data Analytics can be referred to as technological initiative for imparting richer and accurate insights to the business entities with the ultimate aim to remain competitive with other companies. Organizations can make timely decisions fast and efficiently track the trends that emerge. Big Data[3] analytics is a crucial tool that is emerging in order to enhance the quality and efficiency in organization. The significance of Big Data analytics is set to increase in the years to come.

The key policies to deal with big data include that of sampling and using distributed systems. If the available dataset is enormous obtaining approximate solution using subset of samples is referred to as sampling. A sampling method will be termed as a good one if best instances are chosen that exhibits better performance utilizing less memory and time. Using of probabilistic technique is an alternative to sampling.

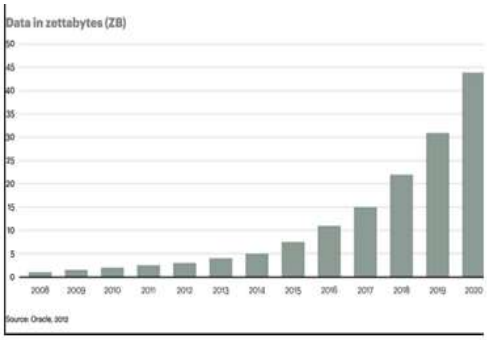


Fig2. Data Growing

In the high speed business world acquiring quick results is a key factor for "Big Data". While analyzing large amount of data well formed theory and substantial experiments from statistics carrying out thorough sampling on data and then computing early results from those samples provide a fast and successful way to acquire approximate results within the stipulated level of accuracy for considerable number of applications

Approximate results that are based on samples quite a number of times offer the only possible way by which advanced analytical applications on Big Data can satisfy their time and resource constraints. Methods and tools for computation of accurate early results are not currently supported in big data systems.

The objective of Early Accurate Result Library Framework (EARL) [9] is to bridge the gap between rapidly increasing data sizes and response time requirements. The powerful and suitable methods and models of statistics are explored and applied to estimate results and accuracy obtained from sampled data.

Early Accurate Result Library Framework works by predicting learning curve and choosing the appropriate sample size for achieving the desired error bound specified by the user. The error estimates are based on a technique called bootstrapping which is used and validated widely.

The demonstration of the EARL [9]elucidates its functionality which includes an intuitive GUI interface for better user understandability on the accuracy obtained by increasing sample sizes and the learning curve. To achieve scalability, EARL uses Hadoop for error estimation. EARL is a simple framework for estimating results and error s for mining algorithm. Furthermore EARL[9] provides

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

the learning curve which can often predict the required sample size for a given error bound without having to carry out the actual computation.

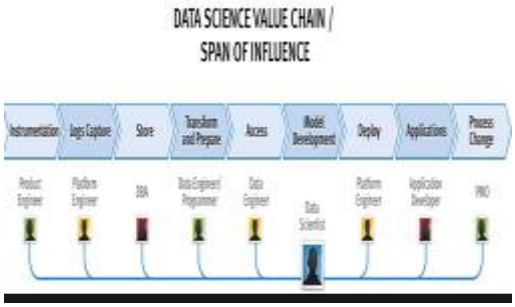


Fig3 Value Creation for Big-Data

It was inferred from the demonstration of EARL that it is not required to use the whole dataset, and in most cases it is sufficient to use 1% of data to achieve similar results compared to the execution over the entire dataset. Various optimizations done to the re-sampling methods have made the framework more attractive.

Over the past few years the enthusiasm of the web companies to use large scale data analytics has been increasing. This is in fact facilitated by the cheap storage facility available using which the user behavior and log data could be stored. So as to derive the value out of the data, statistical and machine learning methods and many more have to be tried.

The conventional methods that are available for managing data such as SQL and relational database systems however do not support the modern analytical methods. In order to get on with this issue multiple trends are evolving.

To tackle this issue database community began to incorporate the analytic capabilities to the existing relational databases by taking advantages of the extensibility features of SQL like User-Defined Functions (UDF) and User-Defined Aggregates (UDA)

A scalable distributed system for data analytics performed at large scale called GLADE [2] is introduced. The system makes use of the analytic functions that are expressed via User-Defined Aggregate (UDA) interface and executes them efficiently on the input data. A single class in which the entire computation is being encapsulated which needs the definition of four methods Init, Accumulate, Merge and Terminate. The user code is taken at runtime and executes right near the data by using the parallelism available within a single machine or across a cluster of computing nodes.

The architecture of the GLADE is described and with the help of analytical functions the processing performed is presented. The system GLADE is compared and contrasted with a relational database (PostgreSQL) enhanced with UDAs and Map-Reduce (Hadoop). The expressive, scalability, and running time efficiency are compared when the analytical functions are coded into each system.

The result of experiment illustrates evident signs of better running time performance of GLADE compared to Hadoop at a constant factor of 30. In a single node scenario an instance of User Visits (20GB) where within the local laptop, the system and queries are placed. GLADE takes the leading edge due to the columnar storage and extensive use of thread level parallelism, hence showcases considerably shorter execution time. In distributed clustered environment which represents big data analytics, same queries are run over an 8TB User

Visits at UC Merced. Thus when the number of nodes is scaled up GLADE showcases a better runtime performance comparably to Hadoop. This is due to the following reasons – GLADE makes use of columnar storage and reads the data required by executed query, only point-to-point communication between the nodes, stores only GLA state and makes use of vectorized processing model.

Big Data is dramatically increasing and we require an efficient large scale analytic system to tackle the data explosion. In order to create and extract value from the available data, large scale analytic system such as GLADE is indispensible. The GLADE[2] showcases better running time performance but more systems that exhibit further better capabilities in terms of value creation, system capability, algorithmic designs have to be improved.

Value creation

Global Pulse is a United Nations initiative, launched in 2009, that functions as an innovative lab and that is based in mining Big Data for developing countries. They pursue a strategy that consists of 1) researching innovative methods and techniques for analyzing real-time digital data to detect early emerging vulnerabilities; 2) assembling free and open source technology toolkit for analyzing real-time data and sharing hypotheses; and 3) establishing an integrated, global network

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

of Pulse Labs, to pilot the approach at country level. Global Pulse describe the main opportunities Big Data offers to developing countries in their White paper "Big Data for Development: Challenges & Opportunities"[6]:

- Early warning: develop fast response in time of crisis, detecting anomalies in the usage of digital media
- Real-time awareness: design programs and policies with a more fine-grained representation of reality
- Real-time feedback: check what policies and programs fails, monitoring it in real time, and using this feedback make the needed changes

The Big Data mining revolution is not [5]restricted to the industrialized world, as mobiles are spreading in developing countries as well. It is estimated that there are over five billion mobile phones, and that 80% are located in developing countries.

Vendor in bigdata management

A consumer bank holds the primary source of the consumer identity for all financial, and non-financial transactions. Banks were in firm control of customer relationship, and the relationship was for subsequent practical purposes as long as the bank wanted it to be and it is not the current scenario. Consumers now have transient relationships with multiple banks: a current account at one that charges no fees, savings accounts with a bank that offers high interest, a mortgage with a one offering the best rate, and a brokerage account at a discount brokerage.

New entrants who offer peer-to-peer services; and the Paypal, Amazon, Google and Walmart of the world – have had the effect of disinheriting the banks. Banks no longer have a complete view of their customer preferences, buying patterns and behaviors. This problem is exacerbated by the fact that social networks now capture very valuable psychographic information – the consumer's interests, activities and opinions.

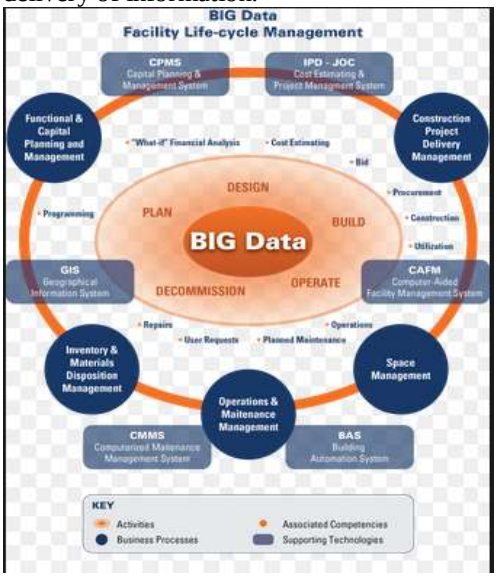
Bringing together transactional data in CRM[4] systems and payments systems, and unstructured data both from within and outside the firm requires new technologies for data integration and business intelligence to argue the traditional data warehousing and analytics approach. Big Data technologies therefore play a pivotal role in enabling customer centricity in this new reality. Efficient allocation of capital is now seen as a major competitive advantage,

Advancements in technology and digital devices' affordability have led to Big Data which encompasses dramatic increase in quantity and diversity of high frequency digital data. This prospective data assists the decision and policy makers.

Turning Big Data—call logs, mobile-banking transactions, online user-generated content
such as blog posts and Tweets, online searches, satellite images, etc.—into actionable information requires using computational techniques to unveil trends and patterns within and between these extremely large socioeconomic datasets. New insights gleaned from such data mining should complement official statistics, survey data, and information generated by Early Warning Systems, adding depth and nuances on human behaviors and experiences—and doing so in real time, thereby narrowing both information and time gaps.

and risk-adjusted performance calculations require new points of integration between risk and finance subject areas. Traditional BI[8] systems work extremely well when the questions to be asked are known. But business analysts frequently do not know all the questions they need to ask. Self-service ability to explore data, add new data, and construct analysis as required is an essential need for banks driven by analytics.

More real-time analytical decisions: whether it is a front office trader or a back office customer service representative, business users demand real-time delivery of information.



ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Fig 4 Big-Data Facility Life-Cycle Management

Event processors, real-time decision making engines and in-memory analytical engines are crucial to meeting these demands. Rapid growth in structured and unstructured data from both internal and external sources requires better utilization of existing technologies and new technologies to acquire, organize, integrate and analyze data.

We define the structure in 'structured data' in alignment with what is expected in relational technologies – that the data may be organized into records identified by a unique key, with each record having the same number of attributes, in the same order. Because each record has the same number of attributes, the structure or schema need be defined once as metadata for the table, and the data itself need not have metadata embedded in it.

Semi-structured data also has structure, but the structure can vary from record to record. Records in semi-structured data are sometimes referred to as jagged records because each record may have variable number of attributes. By unstructured data, we mean data for which structure does not conform to the two other classifications discussed above. Strictly speaking, unstructured text data usually does have some structure.

Each technology enjoys a set of distinct advantages depending on the phase in the lifecycle of data management and on the degree of structure within data. Data needs to be organized for analysis, but the organized data may reside on any suitable technology for analysis.

That structured data is always best handled in a relational database. That is not the case always, and the section on handling structured data explains what other technologies may come into play when we consider additional dimensions for analysis.

Fig 5 Big-Dta for Vendor Creation

Semi-structured data within the bank may exist as loan contracts, in derivatives trading systems unstructured data, semi-structured data contains tags to mark significant entity values contained within it. These tags and corresponding values are key-value pairs. If the data is in such a format that these key-value pairs need to be extracted from within, it may need to be stored on a distributed file system for later parsing and extraction into key-value databases.

Banks have applications that generate many terabytes of structured data and have so far relied almost exclusively on relational technologies for managing this data. Systems have to choose two from among the three properties of Consistency, Availability and Partition Tolerance. And most implementations choose to sacrifice Consistency, the C in ACID, thereby redefining themselves as BASE systems (Basically Available Soft-state Eventually consistent).

The rigidity enforced by the relational schema model requires more upfront planning and may make the schema more brittle in the face of changing requirements.

Computational model for big data

Big Data is new concept in Information Technology[7]. It's mainly focusing nature of data: randomly enhancing volume of data, increase of processing velocity and also focusing variety of data types. Computational model is a mathematical model in computational science. It will be the mathematical analytical solution to the problem. It is an agent based model. Business value that gives organization a compelling advantage, due to the ability of taking decisions based in answering questions that were previously considered beyond reach.

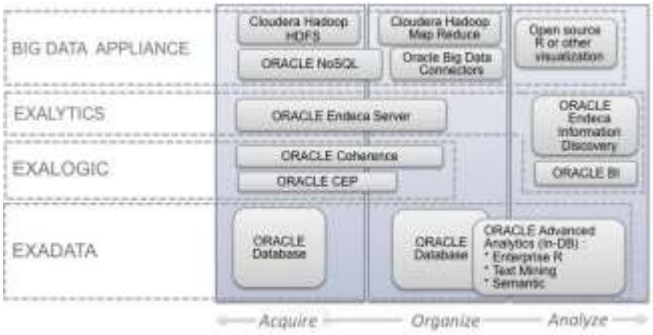


Fig.6: Oracle products for Big Data Management

Here there is no need for strict ACID compliance; availability needs are less say, a payment transaction

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

system; there are no complex queries to be executed against this data; and it would be more efficient for the application that generates the data (the Monte Carlo runs) to have local data storage. Although the data is structured, a relational database may not be the optimal technology here.

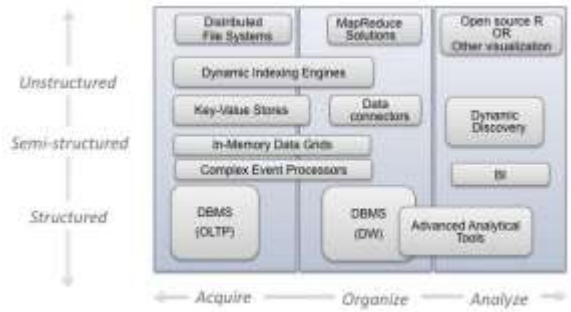


Fig.7 Big Data Technology continuum

A computation is a process evoked when a computational model (or computational agent) acts on its inputs under the control of an algorithm to produce its results.

Theoretical computer science[6] has now undergone several decades of development. The 'classical' topics of automata theory, formal languages and computational complexity have become firmly established and their importance to other theoretical work and practice is widely recognized. However, in many areas like multi-agent interactive systems, web-distributed programming, robotics, reactive systems, client-server models, distributed databases, has turned out that algorithms do not describe well their behavior.

Conclusion and futurework

Big Data is growing at dramatic rate every instant and every time a new controversy also arises with it. The future research should be in such a way that the controversies are eliminated and significant challenges in leveraging large amount of data including, system capabilities, algorithmic design and business models are dealt with in a better manner.

References

1. Wei Fan, Albert Bifet: Mining Big Data: Current Status, and Forecast to the Future. In SIKDD 2013
2. Yu Cheng, Chengjie Qin, Florin Rusu GLADE: Big Data Analytics made easy. In SIGMOD 2011
3. Y.Zhang, H.Herodotou, and J.Yang. Riot:I/O-efficient numerical computing without SQL.InCIDR2009
4. Financial Services Data Management Big Data Technology in Financial Services June 2012
5. D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.
6. E. Letouz_e. Big Data for Development: Opportunities & Challenges. May 2011.
7. J. Cohen and al. MAD Skills: New Analysis Practices for Big Data. In VLDB 2009.
8. F. Rusu and A. Dobra. GLADE: A Scalable Framework for Efficient Analytics. In LADIS 2011.
9. Nikolay Laptev, Kai Zeng, Carlo Zaniolo: Very Fast Estimation for Result and Accuracy of Big Data Analytics: the EARL System. In ICDE 2013.
10. T.K.Das, P.Mohan Kumar: Big Data Analytics: A Framework for Unstructured Data Analysis.