A Way Ahead Towards Efficient Big Data Analytics: Prime Utilization in Businesses Moving Towards Cloud

Rupali Sathe
University of Mumbai
Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Rasayani

Sandip Raskar
University of Mumbai
Department of Computer Engineering, Pillai HOC College of Engineering and Technology, Rasayani

ABSTRACT

Nowadays Businesses have greatly benefited from data analytics. Companies analyze data from various activities like fraud, sales, risk management, marketing, inventory optimization, and customer support to improve their strategic and tactical business decisions. However, analyticsis powerful enough to work with big data which is too complex, expensive, difficult for computation and resource-intensive for smaller companies to use. However, all these businesses have not been able to benefit from high powered analytics and therefore cannot make the most out of their information. Big data administration generally require more no of IT staff. It also uses many expensive servers with high configuration and includes software that is very difficult to set up and maintain. Organizations require innovative technology or systems that should be able to handle complex data to get the appropriate output. Smaller companies are facing trouble in finding employees capable of working with big analytics. This field deals with advanced and complex technology and new area of technology growing rapidly. All above mentioned factors made big data analytics fitted only to the large organizations. The above requirements are accomplished by proposing a system which performs adopting cloud as a platform to work with big data, which will help to make big analytic easier to handle the analytics and provides on demand cost efficient platform with great horizontal scalability. This computational methodology and algorithm for big data in the cloud environment make their platform more accessible. This new paradigm will play a leading role in the near future.

Keywords

Scalability,Cloud, Analytics.

1. INTRODUCTION

Various organizations can have the system resources as well as proficiency to manage huge bulk of data but quick streaming of data and size of data is increasing so they lack the capacity to mine it and extract the intelligence in a better way. Not only the size is increasing too rapidly but also the velocities in which it arrives and the variety of data types require new data processing techniques are changing for analytic solutions. There are many new data types according to its complexity, like structured and unstructured which are processed to find intelligence into a business or condition. Result of it big data does not always fit into neat tables of columns and rows. Hadoop is an open source technology framework which allows the geographical processing on huge data over clusters of valued servers, which is developed to rate up from a single server machine to thousands of servers, having a high grade of fault tolerance. Rather than depending on high-end hardware for big data processing, the elasticity of these commodity hardware clusters gained by the software's ability to detect and handle failures at the application layer. Important predictions which organization are finding can be made by searching through and analyzing Big Data. Around 82 percent data is "unstructured", it must be formatted or converted into structured to make it capable for data mining. The small and medium sized organizations could profit from analytics as much as large one. Such companies are facing issues with increasing data size against data storage capacity. The platform to structure Big Data is Hadoop it has the ability to solve the problem faced during analysis to make it useful for analytics purposes. Big data is increasing 62 percent per annum according to market research firm. These businesses have not been able to avail great powered analytics and because of that it cannot find the most appropriate value of information assets.

2. BACKGROUND OVERVIEW

Smaller organizations could take the benefits from analytics as much as large one. Such companies are facing so many problems with data volumes rising up to 63per cent per annum, according to market research firm[3]. However, businesses could not take the benefit from high powered analytics and so that it is difficult to find the most out of their information [3]. Gaining advantage of big data frequently includes an advancement of cultural and technical changes throughout the business. Exploring new business opportunities to expand sphere of inquiry, exploit new insights by merging traditional and big data analytics. Traditional enterprise data and tools discover insights from domain of sales forecasts to inventory levels. Usually data resides in a data warehouse when it is high volume data and then it is analyzed by using SQL oriented business intelligence (BI) tools. The data which is stored in the data warehouse comes from business processing transaction which is stored in online transaction processing (OLTP) database. These types of application like planning resources and forecasting can take advantages from big data and for thatbusiness organizations required advanced methodology of analytics to make this in a reality.

Companies to have more advanced data analysis like statistical analysis, data mining, predictive analytics, and text mining, need to move the data to dedicated servers for analysis. Bring out the data from data warehouse, generating copies of data in the external servers for analytics and perform analytics to derive insights and predictions. This is time consuming process. There is a need of duplication of data, storage environments and data analysis methodology related skills. Once predicate model successfully developed by using predictive model with production of data rewrite the model by analyzing complexity. It involves the extra movement of large size of data from a data warehouse to an data analysis server which is externally located. At this phase the data scoring occurred and is called as data "scored". After this the results are fetched back to the data warehouse. This process of transforming and rewriting data to generate actionable intelligence from information, this can take days, weeks or even months to successfully complete. Many organizations now are in a position to have proficiency in exploiting their data through data analysis; these organizations are still at the early stages of developing an analytical solution to generate a model that are capable to deliver value of real business from big data. The main obstructions are these time consuming slow and complex processes to enable direct access and timely available information to corporate data.

3. APPLICATION AREAS

We have recently crossed a threshold beyond which mobile technology and social media are generating datasets which are very difficultfor humans to comprehend. According to IDC Digital Universe studythe growth of data will never stop.

1) Social Media
2) Sensor Network
3) Mobile Technology
4) Commercial industries

Organizations are increasingly producing huge size of data. Such as monitoring user activity,sensors, instrumented business activities, click streams, finance, and accounting. With theinnovation in the web technology, social network, data of user is recorded and created by dailyposting details of activities, events, places, pictures, and things. This data tactics is oftenreferred as Big Data. The term big data that raises challenges that are based on previouslyexisting infrastructure that related to storage capability, interoperability, administration,management, and analysis.Nowadays, weare surrounded by data. People upload videos, take pictures on their cell phones, text friends,update their Facebook status, leave comments around the web, click on ads, and so forth. Machines, too,are generating and keeping more and more data. Web log files track the movement of visitors to a website, revealing who clicked where and when. Thisdata can reveal how people interact with your site. Social media helps to understand what people are thinking or how they feel about something. It can be derived from web pages, social media sites, tweets,blog entries, email exchanges, search indexes, click streams, equipment sensors, and all types of multimedia files including audio, video, and photographic. This data can be collected not only from computers, but also from billions of mobile phones, tens of billions of social media posts, and an ever-expanding array of networked sensors from cars, utility meters, shipping containers, shop floor equipment, point of sale terminals and many other sources.

12+ TBs of tweet data
? TBs of data every day
25+ TBs of log data every day
30 billion RFID tags today (1.3B in 2005)
4.6 billion camera phones world wide
100s of millions of GPS enabled devices sold annually
76 million smart meters in 2009... 200M by 2014
2+ billion people on the Web by end 2011
http://www.
Google
talk
Google Analytics
YouTube

Fig:1 Big Data Sources

4. RESEARCH PROBLEM

Big data administration generally require more no of IT staff. It also uses many expensive servers with high configuration and includes software that is very difficult to set up and maintain. Organizations require innovative technology or systems that should be able to handle complex data to get the appropriate output. Smaller companies are facing trouble in finding employees capable of working with big analytics. This field deals with advanced and complex technology and new area of technology growing rapidly. All above mentioned factors made big data analytics fitted only to the large organizations. The first challenge of the big data is to break down data silos to access all data stored in distributed manner. Data is getting difficult to process using traditional database system and software methods because of its growth rate. A second challenge is to develop a processing platform that can manipulate an unstructured complex data as fluently as structured data. The above requirements are accomplished by proposing a system which performs adopting cloud as a platform to work with big data, which will help to make big analytic easier to handle the analytics and provides on demand cost efficient platform with great horizontal scalability. This computational methodology and algorithm for big data in the cloud environment make their platform more accessible. This new paradigm will play a leading role in the near future.

5. IMPLEMENTED SYSTEM

In this section, a new model has been proposed which avail the sufficient storage capacity, required computational power and better visualization within a tolerable time and cost. This system proposes a new solution to the major issues faced by masses which are storage management and analytics of big data. This provides on-demand cost efficient computing platform. To handle big data task, networks with greater capacity, data storage, innovative analytical techniques, and considerable computing power is required. The cloud technology will avail big analytics for smaller businesses by adapting the solution to perform the analytics.

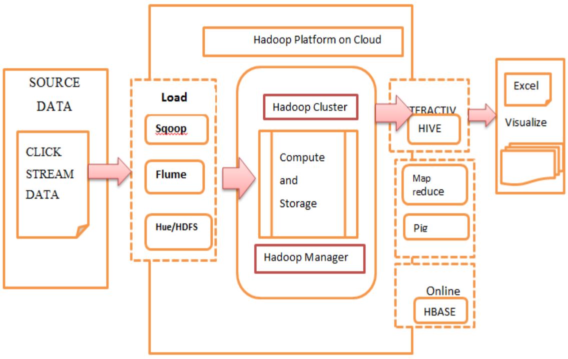

Fig:2 Architecture Diagram

Big data analytics solution named as cloud based Hadoop platform is applied on click stream data. Click stream data is a semi structured website log file which contains the data about click paths of customers. Process is classified into three main modules such as uploading data, refining data, processing data and visualizing data. Data is gathered and interpreted by different sources. First Hadoop and cloud infrastructure is developed by implementing Hadoop cluster on cloud. Client can upload the data directly on the Hadoop cluster which is hosted on cloud. The cluster is a group of servers implemented in master slave architecture, and each machine work parallel to process the data. Whole task is divided into small size task and then it is assigned to individual nodes to process simultaneously by Map Reduce algorithm. Cloud-based Hadoop analytics services can become the more popular route because of its scalability and affordability. It basically divides into three layer architecture.

Infrastructure layer

Computational layer

Application Layer

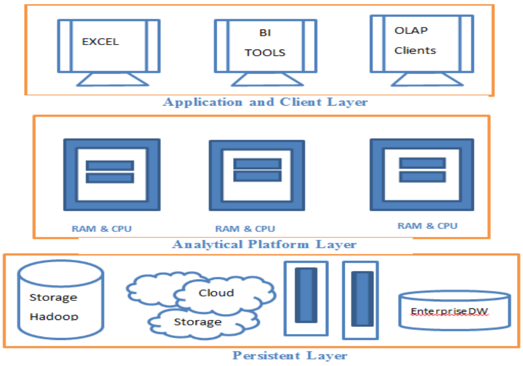

Fig:3 System Layers

6. PROJECT METHODOLOGY

The following will be development steps so as to achieve the working Prototype Model of the above

Cloud (AWS/Rackspace)

The system will be consisting of the following major sections

1. Set up Hadoop Development Environment
2. Hadoop Cluster Development
3. Installation of Pig on Hadoop Custer
4. Configure Hadoop Cluster with Pig
5. Cloud Virtual Image development
6. Uploading Data to hadoop Cluster
7. Upload hadoop Cluster on Cloud
8. Install and Configure Hue
9. Access data from on cloud to External Web site
10. Perform Click Stream Analytics on Big Data

7. RESULTS ANALYSIS

Experiments were conducted on 2 different kinds of VMs of one Hadoop clusters. Cloud environment formed with Cloudera VM. There were two types of VM machines used that are slightly different from each other. Types of machines used in the Experimental Cluster named as CDH4. All the services were running in Cloudera distribution of Hadoop. File system used in all the experiments was HDFS.In this section, in order to demonstrate that the proposed method used to measure the nodes performance in Hadoop cloud environment, experiments on a real-world Hadoop cluster are conducted. In experiment, first select the appropriate dataset to measure system performance value, and then run analytics programs in Hadoop cluster.

Table 1.Comparison of traditional of classification algorithm with Map reduce model with cloud

Analytics	**Cost**	**Storage**	**Elasticity of cluster**
Traditional Approach (Classification)	High Computational Cost	Less Storage	Less
Hadoop on Cloud (Map Reduce)	Less Computational Cost	Large Storage	High

This section reports some experiments conducted upon on click stream dataset which is a semi structured log file. Which gone through the various phases of our cloud based analytical model. Performance is calculated according to the traditional approach and our new approach on cloud. Implemented technique provides an excellent visualization approach using business

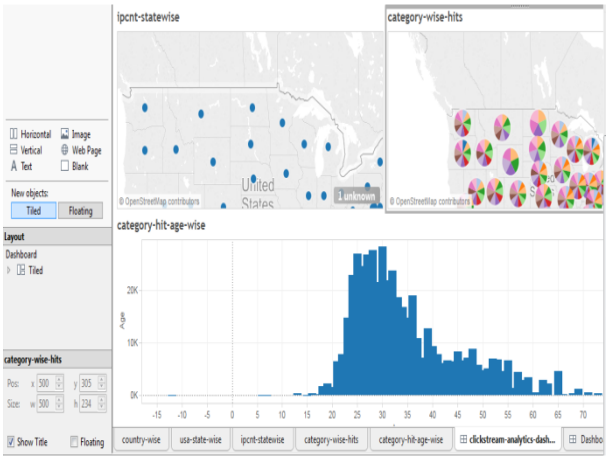

Fig:4 Click stream Analytics Results

8. CONCLUSION

The approach of cloud computing makes easier to meet and solve above problems issues by providing required resources on-demand and the cost is corresponding to the actual usage. Cloud infrastructure is a great place for running Hadoop, as it allows you to easily scale to hundreds of nodes and gives you the financial flexibility to avoid upfront investments. Additionally the system made it possible to scale up and down quickly the system infrastructures. Cloud based infrastructure offered more resilient capacity to Hadoop cluster to gain computational power by adding nodes when required. This work concluded with business models with algorithm to perform analytics on structured and complex unstructured data by giving Cloud-based data analytics solution. This system has been able to achieve less dependency on administrative requirements to manage big data. With this approach there is no need to upgrade anything, no need to fix or replace hardware, and also no need to apply any patches. It is achieved that companies can use their IT resources on strategic initiatives instead of application maintenance, and for businesses that expect to decrease the dependency on information management resources; because of this the Hadoop cluster hosted on cloud can be a precious solution

9. REFERENCES

[1] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, 2013 “Big Data: Issues and Challenges Moving Forward”, IEEE, 46th Hawaii International Conference on System Sciences.

[2] Yuri Demchenko, Zhiming Zhao, Paola Grosso, AdiantoWibisono, Cees de Laat,2012 “Addressing Big Data Challenges for Scientific Data Infrastructure”, IEEE „4th International Conference on Cloud Computing Technology and Science.

[3] R. Ranjan, May 2014 “Streaming Big Data Processing in Datacenters Clouds”, IEEE Cloud Computing, Blue Skies Column, Vol. 1, No. 1, Pp. 78–83.

[4] Sachidanand Singh, Nirmala Singh,2012 “Big Data Analytics”, IEEE, International Conference on Communication, Information & Computing Technology.

[5] S.Loughran, J.AlcarazCalero, A. Farrell, J.Kirschnick, and J.Guijarro,Nov.2012 “Dynamic cloud deployment of a map reduce architecture,”IEEEInternetComput.Vol.16.

[6] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears,2010 ``Benchmarking cloud serving systems with YCSB," in Proc. ACM Symp.Cloud Computing. Pp. 43_154.

[7] Michael, K. and Miller, K.W. (2013) Big Data: New Opportunities and New Challenges. Journal of IEEE Computer Society, 46, 22-24

[8] G. Jung, N. Gnanasambandam, T. Mukherjee, 2012, Synchronous Parallel Processing of Big-Data Analytics Services to Optimize Performance in Federated Clouds, in: Proceedings of the IEEE 5th International Conference on Cloud Computing (Cloud 2012Pp. 811–818.

[9] Zulkernine, F., Martin, P., Ying Zou, F.; Aboulnaga, A., "Towards Cloud-Based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud," Big Data (Big Data Congress), 2013 IEEE International Congress on Big Data, Vol., No., Pp.62, 69

[10] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P.Wyckoff,and R. Murthy. 2009 Hive: a warehousing solution over a map-reduce framework. Proc.