Dynamic Bigdata and Security with Kerberos

Sachin Choudhary[1], Sandesh Manohar[2], Sunil Salunkhe[3]
[1]*Department of Computer Engineering, MGMCET, Navi Mumbai*
[2]*Master of Computer Applications, IMCOST, Thane*
[3]*Master of Computer Applications, IMCOST, Thane*

Abstract— "Dynamic Big data" – is an emerging concept which denotes the large amount of data and its dynamic processing for analysis. It helps many organization to easily predict the behavior of market which is helpful for their success in business. We are entering into the age of "Dynamic Bigdata". Dynamic big data is the fastest processing of data, accumulated from various sources which gives immediate result after the analysis.Security in big data is developing at a rapid pace which includes information security, data privacy , protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools. With the right solutions, organizations can dive into all data and gain valuable insights that were previously unimaginable.

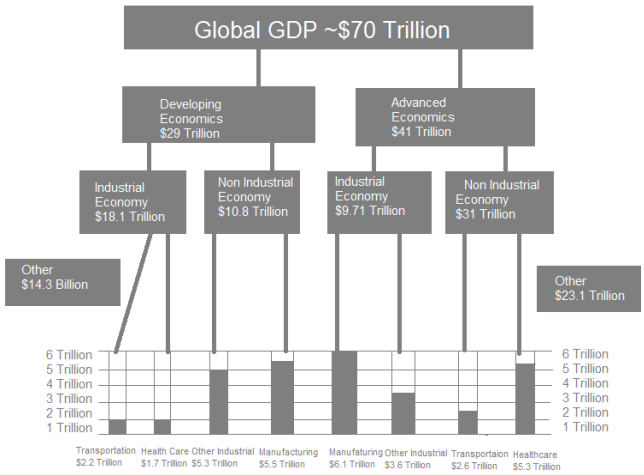
***Index Terms*—Dynamic Bigdata, Bigdata analysis, Distributed data, Bigdata Security**

I. INTRODUCTION

Big data is a very large amount of data which has the size beyond the commonly used software applications eg. Database management tools, processing applications that capture, manage and analyze .
With the growth of internet the size of bigdata is deliberately increasing. This data increases from few dozens of terabytes to many petabytes. This increasing data is to be analysed and processed. To meet the demands of handling very large a+9mount of data, many tools for the big data are being developed.

"Big data are high volume, high velocity, and/or high variety information assets which require new forms of processing technique to enable enhanced decision making, insight discovery and process optimization."
Entry into new technology for Dynamic Big Data, now the trend toward larger data sets is due to the extra information derivable from analysis of a single large set of related data, as compared to separate smaller sets.
As different types of data ie. structured and unstructured are stored in database. This data may come from many different organizations which continuously increases their data day by day. Area mostly in which bigdata can be work as dynamic big data are scientific research centre, traffic management, health science etc. Many organization collect many as data which has to process fast for the successful of an organization.

According to GE research below…the economic impact of this is pretty significant and sizable



Industrial internet opportunity ($32.3 Trillion)
46% of Global economy today

II. DYNAMIC BIGDATA

Dynamic Bigdata is the dynamically processing of information data as it is accumulated from the various information sources in the small block of data storage.
Dynamic data or transactional data denotes information that is asynchronously changed as further updates to the information become available. Dynamic data is also different from streaming data. in that there is no constant flow of information. Rather, updates may come at any time, with periods of inactivity in between. "dynamic data" would be reused or changed frequently and therefore needs to be kept in office proper ("near" storage).
Data is increasing day by day for immediate result of the changing market behavior regular analysis of data is necessary. With large data to be processed in the short span of time new technology has to be evolved which give results in a visualization manner. Hadoop uses the large data set to be stored in the huge clusters of database. With some improvement it can process the

ISSN: 2278 – 1323

continuously changing large volume of data in terms of dynamic big data.
Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. organization. Big Data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, alternatively we use "massively parallel software running on tens, hundreds, or even thousands of servers." What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

III. HADOOP TECHNOLOGY IN DYNAMIC BIGDATA

A. Hadoop:

Hadoop is a system which compromises of distributed filesystem and a framework which are used for analysis and transformation of very huge amount of data using the MapReduce. The phases of HDFS is designed after the Unix filesystem, which has the power to improve performance for the applications.

The best part of hadoop which make it as different is partitioning and computation of many hosts and execution of applications in parallel. This system measures I/O bandwidth, capacity of computation and storage by adding many servers.

These ensures that this Hadoop cluster is highly functional and available:

They assigns a physical location of the node when there is storage allocation and task scheduling, done by rack awareness.

Mapreduce is the program which computes the processes to the data on HDFS. There is a occurrence of processing tasks on the physical node of the data residence. Reduction of significantly network I/O patterns and all of the the I/O in the same rack which deliberate high aggregate read/write bandwidth is the minimal data motion

Determination of the health of system and balancing of the data on different nodes is done by its utilities.The system which bring back the previous version of HDFS by the error occurred in processing is rollback.System which provides redundancy and gives high availability is in standby namenode.System which handles different clusters requires operator to perform the task is highly operable. It allows one operator for the system maintainance of many nodes is highly operable.

The benefit of hadoop is its cost over many system. Earlier system have certain workloads which were not designed as engineered with the need of big data. It will be then too expensive of general use using very large amount of data.Hadoop lets the benefit of cost because it depend on the internally redundant data structure and it is deployed on industry standard server instead of very costly data storage system. It is understandable that once data is on tape, it is same as if it had deleted only in extreme circumstances.

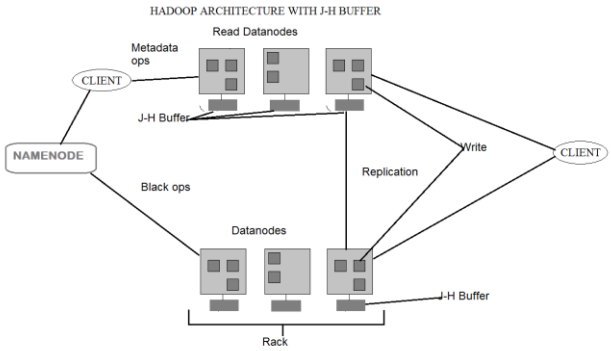
The growth of data is rapidly increasing and the 90% of accounting data is rising. Now the time has come for the enterprises to check their approach on data storage, analytics and management. For specific high value, low volume workload legacy systems will remain necessary and complement the use of Hadoop is optimizing the data management structure in organization by putting the right big data workload in the right system.The cost-effectiveness, scalability, and streamlined architectures of Hadoop will make the technology more and more attractive.

B. J-H Buffer(Joint Hadoop Buffer)

Joint Hadoop buffer is the clusters of buffer where the initial data is stored in a small space and is being processed for analysis. It can be used with hadoop architecture.

Initially information accumulated from the various sources are stored in the buffer of hadoop architecture. As there is a cluster of buffer near the datanode, each buffer has data processing system where the block of buffer data is being processed for analysis. Thus calculating the average of result came from data processing the final outcome is analysed.

J-H buffer in Hadoop Architecture



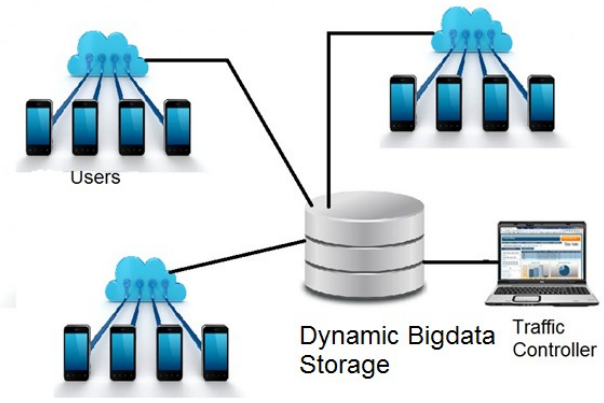
Namenode: Meta data is the data about data. This metadata is file system metadata which is stored by namenode i.e. file is maps to what block locations and which blocks are stored on which datanode. This name node preserve 2 in memory tables, block of datanode is map by one in memory table ie. one block maps to three datanodes for a replication value of 3 and on the other hand a datanode to block number mapping. When block of datanode reports a disk failure of a particular block, the first table gets updated and whenever datanode is detected to be dead (because of a node/network failure) both the tables get updated. This updation lead to information maintainace in the table.

2) Datanode
The residence of the actual data is data node. Some interesting think about data node is : Heartbeat message is send by datanode after 3 seconds to the namenode as a communication part that they are alive. If heartbeat is not received for 10 minutes by the namenode it assumes that the datanode is dead and initiates replication of blocks which were hosted on that data node to be hosted on some other datanode.
The communication between them take place to rebalance data, move and copy data around and keep high replication.
Checksum is maintained by the datanode when block of information is stored on the data node. The n updation of the namenode takes place with the block of information periodically and verify the checksum before update. If the checksum is incorrect for a particular block i.e. there is a disk level corruption for that block, it skips that block while reporting the block information to the namenode. Thus namenode is a wave of the disk level corruption on that datanode and takes steps accordingly.

IV. DYNAMIC BIGDATA IN VEHICULAR TRAFFIC CONTROLLING AND MONITORING SYSTEM

Infrastructure in road construction is increasing in the greater pace. With this increasing infrastructure it led to increase in the traffic. For well management of this traffic a system will require which gives the account of every details of traffic occur. As many details of data such as images, controlling of signals, videos were accumulated from various sources. This data may increase from 10 terabytes to 15 terabytes of data daily. There is an increase of 65% of storage data per year. This storage of data has to force it for a short span of time. A well developed storage system will enable the system to scale according to increasing traffic data.

This development can be seen the city, as it continue to develop. City police department can use this system with efficient for proper management of traffic and control. It is also beneficial for the travelers travelling on the road and to get the information about the traffic on their way. For giving this information system has to work very fast to give analysis in a very short time and give the desired result to the travelers. Every checkpoint has to be mounted with electronic equipment for capturing huge amount of images and videos.



If any user want to travel some distance if he may get an early warning of the traffic. This help the user to divert the path and save his time.
Through dynamic bigdata it can be possible for system to work efficiently. First the captured data has to stored in the storage space and the analysis is made immediately for executing the fast process. Hadoop system with J- H buffer break down the data into small chunks of small data which is then processed by every buffer. Hadoop provides a good platform for the distributed system to work in parallel. When the data is read, it is automatically verified and if a verification error is found, the operation is repeated in the dynamic bigdata storage.
This system will reduce the traffic to a greater extent. The data analysis can be used for effective

ISSN: 2278 – 1323

decision-making. With using the updated technology we can provide information for better traffic analysis.

V. SECURITY ISSUE

As the data increases, the challenges to keep the data safe also increases. Data security has to maintain for the users privacy policy, unknown threats. Dynamic bigdata are more difficult to secure in a word, variety.

But the business won't wait, some analysis shows that business intelligence and information management analysis of big data is a top priority. 47% are very or extremely interested vs 9% with no interest in big data analysis. This are driving by the many sectors which depend on 44% customer behavior, 43% cite finding correlations across multiple, disperate data sources. There is a 27% expect to predict fraud or financial risk.

But the big data landscape diverse across three areas.

Data form may be structured, like databases and transactional data, or it could be unstructured- think office documents, images and raw data stored as flat files.

Data sources include financial accounting applications, sales and product data, CRM applications, email files, sever logs, office files, images, mobile device data including geolocation and much more.

Data consumers range from department level analysis to senior business managers to IT and infosec teams to partners, customers and sundry business users.But across the board database security is shaky. 18% do no encrypt databases that contain sensitive information: 28% encrypt only some. 20% admit to breaches, dodge the question, or say they can't be sure. 24% do not security assessments, even as many shops add new databases like PostgreSQL, MonogoDB and Amazon DynamoDB.

It must accepts a few hard facts: pure perimeter and end points centric security is over. We need to focus on data. Many new big data analysis tools threat security as a secondary or tertiary requirement. Business users want constant and flexible access. To forget about that iron fisted control we had in traditional data warehouse/business intelligence models. There is significant challenge to establish a ownership of information. We should built a trust boundary to establish between data owners and data storage owners. For protecting the data adequate access control mechanisms will be required. Operating system provide the traditional access control or restrict access to the information which typically exposes all the information if the system is hacked. To protect the information using encryption which need the decryption only if system trying to access the information which is authorized by access control policy.

The problem of big data is to store the software in bigdata such as Hadoop, which does give authentication. It results in access control in worse condition as a default it would leave the information open to unauthenticated users. Big data depend on the firewall and the application layer which restrict access to the information.

VI. SECURITY WITH KERBEROS

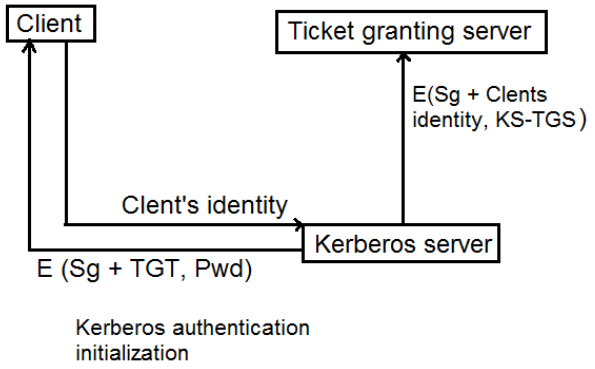
We need the security of dynamic bigdata while securing the hadoop cluster. It includes the authentication which is different from authorization. Hadoop uses the Kerberos which is bundled with Hadoop to provide authentication.

Kerberos is a centralized authentication service whose function is to authenticate users to servers and vice versa. Kerberos is an authentication service developed at Massachusetts Institute of Technology, USA for open network computing environments. It is based on some data such as when we log in through Kerberos, central server uses our user ID and password to create a token, which is matched against a private token on the server which we are authenticating. These token are called as tickets.
Kerberos tickets: A ticket is unforgettable, non replay and an authenticated message sent to requesting application. Once the Kerberos grants the ticket, we do not need to login again every time we communicate with the server. Kerberos uses two types of tickets in authentication process:
1) Ticket Granting Ticket(TGT)
2) Service Tickets.
Kerberos authentication model is available in two versions:V4 and V5

Authentication Process: At the initial stage, a session is established with the Kerberos server. The client authenticates itself to the Kerberos server which forwards the user's identity to a key distribution center(KDC). If the user is authorized, Kerberos server issues two messages:
1) A session key(Sg) which is used in communication with ticket granting server (G) and a ticket granting ticket (TGT) for ticket granting server, which is time stamped. This is encrypted under user's password, ie. (Sg + TGT, Pwd) and sent to user's workstation.

2) A copy of the session key(Sg) and user's identity is sent to the ticket granting server. This is encrypted using a key shared between Kerberos server and ticket granting server.
Now, the user is authenticated if and only if user can decrypt E (Sg + TGT,Pwd) by Pwd, the password of user. User Id and password remain secure, since they are never sent over the network.

This authenticates process is shown as follows:



Now, whenever user wants some services from distributed system, the session key Sg can be used. For eg. Using this Sg, user requests a ticket from ticket granting server to access a file. The ticket granting server the verifies if user is authorized and returns a ticket along with session key (Sf) for the file server. The ticket contains authentication information about user and the information about file which is to be accessed. It also contains an expiration date to prevent replay attack.

Benefits of Kerberos:
The Kerberos authenticates system offers many advantages over traditional authenticates system as follows:
Prevent eavesdropping: Kerberos prevents eavesdropping of password since user's passwords are never sent across the network. These passwords are stored at the Kerberos server. Only the secret keys are passed in an encrypted form across the network.

Provide mutual authentication: Clients and server are mutually authenticated in every process. Because of mutual authentication, continuous authenticity is provided.
Prevent from brute force and replay attack: The tickets passed between client and server include timestamp and lifetime information. Hence, Kerberos prevents brute force attack and replay attack.
Uses single sign-on method: Kerberos uses single sign-on method for authentication. A user needs to authenticate to Kerberos system only once and then authenticate to different services across the network without re-entering the password.
Prevents spoofing:
Protecting against spoofing is provided, since all requests are passed through ticket granting server.
Perform timely transactions: Kerberos has strict time requirements. If the host clock is not synchronized with Kerberos server clock, authentication fails.

This Kerberos system has to be implement while data is being processed which gives authentication and H-Hive to protect the data. With the implementation of dynamic big data this system has greatest advantage in securing data. Many organization has started to implement and know the importance of security in the market.

VII. CONCLUSION

Dynamic big data can deliver better products, but to effectively achieve this, it should be used to test the questions that have previously been impossible to answer until after the product has gone to market. But what was impossible five years ago is now mundane in terms of computing power and capability. It could help many organizations to successfully run their business. The application of J-H buffer in the hadoop will results in processing of data in simplest form. In the future, significant challenges need to be tackled by industry and academia. It is an urgent need that computer scholars and social sciences scholars make close cooperation, in order to guarantee the long-term success of big data and collectively explore new territory.

REFERENCES

[1] "Enhance Big data Security"- Advantech
[2] Hadoop Distributed File System- Robert Chansler, Hairong Kuang, Sanjay Radia, Konstatin Shvochko and Suresh Srinivas
[3] Towards Scalable Systems for Big Data Analytics: A Technology : IEEE paper – Han Hu,Yonggang Wen, Tat seng, Chua, XuelongLi
[4] Hadoop and Big data by Cloudera
[5] Improving Traffic Management with big data analytics: Intel Case Study
[6] "Security Issues associated with big data in cloud computing"- Venkata Narasimha Inukollu1 , Sailaja Arsi1 and Srinivasa Rao Ravuri
[7] "Big Data" – Big gaps of knowledge in the field of internet science : Chris Snijder , Uwe Matzat , Ulf-Dietrich Reips

ISSN: 2278 – 1323

Sachin Choudhary is currently studying in Mumbai University at MGMCET, Kamothe Navi Mumbai since 2011, currently pursuing B.E. in Computer Science and Engineering, with excellent academics . He is having interest in Studying New Technologies, and Student Member of CSI.

Sandesh Manohar is currently studying in Mumbai University at IMCOST, Thane since 2012, currently pursuing MCA with excellent academics . He is having interest in Studying New Technologies.

Sunil Salunkhe is currently studying in Mumbai University at IMCOST, Thane since 2012, currently pursuing MCA with excellent academics . He is having interest in Studying New Technologies.

