"Consideration Of Network Infrastructure For The Implementation Of Big Data Using Real Time Analysis "

Vilas D. Ghonge1, Vijay S. Gulhane2, Nitin D. Shelokar3

1ME (CSE) Scholar, Department of CSE,
Sipna College of Engg. & Tech. Amravati,

2Associate Professor, Department of CSE,
Sipna College of of Engg. & Tech. Amravati ,

3Assistant Professor, Department of CSE,
Sipna College of of Engg. & Tech. Amravati .

Abstract :-

Due to the explosive growth of data volume by mobile devices and SNS(Social Networking Service), Big Data has recently become one of the important issues in the networking world. Big traffic is generated as Big Data processing steps and multiple regionally distributed data centers are included, and data are delivered among clusters for the purpose of storage hierarchy management**.** Big Data produces big traffic and thus results in the significant burden to the network infrastructure. Therefore, the enterprise network should be optimized to support a strong foundation in terms of volume, speed, and accessibility of data for both traditional transaction-oriented RDBMS and diverse applications such as Big Data.

IJERT

Introduction

A big data is a collection of data set which is so large and complex that it becomes difficult to process using traditional data processing application. In general the big data is to be measured in Zettabyte. It means that it is 10 gigabytes i.e one trillion gigabytes[10] .The volume of data is increasing rapidly due to the tremendous network use also mobile communicating devices and social networking sites are increasing rapidly. Huge data is becoming an important issue in computer science field. It was seen that such a huge amount of data is generated due to the following main elements.

1.Mobile communicating devices

2.Social networking sight, various application of internet.

General Uses of big data.

There are many examples of big data not only in technology but also in the industry.

1) Media and entertainment companies are using big data in order to show their programme and to provide more focused marketing and customer analytics.
2) Healthcare providers are also using big data as they stored patient electronic health records from multiple sources such as imagery, treatment, biosignals, graphs and also pharmaceutical companies and regulatory agencies are creating big data solution to make their services efficient.[9]
3) Banks, financial services are adopting big data network infrastructure to help determine eligibility for equity capital, insurance, mortgage or credit.
4) Transport sector like railways, truck companies, airlines are using big data to track fuel consumption and their regular record in order to improve efficience.

Due to the big data generated from social networking site, mobile communicating devices create the huge amount of traffic with significantly increased real-time and workload transaction.

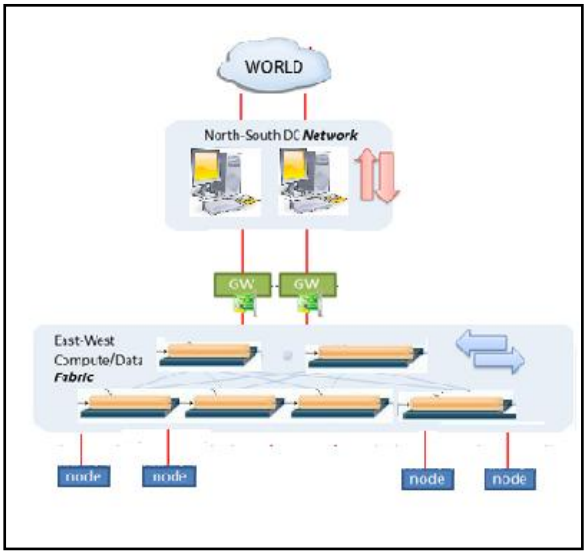

Fig1. Illustrating big data/cloud appliance.

There is a requirement of hadoop clusters in order to handle such a huge amount of data from the network .The hadoop structure requires high speed network infrastructure in order to proceed such a huge amount of data. The network infrastructure should be constructed in order to provide efficient, scalable, easily accessible data. Supporting the object oriented database system and various application such as big data[10]. The discussed network is very useful in order to reduce the workload at the given system, here we have given an input of 1 TB and then the output is 1 MB from that data. Considered network is also very useful in order to upload, download, extract, transform the given data[10].

IJERT

A Comparative Study between big data and traditional data.

The traditional data is small in size hence it is to be measured in Terabytes. In comparison with big data it is to be measured in Petabyte or Exabyte as it is large in nature. The architecture used for distribution in traditional data is centralized in nature whereas in case of big data it is distributed in nature. The data type used for traditional data is structured or traditional whereas the datatype for big data is unstructured or semi-structured in nature. Traditional data consist of known relationship whereas big data consist of complex and unknown relationships. Traditional data uses fixed schema data model whereas big data uses schema-less data model. This is the difference between big data and traditional data[9].

Networking Requirement.

The phenomenon of traffic shift from server-to-client pattern to server-to-server traffic flow among data center network fabrics due to the change of traffic sources and patterns by the big data[9].The hyperscale server architectures may consist of thousands of nodes which have many processors and disks. Therefore, the networking infrastructure which connects these nodes must be scalable and resilient for the optimal performance, especially when the data are shuffled among them during a certain application phase[10]. Big Data imposes its own computing infrastructure requirements and should incorporate essential functions such as creation, collection, storage and analysis of data[10].

IDC white paper[8] denotes that the network is an essential foundation for transactions between massively parallel servers within Hadoop or other architectures and between the server cluster and existing enterprise storage systems. In [8], the hyperscale network architecture mentioned above is called as "holistic network". The advantages to the holistic network approach were described as the following:
- Ability to minimize duplicative costs whereby one network can support all workloads,
- Multitenancy to consolidate and centralize Big Data projects,
- Ease of network provisioning where sophisticated intelligence is used to manage workloads based on the business priorities,
- Ability to leverage network staffing expertise across the data center.

Other factors affecting the design and implementation of networks was noted as governance or regulation requirements[8]. For example, in the application of health care, the separation of data plane may be necessary to meet the privacy requirements of sensitive data in the application.

Implementing big data infrastructure and analytics.

Big data requires infrastructure components, solutions and processes to address the following general challenges. Implementing a scalable network infrastructure for big data. Nowdays point to point switching fabric network is one of the best solution for big data.

The best way to describe a network switching fabric and its benefits is to refer to

Wikipedia's definition:

"Switched fabric, switching fabric, or just fabric, is a network topology where network nodes connect with each other via one or more network switches (particularly via crossbar switches, hence the name). The term is popular in telecommunication, Fibre Channel storage area networks, and other high-speed networks. The term is in contrast to a broadcast medium such as early forms of Ethernet. Switched fabrics can offer better total throughput than broadcast networks because traffic is spread across multiple physical links.

"A switched fabric should be able to function as a simple point-to-point interconnect and also scale to handle thousands of nodes. The term fabric derives its name from its topological representation. As the data paths between the nodes of a fabric are drawn out, the lines cross so densely that the topology map is analogous to a cloth.

"The advantage of the switched fabric is typically one of overall system bandwidth and performance versus connectivity between individual devices." The switching fabric architecture create a point to point connection between nodes with a single hop due to this there is significantly reduce in latencies between nodes. This network also consist of virtualizing the switching fabric, as it allows multiple networking components to behave as a single component.

Big Data Processing in real-time.

A huge amount of data collected from social sites, cloud services, traditional services like internet are passed through high speed low latency Ethernet. After passing through high speed low latency Ethernet it passes through various processes like operational systems, business analytics, indexing and metadata, big data analytics, governance system, flash appliances.

Backup data management of operational systems is to be done in shared databases, active indexes, shared metadata, archive data and metadata. For the implementation of big data there is a requirement of software infrastructure called hadoop. Hadoop is the big data management software infrastructure used to distribute , Cataog, manage and query data across multiple ,horizontally scaled server nodes.

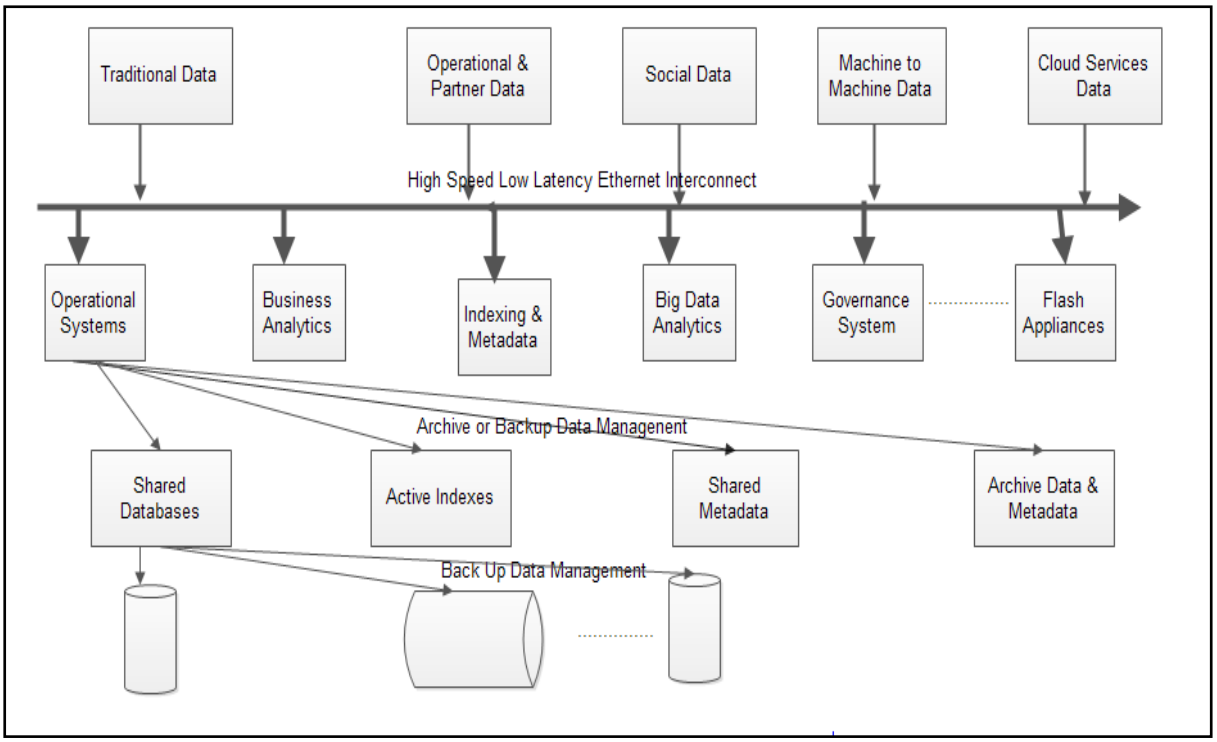

Fig2. Big Data Processing in Real-Time(Source: wikibon 2012)

IJERT

Output of the consider network

Hadoop, which includes a distributed file system known as Hadoop Distributed File System that provides a scalable file system infrastructure[9] . The considered network is useful in reduce the data having workload 1 terabyte data as an input and output it as 1 MB data. The considered network is also helpful in transformation of data from one format to the another format i.e the operation like extraction, transformation and load can be easily accomplished.

A Hadoop implementation creates four unique node types for cataloging, tracking, and managing data throughout the infrastructure[9].The data nodes are the repositories for the data, and consist of multiple smaller database infrastructures that are horizontally scaled across compute and storage resources through the infrastructure. Larger big data repositories will have numerous data nodes. The critical architectural concern is that unlike traditional database infrastructure, these data nodes have no necessary requirement for locality to clients, analytics, or other business intelligence[9]. The client represents the user interface to the big data implementation and query engine. The client could be a server or PC with a traditional user interface[9]. The name node is the equivalent of the address router for the big data implementation. This node maintains the index and location of every data node. The job tracker represents the software job tracking mechanism to distribute and aggregate search queries across multiple nodes for ultimate client analysis[9].

Conclusion

Big data is becoming an one of the important issue in recent years as big traffic is increasing rapidly. In this paper, networking infrastructure consideration to support big data is to be studied . Big data produces large amount of traffic in the network infrastructure.

The considered network helps in support to reduce data traffic and reduce the burden on the network .In order to use big data efficiently there is a requirement of hadoop infrastructure .

References

[1] Apache Hadoop. http://hadoop.apache.org/.
[2] Sung-Choon Lee, "The viewpoint on the Big Data utilization and communication industry", http://www.ktoa.or.kr, pp. 6-11, Vol. 60, Spring 2012.
[3] Sung-Choon Lee, Yang-Soo Lim, Min-Jee Ahn; Big Data: The key to open the future, KT Economics and Management Research Center Report, July 2011.
[4] Cisco White Paper, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2010-2015, Feb. 2011.
[5] Cisco White Paper, Big Data in the Enterprise: Network Design Considerations, 2011.
[6] Eung-Yong Lee, "The USA government Big Data R&D strategy", KISA, Internet and Security Issue, pp.3-26 August, 2012.
[7] Internet Research Group, Big Data, Big Traffic and the WAN, Jan. 2012.
[8] Lucinda Borovick and Richard L. Villars, The Critical Role of the Network in Big Data Applications, IDC White Paper, April 2012.
[9]JUNIPER NETWORKS, "Introduction to Big data: Infrastructure and Networking Considerations," White Paper.
[10] Yong-Hee Jeon , "Impact of Big Data : Networking Considerations and Case Study. IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.12, December 2012.

IJERT