IJARMATE
Your uh-MATE Research Paper !!!

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 2, Special Issue 12, April 2016.

Big Data Analytics on Virtual Private Network

Dr.C.Sathiyakumar[1], N.C.Santhosh[2], N.Vinoth[3]

[1]Department of CSE, K.S.Rangasamy College of Technology,Tiruchengode Email: sathiyakumar@ksrct.ac.in
[2]Department of CSE, K.S.Rangasamy College of Technology,TiruchengodeEmail: ncsanthosh7@gmail.com
[3]Department of CSE, K.S.Rangasamy College of Technology,Tiruchengode Email: nvinothcse@gmail.com

Abstract—*Big Data is data whose volume, variety, velocity, and/or timeliness need the use of fresh technical architectures and analytics to enable insights that reveal new sources of business value. Big data is for large or complex data sets that traditional data processing applications are insufficient. With the hasty development of the Internet world, network service has become one of the most habitually used computer applications. Search engine, web mail, and social network services are currently indispensable data-intensive applications. Because increasingly more people use web services, processing a large amount of data efficiently can be an important problem. Currently, the method for processing a large amount of data involves adopting parallel computing. The above existing system lacks in some of the concepts like data analytics in mobility. It also lacks in the configuration area. The proposed system aims on providing the data analytics process in mobile environment with faster access. And also it supports heterogeneous configuration among different nodes. In mobile environment the nodes are connected with server and with other nodes via a network which is created virtually. The proposed system helps in easy and simple access of big data. Shiny package by Rstudio provides a GUI for the R language. R language and Shiny package provides an environment for data processing and visualization.*

Keywords: Big data**,** GUI,Rstudio,Shiny,Virtual Private Network,Analytics, Multi node, Mobile platform, Heterogeneous environment

1. Introduction:

Big Data is data whose volume, variety, velocity, and/or timeliness need the use of fresh technical architectures and analytics to enable insights that reveal new sources of business value. Big data is for large or complex data sets that traditional data processing applications are insufficient. Challenges include analysis, capture, sharing, visualization, and information secrecy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and occasionally to a particular size of data set. Accuracy in big data may lead to more assured decision making. And better decisions can mean better operational efficiency, cost reduction and reduced risk.

Analysis on several data sets can find new connection, to "spot business trends, prevent diseases, crimes and etc." Scientists, business executives, government and in many fields regularly meet difficulties with large volume of data sets in areas including finance and business informatics. Scientists meet curbs in e-Science work, including meteorology, multifaceted physics simulations, and biological and environmental study.

Big data usually includes data sets with sizes beyond the capacity of generally used software tools to capture, and process data within an allowable pass by time. Big data "size" is a constantly moving from a few dozen terabytes to large petabytes of data. Big data is a set of techniques and technologies that need new forms of integration to expose large buried values from large datasets that are diverse, complex, and of an enormous scale. In addition, the data, due to its volume or level of structure, cannot be efficiently examined using only old databases.

These kinds of big data problems require new tools/technologies to store, manage and realize the business benefit. The new architectures it necessitates are supported by new tools, processes and procedures that enable organizations to create, manipulate and manage these large data.

2.Literature Review:

Big data analytics is the process of gathering, unifying and examining large sets of data to discover patterns and other useful information. Analysts working with big data basically want the acquaintance that comes from analyzing the data.With today's technology, it's possible to analyze your data and get solutions from it almost instantaneously an effort that's slower and less efficient with additional traditional business intelligence solutions.

Chia-wei et al performed dynamic data placement strategy for hadoop in heterogeneous environment [1]. The technique achieves the big data analytics in multimode clustering also using HDFS and Mapreduce. The concept had performed successfully by using Dynamic Data Policy statement.However, in a heterogeneous environment, a load imbalance will occurs and leads to create the necessity for spend additional overhead.

A prominent parallel data processing tool MapReduce is gaining necessary momentum from both industry and academia as the size of data to analyze grows rapidly [2]. While MapReduce is used in many areas where massive data analysis is necessary, there are still discussions on its performance, efficiency on each node, and simple abstraction. This survey intends to assist the database and open source groups in understanding various technical aspects of the MapReduce framework.
Approximate results based on samples often provide the only way in which advanced analytical applications on very massive data sets can satisfy their time and resource constraints. Unfortunately, methods and tools for the computation of accurate early results are currently not supported in big data systems [6].In fact, graphs are ubiquitous in the Internet and in our lives: as examples, take the connections between friends, the dependencies between providers and suppliers in today's complex business world [9].
Because of the less cost in the overall investment and greatest flexibility on the data provided by the cloud, all the companies are nowadays migrating their applications towards cloud platform. Cloud provides the larger volume of space for the storage and different set of services for all kind of applications to the cloud users without any delay and not required any major changes at the client level.The existing data mining techniques are inadequate to analyze those massive data volumes and identify the common services accessed by the cloud users. In this proposed scheme trying to provide an optimized data and service analysis based on Map-Reduce algorithm along with Big Data analytics techniques [10].

3.Proposed System

In the proposed system, the analytics processes are performed by creating the spontaneous network among the available nodes. The nodes in the network may have different operating systems, different configuration and so on. In this project, analytic concept of cloud environment (i.e. heterogeneous network) is implemented in the Virtual Private Network (VPN).

3.1 Existing System

The existing scheme is made up of traditional database methodologies like SQL. It can able to handle only structured data with the limited amount of data like gigabytes of data. In the existing system there is a need of dedicated centralized server with high configuration and the connected nodes are in same working environment with homogenous configuration. The analytics processes are performed on the preconfigured area. Users or clients should be present at that time in the environment.

3.2 Data Analytics

Data analytics (DA) is the science of scrutinizing raw data with the purpose of illustrating conclusions about thatdata sets. Data analytics is used in many industries to allow companies to make better business results and in the sciences to verify or contradict existing models or theories. Analysis of data is a process of examining, converting, and modeling data with the goal of realizing useful information, suggesting inferences, and supportive decision-making. Data analysis has multiple sides and approaches, encompassing varied techniques under a variety of names, in different business, science, and social science domains.

3.3Virtual Private Network

A Virtual Private Network (VPN) extends a private network across a public network, such as the Internet. It enables users to transmits the data across shared or public networks as if their computing machines were directly associated to the private network, and thus are benefiting from the functionality, security and management policies of the private network. A VPN is created by establishing a virtual peer-to-peer connection through the use of faithful connections, virtual tunneling protocols.

A VPN spanning the Internet is similar to a wide area network (WAN). From a user perception, the extended network resources are call up in the same way as resources available within the private network. Traditional VPNs are deliberated by a point-to-point topology, and they do not tend to support or connect broadcast domains. Therefore, communication, software, and networking, which are based on OSI layer 2 and broadcast packets, such as NetBIOS used in Windows networking, may not be fully supported or work exactly as they would on a local area network (LAN). VPN variants, such as Virtual Private LAN Service (VPLS), and layer 2 tunneling protocols, are aimed to overcome this limitation.

VPNs allow employees to securely access the corporate intranet while traveling outside the office. Similarly, VPNs firmly connect geographically separated offices of an organization, creating one ordered network. VPN technology is also used by individual Internet users to secure their wireless transactions, to by-pass geo restrictions and censorship, and to connect to proxy servers for the purpose of shielding personal identity and location.

3.4Protocols Used in VPN

- **Point-to-Point Tunneling Protocol (PPTP)** is the minimumprotected VPN method, but it’s a great early point for your first VPN because almost every operating system supports it.

- **Layer 2 Tunneling Protocol (L2TP)** and **Internet Protocol Security (IPsec)** are more secure than PPTP and are almost as broadly supported, but they are also more difficult to set up and are vulnerable to the same connection issues as PPTP.

- **Secure Sockets Layer (SSL)** VPN systems provide the same level of security that you trust when you log on to banking sites and other sensitive domains. Most SSL VPNs are referred to as "clientless," since you don't need to be running a dedicated VPN client to connect to one of them.

3.5 Ubuntu

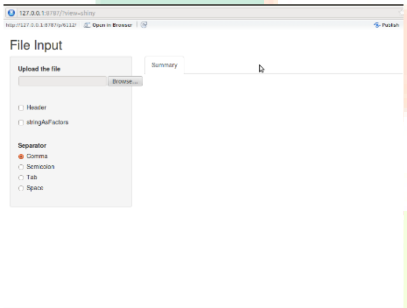
Ubuntu is a Debian-based Linux operating system and used in distribution for personal computers, smartphones and network servers. It uses Unity as its default desktop environment. It is an open source operating system and freeware. It named after the Southern African philosophy of Ubuntu.Ubuntu is built on Debian's architecture and infrastructure, to provide Linux server, desktop, phone, tablet and TV operating systems

3.6 RStudio

RStudio is a free and open source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface.

3.7 R-Shiny Package

Shiny is a web application framework for R language. It is a package that turn our analyses into interactive web applications. For this we don’t need any programming languages like HTML, CSS or JavaScript. It is a package from RStudio that makes it incredibly easy to build interactive web applications with R. The features oh Shiny are: This has an attractive default UI theme based on Bootstrap. It uses a reactive programming model that eliminates messy event handling code, so we can focus on the code that really matters.



File Input

Upload the file

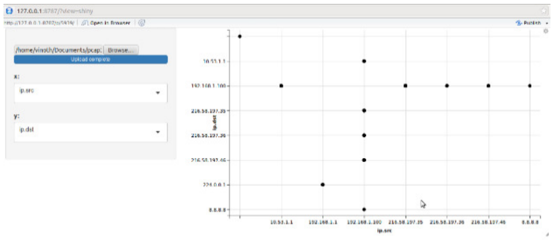
□ Header
□ stringAsFactors

Separator
● Comma
○ Semicolon
○ Tab
○ Space

Fig. 1 File Upload Page in R Shiny application

3.8 Data Visualization

Data visualization is the demonstration of data in a graphical format. It enables decision makers to see analytics presented visually, so they can understand difficult concepts or identify different patterns. With interactive visualization, anyone can take the concept a step additional by using technology to compute problems into charts and graphs for detail understanding and interactively changing what data you see and how it’s processed. It helps people to know the significance of data and its importance, also enables different dimensions of thinking and computing in the future for betterment of the result.



/home/vinoth/Documents/pcap:

x: ip.src

y: ip.dst

Fig. 2 Visualized Data set in R Shiny application

3.9 Apache Hadoop

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very huge data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (MapReduce). Hadoopdivides files into large blocks and distributes them amongst the nodes in the cluster. To process the data, HadoopMapReduce transfers wrapped code for nodes to process in parallel, based on the data each node needs to process. This approach takes benefit of data locality nodes operatingthe data that they have on hand to allow the data to beprocessed quicker and more efficiently than it would be in a more conventional supercomputer architecture that dependson a parallel file system where computation and data are connected via high-speed networking.

4. Conclusion

The main objective of the proposed methodis to analyze the big data sets based on the user choice. This visualization is done and obtain a result in the format ofSVG Image. The exactresults depends on the selection of columns, processing methods and variety of datasets obtained from the users. Series of observations are made on obtained datasets using their analysis and the results obtainedare evaluated to be more accurate and robust. The results would share to the user on the virtual private network. The resulted process would certainly enhance and help the current big data analytic techniques.

References:

[1] Chia-Wei Lee , Kuang-Yu Hsieh , Sun-Yuan Hsieh , Hung-Chang Hsiao,(2014),"A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments", Science Direct Article belongs to Scalable Computing for Big Data,pp.2-9.

[2] Kyong-ha Lee, Yoon-joon Lee, Hyunsik Choi, Yon Dohn Chung, Bongki Moon, (2011)," Parallel Data Processing with Map Reduce:ASurvey",IEEE Transaction on MapReduce,pp.11-18.

[3] Cohen.J, (2009)" Graph Twiddling in a MapReduceWorld",IEEE Transaction on Computing in Science & Engineering 11,pp.29-41.

[4] Kyuseok Shim, (2012),"Mapreduce Algorithms for Big Data Analysis",IEEE Journal on Big Data 5,pp.2016-2017.

[5] Reka Albert, HawoongJeong, Albert-LeszleBarabasi, (2000),"Error and attack tolerance of complex networks", IEEE Transaction on Networks 406,pp.378-382.

[6] Laptev.N, Kai Zeng, Zaniolo.C, (2013)," Very fast estimation for result and accuracy of big data analytics: The EARL system", IEEE Transaction on Data Engineering (ICDE), pp.1296-1299

[7] Xin Luna Dong, DiveshSrivastava, (2013)," Big Data Integration",IEEE Transaction on Data Engineering(ICDE),pp.1-3

[8] Aboulnaga, A, Babu.S, (2013)," Workload management for Big Data analytics",IEEE Transaction on Data Engineering (ICDE),pp.1249

[9] Elser.B, Montresor.A, (2013)," An evaluation study of BigData frameworks for graph processing",IEEE Transaction on Big Data,pp.60-67.

[10] Ramamoorthy.S, Rajalakshmi.S, (2013)"Optimized data analysis in cloud using BigData analytics techniques",IEEE Transaction on Computing, Communications and Networking Technologies,pp.1-5.

[11] Mukhopadhyay.S, Panigrahi.D, Dey, S, (2004)" Data aware, low cost error correction for wireless sensor networks", IEEE Transaction on Wireless Communications and Networking Conference 4, pp.2492-2497.

[12] Lin Gu, DezeZeng, Peng Li, Song Guo,(2014)"Cost Minimization for Big Data Processing in Geo-Distributed Data Centers", IEEE Transaction on Emerging Topics on Computing 3,pp.314-323.