Efficient Data Processing for Balancing Performance and Power Consumption with Big Data

[1]Mr. P. Dileep Kumar Reddy,[2]B. Muni Lavanya
[1]Associate Professor, [2]Assitant Professor (Adhoc)
[1,2]CSE Department
[1]SVCE, Karakambadi Road, Tirupati, Andhra Pradesh, India
JNTUACEP, Pulivendala, Andhra Pradesh, India

AbstractBig Data is one of the most emerging field in computer science and a broad area for analyzing, pre-processing and visualizing the large amount of data. In today's world, large volume of data is generating day by day and handling this type of data is too critical. Data with large volume, variety and veracity is generated. The existing methods worked on performance analysis but the result is not much optimized. This paper objective is to improve the performance of big data processing and in parallel it allows to consume less power. The techniques of parallel computing is introduced for increasing performance as well as reduction in power consumption. This kind of techniques allows system to process data in parallel and simultaneously visualizing the big data for business decisions. This is one of the major challenge facing with big data in real time scenarios like data centers and social media servers.

Keywords: Big Data, Parallel Computing, Performance, Power Consumption, Pre-Processing

---

I. INTRODUCTION

Big Data means large volume of data. The definition of big data depends on the storage device. For example, if a chip is capable of storing 1MB then 2MB is Big Data for that particular chip. In order to make easier to access this 2MB of data for that particular chip, the data can be partitioned into chunks and accessed easily within in a shorter span of time.

Big Data system contains functional block for acquiring data from the storage device. There are some pre-processing mechanisms used to cleanse the raw data [1].

The pre-processing steps are grouped together and called as Data Acquisition block in Big Data System. They are Data cleaning, Data Integration, Data Transformation, Data Reduction and Data Discretization. These are well known methods used for pre-processing the big data.

In order to make more efficient, Big Data can be accessed through online that is by using cloud services. This is also an emerging technology and most widely used in all organizations [2].

Cloud Computing (CC) is defined as accessing resources over a network channel. The resources may be any of these such as Data from the Data Center, Software, Hardware or an Infrastructure and also may be data of various types [3].

There are different formats of data. The raw data may be structured or unstructured format. Big Data with Cloud Computing is used in today's world for efficient access of data. The resources also called as services. These services are given by the cloud service provider [4]. There is Service Level Agreement (SLA) between the service provider and the client or end user.

The cloud technology with big data can be provided for days, weeks, months or yearly as an agreement. The use of this technology helps IT Sector to reduce its cost, time and also increased performance with optimized power consumption. This technology works most efficiently with the help of internet. If there is poor network connectivity then the service may not be delivered properly to the end user.

Data Acquisition in Big Data plays a critical role in performance evaluation. There are parameters like disk access time, response time and availability and many more. Some of them are considered for performance evaluation and results shown as improved performance in Big Data Systems [5].

The different formats of data that is structured or unstructured data can be categorized into different classes of data. The data can be private, and public. Whenever client is accessing some resource over a network channel, client must be authenticated with their credentials which are given at the time of service level agreement [6].

Security is provided while accessing data over a network. All the data in a cloud centre is encrypted at the time of storage. Sometimes, if it is a public cloud, then the data is not encrypted. The data classes in any cloud can't be accessed until and unless the successful authentication of the client. Re-Authentication is required, if the client needs to access another class of data. Data classes of one type will be accessed many times per day or some other classes of data may not be accesses at all. For optimizing performance, data is partitioned using different existing algorithms [7].

In addition to its generic property (e.g., its rate of generation), big data sources are tightly coupled with their generating domains. In fact, exploring datasets from different domains may create distinctive levels of potential value [8]. However, the potential domains are so broad that they deserve their own dedicated survey paper. In this survey, we mainly focus on datasets from the following three domains to investigate big data-related technologies: business, networking, and scientific research. Our reasons of choice are as follows. First, big data is closely related to business operations and many big data tools have thus previously been developed and applied in industry. Second, most data remain closely bound to the Internet, the mobile network and the Internet of Things. Third, as scientific research generates more data, effective data analysis will help scientists reveal fundamental principles and hence boost scientific development. The three domains vary in their sophistication and maturity in utilizing big data and therefore might dictate different technological requirements.

II LITERATURE SURVEY

Big data differs when data volumes, number of transactions and the number of data sources are so big and complex that they require special methods and technologies. [12] This also forms the basis for the most used definition of big data, the three V: Volume, Velocity and Variety as:

Volume: Large amounts of data, from datasets with sizes of terabytes to zettabyte.
Velocity: Large amounts of data from transactions with high refresh rate resulting in data streams coming at great speed and the time to act on the basis of these data streams will often be very short. There is a shift from batch processing to real time streaming.
Variety: Data come from different data sources. For the first, data can come from both internal and external data source. More importantly, data can come in various format such as transaction and log data from various applications, structured data as database table, semi-structured data such as XML data, unstructured data such as text, images, video streams, audio statement, and more. There is a shift from sole structured data to increasingly more unstructured data or the combination of the two. Organizations have a long tradition of capturing transactional data. Apart from that, organizations nowadays are capturing additional data from its operational environment at an increasingly fast speed. Some example are listed here.
Web data: Customer level web behavior data such as page views, searches, reading reviews, purchasing, can be captured. They can enhance performance in areas such as next best offer, churn modelling, customer segmentation and targeted advertisement.
Text data (email, news, Facebook feeds, documents..) is one of the biggest and most widely applicable types of big data. The focus is typically on extracting key facts from the text and then use the facts as inputs to other analytic process.[9]
Time and location data: GPS and mobile phone as well as Wi-Fi connection makes time and location information a growing source of data. At an individual level, many organizations come to realize the power of knowing when their customers are at which location. Equally important is to look at time and location data at an aggregated level. As more individuals open up their time and location data more publicly, lots of interesting applications start to emerge. Time and location data is one of the most privacy-sensitive types of big data and should be treated with great caution [12].
Smart grid and sensor data: Sensor data are collected nowadays from cars, oil pipes, windmill turbines, and they are collected in extremely high frequency. Sensor data provides powerful information on the performance of engines and machinery. It enables diagnosis of problems more easily and faster development of mitigation procedures.
Social network data: Within social network sites like Facebook, LinkedIn, Instagram, it is possible to do link analysis to uncover the network of a given user. Social network analysis can give insights into what advertisements might appeal to given users. This is done by considering not only interests the customers have personally [10].

III. EXISTING SYSTEM

Big Data is being capturing, storing and accessing from different locations. [5] Traditionally, data are stored in relational database (for example a CRM system for customer data, a supply chain management software for vendor related information) and some of these data are extracted periodically from the operational database, transformed and loaded into data warehouse for reporting and further analysis.
This is typically in the realm of Business Intelligence. Such process and tool set fall short when dealing with big data. For instance, one of the largest publicly discussed Hadoop cluster (Yahoo's) was at 455 petabytes in 2014 and it's grown since then. There simply is no parallel relational databases or data warehouse that have come even close to those kinds of numbers. Another sweet spot for Hadoop (over relational technology) is when data comes in unstructured format, such as audio, video, text [10].
It is worthwhile to mention that there is a general misconception that new technology, such as Hadoop is replacing other technologies, such as relational database. It is not the case. It is more likely that they are being added alongside each other. The sweet spot for a massively parallel relational platform for instance, is dealing with high-value transactional data that is already structured, that needs to support a large amount of user and applications that ask repeated questions of known data (where a fixed schema and optimization pays off) with enterprise level security and performance guarantee [3].
It is often called the Hadoop eco-system when discussing the various lays of technologies used to deal with big data. For a complete list, please refer to https://hadoopecosystemtable.github.io/.

An example stack might look like [4]:

- Amazon web service for infrastructure (in the Cloud and pay as you go.
- Apache HDFS (Hadoop Distributed File System) for distributed file system
- MapReduce or Spark for distributed programming model.
- Cassandra or HBase for non-relational distributed database management system [12].

Big Data Challenges are represented below:
Designing and deploying a big data analytics system is not a trivial or straightforward task. As one of its definitions suggests, big data is beyond the capability of current hardware and software platforms. These platforms in turn demand new infrastructure and models to address the wide range of challenges of big data. Recent works [2],[3] have discussed potential obstacles to the growth of bigdata applications. This paper strive to classify these challenges into three categories: data collection and management, data analytics, and system issues. Data collection and management addresses massive amounts of heterogeneous and complex data. The following challenges of big data must be met:
• Data Representation: Many datasets are heterogeneous in type, structure, semantics, organization, granularity, and accessibility. A competent data presentation should be designed to reflect the structure, hierarchy, 658 VOLUME 2, 2014 H. Hu et al.: Toward Scalable Systems for Big Data Analytics and diversity of the data, and an integration technique should be designed to enable efficient operations across different datasets.
• Redundancy Reduction and Data Compression: Typically, there is a large number of redundant data in raw datasets. Redundancy reduction and data compression without scarifying potential value are efficient ways to lessen overall system overhead.
• Data Life-Cycle Management: Pervasive sensing and computing is generating data at an unprecedented rate and scale that exceed much smaller advances in storage system technologies. One of the current challenges is that the current storage system cannot host the massive data. In general, the value concealed in the big data depends on data freshness; therefore, we should set up the data importance principle associated with the analysis value to decide what parts of the data should be archived and what parts should be discarded.
• Data Privacy and Security: With the proliferation of online services and mobile phones, privacy and security concerns regarding accessing and analyzing personal information is growing. It is critical to understand what support for privacy must be provided at the platform level to eliminate privacy leakage and to facilitate various analyses. There will be a significant impact that results from advances in bigdata analytics, including interpretation, modelling, prediction, and simulation.

IV. PROPOSED SYSTEM

This paper objective is to improve performance and to optimize power consumption. The following figure depicts the architecture of proposed model:

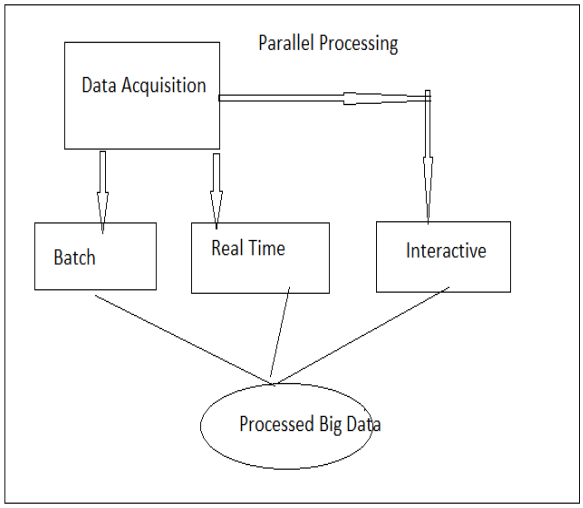

Fig: Architecture for Proposed Model

The techniques of parallel processing is applied in data pre-processing stage for power consumption reduction an also to improve speed that is directly relate to the performance of the big data application.

The processes are categorized into three different types. They are Real time, batch and interactive processes. These can run on background as well as foreground as an application. The access time is reduced with this kind of separation and can be processed at a time.

The time of application execution reduces, so that the performance of the application also increases and also balances the utilization of power consumption.

The collected data sets may have different levels of quality in terms of noise, redundancy, consistency, etc. Transferring and storing raw data would have necessary costs. On the demand side, certain data analysis methods and applications might have strict requirements on data quality. As such, data pre-processing techniques that are designed to improve data quality should be in place in big data systems. In this section, research efforts for three typical data pre-processing techniques which are already available in the market.

They are Integration, Cleansing and Redundancy Elimination.

Data Integration: The extraction step involves connecting to the source systems and selecting and collecting the necessary data for analysis processing. The transformation step involves the application of a series of rules to the extracted data to convert it into a standard format. The load step involves importing extracted and transformed data into a target storage infrastructure.

Data Cleansing consists of five complementary steps:

- Define and determine error types;
- Search and identify error instances;
- Correct the errors;
- Document error instances and error types; and
- Modify data entry procedures to reduce future errors.

Data Redundancy: Data redundancy is the repetition or superfluity of data, which is a common issue for various datasets. Data redundancy unnecessarily increases data transmission overhead and causes disadvantages for storage systems, including wasted storage space, data inconsistency, reduced reliability and data corruption.

V. PERFORMANCE VISUALIZATION

The following visualization of graphs indicates the performance measured with the help of various parameters. The analytics of accessing, analyzing, storing and retrieving data is shown in the form of graph as follows:

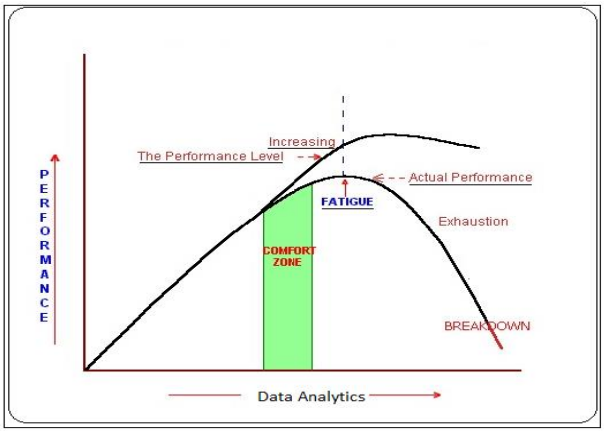

Fig: Performance Visualization

VI. CONCLUSION

In this paper, Parallel Processing and Big Data with cloud Computing is introduced for ease of accessing data and efficient improvement in performance and balance in power consumption of Big Data application. This proposed scheme helps the client to access the big data applications over online with more security.

REFERENCES

[1] Mark A. Beyer and Douglas Laney. "The Importance of 'Big Data': A Definition". Gartner, 2012

[2] Bill Franks. "Taming the big data tidal wave". Wiley, 2012

[3] David R. Hardoon and Galit Shmueli. "Getting started with business analytics – insightful decision making". Talor & Francis Group.2013 Foster Provost and Tom Fawcett. "Data science for business". O'Relly, 2013

[4] Thomas H. Davenport and D.J. Patil . "Data Scientist: The Sexiest Job of the 21st Century", Harvard Business Review, 2012

[5] Mr. P.Dileep Kumar Reddy , Dr. R. Praveen Sam, Dr. C. Shoba Bindu, "A Tripartite Partite Key Assignment Scheme For Security Of Cloud Dataclasses", Journal of Theoretical and Applied Information Technology, 15 th July 2017. Vol.95. No 13.

[6] WassimItani; AymanKayssi; Ali Chehab "Privacy as a Service: Privacy- Aware Data Storage and Processing in CloudComputing Architectures", 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Pages: 711 & 716, DOI: 10.1109/DASC.2009.139.

[7] Md. Rafiqul Islam; Mansura Habiba, "Agent based frame work for providing security to data storage in cloud" 2012 15th International Conference on Computer and Information Technology (ICCIT), Pages: 446 451, DOI: 10.1109/ICCITechn.2012.6509712

[8] Sandeep K. Sood , "A combined approach to ensure data security in cloud computing', ELSIVER, Journal of Network and Computer Applications, Volume 35, Issue 6, November 2012, Pages 1831–1838.

[9] LeinHarn Hung-Yu Lin "A cryptographic key generation scheme for multilevel data security", ELSIVER, Computers & Security, Volume 9, Issue 6, October 1990, Pages 539-546.

[10] DongyangXu; FengyingLuo; Lin Gao; Zhi Tangfine grained document sharing using attribute- based encryption in cloudservers" Third International Conference on Innovative Computing Technology (INTECH 2013).

[11] Yi-Ruei Chen, CHU Cheng-Kang, Wen- GueyTzeng, Zhou Jianying " CloudHKA: A Cryptographic Approach for Hierarchical Access Control in Cloud Computing", International Conference on Applied Cryptography and Network Security (ACNS), 26 Jun 2013.

[12] Hadoop-The Definitive Guide by Tom White.