Analysis of Big Data: Challenges and Fundamentals in the Computing System

Department of Business Administration, Adjunct Professor, Theology and Sciences Institute of Florida, USA
Department of Business Administration, Doctoral Student, Theology and Sciences Institute of Florida, USA
Advanced Information Science and Engineering, Master of Engineering, Ritsumeikan University, Japan,
Department of Computer Science, Bachelor of Science Computer Science, Dadabhoy Institute of Higher Education, Karachi, Pakistan
Faculty of Education, Bachelor of Education, Preston University Karachi, Pakistan
Faculty of Arts, Bachelor of Arts, University of Karachi, Karachi, Pakistan
Sr. Full Stack Developer, T-Mark Inc., Tokyo, Japan
muhammad.suhailt-mark.co.jp

Abstract
We are living in Digital universe with data prolife ring by Individuals, Institutions and Machines at an extremely high rate. This data is grouped as "Big Data" due to its Volume, Velocity and Variety. Most of the data is not structured, quasi structured or semi structured and it is mixed in nature. The capacity and the heterogeneity of information with the speediness is generated, makes it challenging for the current computing structure to administer Big Data. The development of Big Data and its increasing ability to process and store a wide array of data has become more efficient with the introduction of High-Performance Computing (HPC) technologies. Its integration with Big Data workloads ensures reliability and safety of data for large corporate sectors. However, with a massive growth of wide spectrum of data that is generated through supercomputing, its efficient usage has become time-demanding and challenging for computational space to warrant successful analysis and processing of data. Though technologies such as Hadoop and Predictive Analysis necessitate the indispensability of Big Data, some software like Apache and YARN often disturb the analysing process of HPC for Big data as a result of which performances of such systems gets disruptive. This paper will entail the resolution of such problems and a subsequent implementation of Big Data system in HPC technologies in a more efficient way.

Keywords: Big Data, Computing System, Hadoop, Predictive Analysis

Introduction
Traditional data management, warehousing and analysis systems fall short of tools to analyse this data. Due to its specific nature of Big Data, it is stored in distributed file system architectures. Hadoop and HDFS by Apache is widely used for storing and managing Big Data. Analysing Big Data is a challenging task as it involves large distributed file systems which should be fault tolerant, flexible and scalable. Map Reduce is widely been used for the efficient analysis of Big Data. Traditional DBMS techniques like Joins and Indexing and other techniques like graph search is used for classification and clustering of Big Data. These techniques are being adopted to be used in Map Reduce

<table>
<thead>
<tr>
<th>Big Data Origin</th>
<th>Big Data Target Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Science</td>
<td>(a) Scientific discovery</td>
</tr>
<tr>
<td>2. Telecom</td>
<td>(b) New technologies</td>
</tr>
<tr>
<td>3. Industry</td>
<td>(c) Manufacturing, process control, transport</td>
</tr>
<tr>
<td>4. Business</td>
<td>(d) Personal services, campaigns</td>
</tr>
<tr>
<td>5. Living Environment, Cities</td>
<td>(e) Living environment support</td>
</tr>
<tr>
<td>6. Social media and networks</td>
<td>(f) Healthcare support</td>
</tr>
<tr>
<td>7. Healthcare</td>
<td></td>
</tr>
</tbody>
</table>

Figure 01: Big Data Origin and Target Domains

Technological ascendency has played a pivotal role in the growth and success of digital data. It has waned the technological barriers by reducing communication gaps and increasing availability of data in digital form (Gandomi and Haider, 2015) [1]. Since huge data lacks effective handling by utilizing traditional models, the invention of High-Performance Computing (HPC) proves to be a significant breakthrough in the
technological world. According to Hashem and Khan, 2015 [2], the traits of Big Data encompass veracity, authenticity, velocity and value. Most of the current systems and applications used by both public and private sectors are dependent upon collecting large-scale data for instance, Hadoop. IBM has defined Big Data Analytics by referring it to the usage of advanced techniques against different data sets which include different sizes and forms. For better and well understanding of Big Data like what’s the big data and what’s the traditional data? the table below represents a comparison of both.

<table>
<thead>
<tr>
<th>Traditional Data</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>Photos</td>
</tr>
<tr>
<td>Finances</td>
<td>Audio and Video</td>
</tr>
<tr>
<td>Stock Records</td>
<td>3D Models</td>
</tr>
<tr>
<td>Personnel files</td>
<td>Simulations</td>
</tr>
<tr>
<td></td>
<td>Location data</td>
</tr>
</tbody>
</table>

Figure 02: Big data understanding

The genesis of big data and its consequent development has led to the beginning of the era of manifold technologies, which acquires the ability to process, analyze and store enormous volume of data with high velocity. Taking this perspective into view, the transformation of High-Performance Computing from resolving hackneyed problems to newer ones related to healthcare, finance and food technology necessitates involvement of Big Data analytics for better and credible handling of technological data (Gandomi et al., 2015) [1]. The analytical requirements of Big Data can be fulfilled with high performing computing technologies.

Related Work

(Fox and Kamburugamuve, 2015) [3] is of the view that large volume of generated data can be stored and retrieved with the help of web page content and emails, which is facilitated by several software products. A brief discussion on some of the products are articulated below:

The Hadoop Ecosystem

Hadoop is regarded as the most easily available program which acquires the ability of processing large volume of data in an eco-friendly computing environment. With this program, large amount of data running over clusters of applications and servers can be analysed. It occurs as a result of involvement of several thousand nodes within which terabytes of information is available (Kambatla and Grama, 2014)[4]. In order to store the data in blocks, a Hadoop Distributed File System (HDFS) is used, which runs on the ‘Moving Computing is cheaper than Moving Data’ dogma. In addition, ‘MapReduce’ programming, a model for generating big data will distributed algorithm. It is a part of Hadoop ecosystem which performing functions like sorting and filtering (Kambatla et al., 2014)[4]

HPC and Environment for Big Data

As per the latest technical reports released by scientific society of the UK, data that is produced presently is quite large in size. Sometimes, it reaches within the ambit of Petabytes. In order for HPC system to provide high-performance big data processing, its programming environment needs to be preserved. The existing models and software for HPC that are designed do not comply with the environmental measures to prevent any malfunctioning within the big data application (Kambatla et al., 2014) [4]. MPI is the most credible program that supports multiple programming features, such as bulk modulation.

Evaluation of High-Performance Computing Architecture

There are many techniques through which HPC environment can support big data applications. Data locality and load balancing One of the types of load balancing technique is work stealing which is utilized for scheduling systems. They have, in turn, multiple schedules and decisions to mark upon. Different tasks are assigned to loaded schedules from idle ones. It is necessary for HPC environment to increase work stealing technique via dedicated queues (Hashem et al., 2015) [2].

There are certain frameworks brought into such HPC programming. Mesos is one of the credible online resource sharing platform. It can work for numerous computing framework. It deploys scheduling system and lets the framework acquire data from long
distances. Besides, Charm++, Dryad and others are important independent parallel programming frameworks used for load balancing techniques.

<table>
<thead>
<tr>
<th>Framework</th>
<th>Features</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mesos</td>
<td>Resource-sharing model for task scheduling. Features include acquiring data locality and reading, coupled with deployment of delay scheduling for storing nodes.</td>
<td>Takes much time to generate any data, particularly for large data set.</td>
</tr>
<tr>
<td>Charm++</td>
<td>Used for parallel programming. Within it, the whole statistical load balancing occurs in a distributed manner</td>
<td>Penurious scalability</td>
</tr>
<tr>
<td>Dryad</td>
<td>Regarded as execution system for parallel applications. Supports the application by running them, like the Hadoop ecosystem</td>
<td>Neither versatile nor rational mapping process through centralization scheduling</td>
</tr>
</tbody>
</table>

Table 01: Features of the frameworks used for HPC Programming

Quality of Service for HPC

Since HPC has become a key area of interest for big data application, the utilized applications require rigid service quality for communication and interpretation for large amount of network resources. The challenges to handle better quality service include optimization of power consumption, fault-tolerant network, light-weight applications, and others (Nystrom and Scott, 2015) [8]. Proper allocation of resources and charging methods are necessary to be learned and used to balance proper consumption of the programs.

Predictive Analysis

Predictive Analysis is another big data analytics program that forecasts imminent behaviour of a computer based on previous encrypted data. It depends upon machine learning technologies to predict what will be going to happen. The most effective use of predictive analysis is detecting any fraud or malfunctioning of the data (Jagadish and Shahabi, 2014) [6]. The recent advancement of artificial intelligence has led to the necessity of predictive analytics solution. Numerous renowned companies such as IBM and Microsoft rely upon this program.

![Figure 04: Predictive Analytics (Source predictiveanalyticstoday.com)](image)

Models used for High-Performance Computing

Since different techniques for high performance computing have been developed, this paper encompasses three main parallel models based on parallel computing models, which are used for high-performing computing technology (Fox, 2015) [3]. The features used in these models entail OpenMP, MPI and MapReduce; each have their own distinctive features.

1. OpenMP
This model is regarded as multiprocessing Application Program Interface (API) that is used for parallel programs of shared memory. It provides certain directives to the program to manage the shared memory, create threads, and synchronize running operations. This model uses a block-structured approach to switch between parallel programs. At the edge of a parallel block lies a single thread of control, which controls other number of threads. Gandomi (2015) [1] described some extensions to OpenMP that implement certain features which are needed for non-uniform memory access.

OpenMP Syntax

```c
#include “omp.h”
int main ()
{
 int var1, var2, var3;
//Serial code
 . . .
//Beginning of parallel section.
#pragma omp parallel private(var1, var2)
shared(var3)
{
//Parallel section executed by all threads
 . . .
//All threads join master thread and disband
}
//Resume serial code . . .
}
```

Figure 05: Source Code Example

OpenMP Programming Model

• Distributed memory, thread-based parallelism
  – OpenMP is based on the presence of numerous threads in the shared memory programming paradigm.

• A distributed memory procedure consists of several threads.
  • Explicit Parallelism
    – Programmer has complete control over parallelization. OpenMP is not an automatic parallel programming model.
  • Compiler directive based
    – Most OpenMP parallelism is stated across the use of compiler directions which are embedded in the source code.

2. MPI

MPI is a message protocol for programming parallel computers. MPI’s aims are high functioning, scalability, and manageability. The MPI interface is intended to deliver important synchronization, virtual topology, and interaction functionality amid a set of processes (that have been mapped to nodes/servers/computer instances). MPI has been widely used in High-Performance Computing. This model, in contrast, is a type of message which passes library specification, which in turn defines message passing model for parallel programming on the computing environment. Some researchers proposed algorithm ideas to implement this sort of model to analyse big data applications from far-fetched localities. At times, Intel core computers are used for implementing MPI by using the approaches of array #D and string vectors. Afterwards, the effects on the performance of computers is analyzed (Gandomi, 2015) [1].

The issue of scalability in MPI is a concern for current big data producers and there is a need for addressing the issue in order to be scalable. However, the main advantage of MPI is that it is a fine match for applications in which space and time are significant.

3. Map Reduce

MapReduce is a programming model and related to implementation for managing and producing large data sets with a parallel, distributed algorithm on a cluster MapReduce is program used for Hadoop ecosystem. It is regarded as a big data processing framework. Usually, Hadoop system is a cluster of numerous commodity computers used for distributed file system, also called Hadoop Distributed File System (HDFS). It has various matches with existing distributed file systems, Earlier MapReduce was created to develop application on datacentres along with hundreds of servers. It is, unlike MPI, a promising scalable model on shared memory systems (Fox ,2015) [2]. It was recently improved to a new model called Map-Reduce-Merge, where an already partitioned data can be merged by map.
Comparative Analysis of the Models

On a sheer distributed memory system, these models are considered to be imperative for big data analytics using HPC. The table below articulates the traits for the implementation of the programming models.

<table>
<thead>
<tr>
<th>Operation</th>
<th>OpenMP</th>
<th>MPI</th>
<th>MapReduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Architecture</td>
<td>Shared Memory</td>
<td>Distributed and Shared Memory</td>
<td>Share Memory</td>
</tr>
<tr>
<td>Programming Model</td>
<td>Compiler</td>
<td>Compiler and Filer</td>
<td>Library</td>
</tr>
<tr>
<td>Synchronization</td>
<td>Implicit</td>
<td>Implicit or Explicitly</td>
<td>Implicitly</td>
</tr>
</tbody>
</table>

Table 02: Programming Model Features

Out of the above three models, MapReduce based program is considered as the best choice among rest of the models for HPC since it aligns with the necessity of using HPC devices for big data analytics. In contrast, OpenMP and MPI models have also been experienced as remain non-updated with the passage of time for better handling of data.

The overhead table displays the comparative analysis aspects of the dual chosen tools in big data established on compatible operating system and. The key objective of this assessment is not to criticize which is the greatest tool in big data, but to determine the usage and to construct attentiveness in several fields.

Challenges in Big Data

Big Data is a comprehensive term for large and complicated datasets where conventional data managing applications are inadequate. The combination of this massive data sets is completely complex. There are numerous challenges one can face throughout this integration like data analysis, data curation, sharing, capture, search, visualization, information, storage and privacy. The management of big data is extremely complex. Several challenges tackled during its integration contain ambiguity of data Management, big data ability gap, receiving data into a big data structure, syncing throughout data sources, receiving suitable information out of the big data, , skill availability, volume, velocity, solution, cost and data complexity.

Graph 01: Challenges involved in Big Data Analytics

(Katal and Goudar, 2013)[5] stated that as large-scale data analytics are rapidly becoming significant tools within the technological community, the present stacks related to big data frequently grapple with the maintenance of reasonable performance of high performing computing systems. The software used in the systems do not always align with the workload of supercomputers and as a result, the system gets unable to collect and store sets of data for corporations.

However, it has already been experience by technological companies’ data big data still lags behind basic employability on some computing resources. The problem exacerbates. Since supercomputing systems have become heterogeneous, the problem will exacerbate to a certain degree (Katal et al., 2013)[5].

It was also argued that in the recent years, tools for big data analytics have proved to be important for enabling the computing systems to process and analyse data in a smooth manner. Many have forestalled the convergence of big data workloads with high performing computing (Jagadish et al., 2014)[6].

This gap in the performance of HPC is not only limited to lack of effort from big data tools. Instead, the visualization applications within HPC have provided non-HPC systems some foothold to reckon it among high performing systems. While visualization has enabled big data analytics to have
bearings upon workloads, some performance issues within the system have been witnessed (Singh and Reddy, 2015). A huge stockpile of big data software assumes to have effectively level of control over different operating environment, some unique tools like Apache and YARN are unable to interact with HPC tools and querying systems (Kambatla et al., 2014).

Another glaring instance of the non-alignment of big data analytics with HPC entail interconnect utilization. Usually big data operates on cloud-based system which utilize Transfer Control Protocol (TCP) their employment with supercomputing is limited. Though current HPC interconnects render lower latencies and more bandwidth, they lose their leverage when TCP appears to be emulated (Qiu and Fox, 2014). There is no second opinion that new HPC technologies exert considerate resources for message-based communication with Message Passing Interface (MPI) customization, they are not regarded as compatible with big data technology.

The issue of heterogeneity of HPC is also witnessed in the storage and memory of big data system. As new systems have designed to help HPC better cope with such diversity of the big data, there is a scarcity of thought for more potential benefits to the analytics-related platforms (Katal et al., 2013)[5]. Not only a challenge of performance or storage of large amount of data, Big Data Analytics also sometimes grapple with the uses of cost and human skills involved in maintaining Big Data.

Resource Utilization

The research on how HPC or other heterogeneous sources can effectively be utilized necessities the implementation of resource management and meta-scheduling tools that both enhance data analytics system as well as develop the ability to transfer implementations from either Apache or HPC-related schedulers (Hashem et al., 2015). Moreover, a set of familiar and open-source libraries must be developed for different computational classes such as memory system, GPUs and others (Waller and Fawcett, 2013)[7]. Lastly, despite visualization research creates some performance issues, it is still significant for the provision of efficient ecosystem access to all available hardware.

Decomposition Techniques and Parallel Programming

Decomposition techniques for HPC are designed for multiple CPU solving problems. Usually, performance goals vary according to scalability, which is a computation number for gauging per unit time within which a certain process is occurred. Looking through this perspective, parallel programming improves responses of such decomposition techniques. The main impediment for parallelism is the Java language can exploit the CPU only when it is present as well as work on shared multi processes (Nystrom et al., 2015)[8]. Other decomposition techniques involved Join or Fork decomposition that depends on parallel notions of techniques which are associated with algorithm design solutions. They encompass matrix, merge sort and many other processing algorithms. For maximum parallelism, task should be smaller so gain more opportunities for it (Nystrom et al., 2015)[8]. It gradually improves locality, load balancing and neutralized the time percentage are certain CPU wait for another for greater process.

It is necessary to note that different decompositions and mapping emanate better performance for a given problem.

Study Outcome

Big data makes a difference in the leverage of companies to drive their business performances. A number of companies range from smaller business corporations to tech giants like Amazon or Google efficiently handle their workloads by involving Big Data Analytics system (Singh et al., 2015)[9].

The influence of Google in the present technological world cannot be analysed without defining its achievements vis-a-vis big data. As big data and big businesses go cheek by jowl, Google’s innovation of Big data has helped it expand its basic search capability. As a result, according to a general estimation, as many as 3.5 billion queries are calculated each day, and the number is rapidly increasing (Singh et al., 2015)[9].

The way in which Google handles its data involve several bots which creeps the web, copy down what has been searched and take it back in the index’s database. What put Google at the forefront against other search engines is its ability to quickly read a wider data instantly. Google launched its universal search engine in the year of 2007, which consequently pulled thousands of sources like historical dates, financial data and language databases into its functions (Kambatla et al., 2014) [4].

Presently, the on-going newly designed big data system Google has been working on is the ‘self-driving car’. With the help of generating large amount of data from various cameras and sensors, coupled with the feature of real-time analysis, it is going to allow the Google car to drive on the roads safely without any physical endeavour from a human driver. As such machines are becoming good learners of
physical tasks, big data analytics are becoming more imperative (Singh et al., 2015). It is not doubted that Google may easily predict the future with the help of such self-driving cars, its current positions as an influential innovator in the big data space appear to be a credible bet.

Conclusion

The volume of data that is drifting across the internet nowadays, not individual that is large, but it is very complex as well for Universities, institutions, Organizations and healthcare system. However Big Data offers a lot of benefits, it arrives with its personal set of issues. This is a new set of complicated technologies, while still in the growing stages of development and evolution. In this paper, a comprehensive description of the concept of Big Data and its requisites in the computing system is discussed. Further, the HPC techniques related to sustainability of big data has also been studied. It has been conceded that the analyses of large data still remain a challenging conundrum (Waller et al., 2013). Each kind of Big data has its various processing and analysing techniques, which are used for cloud computing, data streaming and machine learning. It is undeniable that further research will be taken into consideration to resolve existing problems in the Big Data system for its more efficient implementation (Qiu et al., 2015) [3].

References


AUTHORS PROFILE

MUHAMMAD SUHAIB

Work as Sr. Full Stack Developer at T-Mark Inc., Tokyo, JP, Experienced Software Engineer with a demonstrated history of working in the information technology and services industry. Born 1988 in Pakistan, Pursuing a Doctoral Degree in the area of Business and HealthCare Management Information Systems from Theology and Sciences Institute of Florida, USA , Graduated from Department of Advanced Information Science and Engineering, Master of Engineering, Ritsumeikan University, Japan, Graduated from Department of Computer Science, Bachelor of Science Computer Science, Graduated from Faculty of Education, Bachelor of Education, Preston University Karachi, Graduate from Faculty of Arts, Bachelor of Arts, University of Karachi, Pakistan.