Key-Value Combination for Scholarship Disbursal Analysis through Map-Reduce

Rajiv Pandey\textsuperscript{1}, Archana Sahai\textsuperscript{2} and Namrata Nagpal\textsuperscript{3}
\textsuperscript{1}Sr. member IEEE, Asst. Professor, Amity University Uttar Pradesh, Lucknow-India.
\textsuperscript{2}Asst. Professor, Amity University Uttar Pradesh, Lucknow-India.
\textsuperscript{3}Asst. Professor, Amity University Uttar Pradesh, Lucknow-India.
\textsuperscript{1}rpandey@lko.amity.edu, \textsuperscript{2}asahai@amity.edu, \textsuperscript{3}nnagpal@lko.amity.edu

Abstract. Big Data includes both structured and non-structured contents. Big data is being generated by devices and increasing in volumes. The social and digital landscape produces this colossal data, although devices, handheld, and IoT devices add to the size. Data set with respect to a Central University scholarship domain in a country like India can make the dataset still larger and complex. The complex dataset in the coming years is sure to certify as Big data in terms of velocity, volume, variety, veracity, and value, thus various Bigdata analytical tools are needed for analytics of this huge dataset. The purpose of research on such a dataset is to retrieve meaningful information that facilitates decision-makers for quick and flawless decision making and redirecting scholarships to the deserving candidates across the country. This paper explores the storage of big data and map reducing through a key-value analysis for scholarship disbursal decision making thus catalyzing Digital India vision.

Keywords— Big data, Map Reduce, scholarship disbursal analysis. Hadoop, IoT

1. Introduction

Big Data is a includes both structured and unstructured data, originating from digital ecosystems that may exhibit categorical and quantitative component, the size of data is growing manifold with every minute. IoT (Internet of Things) is significantly contributing to the task of generating data. Social networks and pervasive computing are a major source of data generation. The Digital India Vision will quantify the data set across Central universities of India to Big Data, as in the future all records, mark sheets, and other documents(images), queries through Facebook and twitter are sure to be the contents of the said data set. Big Data thus depicts the datasets whose sizes are beyond the ability of typical database software tools to capture, store, manage, and analyze [1].

Big data representation assumes an OR-ing operation of the various features like V’s and C’s i.e. velocity, volume, variety, veracity, value and Complexity, cardinality, continuity, respectively. Managing and organizing data is an important task in rapidly growing economies and developing geography like India. Data is useless unless it is synchronized and brings some value to the end-user, organization, or enterprise. Big Data emerges as a technological upgrade to the datasets, which in time improves processes and make real-time outputs and quality decisions. Even small organizations and enterprises use data in terabytes to certain petabytes, every day. Social media, machine transactions, digital entertainment such as videos, images, audio, etc. are likely important results for aiding of Big Data.
Data on the internet and organizations is accumulating across the mix of structured, semi-structured, and unstructured data that needs to be analyzed, managed, and optimized to give a value greater than that of the conventional techniques. This can also be helpful in processing and flow of the data with ease. A dataset such huge and diversified needs and appropriate file structure to store the data and appropriate programming algorithms to process it.

This paper describes the evolution, need, and challenges with big data and it also determines the growing important uses of big data. This paper explores the key-value representation at map-reduce stages to enable processing concerning scholarship disbursal analysis. The paper is organized in the undermentioned six sections: 1) Introduction 2) Big data defined and illustrated 3) Requirement of Big data 4) Scholarship Data Is It Big? 5) Challenges to Big Data 6) Mining of Big Data 7) Figuring out Map – Reduce and 8) Conclusion.

2. **Big data defined and illustrated**

“In huge databases, interesting information is hidden called KDD (Knowledge Discovery from databases)” [3]. “KDD covers a variety of analysis methods such as distributed ecosystems, pattern and image recognition, data warehousing and mining, NLP, sentiment analysis, statistical and vision analysis and HCI” [4].

Interesting information needed for the human understanding ability of Big Data cannot be discovered with traditional analysis methods because Big Data exhibits features as “Data Volume, Velocity and Variety and the aspects of Complexity” [5]. There do also exist various V’s differentiating small data and big data such as:

**A. Volume**

Volume estimates indicate more than many Zeta bytes of data exist in the digital universe. On Facebook billions of users comment and upload more than millions of images each second. The volume of data that is being created and collected every day; Fig 1 is an immediate challenge to the businesses.

![Fig. 1 Rapid Growth of Data](image)

**B. Velocity**

Velocity refers to the speed at which the data is being created, stored, analyzed, and visualized.

**C. Variety**
The varied form of data requires various storage technique.

D. Veracity
Volume and Variety can be a big problem if the data is Incorrect or faulty i.e. the mass production of data with different varieties of wrong data is useless in other words Garbage in Garbage Out (GIGO).

E. Value
Value of data and its analytical significance to each organization. Big data has a potential annual value of € 250 billion to Europe’s public sector administration,[6] that is a lot of value.

3. Requirement of Big Data
Uses of Big data are different for every different industry and organization. Some would need to know the customers well while others need to know the popularity and acceptance of customers. The generic needs of big data for different industries and organizations [12] are:

- Rendering effectiveness to e-governance
- Real-time customer analysis
- Product Innovation and Renovation
- Determining the risk factors
- Examining new Markets and Business
- Staying ahead of customers

4. Scholarship data is it big?
India is a diverse country with a federal backbone. Scholarships and student financial initiatives are rendered by both Central and state governments some time independently and some collectively. The data set thus needs to be seen unified as well as distributed.

It is about Variety and Volume: The focus to collect every single useful information in such a way to be utilized in future demands of citizens/students. It demands an ability to handle a new type of data: voice, text, images, and video Futuristic Central university dataset would contain:

- The virtualized and integrated information of all students across India (Centre and State on need basis)
- Add to variety in the dataset as inputs may be images, audio/video queries, and social media inputs
- Tools (Map Reduce) for real-time data set analysis incorporated.

The scholarship dataset would thus display the features that pertain to a Big data thus we will have to treat the dataset with appropriate Bigdata analytical tools this paper looks to make a beginning by suggesting the key-value representation in context to Hadoop, Map-reduce framework. Above parameters are only considered from the scholarship domain but the context in real implementation would be diverse.
5. **Challenges to Big Data**

We need to ponder and decide whether the data has to be stored and analyzed through KDD or in Hadoop architecture?

- KDD/Data warehouse: structured data, data “trusted”
- Hadoop: semi-structured and unstructured data. Data “not trusted”

Big Data needed by all organizations where data is generated in volumes and variety. Such that to store data and to retrieve the meaningful information for their respective organization to generate more revenue and to refer to the related product to the specific customer.

Big Data can have a major impact in our context in the undermentioned dimensions were by we can:

- Discover hidden patterns
- Enhance decision making by providing summarized graphically represented data.
- Automate business processes to enhance KDD.
- Quick error-free disbursal of scholarship and other financial incentives.

6. **Mining of Big Data**

The big deal of mining, specific information from the ocean of varieties of data is described by Cluster computing. The classical way of data mining is based on the single-node architecture which includes a processing unit, a memory, and a disk. The processing of an algorithm running requires the data that needs to be fetched from the disk to the memory.

![Fig. 2 Single node classical Data mining](image)

The role of Cluster Computing begins when the data exceeds the memory need to be processed inside the CPU. The program is executed in portions in such a case therefore transfer of large data is a big issue. Another flaw is the bandwidth, or the speed needs to carry data is very low comparing to the data sets of terabytes to certain petabytes from disk to memory.

Cluster Architecture is a combination of multiple racks where each rack contains 16 to 64 commodity Linux nodes and each node is essentially working as an independent personal computer, joined together through a switch, and further, the switches are connected to a higher bandwidth backbone switches and soon.
The cluster computing comes with its limitations which are

- Data persistency and availability: The nodes may fail within the data center after a period or due to the heavy and long computational process.
- Network Bottle Neck: The transfer of huge data throughout the network of nodes can delay down the computational process.
- Distributed programming: Distributed Programming becomes very complicated.

7. **Figuring Out Map Reduce**

Since cluster computing has major limitations, the need was felt to have a file system to store distributed data and to develop a programming model where the code moves to the data. Thus, file systems like GFS and HDFS evolved and a programming model was introduced named MapReduce. MapReduce removes the challenges of cluster computing and makes it a powerful tool in distributed processing. It addresses the challenges with help of three main techniques. MapReduce can be divided into two stages [14]:

- Map stage: the master node data breaks the problems into many small subproblems. Each worker node processes a part of the problems by the supervision of the JobTracker node and stores the result in the HDFS, where a reducer can access it.
- Reduce Step: This stage makes many to one on certain predefined rules.

A storage and computation framework that shall enhance the data storage and analysis can be perceived as below:
Fig. 4 Hadoop and its components.

Fig. 5 framework will enable job distributions on the slave nodes and the job tracker supervision on the master node.

Storing data redundantly in multiple nodes which will protect and make available the data even if any node fails. In simple terms Distributed File System (DFS). A distributed file system helps in creating redundant data stored in chunks in machines, and each chunk will have a replica which is stored different machine which helps in protecting and availability of data. The machines are often known as chunk servers. Beyond the storing of files, a master node is created which stores the data about data (Metadata) i.e. where and when it is stored or updated. And master nodes are even replicated and stored safely. The popular examples of DFS are Google GFS and Hadoop HDFS.

- Moving the computation close to the data instead of mass amount of data transferred to be computed. And direct contact with chunk servers fixed.
- The simplicity of the programming model. It simplifies the algorithms for distributed processing.

In context to a scholarship distribution analysis framework, a typical Hadoop-MapReduce model is described in Fig. 5 and Fig. 6 which will help in the analysis of meritorious students across campuses of Central Universities across India situated at various locations for the purpose of merit.
scholarships based on grades and parameters of concern. The data items and their respective key-value pairs at various stages of MAP –REDUCE are as follows
- Map Input (K, V) = (campus-state, program)
- Shuffle (K, V) = (Program, student list)
- Sort (K, V) = (sorted/grouped program, no of students)
- Reduce (K, V) = (Program, aggregated count for scholarship)

The resultant output after the reducing stage will now give us the top-grade list of student program wise across all campuses of the central university. The Government can easily decide on the number of scholarships to be disbursed in each program.

8. Conclusion
The paper has suggested the use of Hadoop, Map-Reduce environment for effective decision making through system supported data analytics in context to a model suggested for all central universities which may have its campuses located geographical across the country, and the management wishes to disburse scholarships to students based on a collective grading across the universities, not taking into consideration the geographical separation. The suggested computation/interpretation could be done by a RDBMS but when the size grows, and variety adds to the data sets the computation time is huge such analytics if futile and not desirable to any organization.

References
[1] McKinsey Global Institute (MGI), Big Data: The next frontier for innovation, competition, and productivity. Report, June, 2012.
[2] http://smalldatagroup.com/2013/10/18/defining-small-data/
[3] Miller, H.J. (2010) “The data avalanche is here. Data Avalanche of Regional Science” page 9-10
[4] SAGIROGLU S. ,SINANC D “Big Data: A Review” 978-1-6404-1/13 ,IEEE 2013
[5] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety” (6 Feb 2001)
[6] https://ezinearticles.com/?How-Big-Data-Can-Help-Your-Organization-Outperform-Your- Peers&id=7917125
[7] Gray, Jim. " Data Management: Past, Present, and Future," June 1996
[8] Fremont Rider, The Scholar and the Future of the Research Library 26 (1944)
[9] https://gcnc.com/Articles/2013/05/30/GCN30-Timeline-Big-Data.aspx?Page=2
[10] Peter Lyman and Hal Varian, UC Berkeley “How Much Information?” , 2011 IDC Digital Universe Study
[11] O'Reilly, 2005, "What is Web 2.0", (2005 )
[12] Thomas H. Davenport and Jill Dyché, “Big Data in Big Companies”,, International Institute for Analytics and SAS, May 2013.
[13] http://invent.ge/uah9yo
[14] K Bakshi, “Considerations for Big Data: Architecture and Approach” Aerospace Conference IEEE, Big Sky Montana, March 2012
[15] C. Verma and R. Pandey, "Big Data representation for grade analysis through Hadoop framework," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 312-315, doi: 10.1109/CONFLUENCE.2016.7508134.
[16] Pandey, Rajiv, Nidhi Srivastava, and Shahnaz Fatima. "Extending R Boxplot Analysis to Big Data in Education", 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015.CSNT 2015.
[17] Rajiv Pandey, Manoj Dhoundiyal. "Quantitative Evaluation of Big Data Categorical Variables through R", Procedia Computer Science, 2015