International Journal of Engineering & Technology, 7 (4) (2018) 5633-5639

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.24404
Research paper

Securitizing big data characteristics used tall array and mapreduce

Wael Jum'ah Al_Zyadat [1] *, Faisal Y. Alzyoud [1], Aysh M. Alhroob [2], Venus Samawi [2]

[1] Assistance Professor. Faculty of Information Technology, Isra University, Jordan
[2] Associate professor. Faculty of Information Technology, Isra University, Jordan
**Corresponding author E-mail: waael.alzyadat@iu.edu.jo*

Abstract

Volume, velocity, variety, veracity, and value are the main characteristics of big data; researchers consider them in the classification process. This study contemplates two of these characteristics (Data Volume and Veracity), as major attributes; the scale of data and accuracy proved to be issued in relation to varying boundaries. In the scenarios discussed by two methods, Tall array and MapReduce are used; as they were used to work with out-of-memory data. Tall array subdivides the data sets into small chunks that individually fit in memory, while MapReduce uses parallelization and distribution by enabling mapper function and reduce function respectively. Theoretical Model and Experimental simulation show that tall array method is more efficient compared to MapReduce as per F-Measure and Arithmetic Mean calculations; in tall array method, veracity is improved by 0.09 and 0.15 in respect to F-Mean and Arithmetic Mean, meanwhile volume is improved by 0.06 and 0.13.

Keywords: *Big Data; MapReduce; Tall Array; Veracity and Volume.*

1. Introduction

Digital Data term comes from digital technology revolution (e.g. mobile phone, sensors, and GPS devices) that provides connectively and interdependent anytime and anywhere (where the internet available) that support the services rely on technology that produces the data-rich; it's strengthen from all parts of society by international development agencies, government, and non-governmental organization. In fact, data-rich is stenography to Big Data (BD). Especially, in the first component named Data generation that refers to the immense volume of data; the Data analysis component operates from multiple resources of data integrate and organization. Furthermore, data ecosystem component which is bridge big data and small data they are combined quantities (Data collection) and qualities to act analysis [1].

A large amount of data or data complexity leads to a term of BD. Moreover, when the data processing in traditional method is insufficient to deal with the capturing data, data analysis, transfer data, search, visualization, updating, information privacy and data storage these are main challenges and disruption in BD. As well as, inability to understand and manage this data [2].

The survey report done by World Economic Forum, (2015) discussed shift 11: BD for Decision the first government to replace its census with BD source expected on 2023; the purpose to automated decision-making faster and better. On other hands, impacts are positive and negative, the positive impact envisions a real-time decision-making and cost saving; the negative impacts pretended in implementation phase especially algorithms are owner and battles. Seemingly BD permeates from a product based on services based [3].

The BD characteristics differ from traditional data; which is commonly unstructured and require real-time analysis that generating unprecedented quantities of data by online platform, undoubted called data deluge era [4]. Demystifying BD through two sides are inside and outside both effects on quality, quantity, and accessibility to acquiring meaningful data. Inside data requires significant planning and effort to scale up and is often difficult to scale down such as relevance data, matching data as well as classifier data (Centre Data). Outside BD able to change and evolve in the structure without requiring data schema re-design, data migration, or new data repositories [5].

The rate with the utilization of internet has expanded as of late, additionally brought the making of BD. The amount of data display is huge to the points, that examining and using those utilizing customary strategies for information examination inconceivable. This has prompted the appearance of BD, and from that point forward it has changed the way the world dealt with this heaping amount of data. Furthermore, there are many benefits of capturing and controlling the BD such as cost reduction, better decision making, and innovative product, amending strategies actions, accurate risk management, and service launches. As observed, the BD affects a different human activity domain empowered by significant growth of the computer power, everywhere availability of computing and storage resources [6].

Delve on storage resources by digital technology side is split into Gigabyte to Terabyte, which is underlying the hardware architecture embark on different perform in a parallel database such as a shared-memory database, shared-disk database, and shared-nothing database, these treated with structured data. The internet era involved the all kinds of data, structured data, unstructured data, semi-structured data and mixed structured data these shifting to stage Terabyte to Petabyte has attained a high degree of industry companies' penetration, which ignited indexing and query data to overhauling kind of data, that consider cover sheer volume foster the development advanced management BD [4]

The significance of BD is that it's big to the point that it can't practically be devoured by any current framework in a span genuinely

Copyright © 2018 Wael Jum'ah Al_zyadat et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sufficiently short to make a significant expectation. The request to process every bit of it is because the exactness of the forecasts is enhanced significantly more by expanding the measure of information rather than by expanding the unpredictability of the handling calculation on a smaller data set. To overcome these points, the BD is characterized as follows to describe the main dimensions of BD analysis as shown in Figure below.

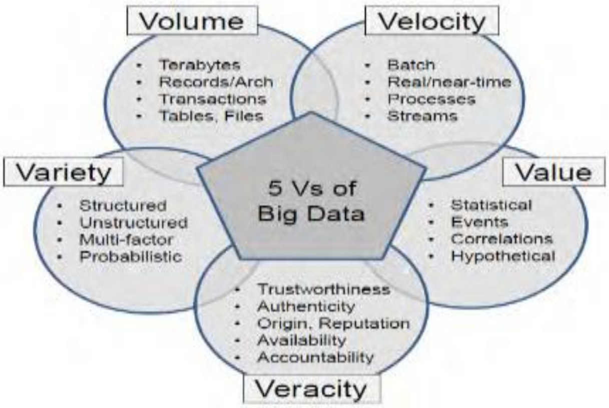

Fig. 1: 5vs. of Big Data.

i) Volume: is concerned about the scale of data immense, the primary attribute of BD as well as can also be quantified by counting records, transactions, tables, or files [7, 8].

ii) Variety: many important sources of BD are relatively new, such as a social network, GPs, and sensors.

iii) Velocity: the term refers to the growing speed at which this data is created. Furthermore, it indicates the ability to accept and responding to actions as they occur.

iv) Veracity: when data are big in a manner of volume, velocity, and variety, it is facing to get fully corrected data. The data accuracy and correctness of analysis turn on the source data.

v) Value: due to the huge of the potential value of BD, the value dimension is the most important aspect, which transforming the quantities of the data volume to understanding.

In this study, address the challenge of unstructured data which pretended clearly at dimensions 5Vs of BD analysis. Furthermore, the data structure data and Relational Database Management System (RDBMS) both are only treated with structured and semi-structured data that demands to determine the suitable mechanism, tools, and framework via comparison under 5Vs.

BD frequency and size are considered important parameters in determining the storage mechanism, format, and the pre-processing tools [9]; as well depend on the data sources which can be classified as follows: on demand as the data used with social media data, continuous feed data as the BD resulted from traffic, real-time data as weather data, and time series data as time-based.

The contemporary BD technologies provision of structured data revolves around key-value databases, column-oriented document-oriented databases, and graph-based. Despite the different nature of unstructured data and/or semi-structured request to be in structured data to perform visualization, predicted, decision making and meaningful scale [8].

2. Theoretical frameworks

In this section provides three frameworks were different domains that shown dimensions of BD were applied capture (data collection) and enables efficient to the life cycle of data throughout phases such data collection, data integrity, data analytics, and data management utilization for operational.

2.1. Big data framework for electric vehicle range estimation

The main factor that limits the distribution of the electric vehicle is the estimation of energy power that vehicle can run before the next charging. An Example of real-time data is jam caused by accident sudden rain. Historical data and real-time data manipulation are an obvious challenge, a combination between real-time data and historical data can be used to adjust future driving prediction range based on standard data, historical data, and real-time data. The proposed framework that collects: route information, weather data, driving behavior data, electrical vehicle modeling data, and battery modeling data. These BD sets that have been introduced to estimate the remaining driving range of electric vehicles based on the power consumption. The collected data is based on different standards, historical and real-time data from different sources as shown in Figure 2 [7].

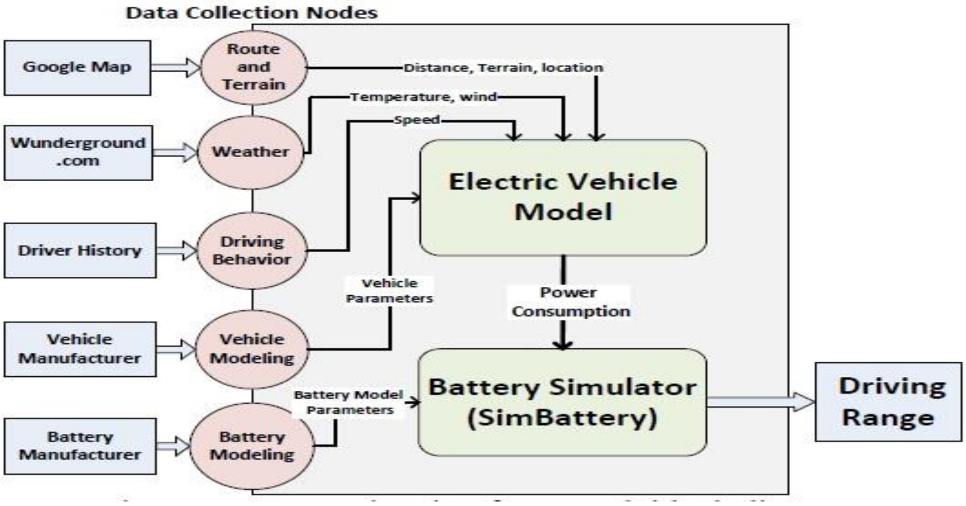

Fig. 2: Range Estimation Framework Block Diagram.

The collected data presents the volume dimension that holds on physical storage with a huge capacity for stabilization. Meanwhile, the source of data that collected from different platforms as an online platform is wunderground.com node; sensor platform applied via google map node. Furthermore, real-time platform that presents in speed calculator from driver history node all of them belongs to a variety dimension.

Through estimating electric vehicle model using several parameters such as location, distance, temp, and speed these requests make action toward to velocity dimension. Consequently, veracity dimension is displayed in power consumption which combined among volume, variety, and velocity.

Value dimension aims to this framework that pinpoints to the prediction for whole parameters that coverage by mixture type of the data. Firstly, is standard data and historical data that are fully structured in physical storage, secondly is real-time data that unstructured so its named raw data; the issue is how to analyses both types of data, as well as the estimation among the platforms through inner parameters under scale conditions, are time, accuracy and performance. This framework under the type of BD-continue feed, BD-real-time, and BD-time series.

2.2. Car stream: an industrial system of big data processing for internet-of-vehicles car stream

The CarStream applied through Internet-of-Vehicles which scalable by big data especially an issue for large volume with low value and quality [10]. Figure 3 shows the volume dimension presented collected data in Carstream are Vehicle status, passenger order, driver activity, and trajectory.

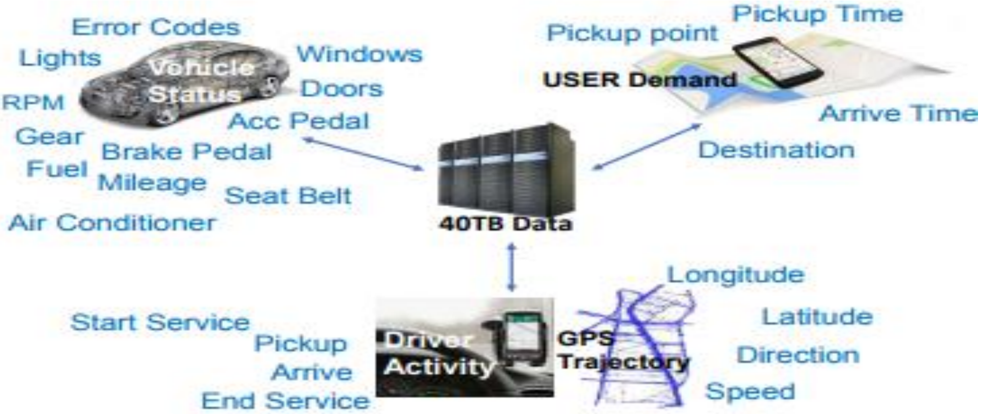

Fig. 3: The Data Processed in Carstream.

The data receiving with multiple structures and differing features that request both stream processing (online) for real-time data and batch processing (offline) for user information that use in data management layer via key-value structure use SQL and NOSQL to provide value dimension.

The variety dimension produces in back-end services Trajectory compression by a change amount of data in time status results (delay or disordered); the control data in multiple functions whatever type online or offline that request pre-process data, monitoring, and tracking that covered by velocity dimension.

The decision making, predictive, allow update in storage in-memory cash that conclude in a purpose of value dimension, the Car Stream use KAFKA apache to apply distributed Messaging System, for real-time stream processing use Jstorm Cluster. On another hand; Hadoop for batch processing.BD management uses the HBase, HDFS RDBM, and NOSQL.[11]

2.3. Big data collection and utilization for operational support of smarter social infrastructure

Hitachi is currently seeking to perform smart city platform from different social infrastructure data interdependency on BD, the combination between new and old infrastructure demand coordinate the exchanging between data from sources as shown in Figure 4.[12], [13].

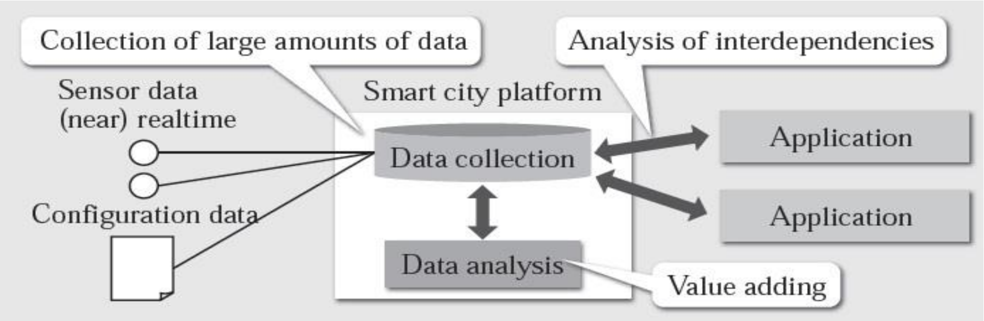

Fig. 4: Roles of Smart City Platform.

Data through Electric Vehicles (EV) use sensors such as Home Energy Management (HEMS) and Building and Energy Management Systems (BEMS) which belongs to historical data. Journal data is data move by topology on equipment operations or alarm, in cooperation use information-control platform to perform the interdependency and interpolation data. Data Analysis: adjusts three roles to achieve high-level decision and prediction application; first role data required from the application, second inherit the function among application, the third role is knowledge acquisition which required data in the storage mood to determine prediction parameters [14].

The dimensions present at smart city platform firstly, volume dimension generate from historical data, journal data, and transaction through topology; the source of data came from power planet by sensors, information that feed on applications that mean the variety dimension. consequently, the life cycle of data that collected from sensor till transaction through topology to storage as information which allow the function information control platform tread with it is velocity dimension.

The knowledge acquisitions analyses information that comes from different boundaries with alarms to find out the high-level decision regarding roles named veracity dimension. On the other hand, the prediction availability whatever from parameters or application with the correct result and face malfunction is valuing dimensions. That combined BD-continue feed, BD-real-time, and BD-time series.

In this section, explained the significance of data collected based on BD characterizes which aim to convert unstructured to structured data with scrutinizing the item, location, and content data.

3. Tall array and mapreduce two sides for one coin

3.1. Tall arrays

Tall arrays were released since 2016 by R2016b MATLAB; they were used to work with out-of-memory data by treating data in multiple files as one large stack array. Tall arrays enable users to work with large data by sets in small chunks that individually fit in memory, tall arrays enable users to work with different types of data in which MATLAB can work on including numeric data, cell arrays, strings, dates and categorical. In tall array data that lives on disk, may be distributed among different disks (distributed file-system). Furthermore, the calculation is performed by stepping data through different files using different defined methods; the following steps illustrated the architecture of tall array in Figure 5.

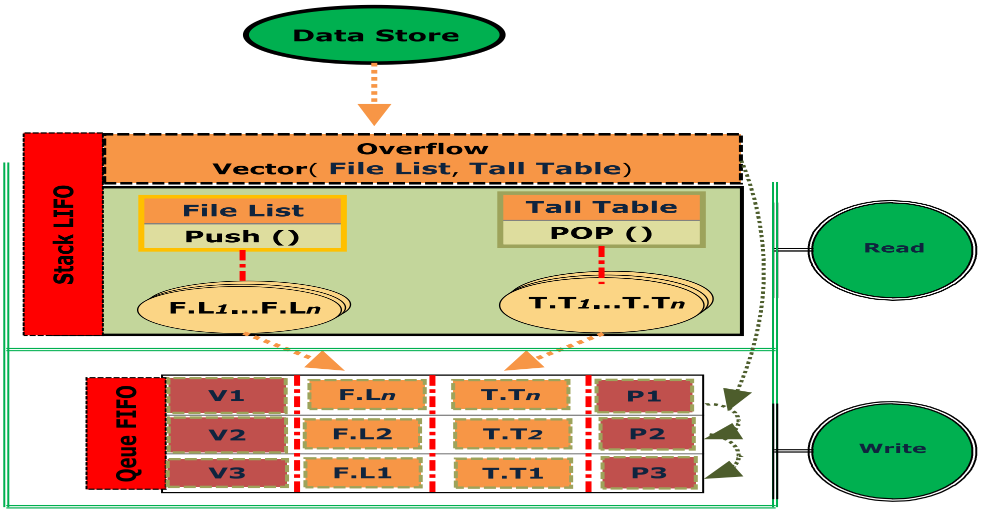

Fig. 5: Architecture Tall array.

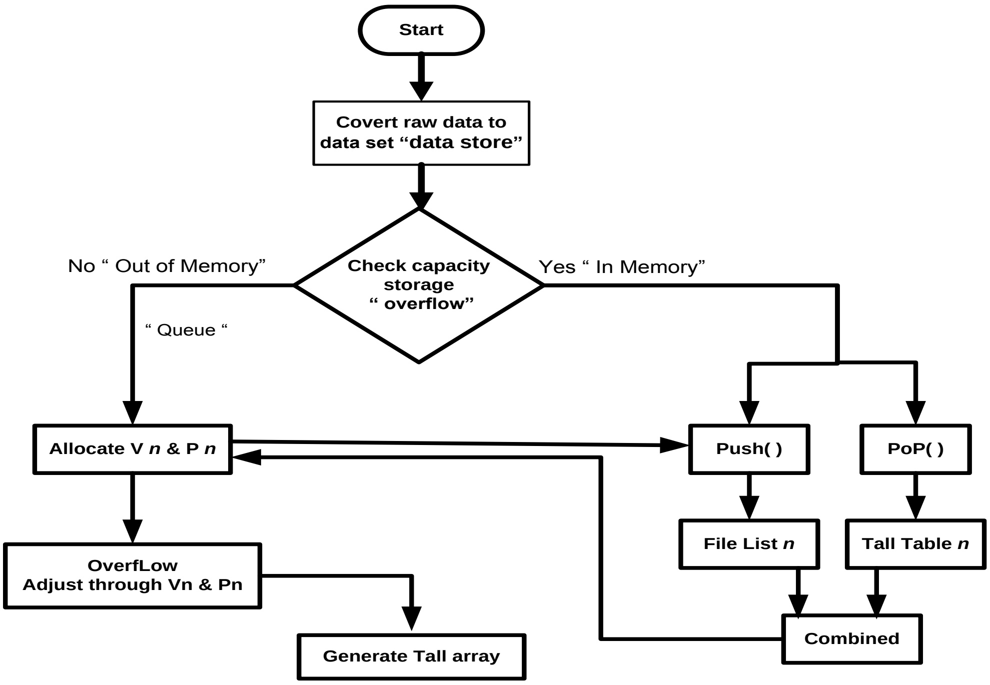

Fig. 6: Flow Chart Tall Array.

When the data is collected from a source as raw data, the first method is to store data using data store, whatever the capacity is available to data (In-memory) or not (Out of memory), in both statuses able to treat with use the overflow operation as an initial step in stack phase, here, overflow important to determine the memory availability and sorting for vectors in creating tall array. A consequence, it is a major operation stack (read) and queue (write).

Since a stack is a list with the two restrictions. Firstly, an insertion equivalent to push operation dwell on file list; secondly deletion equivalent to a pop operation carried out in dynamically allocated (position data in a table) supported via top to consider tall table.

A queue as well list, involve two operations are enqueued operation that adds (V and P elements in each vector in tall array) and an insert focus on elements are output from stack (File List (F.L) and Tall Table (T.T)); the second operation is dequeue which implements delete when the vectors are fully moving to next vectors in tall array. The pop and dequeue operations generally are similar implement by deletion. But in precise have different to find out at two concept terms are abstract data type (ADT) stack via Last in First out (LIFO) and queue via First in First out

(FIFO); second term content data the stack considers in the data store. Meanwhile, perform such read and monitor dataset. On the other hand, the queue manipulates by two extra items vectors and partitions.

The overflow operation is constants stack and queue to establish the allocated raw data to a tall array which expand location and content in vectors.

The flow chart as shown in Figure 6 presents the processes of tall array, as follows:

Step1: Start determining the source of data to collect the primary data, use the path data.

```
Title: Data store
File Is "C:\Temp\.csv"
Analysis: ReadSize of Data (RSD)
Output: Data store
```

Step 2: Convert Raw data to the data set. Prepare the raw data (primary data) which determine the hold location data and size of data.

Step 3: Check Capacity storage "Overflow"

```
Read RSD
If RSD $\leq$ available memory size
 Forward to Stack
Else
 Forward to Queue
```

Step 4: Stack, the two operations push and pop running parallel.

```
Push ()
Loop (F.L$_n$)
Loop (Attributes$_m$andValues$_k$)
WriteValues$_k$=Distinct Attributes
End
Store F.L$_n$=Values$_k$ with Distinct Attributes
End
Pop ()
Loop (T.T$_n$)
Loop (Attributes$_m$)
If (Attributes$_m$= = Null or Empty)
Delete ()
Else
Write Attributes via a new line
End If
End
End
```

Step 5: Combined

Match F.l and T.T to produce the v and p, throughout location and content use queue sorting by enqueue and dequeue operations. If out of memory use firstly queue to sorting, then forward to stack.

3.2. Map reduce

Map Reduce is a programming paradigm that allows parallel massive data processing, it has been proven to be an effective tool to process large data sets [15].

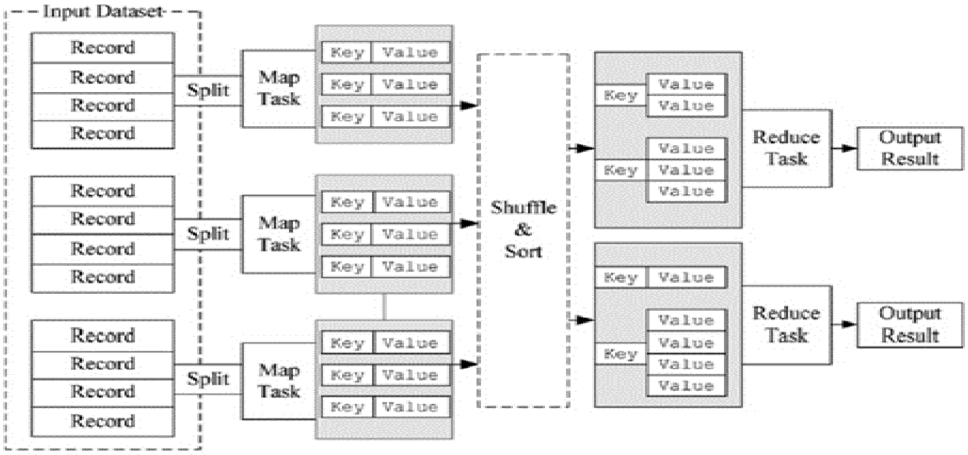

Fig. 7: Architecture MapReduce.

Map function first reads data and transforms records into a key value format. Transformations in this phase may apply any sequence of operations on each record before sending the tuples across the network.

Output keys are then shuffled and grouped by key value so that coincident keys are grouped together to form a list of values. Keys are then partitioned and sent to the Reducers according to some key based scheme previously defined. Finally, the Reducers perform fusion on the lists to eventually generate a single value for each pair. As a further optimization, the reducer is also used as a combiner on the map outputs. This improvement reduces the total amount of data sent across the network by combining each word generated in the Map phase into a single pair.

4. Experiment

The experiments used multivariate Diabetic Retinopathy Debrecen dataset, it contains feature extracted from the Messidor image set to predict whether an image contains signs of diabetic retinopathy or not, the attributes are 50 attributes and the total record is 101766 [16].

The standards theoretical analysis to quantify the amount of data based on precision and recall [17, 18], the purpose of using these terms to scale the volume and veracity of Big data [19].

To compare between Tall array and Map Reduce in term of precision and recall we use two measures; the first measure is F-Measure which is tested using the Formula 1:

$$F_{\beta} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \quad \text{Where} \, 0 \leq \beta \leq \infty \tag{1}$$

And β is a factor that controls the balance between P (Precision) and R (Recall)

When $\beta = 1$ F-Measure is equivalents to the harmonic mean

When $\beta = 0$ F-Measure becomes more precision oriented

When $\beta > 1$ F-Measure becomes more Precision oriented

In this paper, we use F-Measure as harmonic mean i.e. when $\beta = 1$

$$F_1 = \frac{(1^2+1)PR}{1^2P+R} = \frac{2PR}{P+R}.$$

The second Measure is the Arithmetic mean (A) which given by the Formula 2:

$$A = \frac{1}{2}(P+R) \qquad (2)$$

The diabetic datasets through Map Reduce show result into two indicators are precision and recall. The precision for overall is 0.539 as well as the weight is 0.291. Meanwhile, the recall for the overall round is 1.00 but in weight, is minimized to 0.539. The F-measure for overall datasets rely on precision and recall is 0.701. The F-measure for weight precision and recall is 0.378.

Table 1: Map Reduce with F1-Measure and Arithmetic Mean

	precision	Recall	F1-Measure	Arithmetic mean
Overall	0.539	1	0.701	0.7695
Weight	0.291	0.539	0.378	0.415

Regarding the results for F-measure; the act accuracy ambiguity was among whole precision and recall. F-measure indicators the difference between the F-measure for overall and weight is 0. 323, that means the key generate in Map Reduce randomly which have an influence on accuracy.

Tall Array applied diabetic datasets by precision and recall as shown in Table.2, precision indicator result to the overall dataset is 0.66, as well as weight of it, is 0.436. Furthermore, recall indicator is 1.00 and the weight of it is 0.66. F-measure overall is 0.795 and the weight F-measure 0.525.

Table 2: Tall Array Algorithm with F1-Measure and Arithmetic Mean

	precision	Recall	F1-Measure	Arithmetic mean
Overall	0.66	1	0.795	0.83
Weight	0.436	0.66	0.525	0.548

From Table 1, and Table 2, we can notice that Tall Array is efficient than Map Reduce in term precision since the F1-measure has value 0.795 for all overall and 0.525 for weights. However, Map Reduce achieves 0.701 for overall and weight 0.548.

Using Arithmetic mean we can notice that Tall Array is superior to map-reduce; since it has an arithmetic mean 0.83 for overall and 0.525. On the other hand, we have 0.7695 for overall and 0.415 for weights. Thus, we can conclude that a tall array is better than map reduces in term of precision and recall.

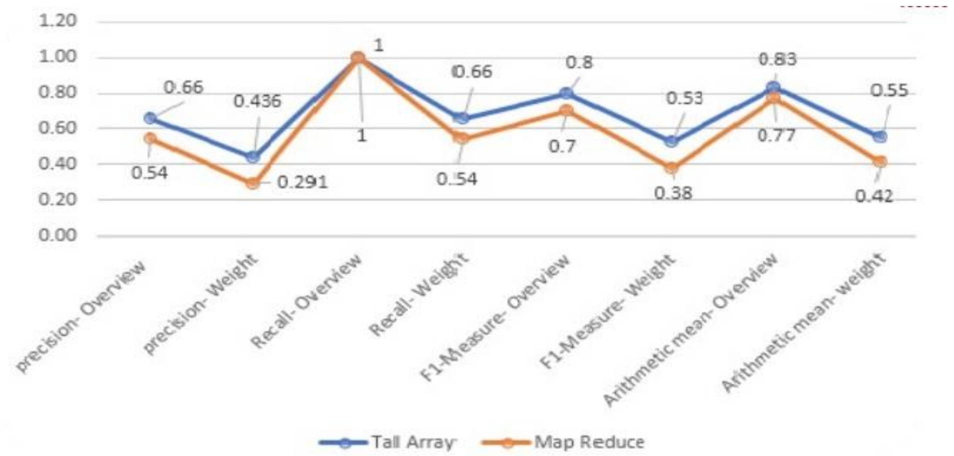

Fig. 8: Comparison between Tall Array and MapReduce.

Veracity characters from big data present the accuracy and correctness as we discussed. Back to experiment F1-measure which expresses an accuracy and correctness used two indicators, precision in overall and weight. Regarding result, the differentiation between Map Reduce and Tall Array is 0.12 and 0.15. On another hand, recall differentiation only in weight difference is 0.12.

Attempt to veracity in big data concept summaries that F1-measure can determine the Tall Array is more efficient than Map Reduce in overall and weight by 0.09 and 0.15.

Volume aspects a major character in big data, in experiment use Arithmetic, mean in overall and weight to detected optimized way treated with it. The results show the different are 0.06 and 0.13 which mean the Tall Array is better than Map Reduce in volume aspect.

5. Conclusion

Map Reduce and Tall Array techniques are analyzed as a solution available for Volume and Veracity using the enabling technologies. Furthermore, have reached the structure of Big data is significant to acquire accuracy and meaningful data. Regarding results the significance of structure data is positive efficiency to Merage BD acritude in two procedures forwards and backward which able to apply data lacked with defragmenting data.

The future work will combine Tall Array with Reduce called Tall Array Reduce to treat BD as a database management system in both statuses internal (in-memory) and external (out-memory).

References

[1] Bamberger, M., Integrating Big Data Into The Monitoring And Evaluation Of Development Programmes, 2016, UN UN Global Pulse, 'Integrating Big Data into the Monitoring and Evaluation of Development Programmes,' 2016. p. 143.

[2] Raj, P., et al., Big and Fast Data Analytics Yearning for High-Performance Computing, in High-Performance Big-Data Analytics2015, Springer. p. 67-99. https://doi.org/10.1007/978-3-319-20744-5_3.

[3] Forum, W.E., Deep Shift Technology Tipping Points and Societal Impact, 2015: World Economic Forum. p. 44.

[4] Hu, H., et al., Toward scalable systems for big data analytics: A technology tutorial. IEEE access, 2014. 2: p. 652-687. https://doi.org/10.1109/ACCESS.2014.2332453.

[5] Russom, P., Big data analytics. TDWI Best Practices Report, Fourth Quarter, 2011: p. 1-35.

[6] Demchenko, Y., C. Ngo, and P. Membrey, Architecture framework and components for the big data ecosystem. Journal of System and Network Engineering, 2013: p. 1-31.

[7] Rahimi-Eichi, H. and M.-Y. Chow. Big-data framework for electric vehicle range estimation. in Industrial Electronics Society, IECON 2014-40th Annual Conference of the IEEE. 2014. IEEE.

[8] Assunção, M.D., et al., Big Data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing, 2015. 79: p. 3-15. https://doi.org/10.1016/j.jpdc.2014.08.003.

[9] Arun, K. and D.L. Jabasheela, Big data: review, classification and analysis survey. International Journal of Innovative Research in Information Security (IJIRIS), 2014. 1(3): p. 17-23.

[10] Zhang, M., et al., SafeDrive: Online Driving Anomaly Detection from Large-Scale Vehicle Data. IEEE Transactions on Industrial Informatics, 2017. https://doi.org/10.1109/TII.2017.2674661.

[11] Zhang, M., et al., CarStream: an industrial system of big data processing for internet-of-vehicles. Proceedings of the VLDB Endowment, 2017. 10(12): p. 1766-1777. https://doi.org/10.14778/3137765.3137781.

[12] Iwamura, K., et al., Big Data Collection and Utilization for Operational Support of Smarter Social Infrastructure. Hitachi Review, 2014. 63(1): p. 18.

[13] Meijer, A.J., J.R. Gil-Garcia, and M.P.R. Bolívar, Smart City Research: Contextual Conditions, Governance Models, and Public Value Assessment. Social Science Computer Review, 2016. 34(6): p. 647-656. https://doi.org/10.1177/0894439315618890.

[14] Morioka, M., et al., City management platform using big data from people and traffic flows. Hitachi Review, 2015. 64(1): p. 53.

[15] Ramírez-Gallego, S., et al., Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. Information Fusion, 2018. 42: p. 51-61. https://doi.org/10.1016/j.inffus.2017.10.001.

[16] Balint Antal, A.H., Diabetic Retinopathy Debrecen Data Set Data Set A. 2014, Editor 2017, the Messidor database:https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set#.

[17] Wang, Y., L. Kung, and T.A. Byrd, Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 2018. 126: p. 3-13. https://doi.org/10.1016/j.techfore.2015.12.019.

[18] Madasamy, K. and M. Ramaswami, Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective. International Journal of Computational Intelligence Research, 2017. 13(9): p. 2267-2281.

[19] Juba, B. and H.S. Le, Precision-Recall versus Accuracy and the Role of Large Data Sets. 2017.