International Journal of Engineering & Technology, 7 (4.37) (2018) 86-89

International Journal of Engineering & Technology

Website: Www.Sciencepubco.Com/Index.Php/IJET

Paper Research

A Survey on Intermediate Data Management for Big Data and Internet of Things

Marwah Nihad[1]*, Alaa Hassan[2], Nadia Ibrahim[3]

1,2,3 Department of Computer Science, College of Science, University of Kirkuk, Kirkuk, Iraq
**Corresponding author E-mail: marwah.nihad@uokirkuk.edu.iq*

Abstract

The field internet of things and Big Data has become a necessity in our everyday lives due to the broadening of its technology and the exponential increase in devices, services, and applications that drive different types of data. This survey shows the study of Internet of Things (IoT), Big Data, data management, and intermediate data. The survey discusses intermediate data on Big Data and Internet of Things (IoT) and how it is managed. Internet of Things (IoT) is an essential concept of a new technology generation. It is a vision that allows the embedded devices or sensors to be interconnected over the Internet. The future Internet of Things (IoT) will be greatly presented by the massive quantity of heterogeneous networked embedded devices that generate intensively "Big data". Referring to the term intermediate data as the information that is provoked as output data along the process. However, this data is temporary and is erased as soon as you run a model or a sample tool. Also, the existence of intermediate data in both of the Internet of Things (IoT) and Big Data are explained. Here, various aspects of the internet of things, Big Data, intermediate data and data management will be reviewed. Moreover, the schemes for managing this data and its framework are discussed.

Keywords: *Big Data, Data Management, Intermediate Data, Internet of Things (IoT).*

1. Introduction

This generation is extremely dependent on Big Data. There is a rapid rise in data types and amount owing to the surfacing of recent utilities such as social networks, cloud computing, and internet of things. Today anything and everything is making use of data, from the simplest objects, to the most massive systems. Today's digital world is in need of management and utilization techniques for this great amount of data. Maturation and transformation are being awaited on database research for Big Data.

In contrast to the data of the classical digital era, Big Data attributes to enormous expanding data sets that consist of heterogeneous formats including structured, unstructured and semi-structured data. Due to its complexity, Big Data, demands robust technologies and state of the art algorithms. Hence, the conventional static Business Intelligence tools are now considered inefficient when it comes to their application in Big Data (Oussous et al., 2017; Wang et al.,2018). Nevertheless, the utilization of the Internet of things (IoT)(Chen et al.,2014; Gubbi et al., 2013) serves as the basis for Big Data applications. There appears to be a constant renewal in the indications of IoT due to the wide range of its objects. Logistic enterprises have recently been backed up by the different Big Data applications. Through the addition of sensors, Global Positioning Systems (GPS) and wireless adapters, tracking vehicle locations has become a reality. By utilizing this data into such uses, companies are now able to perfect the delivery course along with being able to watch and administer its workers. This is all made possible by putting together all the bits of information. Another hot research topic which utilizes IoT data is the smart city. The summarization of data management and intermediate data management from two points of view; Big Data and Internet of Things, will be conveyed in this review paper. The paper will emphasize on the meanings of Big Data, Internet of Things (IoT), intermediate data and data management and discusses the related topics. Analysis regarding the styles to manage data in Big Data and Internet of Things (IoT) will be conducted. Multiple definitions are summarized in Section 2. The modes of managing data in Big Data and Internet of Things (IoT) are broken down in Section 3. Internet of Things (IoT) is examined in section 4. Big Data is considered in section 5. Intermediate Data management is looked at in section 6. Finally, conclusions are given in Section 7.

2. Materials and Methods

In the following sections, some definitions have been considered as very important subjects in our scientific research area.

Big Data: It is data groups that are so extensive and sophisticated that the conventional data-processing application software are incompetent to be good aids at handling them ("Big data . https://en.wikipedia.org/wiki/Big_data,") (see Fig. 1).

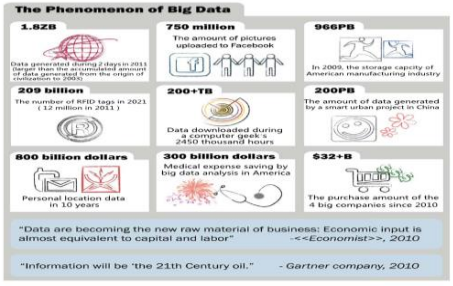

Fig. 1: The Continuously Increasing Big Data (Chen et al., 2014).

The IoT: The term "Things" in IoT is defined as subsystems and individual physical and virtual objects which are uncontrolled, conformable, and accountable. Therefore, IoT is a universal effective network framework. Through reciprocating information provoked by sensing, "things" are known to collaborate between themselves and connect with the surroundings. This in turn will enable them to acknowledge different circumstances and bring about behaviour to physically regulate physical systems (Abu-Elkheir et al., 2013).
Intermediate Data: It refers to the generation and transmission of information along the phases of computing (Moise et al., 2011).
Data Management: Generally, this is an expansive approach comprising the engineering and its methods utilized for appropriate executives for the data collection stages of a specific setup. Through the use of data management, there should be a barrier between the object entities and the main devices providing the information, as well as isolation of the programs examining this data. An arrangement of the devices themselves can be carried out yielding subdivisions with uncontrolled power and an inherent hierarchical administration (Pujolle, 2006). Dependent on the degree of privacy constraints required from the subsystem's holders, the performance and data granted could be adjusted on the IoT network.

2.1. Types of Data Management in Big Data and Internet of Things (Iot):

Management of Big Data: Various obstacles are being faced by data experts, when it comes to handling Big Data. One such obstacle is how to gather, assimilate and store large amounts of data sets hatched from scattered sources. Being able to maintain data is very critical and is the skeleton for Big Data analysis. This is of major importance in aiding the derivation of a dependable comprehension and hence enhancing the expenses. To manage Big Data means to gather information coming from various sources, to encode the information for privacy and precautionary reasons, and to rearrange and/or erase data for accuracy. Adequate storage and an easy entrance to various assigned end points are also provided by Big Data management (Abdullah et al., 2014; Najafabadi et al., 2015; Sahal et al., 2018).
Management of IoT Data: In addition to storage, logging, and auditing facilities for offline studies, a data management system must compile data online. This elaborates on the perception of offline storage, inquiries, and the management of activities from the perspective of online-offline communication/storage dual operations (Abu-Elkheir et al., 2013).

2.2. Internet of Things (Iot)

Through IoT, the inter-communication stretches from being between person to person, from person to object, as well as being between object to object, until a well-balanced environment is achieved. The assimilation of devices such as sensors, Radio Frequency Identification (RFID) tags, and other objects are conducted by the Internet of Things (IoT). Some threats are brought to our mind, due to the divergence of the different apparatus and technologies utilized. This is true specifically when dealing with heterogeneity data comprising numerous traits (Fan et al., 2010).
The IoT Vision: The perception of IoT is to attain a generalized strong base for flourishing numerous helpful and critical applications. This is made possible by utilizing the "Things" and their sub-systems. Data can be collected from a mixture of the information collected from real-time as well as data stored in permanent repositories. Wide research topics, novel inventions and hence state of the art applications can be yielded from this data collecting. The objects that comprise IoT are utilized to form a complete management framework (see Fig. 2) (Abu-Elkheir et al., 2013).

According to the IoT information life process, IoT data management system breaks down into two parts: an online real-time front-end and an offline back-end. The front-end collaborates precisely with the correlated IoT items while the back-end breaks down the IoT data and acts as a storehouse. The front-end involves the generation of inquiries back and forth to and from the sensors and other parts of the smart system. These sensors are hooked up to the system via intermediate data aggregator or concentrator. The best example for this is a smart phone belonging to a patient. As mentioned, the back-end is completely for storage. Here, the information is stored in the memory, to be processed and analyzed in depth at a later time. There are constant interactions between the storage particles which are located on the back-end with the front-end through regular updates and this is known as being online. Within the different stages, the uncontrolled boundaries are expressed as being more communication-intensive than storage-intensive, due to the fact that they grant real-time data to specific requests (Abu-Elkheir et al., 2013; Hassanalieragh et al., 2015).
A Framework for IoT Data Management: A data management framework is multilayered system and is comprised of federated model and data-centricity and it falls under an independent IoT subsystem due to its characteristics such as its ductile, versatile, and coherent nature. The propagation of data takes place in the "things" layer and is generated by all entities and subsystems embedded within the framework. Next, transportation by the communications layer occurs.
Here, raw data or the basic aggregates are transferred to data repositories. Depending on if they belong to common people or institutions, the data repositories are placed on the cloud or specialized servers. Through utilizing query and federation layers, the intended user can be connected to these repositories.
These layers have tasks of interpret inquiries and analysis tasks, determine which repositories will hold the information, and achieve the information by negotiating cooperation. The federation layer takes care of the real-time or context-aware issues. This is made possible by a different layer known as the sources layer which immaculately processes the analysis and obligations of various sources of data. Therefore, both, the questioning in the data is answered and a production of the results is made within the framework. In other words, the current information received will be processed and the solutions to the inquiries will be forwarded to the user, yielding realistic proficiencies for long-term trend discovery and analysis.
This in the long run, boosts the users in the direction of opportunities for discovering a valuable divergence with their data being processed (Abu-Elkheir et al., 2013).
The Aggregation Modules: There is an expansion in the implementation of grouping and outlines of the intermediate data where only the basic value is desirable.
This takes place in the 'things' layer known as the aggregation modules. The arrangement of the aggregation points is carried out in a way so that they are neighbouring the intermediate data sources. This is done for cost effectiveness. Intermediate data from various articles/things is gathered and compiled by the modules themselves. The "things" can be either compatible or associated to a single system. Depending on the requirements and preparations of the related subsystem, the utilization of the aggregation points is determined.
Here, promptness to eliminate any trade-offs in the system's real-time performance is needed (Ray & Koopman, 2009; Roe et al., 2005).
BIG DATA
Big Data generally refers to the information that goes beyond the common storage, computing dimensions and processing of traditional databases and data analysis methods. Big Data calls for devices and procedures that can be utilized to evaluate and cite the arrangements from broad ranged data. As a result of characteristics such as an expanded data storage potential, heightened computational processing aptitude and its ability to enclose enormous amounts of information, Big Data is racing up the charts. To

summarize, Big Data is correlated with complex tasks, specifically being the four V's: Volume, Variety, Velocity, and Veracity (Alpaydin, 2014; Dumbill, 2012; Grobelnik, 2012).

Big Data Analytics Adverse Effects: Just like with any other emerging technology, with these advantages comes some drawbacks. Some adverse points include data attributes and validation, data cleaning, high-dimensionality and data degradation, data reduction, data representations and distributed data sources, data sampling, data visualization, parallel and distributed data processing, real-time analysis and judgment, crowdsourcing and semantic input for improved data analysis (Abdullah et al., 2016). Other points include scalability of algorithms, tracing and analyzing data provenance, data discovery and integration, parallel and distributed computing, exploratory data analysis and interpretation, integrating heterogeneous data, feature engineering and promoting new models for massive data computation (Najafabadi et al., 2015). Nevertheless, the advantages surpass the negative aspects.

2.3. Intermediate Data Management:

The information system enables different information processes to be extended to share the intermediate data prior to their completion without sacrificing data integrity controls.

Intermediate Data Management Challenges: The challenges with intermediate data management are not explored yet and are confined locally. This means that the local file system is entrusted here. Firstly, the information is recorded locally on the node of its origin.

Then it is interpreted remotely by the following node that demands it. Here, the storage systems do not offer any help and all issues are managed within the framework. Distressed assignments are revised and in that way the correct intermediate data will be attempted to be achieved once more (Ko et al.,2009).

Intermediate Data Management in MapReduce Computations: Refinement of the input information is done by MapReduce (Dean & Ghemawat, 2008) applications along with other cloud data flows. The resultant output is eventually displayed as well. This takes place through various steps of calculations. Intermediate data has unique characteristics when compared to meaningful data. Its journey works through its production by each calculation and then being passed on to be handled by the adjacent stage. Intermediate data is short-term information that is recorded a single time by an individual stage and translated a single time by the following stage. This is dissimilar to input and output information which remains constant and are scanned several times. Being in the MapReduce form gives intermediate data the design of the key/value pairs which are provoked by the map stage of the application. Being associated with the same intermediate key, all intermediate codes are gathered into one aggregate and sent to the reduce function. The remainder of this text spotlights the Hadoop project , the resourceful practices of the MapReduce paradigm ("The Hadoop Map/Reduce Framework.http://hadoop.apache.org/mapreduce/
") and finally the methods of management of the intermediate data, within the Hadoop framework, are studied (see Fig. 3) (Moise et al., 2011).

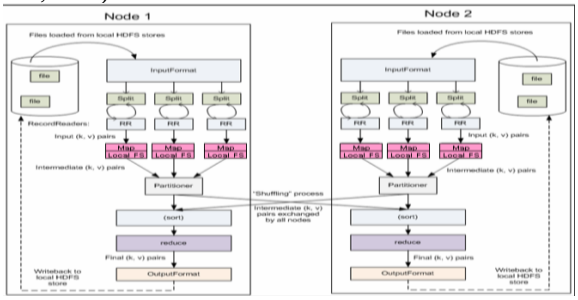

Fig. 3: Hadoop Map/Reduce Data Flow (Moise et al., 2011)

Hadoop Intermediate Data Management: The map function described by the user is performed by an individual function tracker on its designated data portion. Next, sorting by key of the output takes place and is then transformed into an input by designated reducers. This is a crucial stage within the Hadoop core and is known as the shuffle phase. Hence, the shuffle phase is the procedure of by which the information is organized and channelled throughout the mappers until it reaches the reducers. Recording of the outputs to the nearest task tracker's file system running the map function is carried out. Once, a successful achievement of a map function is reached, the task tracker notifies the job tracker. Hence, there is awareness upon the mapping of the map outputs and the corresponding storing nodes.

A partition incorporates value/key combinations lodged on the local disk of various task trackers throughout the bunch. During the reduce phase, each reducer is accredited a partition of keys to operate on.

The processing of the mappers occurs at various times and the reduce task begins duplicating the outputs alongside the entire process. Due to the possibility of the reducer declining, task trackers do not erase map outputs from disk as soon as the reducer has restored them.

Alternatively, after the execution of the task, the job tracker commands the task trackers to abolish them. The distributed file system then receives the output of the reduce phase. (by default, HDFS, which is a Java-based file system that offers scalable and dependent data storage ("HDFS. The Hadoop Distributed File System.
http://hadoop.apache.org/common/docs/r0.20.1/hdfs_design.html
")) (see Fig. 4) (Moise et al., 2011).

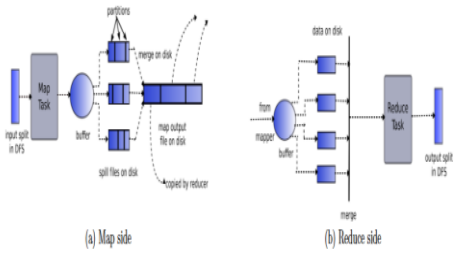

Fig. 4: Intermediate Data in the Original Hadoop MapReduce Framework (Moise et al., 2011)

Intermediate Data Generation through Map and Reduce Functions: The two kinds of intermediate data are intermediate data on the reducer and intermediate data on the mapper. A broad range of values of a certain set of keys partitioned by the key, make up the intermediate data on reducer. On the other hand, the data located on the mapper is recorded from key to value. When there appears a situation with an enormous amount of information (regarding a specific key) to be stored within one file, in this case it has to be stored in multiple files. Contradictory to this, when the reducer is called on a certain key, one must get admittance into all of them (all values of any key)(Wan et al.,2016).

The Framework of MapReduce: The creation of a single group of individually varying developed keys is made possible by The MapReduce framework. This takes place by scanning all lists and grouping all couples with the identical key accordingly. A connection between the diverse nodes is essential in this part of the process. It is highly cost-effective compared to transferring the initial information from place to place. A condensed outline of the information that goes through the node is produced by the map stage. The specific employment of the MapReduce utilized and the particular essence of the distributed data, determines the definite exertion of this part of the process. Within the IoT framework, in the majority of instances, the data may be assigned geographically, given that the information was formed at that particular station. In a small amount of the cases, the data is scattered along a re-

gional cluster of computers (through utilizing a system such as Hadoop). It should be noted that gathering the intermediate results from the numerous Map steps follows a series of stages and is highly dependent on the plot and unique exertion by which the MapReduce framework is handled. A selection of values that range with an equivalent domain (Aggarwal et al., 2013) are manufactured by the Reduce function which is applied in parallel to each group.

3. Conclusions

Big Data and the Internet of Things are the two most-talked-about technology topics of the last few years. The Internet of Things (IoT) denotes spreading the Internet of physical objects, or another human sensing objects. They can be detected, determined, and accessed by devices like sensors, actuators or other smart devices.
As the big increase in existing devices, actuators, sensors and network communications, Big Data has been generated, intermediate data and nodes are used for the spread of data.
The definition and procedures for the operation of data in Big Data and that for the Internet of Things are summarized. Challenges for analytical Big Data are conveyed. How intermediate data is achieved in the map program and how reduce functions work within the MapReduce framework. Management of intermediate data and its drawbacks in the MapReduce and Hadoop programs are reviewed as well. The perception for the internet of things and its framework are discussed in detail elaborating on its ability to regulate and group data. Finally, this survey tries to clarify the existence of intermediate data in each of Internet of Things (IoT) and Big Data and it is important and necessary to highlight in future scientific research in a manner that leads to the efficiency of applications that.

References

[1] Abdullah, M. N., Hassan, A., & Naef, N. (2016). Knowledge-Based Analysis of Web Data Extraction. Paper presented at the The Fifth International Conference on Informatics and Applications (ICIA2016).
[2] Abdullah, M. N., Khafagy, M. H., & Omara, F. A. (2014). Home: Hiveql optimization in multi-session environment. Paper presented at the 5th European Conference of Computer Science (ECCS'14).
[3] Abu-Elkheir, M., Hayajneh, M., & Ali, N. A. (2013). Data management for the internet of things: Design primitives and solution. Sensors, 13(11), 15582-15612.
[4] Aggarwal, C. C., Ashish, N., & Sheth, A. (2013). The internet of things: A survey from the data-centric perspective Managing and mining sensor data (pp. 383-428): Springer.
[5] Alpaydin, E. (2014). Introduction to machine learning: MIT press.
[6] Big data. https://en.wikipedia.org/wiki/Big_data. Retrieved from https://en.wikipedia.org/wiki/Big_data
[7] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.
[8] Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). Big data: related technologies, challenges and future prospects: Springer.
[9] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM,51(1),107-113.
[10] Dumbill, E. (2012). What is big data? An introduction to the big data landscape. oreilly. com, http://radar. oreilly. com/2012/01/what-is-big-data. html.
[11] Grobelnik, M. (2012). Big data tutorial. Kalamaki: Jožef Stefan Institute.
[12] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. Future generation computer systems, 29(7), 1645-1660.
[13] The Hadoop Map/Reduce Framework.http://hadoop. apache. org/ mapreduce/
[14]
[15] Hassanalieragh, M., Page, A., Soyata, T., Sharma, G., Aktas, M., Mateos, G., . . . Andreescu, S. (2015). Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges. Paper presented at the Services Computing (SCC), 2015 IEEE International Conference on.
[16] HDFS. The Hadoop Distributed File System. http://hadoop.apache.org/common/docs/r0.20.1/hdfs_design.html
[17] Ko, S. Y., Hoque, I., Cho, B., & Gupta, I. (2009). On Availability of Intermediate Data in Cloud Computations. Paper presented at the HotOS.
[18] Moise, D., Trieu, T.-T.-L., Bougé, L., & Antoniu, G. (2011). Optimizing intermediate data management in MapReduce computations. Paper presented at the Proceedings of the first international workshop on cloud computing platforms.
[19] Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1.
[20] Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2017). Big Data Technologies: A Survey. Journal of King Saud University-Computer and Information Sciences.
[21] Pujolle, G. (2006). An autonomic-oriented architecture for the internet of things. Paper presented at the Modern Computing, 2006. JVA'06. IEEE John Vincent Atanasoff 2006 International Symposium on.
[22] Ray, J., & Koopman, P. (2009). Data management mechanisms for embedded system gateways. Paper presented at the Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on.
[23] Roe, B., & Beech, R. Intermediate and continuing care–policy and practice.
[24] Sahal, R., Nihad, M., Khafagy, M. H., & Omara, F. A. (2018). iHOME: Index-Based JOIN Query Optimization for Limited Big Data Storage. Journal of Grid Computing, 1-36.
[25] Wan, J., Humar, I., & Zhang, D. (2016). Industrial IoT Technologies and Applications: Springer.
[26] Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change,126,3-13.
[27] Yu, J., & Buyya, R. (2005). A taxonomy of scientific workflow systems for grid computing. ACM Sigmod Record, 34(3), 44-49.