Application of Big Data Analytics via Cloud Computing

Yunus Yetis, Ruthvik Goud Sara, Berat A. Erol, Halid Kaplan, Abdurrahman Akuzum
and Mo Jamshidi Ph.D
The Department of Electrical and Computer Engineering
The University of Texas at San Antonio
San Antonio, TX, USA
yunusyetis68@hotmail.com, ruthvik.goud@gmail.com, berat.erol@utsa.edu, halid88@gmail.com
a.akuzum@gmail.com, moj@wacong.org

***Abstract*— Advances in sensor technology, the Internet of things (IoT), social networking, wireless communications and huge collection of data from years have all contributed to a new field of study Big Data is discussed in this paper. The System of Systems (SoS) integrates independently operating, non-homogeneous systems to achieve a higher goal than the sum of the parts. Recently, management of data has become strenuous and SoS helps in solving the problems and providing solutions, with the new approaches in Data Analytics. Data Analytics uses both statistical and cloud computing using machine learning or computational intelligence to reduce the size of Big Data to a manageable size to extract information, build a knowledge base using the derived data, and eventually develop a nonparametric model for the Big Data. In this research, the approaches towards the cloud environment for Data Analytics is discussed which is one of the key application areas of Big Data. Through this analysis and survey, we provide recommendations for the research community on future directions on providing data-based decisions for cloud-supported Big Data computing and analytic solutions.**

Key words: Cloud Computing, Data Analytics, MapReduce

I. INTRODUCTION

The System of Systems (SoS) is a integrated environment in which independent operating systems work in a cooperative mode to achieve a higher performance. A detailed literature survey of definitions of applications of SoS and many applications can be found in the recent works by Dr. Jamshidi [1], [2]. The application areas of SoS are vast indeed. They are varied in the fields of software systems in the areas of cloud computing, secure systems, medical and health care and also the cyber-physical systems more specifically application areas include energy harvesting, military research, and transport optimization. Data Analytics uses statistical analysis and cloud computing, such as evolutionary computation methods in solving the problems and also has it's own applications in forecasting of SoS. SoS's are generating Big Data which makes modeling of such complex systems a challenge indeed [3]. Big data is the term for huge complicated data that are difficult to process using traditional data processing techniques and management tools. The analysis of data and the identification of the trends are the key considerations to securely store, manage and share large amounts of complex data [17] . On the one hand, the cloud comes with many challenges mainly concerning security, challenging the data ownership and dependency. Hadoop Distributed File System (HDFS) is evolving as a superior software component for cloud computing combined with integrated parts such as Map Reduce [4]. Hadoop, which is an open-source implementation of Google Map Reduce, which includes a distributed file system, provides the application programmer the abstraction of the map and the reduce [5]. MapReduce paradigm is derived from two words map and reduce which can be built in many programming languages [6]. With Hadoop, it is simplified for organizations to easily access and analyze large volumes of data which are generated every day. However, the problems associated with data such as security, data management, monitoring and data dependency continue. MapReduce is helpful in sorting, accessing the web logs, statistical computations, machine learning and also distributed pattern search. MapReduce models adapt to many environments varying from cloud, multi core systems and also mobile environments [7], [8].

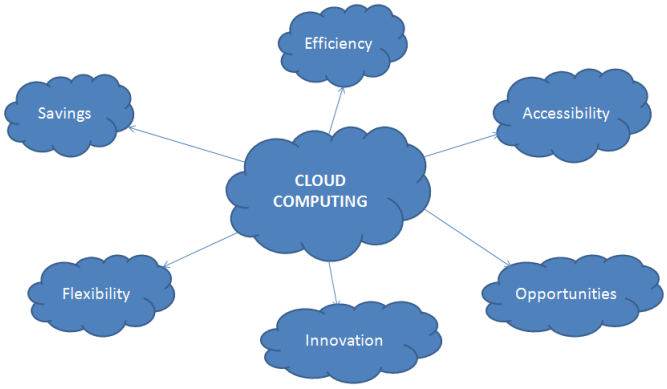

Fig. 1: Cloud Computing Framework

In cloud computing, the word cloud implies The Internet, so cloud computing implies a kind of registering in which administrations are conveyed through the Internet. The objective of cloud computing is to find the solutions efficiently by executing a huge number of guidelines every second. Cloud computing utilizes systems of a large gathering of servers with particular associations with conveying

*This work was supported by Grant number FA8750-15-2-0116 from Air Force Research Laboratory and OSD through a contract from NCA&T State University .

information prepared among the servers. Cloud computing comprises a front end and back end. The front end client's PC and programming required to get to the cloud system. The back end comprises of different PCs, servers and database frameworks that make the cloud. The client can get to applications in the cloud system by interfacing with the cloud utilizing the Internet. Fig. 1 shows that the user can access applications in the cloud network by connecting to the cloud using the Internet which is an example for some of the real time applications [8], [14].

Cloud computing has three principal sorts that are usually alluded to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). There is a completely distinctive "cloud" with regards to business. Most of the business organizations execute Software-as-a-Service (SaaS), where the business accepts to an application or a software which gets to over the Internet.

II. BIG DATA

Big data is the term for so extensive and complicated data sets that it gets to be hard to process using conventional data management tools and processing techniques. The data and the information are used to recognize the trends and patterned structures it is critical to safely store, manage and share a lot of complex data. The cloud is accompanied by an explicit security challenge, i.e. the data owner won't have any control of where the data is placed. Apaches Hadoop Distributed File System (HDFS) is developing as a prevalent programming segment for cloud computing joined alongside incorporating parts, for example, Map Reduce. Hadoop, which is an open-source usage of Google MapReduce, including a distributed file system, gives to the application developer the reflection of the map and the reduce.There are many outcomes with the Big Data Analytics like cost reduction, hadoop and cloud computing has a significant cost benefits. It is also helping faster and better decision making as the future depends mostly on data driven decisions and also a lot of space in improvement either with new applications or services.

Big data refers to exponentially growing structured or unstructured data. The production of big data is created by businesses, the Internet, society, and cyber-physical systems and deriving intelligence from the data and providing the solutions [15]. Another possible definition of big data refers to those data sets that are complex and large which makes it difficult to process with available management tools or traditional paradigms. One of the most promising paradigms to manage big data has been Data Analytics [9]. Big Data is distinguished by increasing volumes of data of different types varying from structure, semi structured and the unstructured, with the increasing in sources it is very essential to plan and pre process the data before performing any kind of analysis on the data. Unstructured data is growing at a very high rate , which can be useful in predictions and forecasting. At the same time academia and industry has focused mainly on the research areas of Big Data Analytics. There are many open source providers with variety of services for different applications are available in support of big data [10], [11].

Data analytics refers to the analysis through inspection, cleaning, transformation, models and verification working towards the creation of conclusions and decision making on the true meaning of the data [12].

Hadoop is an open-source programming structure for processing and storing big data information in an appropriated style on vast groups of item equipment . Basically, it achieves two assignments: enormous information stockpiling and speedier handling. Open-source software: Open source programming varies from business programming because of the expansive and open system of designers that make and deal with the projects [16]. Customarily, it's allowed to download, utilize and add to, however, more commercial and research adaptation of Hadoop are getting to be accessible [13].

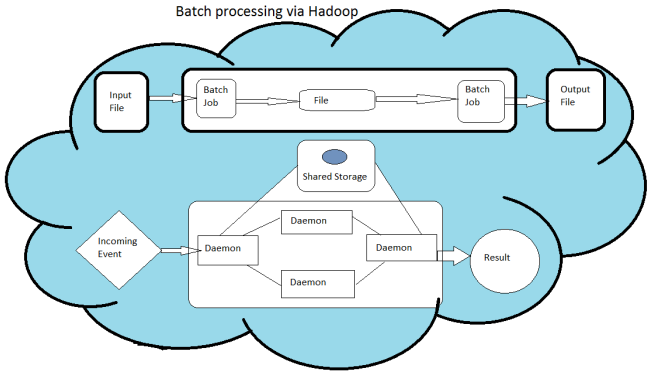

Fig. 2: Structure of Hadoop

- Framework:In this case, it implies all that you have to create and run your product applications is given programs, device sets, associations, and so. [Fig.2]

The Hadoop brand contains a wide range of tools. Two of them are center parts of Hadoop;

- Hadoop Distributed File System (HDFS) is a virtual document framework that resembles some other record framework with the exception of than when you move a record on HDFS, this document is part into numerous little documents, each of those records is reproduced and put away on (ordinarily, might be altered) three servers for adaptation to internal failure requirements.
- Hadoop MapReduce is an approach to part every solicitation into littler solicitations which are sent to numerous little servers, permitting a really adaptable utilization of CPU power.

III. DATA ANALYTICS AND IMPLEMENTATION

For this part, we try to explain the relationships of the theoretical and implementation parts of Cloud Computing. They can be explained as following;

- Understand what defines Cloud Computing and be able to explain the nature and make up of typical cloud scenarios
- Understand the usage of MPI programming with Python

- Understand NoSQL database structure and theory, Map/Reduce algorithm, and implementations such as Hadoop

This approach would be reflecting several skills, such as programming, implementing a program for a particular problem, data gathering, project management, and some other workflows.

On the other hand, establishing the study was challenging while performing the feasibility studies. We have faced with these constraints. We were able to manage our plan and started to handle these obstacles in time. Furthermore, we were gathering the data from the City of Austin (https://data.austintexas.gov) that includes several important data sets, such as water quality samplings, restaurant sampling records, APD crime summaries, etc. For this experiment, we picked the historical crime data that entered by the officials. These data sets include several data fields and different attributes as can be seen on Table-I. Moreover, for building the best approach we were back through 2008 to 2011, and added the most recent entries to make a precise comparison as described year-to-date for 2014.

TABLE I: DATA FIELDS

Fields	**Attributes**
Incident Report Number	Report Number
Crime Type	100 Different Types
Date	mm/dd/yyyy
Time	24 Hour
Location Type	Blank
Address	Reported Address
Longitude	Received Data
Latitude	Received Data
Location 1	Blank

For each year, fields and related attributes that received based on the report generated every day are shown in Table-I. Since these entries are collected every hour and every day, filtering or working on a particular attribute can be hard. On the other hand, the amount of the data entered in a specific time interval or geographical location is not bounded. That means on a particular day and time, there will be a different kind of entries that tagged by different report number.

Therefore, this part requires an implementation of discussing topics that includes a development in Python programming language. Moreover, we were using the data sets gathered from the citys database that provided for a year for addressing the most crime-centered locations of the city. Based on the received results, we tried to conclude a work that shows the most occurred, crime type, the address for the most crime traffic, and the total of the crime incidents that happened in the city by using Map-Reduce approach.

IV. ANALYSIS

The design case was loaded into ***.csv*** file to see differences among those received data from the city. Each file named as the year which was to be analyzed. Therefore, we had five ***Year.csv*** files that include the data for different fields from the Table-I. It is obvious that for making the project outcomes reliable and crucial, we have to come up with the reasonable conclusions from these data.

Our approach was the following; pointing out the total number of the crime in the entire city during the year, gathering the most happened crime type in the city, matching the specific crime type with the local data, relating them to the address attributes; then, concluding with a match with a crime type and the address. These preliminary design steps are finalized by different concepts for performing reliably and fast data storing and visualization, which focusses our priorities for this project.

In the implementation part of this paper, a Map Reduce algorithm was developed to sort all gathered data based on the years. Then, the program was designed in a way that will be fulfilling the priorities from crime based approach. The following samples can be considered as a sneak peak for both mapping and reduce part of the paper. After entering the data into the database by the officials, the server gathers the data and updates the database and it is open to the public. These data sets have been utilized in building the project. Therefore, the Map/Reduce algorithm is executed to filter the data that attract the attention and focus on the crime analysis, the next intent then is storing those data to visualize the results.

Therefore, not only the gaps between the data and complexity to match different crime types and location that occurred among others, can be reduced but also by modeling these results to build a better design to prevent those crimes for public safety.

After running our algorithm, we received following results for each particular attributes as seen in Table-II. It is easy to see that we sort the data results based on the years and their priorities. First column shows the maximum number of the crime along with its type of the crime in the second column that reported in the system. Then, third column shows the location where the most crime traffic occurs for the past year with number of crimes occurred in the location. Finally, last two columns reveal that most common crime type in that particular address along with its occurrence.

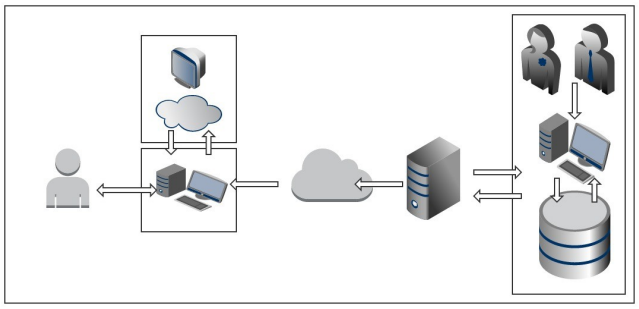

Fig. 3: Flow Model view of the system based on the sequential approach. Data access made available by law enforcements to public, a database has created based on gathered data.

TABLE II: DATA RESULTS

Year	# of most occurred crime	Type of the crime	Location for the most crime traffic	# of the crime	Most common crime	# of occurrence
2008	14789	Theft	3600 Bl Presidential Blv	648	Salvage Insp.	109
2009	16990	Bulg. Of Vehc.	700 Bl E 8th St	772	Reg. of Sex,Offend.	135
2010	14437	Bulg. Of Vehc.	700 Bl E 8th St	775	Reg. of Sex,Offend.	110
2011	12903	Bulg. Of Vehc	700 Bl E 8th St	960	Reg. of Sex,Offend.	230
2014	10499	Theft	410 Bl Guadalupe,St	1071	Salvage Insp.	123

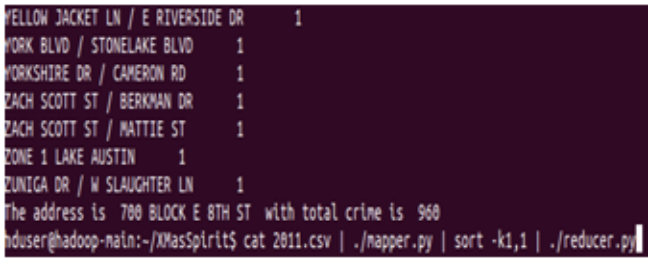

Fig. 4: The address with highest number of incident is 700 BLOCK E 8TH ST with total 960 crime. In this address, most happened crime is REG. SEX OFFENDER INFORMATION with 230 times.

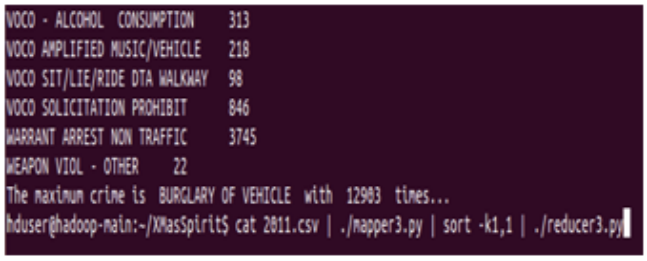

Fig. 5: The type of crime incident that happaned in maximum is BURGLARY OF VEHICLE with the call of 12903 number of times in entire city.

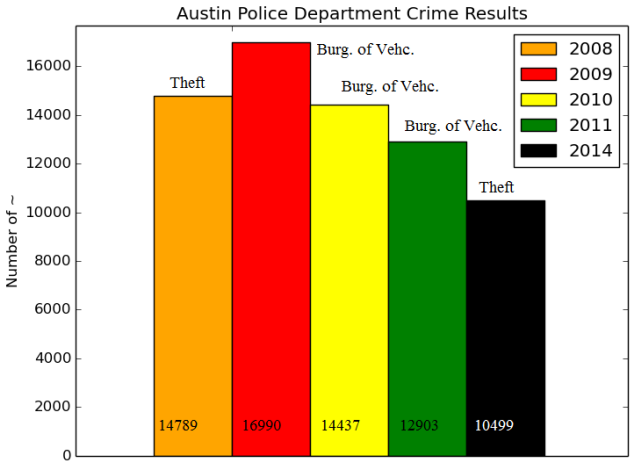

Fig. 6: Number of the Most Occurred Crime and its Type

V. CONCLUSIONS

Big Data Analytics is widely growing area for the analysis and forecasting of trends and also to derive decisions based

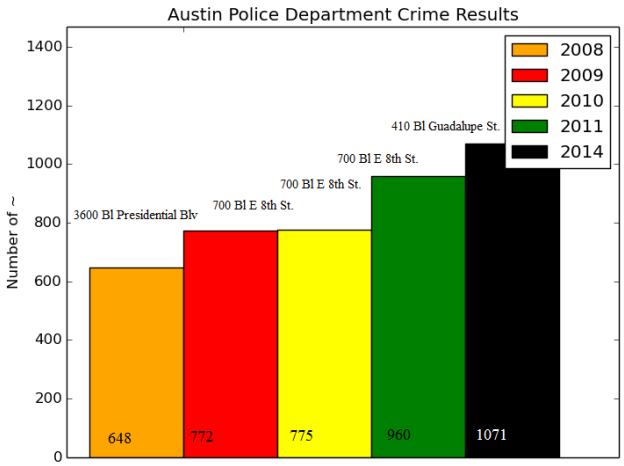

Fig. 7: Number of the crime at the location with highest crime rate

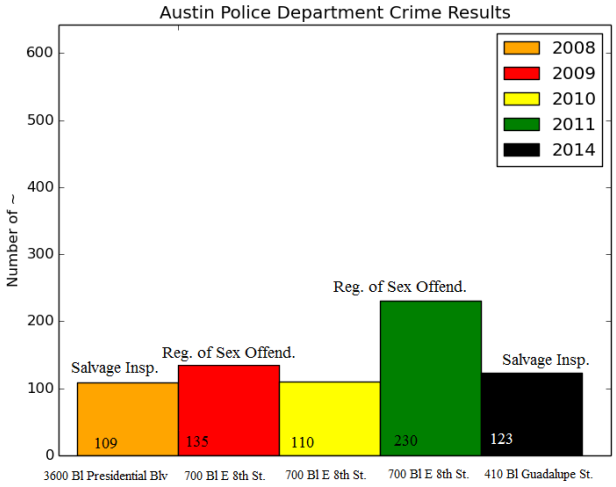

Fig. 8: Crime types with the highest rates

on the data which is used in both industry and research aspects; In this paper an application of Data Analytic has been discussed where the use of Data helps in identifying the trends of the crimes occurred, as security is an important aspect for organizations. Using proposed approaches and big data tools to analyze the massive amount of threat data received daily, and correlating the different components of an attack, allows a security vendor to continuously update their global threat intelligence and equates to improved threat

knowledge and insight.Through Big Data Analytics fraud can be identified the moment it happens and appropriate measures can be taken to constrain the harm. Customers are benefited through improved, faster, and broader threat protection by Data Intelligence.

REFERENCES

[1] M. Jamshidi (ed.), Systems of Systems EngineeringPrinciples and Applications (CRC/Taylor & Francis, London, 2008) (also in Mandarin language, China Machine Press, ISBN 978-7- 111-38955-2, Beijing, 2013)

[2] Jamshidi, M., Tannahill, B., Ezell, M., Yetis, Y., and Kaplan, H. (2016). Applications of Big Data Analytics Tools for Data Management. In Big Data Optimization: Recent Developments and Challenges (pp. 177-199). Springer International Publishing

[3] Jamshidi, Mo, Barney Tannahill, Yunus Yetis, and Halid Kaplan. "Big Data Analytic via Soft Computing Paradigms." In Frontiers of Higher Order Fuzzy Sets, pp. 229-258. Springer New York, 2015.

[4] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices. Noida: 2013, pp. 404 409, 8-10 Aug. 2013.

[5] White and Tom,Hadoop: The Definitive Guide,2009 ,1st edition by O'Reilly Media, Inc.

[6] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," Commun ACM, 51(1), pp. 107-113, 2008

[7] S. Sakr, A. Liu and A. Fayoumi, "The family of mapreduce and largescale data processing systems," ACM Computing Surveys, 46(1), pp.1-44, 2013.

[8] Y, Amanatullah, Ipung H.P., Juliandri A, and Lim C. "Toward cloud computing reference architecture: Cloud service management perspective. Jakarta: 2013, pp. 1-4, 13-14 Jun. 2013.

[9] Gczy, P., Izumi, N., & Hasida, K. (2012). Cloudsourcing: Managing cloud adoption. Global Journal of Business Research, 6(2), 57-70.

[10] Ghazal, Ahmad and Rabl, Tilmann and Hu, Minqing and Raab, Francois and Poess, Meikel and Crolotte, Alain and Jacobsen, Hans-Arno. "BigBench: Towards an Industry Standard Benchmark for Big Data Analytics" Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data.

[11] S.Ramamoorthy and S.Rajalakshmi "Optimized Data Analysis in Cloud using BigData Analytics Techniques" 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013.

[12] Tannahill, B. K., Maute, C. E., Yetis, Y., Ezell, M. N., Jaimes, A., Rosas, R., . & Jamshidi, M. (2013, June). Modeling of system of systems via data analyticsCase for Big Data in SoS. In System of Systems Engineering (SoSE), 2013 8th International Conference on (pp. 177-183). IEEE.

[13] A. Navaz, G. Velusam and D. Gurkan, "Experiments on Networking of Hadoop," 2014 IEEE 22nd International Conference on Network Protocols, Raleigh, NC, 2014, pp. 544-547.

[14] A. Sahba, R. Sahba, and W.-M. Lin, "Improving IPC in Simultaneous Multi-Threading (SMT) Processors by Capping IQ Utilization According to Dispatched Memory Instructions," presented at the 2014 World Automation Congress, Waikoloa Village, HI, 2014.

[15] A. Sahba, Y. Zhang, M. Hays and W.-M. Lin, "A Real-Time Per-Thread IQ-Capping Technique for Simultaneous MultiThreading (SMT) Processors", In the Proceedings of the 11th International Conference on Information Technology New Generation (lTNG 2014), April 2014.

[16] Ghanat Bari, M., Ramirez, N., Wang, Z., and Zhang, J. M. (2015). MZDASoft: a software architecture that enables largescale comparison of protein expression levels over multiple samples based on liquid chromatography/tandem mass spectrometry. Rapid Communications in Mass Spectrometry, 29(19), 1841-1848.

[17] Bari, Mehrab Ghanat, Xuepo Ma, and Jianqiu Zhang. "PeakLink: a new peptide peak linking method in LC-MS/MS using wavelet and SVM."Bioinformatics (2014): btu299.

[18] Salekin, S., Bari, M. G., Raphael, I., Forsthuber, T. G., and Zhang, J. M. (2016, February). Early disease correlated protein detection using early response index (ERI). In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 569-572). IEEE.

[19] M. Bagheri, M. Madani, R. Sahba, and A. Sahba, "Real time object detection using a novel adaptive color thresholding method", International ACM workshop on Ubiquitous meta user interfaces (UbiMUI'11), Scottsdale, AZ, November 2011.