Application of Big Data Analytics via Cloud Computing

Yunus Yetis[1], Ruthvik G. Sara[2], B. Erol[3], H. Kaplan[4], A. Akuzum[5] and Mo Jamshidi, Ph.D[6]
The Department of Electrical and Computer Engineering
The University of Texas at San Antonio
San Antonio, TX, USA
yunusyetis68@hotmail.com[1], ruthvik.goud@gmail.com[2], moj@wacong.org[6]

***Abstract*— On this paper, large sets of data have been accumulating in all aspects of our lives for a long time. Advances in sensor technology, the Internet, social networks, wireless communication, and inexpensive memory have all contributed to an explosion of Big Data. System of Systems (SoS) integrate independently operating, non homogeneous systems to achieve a higher goal than the sum of the parts. Todays SoS are also contributing to the existence of unmanageable Big Data. Recent efforts have developed a promising approach, called data analytic, which uses statistical and cloud computing to reduce using 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. This research discusses approaches and environments for carrying out analytics on Clouds for Big Data applications. It revolves most important areas of analytics and Big Data. Through a detailed survey, we provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.**

Key words: Cloud Computing, Data Analytics, MapReduce

I. INTRODUCTION

System of Systems (SoS) are integrated, independently operating systems working in a cooperative mode to achieve a higher performance. A detailed literature survey on definitions to applications of SoS and many applications can be found in recent texts by Jamshidi [1], [2]. The application areas of SoS are vast indeed. They are software systems like the Internet, cloud computing, health care, and cyber physical systems all the way to such hardware-dominated cases like military, energy, transportation, etc. Data analytic and its statistical and cloud computing such as evolutionary computations have their own applications in forecasting of SoS. SoSs are generating Big Data which makes modeling of such complex systems a challenge indeed [3]. Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. The data and to identify patterns it is very important to securely store, manage and share large amounts of complex data. On one hand, cloud comes with an explicit security challenge, i.e. the data owner may not have any control of where the data is placed. Hadoop distributed file system (HDFS) is evolving as a superior software component for cloud computing combined along with integrated parts such as Map Reduce [4]. Hadoop, which is an open-source implementation of Google Map Reduce, including a distributed file system, provides to the application programmer the abstraction of the map and the reduce. With Hadoop it is easier for organizations to get a grip on the large volumes of data being generated each day, but at the same time can also create problems related to security, data access, monitoring, high availability and business continuity. Recent progress on classic big data networking technologies, e.g., Hadoop and Map Reduce, big data technologies in cloud computing, big data benchmarking projects, and mobile big data networking [5], [6].

*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 .

II. WHAT IS CLOUD COMPUTING ?

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 make utilization of expanding figuring energy to execute a huge number of guidelines every second. Cloud computing utilizes systems of a huge gathering of servers with particular associations with convey information preparing among the servers [7]. Cloud computing comprises of a front end and back end. The front end client's PC and programming required to get to the cloud system. 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.Cloud computing has three principle 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. A few organizations execute Software-as-a-Service (SaaS), where the business subscribes to an application it gets to over the Internet. Fig. 1 is shown that the user can access applications in the cloud network by connecting to the cloud using the Internet some of the real time applications [8].

III. BIG DATA

Big data is the term for so extensive data sets and complicated that it gets to be hard to process using conventional data management tools or processing applications. The data and the information is to recognize patterns it is critical to safely store, manage and share a lot of complex data. Cloud accompanies an explicit security challenge, i.e. the data owner won't not have any control of where the data is

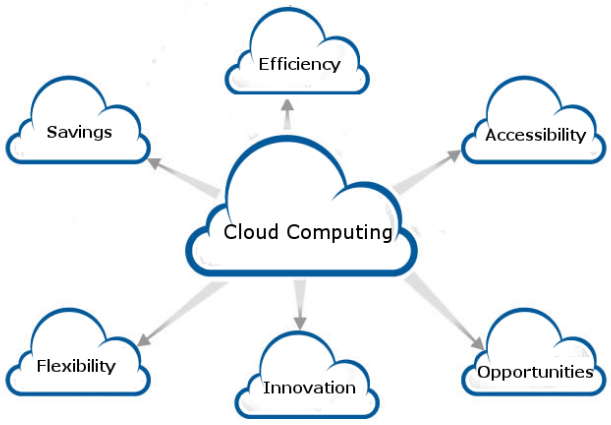

Fig. 1. Cloud Computing Framework

placed. Apaches Hadoop distributed file system (HDFS) is developing as a prevalent programming segment for cloud computing joined alongside incorporated 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. With Hadoop it is simpler for organizations to take a few to get back some composure on the huge volumes of information being produced every day, except in the meantime can likewise make issues identified with security, data access, observing, high accessibility and business continuity. Recent progress on classic big data networking technologies, e.g., Hadoop and MapReduce, big data technologies in cloud computing, and mobile big data networking.

IV. BIG DATA OPTIMIZATION

Big data refers to exponentially growing structured or unstructured data. Production of big data is created by businesses, the Internet, society and cyber physical systems. Another possible definition of big data refers to those data sets that are complex and large which makes it difficult to process available management tools or traditional paradigms. One of the most promising paradigms to manage big data has been data analytic [9]. Data analytic refers to the analysis through inspection, cleaning, transformation, models and verification working towards creation of conclusions and decision making on the true meaning of the data. The showing Fig.2 depicts the principles of data analytic.

Fig. 2. Data Analytic Tools for Big Data Management

V. HADOOP

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. Customarily, it's allowed to download, utilize and add to, however more business adaptations of Hadoop are getting to be accessible.

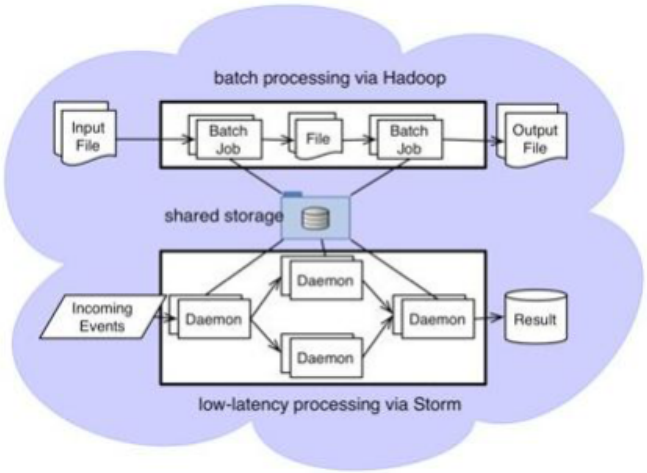

Fig. 3. Structure of Hadoop

- 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.3]

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.

VI. 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 how to use 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 sheets, 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. This data sheets includes 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. The data files can be obtained from the following links.

- https://data.austintexas.gov/api/views/r6sg-xka2/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/ei2n-fehk/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/4c6h-tv2y/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/gr59-ids7/rows.csv?accessType=DOWNLOAD
- https://data.austintexas.gov/api/views/b4y9-5x39/rows.csv?accessType=DOWNLOAD

For each year fields and related attributes that received based on report numbers in every day 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 in a particular day and time, there will be different kind of entries that tagged by different report number.

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

Therefore, this part requires an implementation of discussed topics that includes a development in Python programming language. Moreover, we were using the data sheets 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.

VII. 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 those interests for it. Therefore, we had five ***Year.csv*** file that includes 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 those data.

Our approach was in this way; pointing out the total number of the crime in 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 with 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 reliable and fast data storing and visualization, which are points our priorities for this project.

VIII. IMPLEMENTATION

For implementation part of this paper, we tried to develop a Map/Reduce algorithm to sort all gathered data based on the years. Then, we designed the program in a way that will be fulfilling our priorities from crime based approach. The following samples can be considered as a sneak peak for both mapping and reducing part of the paper. After entering the data to the database by the officials, the server gathers the data and updates the database that it offers to the public freely. Then, we are simply requesting the data sheets for each year from the internet and downloading them into our directory in the cluster. Therefore, we run the Map/Reduce algorithm to filter the data that attracts our attention. Our 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 those 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 can be 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

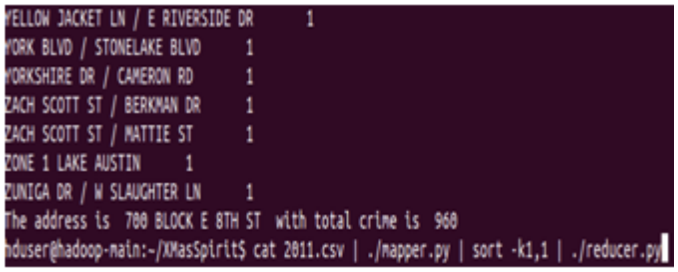

Fig. 4. The address is 700 BLOCK E 8TH ST with total crime is 960

common crime type in that particular address along with its occurrence.

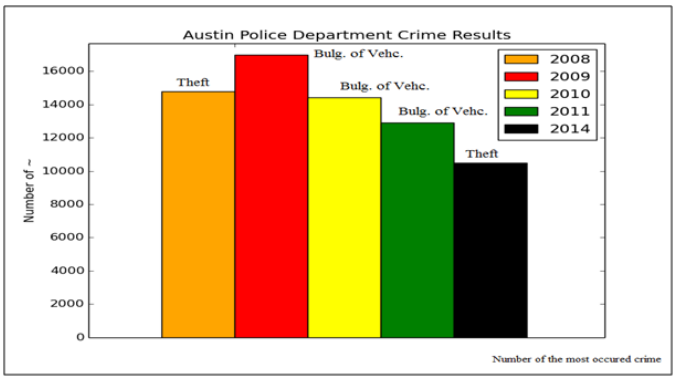

Fig. 7. Number of the Most Occurred Crime and its Type

TABLE II
DATA RESULTS

	# 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. 5. In This address, most happened crime is REG. SEX OFFENDER INFORMATION with 230 times.

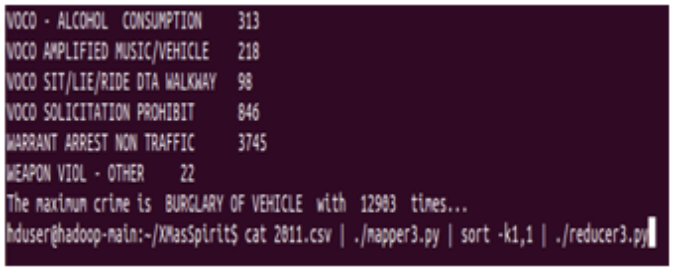

Fig. 6. The maximum crime is BURGLARY OF VEHICLE with 12903 times.

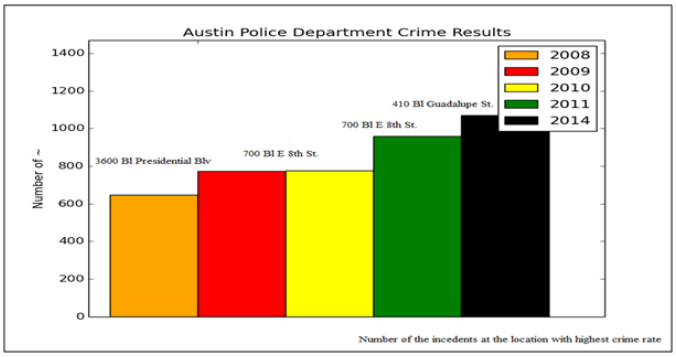

Fig. 8. Number of the crime at the location with highest crime rate

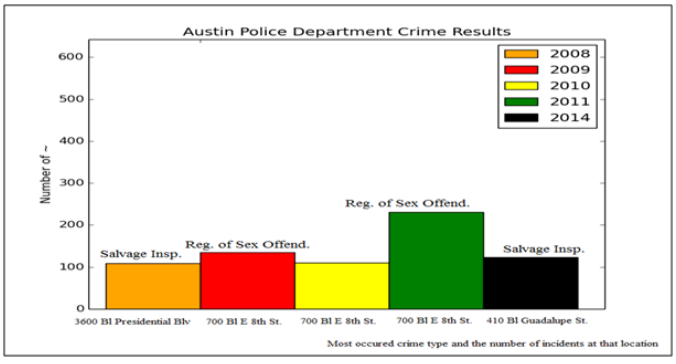

Fig. 9. Crime types with the highest rates

IX. CONCLUSIONS

Cloud environment is widely used in industry and research aspects; therefore security is an important aspect for organizations running on these cloud environments. Using proposed approaches, cloud environments can be secured for complex business operations. Using 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 benefit through improved, faster, and broader threat protection.

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] M. Jamshidi (ed.), System of Systems EngineeringInnovations for the 21st Century (Wiley, NewYork, 2009)

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

[6] 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.

[7] 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.

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

[9] 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.