*International Journal of Latest Trends in Engineering and Technology*
*Vol.(8)Issue(4-1), pp.223-227*
*DOI: http://dx.doi.org/10.21172/1.841.38*
*e-ISSN:2278-621X*

# CHALLENGES AND OPPORTUNITIES OF BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY

Khushboo (Research Scholar )[1], Dr. V. K. Joshi(Professor & Head of Department)[2]

**Abstract- To describe the promise and potential of big data analytics in healthcare. The paper describes the nascent field of big data analytics in healthcare, discusses the benefits, outlines an architectural framework and methodology, describes examples reported in the literature, briefly discusses the challenges, and offers conclusions. This paper provides a broad overview of big data analytics for healthcare researchers and practitioners. Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.**
**Keywords—Big Databases, Healthcare Analytics, Encryption, Masking, De-Identification.**

## 1. INTRODUCTION

Big Data is an emerging trend which is characterized by any kind of data source that has large volumes, high velocity and wide variety of data. It enables the organizations to gather, store, manage and analyze vast amounts of data at the right speed, at the right time, to gain the right insights from complex, noisy, heterogeneous, longitudinal data. Today it has become a significable challenge and a necessity for numerous application domains. Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data which scales to terabytes, petabytes and ex bytes in search of valuable information and insights. Big data initiatives have the potential to transform healthcare, as they have revolutionized other industries. Healthcare organizations use big data analytics to transform data into actionable information by producing data-driven insights to smarter business and clinical decisions like reducing readmissions, cut hospital-contracted conditions, identifying and eliminating waste, improved clinician workflow etc.

Big Data platform must embrace multiple layers of security for data at rest and the data in flight. It should be able to encrypt data at rest within the data repository at least for personal and sensitive data. All communications between data sources, data consumers and the Big Data warehouse should be encrypted to provide security to the data.

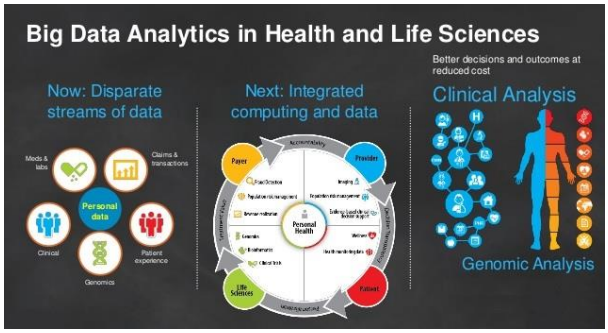

Fig 1: Big Data Healthcare Analytics

Big data analytics in healthcare Health data volume is expected to grow dramatically in the years ahead. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits. What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as "large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information". Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalyzed) patient-related

[1] *Department of Computer Science and Engineering, Ludhiana College of Engineering & Technology, Katani Kalan, Ludhiana, Punjab, India*
[2] *Department of Computer Science and Engineering, Ludhiana College of Engineering & Technology, Katani Kalan, Ludhiana, Punjab, India*

health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient. Advantages to healthcare by digitizing, combining and effectively using big data, healthcare organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant benefits. Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics. Certain developments or outcomes may be predicted and/or estimated based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely will not benefit from surgery; complications; patients at risk for medical complications; patients at risk for sepsis, MRSA, C. difficile, or other hospital-acquired illness; illness/disease progression; patients at risk for advancement in disease states; causal factors of illness/disease progression; and possible comorbid conditions (EMC Consulting).

***Big data can improve the safety level of traffic***

The real-time processing capabilities of big data can accurately probe traffic accidents, its predictive ability can effectively predict the occurrence of traffic incident, using microwave detection systems, video surveillance systems, mobile detection system, and they can build an effective security model to improve the safety of vehicles. When security incidents happened, and emergency rescue needed, Because of its comprehensive processing and decision-making capability, rapid response capability, big data can greatly improve the ability of emergency rescue, and reduce casualties and property losses.

## 2. LITERATURE REVIEW

**Jie Xu, et.al [2015]** proposed a novel online framework that could learn from the current traffic situation (or context) in real-time and predict the future traffic by matching the current situation to the most effective prediction model trained using historical data. As real-time traffic arrives, the traffic context space is adaptively partitioned in order to efficiently estimate the effectiveness of each base predictor in different situations. They obtained and proved both short-term and long term performance guarantees (bounds) for thier online algorithm. The proposed algorithm also works effectively in scenarios where the true labels (i.e. realized traffic) are missing or become available with delay. Using the proposed framework, the context dimension that is the most relevant to traffic prediction can also be revealed, which can further reduce the implementation complexity as well as inform traffic policy making. Their experiments with real-world data in real life conditions show that the proposed approach significantly outperforms existing solutions [1].

**Muhammad Tayyab et.al (2012)** proposed unsupervised learning methods, such as k-means clustering, Principal Component Analysis (PCA), and Self Organizing Maps (SOM) to mine spatial and temporal performance trends at both network level and for individual links. They performed prediction for a large, interconnected road network, for multiple prediction horizons, with SVR based algorithm. They showed the effectiveness of the proposed performance analysis methods by applying them to the prediction data of SVR [2].

**Yufei Han et.al (2014)** proposed a new methodology for extracting spatio-temporal traffic patterns, ultimately for modeling large-scale traffic dynamics, and long-term traffic forecasting. They attacked this issue by utilizing Locality-Preserving Non-negative Matrix Factorization (LPNMF) to derive low-dimensional representation of network-level traffic states. Clustering is performed on the compact LPNMF projections to unveil typical spatial patterns and temporal dynamics of network-level traffic states. They have tested the proposed method on simulated traffic data generated for a large scale road network, and reported experimental results validate the ability of their approach for extracting meaningful large-scale space-time traffic patterns. Furthermore, the derived clustering results provide an intuitive understanding of spatial-temporal characteristics of traffic flows in the large-scale network, and a basis for potential long-term forecasting [3].

**Nitesh V. Chawla et.al (2013)** presented the foundations of work that takes a Big Data driven approach towards personalized healthcare, and demonstrate its applicability to patient-centered outcomes, meaningful use, and reducing re-admission rates [12].

**J. Archenaa et.al (2015)** proposed an insight of how they can uncover additional value from the data generated by healthcare and government. Large amount of heterogeneous data is generated by these agencies. But without proper data analytics methods these data became useless. Big Data Analytics using Hadoop plays an effective role in performing meaningful real-time analysis on the huge volume of data and able to predict the emergency situations before it happens. It describes about the big data use cases in healthcare and government [13].

**Uma Srinivasan et.al (2013)** described two novel applications that leverage big data to detect fraud, abuse, waste, and errors in health insurance claims, thus reducing recurrent losses and facilitating enhanced patient care. The results indicate that claim

anomalies detected using these applications help private health insurance funds recover hidden cost overruns that aren't detectable using transaction processing systems. This article is part of a special issue on leveraging big data and business analytics [10].

**Joachim Roski et.al (2014)** explained the potential of big data to create significant value in health care by improving outcomes while lowering costs. Big data's defining features include the ability to handle massive data volume and variety at high velocity. New, flexible, and easily expandable information technology (IT) infrastructure, including so-called data lakes and cloud data storage and management solutions, make big-data analytics possible. However, most health IT systems still relies on data warehouse structures. Without the right IT infrastructure, analytic tools, visualization approaches, work flows, and interfaces, the insights provided by big data are likely to be limited. Big data's success in creating value in the health care sector may require changes in current polices to balance the potential societal benefits of big-data approaches and the protection of patients' confidentiality. Other policy implications of using big data are that many current practices and policies related to data use, access, sharing, privacy, and stewardship need to be revised [11].

## 3. ARCHITECTURE OF INTELLIGENT TRANSPORTATION ON BIG DATA PLATFORM

Intelligent transportation system on big data platform is a combination of multiple systems, models, department and technology. It can be said, it is a comprehensive system of system science, management science, mathematics, economics, behavioral science, and information technology. From the architecture, the platform includes basic business layer, data analysis layer and information publishing layer. As shown in Figure 2.

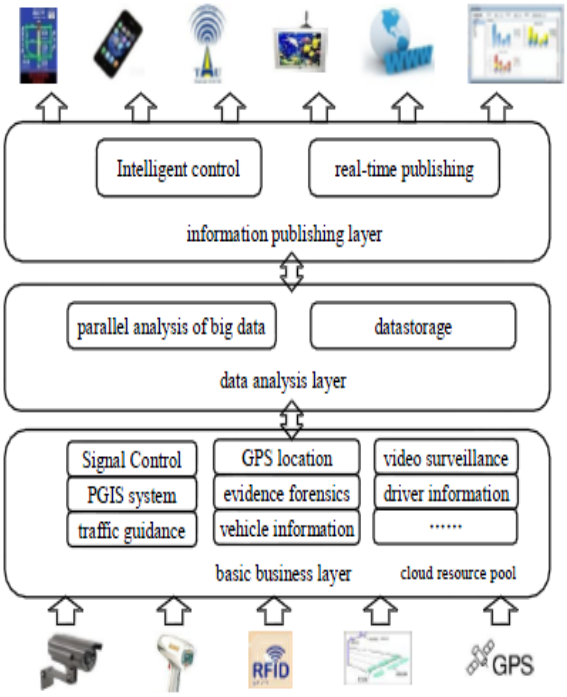


Fig 2: Architecture of Intelligent Transportation on Big Data Platform

The basic business layer is the foundation of data analysis layer and information publishing layer, its main function is to complete the basic work of the various business units, and to produce basic business data. It includes traffic information collection system, signal control systems, video surveillance systems, illegal evidence forensics system, 122 alarm receive and dispose system, GPS vehicle location tracking system, traffic guidance system, vehicle information management system, driver information management system, PGIS system, and so on. the service of basic business layer is the basis for the work of the various business units, its data comes from data acquisition system mentioned above, storage and handling of data is very important. Therefore, cloud computing technology can be used on the basic business layer, decentralized system can be integrated into the cloud, this will ensure the security and stability of the application system, and provide an efficient computing environment.

According to the information of the road network, the demand of public travel and comprehensive analysis of data, data analysis layer uses big data technology, data mining technology, combines with a variety of mathematical models for real-time effective analysis. It can grasp the condition of the transportation system in any time, such as road congestion degree, average speed, saturation, occupancy rate, interrupt rate. It can make further congestion warning, traffic guidance and other intelligent

transportation behavior. Data analysis layer is built on Hadoop ecosystem, use commercial cheap server as hardware platform, use the open-source Linux as operating system, use HDFS as file system for big data storage, use Map Reduce as a parallel computing model, use H Base as the database for processing the data, use Hive as data warehouse, use Sqoop and Flume as tools for data integration.

The information publishing layer according the result of the data analysis layer, publishes traffic conditions to public, business units, industry executives, etc. by internet, mobile terminal, and desktop application, report, for their travel and business decisions. It is necessary for friendly interface, operating easily, rich feature. The information published includes traffic condition, traffic warning, data charts for decision. With the development of the times, publishing channels become diversified, changed from traffic radio and information bulletin board to today's traffic radio, mobile TV, micro blog, Wechat, information bulletin board and other forms and channels.

## 4. THE KEY TECHNOLOGY ON DATA ANALYSIS LAYER

The difference between intelligent transportation systems(ITS) and the traditional traffic control system lies in its intelligent features, ITS can carry out intelligent control based on traffic condition. Hadoop ecosystem has a natural advantage in dealing with traffic big data. ITS has a variety of data sources, complex data types, the huge amount of data, the traditional relational database is incompetence for big data. The bayonet in the city is equipped with video surveillance system, and it collects traffic HD video data and vehicle information, Including vehicle pictures, vehicle type, vehicle color, passing time, speed, bayonet number, lane number, direction of travel and so on. The data generated by the GPS location tracking system includes the plate number, time, location coordinates; Internet of things produced the similar data. According to these basic business data, ITS can report traffic conditions, for example: traffic flow in the bayonet, average speed, degree of congestion.

***Calculation of bayonet traffic flow:***

ITS can calculate the traffic flow of every bayonet in a certain time interval, such as 5 minutes, 10 minutes, 15 minutes, or other period of time, and push the calculated data to data publishing layer, report to the traveler, policy makers, business supervisor. Statistical analyzes were performed using Hadoop Map Reduce parallel model, which is the most efficient way. The data got from H Base database including bayonet ID, direction ID, passing time. The key in map() function is bayonet ID and direction ID, the value in map() function is passing time.

The output <key-value> pair of Map() function is <key, one>, the key include bayonet ID, direction ID and passing time, the value is one.

Reduce () function can calculate the sum of one direction of traffic flow, between the start time and end time in a bayonet; the output <key-value> pair is <bayonetID_ directionID_passingtime, count>.

***Calculation of average speed of a road :***

The average speed of a road is an important indicator of the efficiency of road traffic, in general, the higher speed of traffic, the higher the traffic efficiency. The average speed is not the speed measured by a radar at a place and a time point. Because it can only represent a point, but can't represent the whole road.

We look the time measured the same car passing the adjacent bayonets as the spent time, distance between the adjacent bayonets as distance that the vehicle has traveled. The average speed may be represented by the following formula:

$$\bar{v} = \frac{n \times s}{\sum_{i=1}^{n}(t_{end} - t_{start})}$$

where, s is the distance between adjacent bayonets, tend is the time the vehicle run out the road segment, tstart is the time the vehicle run into the road segment, tend and tstart must be the time that the same vehicle run into and run out the road segment. The vehicle leaving or entering in the middle section of the road does not included.

In Map Reduce model, to simplify programming, we use two Map Reduce process to calculate the average speed. At the first round of calculation, map() function get the information that the vehicle pass the bayonet from HBase, including plate_num, bayonetID, directionID, passing time. Reduce() function can calculate the time the vehicle has spent through the road.

At the second round of calculation, in the map() function, the key is bayonetID and directionID , the value come from the map() function's output at the first round. The reduce() function count the sum of the spent time and calculate the average speed.

***Querying the travel path of a vehicle :***

Querying the travel path of a vehicle has an important role in the public security investigation work at a specific period of time. This work requires a lot of manpower, to search the surveillance video day and night, to look for suspicious vehicles manually, then the travel path of the vehicle is drawn manually. Now, ITS can resolve this problem efficiently, the bayonet can identify and record the plat number of the passing vehicle, save it into H Base, index on plat number and passing time, when querying the travel path, enter the start time and end time, Then an ordered data set is returned, now we can draw the travel path very fast according it, it can reach the second level.

***Checking and controlling the fake vehicles :***

The fake vehicles we call it clone vehicle, its plate number, type, color , even credentials are the same as the true vehicle, and its harmfulness is obvious. The police carried out its strict management and control to identify the fake vehicle, fully relying on personal experience before, the police can touch the plat, enquire the driver, query the information of the vehicle and driver.

Now, we can use big data technology to identify the fake vehicle. Its principle is shown as: ITS can query the information, and calculate the time difference between different bayonets, if the time difference is less than 5 minutes, even 2 minutes, and the vehicle maybe is a fake vehicle, because the vehicle can't travel the distance between the two different bayonets within the certain time. In Map Reduce model, the key of reduce() function is plat_num, its value is bayonetID+","+ pastime, then query the plat number that the time difference of it is less than 5 minutes between two different bayonets. Because opportunities that the vehicle with same plate number are limited, the function of map() and reduce() can calculate efficiently.

## 5. CONCLUSIONS

Healthcare is moving towards big data, with patient information residing in multiple locations that must be accessed rapidly. Big data holds much promise for healthcare. Healthcare Analytics focuses on tasks such as giving physicians more information at the point of care, reducing hospital readmissions and better treating chronic diseases. Healthcare data is also extremely sensitive, with confidentiality and integrity a key attribute. So, in healthcare, big data security is vital. Also for the best care to be provided by a doctor there must be fast access to a patient's medical history, securely and as fast as possible.

Security solutions should ensure safeguarding analytics and securing Big Data Frameworks. Designing the right technical foundation in place is a prerequisite for successful data analysis.

## 6. REFERENCES


[1] Jie Xu, Dingxiong Deng, Ugur Demiryurek, Cyrus Shahabi, Mihaela van der Schaar ," Mining the Situation: Spatiotemporal Traffic Prediction with Big Data", 1932-4553 (c) 2015 IEEE.

[2] Muhammad Tayyab Asif, Justin Dauwels, "Spatial and Temporal Patterns in Large-Scale TrafficSpeed Prediction", This work has been presented in part at the IEEE Intelligent Transportation Systems Conference, Hilton, Anchorage, AK, USA, 2012.

[3] Yufei HAN, Fabien Moutarde, "Statistical Traffic State Analysis in Large-scale Transportation Networks Using Locality- Preserving Non-negative Matrix Factorization", CAOR, MINES-ParisTech, 60 Boulevard Saint-Michel, 75006, Paris

[4] ZongtaoDuan, Ying Li, Xibin Zheng, Yan Liu, Jiting Dai, Jun Kang.Traffic Information Computing Platform for Big Data.AIP Conference Proceedings.2014,1618(1):464-467.

[5] JemalAbawajy. Comprehensive analysis of big data variety landscape.International Journal of Parallel, Emergent and Distributed Systems.2015,30(1):5-14.

[6] Ana L.C. Bazzan, FranziskaKlügl. Introduction to Intelligent Systems in Traffic and Transportation.Synthesis Lectures on Artificial Intelligence and Machine Learning. 2013,7(3).

[7] EmadFelemban, Adil A. Sheikh. A Review on Mobile and Sensor Networks Innovations in Intelligent Transportation Systems.Journal of Transportation Technologies.2014,4(3):196-204.

[8] Wei Shi, Jian Wu, Shaolin Zhou, Ling Zhang. Variable message sign and dynamic regional traffic guidance.Intelligent Transportation Systems Magazine, IEEE. 2009,1(3):15-21.

[9] EPJ Data Science. Personalized routing for multitudes in smart cities.EPJ Data Science.2015,4(1).

[10] Srinivasan, U. and Arunasalam, B., 2013. Leveraging big data analytics to reduce healthcare costs. IT Professional, 15(6), pp.21-28.

[11] Roski, J., Bo-Linn, G.W. and Andrews, T.A., 2014. Creating value in health care through big data: opportunities and policy implications. Health Affairs, 33(7), pp.1115-1122.

[12] Chawla, N.V. and Davis, D.A., 2013. Bringing big data to personalized healthcare: a patient-centered framework. Journal of general internal medicine, 28(3), pp.660-665.

[13] Archenaa, J. and Anita, E.M., 2015. A survey of big data analytics in healthcare and government. Procedia Computer Science, 50, pp.408-413.