Big Data and Transportation

Syam Sundar Herle
Indiana University
711 N Park Ave
Bloomington, Indiana 47408
syampara@iu.edu

ABSTRACT

With the rise of population in cities, commutation by road or rail have become hard for people. Transit plays major role in public and private day to day life, but there are limited number of system to address the transportation issues. On other hand Big Data have proved to be more effective and helpful in most of the sectors and business. Big data is all about realizing full potential of large data set by acquiring, storing and managing by advanced technologies and optimizing techniques. With the help of new era of technologies like internet, social media, traffic camera,feeds and smart-phones one can have access to more real-time data which contains abundant of information, which can be used in transportation sector. We can use advance Big Data technologies like Spark on those real-time data set to address transportation issues to build next level Intelligent Transportation Systems (ITS).

KEYWORDS

i523, HID219, Big Data, Intelligent Transportation Systems

1 INTRODUCTION

In modern era, technologies plays a major role in our day to day life changing the way of our interaction and livelihood. Smart-phones have become unsaid finger for our hand and internet have pushed communication to next level. The revolution made by technologies and internet have resulted in collection of large real-time data. Real-time data coming from transport camera's, mobile devices, social media's and other sources need to be dealt to provide easy decision making system for commuter in order to set a effective transportation system. According to [4] "the volume and speed at which data are generated, processed and stored is unprecedented". Said that, Big data is all about realizing full potential and revealing hidden patterns of large data set by acquiring, storing and managing by advanced technologies and optimizing techniques.

The overwhelming real-time data collected from mobile devices and social media like Facebook, Twitter while travelling contains not only the location based information of the users but also the nature of traffic in hidden manner. These real-time data collected from social media applications and traffic camera can be used to predict the nature of road ahead, optimized route, traffic forecasting and predict them before they occur for the public by Intelligent Transportation system(ITS)[5]. The nature of the data are in diverse manner ranging from structured to unstructured, as these are in real-time the consumed data size would be in terabytes. Because of the huge amount of big data and its nature there is a urgent need of addressing the issues like storing, managing and analyzing which are beyond the capacity of the traditional tools which are used for analysis.

2 WHAT IS BIG DATA

As defined earlier, Big data is all about realizing full potential and revealing hidden patterns of large data set by acquiring, storing and managing by advanced technologies and optimizing techniques. According to [6] Big Data has also been defined as by the four V's:

2.1 Volume

The mass of amount of the data, which indicates to the data in low level. Lot's of data collected from social media platform are variety and huge amount in nature, which consists of structured and unstructured. Looking up the unstructured data will be futile and we cannot mine any useful information from the raw unstructured data. Advance technologies like Hadoop, Map Reduce [1, 8] needs to employed to manage storage and mine the unstructured data.

2.2 Velocity

This can be defined by the rate of pace at which the data are accumulated. In some real-time applications like Internet of Things (IoT) related to health care systems and transportation systems there is a need to evaluate real-time data. Instead of writing high streams of data we need to store it directly to memory to evaluate them.

2.3 Variety

Usually variety can be defined as mixture, when coming to data there will be structured, semi-structured and unstructured. Data like text, audio and video will be unstructured in nature, we need to process to make the unstructured data meaningful and we need to tag the data in same format of the structured data and apply the same techniques which are applied on the structured data. In the case of real time data some of the structured data may change to unstructured or semi-structured without notice.

2.4 Value

All data has hidden value, which must be revealed by employing range of technique's depending on the nature of value of the data. The technique's employed usually vary from the quantitative and investigative to discover the real value of the data ranging from expression of the consumers to financial status of the consumers.

Big Data analytic techniques are applied on large data to generate knowledge and reveal hidden patterns which can be harnessed to make useful in customer point of view. Now a days, Big Data analytic are used in many areas like DNA analysis, machine learning, deep learning, robotics and data classification. Usually the Big Data process can be visualized as in the Figure 1.

The Big Data process can be broken down to two major component, Data Management and Data Analytic. In Data management segment the large set of data are acquired and recorder as first step,

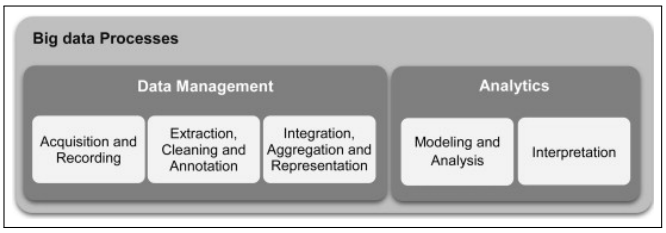

Figure 1: Big Data Process [6].

later the data are extracted from the storage or recorded devices and cleansed and annotated to meta file in the case of the unstructured data like video, text and audio. After cleaning and annotation the data are aggregated and presented in order to extract useful information. The unstructured data aggregated and integrated in this step to represent it as a structured data. In the next step Big Data Analytic are employed on the aggregated data, the first step of the analytic is the creation of model to apply on the aggregated data to perform analysis to reveal hidden value and pattern. In the next step visualization of the hidden pattern in the data is done and then the pattern are interpreted to make useful decision making to make a effective system for the consumer needs.

3 OVERVIEW OF BIG DATA ANALYTIC

In the following subsection we will look into the available Big Data analytic for structured and unstructured data.

3.1 Text Analytic

Extracting information from text data accumulated from social media, online questionnaire, survey and online feeds is known as Text Analytic. A number of analysis and model involved in text analytic like machine learning algorithm, natural language processing and linguistics. The following are the mining techniques are applied in the section of text analytic:

3.1.1 Information Extraction. These techniques extract structured meaning from the unstructured data. These techniques work in two sub task which are Entity Recognition (ER) [6] and Relation Extraction (RE) [6].

3.1.2 Text Summarizing. As the name suggests these techniques create key points of single or multiple document in a succinct way. This analytic technique are employed in two ways, "Extractive Approach"[6] and "Abstractive approach" [6]. In the "Extractive Approach" summary is created from original statements and these summaries are subset of the original document by determining the salient points and putting them together. While the "Abstractive Approach" does the extraction of the semantic information from the text by employing Natural Language Processing techniques.

3.1.3 Questionnaire. These techniques provides answers to the questions, some of the edge of day applications are Apple Siri and Google Home [6]. Machine learning algorithm like deep learning and speech recognition and artificial intelligence are employed in the Questionnaire technique.

3.1.4 Opinion Mining. Opinion Mining are sentiment analysis, where Natural Language Processing and machine learning are employed to mine and classify emotion of the consumer in related to product or governance of a government. Usually text generated from online review portal or social media application like Twitter, Facebook are used to do sentiment analysis.

3.2 Audio Analytic

Audio Analytic are employed on the structured and unstructured audio data to extract information, which are used to evaluate consumer needs and evaluating turn over of a company product and to evaluate agent performances. Most of the call centers use the Interactive Voice Response (IVR) [7] platforms to identify and handle frustrated customers. In health-care sector, audio analytic are employed to evaluate patients with certain medical condition like depression, schizophrenia [6]. Audio Analytic are done by two ways large-vocabulary continuous speech recognition (LVCSR) [6] and phonetic based approach [6].

3.3 Video Analytic

Video Analytic employs a variety of process like monitoring, analyze and extract useful information from video data. Growth of CCTV usage and video sharing portal usage have increased the need for effective and efficient application and technologies to store and clean and analyze the video data. Text Analytic and Audio Analytic can be used to index the video data for the easy retrieval. Extracted information from the video data can be useful in relating consumer needs based on demographic locations and to derive decisions of the product endorsement in retail sector. The system architecture for video analytic can be done in two ways, Server-based architecture and Edge based architecture.

3.4 Social Media Analytic

Analysis of the structured and unstructured data from social media is known as social media analytic. Social media can be categorized as social network (Twitter, Facebook), micro-blog (Reddit) and video sharing (YouTube). The two sources of social media analytic are content shared by users (videos,text and audio) and communication between entity and networks (people and organization). Social Media analytic can be categorized as:

3.4.1 Content based analytic [6]. The content based analytic are the analyze done on the content posted by people or participant of the social media like Facebook, Twitter. These content based data are large set of unstructured data such as videos, text and audio files. The Text Analytic, Audio Analytic and Video Analytic are usually employed on these type of data-sets to mine useful information.

3.4.2 Structure based analytic [6]. Structure based analytic usually related to the extraction of relation among the entities or participant of the social media. Social network are modelled through edge and nodes to represent the entities and the relationship among the entities.

3.5 Predictive Analytic

Prediction of future outcome using data is known as Predictive Analytic. The data for these type of analytic can be categorized as,

real-time and historical data. Primarily predictive analytic is based on statistical models applied on data to predict the outcome of future or predict the relationship among the data. When predicting the outcome of future techniques such as moving average are employed on the historical data. On the real-time data techniques like linear regression are applied to predict the relationship among the data. Predictive analytic are applied in two ways, statistical model applied on sample population of the data which is further applied to full scale data. The other way is usage of computational techniques applied on full scale big data to do the predictive analytic.

4 TRANSPORTATION

Transportation is a means of transporting goods and people between different location, it is a vital element of modern society. Since the earliest days of the industrial revolution, transportation has facilitated economic development by moving materials, resource, products and people [3]. Transportation around the world have some common problems, which causes disruption of not only road traffic but also it has effects on the economic and as well as ecosystem of country.

Common problems in transportation,

- Road traffic
- Environmental Pollution
- Accidents
- Inability of forecasting traffic
- Road Maintenance
- Road Construction/Maintenance

Creation of a transportation system need to address the above problems, when implementing a transportation, government needs to address to avoid the future traffic congestion in case of road maintenance of road construction. Road accidents costs more than life cost, like insurance, rehabilitation cost, property damage cost and so on. Pollution impact also need to be addressed when designing and implementing a effective transport system, if CO2 emission is unchecked it may lead to health related problem and global warming.

Apart from mentioned problem's, creation of a transport system also needs to address some to the issues related to city planning also, the problems can be categorized in the following subsections,

(1) City Plan
    The plan of the city is very crucial, specifically in cities as the traffic congestion as to be addressed.
(2) Police and Law Enforcement
    Consideration of traffic law is also very important, speed limits are also important to taken into consideration.
(3) Event Gathering
    During specific social event gathering, road block and alternative route have to be taken into consideration.

The above issues highlights specifically the situation of the road congestion and road traffic, so the government or transport system planning body should take necessary considerations in the above stated problems by investing sufficient time and money for installing the surveillance and tools.

5 EXISTING TRANSPORTATION SYSTEM USING BIG DATA ANALYTIC

Different countries have employed different big data technologies for their transportation systems project to ease the traffic and make an efficient and social impact transportation. Some of the projects are as follows:

5.1 Public Transit System, Ireland

The country of Ireland came up with an Public Transit System project [2]. IBM big data analytic helped them to come up with an intelligent public transportation system and ease the traffic congestion. The data sources of this were GPS data, speed data, Stop data and fare data collected from Bus System of Ireland and real-time video stream data from CCTV footage and Ireland road weather condition and road work data as an CSV formatted files. real-time road sensor data were also used to build Public Transit System. Advanced analytic concepts were applied on the data to identify traffic congested areas and alternative in the case of road maintenance.

5.2 Real-time vehicle monitoring system, India

The government of India came up with Real-Time vehicle monitoring program [2] using advanced analytic tools to ease operational complexion in logistic and transportation. The data collected from vehicle sensors and GPS devices are used to study about the vehicles fuel status, speed, acceleration and location. All the unstructured data are stored in HDFS system and monitored on consistent time interval to improve logistic productivity.

5.3 Air travel customer analysis system, United Kingdom

Air travel customer analysis system [2] was developed by United Kingdom to understand the needs of the clients of the aviation industries. The data collected from social media, call center and Smart-phone device were used to do Predictive Analytic and Social Media Analytic to identify the customer needs and problems to improve the standard of service provided in the Aviation industry of United Kingdom.

5.4 Real-time Intelligent Transportation system, Sweden

Sweden created a project based on IBM InfoSphere [2] to improve the transportation network in their city of Stockholm. For this project the system collected data from taxi vehicle GPS along with link related to location information to create a Real-Intelligent Transportation system to predict future traffic condition and shortest-route to destination for public and law enforcement agencies

6 CONCLUSION

We came across what is big data and type of big data analytic employed based on different nature of data accumulated from different data sources and some of the existing transportation system employed in different countries using Big Data Analytic. Even-though there are good transportation system employed using Big Data

Analytic, with the current advancement in integration and penetration of machine learning, artificial intelligence and Big Data Analytic tools like Apache Hadoop and Spark, an edge of the day Intelligent Transportation System can be developed to use in areas like congestion management, traffic routing and scheduling.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Gregor von Laszewski and his teaching assistants for providing helpful feedback.

REFERENCES

[1] IBM Analytic. 2017. Map Reduce. (2017). https://www.ibm.com/analytics/us/en/technology/hadoop/mapreduce/

[2] A. Ben Ayed, M. Ben Halima, and A. M. Alimi. 2015. Big data analytics for logistics and transportation. In *2015 4th International Conference on Advanced Logistics and Transport (ICALT)*. 311–316. https://doi.org/10.1109/ICAdLT.2015.7136630

[3] R. P. Biuk-Aghai, W. T. Kou, and S. Fong. 2016. Big data analytics for transportation: Problems and prospects for its application in China. In *2016 IEEE Region 10 Symposium (TENSYMP)*. 173–178. https://doi.org/10.1109/TENCONSpring.2016.7519399

[4] Big Data, Transport: Understanding, and assessing options. 2015. OECD/ITF. (2015). https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf Accessed:30-Mar-2017.

[5] United States Department of Transportation. 2017. Intelligent Transportation System. (2017). https://www.its.dot.gov/

[6] Oracle. 2017. Big Data. (2017). https://www.oracle.com/big-data/index.html

[7] Qubole. 2017. Big Data and Customer Service. (2017). https://www.qubole.com/blog/call-center-analytics/

[8] SAS. 2017. Hadoop. (2017). https://www.sas.com/en_us/insights/big-data/hadoop.html