

ANALYSING THE PRIVACY CONCERNS IN BIG DATA INTEGRATION TO DEVELOP A FRAMEWORK FOR ITS EFFECTIVE AND OPTIMISED USAGE IN SAFEGUARDING PRIVACY ISSUES

Aryan Grover

ABSTRACT

Background/Objectives: The sources of big data are social media, enterprise data, unstructured data, and sensor and clickstream data. The objective is to integrate this variety of data at one platform for processing the big data and find privacy concerns. Methods: The privacy concerns are raised due to unauthorized data extraction, collection and sharing information about user. For integrating and processing of big data; different tools and techniques are available. Findings: General framework for privacy preserving is discussed. Advancements in the big data analytics methods have posed different challenges in front of user. Due to large volume and variety of big data many organizations cannot process the data and needs to outsource it. While sharing such data for processing; there is need to apply proper privacy preserving measures. Application/Improvements: Privacy preserving techniques have applications in electronic health record processing, government surveys, outsourcing enterprise data for processing.

1. INTRODUCTION

Due to advancement in microprocessor electronics and availability of high performance communication networks abundant information is available. The data is getting generated in large quantity from number of sources. Data generation is estimated up to 2.5 Exabyte (1 Exabyte = 1,000,000 Terabytes) of data per day¹. Figure 1 shows the exponential growth of the data. The sources for the data can be categorized in internal and external sources broadly. Figure 2 shows different sources of big data. The internal sources are application log, machine generated data, click stream data, sensor data etc. External sources of big data generation are social media, enterprise data such as transactions, emails, contracts. It also includes weather data, sensor generated data for vehicle, traffic, cell phone GPS signals. New York stock exchange generate 1 TB data; twitter generates 10TB data every day. This can be fed to sentiment analysis and based on this it can be discovered what people feel about various products and events. Volume is important to consider for example power meters are generating billions of reading every year and it is necessary to analyse this data to optimize the energy and actually see usage at energy per man.

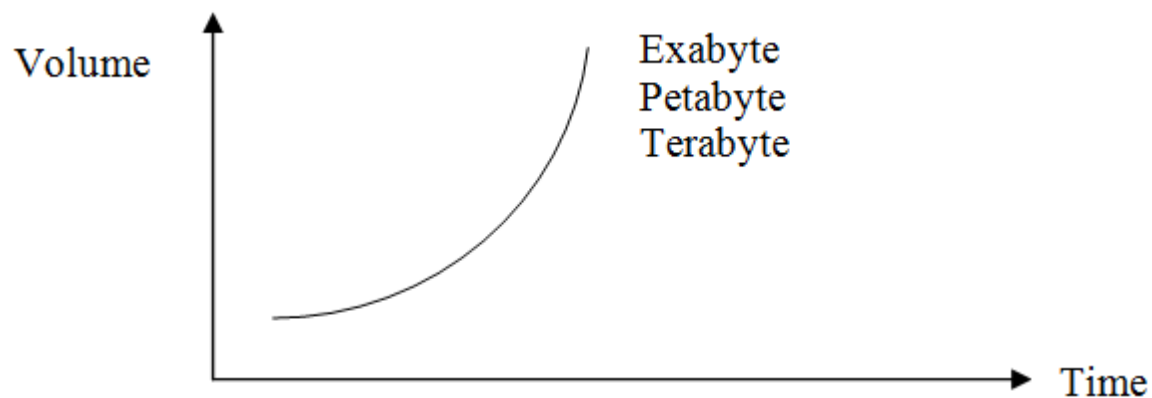


Figure 1. Big data generation at exponential growth.

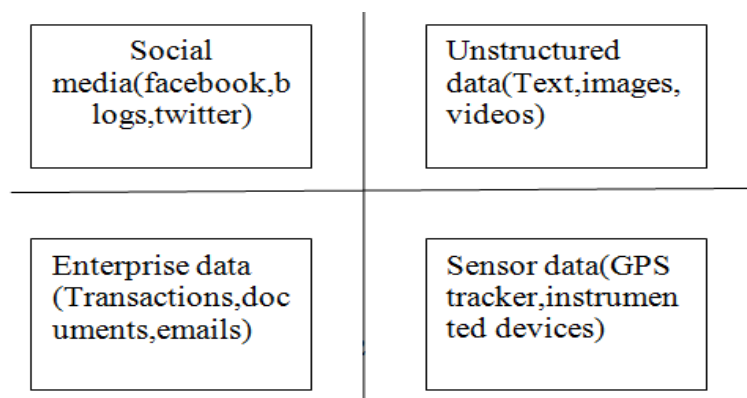


Figure 2. Big data sources.

Velocity is another important characteristic, for various time sensitive activities. For example, to decide fraud, seconds can be decisive in being successful or not. The other aspect of the velocity is that the combining of data which is real time with the default should be possible. In other example a modern car is having 100s sensors and sensors generating large volume of data arriving in very rapid way. 95% of the data is being generated in unstructured or semi structured format. As the population is increasing this uses smart phones. Business generate transaction data, but now users being on the internet generate tremendous amount of data images, videos text and it is need to process all of them. The number of smart phone users is increased to 75% up from 35% in 2011 in United States. This availability of mobile devices made many things come into reality. The large content of information is available. The people get connected with others for communication virtually.

2. BIG DATA DEFINITIONS

Data generating from different sources have different characteristics. In the definitions available in the literature are focusing on the large volume, variety, velocity of the data. The emphasis is on the processing capabilities or infrastructure availabilities available for processing, otherwise which was impossible with traditional framework. The definitions for big data are leveraging on the ability of

business intelligence, competitive intelligence, enhanced insight and decision making. In 2001 definition given by Laney⁴ as: “high-volume, high-velocity and high variety information assets that demand cost-effective, innovative forms of information process for enhanced insight and decision making”. In 2012 the definition is updated in⁵ as “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”. The above definitions are emphasizing on 3V model i.e. high volume, high variety and high velocity. Few organizations added the term value in definition to make it 4V model. Afterward veracity term is added for big data to call it 5V model. In⁶ addition of ambiguity, viscosity, and virility the 3V model is discussed. Lack of metadata causes the ambiguity for example in the large volume of data M and F can be taken for March and February instead of male and female. Viscosity is the measure of resistance. Viscosity for example resistance in data flow, business rules and technology may cause loss of business. Virality measures how fast data can spread. For example, re-tweets on a tweet. The ambiguity, viscosity and virality characteristics are useful from the point of analysis.

From the point of scalability to big data analytics the definition are suggested in⁷ as attributive definition and architectural definition. In attributive definition it says that according to a 2011 report that was sponsored by EMC (the cloud computing leader)⁸: Big data technologies require new platforms to store and process the data and derive the value from large volume and different forms of data. In architectural definition the National Institute of Standards and Technology (NIST)⁹ suggests that, due to limitations of traditional relational approaches, processing of big data in large volume and variety of data which is coming at varying velocity, the need of scalability in the processing is required.

3. BIG DATA PROCESSING

In processing of big data, we have to consider diversity of data. The data is taken to one platform. Based on the internal and external sources steps can be identified as 1) acquisition of data from different sources, 2) processing 3) visualize 4) intelligence (Figure 3)

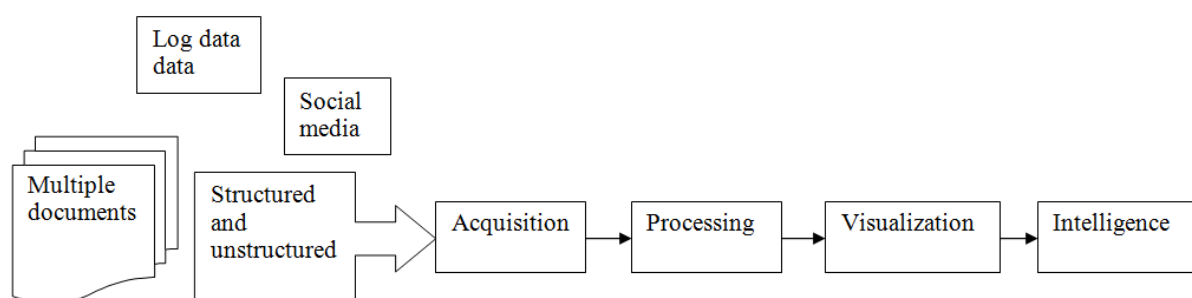


Figure 3. Big data processing flow.

3.1 Acquisition

Acquisition: data from different sources such as socialmedia, application logs, clickstreams, emails, documents,SMS and phone calls are aggregated. Data integration toolscan be helpful to integrate structured and unstructureddata.

- Flume is distributed reliable tool for efficiently collecting and aggregating the log data. It works on streaming data flows¹⁰. It is fault tolerant and reliable and support real-time analytic⁶.
- Sqoop: For acquisition and integration of the data from RDBMS platform to Hadoop platform for processing sqoop connector can be used. Enterprise data like transactions, metadata, data warehouse, data from enterprise system is taken to Hadoop platform and processed in batch^{10,6}.

3.2 Processing

Processing: In data processing data collected in largevolume is processed. Variety of data present two basicprocessing types. First is batch processing to process largevolume of recorded data in the form of file. Second isreal time processing to process large volume of data inthe form of stream. Such data in large volume is stored innode of clusters on Hadoop distributed file system. HDFSis scalable, fault tolerant framework for storing data. Ituses data nodes and name nodes to provide the reliabilityusing replication of data among distributed node incluster.

3.3 Visualization

It helps to get 360-degree view of social issue. Visualizationis useful to draw the inferences and test the hypotheses.JavaScript's and different open source tools are used tovisualize the response of followers in case of social media.Authors in¹⁸ showed emotions of viewers can be expressedon twitter and changes on incident. Joy, sadness andneutral views can be visualized. In case of reality showsto find the impact of show on national and global level.Understand the views of audience and summarize andrepresent in understandable format.

3.4 Intelligence

Enterprise top management can take smart decisionfrom the visualization and patterns come out of big dataanalysis. For customer sentiment analysis can be helpfulfor marketing and product development. Email analysisis useful to target key customers and their perceptions.Customer reviews can be analysed to find satisfaction ofcustomers. Attrition modelling helps to understand moodof customer and take the moves in business. Response modelling is similar to attrition modelling. By predicting anegative behaviour of customer the corrective actions canbe taken for purchase or response.

4. BIG DATA ANALYTICS

Based on the data the analytics technique is applied to make the inferences. Table 1 shows different techniques available in the text analysis. It includes text mining, data mining, machine learning, information retrieval, and natural language processing and sentiment analysis. As big data comprised of images, audio, video the techniques for audio analysis and video analysis are shown. The applications of the text analytics include Stock market prediction, healthcare, finance marketing, education, political, social sciences. In social media analytics in content based analytics content filtering, ranking and tagging is done. Quick insight from existing database is possible. Using structure based analytics; analysis of large data over billions of records is possible. Using social graph and graph analytics identification of most influential accounts is done. Using activity graph identification of strong connectedness from large records is done. After finding such most influential people from the graph analytics from social media, special offers can be designed to those customers. Audio analytics use transcription based and phonetic based approaches to analyse the audio contents. The application of this is customer care analysis and satisfaction analytics. Video analytics applications include automated security and surveillance systems. It also includes the application in retail industry. By observing the videos from customer's interaction in supermarket the items can be placed.

Big data analytics has application in variety of areas. In real time monitoring of businesses it plays an important role. To run competitive business and respond to continuously changing business environment, real time big data analytics is required. Highly transactional businesses produce vast amount of event data that can be managed by the cloud based architecture, which can process big data in real time²³.

Table 1. Big data analytics and applications

Paper	Type	Technique	Subtasks	Advantage	Application
		Information ex- traction	Entity recognition Relation extraction	Evidence-based decision-making	Stock market Prediction
2,3,19,20	Text analytics	Text summarization	a.extractive approach (1.identify main units in text and relationship in them 2. location and frequency of text) b.abstractive approach.(extract semantic info)	Report writing	scientific and news articles, advertisements, emails, and blogs.
		Question answering	1.IR based approach 2. knowledge based approach 3.Hybrid approach	Reduction in response time	healthcare, finance, market- ing, and education
1,3,21	Text analytics	Sentiment analysis	document-level, sentence-level, and aspect-based	Finding positive or negative emotions	Marketing, finance, and the political and social sciences
22,3,12,18	Social media analytics	Content based ana- lytics Structure based analytics	In structure base analytics 1) social graphs 2)activity graphs	Community de- tection Social influence analysis, Link prediction	Quick insights into the public perception 360 degree view of the social issue Ex. Facebook's "People You May Know" YouTube's "Rec- ommended for You",
3	Audio ana- lytics		1.transcript-based approach 2. phonetic based approach	Feedback from customers or agents To handle frustrat- ed callers	customer call centers and healthcare
3	Video ana- lytics		Server-based architecture Edge- based architecture	Placement of items	automated security and surveillance systems, retail industry

In 12 applications based on social big data are reconsidered. The social big data applications are divided in social big data applications related to marketing area, crime analysis area, health care area and user experiences based visualization. In 18 social media twitter is used to find emotions of the users based on the tweet. In this sentiment analysis is used to find the emotions. In 19 content analysis is used to find the environmental disaster situations in the newspaper archives. It describes the system which takes the archives of the newspapers as input and generates useful event summaries from unstructured text. It extracts geographic positions for the event and store in online database that can be searched and visualized using an interactive map. In 20 tweet analysis of academic libraries is done. The most frequently occurred words, bigrams, trigrams are found using text mining methods. Text mining and data mining methods are used to understand importance of social data in academic libraries to help in decision making and strategic planning. Big data analytics applications in 21 include Marketing, finance, and the political and social sciences.

Growing popularity and development in the big data analytics has provided advantage in many of the applications. The applications include retail industry, telecom industry, finance sector, medical diagnosis, banking, manufacturing etc. It provides the excellent result in big data analytics, by processing on large volume of data. At the same time the privacy concern about user is increased. In data mining, emerging topic is privacy preserving data mining. In recent years lot of research is undertaken on this area. PPDM is all about reducing the risk of data mining operations. It focuses on avoiding unwanted disclosure of the sensitive information in the different operations of

knowledge data discovery. The operations include data collecting, pre-processing, publishing and information delivering. The aim of the PPDM is to protect the information for secondary usage by unsanctioned disclosure. But at the same time utility of the data should be intact after applying privacy preserving techniques. While applying the PPDM techniques sensitive information should not be used directly. In themining if the results are of sensitive data, it should be excluded.

5. BIG DATA CHALLENGES

5.1 Challenge 1

To make the business and personalized service the data is collected but which is unknowingly breaching the privacy of the people. For ex. In retail industry in a smart the collection of videos where customer has spent a lot of time, which objects are handled by the customer from this preference of the customer can be known. Even detailed analysis of video and speech or audio of conversation captured while the family is purchasing in a mart can be done. This is helpful for the retailer for making the preference model, next best offer, discounts and placement of the products etc.

Challenge: - Such analysis can raise the privacy concerns also.

5.2 Challenge 2

Same is the case about data generated in terms of videos. The use of CCTV for security has increased the need of analysis of video contents. The videos generated and shared among the groups or individuals on social media have increased quantity of the data. On some social media sites, video content uploading limit for users is increased up to 72 hours per minute. The high resolution video content of one second is equivalent to 2000 text pages. The main problem is integrating this variety of data and management of this data. The extraction of useful information from such data sources is a challenging task. Challenge: - Such large volume of content requires need of scalability in storage systems.

5.3 Challenge 3

95% of the data is being generated in unstructured or semi structured format. As the population is increasing this uses smart phones. According to the number of smart phone users is increased to 75% up from 35% in 2011 in United States. This availability of mobile devices made many things come into reality. The large content of information is available. The people get connected with others for communication virtually.

Challenge: - This connectedness exposes their information to third parties also.

5.4 Challenge 4

Interesting characteristic of big data is veracity; can we trust the data that we have? It is interesting that Van Paul, business leader stated that about one third of big data available in the organization is not trustworthy. Challenge: - So determining the data is truthful is a very important challenge for big data.

5.5 Challenge 5

Data is available in different formats such as structured and unstructured. Much of the unstructured data includes word and excel sheets, messages, tweets, images, audio, video. Few contents of this information may be sensitive in nature²⁶.

Challenge: - In such data personally identifiable information and intellectual property right violation may take place.

6. PRIVACY CONCERNS IN BIG DATA

The information extraction policies of organization have increased the concerns of users about their privacy. The terms user and consumer is used interchangeably in the privacy section. The abundant information coming from sensors, location trackers, GPS, clickstream, log data can be treated as big data. Capturing and sharing such information may be the concern of users. While collecting the user related data there are number of privacy pitfalls, considered in²⁷. Privacy related data is extracted in social media. In²⁸ showed that it is possible to show or identify the location of user from the tweets made by user. The basic machine learning and geotagged information is used for that. Also²⁹ showed that from geotagged twitter information the geographic coordinates can be extracted and it can be extended up to city of user or zip code of location. In³⁰ proposed that image and structural analysis combined with content analysis on geotagged photos with textual tags collected from flicker can be used for finding location. In³¹ authors have considered likes and dislikes which shows interest on Facebook can reveal information about hidden information like location, feelings, relationship status. Considering the above points, the private information or collected data from social media should be manipulated so that risks can be reduced. Due to such privacy concerns about data collected on social media, the users of social media are reluctant to give correct information. Such problem is called as blackhole³².

Big data characteristics like volume, velocity and variety are related to privacy concerns. Large amount of data means the breach of security and violations in the privacy. This leads to dishonesty with the consumer. High velocity data means data coming from sensors, GPS, clickstream. For such data real time analysis is required. This analysis can be used for short term prediction²⁶. The organizations which don't have capability to store the big data, such organizations cannot handle the volume, velocity and complexity of big data. This data is produced at certain time and need to be outsourced. The cloud service providers are providing scalable storage capabilities as per demand^{33,34}. But at the same time the privacy constraints should be applied while handing over this data to cloud service providers. Variety characteristic of big data suggest that data comes in different formats such as csv, images, videos, instant messages, signals. This structured and unstructured information may contain personally identifiable information and intellectual property. Such information capturing and sharing may lead to privacy violations²⁶. According to surveys many organizations lack of comprehensiveness for addressing security and privacy issues. As per the EMC sponsored study conducted by IDC, only one third of the businesses have made the distinction in big data from traditional non big data and adapted tools and management

approaches accordingly. Still many organizations use traditional databases as their main tools of handling data. The consumers have expressed deep concern about dishonesty among the businesses and misuse of personal

information. So consumers are reluctant to give the correct information. Many consumers have taken actions such as turning off information collecting system such as location tracking feature. Consumers are opposing these secondary uses of the data collected for different use³⁵.

7. PRIVACY PRESERVING METHODS

To apply the privacy preserving techniques we have to consider the different dimensions. In multidimensional dataset to find sensitive attributes, quasi identifiers and non-sensitive attributes; different attribute selection methods should be applied³⁶. These methods include Information Gain, Gain ratio, Pearson Correlation, Gini Index. After selection of key identifiers; these identifiers should be modified such that information will not be released to unauthorized user but at the same time utility of data will remain unchanged (Figure 4).

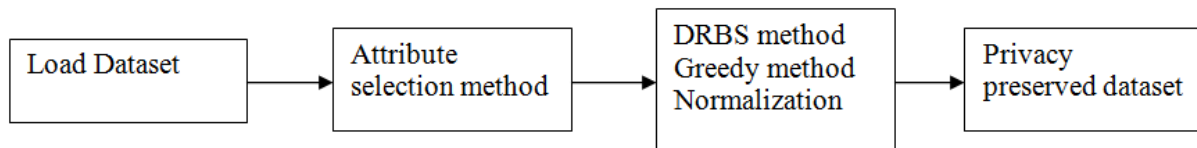


Figure 4. Privacy preserving flow graph.

The methods available for perturbation of key identifiers are data relocation based sub clustering (DRBS)³⁷, Greedy method, Normalization³⁸. In clustering based method; the clusters are found with centroid. Again clustering is applied to find sub clusters. Then distance between the centroid of cluster and parent cluster is found and based on distance sub clusters are arranged. The elements are rotated to neighbour cluster until last element is visited³⁹. In normalization method for perturbation the key identifier values are normalized.

8. CONCLUSION

In this paper need of big data processing is addressed. Advancement in big data analytics is useful for drawing inferences; at the same time, it is main reason for increasing privacy concerns of user. Framework for privacy preserving is discussed.