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 andclickstream data. The objective is to integrate this variety of data at one platform for processing the big data and findprivacy concerns. Methods: The privacy concerns are raised due to unauthorized data extraction, collection and sharinginformation 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 poseddifferent challenges in front of user. Due to large volume and variety of big data many organizations cannot process thedata and needs to outsource it. While sharing such data for processing; there is need to apply proper privacy preservingmeasures. Application/Improvements: Privacy preserving techniques have applications in electronic health recordprocessing, government surveys, outsourcing enterprise data for processing.

1. INTRODUCTION

Due to advancement in microprocessor electronicsand availability of high performance communicationnetworks abundant information is available. The datais getting generated in large quantity from number ofsources. Data generation is estimated up to 2.5 Exabyte(1Exabyte=1,000,000 Terabytes) of data per day1. Figure1 shows the exponential growth of the data. The sourcesfor the data can be categorized in internal and externalsources broadly. Figure 2 shows different sources of bigdata. The internal sources are application log, machinegenerated data, click stream data, sensor data etc. Externalsources of big data generation are social media, enterprisedata such as transactions, emails, contracts. It alsoincludes weather data, sensor generated data for vehicle,traffic, cell phone GPS signals. New York stock exchangegenerate 1 TB data; twitter generates 10TB data every day.This can be fed to sentiment analysis and based on this itcan be discovered what people feel about various productsand events. Volume is important to consider for examplepower meters are generating billons of reading every yearand it is necessary to analyse this data to optimize theenergy and actually see usage at energy per man.

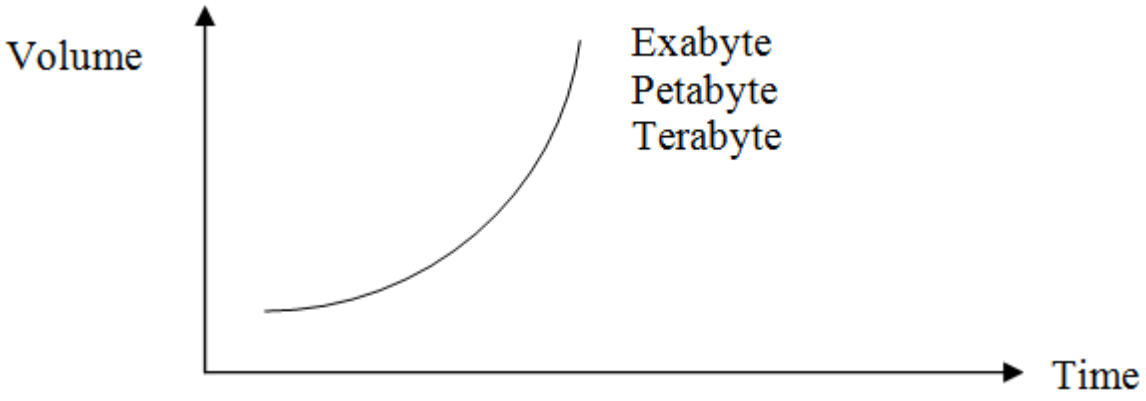

Figure 1. Big data generation at exponential growth.

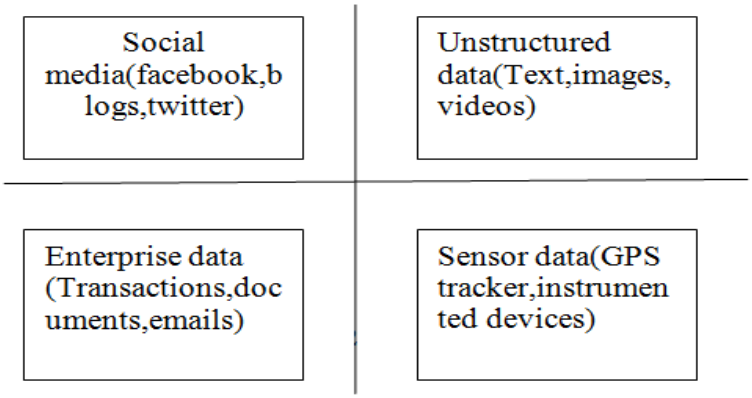

Figure 2. Big data sources.

Velocity is another important characteristic, forvarious time sensitive activities. For example, to decidefraud, seconds can be decisive in being successful ornot. The other aspect of the velocity is that the combingof data which is real time with the default should bepossible. In other example A modern car is having 100s sensors and sensors generating large volume of dataarriving in very rapid way.95% of the data is beinggenerated in unstructured or semi structured format2.As the population is increasing this uses smart phones.Business generate transaction data, but now users beingon the internet generate tremendous amount of dataimages, videos text and it is need to process all of them.The number of smart phone users is increased to 75% upfrom 35% in 2011 in United States3. This availability ofmobile devices made many things come into reality. Thelarge content of information is available. The people getconnected with others for communication virtually.

2. BIG DATA DEFINITIONS

Data generating from different sources have differentcharacteristics. In the definitions available in the literatureare focusing on the large volume, variety, velocity ofthe data. The emphasis is on the processing capabilitiesor infrastructure availabilities available for processing,otherwise which was impossible with traditionalframework.The definitions for big data are leveraging on theability of

business intelligence, competitive intelligence,enhanced insight and decision making. In 2001 definitiongiven by Laney4 as: "high-volume, high-velocity and highvarietyinformation assets that demand cost-effective, innovative forms of information process for enhancedinsight and decision making". In 2012 the definition isupdated in5 as "Big data is high volume, high velocity,and/or high variety information assets that requirenew forms of processing to enable enhanced decision

making, insight discovery and process optimization". Theabove definitions are emphasizing on 3V model i.e. highvolume, high variety and high velocity. Few organizationsadded the term value in definition to make it 4V model.Afterward veracity term is added for big data to call it 5Vmodel. In6 addition of ambiguity, viscosity, and virility the 3V model is discussed. Lack of metadata causes theambiguity for example in the large volume of data M and Fcan be taken for March and February instead of male andfemale. Viscosity is the measure of resistance. Viscosityfor example resistance in data flow, business rules andtechnology may cause loss of business. Virality measures how fast data can spread. For example, re-tweets on atweet. The ambiguity, viscosity and virality characteristics are useful from the point of analysis.

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

3. BIG DATA PROCESSING

In processing of big data, we have to consider diversityof data. The data is taken to one platform. Based on theinternal and external sources steps can be identified as 1)acquisition of data from different sources, 2) processing3) visualize 4) intelligence (Figure 3)

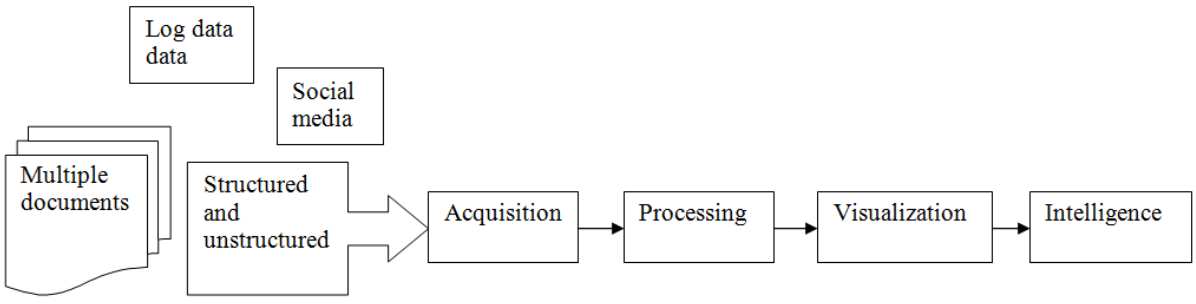

Figure 3. Big data processing flow.

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 batch10,6.

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. Responsemodelling 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 tomake the inferences. Table 1 shows different techniquesavailable in the text analysis. It includes text mining, datamining, machine learning, information retrieval, andnatural language processing and sentiment analysis. Asbig data comprised of images, audio, video the techniquesfor audio analysis and video analysis are shown. Theapplications of the text analytics include Stock marketprediction, healthcare, finance marketing, education,political, social sciences.In social media analytics in content based analyticscontent filtering, ranking and tagging is done. Quickinsight from existing database is possible. Using structurebased analytics; analysis of large data over billions ofrecords is possible. Using social graph and graph analyticsidentification of most influential accounts is done. Usingactivity graph identification of strong connectedness fromlarge records is done. After finding such most influentialpeople from the graph analytics from social media, specialoffers can be designed to those customers.Audio analytics use transcription based and phoneticbased approaches to analyse the audio contents. Theapplication of this is customer care analysis and satisfactionanalytics. Video analytics applications include automatedsecurity and surveillance systems. It also includes theapplication in retail industry. By observing the videosfrom customer's interaction in supermarket the items canbe placed.

Big data analytics has application in variety of areas. Inreal time monitoring of businesses it plays important role.To run competitive business and respond to continuouslychanging business environment, real time big data analytics is required. Highly transactional businessesproduce vast amount of event data that can be managedby the cloud based architecture, which can process bigdata in real time23.

Table 1. Big data analytics and applications

Paper	Type	Technique	Subtasks	Advantage	Application
[2,3,19,20]	Text analytics	Information extraction	Entity recognition Relation extraction	Evidence-based decision-making	Stock market Prediction
		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, marketing, 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 analytics Structure based analytics	In structure base analytics 1) social graphs 2)activity graphs	Community detection 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 "Recommended for You",
[3]	Audio analytics		1.transcript-based approach 2. phonetic based approach	Feedback from customers or agents To handle frustrated callers	customer call centers and healthcare
[3]	Video analytics		Server-based architecture Edge-based architecture	Placement of items	automated security and surveillance systems, retail industry

In12 applications based on social big data areconsidered. The social big data applications are dividedin social big data applications related to marketing area,crime analysis area, health care area and user experiencesbased visualization. In18 social media twitter is used tofind emotions of the users based on the tweet. In thissentiment analysis is used to find the emotions.In19 content analysis is used to find the environmentaldisaster situations in the newspaper archives. It describesthe system which takes the archives of the newspapersas input and generates useful event summaries fromunstructured text. It extracts geographic positionsfor the event and store in online database that can besearched and visualized using an interactive map. In20tweet analysis of academic libraries is done. The mostfrequently occurred words, bigrams, trigrams are foundusing text mining methods. Text mining and data miningmethods are used to understand importance of socialdata in academic libraries to help in decision making andstrategic planning. Big data analytics applications in21 include Marketing, finance, and the political and socialsciences.

Growing popularity and development in the bigdata analytics has provided advantage in many of theapplications. The applications include retail industry,telecom industry, finance sector, medical diagnosis,banking, manufacturing etc. It provides the excellentresult in big data analytics, by processing on large volumeof data. At the same time the privacy concern about useris increased. In data mining, emerging topic is privacypreserving data mining. In recent years lot of researchis undertaken on this area. PPDM is all about reducingthe risk of data mining operations. It focuses on avoidingunwanted disclosure of the sensitive information inthe different operations of

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

5. BIG DATA CHALLENGES

5.1 Challenge 1

To make the business and personalized service thedata is collected but which is unknowingly breachingthe privacy of the people. For ex. In retail industry in amart the collection of videos where customer has spentlot of time, which objects are handled by the customerfrom this preference of the customer can be known2.Even detailed analysis of video and speech or audio ofconversation captured while the family is purchasingin mart can be done. This is helpful for the retailer formaking the preference model, next best offer, discountsand placement of the products etc.

Challenge: - Such analysis can raise the privacy concernsalso.

5.2 Challenge 2

Same is the case about data generated in terms of videos.The use of CCTV for security has increased the needof analysis of video contents. The videos generated andshared among the groups or individuals on social mediahave increased quantity of the data. On some social mediasites, video content uploading limit for users is increasedup to 72 hours per minute22. The high resolution videocontent of one second is equivalent to 2000 text pages.The main problem is integrating this variety of dataand management of this data. The extraction of usefulinformation from such data sources is challenging task24.Challenge: - Such large volume of content requires needof scalability in storage systems.

5.3 Challenge 3

95% of the data is being generated in unstructured orsemi structured format2. As the population is increasingthis uses smart phones. According to3 the number ofsmart phone users is increased to 75% up from 35% in2011 in United States. This availability of mobile devicesmade many things come into reality. The large content ofinformation is available. The people get connected withothers for communication virtually.

Challenge: - This connectedness exposes theirinformation to third parties also25.

5.4 Challenge 4

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

5.5 Challenge 5

Data is available indifferent formats such as structuredand unstructured. Much of the unstructured data includesword and excel sheets, messages, tweets, images, audio,video. Few contents of this information may be sensitivein nature26.

Challenge: - In such data personally identifiableinformation and intellectual property right violationmay take place.

6. PRIVACY CONCERNS IN BIG DATA

The information extraction policies of organization haveincreased the concerns of users about their privacy. Theterms user and consumer is used interchangeably in theprivacy section. The abundant information coming fromsensors, location trackers, GPS, clickstream, log datacan be treated as big data. Capturing and sharing suchinformation may be the concern of users. While collectingthe user related data there are number of privacy pitfalls,considered in27. Privacy related data is extracted in socialmedia. In28 showed that it is possible to show or identifythe location of user from the tweets made by user. Thebasic machine learning and geotagged information isused for that. Also29 showed that from geotagged twitterinformation the geographic coordinates can be extractedand it can be extended up to city of user or zip code oflocation. In30 proposed that image and structural analysiscombined with content analysis on geotagged photoswith textual tags collected from flicker can be used forfinding location. In31 authors have considered likes anddislikes which shows interest on Facebook can revealinformation about hidden information like location,feelings, relationship status.Considering the above points, the private informationor collected data from social media should be manipulatedso that risks can be reduced. Due to such privacy concernsabout data collected on social media, the users of socialmedia are reluctant to give correct information. Suchproblem is called as blackhole32.

Big data characteristics like volume, velocity andvariety are related to privacy concerns. Large amountof data means the breach of security and violations inthe 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. Theorganizations which don’t have capability to store thebig data, such organizations cannot handle the volume,velocity and complexity of big data. This data is producedat certain time and need to be outsourced. The cloudservice providers are providing scalable storage capabilityas per demand33,34. But at the same time the privacyconstraints should be applied while handing over thisdata to cloud service providers. Variety characteristicof big data suggest that data comes in different formatssuch as csv, images, videos, instant messages, signals. Thisstructured and unstructured information may containpersonally identifiable information and intellectualproperty. Such information capturing and sharing mayleads to privacy violations26.According to surveys many organizations lack ofcomprehensiveness for addressing security and privacy issues. As per the EMC sponsored study conducted byIDC, only one third of the businesses have made thedistinction in big data from traditional non big data andadapted tools and management

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

information. So consumers are reluctant to give thecorrect information. Many consumers have taken actionssuch as turning off information collecting system such aslocation tracking feature. Consumers are opposing thesecondary uses of the data collected for different use35.

7. PRIVACY PRESERVING METHODS

To apply the privacy preserving techniques we have toconsider the different dimensions. In multidimensionaldataset to find sensitive attributes, quasi identifiers andnon-sensitive attributes; different attribute selectionmethods should be applied36. These methods includeInformation Gain, Gain ratio, Pearson Correlation, GiniIndex. After selection of key identifiers; these identifiersshould be modified such that information will not bereleased to unauthorized user but at the same time utilityof data will remain unchanged (Figure 4).

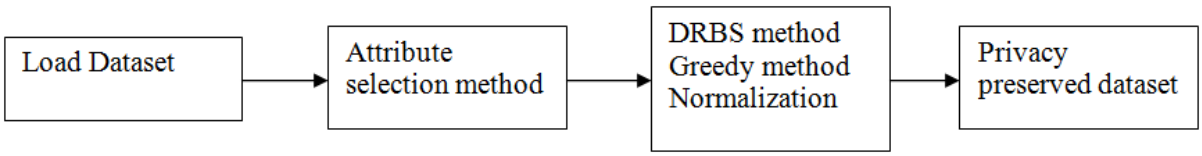

Figure 4. Privacy preserving flow graph.

The methods available for perturbation of key identifiersare data relocation based sub clustering(DRBS)37, Greedymethod, Normalization38. In clustering based method;the clusters are found with centroid. Again clustering isapplied to find sub clusters. Then distance between thecentroid of cluster and parent cluster is found and basedon distance sub clusters are arranged. The elements arerotated to neighbour cluster until last element is visited39. Innormalization method for perturbation the key identifiervalues normalized.

8. CONCLUSION

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