ISSN 2320 - 2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol. 4 No.1, Pages : 22 –31 (2015)
Special Issue of ICCSIE 2015 - Held during January 16, 2015, Bangalore, India
http://warse.org/pdfs/2014/iccsie2015sp04.pdf

Case Study in Identifying frauds using Big Data Analysis

Ch.Hephzibah , Faculty, P.G. Dept of Computer Science, TJPS College, Guntur, c.hepsiba@gmail.com
L.Padmavathi ,Faculty P.G. Dept of Computer Science, TJPS College, Guntur, padmavathi.tjps@gmail.com

ABSTRACT

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing tools and methods .This paper investigates the analysis of Big Data technology which can be applied to the case of fraud detections in every industry.

Keywords: Big Data, Data Mining,FraudDetectionMining.

INTRODUCTION

Big data is certainly one of the biggest buzz phrases in IT today. Combined with virtualization and cloud computing, big data is a technological capability that will force data centers to significantly transform and evolve. Numerous technological innovations are driving the dramatic increase in data and data gathering. This is an area in which the capabilities of the technology and the range of potential applications are evolving rapidly and there is ongoing discussion of the implications of big data .This is why big data has become a recent area of strategic investment for IT organizations. Big data is currently a major topic of discussion across a number of fields, including management and marketing, scientific research, national security, government transparency and open data. Both public and private sectors are making increasing use of big data analytics.

Rapid advances in digital sensors, networks, storage, and computation along with their availability at low cost are leading to the creation of huge collections of data - dubbed as Big Data. This data has the potential for enabling new insights that can change the way business, science, and governments deliver services to their consumers and can impact society as a whole.

To realize the full potential of Big Data Computing, we need to address several challenges and develop suitable conceptual and technological solutions for dealing them. These include life-cycle management of data, large-scale storage, flexible processing infrastructure, data modeling, scalable machine learning and data analysis algorithms, techniques for sampling and making trade-off between data processing time and accuracy, and dealing with privacy and ethical issues involved in data sensing, storage, processing, and actions. In recent years frauds poses prominent role, in every industry. In order to identify the frauds the availability of data sets became insufficient for traditional data processing tools and methods. Big data is a popular term used to describe the exponential growth and availability of data for analyzing the fraud detections. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data ("big data") to discover patterns and other useful information in many ways. Every big data source has different characteristics, including the volume, variety and velocity of data.

This paper is intended to give an overview of the issues as we see them and contribute to the debate on big data. In the paper we refer to a number of examples of big data applications used by companies and cite reports and other publications from companies.

There are several sources of big data and the corresponding mining techniques that might be applied.

1. **Social network profiles**—Tapping user profiles from Facebook, LinkedIn, Yahoo, Google, and specific-interest social or travel sites, to cull individuals' profiles and demographic information, and extend that to capture their hopefully-like-minded networks. (This requires a fairly straightforward API integration for importing pre-defined fields and values – for example, a social network API integration that gathers every B2B marketer on Twitter.)

2. Social influencers—Editor, analyst and subject-matter expert blog comments, user forums, Twitter & Facebook "likes," Yelp-style catalog and review sites, and other review-centric sites like Apple's App Store, Amazon, ZDNet, etc. (Accessing this data requires Natural Language Processing and/or text-based search capability to evaluate the positive/negative nature of words and phrases, derive meaning, index, and write the results).

3.**Activity-generated data**—Computer and mobile device log files, aka "The Internet of Things." This category includes web site tracking information, application logs, and sensor data – such as check-ins and other location tracking – among other machine-generated content. But consider also the data generated by the processors found within vehicles, video games, cable boxes or, soon, household appliances. (Parsing technologies such as those from Splunk or Xenos help make sense of these types of semi-structured text files and documents.)

4. **Software as a Service (SaaS) and cloud applications**—Systems like Salesforce.com, Netsuite, SuccessFactors, etc. all represent data that's already in the Cloud but is difficult to move and merge with internal data. (Distributed data integration technology, in-memory caching technology and API integration work may be appropriate here.)

5.**Public**—Microsoft Azure MarketPlace/DataMarket, The World Bank, SEC/Edgar, Wikipedia, IMDb, etc. – data that is publicly available on the Web which may enhance the types of analysis able to be performed. (Use the same types of parsing, usage, search and categorization techniques as for the three previously mentioned sources.)

6.**Hadoop MapReduce application results**—The next generation technology architectures for handling and parallel parsing of data from logs, Web posts, etc., promise to create a new generations of pre- and post-processed data. We foresee a ton of new products that will address application use cases for any kinds of Big Data – just look at the partner lists of Cloudera and Hortonworks. In fact, we won't be surprised if layers of MapReduce applications blending everything mentioned above (consolidating, "reducing" and aggregating Big Data in a layered or hierarchical approach) are very likely to become their own "Big Data".

7. Data warehouse appliances—Teradata, IBM Netezza, EMC Greenplum, etc. are collecting from operational systems the internal, transactional data that is already prepared for analysis. These will likely become an integration target that will assist in enhancing the parsed and reduced results from your Big Data installation.

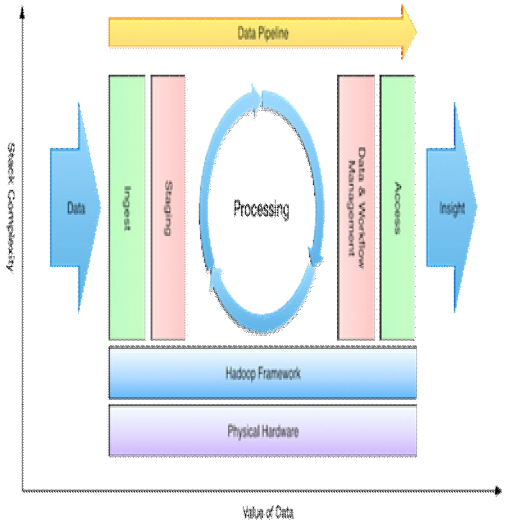
8. **Columnar/NoSQL data sources**—MongoDB, Cassandra, InfoBright, etc. – examples of a new type of map reduce repository and data aggregator. These are specialty applications that fill gaps in Hadoop-based environments, for example Cassandra's use in collecting large volumes of real-time, distributed data.

9.Network and in-stream monitoring technologies—Packet evaluation and distributed query processing-like applications as well as email parsers are also likely areas that will explode with new startup technologies.

10.Legacy documents—Archives of statements, insurance forms, medical record and customer correspondence are still an untapped resource. (Many archives are full of old PDF documents and print streams files that contain original and only systems of record between organizations and their customers. Parsing this semi-structured legacy content can be challenging without specialty tools like Xenos.)

Big Data Architecture

Big Data architecture is premised on a skill set for developing reliable, scalable, completely automated data pipelines. That skill set requires profound knowledge of every layer in the



stack, beginning with cluster design and spanning everything from Hadoop tuning to setting up the top chain responsible for processing the data. The following diagram shows the complexity of the stack, as well as how data pipeline engineering touches every part of it. A good first step is to classify the big data problem according to the format of the data that must be processed, the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyze and store.

To simplify the complexity of big data types, we classify big data according to various parameters and provide a logical architecture for the layers and high-level components involved in any big data solution. Next, we propose a structure for classifying big data business problems by defining atomic and composite classification patterns. These patterns help determine the appropriate solution pattern to apply. We include sample business problems from various industries. The following steps explain how to classify big data and defining a logical architecture of the layers and components of a big data solution.

- Understanding atomic patterns for big data solutions
- Understanding composite (or mixed) patterns to use for big data solutions
- Choosing a solution pattern for a big data solution
- Determining the viability of a business problem for a big data solution
- Selecting the right products to implement a big data solution

Using big data type to classify big data characteristics

It's helpful to look at the characteristics of the big data along certain lines — for example, how the data is collected, analyzed, and processed. Once the data is classified, it can be matched with the appropriate big data pattern:

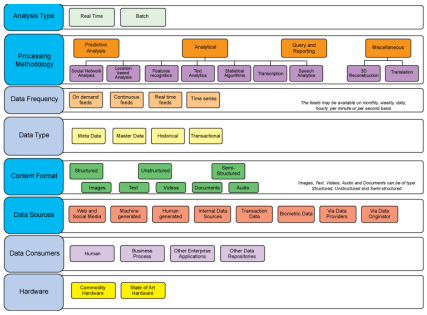
- **Analysis type** — Whether the data is analyzed in real time or batched for later analysis. Give careful consideration to choosing the analysis type, since it affects several other decisions about products, tools, hardware, data sources, and expected data frequency. A mix of both types may be required by the use case:
  - Fraud detection; analysis must be done in real time or near real time.
  - Trend analysis for strategic business decisions; analysis can be in batch mode.
- **Processing methodology** — The type of technique to be applied for processing data (e.g., predictive, analytical, ad-hoc query, and reporting). Business requirements determine the appropriate processing methodology. A combination of techniques can be used. The choice of processing methodology helps identify the appropriate tools and techniques to be used in your big data solution.
- **Data frequency and size** — How much data is expected and at what frequency does it arrive. Knowing frequency and size helps determine the storage mechanism, storage format, and the

necessary preprocessing tools. Data frequency and size depend on data sources:

- On demand, as with social media data
- Continuous feed, real-time (weather data, transactional data)
- Time series (time-based data)

- **Data type** — Type of data to be processed — transactional, historical, master data, and others. Knowing the data type helps segregate the data in storage.
- **Content format** — Format of incoming data — structured (RDMBS, for example), unstructured (audio, video, and images, for example), or semi-structured. Format determines how the incoming data needs to be processed and is key to choosing tools and techniques and defining a solution from a business perspective.
- **Data source** — Sources of data (where the data is generated) — web and social media, machine-generated, human-generated, etc. Identifying all the data sources helps determine the scope from a business perspective. The figure shows the most widely used data sources.
- **Data consumers** — A list of all of the possible consumers of the processed data:
  - Business processes
  - Business users
  - Enterprise applications
  - Individual people in various business roles
  - Part of the process flows
  - Other data repositories or enterprise applications
- **Hardware** — The type of hardware on which the big data solution will be implemented — commodity hardware or state of the art. Understanding the limitations of hardware helps inform the choice of big data solution.

Figure 1 depicts the various categories for classifying big data. Key categories for defining big data patterns have been identified and highlighted in striped blue. Big data patterns, defined in the next article, are derived from a combination of these categories.



General issues concerning big data (are discussed) are as follows:

In **clinical medicine**, in the early stages of a BPS relationship, we agree on the scope and range of the work to be done by the CGI team. Included in this scope is the definition of the data to which we will have access, the privacy and confidentiality controls, and the responsibilities for reporting and analysis each partner will assume. Another important step at this stage is identifying the initial set of edits. There are two sources of input for the starting edits: the client organization's policies and CGI's knowledge of edits that have worked in the past for similar organizations. Once the process is underway, data begins to accumulate showing actual results of edits and various levels of audit and recovery actions applied to the wide variety of claims. At this point, the data scientists can go to work, using their knowledge of the data, the clinical domain, the existing edit rules, and their ability to creatively construct new hypotheses to test. Their analysis includes the following:

- Details of claims and recoveries

- Trends in the effectiveness of existing edits, especially in terms of the tendency for them to "wear out" as providers change their behavior
- Medical records (scanned paper files and newer electronic medical records, which are used in the audits themselves and in analytics)
- News stories about health care fraud, which give ideas for new hypotheses and may also inspire copycats
- Anonymous tips (these are routed to the data scientists, so they can aid in the follow-up investigations)

Improper medical claims fall on a spectrum from simple errors, to mild types of claim inflation, all the way to large scale and sophisticated fraud operations. These different kinds of claim situations call for different analytical approaches.

- **Erroneous claims:** These are often identified by logical consistency and policy rules.For example, root canals are not performed by ophthalmologists. Because health diagnoses and medical practice are complex by nature, analysis is often needed to discover these relationships.
- **Ordinary claim inflation:** This falls in a grey area between error and intentional misrepresentation. Inflated claims usually make logical sense on paper, but are improper because the claim does not match the reality of the care situation. An example is a set of codes indicating a physician office visit. Sometimes a longer or more intensive visit is claimed, where a short and uncomplicated visit was actually delivered. This is an area where predictive models can be employed to great effect to find these improper payments, where claim characteristics are correlated to a high probability of a recoverable payment.
- **Fraud:** Serious cases of intention misrepresentation are difficult to detect, and fall outside the normal activity of claims audit. When situations arise that present reasonable suspicion of criminal activity, those cases are turned over to a special unit for further investigation and possible legal action. The analytical techniques needed to uncover
- Potential fraud is equally sophisticated, including pattern recognition and social network analysis. Even then, data and models can't prove intent, only uncover the evidence.

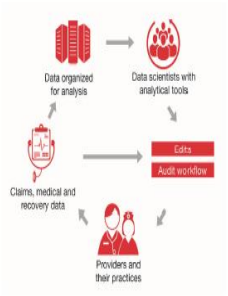
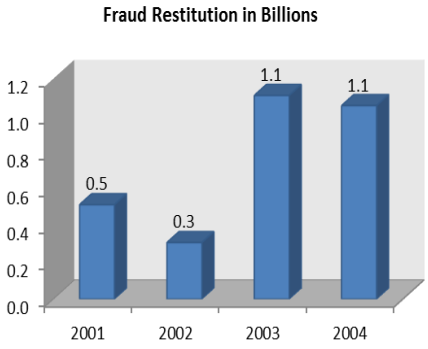


Figure 2: Schematic diagram of the claims audit and data analysis functions integrated to form a closed-loop process

Conclusion: In this paper we have described how the combination of business process workflow, data analytics and smart people with the right skills can produce sustained and measurable business benefits. Big data is expected to play an important role in identifying causality of patient symptoms, in predicting hazards of disease incidence or reoccurrence, and in improving primary-care quality.

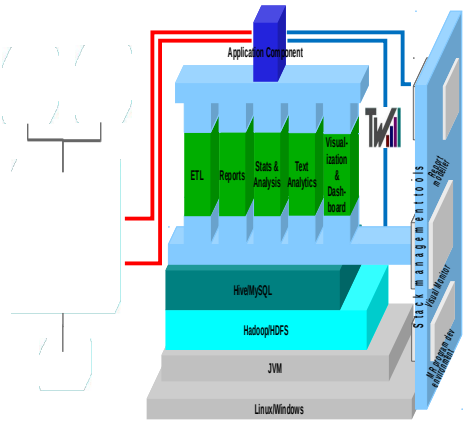


In **Health Care Industry, Healthcare insurance suffering from Fraud,** Current form of Fraud Management is based on enterprise data and heuristics**.** The healthcare domain has been an easy target for people who seek easy money by using fraud methods. Health budgets are a common target of fraudulent practices. Due to the complicated nature of medical processes, frauds have always found a favorable environment in the health insurance. Healthcare fraud is expected to continue to rise as people live longer. This increase will produce a greater demand for Medicare benefits. Additionally, fraudulent billings and medically unnecessary services billed to health care insurers are prevalent throughout the world. The quality of the outcome of the predictive analysis depends on the quality of historical data. A data set which covers a wide range of cases can made better predications. Also, analytical processing with more influencing factors results with higher confidence. In order to exploit both of these, one needs a more powerful computing platform. Big Data platform is not only capable of processing terabytes or megabytes of data but also supports massively parallel processing.

Advent of Big Data analytics makes Healthcare fraud detection more reliable and quicker

Healthcare claim system leverages the power of Big Data platform by delegating its analytical needs. A Big Data platform (as shown in the figure) has ability to sift through a huge amount of historical data in relatively shorter amount of time, so that the business transactions can use fraud detection on real time.

Typical, a Big Data platform based on FMS.The FMS service provider manages the entire infrastructure with an assured quality of service (QoS). These services are, usually, available as software-as-a-service (SaaS) on pay as you go basis.



Following are the benefits of a BigData based FMS over a traditional process

1. Can process a large volume of historical data using complex algorithms
2. Business process automation can run almost real-time analysis before proceeding to the next step
3. Analysis is based on both enterprise and unstructured information from the Net.
4. Service is available online as a web-service which does not mandate software installation on the client premise - “Pay as you use”.

FMS undergoes a continuous self-learning on the basis of transaction made on an ongoing basis Traditional Healthcare Fraud Management analytical solutions are based on enterprise data which is limited in many respects - variety of data, processing speed and analytical algorithms. This paper provides one of many approaches to utilize fraud management solution to detect potential frauds. The solution is based on a high volume of historical data, predictive statistical models.

Highly Targeted Selling, Discrimination based on big data models has the potential to permeate all areas of the market. For business executives in multiple functions, across many industries and geographies, “big data” presents tremendous opportunities. Big Data, although poorly defined, has created considerable interest amongst both vendors and communication service providers (CSPs) and Big Data is now on the agenda of most CSPs at board level. This paper

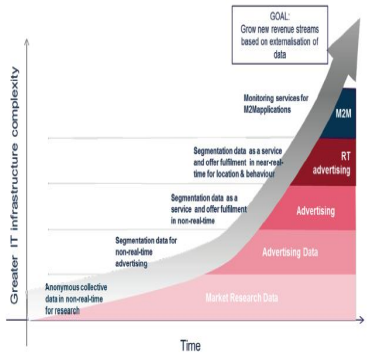
presents the results of an operator survey, which highlights the use of Big Data to enable context - aware marketing campaigns and to help improve customer care as two areas of immediate interest. The use of Big Data analytics to build new data services to be sold to external partners is immature, but holds the potential for some significant revenue streams. The report looks as some of the market drivers that makes Big Data and associated analytics tools important for CS This combination of Big Data characteristics is driving substantial changes within IT requirements, with the greater use of particularly unstructured or semi-structured data changing storage and modeling requirements.

Big Data also makes use of tracking transient data, sometimes referred to as data in motion, which has a much higher value if it is analyzed and acted on quickly.

This requirement for near-real-time analytics to be performed on large data sets, combined with a need to act on the results for huge volumes of data, is moving BDA to need tight integration into a business process management engine as an integral part of the system. In the past most processes were done off-line and used manual interventions, but new uses need to act faster, at a lower cost and on larger number of insights, driving the need for automation.

Figure 1 shows the move to process driven and in-line analytics where processes are automated from insights derived from analytics tools and systems.

Figure 1: Big Data IT infrastructure requirements [Source: Analysys Mason, 2013]



Conclusion: This paper analyses the impact of Big Data on communications service providers and uses primary research through an industry survey to understand how CSPs are currently using Big Data and what their plans are in the future for preventing the frauds in industry.

CREDIT CARD FRAUD

Due to the theatrical increase of fraud which results in loss of dollars worldwide each year, several modern techniques in detecting fraud are persistently evolved and applied to many business fields. In recent decades have seen a gigantic expansion in the use of credit cards as a true transactional medium. Data mining is rising as one of the chief features of many homeland security\initiatives. Often, it is used as a means for detecting fraud, assessing risk, as well as product retailing. Data mining is becoming increasingly common in both the private as well as Public sectors. Data mining involves the use of data analysis tools to find out formerly unknown, believable Patterns and relationships in large data sets. Credit card offers a number of secondary benefits unavailable from cash or checks. Credit cards are safer from theft than is cash. Fraud detection involves monitoring the behaviour of populations of users in order to estimate, detect or avoid unwanted behaviour.

Data mining involves the use of complicated data analysis tools to discover previously unknown, valid patterns and relationships among large data sets. Data mining applications can use a range of parameters to observe the data. This includes association, classification, sequence or path analysis, clustering and forecasting. When using normal measure, detection of credit card fraud is a tricky task.

Introduction of new technologies such as telephone, automated teller machines (ATMs) and credit card systems have enlarged the amount of fraud loss for many banks. Analyzing whether each transaction. Being processed is legitimate or not is very expensive is another task to determine transaction genuinely. Credit Card Fraud Detection domain presents a number of challenging issues for data mining as well

- There are millions of credit card transactions processed each day. Mining of such massive amount of data requires highly efficient techniques that scale data efficiently
- Highly skewed -data,
- Each transaction record has a different dollar amount and there is a chance of variable potential loss

Problem of detecting fraudulent transactions occurs after they have been focused to fraud prevention. Methods and relevant processes. Credit cards create fascinating problems since generally no pin is required for their use; only the name, expiration date and account number is required. Popular means of criminally transacting with credit cards is by stealing someone's identity & in some cases, creating a new fake identity

Different Type of Fraud Techniques: There are many ways in which fraudsters bring out credit card frauds. The technology changes, so does the technology of fraudsters varies and thus the mode in which fraudsters go about carrying out fraudulent activities. Frauds can be broadly categorized into three stages i.e**., traditional card related frauds, merchant related frauds and Internet frauds**. Different types of methods for committing credit card frauds are: **Merchant Connected Frauds (MCF), Merchant Collusion, Triangulations, Internet Associated Frauds (IAF)**.

Credit card frequently used as a necessary mode of payments in today's society. People used credit card for a range of reason such as obtaining credit facility, cash advance, easy payment, charge card. There are some controversial issues that have been addressed not only in terms of the numbers of credit flooding the nation's economy, but the amount transactions that end up with payment default and the numbers of credit card fraud as been recorded which endangered the economy should be seriously paying attention

Conclusion: Banks, credit card brands, payment processors, and e-commerce companies launching growth initiatives are dramatically reducing fraud by as much as 80 percent while reducing false alarms with Feedzai Fraud Prevention. Feedzai' s simple installation process is yet another way it differs from other commonly used enterprise fraud management solutions that require to deploy and refine rules, in addition to periodic service engagements for further tuning. No other enterprise fraud management solution on the market today is built on big data technology and can analyze as much historical data (KPI). .

Illegal Use Illegal practices of tapping into data in new ways have caused quite a scare among those who value their privacy. This document describes how the incorporation of Big Data is changing security analytics by providing new tools and opportunities for leveraging large quantities of structured and unstructured data.

Big Data analytics –the process of analyzing and mining Big Data–can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

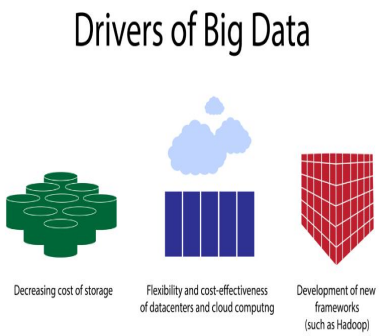


Figure 2. Technical factors driving Big Data adoption

1. Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends.

2. Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics.
3. Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use
4. Predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

New Big Data technologies, such as databases related to the Hadoop ecosystem and stream processing, are enabling the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. These technologies will transform security analytics by:

(a) Collecting data at a massive scale from many internal enterprise sources and external sources such as vulnerability databases;
(b) Performing deeper analytics on the data;
(c) Providing a consolidated view of security-related information
(d) Achieving real-time analysis of streaming data.

It is important to note that Big Data tools still require system architects and analysts to have a deep knowledge of their system in order to properly configure the Big Data analysis tools.

This section describes examples of Big Data analytics used for security purposes like: Network Security, Enterprise Events Analytics, Advanced Persistent Threats Detection (API's)

Conclusions: The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential.

Finally the contribution of this paper is to provide an analysis of the available literature on big data analytics. Accordingly, some of the various big data tools, methods and technologies which can be applied are discussed and their applications and opportunities provided in several decision domains are portrayed. The literature was selected based on its novelty and discussion of important topics related to big data, in order to serve the purpose of research.

REFERENCES

- http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics
- *www.sas.com/en_us/insights/**analytics**/**big-data-analytics**.html*
- *www.ibm.com/**big-data**/us/en/**big-data**-and-**analytics**/*
- *www.edureka.co/**Big-Data**-and-Hadoop*
- http://tdwi.org/portals/big-data-analytics.aspx
- *www.workday.com/applications/**big_data_analytics**.php*
- Dembosky A: "Data Prescription for Better Healthcare." Financial Times, December 12, 2012,
- Feldman B, Martin EM, Skotnes T:"Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie.
- Fernandes L, O'Connor M, Weaver V: Big data, bigger outcomes. JAHIMA 2012, 38-42.
- Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. 2013.
- Wullianallur Raghupathi and Viju Raghupathi Big data analytics in healthcare: promise and potential
- A. Shen, R. Tong, and Y. Deng, "Application of classification models on credit card fraud detection," June 2007
- Abhinav Srivastava, Amlan Kundu, Shamik Sural, Arun K. Majumdar, "Credit Card Fraud Detection using Hidden Markov Model," *IEEE Transactions*

On Dependable And Secure Computing, vol. 5, Issue no. 1, pp.37-48, January-March 2008.

- Aihua Shen, Rencheng Tong, Yaochen Deng, Application of Classification Models on Credit Card Fraud Detection, 2007 IEEE.
- Amlan Kundu, Suvasini Panigrahi, Shamik Sural and Arun K. Majumdar, "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning," *Special Issue on Information Fusion in Computer Security*, Vol. 10, Issue no 4, pp.354- 363, October 2009.
- CLIFTON PHUA1*, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2**A** Comprehensive Survey of Data Mining-based Fraud Detection Research
- **Alperovitch, D.** (2011). Revealed: Operation Shady RAT. Santa Clara, CA: McAfee.
- **Bilge, L. & T. Dumitras.** (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh, NC.
- **Bryant, R., R. Katz & E. Lazowska.** (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science and society. Washington, DC: Computing Community Consortium.
- **Camp, J.** (2009). Data for Cybersecurity Research: Process and "whish list". Retrieved July 15, 2013, from http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf.