BIG DATA ANALYTICS

Module I :

- **Introduction:** Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011, and in every ten minutes in 2013. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it is being neglected.
- 90% of the world's data was generated in the last few years.

Big data and its importance:

Big data:

Big data means really a big data , it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

(or)

Big data refers to datasets whose size is beyond the ability of typical database software tool to capture, store, managed and analyze.

Here 's our standard answer in three parts:

1.Computing perfect storm: Big Data analytics are the natural result of four major global trends: •Moore's Law (technology always gets cheaper),

•Mobile computing(that smart phone or mobile tablet in your hand),

•Social networking(Facebook, Foursquare, Pinterest, etc.), and

•Cloud computing(you don't even have to own hardware or software anymore;).

2.Data perfect storm: Volumes of transactional data have been around for decades for most big firms, but now:

•Gates have now opened with more volume, and the velocity and variety—the three Vs

•Three Vs makes it extremely complex and cumbersome with the current data management and analytics technology and practices.

3.Convergence perfect storm: New alternatives for IT and business executives to address Big Data analytics; Merging of

•Traditional DBMS and analytics software and hardware technologies,

•Open-source technology, and

•Commodity hardware

Aside from the changes in the actual hardware and software technology, there has also been a massive change in the actual evolution of data systems: (Misha Ghosh - Innovator)

- Dependent (Early Days). Data systems were fairly new and users didn't know quite know what they wanted. IT assumed that "Build it and they shall come."
- Independent (Recent Years). Users understood what an analytical platform was and worked together with IT to define the business needs and approach for deriving insights for their firm.
- Interdependent (Big Data Era). Interactional stage between various companies, creating more social collaboration beyond your firm's walls

A flood of mythic "start up" proportions:

- People who compare the amount of data produced daily to a deluge of mythic proportions are entirely correct. This fl ood of data represents something we've never seen before. It 's new, it 's powerful, and yes, it 's scary but extremely exciting
- This is something that every entrepreneur takes to heart as they evangelize their start-up 's big idea that they know will impact the world! This is also true with Big Data and the new technology and approaches that have arrived at our doorstep.
- Over the past decade companies like Facebook, Google, LinkedIn, and eBay have created treasured fi rms that rely on the skills of new data scientists, who are breaking the traditional barriers by leveraging new technology and approaches to capture and analyze data that drives their business.
- It 's all about finding the right home for the new approaches and making them work for you!

Big data is more than merely big why now?

- Defines Big Data as "data that becomes large enough that it cannot be processed using conventional methods."
- Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective.
- Technology advances over time, the size of datasets that qualify as big data will also increase. Depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry.
- The real challenge is identifying or developing most cost-effective and reliable methods for extracting value from all the terabytes and peta bytes of data now available. That's where Big Data analytics become necessary.

Why Now?

Figure 1.1 shows a timeline of recent technology developments.

- If you believe that it 's possible to learn from past mistakes, then one mistake we certainly do not want to repeat is investing in new technologies that didn 't fi t into existing business frameworks.
- During the customer relationship management (CRM) era of the 1990s, many companies made substantial investments in customer-facing technologies that subsequently failedto deliver expected value.
- The reason for most of those failures was fairly straightforward: Management either forgot (or just didn 't know) that big projects require a synchronized transformation of people, process, and technology.
- We can avoid those kinds of mistakes if we keep our attention focused on the outcomes we want to achieve. The technology of Big Data is the easy part—the hard part is fi guring out what you are going to do with the output generated by your Big Data analytics

A convergence of key trends, a wider variety of data:

- The difference between "Old Big Data" and "New Big Data" is accessibility. Here 's a brief summary of our interview:
- Companies have always kept large amounts of information. But until recently, they stored most of that information on tape. While it 's true that the amount of data in the world keeps growing, the real change has been in the ways that we access that data and use it to create value.
- Today, you have technologies like Hadoop, for example, that make it functionally practical to access a tremendous amount of data, and then extract value from it. The availability of lower-cost hardware makes it easier and more feasible to retrieve and process information, quickly and at lower costs than ever before.
- So it 's the convergence of several trends—more data and less expensive, faster hardware—that 's driving this transformation. Today, we 've got raw speed at an affordable price. That cost/benefit has really been a game changer for us.

A Wider Variety of Data

- The variety of data sources continues to increase
- Traditionally, internally focused operational systems, such as ERP (enterprise resource planning) and CRM applications, were the major source of data used in analytic processing.
- wider variety of data sources such as:
  - Internet data (i.e., clickstream, social media, social networking links)
  - Primary research (i.e., surveys, experiments, observations)
  - Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)
  - Location data (i.e., mobile device data, geospatial data)
  - Image data (i.e., video, satellite image, surveillance)
  - Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information)
  - Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

The expanding universe of unstructured data

- structured data (the kind that is easy to defi ne, store, and analyze)
- unstructured data (the kind that tends to defy easy defi nition, takes up lots of storage capacity, and is typically more diffi cult to analyze).
- *Unstructured data* is basically information that either does not have a predefi ned data model and/or does not fi t well into a relational database.
- Unstructured information is typically text heavy, but may contain data such as dates, numbers, and facts as well.
- The term *semi-structured data* is used to describe structured data that doesn ’t fi t into a formal structure of data models.
- T he amount of data (all data, everywhere) is doubling every two years.
- Our world is becoming more transparent. We, in turn, are beginning to accept this as we become more comfortable with parting with data that we used to consider sacred and private.

- Most new data is unstructured. Specifi cally, unstructured data represents almost 95 percent of new data, while structured data represents only 5 percent.
- Unstructured data tends to grow exponentially, unlike structured data,which tends to grow in a more linear fashion.
- Unstructured data is vastly underutilized. Imagine huge deposits of oil or other natural resources that are just sitting there, waiting to be used. That ’s the current state of unstructured data as of today. Tomorrow will be a different story because there ’s a lot of money to be made for smart individuals and companies that can mine unstructured data successfully.
- **Big Data analytics uses a wide variety of advanced analytics, as listed in Figure 1.2 , to provide:**

■ **D eeper insights.** Rather than looking at segments, classifi cations,regions, groups, or other summary levels you ’ll have insights into *all*the individuals, *all* the products, *all* the parts, *all* the events, *all* thetransactions, etc.

■ **B roader insights.** The world is complex. Operating a business in a global, connected economy is very complex given constantly evolving and changing conditions. As humans, we simplify conditions so we can process events and understand what is happening. Big Data analytics takes into account all the data, including new data sources, to understand the complex, evolving, and interrelated conditions to produce more accurate insights.

■ **Frictionless actions.** Increased reliability and accuracy that will allow the deeper and broader insights to be automated into systematic actions

Industry examples of big data: Digital marketing and the online world

Industry Examples of Big Data

I) Digital Marketing

- Introduction
- Database Marketers, Pioneers of Big Data Big Data &
- New School of Marketing Cross Channel Life cycle
- Marketing Social and
-
-

Introduction

- Digital marketing encompasses using any sort of online media (profit or non-profit) for driving people to a website, a mobile app etc. and retaining them and interacting with them to understand what consumers really want.
- Digital \marketing is easy when consumers interact with corporate` primary platform (ie. The one the corporate own) because corporate get good information about them. But corporate get very little information once people start interacting with other platforms (eg. Face book, Twitter, Google +).
- One of the problems that have to be dealt with in a digital existence, is that corporate do not have access to data for all consumers (ie. There is very little visibility about people when they interact in social networking sites). So corporate lose control of their ability to access the data they need, in order to make smart, timely decisions.

- Big data on the web will completely transform a company's ability to understand the effectiveness of its marketing and the ability to understand how its competitors are behaving. Rather than a few people making big decisions, the organizations will be able to make hundreds and thousands of smart decisions every day.

Database Marketers, Pioneers of Big Data:

Database marketing is concerned with building databases containing info about individuals, using that information to better understand those individuals and communicating effectively with some of those individuals to drive business value. Marketing databases are typically used for

- Customer acquisition
- Retaining and cross-selling to existing customers which reactivates the cycle

- As companies grew and systems proliferated, a situation where there was one system for one product and another for another product etc. was landed up (silos).

- Then companies began developing technologies to manage and duplicate data from multiple sources companies started developing software that could eliminate duplicate customer info (de-duping).

- This enable them to extract customer information from silos product systems, manage the info into single database, remove all the duplicates and then send direct mail to subsets of the customers in the database.

- Companies such as Reader's Digest and several other firms were early champions of this new kind of marketing and they used it very effectively. By the 1980's marketers developed the ability to run reports on the info in their databases which gave them better and deeper insights into buying habits and preferences of customers.

- Telemarketing became popular when marketers figured out how to feed information extracted from customer databases to call centers. In 1990's email entered the picture and marketers saw opportunities to reach customers via Internet and WWW.

- In the past five years there has been exponential growth in database marketing and the new scale is pushing up against the limits of technology

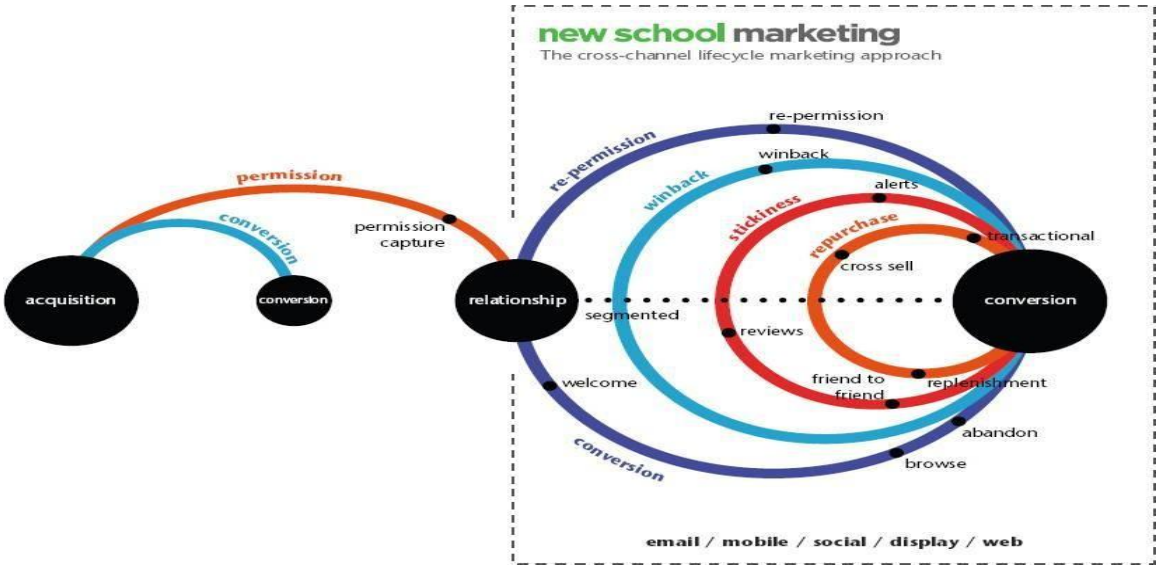
Big Data & New School of Marketing: New school marketers deliver what today's consumers want ie. Relevant interactive communication across digital power channels Digital power channels: email, mobile, social display and web. Consumers have changed so must marketers.

Right approach – Cross Channel Lifecycle Marketing

- Cross-Channel Lifecycle Marketing really starts with the capture of customer permission, contact information, and preferences for multiple channels.
- It also requires marketers to have the right integrated marketing and customer information systems, so that (1) they can have complete understanding of customers through stated preferences and observed behavior at any given time; and (2) they can automate and optimize their programs and processes throughout the customer lifecycle.
- Once marketers have that, they need a practical framework for planning marketing activities.

- The various loops that guide marketing strategies and tactics in the Cross- Channel Lifecycle Marketing approach: conversion, repurchase, stickiness, win-back, and re-permission are shown in the following figure.

II) Financial Services



i) Fraud & Big Data

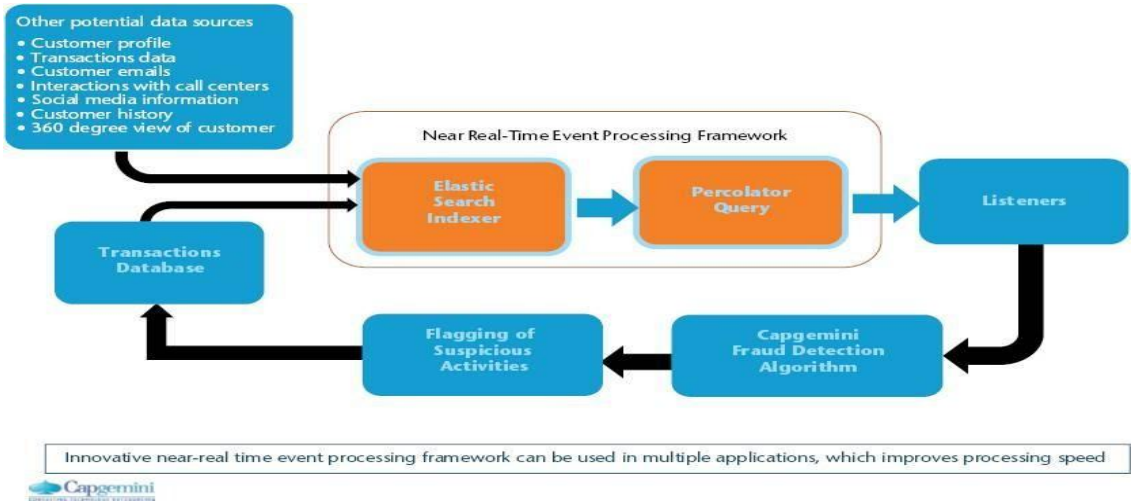
- Fraud is intentional deception made for personal gain or to damage another individual.
- One of the most common forms of fraudulent activity is credit card fraud.
- Social media and mobile phones are forming new frontiers fraud.
- Capegemini financial services team believes that due to the nature of data streams and processing required BIG Data Technologies provide an optimal technology solution based on the following three Vs :

1. **High volume:** Years of consumer records and transactions (150 billion +
2. records per year).
3. **High velocity:** Dynamic transactions and social media info.
4. **High variety:** Social media plus other unstructured data such as customer E-mails, call center conversations as well as transactional structured data.

Fraud Detection Powered by Near Real-Time Event Processing Framework

- The near real-time event processing framework can be used in multiple applications which improves processing speed.

- This fraud detection system uses an open source search server based on Apache Lucene.

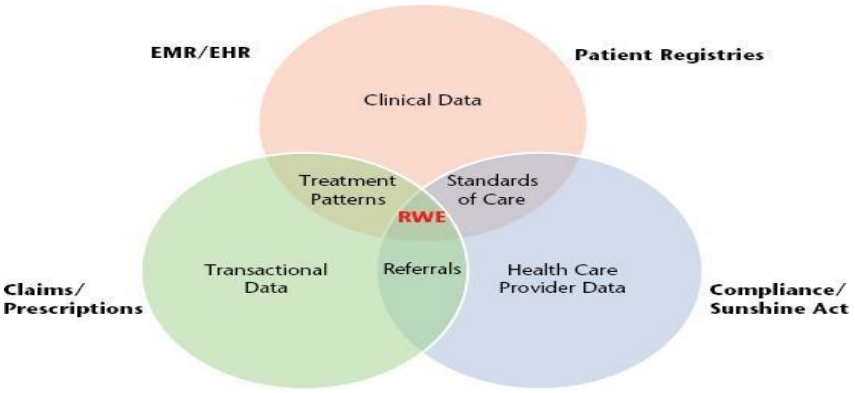


- It can be used to search all kind of documents at near real-time. The tool is used to index new transactions which are sourced in real-time, which allows analytics to run in a distributed fashion utilizing the data specific to the index.
- Using this tool, large historical data sets can be used in conjunction with real-time data to identify deviation from typical payment patterns.
- The big data component allows overall historical patterns to be compared and contrasted and allows the number of attributes and characteristics about consumer behavior to be very wide with little impact on overall performance.
- Percolator query performs the function of identifying new transactions that have raised profiles. Percolator query can handle both structured and unstructured data.
- This provides scalability to the event processing framework and allows specific suspicious transactions to be enriched with additional unstructured information (E.g. Phone location/geospatial records, customer travel schedules and so on).
- This ability to enrich the transaction further can reduce false positives and increase the experience of customer while redirecting fraud efforts to actual instances of suspicious activity.
- Capegemini's fraud Big Data initiative focuses on flagging the suspicious credit card transactions to prevent fraud in near real-time via multi-attribute monitoring.
- Real- time inputs involving transaction data and customers records are monitored via validity checks and detection rules.

III) Big data and Healthcare

- Big data promises enormous revolution in healthcare, with the advancements in everything from the management of chronic disease to the delivery of personalized medicine.
- In addition to saving and improving lives, Big Data has the potential to transform the entire healthcare system by replacing guesswork and institution with objective data-driven science.
- The healthcare industry now has huge amount of data: from biological data such as gene expression, Special Needs Plans (SNPs), proteomics, metabolomics, and next-generation gene sequence data etc. The exponential growth in data is further accelerated by the digitization of patient level data stored in Electronic Medical Records (EMRs) or Electronic Health Records (EHRs) and Health Information Exchanges (HIEs) enhanced with data from imaging and test results, medical and prescription claims and personal health devices.
- In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science (see Figure).

Figure: Data in the World of Healthcare



- The healthcare system is facing severe economic, effectiveness and quality challenges. These factors are forcing transformation in pharmaceutical business model. Hence the healthcare industry is moving from traditional model built on regulatory approval and settling of claims to medical evidence and proving economic effectiveness through improved analytics derived insights. The

success of this model depends on the creation of robust analytics capability anharnessing integrated real-world patient level data.

Advertising and Big Data

Big Data is changing the way advertisers address three related needs.

(i) How much to spend on advertisements.

(ii) How to allocate amount across all the marketing communication touch points.

How to optimize advertising effectiveness. Given these needs advertisers need to measure their advertising end to end in terms of Reach, Response & Reaction.

LAST YEAR QUESTION PAPER:

1.What are real-time industry applications of Hadoop?

- Hadoop, well known as Apache Hadoop, is an open-source software platform for scalable and distributed computing of large volumes of data.
- It provides rapid, high performance and cost-effective analysis of structured and unstructured data generated on digital platforms and within the enterprise. It is used in almost all departments and sectors today. I would like to share a interesting video related to hadoop.

Some of the instances where Hadoop is used:

• Managing traffic on streets
• Streaming processing
• Content Management and Archiving Emails
• Processing Rat Brain Neuronal Signals using a Hadoop Computing Cluster
• Fraud detection and Prevention
• Advertisements Targeting Platforms are using Hadoop to capture and analyze click stream, transaction, video and social media data
• Managing content, posts, images and videos on social media platforms
• Analyzing customer data in real-time for improving business performance
• Public sector fields such as intelligence, defense, cyber security and scientific research
• Financial agencies are using Big Data Hadoop to reduce risk, analyze fraud patterns, identify rogue traders, more precisely target their marketing campaigns based on customer segmentation, and improve customer satisfaction
• Getting access to unstructured data like output from medical devices, doctor's notes, lab results, imaging reports, medical correspondence, clinical data, and financial data.

What is the importance of unstructured data in real word? Give Examples?

unstructured data: It might be clear to many, but just so we are all on the same page, unstructured data is images, video, sound and documents .