**UNIT 1**

Big Data

* Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

**Classification of Big Data:**

**1. Social Networks (human-sourced information)**:

Data are loosely structured and often ungoverned.

  - Social Networks: Facebook, Twitter, Tumblr etc.

  - Blogs and comments

  - Personal documents

  - Pictures: Instagram, Flickr, Picasa etc.

  -. Videos: Youtube etc.

  - Internet searches

  -Mobile data content: text messages

  - User-generated maps

  - E-Mail

**2. Traditional Business systems (process-mediated data)**:

Traditional business data is the vast majority of what IT managed and processed, in both operational and BI systems. Usually structured and stored in relational database systems.

-Data produced by Public Agencies

 - Medical records

 - Data produced by businesses

 - Commercial transactions

 - Banking/stock records

 - E-commerce

**3. Internet of Things (machine-generated data)**:

derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured.

- Data from sensors

 - Fixed sensors

 - Home automation

 - Weather/pollution sensors

 - Traffic sensors/webcam

 - Scientific sensors

 - Security/surveillance videos/images

 - Mobile sensors (tracking)

 -  Data from computer systems

**Types of Big Data:**

Three concepts come with big data : structured, semi structured and unstructured data.

**Structured Data**

It concerns all data which can be stored in database SQL  in table with rows and columns. They have relational key and  can be easily mapped into pre-designed fields. Today, those data are the most processed in development and the simplest way to manage informations.

But structured data represent only 5 to 10% of all informatics datas.

**Semi structured data**

Semi-structured data is information that doesn’t reside in a relational database but that does have some organizational properties that make it easier to analyze. With some process you can store them in relation database (it could be very hard for some kind of semi structured data), but the semi structure exist to ease space, clarity or compute…

Examples of semi-structured :CSV but  XML and JSON documents are semi structured documents,  NoSQL databases are considered as semi structured.

But as Structured data, semi structured data represents a few parts of data (5 to 10%) so the last data type is the strong one : unstructured data.

**Unstructured data**

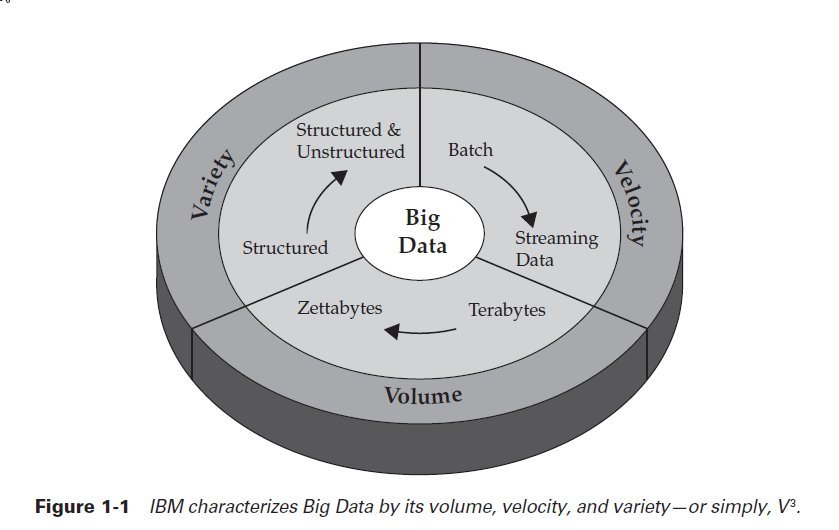
Unstructured data represent around 80% of data. It often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. Note that while these sorts of files may have an internal structure, they are still considered « unstructured » because the data they contain doesn’t fit neatly in a database.

Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine generated or human generated.

Here are some examples of machine-generated unstructured data:

* **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery. Just think about Google Earth, and you get the picture.
* **Scientific data:** This includes seismic imagery, atmospheric data, and high energy physics.
* **Photographs and video:** This includes security, surveillance, and traffic video.
* **Radar or sonar data:** This includes vehicular, meteorological, and oceanographic seismic profiles.
* **Elements of Big Data:**
* **Volume**

Big data implies enormous volumes of data. It used to be employees created data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive. Yet, Inderpal states that the volume of data is not as much the problem as other V’s like veracity.

* **Variety**
* Variety refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data. Jeff Veis, VP Solutions at HP Autonomy presented how HP is helping organizations deal with big challenges including data variety.
* **Velocity**
* Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and ROI if you are able to handle the velocity. Inderpal suggest that sampling data can help deal with issues like volume and velocity.
* **Veracity**
* Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Inderpal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity. In scoping out your big data strategy you need to have your team and partners work to help keep your data clean and processes to keep ‘dirty data’ from accumulating in your systems.
* 

**Web Analytics**

**Web analytics** is the measurement,collection analysis and reporting of web data for purposes of understanding and optimizing web usage. However, Web analytics is not just a process for measuring web traffic but can be used as a tool for business and market research, and to assess and improve the effectiveness of a website. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It helps one to estimate how traffic to a website changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a website and the number of page views. It helps gauge traffic and popularity trends which is useful for market research.

**Big Data and Marketing**

Having big data doesn’t automatically lead to better marketing – but the potential is there. Think of big data as your secret ingredient, your raw material, your essential element. It’s not the data itself that’s so important. Rather, it’s the insights derived from big data, the decisions you make and the actions you take that make all the difference.

By combining big data with an integrated marketing management strategy, marketing organizations can make a substantial impact in these key areas:

* **Customer engagement.** Big data can deliver insight into not just who your customers are, but where they are, what they want, how they want to be contacted and when.
* **Customer retention and loyalty.** Big data can help you discover what influences customer loyalty and what keeps them coming back again and again.
* **Marketing optimization/performance.** With big data, you can determine the optimal marketing spend across multiple channels, as well as continuously optimize marketing programs through testing, measurement and analysis.

**Fraud and Big Data**

A Fraud can be defined as false representation of facts, leading to concealment or distortion of the truth. Frauds can be committed by both words and conduct and is intended to deceive the other party, generally to gain an advantage over it in some manner. Frauds that occur frequently in financial institutions, such as banks or insurance companies, or involve any type of monetary transactions, such as in the retail industry, are called financial frauds. In such fraudulent cases ,online retailers, such as Amazon, eBay and Groupon , tend to incur huge expenses and losses.

**Common types of financial frauds are:**

**1.Credit Card Fraud- It** is a wide-ranging term for [theft](https://en.wikipedia.org/wiki/Theft) and [fraud](https://en.wikipedia.org/wiki/Fraud) committed using or involving a payment card, such as a credit card or debit card, as a fraudulent source of funds in a transaction. The purpose may be to obtain goods without paying, or to obtain unauthorized funds from an account. Credit card fraud is also an adjunct to identity theft.

**2.Exchange or return policy fraud-**An online retailer always has a policy allowing the exchange and return of goods, and sometimes people take advantage of this policy. These people buy a product online, use it, and then return it back as they are not satisfied with the product.

**3. Personal Information Fraud-**The theft of Personal Information, otherwise known as Identity Theft, is one of the leading contributors to a successful fraud. Identity Theft is a type of fraud which involves stealing money or gaining other benefits by acquiring Personal Information and pretending to be someone else.

**Methods to prevent Fraud:**

Analyzing Big Data allows organizations to:

* Keep track of and process huge volumes of data
* Differentiate between real and fraudulent entries
* Identify new methods of fraud and add them to the list of fraud- prevention checks
* Verify whether a product has actually been delivered to the valid recipient.
* Determine the location of the customer and the time when the product was actually delivered.
* Check the listings of popular retail sites, such as e- Bay, to find whether the product is up for sale somewhere else

RFID**(radio frequency identification)**

RFID (radio frequency identification) is a technology that incorporates the use of electromagnetic or electrostatic coupling in the radio frequency (RF) portion of the electromagnetic spectrum to uniquely identify an object, animal, or person. RFID is coming into increasing use in industry as an alternative to the bar code. The advantage of RFID is that it does not require direct contact or line-of-sight scanning.

An RFID system consists of three components: an antenna and transceiver (often combined into one reader) and a transponder (the tag).

The antenna uses radio frequency waves to transmit a signal that activates the transponder. When activated, the tag transmits data back to the antenna. The data is used to notify a programmable logic controller that an action should occur. The action could be as simple as raising an access gate or as complicated as interfacing with a database to carry out a monetary transaction. Low-frequency RFID systems (30 KHz to 500 KHz) have short transmission ranges (generally less than six feet). High-frequency RFID systems (850 MHz to 950 MHz and 2.4 GHz to 2.5 GHz) offer longer transmission ranges (more than 90 feet). In general, the higher the frequency, the more expensive the system.

|  |
| --- |
| **Inventory Control and RFID Technology:** |
| Effective inventory management depends upon consolidating, integrating, and analyzing data collected from many sources such as, distribution centers and warehouses. Conventional tracking systems require manual intervention, which is labor intensive, time consuming, and error-prone. On the other hand, the use of RFID technology has significant advantages over the conventional methods; these are discussed below: |
|  |
| **Inventory monitoring and asset visibility – Smart Shelves :** |
| In a replenishment-based system, whenever the total inventory at a warehouse or distribution center drops below a certain level, the RFID enabled system could place an automatic order. RFID-tagged products will allow stores to track the location and count of inventories in real time. This will better monitor demand for certain products and place orders to prevent an out-of-stock situation. The high levels of inventory monitoring obtained using RFID can particularly benefit FMCG industries.  http://www.decisioncraft.com/dmdirect/images3/mar.gif |

* **Risk and big data**
* As with any business initiative, a big data project involves an element of risk. Any project can fail for any number of reasons: bad management, under-budgeting, or a lack of relevant skills. However, big data projects bring their own specific risks.
* **Data Security:** The bigger your data, the bigger the target it presents to criminals with the tools to steal and sell it. In the case of Target, hackers stole credit and debit card information of 40 million customers, as well as personal identifying information such as email and geographical addresses of up to 110 million people.
* **Data Privacy:** Closely related to the issue of security is privacy. But in addition to ensuring that people’s personal data are safe from criminals, you need to be sure that the sensitive
* Information you are storing and collecting isn’t going to be divulged through less malevolent but equally damaging misuse by yourself or by people to whom you have delegated responsibility for analyzing and reporting on it.
* **Costs:**Data collection, aggregation, storage, analysis, and reporting all cost money. On top of this, there will be compliancy costs – to avoid falling foul on the issues I raised in the previous point. These costs can be mitigated by careful budgeting during the planning stages, but getting it wrong at that point can lead to spiralling costs, potentially negating any value added to your bottom line by your data-driven initiative. This is why “starting with strategy” is so vital.
* **Bad Analytics:**Aka “getting it wrong.” Misinterpreting the patterns shown by your data and drawing causal links where there is in fact merely random coincidence is an obvious pitfall.
* **Bad Data:**Many data projects that start off on the wrong foot by collecting irrelevant, out of date, or erroneous data. This usually comes down to insufficient time being spent on designing the project strategy.
* **Big Data and Healthcare**
* A number of use cases in healthcare are well suited for a big data solution. Some academic- or research-focused healthcare institutions are either experimenting with big data or using it in advanced research projects. Those institutions draw upon data scientists, statisticians, graduate students, and the like to wrangle the complexities of big data. In healthcare, we do have large volumes of data coming in. EMRs alone collect huge amounts of data. Most of that data is [collected for recreational purposes](https://intermountainphysician.org/UrologySeminar/Documents/10_James.pdf) according to Brent James of [Intermountain Healthcare.](https://intermountainhealthcare.org/) But neither the volume nor the velocity of data in healthcare is truly high enough to require big data today. Our work with health systems shows
* that only a small fraction of the tables in an EMR database (perhaps 400 to 600 tables out of 1000s) are relevant to the current practice of medicine and its corresponding analytics
* use cases. So, the vast majority of the data collection in healthcare today could be considered recreational.

**Big Data and Medicine**

The data generated by medical care and medically relevant research are rapidly becoming bigger and more complex, particularly with the advent of new technologies. Our ability to advance medical care and efficiently translate science into modern medicine is bounded by our capacity to access and process these big data. From human genetics and pathogen genomics to routine clinical documentation, from internal imaging to motion capture, from digital epidemiology to pharmacokinetics, and from treatment pathways to life course assessment, the big Vs of Big Data - volume, variety, velocity and veracity - abound in medicine. Statistical, mathematical, visualisation, and computational approaches, from a wide range of disciplines, as well systems for innovative ICT-based interventions are needed to keep apace of the complexity in Big Data and to advance medicine.

**Advertising and big data**

The big broadcast nets are beginning to book their first big upfront deals—deals that are showing signs of today's complex media world. Technology innovation continues to drive media fragmentation, muddying the once-simple world of TV, radio and print. Advertisers must now divide their budgets among all the proliferating devices and channels people watch. How can they hope to get the return they are used to in terms of what people buy? And how can the broadcast nets and other media companies offer such a return? Technology innovation will make it possible to trim a great deal of waste out of advertising by making it more precise. As advertising becomes more precise, it will become more efficient, which will drive up its ROI*.* This higher ROI will then lead to more investment in advertising.

**Big Data Technologies / Open Source Technologies**

Open source software is computer software that is available in source code form under an open-source license that permits users to study, change and improve and at times also to distribute the software.

Top emerging technologies that are helping users cope with and handle Big Data in a cost-effective manner-

1.**Column-oriented databases**

Traditional, row-oriented databases are excellent for online transaction processing with high update speeds, but they fall short on query performance as the data volumes grow and as data becomes more unstructured. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge data compression and very fast query times. The downside to these databases is that they will generally only allow batch updates, having a much slower update time than traditional models.

2.**MapReduce**

This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers.

Any MapReduce implementation consists of two tasks:

* The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples;
* The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).

3.**Hadoop**

Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data.

4.**Hive**

Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

5.**PIG**

PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

6.**PLATFORA**

Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

7.**SkyTree**

SkyTree is a high-performance machine learning and data analytics platform focused specifically on handling Big Data.

**Introduction to Hadoop**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

**Hadoop Architecture:**

Hadoop framework includes following four modules:

* **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries providesfilesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
* **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
* **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
* **HadoopMapReduce:** This is YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.



Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

**CLOUD AND BIG DATA**

Big Data and cloud computing go hand-in-hand. Cloud computing enables companies of all sizes to get more value from their data than ever before, by enabling blazing(shining intensely) fast analytics at a fraction of previous costs. This, in turn drives companies to acquire and store even more data, creating more need for processing power and driving a virtuous(morally excellent) circle.

**Features of cloud computing that can be useful to handle Big data:**

1. **On-demand self-service:** A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
2. **Broad network access:** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., **mobile** phones, tablets, laptops and workstations).
3. **Resource pooling:** The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state or datacenter). Examples of resources include storage, processing, memory and network bandwidth.
4. **Rapid elasticity:** Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
5. **Measured service:** Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth and active user accounts). Resource usage can be monitored, controlled and reported, providing transparency for the provider and consumer.

**Services Of Cloud Computing:**

* **1. Software as a Service (SaaS)*–***software runs on computers owned and managed by the SaaS provider, versus installed and managed on user computers. The software is accessed over the public Internet and generally offered on a monthly or yearly subscription.
* **2.Infrastructure as a Service (IaaS*) –***compute, storage, networking, and other elements (security, tools) are provided by the IaaS provider via public Internet, VPN, or dedicated network connection. Users own and manage operating systems, applications, and information running on the infrastructure and pay by usage.
* **3.Platform as a Service (PaaS*) –***All software and hardware required to build and operate cloud-based applications are provided by the PaaS provider via public Internet, VPN, or dedicated network connection. Users pay by use of the platform and control how applications are utilized throughout their lifecycle.

**MOBILE BUSINESS INTELLIGENCE**

Mobile business intelligence is software that extends desktop business intelligence (BI) applications so they can be used on a mobile device.

MBI applications optimizes traditional BI reports so they can be viewed easily on a small screen and is ideal for displaying key performance indicators (KPIs) and alerts on small screens with simple charts, graphs and sparklines.

An additional benefit of MBI is that it allows data that's captured by the mobile device to be integrated on-the-fly so that reports are currents and mobile workers can make informed decisions in real time.

**CROWD SOURCING ANALYTICS**

Crowdsourcing is the process of getting work or funding, usually online, from a crowd of people. The word is a combination of the words 'crowd' and 'outsourcing'. The idea is to take work and outsource it to a crowd of workers.