

# A STUDY ON BIG DATA

**S.Maheswari,**

B.Sc (Computer Technology),  
Department Of Computer Science,  
Sri Krishna Adithya College of Arts and Science,  
Coimbatore,Tamilnadu,India.

**S.Dharani,**

B.Sc (Computer Technology),  
Department Of Computer Science,  
Sri Krishna Adithya College of Arts and Science,  
Coimbatore,Tamilnadu,India.

**V.Gayathri,**

B.Sc (Computer Technology),  
Department Of Computer Science,  
Sri Krishna Adithya College of Arts and Science,  
Coimbatore,Tamilnadu,India.

**Abstract:** the term “big data” is relatively new, but the act of gathering and storing large amounts of information for analysis is ages old. 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. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

**Keywords:** large volume data-big data, volume, velocity , variety, cost reduction-time, reduction-smart decision-tools-hadoop-lifecycle.

## I.INTRODUCTION

“Big data” is the collection of large number of data that are stored in cloud servers. Data contains audios, videos, animations, texts, and documents. This type of large volume data contains both structured and unstructured data therefore it is represented as big data. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

## II. DEFINITION OF BIG DATA

While the term “big data” is relatively new, but the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs: **Volume, Velocity, and Variety.**

**Volume:** Large volume data are called big data. Here volume represents huge number of data. Organizations collect data from a variety of sources. It including business transactions, social media and information from sensor or machine-to-machine data, therefore the huge number of data is collected by that organization.

**Velocity:** Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

**Variety:** Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

## III. IMPORTANCE OF BIG DATA

The importance of big data doesn’t revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product

development and optimized offerings, and 4) smart decision making. When you combine big data with high-powered analytics, you can accomplish business-related tasks such as:

- Determining root causes of failures, issues and defects in near-real time.
- Generating coupons at the point of sale based on the customer’s buying habits.
- Recalculating entire risk portfolios in minutes.
- Detecting fraudulent behavior before it affects your organization.

It’s important to remember that the primary value from big data comes not from the data in its raw form, but from the processing and analysis of it and the insights, products, and services that emerge from analysis. The sweeping changes in big data technologies and management approaches need to be accompanied by similarly dramatic shifts in how data supports decisions and product/service innovation.

## IV. SOURCES OF BIG DATA

We have generated more than 90% of data in the last two years itself. And it is getting generated exponentially day by day with the increasing usage of devices and digitization across the globe. Our current output of data is roughly **2.5 quintillion** bytes a day.

Tweets	98,000+
Status updates	695,000
Instant messages	11million
Google searches	694,445
Sending emails	168million+
Data created	1,820TB
New mobile web users	217

## V. BIG DATA TOOLS

Big data is a term used for a collection of data sets, so large and complex that it is difficult to process using traditional applications/tools. It is the data exceeding Terabytes in size. Because of the variety of data that it encompasses, big data always brings a number of challenges relating to its volume and complexity. A recent survey says that 80% of the data created in the world are unstructured. One challenge is how these unstructured data can be structured, before we attempt to understand and capture the most important data. Another challenge is how we can store it. Here are the top technologies used to store and analyses Big Data. We can categorize them into two ways (storage and Querying/Analysis).

**Apache Hadoop:** Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This also replicates data in a cluster thus providing high availability.

**Microsoft HDInsight:** It is a Big Data solution from Microsoft powered by Apache Hadoop which is available as a service in the cloud. HDInsight uses Windows Azure Blob storage as the default file system. This also provides high availability with low cost.

**NoSQL:** While the traditional SQL can be effectively used to handle large amount of structured data, we need NoSQL (Not Only SQL) to handle unstructured data. NoSQL databases store unstructured data with no particular schema. Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are many open-source NoSQL DBs available to analyze big Data.

**Hive:** This is a distributed data management for Hadoop. This supports SQL-like query option HiveSQL (HSQL) to access big data. This can be primarily used for Data mining purpose. This runs on top of Hadoop.

**Sqoop:** This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.

**PolyBase:** This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access data stored in PDW. PDW is a datawarehousing appliance built for processing any volume of relational data and provides integration with Hadoop allowing us to access non-relational data as well.

**Big data in EXCEL:** As many people are comfortable in doing analysis in EXCEL, a popular tool from Microsoft, you can also connect data stored in Hadoop using EXCEL 2013. Hortonworks, which is primarily working in providing Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. You can use Power View feature of EXCEL 2013 to easily summarize the data. Similarly, Microsoft's HDInsight allows us to connect to big data stored in Azure cloud using a power

query option.

**Presto:** Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named presto which is built to handle petabytes of data. Unlike Hive, Presto does not depend on Map Reduce technique and can quickly retrieve data.

## VI. BIG DATA ANALYTICS LIFECYCLE

Big Data analysis differs from traditional data analysis primarily due to the volume, velocity and variety characteristics of the data being processes. To address the distinct requirements for performing analysis on Big Data, a step-by-step methodology is needed to organize the activities and tasks involved with acquiring, processing, analyzing and repurposing data.

The Big Data analytics lifecycle can be divided into the following nine stages, as shown in Figure 1.1

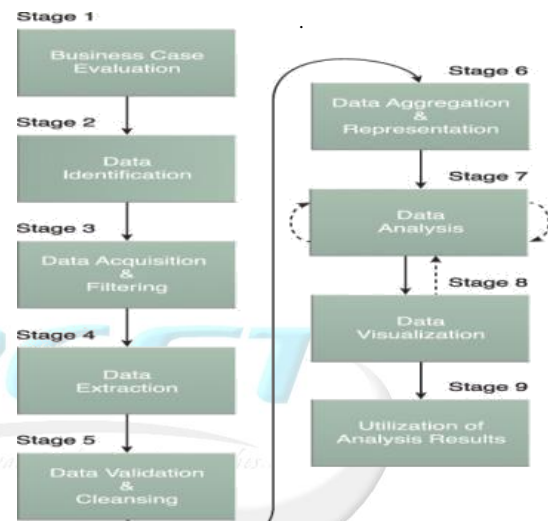


Figure 1.1 the nine stages of the Big Data analytics lifecycle

### Business Case Evaluation (Stage 1):

Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis. The Business Case Evaluation stage requires that a business case be created, assessed and approved prior to proceeding with the actual hands-on analysis tasks. An evaluation of a Big Data analytics business case helps decision-makers understand the business resources that will need to be utilized and which business challenges the analysis will tackle.

### Data Identification (Stage 2):

The Data Identification stage is dedicated to identifying the datasets required for the analysis project and their sources. Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise. In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and matched

against a pre-defined dataset specification. In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are compiled.

**Data Acquisition & Filtering (Stage 3):**

During the Data Acquisition and Filtering stage the data is gathered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives. In many cases, especially where external, unstructured data is concerned, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process. Data classified as “corrupt” can include records with missing or nonsensical values or invalid data types. Data that is filtered out for one analysis may possibly be valuable for a different type of analysis.

**Data Extraction (Stage 4):**

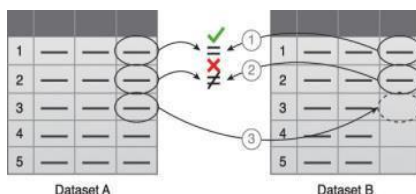
Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. The Data Extraction lifecycle stage is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

**Data Validation & Cleansing (Stage 5):**

Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity. Its complexity can further make it difficult to arrive at a set of suitable validation constraints. Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.

For example, as illustrated in Figure 1.2:

- The first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A.
- If a value is missing, it is inserted from Dataset A.



**Figure 1.2 Data validation can be used to examine interconnected datasets in order to fill in missing valid data.**

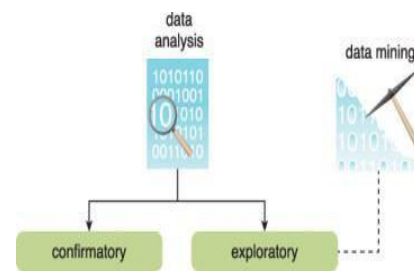
For batch analytics, data validation and cleansing can be achieved via an offline ETL operation. For real time analytics, a more complex in-memory system is required to validate and cleanse the data as it arrives from the source.

**Data Aggregation & Representation (Stage 6):**

The Data Aggregation and Representation stage is dedicated to integrating multiple datasets together to arrive at a unified view. Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconciliation is required or the dataset representing the correct value needs to be determined.

**Data Analysis (Stage 7):**

The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. Data analysis can be classified as confirmatory analysis or exploratory analysis, the latter of which is linked to data mining, as shown in Figure 1.3. Confirmatory data analysis is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand. Exploratory data analysis is an inductive approach that is closely associated with data mining.



**Figure 1.3 Data analysis can be carried out as confirmatory or exploratory analysis.**

**Data Visualization (Stage 8):**

The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users. Business users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback, as indicated by the dashed line leading from stage 8 back to stage 7.

**Utilization of Analysis Results (Stage 9):**

Subsequent to analysis results being made available to business users to support business decision-making, such as via dashboards, there may be further opportunities to utilize the analysis results. The Utilization of Analysis Results stage is dedicated to determining how and where processed analysis data can be further leveraged.

**VII. APPLICATIONS**

Big data applications are creating a new era in every industry. The following are the some examples of Big Data applications in different type of industries.

**Banking and securities :** The Securities Exchange Commission (SEC) is using big data to monitor financial market activity. They are currently using network analytics and natural language processor to catch illegal trading activity in the financial markets. Retail traders, Big banks, hedge funds and other so-called ‘big boys’ in the financial markets use big data for trade analytics used in high frequency trading, pre-trade decision-support analytics, sentiment measurement, Predictive Analytics etc. This industry also heavily relies on

big data for risk analytics including; anti-money laundering, demand enterprise risk management, "Know Your Customer", and fraud mitigation.

**Communications, Media and Entertainment:** Since consumers expect rich media on-demand in different formats and in a variety of devices, some big data challenges in the communications, media and entertainment industry include:

- Collecting, analyzing, and utilizing consumer insights
  - Leveraging mobile and social media content
  - Understanding patterns of real-time, media content usage
- Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to:
- Create content for different target audiences
  - Recommend content on demand
  - Measure content performance

For example, **Amazon Prime**, which is driven to provide a great customer experience by offering, video, music and Kindle books in a one-stop shop also heavily, utilizes big data. Apart from Google, **Facebook** is probably the only company that possesses high level of detailed customer information by using big data technology.

#### **Healthcare Providers:**

Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital. A battery of tests can be efficient but they can also be expensive and usually ineffective. Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease. Obamacare has also utilized big data in a variety of ways.

**Education :** Big data is used quite significantly in higher education. For example, The University of Tasmania, An Australian University with over 26000 students, has deployed a Learning and Management System that tracks among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time. In a different use case of the use of big data in education, it is also used to measure teacher's effectiveness to ensure a good experience for both students and teachers. Teacher's performance can be fine-tuned and measured against student numbers, subject matter,

#### **IX. REFERENCES**

- [1]. <https://intellipaat.com/tutorial/hadoop-tutorial/big-data-overview/>
- [2]. <http://upxacademy.com/big-data-analysis-top-5-challenges/>
- [3]. <http://www.informit.com/articles/article.aspx?p=2473128&seqNum=11>
- [4]. <https://www.simplilearn.com/big-data-applications-in-industries-article>
- [5]. Franks, B. (2012) *Taming the Big Data Tidal Wave*, and New York: Wiley.
- [6]. Gartner (2012) "Gartner Says Big Data Creates Big Jobs: Million IT Jobs Globally to Support Big Data by 2015",

student demographics, student aspirations, behavioral classification and several other variables.

**Manufacturing and Natural Resources :** Increasing demand for natural resources including oil, agricultural products, minerals, gas, metals, and so on has led to an increase in the volume, complexity, and velocity of data that is a challenge to handle. In the natural resources industry, big data allows for predictive modeling to support decision making that has been utilized to ingest and integrate large amounts of data from geospatial data, graphical data, text and temporal data. Areas of interest where this has been used include; seismic interpretation and reservoir characterization. Big data has also been used in solving today's manufacturing challenges and to gain competitive advantage among other benefits.

**Government:** In governments the biggest challenges are the integration and interoperability of big data across different government departments and affiliated organizations. In public services, big data has a very wide range of applications including: energy exploration, financial market analysis, fraud detection, health related research and environmental protection.

Some more specific examples are as follows:

Big data is being used in the analysis of large amounts of social disability claims, made to the **Social Security Administration (SSA)**, that arrive in the form of unstructured data. The analytics are used to process medical information rapidly and efficiently for faster decision making and to detect suspicious or fraudulent claims.

**The Food and Drug Administration (FDA)** is using big data to detect and study patterns of food-related illnesses and diseases. This allows for faster response which has led to faster treatment and less death. The Department of Homeland Security uses big data for several different use cases. Big data is analyzed from different government agencies and is used to protect the country.

#### **VIII. CONCLUSION**

Big data is everywhere and there is almost an urgent need to collect and store whatever data is being generated. There is huge amount of data floating around. Therefore it can be easily analyzed by big data technology. Big Data is a great boon to various industries and organizations, as it is helping them take better decisions, thus profiting the company. Professionals, who are skilled in big data analytics, there is an ocean of opportunities out there.

Gartner Press Release,  
[http://www.gartner.com/newsroom/id/2207915\(currentMarch7,2014\)](http://www.gartner.com/newsroom/id/2207915(currentMarch7,2014)).

- [7]. Harris, D. (2013) "The History of Hadoop: From 4 Nodes to the Future of Data", Gigaom,
- [8]. [http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/\(currentMarch7,2014\)](http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/(currentMarch7,2014)).
- [9]. <http://www.bigdataplanet.info/p/what-is-big-data.html?m=1>
- [10]. <http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/>
- [11]. Healy, M. (2012), "Big Data Lies", Information Week Research Report,