**Different Sources of Data for Data Analysis**

* Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.
* In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data.
* The data which is to be analyzed must be collected from different valid sources.
* The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as “knowledge”.
* Knowledge has many meanings like business knowledge or sales of enterprise products, disease treatment, etc. The main goal of data collection is to collect information-rich data.
* Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection.
* Most of the data collected are of two types known as “qualitative data“ which is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group and another one is “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data.

**Few methods of collecting primary data:**

**1. Interview method:**

* The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

**2. Survey method:**

* The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

**3. Observation method:**

* The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing.

**4. Experimental method:**

* The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.
* **CRD- Completely Randomized design** is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
* **RBD- Randomized Block Design** is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
* **LSD – Latin Square Design** is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
* **FD- Factorial design** is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

**Few methods of collecting secondary data:**

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

**Internal source:**

* These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

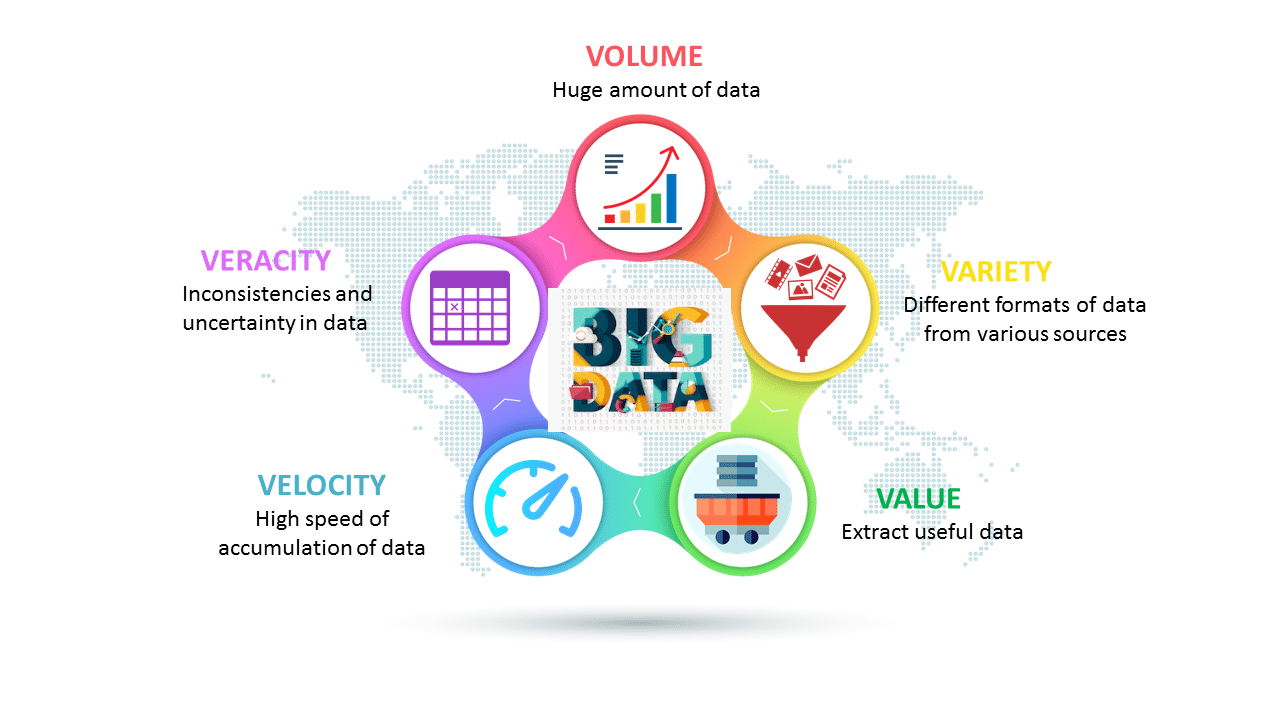
**External source:**

* The data which can’t be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

**Other sources:**

* **Sensors data:**With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
* **Satellites data:**Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
* **Web traffic:** Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

**Following are the big data core characteristics**. Understanding the characteristics of big data is vital to know how it works and how you can use it. There are primarily seven characteristics of big data analytics:



**1. Velocity**

Volume refers to the amount of data that you have. We measure the volume of our data in Gigabytes, Zettabytes (ZB), and Yottabytes (YB). According to the industry trends, the volume of data will rise substantially in the coming years.

**2. Volume**

Velocity refers to the speed of data processing. High velocity is crucial for the performance of any big data process. It consists of the rate of change, activity bursts, and the linking of incoming data sets.

**3. Value**

Value refers to the benefits that your organization derives from the data. Does it match your organization’s goals? Does it help your organization enhance itself? It’s among the most important big data core characteristics.

**5. Veracity**

Veracity refers to the accuracy of your data. It is among the most important Big Data characteristics as low veracity can greatly damage the accuracy of your results.

**6. Validity**

How valid and relevant is the data to be used for the intended purpose.

**7. Volatility**

Big data is constantly changing. The data you gathered from a source a day ago might be different from what you found today. This is called variability of data, and it affects your data homogenization.

**8. Visualization**

Visualization refers to showing your big data-generated insights through visual representations such as charts and graphs. It has become prevalent recently as big data professionals regularly share their insights with non-technical audiences.

**The Evolution of Analytic Scalability**

* Same old methods for handling data just won ’ t work anymore.
* Organizations that don ’ t update their technologies to provide a higher level of scalability will quite simply choke on big data.

**THE CONVERGENCE OF THE ANALYTIC AND DATA ENVIRONMENTS**

* It used to be that analytic professionals had to pull all their data together into a separate analytics environment to do analysis
* None of the data that was required was together in one place, and the tools that did analysis didn ’ t have a way to run where the data set
* The only option was to pull the data together in a separate analytics environment and then start performing analysis.
* Much of the work analytic professionals do falls into the realm of advanced analytics, which encompasses data mining, predictive modeling, and other advanced techniques.
* There ’ s not much difference between a data set as analysts defi ne it and a “ table ” in a database.
* Both a data set and a table contain rows and columns. Often, a single row represents an entity like a customer.
* The columns represent information about the customers such as name, spending level, or status
* Analysts have done “ merges ” of their data sets for years.
* That is the exact same thing as a “ join ” of tables in a database.
* In both a merge and a join, two or more data sets or tables are combined together.
* They are typically merged/joined so that specific rows of one data set or table are combined with specific rows of another.
* Analysts do what is called “ data preparation. ”
* In this process, they pull data from various sources and merge it all together to create the variables required for an analysis.
* In the data warehousing world this process is called “ extract, transform, and load (ETL). ”

**MASSIVELY PARALLEL PROCESSING SYSTEMS**

* Massively parallel processing (MPP) database systems have been around for decades.
* While individual vendor architectures may vary, MPP is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.
* So what is an MPP database, and why is it special?

**IMPORTANCE OF DATA ANALYTICS**

For businesses to achieve a strategic edge, data analytics plays a vital role. Here are a few ways in which we know why is data analytics important for businesses today:-

**1. Product Development**

Data analytics offers both estimation and exploration capability for information. It allows one to understand the market or process’s current state and offers a solid base for forecasting future results. Data analysis helps companies to comprehend the current business situation and change the processor cause the need for a new product creation that meets market requirements.

**2. Target Content**

Learning what consumers wish in advance improves consumer orientation in marketing campaigns. It encourages advertisers to tailor their advertising to a subset of the entire consumer base. It also allows you to figure out which client base group can better respond to the initiative. It also saves money to convince a buyer to buy and increases the overall performance of the marketing activities.

**3. Efficiency in Operations**

Theimportance of data analytics in marketing finds more viable ways to streamline operations or increase benefit levels. It helps to recognize possible issues, avoids the waiting period, and takes action on them.

**Difference between Structured, Semi-structured and Unstructured data**

**Big Data** includes huge volume, high velocity, and extensible variety of data. These are 3 types:

* Structured data,
* Semi-structured data, and
* Unstructured data.

**Structured data –**

* Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data are most processed in the development and simplest way to manage information. Example: Relational data.

**Semi-Structured data –**

* Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyze. With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.

**Unstructured data –**

* Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.

|  |  |  |  |
| --- | --- | --- | --- |
| Properties | Structured data | Semi-structured data | Unstructured data |
| Technology | It is based on Relational database table | It is based on XML/RDF(Resource Description Framework). | It is based on character and binary data |
| Transaction management | Matured transaction and various concurrency techniques | Transaction is adapted from DBMS not matured | No transaction management and no concurrency |
| Version management | Versioning over tuples,row,tables | Versioning over tuples or graph is possible | Versioned as a whole |
| Flexibility | It is schema dependent and less flexible | It is more flexible than structured data but less flexible than unstructured data | It is more flexible and there is absence of schema |
| Scalability | It is very difficult to scale DB schema | It’s scaling is simpler than structured data | It is more scalable. |
| Robustness | Very robust | New technology, not very spread | — |
| Query performance | Structured query allow complex joining | Queries over anonymous nodes are possible | Only textual queries are possible |

**Analytic process and tools**

Ensemble Methods

* + Multiple models are built using multiple techniques. Once the results from all of the models are known, all of the results are combined together to come up with a final answer.
  + The process of combining the various results can be anything from a simple average of each model ’ s predictions to a much more complex formula.
  + It is important to note that ensemble models go beyond picking the best individual performer from a set of models.
  + They actually combine the results of multiple models in order to get to a single, final answer
  + The power of ensemble models stems from the fact that different techniques have different strengths and weaknesses

Commodity Models

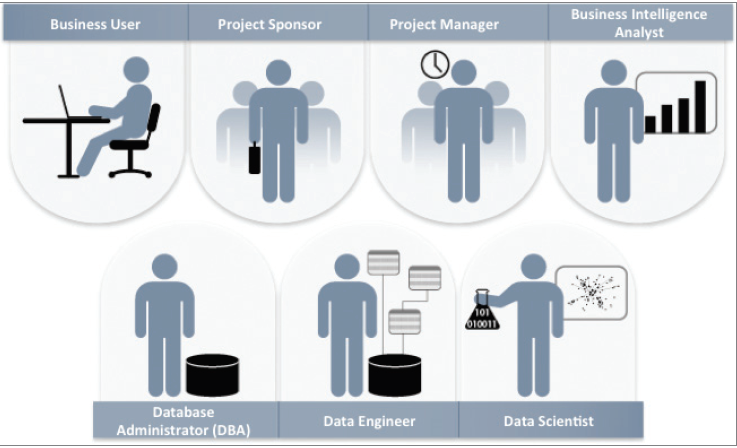
* The goal of a commodity model is not to get the best model, but to quickly get a model that will lead to a better result than if there had been no model at all
* In evaluating a commodity model, the primary concern is that there ’ s a benefit being achieved by using it.
* There may be much room for improvement if more effort was put in.
* But, if a quick model can help in a situation that otherwise wouldn ’ t have a model, it is utilized.

Text Analysis

* One of the most rapidly growing methods utilized by organizations today is the analysis of text and other unstructured data sources.
* A lot of big data falls into these classifications.
* Text analysis, as the name implies, takes some sort of text as input.
* This text can be written material like an e - mail, transcribed material such as a medical dictation, or even text that has been scanned from a hard copy and converted to electric form like old courthouse records.
* The reason text analysis has grown in prominence is because of the wealth of new sources of text data
* Text is a very common type of big data and text analysis tools and methods have come a long way.
* Today there are tools that will help you parse text into its component words and phrases and then assist in determining the meaning of those words and phrases.

**Key Roles for a Successful Analytics Project**

* Figure depicts the various roles and key stakeholders of an analytics project.
* Each plays a critical part in a successful analytics project.
* Although seven roles are listed, fewer or more people can accomplish the work depending on the scope of the project, the organizational structure, and the skills of the participants.
* For example, on a small, versatile team, these seven roles may be fulfilled by only 3 people, but a very large project may require 20 or more people.

****

* Business User: Someone who understands the domain area and usually benefits from the results.
* This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized.
* Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.