

UTILIZING AND HANDLING DATA: DATA ANALYTICS AND MINING IN SOCIAL NETWORKING MEDIA

K Subhash¹ and E N Vihari²

¹Indian Institute of Management (IIM Ranchi), Ranchi, Jharkhand, India

²GITAM Deemed University, Hyderabad, Telangana State, India

Abstract: In the contemporary business connections, unstructured data which forms the world's 85% of the data. This unstructured data which is pooled together for analysis and other processing is as referred to as big data, is generated by either humans or machines. This paper brings out the various delineations of understanding Data and its analytics, encompassing aspects such as Volume, Variety and Velocity. Traditionally, business decisions are based on only 10% of the whole data that is available in structured form. Rest 90% of data was earlier unused. But as the business strategies evolved and the technologies started to advance, the unused data became really important. This research aims at deliberating the factors that come into play in this context and possible areas of implications of data analytics for prediction, analysis and planning of the business concerns. This paper also aims at throwing light upon the data handling capacities between Hadoop and MSBI and MSBI can be used instead of Big data tools to organise and warehouse small to medium sized data sets. The role of Data analytics in social media research and usage of AI for social media data mining shall make a part of discussion in this research paper. The rapid growth of online textual data creates an urgent need for powerful text mining techniques. As an interdisciplinary field, text data mining spans multiple research communities, especially data mining, natural language processing, information retrieval and machine learning with applications in many different areas. Many models and algorithms have been developed for various text mining tasks. The paper concludes a few propositions for future research and implementations in the same connection.

Index Terms: Data, Analytics, Technology, Business.

I. INTRODUCTION

In general, we have 3 types of data namely; structured semi-structured and unstructured data. Here, structured data consists of 5-10% of the whole world's data which is present in RDBMS and other sources and this data is well organised and maintained. Next up is the semi-structured data which sums up to a total of 10-15% of the total world's data and is a mixture of both unorganised and organised data and can be converted into organised data with help of KDD(Knowledge Discovery in Databases).

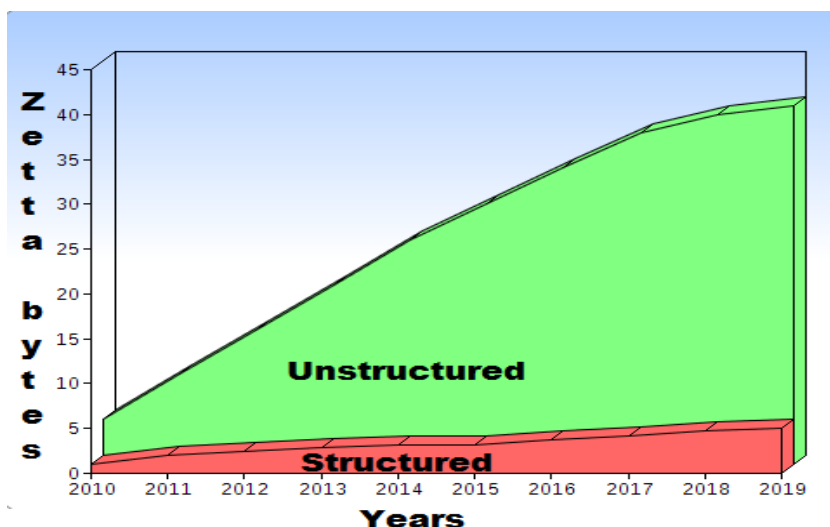


Fig 1. Structured and Unstructured Data statistics. (Source: Google images)

Now comes the major part i.e. unstructured data which forms the world's 85% of the data. One might wonder what this unstructured data which is so huge is. It is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. This unstructured data which is pooled together for analysis and other processing is called as big data and is generated by either humans or machines.

Why has big data evolved in the recent times?

Rapid decrease in the storage costs have dramatically decreased in the past few years and this increases the processing power, flexibilities and effectiveness of the data centres and cloud computing. Volume is the most concerning and alarming dimension as now a days, data generating sources have increased with growing technologies. Here, variety refers to the wide range of data that is being generated by different resources each of them which is quite different from one another. The velocity at which the data is growing is too high for RDBMS's to handle.

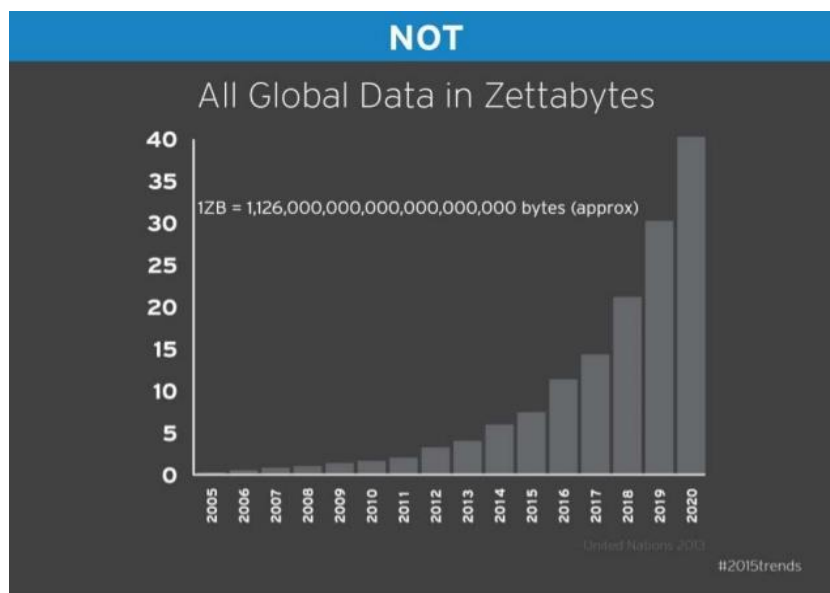


Fig 2. Growth trajectory of Global Data in Zettabytes. (Source: Google images)

What exactly is big data?

Big data is that set of data which has the 3 attributes or 3 V's rule. Only if these are satisfied can a data set be called BIG DATA. The 3 dimensions are:-

1. Volume
2. Variety
3. Velocity

A few more to add to the characteristics are:-

Veracity is a dimension which refers to the noise present in the data which affects the quality and meaningfulness of the data. Validity is another one which refers to the data sets being authentic enough to be processed for various inferences to be made.

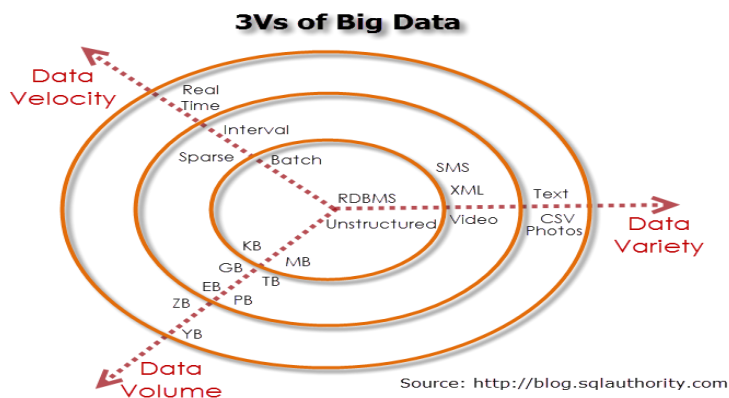


Fig 3. 3V's of Big Data (Source: sqlauthority.com)

The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Data Analytics.

One might wonder why we have put in the above information which is not as much as related to the topic given to us. But the reason that we have put them is that, we should primarily understand the basic facts above which we are going to build upon. Now that we know the various types of data, characteristics of data, and other things, we will be in a better position to understand the BIG DATA analytics as a huge subject. Let's look at the huge data generators:

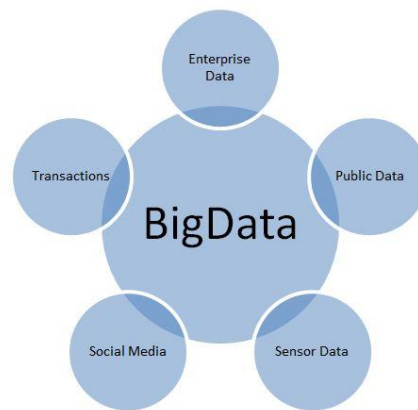


Fig 4. Sources of Big Data (Source: Google images)

- Click
- Advertisements
- Billing events
- Transactions
- Server requests
- Websites
- Phones
- Smart phones
- Social media
- Scientific equipment
- Flights
- Telecom networks
- Log files
- IoT (Internet of Things)
- Sensors
- Road surveillance
- Archives
- Media
- Public web.

Why Big data is important to Business?

Traditionally, business decisions are based on only 10% of the whole data that is available in structured form. Rest all 90% of data was earlier unused. But as the business strategies evolved and the technologies started to advance, the unused data became really important for the following reasons:

- Finding co-relation among various data sets.
- Spot the trends in the business.
- Make better and much predicted decisions
- Increase the revenue with much informed results.
- Win a competitive edge over others.
- To overcome their shortcomings.
- Understand the customer's requirements, etc.

Thus, Big Data Analytics (BDA) has a lot of impacts on running the business. Let us now look into few business applications of Big Data which help us grow our economy and also expand our business by prediction, analysis and planning: -

- **Security:** BDA helps analyze the log files to combat crime, fraud detection, develop spam filters etc.
- **Banking:** BDA helps in investigating and payment fraud detection along with loan defaulters etc.
- **Government:** We use BDA for threat detection & prevention, crime detection & prevention and to trace out the tax compliance (fraud and abuse).
- **Health Care:** It helps prevent spread of epidemics, predict readmissions for better management of hospital, health monitoring and act upon the population health etc.
- **Telecom:** We can use BDA for location-based services, network & call record analysis to detect any suspicious activities and for capacity & pricing optimization.
- **Consumer Companies:** These companies use BDA for building recommendation systems, displaying relevant advertisements and provide better service to the customers with various applications.
- **Retail Organizations:** These use BDA to know the customer preferences and the product perceptions.
- **Resource optimization:** In this the system log files are analyzed for usage statistics.
- **Manufacturing Units:** This utilizes BDA for vibration in machine changes when a unit wears down, vibration data is analyzed to help in predicting the time to optimally replace or repair equipment. To predict the product acceptance and take feedback on products to know the defects and make the necessary changes and improvements and also plan the production in advance depending upon the feedbacks and the product acceptance prediction so as to be ready to face the situations in a very optimized way which will be encountered in the future.
- **Road Traffic:** We use various applications to check out the various possible road ways to reach the destination. BDA analyses the real time roadway traffic conditions and shows us the most optimized ways.
- **Sports:** BDA plays an important role here too by helping to plan effective team strategies.

II. TRADITIONAL DATA ANALYTICS LIFE CYCLE

To give a model to structurally arrange the work needed by an organization and deliver clear insights from Big Data, it's useful to think of it as a cycle with different stages. It is a non-linear, which at all the stages are related with each other.

This cycle has superficial similarities with the more traditional data mining cycle as described in CRISP methodology and SEMMA methodology.

CRISP-DM Methodology: The CRISP-DM methodology stands for Cross Industry Standard Process for Data Mining, is a cycle that describes commonly used approaches that data mining experts use to tackle problems in traditional BI data mining.

SEMMA Methodology: SEMMA is another methodology developed by SAS for data mining modelling, which stands for Sample, Explore, Modify, Model, and Assess. A brief description of its stages are as follows:

- **Sample** – The process starts with data sampling, e.g., selecting the dataset for modelling. Sample stage also deals with data partitioning.
- **Explore** – Understanding of the data by discovering anticipated and unanticipated relationships between the variables, and also abnormalities, with the help of data visualization is done in this stage.
- **Modify** – The Modify phase contains methods to select, create and transform variables in preparation for data modelling.
- **Model** – In the Model phase, the focus is on applying various modelling (data mining) techniques on the prepared variables in order to create models that possibly provide the desired outcome.
- **Assess** – The evaluation of the modelling results shows the reliability and usefulness of the created models.

The main difference between CRISM-DM and SEMMA is that SEMMA focuses on the modelling aspect, whereas CRISP-DM gives more importance to stages of the cycle prior to modelling such as understanding the business problem to be solved, understanding and pre-processing the data to be used as input, for example, machine learning algorithms.

III. DATA HANDLING CAPACITY BETWEEN MSBI AND HADOOP

Data plays a very important role in a company and has always been the heart of the company. The data of any company is not stationary and is ever modified day to day exponentially. Since the start of this era, until 2003 we have at least 5 billion GB's of data, say by 2011 the data has increased enormously in a way that the data got produced in just days' time and further this amount of data is being produced for every couple of minutes. Nowadays, data processing, modelling and reporting have changed dramatically on a wide range of platforms. With the invention of Big data. But the problem is it would be able to handle only one type of data which is structured. Recently it is being handled or trying to handle the unstructured data using the features of MPP. And Big data has a proven experience of handling Unstructured data i.e. with very large data sets.

In this world, data is of various formats. It is never the same. Data is unique and this being the reason, there are lot of extensions and plug-ins being developed for existing tools and technologies. The data is a combination of multiple forms. There is no industry or domain for that matter no company that relies on single type of data. Google will hit ZB's of data very soon. This increasingly enormous amount of data cannot be handled and managed just by these traditional databases like RDMS and cannot be worked simply with SQL. Only little amount of data can be handled and processes. Now No-SQL is ruling the world of unstructured data for storage, warehousing and analytics and Hadoop is being the most relevant technology for handling data. Data management, data warehousing, and data analytics with no doubt are the most important things of 21st century and Microsoft has been a pioneer in it ever since.

IT professionals from Microsoft Business intelligence who use ETL, RDMS and SQL and SSRS for Reporting consider this as an end of the ecosystem. But it is just mainstream of IT circle in the world of data. Business that runs on Hadoop are almost all the top 10 companies in the world like Facebook, Yahoo, Twitter. Hadoop is used extensively by data giants which make use of heavy data, like Microsoft, Informatica and Teradata etc. Challenging part of this whole scenario would be to make use of Hadoop by bringing data into the Mainstream IT which is in general like warehousing the data, analytics performed on it and finally reporting it. This methodology is required to do activities like data modelling, data profiling, data cleansing, ETL, and Data Warehouse (or) data Marts.

How MSBI can be used instead of Big data tools to organise and warehouse small to medium sized data sets?

Regularly we load data warehouses with multiple years' worth of data at a time. If the data is huge, we definitely should look to Big data processes. Inside MSBI stack, SSIS supports batch processing very decent enough with the components already existing in the SSIS toolbox which is present in the integration project. Also, if the data is coming in the form of data files but they are huge in number, we would definitely want to load the data in the order of files being dropped in the source location.

The question here would be as to how it will pick according to the order they have been dropped in the pick-up location. The answer would be Created date or modified date. So, the files to be loaded have to be picked up according to either of those dates. For that purpose, MSBI has given the option to use script task i.e. to use C# code to pick files according to those dates and store them one by one in a local variable and process them hence forth.

The code that we can use to sort the files as per date dropped and then pick up is given in Appendix-A.

IV. DATA MINING IN SOCIAL NETWORKING MEDIA

Why do you think data mining is having such importance in today's world? You have seen astonishing numbers – the volume of data produced is doubling every two years. Unstructured data alone is about 90 percent of the digital universe. But more amount of information does not mean more knowledge or insights.

Data mining allows you to:

1. Sweep through all the chaotic and repetitive noise in your data. Understand what is relevant and then make good use of that information to assess likely outcomes.
2. Accelerate the pace of making informed decisions.

Social Networks: As social networking platforms become larger and more complex data handling platforms, reasoning about social dynamics and data mining via simple statistics is logical explanation though it might not be very intuitive. Visualization provides a natural way to summarize the information in order to make it much easier to understand. In recent times, we have seen an intersection of social networking analytics and visualization blended with data mining which is changing the way analysts understand and characterize social networks. The main goal of data mining is discussed in the context of business understanding and interaction. The chapter also examines how different metaphors are aimed towards elucidating different aspects of social networks, such as structure and semantics. A number of methods are described, where analytics and visualization are interwoven towards providing a better comprehension of social structure and dynamics.

Social Media provides a wealth of social network data, which can be mined in order, to discover useful business applications. Data mining techniques provide researchers and practitioners the tools needed to analyse large, complex, and frequently changing social media data. An overview on the topic of data mining in social media is provided in Chapter 12. This chapter introduces the basics of data mining in the context of social media and discusses how to mine social media data. The chapter also highlights number of illustrative examples with an emphasis on social networking sites and blogs.

Text Mining in Social Networks: Social networks contain a lot of text in the nodes in various forms. For example, social networks may contain links to posts, blogs or other news articles. In some cases, users may tag one another, which is also a form of text data on the links. The use of content can greatly enhance the quality of the inferences which may be made in the context of graphs and social networks. In chapter 13, we present methods for using text mining techniques in social networks in the context of a variety of problems such as clustering and classification.

V. USING AI FOR SOCIAL MEDIA DATA MINING

Social media data mining powered by AI and cognitive technologies can provide even more powerful intelligence from the information gathered from social media. The key is its understanding of language, meaning and context. To capture the unique and personal ways that customers express themselves on social media requires understanding the nuanced locution, influence and consequence expressed in language.

The Cogito cognitive technology understands the unique aspects of the way users communicate on social media such as through slang, jargon, acronyms and abbreviations. In this way, the technology helps users hear and understand the voices on social media through a comprehension of content, context, intent and sentiment expressed in information.

To get an idea of the reach of social media, consider that, in the 2016 US presidential election, more than one billion election related tweets were posted on Twitter from the first presidential debate until the day before the election. Here, it became the go-to place for conversations and reactions about breaking news and sharing opinions. Therefore, it's no surprise that social media data mining software is being applied in many areas. Companies, political parties, social and religious groups and others exploit the conversations and comments shared on social networks to gather information and intelligence to fuel research on markets, competitors, customers, competitors and more.

Recently there has been rapid growth of text data in the context of different web-based applications such as social media, which often occur in the context of multimedia or other heterogeneous data domains. Therefore, a number of techniques have been recently been designed for the joint mining of text data in the context of these different kinds of data domains. For example, the we contains text and image data which are often intimately connected to each other and these links can be used to improve the learning process from one domain to another.

Similarly, cross-lingual linkages between knowledge from one language domain to another. This is closely related to the problem of transfer learning.

VI. CONCLUSION AND PROSPECTIVE SCOPE FOR RESEARCH

The rapid growth of online textual data creates an urgent need for powerful text mining techniques. As an interdisciplinary field, text data mining spans multiple research communities, especially data mining, natural language processing, information retrieval and machine learning with applications in many different areas. Many models and algorithms have been developed for various text mining tasks.

- 1) Scalable and robust methods for natural language understanding.
- 2) Domain adaptation and transfer learning.
- 3) contextual analysis of text data
- 4) Parallel text mining

APPENDIX-A

#region Introduction to the script task

/ The Script Task allows you to perform virtually any operation that can be accomplished in
* a .Net application within the context of an Integration Services control flow.
*
* Expand the other regions which have "Help" prefixes for examples of specific ways to use
* Integration Services features within this script task. */*

/ Expand the other regions which have "Help" prefixes for examples of specific ways to use
* Integration Services features within this script task. */*

#endregion

#region Namespaces

using System;

using System.Data;

using Microsoft.SqlServer.Dts.Runtime;

using System.Windows.Forms;

#endregion

using System.IO;

using System.Linq;

namespace ST_e2be047b755848b5ba47f9e883153087

{

/// <summary>

*/// ScriptMain is the entry point class of the script. Do not change the name, attributes,
/// or parent of this class.*

/// </summary>

[Microsoft.SqlServer.Dts.Tasks.ScriptTask.SSISScriptTaskEntryPointAttribute]

public partial class Script Main: Microsoft.SqlServer.Dts.Tasks.ScriptTask.VSTARTScriptObjectModelBase

{

#region Help: Using Integration Services variables and parameters in a script

/ To use a variable in this script, first ensure that the variable has been added to*

** either the list contained in the ReadOnlyVariables property or the list contained in*

** the ReadWriteVariables property of this script task, according to whether or not your*

** code needs to write to the variable. To add the variable, save this script, close this instance of*

** Visual Studio, and update the ReadOnlyVariables and*

** ReadWriteVariables properties in the Script Transformation Editor window.*

** To use a parameter in this script, follow the same steps. Parameters are always read-only.*

** Example of reading from a variable:*

** DateTime startTime = (DateTime) Dts.Variables["System::StartTime"].Value;*

** Example of writing to a variable:*

** Dts.Variables["User::myStringVariable"].Value = "new value";*

** Example of reading from a package parameter:*

** int batchId = (int) Dts.Variables["\$Package::batchId"].Value;*

** Example of reading from a project parameter:*

** int batchId = (int) Dts.Variables["\$Project::batchId"].Value;*

** Example of reading from a sensitive project parameter:*

** int batchId = (int) Dts.Variables["\$Project::batchId"].GetSensitiveValue();*

** */*

#endregion

#region Help: Firing Integration Services events from a script

/ This script task can fire events for logging purposes.*

** Example of firing an error event:*

** Dts.Events.FireError(18, "Process Values", "Bad value", "", 0);*

** Example of firing an information event:*

** Dts.Events.FireInformation(3, "Process Values", "Processing has started", "", 0, ref fireAgain)*

** Example of firing a warning event:*

** Dts.Events.FireWarning(14, "Process Values", "No values received for input", "", 0);*

** */*

#endregion

#region Help: Using Integration Services connection managers in a script
/ Some types of connection managers can be used in this script task. See the topic*
** "Working with Connection Managers Programmatically" for details.*

```
*
* Example of using an ADO.Net connection manager:
* object rawConnection = Dts.Connections["Sales DB"].AcquireConnection(Dts.Transaction);
* SqlConnection myADONETConnection = (SqlConnection)rawConnection;
* //Use the connection in some code here, then release the connection
* Dts.Connections["Sales DB"].ReleaseConnection(rawConnection);
*
* Example of using a File connection manager
* object rawConnection = Dts.Connections["Prices.zip"].AcquireConnection(Dts.Transaction);
* string filePath = (string)rawConnection;
* //Use the connection in some code here, then release the connection
* Dts.Connections["Prices.zip"].ReleaseConnection(rawConnection);
* */
```

#endregion

```
/// <summary>
/// This method is called when this script task executes in the control flow.
/// Before returning from this method, set the value of Dts.TaskResult to indicate success or failure.
/// To open Help, press F1.
/// </summary>
```

```
public void Main()
{
```

```
// TODO: Add your code here
string partialName = "SAMPLE_NAME_OF_FILE";
var directory = new DirectoryInfo(Dts.Variables["User::src_directory"].Value.ToString());
FileInfo[] files = directory.GetFiles("*" + partialName + "*");
```

```
var recentFile = files.OrderByDescending(f => f.LastWriteTime).ToList().FirstOrDefault();
```

```
var currentDate = recentFile.LastWriteTime;
Dts.Variables["User::TEMPVARIABLE"].Value = recentFile.ToString();
}
```

#region ScriptResults declaration

```
/// <summary>
/// This enum provides a convenient shorthand within the scope of this class for setting the
/// result of the script.
///
/// This code was generated automatically.
/// </summary>
```

```
enum ScriptResults
{
    Success = Microsoft.SqlServer.Dts.Runtime.DTSExecResult.Success,
    Failure = Microsoft.SqlServer.Dts.Runtime.DTSExecResult.Failure
};
```

#endregion

}

So, with the help of such code and processes, we can load small to medium sized data into the destination databases by diving them into batches and loading them according to date of creation or modification hassle free.

REFERENCES

- Harvard Business Review. (2018). *HBR Guide to Data Analytics Basics for Managers*. Harvard Business Review Press. Print.
- Bagha, Ashadeep. (2018, 7 September). *Big Data Analytics: A Hands-On Approach*. VPT India. Print.
- Reed, Jeff. (2017). *Data Analytics: Applicable Data Analysis to Advance Any Business Using the Power of Data Driven Analytics*. Web.
- Goyal, Niraj. 2019. *How Data Analytics Backed By AI And ML Is Transforming The BFSI Sector*. Retrieved from <https://www.analyticsindiamag.com/data-analytics-ai-ml-bfsi/>
- Varone, Marco. (2017, 17 June). Social Media Data Mining. Retrieved from <https://www.expertsystem.com/social-media-data-mining/>