Big Data Analysis

SHREY PAREKH
B.Tech IT,
Smt. Kundanben Dinsha Patel Department of Information Technology.
Rajkot, India.
shreyparekh1199@gmail.com

Abstract

In a world where information is everything and people are making billions only by selling information what is important? And the answer is data which is the source of all the information that is being sold. This is an era where data is being produced in enormous amounts and we need a place to store such a big amount of being produced every second, which is a problem. Then there is a term in the market now a days that is "Big Data". Big data is not just any normal dataset but it is a dataset that consists a huge amount of data that is produced and needs to be handled using proper techniques or else it is of no use. Handling such data can benefit many businesses and help them grow fast with lower amount of risks and improve their decision making. The paper aims to analyze different ways and techniques that can be used on big data to make such decisions.

Keywords: Big data, Analysis, Power Bi, Decision Making.

I. INTRODUCTION

What would happen if we tend to produce data at a speed we are doing and not be able to store the data or handle it in a way it is useful? Yes, that is a problem that mankind is going to face as the amount of data we are producing is increasing at an enormous speed and we are already facing problems handling such a big amount of data.

According to Visual capitalist companies survey these is the amount of data being produced every minute in 2018 and 2019 respectively:

Platform	2018 (Per minute)	2019 (Per minute)
Facebook	973,000 logins	1M+ logins
Messages	18M texts sent	18.1M texts sent
Instagram	174,000 scrolling	347,222 scrolling
Google	3.7M searches	3.8M searches
Netflix	266,000 hours watched	694,444 hours watched

Report 2018

Report 2019

The amount of data we are producing every minute in 2019 is way more than we did in 2018 and we are already in 2020 so just imagine that if the speed at which the data is being produced is increasing every year then we definitely require a different way to handle such huge amount of data. The important thing about data is the information we obtain out of that data and not data itself. So just storing data does not solve problem we actually need to analyse the data and get a meaning out of it that is what we call information. Data can be names, numbers, images etc.. While information is name of customers, their phone numbers, their photographs that adds meaning to our data. Now that we are slave of internet imagine the amount of information that we can extract the information out of the data that we can store and analyse.

II. ANALYSIS

What if we only had "Data" and no sense of what to do next? Then all of the Data that we would be having is useless. So this is where analysis comes into picture if we don't analyse the data we have using proper techniques and method then data is of no use. It is very important to analyse the data properly for example we have alphabets, 2 digit numbers, 10 digit number etc and thousands and thousands of such rows stored in our database and we are not able to figure out the meaning of such data, it is useless. So using proper analysis techniques we can get to know the meaning of what we have alphabets are names of people, 2 digit numbers are age, 10 digit numbers are phone numbers thus we get something out of our information. There are different tools available to handle such enormous amount of data and analyse it at the same time and get information out of it. Analysing such "Big Data" is called "Big Data Analysis". Tools available to analyse data are as follows:

- Power BI
- IBM Cognos Analytics
- Qlik Sense
- TIBCO Spotfire
- So on…

Analysing data is one of the most crucial and important part of Big Data.

III. POWER BI

Power BI is a tool by Microsoft which helps us convert data to information and also makes it easy to even visualise the data more effectively to understand it. There are certain features that are provided by Power BI and they are as follows:

- Data protected using End-to-End protection
- Various services can directly be used
- Data can be used directly from 120+ free connectors
- Collaborate with colleagues directly

Power BI gives so many features that are useful to a greater extent. Using Power BI we can directly import data from any site, we can import live data from the place it is being hosted, there are different sources from where we can get our data from and visualise data in different ways as we wish and create different graphs that can help understand data in a better way. There are features that provides us SQL servers and Data sources as follows:

- DATA SOURCES
    - Excel
    - XML
    - JSON
    - Oracle Database
    - Azure
    - Power BI Datasets
    - Etc..
- ONLINE SERVERS
    - Google analytics
    - Dynamic NAV
    - Facebook
    - GitHub
    - SparkPost
    - Salesforce Reports
    - Etc..

IV. BIG DATA ANALYSIS

"Big Data" is a new term, the size of the datasets in Big data are usually so huge that is very difficult to work with and makes it tough to work with as such a huge data makes ones work tedious. Our traditional methods can not help us making our work easy. The datasets of Big Data are so big that they cannot be stored, managed or capture using our old tools and software.

Big data in terms of size are larger than usual datasets they are in tera-bytes(TB), Petra-bytes(PB) in a single dataset. Others problems such as storage, search etc still prevails. It

becomes so difficult to manage such huge data then it would be so difficult to analyse such a huge dataset. Companies are dealing with such huge datasets nowadays and it becomes difficult for them to manage. Hence, big data needs an advanced technique to be applied on big datasets. This analysis on big dataset help businesses to take better decisions. Thus the business will experience change and can benefit out of that change.

1. Characteristics of Big Data

The term "Big Data" have different diversity, scale, distribution and timeliness that needs to have various techniques and tools to get 100% output from it. There are 3 v's that are very important in Big Data that are:

- Volume
- Velocity
- Variety

Volume: It is the size of datasets and how enormously large the datasets are.

Velocity: It refers to the rate at which the data is being created and changed.

Variety: It states various data formats and types being also the various ways of analysing the data.

The size of Big Data is always in TBs or PBs, the numbers are enormously large. The source of Big Data has also changed, now sources provide various format where data can be in log form, tables, records, clicks, transactions and social platforms etc. Now the data will not be available in common form it will be a mixture of different forms and we need to analyse such data. Thus, the term "Variety" is also important as "Volume" in Big Data.

Frequency at which data is being created or being destroyed also matters a lot in Big data as "Velocity" is also third and an important wheel of 3 V's. There is one more V that is introduced to it Veracity that means the quality of the data i.e whether the data is ambiguous, incomplete, inconsistent, approximate etc.

2. Analytics Tools and Methods

With the increasing multitudes of data that is being produced, transferred or destroyed in organizations every hour we need improved and new ways to analyse such data. Only large number of datasets will not help making decisions.

Handling large datasets is not an easy task to do so we need to move on to some new methods leaving the old techniques behind. We need new tools that are specially designed to manage and analyse Big Data. There are different techniques available now a days one of them is "Big Data Analytics and Decisions" which will help integrate data and methods to analyse it and make it useful for decision making. There are mainly three fields that are to be focused on in Big Data analysis that are:

1. Storage and architecture
2. Analytics processing

A. Storage and Architecture

First and foremost is where and how to store such enormous data? Once the organization acquires the data it needs to manage the data until and unless they analyse it so for that there are traditional methods that they use like data marts and warehouses. Data is uploaded then transformed to fit according to the needs and then they are loaded into the data warehouses or database. Thus data is cleaned, transformed and categorised.

But for Big Data analysis there are different set of techniques that is MAD – Magnetic, Agile, Deep analysis that is different than the traditional methods. Traditional methods require cleansing first which is a difficult task in terms of Big Data as the data is not easy to clean as it contains various formats of data, it also requires a speed that can match the speed of data being produced. Thus here comes Agile database that can easily sync the data and adapt it easily, Then they require to analyse using complex statistical methods and analysts study large data by drilling it up and down.

Using massive parallel processing that provide high query performance and platform scalability. There are non-relational databases which were developed to store unstructured or non-relational data base.

B. Analytics processing

There are many requirements they are fast data loading, fast quey processing, high utilization of space available and last is dynamic workload patterns. This data is no normal data so it needs to be treated in a different way.

There is a method named Map Reduce which is base for Hadoop. The idea is to break the task into sub tasks and parallelly execute those sub tasks which can reduce the time required to perform the task. "Map" will divide the tasks and bring individual output by operating on them and then "Reduce" will then collect all these outputs and provide results. Hadoop operates on that and helps managing data.

References

1. Elgendy, Nada & Elragal, Ahmed. (2014). Big Data Analytics: A Literature Review Paper. Lecture Notes in Computer Science. 8557. 214-227. 10.1007/978-3-319-08976-8_16.

Turnitin Originality Report

Processed on: 31-Oct-2020 22:21 IST
ID: 1430884086
Word Count: 1983
Submitted: 4

final By shrey parekh

Similarity Index	Similarity by Source	
0%	Internet Sources:	0%
	Publications:	0%
	Student Papers:	N/A

SHREY PAREKH B.Tech IT, Smt. Kundanben Dinsha Patel Department of Information Technology. Rajkot, India. shreyparekh1199@gmail.com Big Data Analysis Abstract In a world where information is everything and people are making billions only by selling information what is important? And the answer is data which is the source of all the information that is being sold. This is an era where data is being produced in en ormous amounts and we need a place to store such a big amount of being produced every second, which is a problem. Then there is a term in the market now a days that is "Big Data". Big data is not just any normal dataset but it is a dataset that consists a huge amount of data that is produced and needs to be handled using proper techniques or else it is of no use. Handling such data can benefit many businesses and help them grow fast with lower amount of risks and improve their decision making. The paper aims to analyze different ways and techniques that can be used on big data to make such decisions. Keywords: Big data, Analysis, Power Bi, Decision Making. I. INTRODUCTION Wha t would happen if we tend to produce data a t a speed we a re doing and not be able to store the data or handle it in a way it is useful? Yes, that is a problem that mankind is going to fa ce as the amount of data we a re producing is increa sing at a n enormous speed a nd we a re a lrea dy fa cing problems ha ndling such a big a mount of data. According to Visual ca pitalist companies survey these is the a mount of data being produced every minute in 2018 and 2019 respectively: Pla tform 2018 (Per minute) 2019 (Per minute) Fa cebook 973,000 logins 1M+ logins Messa ges 18M texts sent 18.1M texts sent Instagram 174,000 scrolling 347,222 scrolling Google 3.7M searches 3.8M searches Netflix 266,000 hours wa tched 694,444 hours wa tched Report 2018 Report 2019 The a mount of data we are producing every minute in 2019 is wa y more than we did in 2018 a nd we a re a lready in 2020 so just ima gine that if the speed a t which the data is being produced is increasing every year then we definitely require a different wa y to ha ndle such huge a mount of da ta. The important thing a bout data is the information we obtain out of tha t data and not data itself. So just storing data does not solve problem we a ctually need to a nalyse the data and get a mea ning out of it that is what we call information. Data can be na mes, numbers, images etc.. While information is name of customers, their phone numbers, their photographs that adds mea ning to our data. Now that we are slave of internet imagine the a mount of information that we can extract the information out of the data that we can store a nd a nalyse. II. ANALYSIS Wha t if we only had "Data" and no sense of what to do next? Then a ll of the Data that we would be having is useless. So this is where a nalysis comes into picture if we don't a nalyse the data we have using proper techniques and method then data is of no use. It is very important to analyse the data properly for example we ha ve alphabets, 2 digit numbers, 10 digit number etc a nd thousands a nd thousands of such rows stored in our da tabase and we a re not a ble to figure out the mea ning of such data, it is useless. So using proper a nalysis techniques we ca n get to know the meaning of what we have a lphabets a re names of people, 2 digit numbers a re age, 10 digit numbers are phone numbers thus we get something out of our information. There a re different tools a vailable to ha ndle such enormous amount of da ta a nd a nalyse it at the sa me time and get information out of it. Analysing such "Big Da ta" is ca lled "Big Da ta Ana lysis". Tools a vailable to a nalyse data are as follows: - Power BI

- IBM CognosAnalytics - Qlik Sense - TIBCO Spotfire - So on… Ana lysing data is one of the most crucial a nd important pa rt of Big Da ta. Power BI is a tool by Microsoft which helps us convert da ta to information and also makes it ea sy to even visualise the da ta more effectively to understand it. There a re certain fea tures that are provided by Power BI and they are as follows: - Da ta protected using End-to-End protection - Va rious services can directly be used - Da ta can be used directly from 120+ free connectors - Colla borate with colleagues directly Power BI gives so many features that are useful to a greater extent. Using Power BI we ca n directly import data from any site, we ca n import live data from the place it is being hosted, there a re different sources from where we ca n get our data from a nd visualise da ta in different wa ys a s we wish and crea te different gra phs that can help understand data in a better wa y. There are features that provides us SQL servers and Data sources a s follows: - DATA SOURCES o Excel o XML o JSON o Ora cle Da tabase o Azure o Power BI Da tasets o Etc.. - ONLINE SERVERS o Google a nalytics o Dynamic NAV o Fa cebook o GitHub o Spa rkPost o Sa lesforce Reports o Etc.. III. POWER BI IV. BIG DATA ANALYSIS "Big Da ta " is a new term, the size of the datasets in Big da ta are usually so huge that is very difficult to work with and ma kes it tough to work with as such a huge data makes ones work tedious. Our traditional methods can not help us making our work ea sy. The datasets of Big Da ta a re so big that they ca nnot be stored, managed or ca pture using our old tools and software. Big da ta in terms of size a re la rger than usual datasets they a re in tera -bytes(TB), Petra-bytes(PB) in a single dataset. Others problems such as storage, search etc still preva ils. It becomes so difficult to manage such huge data then it would be so difficult to analyse such a huge dataset. Companies are dea ling with such huge da tasets nowadays a nd it becomes difficult for them to ma nage. Hence, big da ta needs an a dvanced technique to be applied on big da tasets. This a nalysis on big dataset help businesses to take better decisions. Thus the business will experience change a nd can benefit out of that change. 1. Characteristics of Big Data The term "Big Data" have different diversity,scale, distribution and timeliness that needs to have various techniquesandtoolstoget100%outputfromit.Thereare3 v'sthatareveryimportantinBigDatathatare: - Volume - Velocity - Va riety Volume: It is the size of datasets a nd how enormously large the da tasets a re. Velocity: It refers to the ra te at which the data is being created a nd changed. Va riety: It states various data formats and types being a lso the va rious wa ys of a nalysing the data. The size of Big Da ta is a lways in TBs or PBs, the numbers a re enormously la rge. The source of Big Da ta ha s a lso cha nged, now sources provide various format where data can be in log form, ta bles, records, clicks, tra nsactions and social pla tforms etc. Now the da ta will not be a vailable in common form it will be a mixture of different forms and we need to a nalyse such data. Thus, the term "Variety" is a lso important a s "Volume" in Big Da ta. Frequency at which data is being created or being destroyedalsomattersalotinBigdataas"Velocity"isalso thirdandanimportantwheelof3V's.ThereisonemoreV thatisintroducedtoitVeracitythatmeansthequalityofthe data i.e whether the data is ambiguous, incomplete, inconsistent,approximateetc. 2. Analytics Tools and Methods With the increa sing multitudes of da ta tha t is being produced, transferred or destroyed in organizations every hour we need improved and new ways to analyse such data. Only la rge number of datasets will not help making decisions. Ha ndling large datasetsis not an easy task to do so weneed to move on to some new methods leaving the old techniques behind. We need new tools that a re specially designed to ma nage and a nalyse Big Da ta. There are different techniques a vailable now a days one of them is "Big Da ta Analytics and Decisions" which will help integra te da ta a nd methods to a nalyse it a nd make it useful for decision making. There are ma inly three fields that a re to be focused on in Big Data a nalysis that a re: 1. Storage and architecture 2. Ana lytics processing A. Storage and Architecture First a nd foremost is where a nd how to store such enormous da ta? Once the orga nization a cquires the da ta it needs to manage the data until a nd unless they a nalyse it so for tha t there a re traditional methods that they use like data ma rts a nd warehouses. Da ta is uploaded then transformed to fit a ccording to the needs and then they a re loaded into the da ta wa rehouses or da tabase. Thus da ta is cleaned, tra nsformed and categorised. But for Big Da ta a nalysis there a re different set of techniques that is MAD – Magnetic, Agile, Deep analysis that is different than the traditional methods. Tra ditional methods require clea nsing first which is a difficult task in terms of Big Da ta a s the data is not ea sy to clea n a s it

contains various formats of data, it a lso requires a speed that can match the speed of da ta being produced. Thus here comes Agile da tabase that can easily sync the data and adapt it easily, Then they require to a nalyse using complex statistical methods and a nalysts study la rge data by drilling it up a nd down. Using ma ssive parallel processing that provide high query performance a nd pla tform sca la bility. There a re non- rela tional da tabases which were developed to store unstructured or non-relational data base. B. Analytics processing There a re many requirements they a re fast data loading, fa st quey processing, high utiliza tion of space a vailable and la st is dynamic workload patterns. This data is no normal data so it needs to be treated in a different wa y. There is a method named Map Reduce which is ba se for Ha doop. The idea is to brea k the ta sk into sub ta sks and pa ra llelly execute those sub tasks which ca n reduce the time required to perform the task. "Map" will divide the tasks and bring individual output by operating on them a nd then "Reduce" will then collect a ll these outputs and provide results. Ha doop operates on that and helps managing data. References 1. Elgendy, Na da & Elra gal, Ahmed. (2014). Big Data Ana lytics: A Litera ture Review Pa per. Lecture Notes in Computer Science. 8557. 214-227. 10.1007/978-3- 319-08976-8_16.