International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Big Data Analytics

Ms.Nidhi Sharma & Ms. Hina
[1,2]Assistant Professor, Department of Computer Science & Applications, S.D. College (Lahore), AmbalaCantt
[1]nidhi.sdc@gmail.com, [2]hina.harman@gmail.com

Abstract

Decision making on various issues plays very crucial role in business. In business scenario data is gathered and used for decision making, so a large amount of data is produced day by day. Data produced has no specific structure or format so it contributes to big data. To extract the valuable data from very huge amount of datasources available requires Big Data Analysis.In this paper we will discuss the role of Big Data Analysis in gathering effective data from various data sources.

Keyword

Big Data Analysis, Business Case Evaluation, Data Identification, Data Acquisition & Filtering, Data Extraction, Data Validation & Cleansing, Data Aggregation & Representation, Data Analysis, Data Visualization, Utilization of Analysis Results.

1. Introduction

In today's era Technology becomes a boon for business as it enhanced the quality of decision making. Demand of Technology in Business Organisation is increasing rapidly. Also these organisations are working hard to adapt Technology because it is the future. A large amount of data is being produced day by day by these organisations. This data must be analysed for quality data production and benefits of data. Because it is very tedious job to make use of large amount of data available. Data Warehouses are used to store data and data mining techniques are applied to that data to get the quality product or data.

2. Benefits of Big Data and Analytics

Analysis of Business data has become a topic of great interest. Big data and analytics can be beneficial for Business organisations as:

2.1 Business users can improve decision making which leads of greater success rate.
2.2 Transparent data is available so various day to day risks can be managed.
2.3 Various innovative models can be produced because of access to huge amount of data.
2.4 Usage of Technology will give assistance to learn and solve complex issues.
2.5 Effective response will increase the business quality production.

3. Phases of Big Data Analysis

Big Data has mainly 3 characteristics: Volume - large amount of data, Velocity - large amount of data production rapidly, Variety - various kind of data.

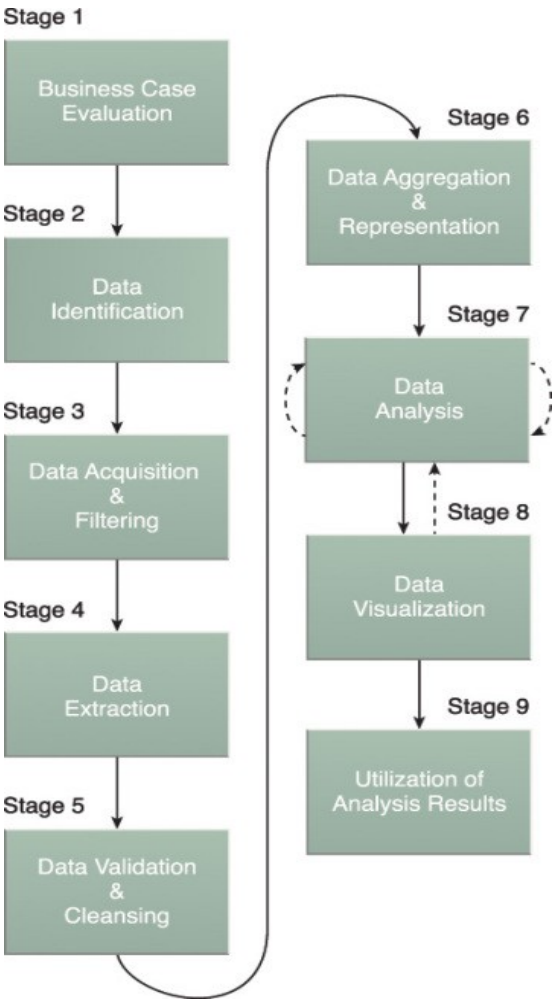

Fig. 1

So, it differs from tradition data hence a step by step methodology is needed to fulfil the requirements for Big

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

Data Analysis. To adopt Big Data in Higher Education training, education and staffing must be considered. The Big Data analytics lifecycle is as follows:

3.1 Business Case Evaluation
3.2 Data Identification
3.3 Data Acquisition & Filtering
3.4 Data Extraction
3.5 Data Validation & Cleansing
3.6 Data Aggregation & Representation
3.7 Data Analysis
3.8 Data Visualization
3.9 Utilization of Analysis Results

3.1 Business Case Evaluation

A well-defined business case is required to start with Big Data analytics lifecycle. A business case helps to recognise the resources to be used and what are the challenges to be tackles. Assessment criteria can be used to guide evaluation of analytic results. A case must be created, assessed and approved before proceeding further analysis.

3.2 Data Identification

This stage tells about the data set requirements and their sources for the project analysis. Collecting data from various sources helps to find the hidden pattern and missing information. It is always favourable to relate data sources to collect more accurate data. The nature of business problems to be addressed can be defined depending upon the business scope and the source of these problems can be internal or external.

- When internal datasets are used then the pre-defined data sets are matched with the list of available datasets from internal sources and compiled.
- When external datasets are used then a list of third-party data providers are compiled. Data from external sources may be obtained from blogs, other types of content-based web sites etc.

3.3 Data Acquisition and Filtering

In this phases Data is acquired from various sources which was identified in last phase. Acquired data is then filtered to remove corrupted data or the data which has no values for analysis objective. Because data is gathered from different sources and is of different forms so filtering is required to remove some or most of the data that may be duplicated or not relevant.

As shown in fig, to improve the data querying, metadata can be added to the data obtained from internal and external sources. Example of metadata includes size, structure, source, date/ time of creation etc. Metadata helps to preserve data quality and accuracy throughout the Big Data analytics lifecycles.

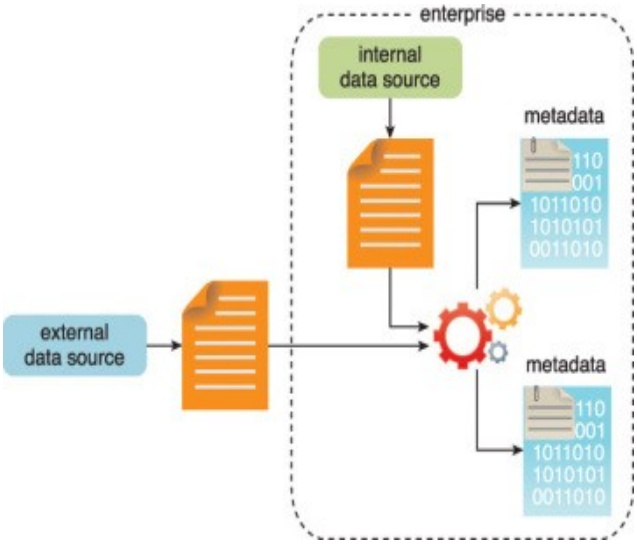

Fig. 2

3.4 Data Extraction

As data is collected from various sources so the data collected may be in incompatible format. There is a need to extract the data from these incompatible format to compatible format. So, the next stage is Data Extraction for extracting data from incompatible sources and transforming it into desired format for the analysis purpose.

Type of analytics of Big Data Solution specifies the extraction and transformation of data. Because if the underlying Big Data solution can access those files directly then there is no need of extracting data. It can be seen in Fig 4 how comments and user ID is extracted from an XML document.

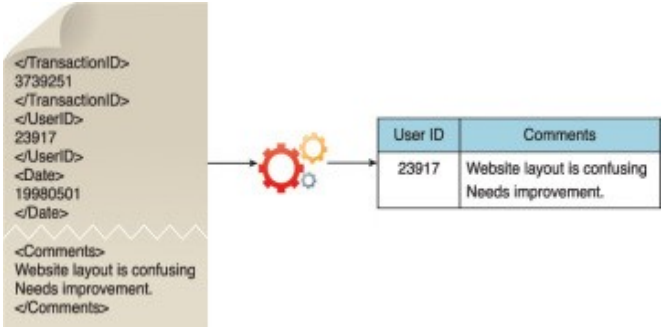

Fig. 3

To separate the data into separate fields, further transformation is required by the Big Data solution.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue-17
December 2017

3.5 Data Validation and Cleansing

Data Validation is one of the necessary steps in analytic process as invalid data may lead to false analysis results. Set of suitable validation constraints are required to validate the data. Big data solution often receive data that is invalid or missing, so this phase is dedicated to validate data of fill in the missing valid data using interconnected data sets. The Data Validation and Cleansing stage is as shown in fig. The Data validation is done in the following steps:

3.5.1 Very first value is Database B is checked against its corresponding value in Database A.
3.5.2 If verified next values are not cross checked.
3.5.3 If any value found missing, its corresponding value is filled from Database A.

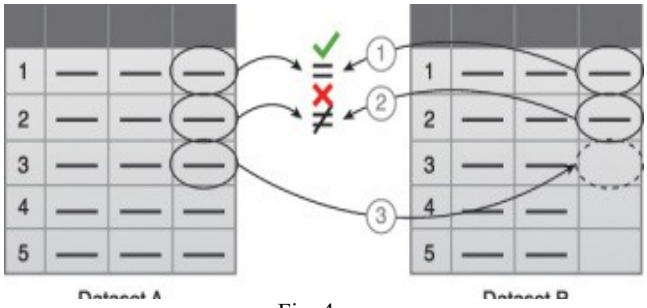

Fig. 4

3.6 Data Aggregation and Representation

Data may be spread across multiple datasets, joined together via some common fields, for example ID used as foreign key in some other data source. In some cases same data is populated in multiple data sets. The method of data reconciliation is required to represent the correct value.Integrating multiple databases together to see unifies view results in Data Aggregation and Representation. This stage is very complicated because of differences in:

1 Data Structure- Data model may be different even if data format is same.
2 Semantics-Same value is represented by different names in different datasets. For example surname in one data set may be defined as last name in other data set.

Data Aggregation is a time consuming and effort-intensive task for large data sets. Data Aggregation is as illustrated in fig.

Fig. 5

Fig 6 represents an example where same piece of data (DOB) is stored in two different formats. Dataset A contains the desired piece of data, but it is part of a BLOB that is not readily accessible for querying. Dataset B contains the same piece of data organized in column-based storage, enabling each field to be queried individually.

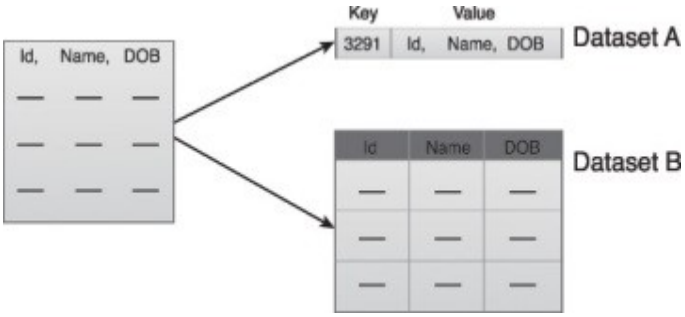

Fig. 7

3.7 Data Analysis

Data Analysis phases is concerned with actual analysis task that involves one or more analysis. This stage may be iterative when case analysis is repeated. According to the type of analytic result required, this stage can be as simple as querying a dataset or as complicated as combining data mining and complex statistical analysis techniques. This phase can be divided into confirmatory or exploratory analysis as sown in fig

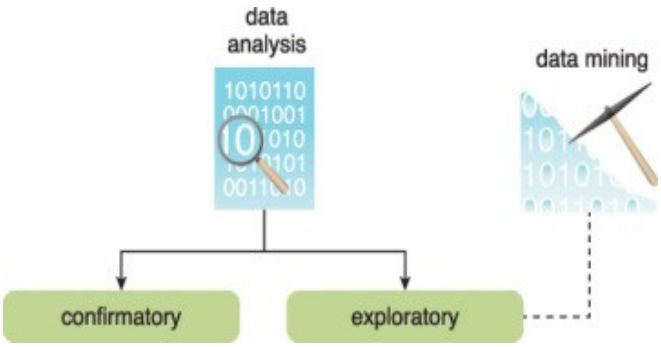

Fig. 8

1 Confirmatory data analysis is a deductive approach where the cause of phenomenon is already assumed and called hypothesis. Then data is analysed to prove or disprove this hypothesis.
2 Exploratory data analysis is an inductive approach where no hypothesis is generated. It is associated with data mining. Data is obtained through analysis to develop the cause of phenomenon.

3.8 Data Visualization

This stage is concerned with using data visualization techniques and tools to represent the analysis results in graphical format for effective interpretation. Users must be able to understand the result to obtain correct and useful value from the analysis provide the feedback (denoted as dotted lines in figure).

3.9 Utilization of Analysis Results

The analysis results are made available to users to support decision making and these results may further be used. This stage is dedicated to determine where and how the results can be used in effective way. Depending upon the nature of analysis problem, the analysis results can produce modules that encapsulates the understanding of relationship within the data analysed.

4. Conclusion

There is no question of doubt that the emergence of big data has a positive impact on our lives. Social media is one such application where big data is being used. Big data analysis are being used not only in business organizations but in other fields too like health sciences, higher education, communication, media and entertainment industry etc. the fundamental component of Big Data in higher education is learning analytics. By performing analysis on student data, predictive models can be created to examine student performance, helping the instructors to adapt or modify their methodology and enabling the continuous assessment. It is being used to analyse large amount of information for faster and efficient decision making in various fields. It also faces many challenges that need to be solved. For e.g. integrating and interoperating big data across different departments and organizations is the biggest challenge in government. Issues like privacy and personal data protection of big data used for educational purposes is another challenge big data is facing today.

5. References

[1] Big Data:The next frontier for innovation, competition and productivity.JamesManiyka, Executive summary ,McKinsey Global Institute ,May 2011, <http://www.mckinsey.com/mgi/publication/big.data/MGI_big_data_exec_summary.pdf>.
[2] IBM,"Real World PredcitiveAnalysis:Putting Analysis into Action for Visible Results",2010,http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03112usen/yt03112USEN.PDF.
[3] Game Changers, Chapter 4 Linda Baer and John Campbell, EDUCAUSE publications May 2011 Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.
[4] Penetrating the fog: Analytics in learning and education, Phil Long and Gorge Siemens, EDUCAUSE Review vol 46,no 5 September/October 2011 .
[5] Kantardzic, M., *Data mining: concepts, models, methods, and algorithms*2011: John Wiley & Sons.
[6] M'Hammed, A., H. Wu, and Y. Cherng- Jyh, *Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade.* Journal of Educational Technology & Society, 2012. **15**(3): p. 77-88.
[7] Ramesh, V., P. Parkavi, and K. Ramar, *Predicting student performance: a statistical and data mining approach.* International Journal of Computer Applications, 2013. **63**(8): p. 35-39.
[8]Agrawal, R., Imielinski, T. & Swami, A. (1993). Mining Associations Between Sets of Items in
Massive Databases. Proceedings of the ACM-SIGMOD 1993 Int'l Conference on Management of Data,
Washington D.C., May 1993.
[9]Ali, L., Adasi, M., Gasevic, D., Jovanovic, J. &Hatala, M. (2013). Factors influencing beliefs for adoption of a
learning analytics tool: an empirical study. *Computers & Education*, *62*, 130–148.
[10] Baepler, P. & Murdoch, C. J. (2010). Academic analytics and data mining in higher education. *International*
Journal for the Scholarship of Teaching and Learning, *4*, 2, 1–9. Retrieved September 17, 2014, from http://
digitalcommons.georgiasouthern.edu/ij-sotl/vol4/iss2/17