IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

Data Modeling and Data Analytics Lifecycle

Manisha R Gupta
Department of Information Technology
Sir Sitaram and Lady Shantabai Patkar College of Arts and Science, Mumbai, India

Abstract: *Big Data insists on high volume, high speed, high veracity and high assortment. It needs to manage two central points of contention, the developing size of the datasets and the increment of data complexity. Since Data Modeling is an intricate science that includes methodizing together corporate data so it fits the requirements of business measures. It requires the plan of logical connections so the data can interrelate with one another and support the business. The cycle gives the heading and strategies to extricate data from the information and continue the correct way to achieve business objectives. The Data Analytics lifecycle guides them all through this cycle. The paper aims to study the data analytics lifecycle and data modeling terminologies that how data can be sorted and complex datasets are handled using big data.*

Keywords: Data Analytics Lifecycle, Big Data, Relational Database , Data Modeling , Large Scale Data.

I. INTRODUCTION

1.1 Big Data

Today the quick development of the web and the huge use of the data have prompted the expanding CPU necessity, speed for reviewing information, a construction for more perplexing data structure the integrity, the dependability and the trustworthiness of the accessible data. This sort of data is called as Large-scale Data or Big Data. It can be considered as an software-utility that aims to solve the problem of growing data requirements through analysing , extracting and processing complex data structures that traditional data processing software could not have done before. For e.g – Social media generates around 500+terabytes of data everyday which includes posting photos , videos and comments on social networking sites like Instagram, facebook , etc.

Statistics shows that a jet engine can generate upto 10+terabytes of data in 30 minutes , with the number of flights per day it can generate upto petabytes of data.

1.2 7V's of Big Data

7V of Big data	Description
Volume	Size of data generated in zettabytes, exabytes, yottabytes .
Velocity	Speed of generated Data in real time, data processing and accessing.
Variety	Structured and unstructured data (different types)
Variability	The data whose value keeps on changing constantly, focuses on interpreting the correct meaning of raw data.
Veracity	It is all about the righteous ,accuracy and the quality of data which can be processed to achieve useful insights.
Visualization	It refers to the presentation of data in human readable , understandable , accessible form using any data presentation tools.
Value	After processing data at each steps, if data cannot turn into a value , it is of no use.

Table [1]: 7V of Big Data

II. DATA ANALYTICS

The Data Analytics Lifecycle is designed specifically for big data problems and data science projects. The iterative depiction of lifecycle is intended to more closely portray a real project, in which aspects of the project move forward

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

and may return to earlier stages as new information is uncovered and team members learn more about various stages of the project.

The data Analytics Lifecycle defines analytics process best practices spanning discovery to project completion. The lifecycle draws from established methods in the realm of data analytics and decision science. This synthesis was developed after gathering input from the data scientist and consulting established approaches that provided input on pieces of the process. There are six phases of the lifecycle which are listed below-

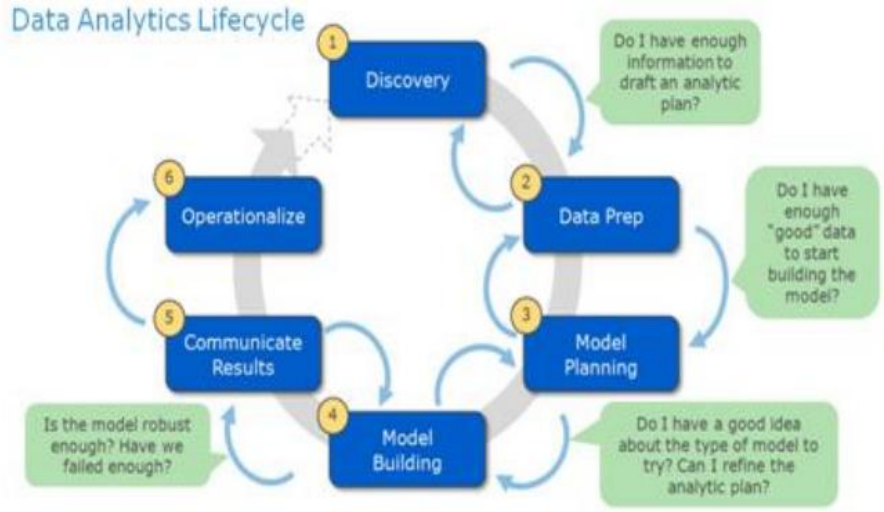

Figure [2]: Data Analytics Lifecycle

1. **Phase 1 – Discovery:** The group learns the business space, including important history, for example, regardless of whether the association or specialty unit has endeavored comparative activities in the past from which they can learn. The group surveys the assets accessible to help the venture as far as individuals, innovation, time, and data
2. **Phase 2 – Data Preparation:** It requires the presence of an insightful sandbox, where the group can work with data and perform examination for the length of the undertaking[2]. The group needs to extract, load, and transform (ELT) or extr act, transform and load (ETL) to get data into the sandbox.
3. **Phase 3 – Model Planning:** In this the group decides the strategies, procedures, and work process it plans to follow for the ensuing model structure stage. The group investigates the information to find out about the connections among factors and hence chooses key factors and the most reasonable models.
4. **Phase 4 – Model Building:** The group creates datasets for testing, preparing, and creation purposes. Moreover, in this stage the group constructs and executes models dependent on the work done in the model planning stage. The group additionally thinks about whether its current devices will get the job done for running the models, or on the off chance that it will require a more powerful climate for executing models and work processes.
5. **Phase 5 – Communicate Results:** The group, as a team with significant partners (Stakeholders), decides whether the after effects of the venture are a triumph or a disappointment dependent on the models created in Phase 1. The group ought to recognize key discoveries, evaluate the business esteem, and build up an account to sum up and pass on discoveries to partners.
6. **Phase 6 - Operationalize:** The group conveys last reports, briefings, code, and specialized archives. Likewise, the group may run a pilot task to carry out the models in a production environment[2].

III. Types of Data Models in Big Data

The data model is a system for arriving at the outline by examining the data being alluded to and getting a significant appreciation of the data set. The strategy to decipher the data pictorially helps the experts and the business with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

appreciating the data and perceive how might it be used. Large Data Models have various possibilities. These lead to variety in both substance and style, which can cause chaos, shock, and logical inconsistency.

3.1 Conceptual Data Model

A Conceptual Data Model is a coordinated perspective on data set ideas and their connections. The purpose behind making a conceptual data model is to set up their properties, and connections. In the data modeling level, there is not really any detail accessible on the genuine data set construction. It has 3 main attributes

1. **Entity:** A real-world thing
2. **Attribute:** Characteristics or properties of an entity
3. **Relationship:** Dependency or association between two entities.

3.2 Physical Data Model

A Physical Data Model depicts an database explicit execution of the data model. It offers database deliberation and creates the pattern. This is a result of the lavishness of meta-data offered by a Physical Data Model. The actual data model likewise helps in imagining data set design by repeating data set segment keys, limitations, files, triggers, and other RDBMS highlights.

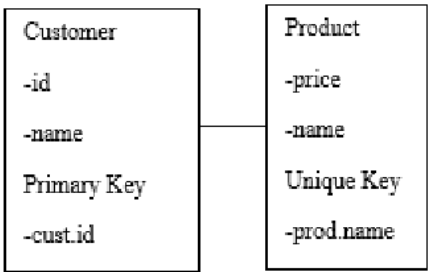

Figure [2]: Example of Physical Data Model

3.3 Logical Data Model

Logical Data models empower to portray the detailed structure of the data segments in a framework and the connection between them. They channel the information parts introduced by a Conceptual data model and frame the establishment of the Physical data model. It has 3 types

1. **Relational data models**– This addresses data as tables or relations.
2. **Network data models**. It addresses data in a form of record. This model additionally depicts a restricted sort of one to numerous relationships called a set type.
3. **Hierarchical data models**– It represents a progressive tree structure. Each branch of the hierarchy shows a various number of related records.

IV. Data Modeling Methods

4.1 Enterprise Data Modelling

An enterprise data model is a kind of model that presents a perspective on all information devoured across the business. It gives an incorporated at this point wide outline of the business information, paying little mind to the information the board innovation utilized. It similarly helps in making other information base parts, for instance, entity relationship graphs, XML diagrams, and data dictionary.

4.2 Star Schema Data Modelling

It is known as a star schema because of its graph that resembles a star, with focuses sending from a center. Star Schema is made for addressing immense informational indexes and are used as a piece of data stockrooms and information stores to help business insights, OLAP 3D shapes, logical applications.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

4.3 Object Oriented Data Modelling

It depend on objects in object-oriented programming (OOP). In OOP, a entity is addressed as an object and objects are put away in memory. objects have properties like fields, properties, and strategies This model portrays an data base as a blend of objects, or reusable programming segments, with related highlights and procedures.

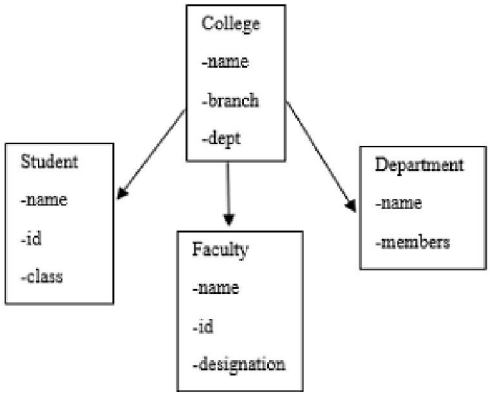

Figure [3]: Example of Object Oriented Data Modeling

4.4 Hierarchical Technique

It is a tree-like structure, A hierarchical model is a model wherein lower levels are arranged under a pecking order of progressively more significant level units. Data is gathered into bunches at least one levels, and the impact of the groups on the data focuses contained in them is considered in any statistical analysis. It stores data in node formats.

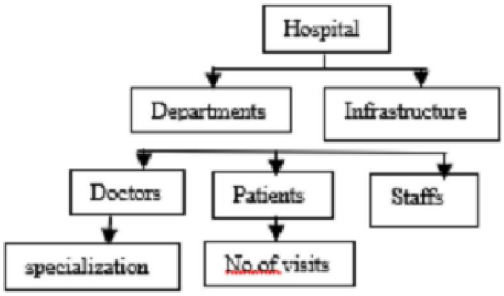

Figure [4]: Hierarchical model

4.5 Entity-Relationship Model

Entity-relationship model is an high level data relational model which is utilized to characterize data components and relationship for the elements in a framework. This conceptual design gives a superior perspective on the information that helps us in better understanding. In this model, the whole database is addressed in a graph called a entity-relationship diagram, comprising of Entities, Attributes, and Relationships.

4.6 Network Technique

The network data set model was a movement from the hierarchical database model and was intended to take care of a portion of that model's issues, explicitly the absence of adaptability. Rather than just permitting every kid to have one parent, this model permits every kid to have various guardians (it calls the kids individuals and the guardians proprietors). It delivers the need to show more unpredictable connections like the orders/parts many-to-numerous relationship given below. In the given fig. M1 has two individuals, N1 and N2. B1. is the proprietor of O1, O2, O3, O4 and O5. In any case, in this model, O2 has three proprietors, P1, P2 and P#.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

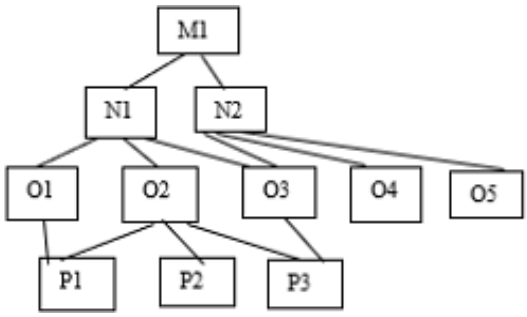

Figure [5]: Network model

4.7 Relational Technique

Relational Model (RM) addresses the database as an assortment of relations. A relation is only a table of Values. Each column in the table addresses an relation of related data values. These columns in the table signify a relationship. The table name and column names are useful to decipher the importance of values in each row. The data are addressed as a bunch of relations. In this model, data are put away as tables. For e.g. DB2 and Informix Dynamic Server - IBM Oracle and RDB – Oracle, SQL Server and Access –Microsoft.

4.8 Dimensional Model

Dimensional data modelling addresses data as cube operation. The impression of Dimensional Modeling was created by Ralph Kimball and is comprise of "facts" and "dimensions" tables. In this, the transaction record is partitioned into either "facts," which are numeric transaction data, or "dimensions," which are the reference data that offers setting to current realities. For instance, a sale transaction can be harm into facts, for example, the quantity of items requested and the cost paid for the items, and into dimensions, for example, request date, client name, item number, request transport to, and bill-to areas, and salesman answerable for accepting the request.

4.9 Non-Relational Database Model

Non-relational data set model is not normal for Relational database model. It doesn't ensure the ACID properties [32]. Non-social information bases may essentially be classified based on method of getting sorted out data as follows.

Key-Value Store: Key Value store permits us to store schema less data. This data comprises of a key which is addressed by a string and the real data which is the value in key -value pair. The data can be any primitive of programming language, which might be a string, a number or a cluster or object. Key-value data sets utilize smaller, proficient record designs to have the option to rapidly and dependably find a value by its key, making them ideal for frameworks that should have the option to recover data in consistent time. For e.g. Redis, Riak, and Oracle NoSQL database are examples of key-value databases.

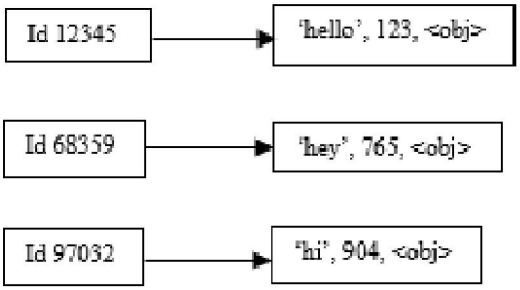

Figure [6]: Key Value store

- **Document Store:** The Document Store are schema less. It gives a system to query collections dependent on multiple attribute value constraints. It is useful for storing texts, emails, XML documents. An archive database

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

connects a unique key with an data structure called a "document." The key is utilized as a simple identifier (ID), ordinarily as a string, a path, or a URI. It tends to be utilized to locate and pull the document from its database.

- Graph Store: A graph model is one whose single hidden data structure is a marked coordinated chart. The Graph Store comprises of single digraph. An database schema in this model is a coordinated graph, where leaves address data and internal nodes address associations between the data. As implementation is concerned, Graph Store may give exceptional capacity graph designs to the representation of structures and the most efficient graph algorithms accessible for acknowledging specific tasks.
- **Column Oriented Database:** A column-oriented database stores data in column order. Segment stores have the benefit that word reference passages may encode various qualities on the double. Data put away in columns is more compressible than data put away in rows. compression algorithms perform better on this data .For model, a data set containing data about companies that have employee name, enrollment number, address, and office. Putting away that data in columns permits the entirety of the name to be put away together. Further, if the data is arranged by one of the sections, that columns will be super compressible.

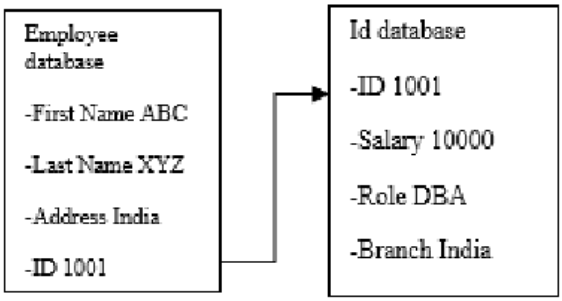

Figure [7]:COD Model

V. BIG DATA MANAGEMENT APPROACHES

There are two main approaches of handling vast amount of data i.e. Map Reduce and Parallel database management system.

5.1 Map Reduce

MapReduce is a software framework and programming model utilized for processing vast data sets. MapReduce program work in two stages, to be specific, Map and Reduce. Map task is associated with splitting and data mapping while Reduce errands shuffle and reduce the data. The data goes through following phases in Map Reduce:

- **Input Splits:** A input to a MapReduce in Big Data work is separated into fixed-size pieces called input splits. It is a lump of the data that is consumed by a single map.
- **Mapping:** This is the absolute first stage in the execution of map reduce system. In this stage data in each split is passed to a mapping method to deliver yield output.
- **Shuffling:** This stage consumes the output Mapping stage. Its errand is to solidify the significant records from Mapping phase output.
- **Reducing:** In this stage, output values from the Shuffling stage are accumulated. This stage joins values from Shuffling stage and returns a single output value. So, this stage sums up the total dataset.

For E.g. Problem Statement

Count the number of words repeated in the sentence:

ABC has green cycle. PQR has blue cycle. XYZ has no cycle no car.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

Solution Statement:

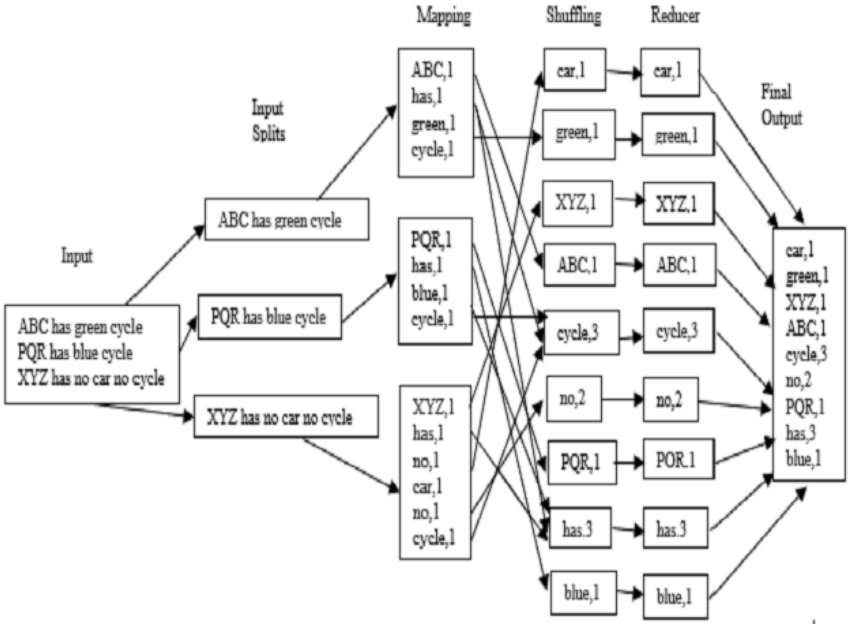

Figure [8]: MapReduce Solution block

A. How MapReduce Works?

The MapReduce algorithm contains two significant tasks, named Map and Reduce. The map task is finished through Mapper Class. The reduce task is finished through Reducer Class. Mapper class takes the data, tokenizes it, guides and sorts it. The output of Mapper class is utilized as input by Reducer class, which thus look through coordinating with sets and decreases them. MapReduce executes different mathematical algorithms to divide task into little parts and allot them to various frameworks. In specialized terms, MapReduce calculation helps in sending the Map and Reduce errands to suitable clusters in a group. Some of the algorithms named - Sorting, Searching, Indexing,TF-IDF.

5.2 Parallel Database Management System

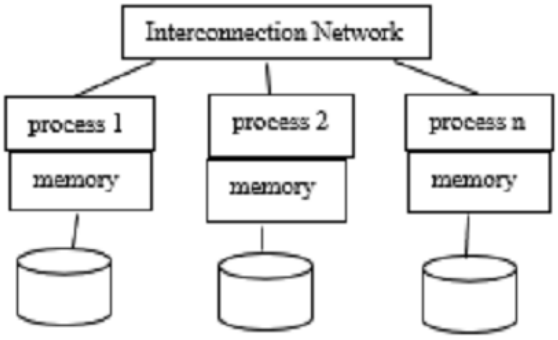

Figure [9]: Architecture of parallel DB

An parallel DBMS is a DBMS executed on a multiprocessor PC. The parallel DBMS implements the idea of horizontal partitioning by disseminating portions of a huge relational table across different multiple nodes to be prepared in parallel. This requires an apportioned execution of the SQL queries, such as SELECT, can be executed in-conditionally on every one of the node. Different assets like CPUs and Disks are utilized in equal. The tasks are performed all the while, rather than sequential handling. A parallel server can permit admittance to a single db by

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 5, Issue 2, May 2021
Impact Factor: 4.819

clients on various machines. It additionally performs numerous parallelization activities like data(loading) stacking, query evaluation, building files(indexes), and query processing. For e.g commercial parallel databases as Teradata, Aster Data, Netezza, DATAllegro, Vertica, Greenplum, IBM DB2 and Oracle Exadata.

VI. Data Modeling Tools

Some of the data modelling tools

- **ER Studio:** E-R Studio is an data displaying programming empowering clients to effectively inventory data resources and sources across various stages, construct and offer data models, and track start to finish data lineage.
- **DBSchema:** DbSchema is a visual databbase originator and supervisor for any SQL, NoSQL or Cloud data set. The device empowers you to plan and communicate with the data set composition, make exhaustive documentation and reports, work disconnected, synchronize the blueprint with the data base, thus substantially more.
- **ConceptDraw:** It offers a scope of business-explicit additional items for making Infographics, graphs, data representation, and flowcharts for the business interaction model.
- **Erwin:** It is an data modelling tool which is utilized to make consistent, physical, and reasonable data models. It is a standout amongst other data modelling devices that assists you with making the actual database from the physical model.
- PgModeler is an open-source tool for making and editing database models with a natural interface. This instrument upholds the making of the most fundamental article like a single column, and the client characterizes process, functions, queries, operators and language.

VII. Conclusion

The paper studies the data modelling techniques along with the data models present and the main approaches of data processing in big data, the paper also have an insight on the data analytics lifecycle. Since with the enormous growth or the hunger of data, the demand arises and the organization has to look out on the data structure and this requires designing and implementing complex and large data set which needs to be professionally and well structured, there are various tools available in the market, some of them are given. With this, we can conclude that big data analytics and its terminologies came up as avital solution to the ever increasing demand of data.

References

[1] https://mkhernandez.wordpress.com/data-analyticslifecycle/#
[2] http://quicktechie.com/cs/data-science-q-a/157-phases-of-data-analytics-lifecycle/#
[3] “Data Modeling and Data Analytics: A Survey from a Big Data Perspective”, André Ribeiro, Afonso Silva, Alberto Rodrigues da Silva, 2015.
[4] “A Study On Big Data Modeling TechniqueS”, B Sai, Singaraju Jyothi, 2020.
[5] “Distributed parallel architecture for big data[j]”, Catalin Boja, Adrian Pocovnicu, Lorena Batagan, 2019.
[6] https://medium.com/edureka/mapreduce-tutorial/#
[7] https://www.journaldev.com/mapreduce-algorithmexample/#
[8] “Analysis of Bigdata using Apache Hadoop and Map Reduce”, Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N, 2019.
[9] “Big Data- solutions for RDBMS problems-A Survey”, S.Vikram Phaneendra & E.Madhusudhan Redd, 2018.
[10] “Big Data And Hadoop: A Review Paper”, Rahul Beakta, 2017.
[11] https://geeksforgeeks/#