Issue 9 vol 3 ISSN: 2321-8134

IJFEAT

INTERNATIONAL JOURNAL FOR ENGINEERING APPLICATIONS AND TECHNOLOGY

ANATOMY OF DATA

Rishabhdev V Shukla[1]

[1]*Computer Science and Engineering, Government College of Engineering, Chandrapur, India*,

rishabhdevshukla13@gmail.com

Abstract

They created the nineteenth century computer for solving complex problems and merely for storing and keeping records. New flavors were added to it from time to time and they never looked back. The entire concept of storage was mesmerizing to the world. The idea of storing large amount of data that could take as much space as a room when stored physically in paper form into a compact device in digitized form was itself a fairy tale coming true experience. Human genius led to the inventions of new storage devices such as CD-ROM, pen drives etc. Later, they felt the need for a mechanism that could not only store but also could retrieve data on demand from time to time which led to the invention of RDBMS. The emerging IT giants back then felt the need of new mechanisms for efficient handling of data. Hence memory management and data handling became the sun of the entire IT industry. With technological advancements, increased the human wants. Orthodox, manual techniques started to be replaced by new efficient technology. Printed bills replaced the manual ones, handy music gadgets such as an IPod carved their places into the market replacing the orthodox tape recorder, cameras, mobile phones and all sort of digital gadgets waived into the market hence digitization was eminent. Virtual money led to the era of cashless economy. The mindset of the IT industry started to shift its gear and there was no looking back. In a very short span of time the industry flourished and its impact gave a boost to each and every sector. Today, all firms rely on data handling centers for services. Data management was neglected by most of the tech giants but due to exponential growth of data over the years, data scientists have come up with new mechanisms for efficient data management. Between 1986 and 2010, world had 1.2 zettabytes (1.3 trillion gigabytes) of data stored on internet, desktop hard drives and other storage devices. And the astonishing fact was that the estimated amount of total data was predicted to be 35 zetta bytes by the year 2020. This tremendous increment was a clear indication to the tech giants to establish new data centers and develop new technologies that could facilitate faster processing and retrieval of data. As a result in the coming decade the world witnessed revolutionary changes in the field of data science. Hence it is important to study about data science and analytics as the world has only 40% computer science engineers of total requirement who work on data. This piece of research work is merely an attempt to put light onto the current data analytical techniques and how did we manage to reach here.

Index Terms: *RDBMS, SQL, PL/SQL, HADOOP (big data), data structures, facts about rapid growth of data, why data management*

--- *** ---

1. WHY DATA MANAGEMENT IS A CONCERN?

With the exponential increment in the number of firms and their increasing reliability on internet and technology there was a common urge to increase the world capacity for data storage and reliable access. Internet banking, online stocks, bitcoin have added fuel to the fire and hence with increase in storage capacity, due attention has to be given to security as much of the customers confidential data is stored on internet. Some facts given below will clear the air

a. Facebook alone stores 300 petabytes of data with an average of 600 terabytes of daily new incoming data.
b. Amazon stores 210 petabytes of customer data alone.
c. On an average, 72 hours of new video is uploaded on YouTube every minute.
d. World had 1.2 zettabytes of data in 2010 and the estimate is that it will become 35 zettabytes by the year 2020.
e. Data production will be 44 times greater in 2020 than it was in 2009.
f. Individuals create more than 70% of the digital universe but enterprises are responsible for storing and managing 80% of it.
g. Walmart collects more than 2.5 petabytes of data every hour from its customer transactions.

The above mentioned points are alarming as the world does not possess sufficient number of data centers to store such a large amount of data. Tech giants have been focusing on this issue over a decade. Google, alone invests 2.5 billion dollars every year to establish new data centers. These alarming facts have forced the world to develop newer algorithms and technologies for compression and storage of such a huge amount of data. The data scientists are heading in this direction by using the primitive approach with modern flavors that adapt to the growing needs.

2. RDBMS

In 1980s the first ever data storage and retrieval tool was developed known to be RDBMS (Relational Data

Base Management System) because the early computer could just store data without effective provisions for retrieving data on demand. RDBMS could store the data in tabular format (structured) and provide relations or links between a set of data by providing primary and foreign keys to associate them. It was a boon to the data management sector and they considered it to be the benchmark. Little did they know that it was just a beginning to an entirely new branch of computer science. With time, it opened the floodgates for new data handling tools as the world started to shift its gears and a common demand for digitization emerged. The growing importance of learning RDBMS lies in the fact that it is the basis to data management and without its knowledge one cannot simply learn about data analytics and other complex mechanisms used for data handling. It is as important as learning C language for learning higher level languages such as java, pearl, python etc. it uses nothing but basic DDL and DML languages to store and retrieve data in a client/server system.

2.1 SQL

RDBMS is the structured or tabular representation of data with relatable attributes. In order to retrieve or access the data from the database a mechanism is required which can interact with the database and provide the required data. The revolutionary step in this direction was the invention of SQL (Structured Query Language) which is considered to be the highest level language as it is almost equivalent to the English language. It was designed for the purpose of providing a user interface to access the database. It had its own pros and cons which later were rectified by a better version of the language. Any set of tuples can be accessed easily as it is a user friendly tool. It basically provides a user interface to interact with the database. The DBMS defines the data in a database using Data Definition Language (DDL) and handles requests to retrieve, update, or delete data using Data Manipulation Language (DML). It is basically a client/server system where client can retrieve or update the data stored by the server using SQL as a tool for the purpose.

2.2 PL/SQL

The SQL was considered to be a benchmark but later with rapid growth of data and user requirements there was an urge to develop a better version of it. SQL couldn’t handle exceptions and had several other drawbacks too. It is simply a query language that can be used to retrieve structured data with no provisions for developing procedures. Hence PL/SQL (Procedure language / Structured Query Language) was developed to overcome the drawbacks of its predecessor and to meet the needs of the growing demands from industry. Today, a considerable number of engineers are employed to work on this software and it has carved out an important position in field of data handling. The distinct feature that makes it so unique is the ability to handle exceptions while making queries and to write step by step procedural coding for data retrieval purpose. The only requirement is that the data must be structured. It does not work well on unstructured data. Earlier when the amount of data was limited, it proved to be a reliable software for data access but with the drastic change in business environments and with growing demand for large data processing, the demand for PL/SQL started to fade which led to the development of newer software that could handle and process very large amount of data in comparatively shorter time period and could also process unstructured data. However, PL/SQL has it own flavors that is the reason why it still sustains in the industry as it can process small data very quickly as compared to the other mechanisms. Hence firms that do not possess very large amount of data still work on PL/SQL.

2.3 HADOOP (BIG DATA)

With the growing reliability over internet with exponential increment in the amount of data in parallel led to the demand for a newer technology that could process and store very large amount of distributed, unstructured data efficiently. The demand was met with the invention of Hadoop. It is a very latest**.** Sensation in the field of data science and analytics. Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment developed in the year 2005. Large data vendors are switching now to Hadoop because of its flavors. The advantages of Java language are carried on by Hadoop and it has its own too. This is the very reason why most of the firms that manage extremely large amount of data are switching to more efficient Hadoop. Engineers who work on Hadoop are gaining importance in the industry and the demand for them is going to increase only in the future too. Hence learning Hadoop is of utmost importance. With the growing demand for big data handlers, world will provide only 30% of the total required engineers working on big data by 2025 hence there will be a shortage of engineers and data analysts who can work on big data. Some facts about big data are mentioned below.

a. Big data will drive $48.6 billion in annual spending by 2019.
b. By 2020 one third of all data will be stored, or will have passed through the cloud by using Hadoop.
c. Retailer using big data to its fullest could increase its operating margin by more than 60%.
d. By 2020, there will be a shortage of big data analytics.
e. USA alone will have a shortage of more than 1.5 million analysts.

3. DATA STRUCTURES AND ALGORITHMS

The part of computer science that deals with internal arrangement and actual implementation of data at the memory level is data structures. The tabular format (rows and columns) is merely for the purpose of human understanding. The computer memory does not work like a human brain. Hence logics have to be developed to efficiently work with the systems memory as it is the key component which decides the efficiency and cost.” Smart data structures and dumb code work way better than the other way round “say the coding geeks. This very saying explains the importance of data structures in the industry. This reason why engineers working at the data level are of so much importance to the system. Their algorithms decide the cost and correspondingly the profit and loss of the firm. Use of appropriate data structures positively influences the efficiency of computer systems by enhancing the ability of the computer to store and retrieve data from any location in its memory. Typically, top coders are considered to be the masters of data structures. Data structures not only provides mechanisms for data storage and arrangement such as stacks, queues, linked lists, trees, graphs, sorting techniques but also provides parameters to determine the effectiveness of the code. These parameters are determine the time and space complexities of a certain code. Time complexity determines the time required by the code to execute the operation while the space complexity determines how much space the code will acquire in the computer memory. The least the values of these complexity parameters, the more optimal the developed code will be.

4. CONCLUSION

As stated in my research paper the science of data processing and handling had been neglected over the years and was not of too much concern as compared to the other branches of computer science but now in this era where technology has spread its roots into the everyday lives of the people, the world is relying over the internet and technology for most of the tasks which were performed manually, earlier. With the growing wants in the past decade has grown the demand for establishment of new data centers and processing and retrieving techniques for faster and secure handling of data. The attempts in this direction over the years have been briefly illustrated in the paper.

ACKNOWLEDGEMENT

It gives me immense pleasure to present the research paper on the topic ANATOMY OF DATA for IJFEAT. The thick and thins of the science that deals with the handling of the data is known to the people working in this very field better than the most. So, I would like to thank my colleagues who are currently associated with multinational giants namely Manhattan and Virtusa Polaris. They were a part of my case study while I was working on my research paper.

REFERENCE

[1] Case study at Virtusa Polaris MNC. Paragraph content goes here.