Chapter 3
A Survey on Tools for Data Analytics and Data Science

Pankaj Pathak
https://orcid.org/0000-0002-5875-0387
Symbiosis International University (Deemed), India

Samaya Pillai Iyengar
Symbiosis International University (Deemed), India

Minal Abhyankar
Symbiosis International University (Deemed), India

ABSTRACT

In the current times, the educational and employment areas are changing at a very fast rate. The change is visible especially in the zone of technology-education. Approximately 4-5 years back, technology education meant coding, using different computer science programming languages. But in the recent times data science and data analytics have become the buzz words. The employment in this area has also undergone a tremendous change effect. Many new employment opportunities have sprung in this area as well with the regular or existing jobs becoming less or extinct. The entire business domain is warming to these buzz words. And the industry preference for these techniques has widened. The chapter discusses both the concepts and the tools being used.

INTRODUCTION AND BACKGROUND

Database is the container of information i.e. processed data. It is used to store the data. The main objective of a database is storage of data. With the Database comes the database management system, the system to create and manage all operations related to the database. Codd (1990).

The timeline for Database is as follows:

1950s and early 1960s: The Data processing and storage of data mainly done with magnetic tapes. The Magnetic tapes could give a sequential access only. For the input process the “Punched cards” were used.

DOI: 10.4018/978-1-7998-3053-5.ch003

Late 1960s and 1970s: The innovation of Hard disk was done. It allowed direct access to the data. In database handling, the network and hierarchical data models were in reputed and used extensively. Ted Codd put forth the concept of relational data model, which is still relevant in today's world. The relational data model enabled better performance in transactions and helped real time transactions.

1980s: Research in the area of relational DB domain has a great commercial value. During this time, SQL became the industrial de facto standard. Parallel and distributed database systems were launched in the commercial arena for usage. They proved to be most useful for organizations. It was during this time that the Object-oriented databases were also featuring as a new concept in the database domain.

1990s: In this era there was a thrust in decision support systems which were huge. The data-mining applications were also launched and developed during this period. Large multi-terabyte data warehouses were designed. It was the new emergence of a concept called "Web commerce".

2000s: Here the XML and XQuery standards were launched and developed. "Automated database administration" started to feature in the organizations simplifying and easing the lives of database administrators.

Data Science

Data science is both art and science of handling data, mostly big data. Data pre-processing needs to done before using the data for analytics, the data gathered can be of historical data or online real time data, once that is received, various algorithms like the KNN algorithms and Multidimensional scaling algorithms were performed to get the required trends. The people working here are called as "data scientist", Van Der Aalst, W. (2016).

Data scientist understands data from a business point of view and provides accurate predictions and insights that can be used to power critical business decisions. Today R and Python are the major tools used in data analytics.

It can be said that "data science" is connected to computer science, but in principle, it is a distinct and separate field. Computer science as a domain consists of forming programs, algorithms and processing data. Data science covers any type and all types of data analysis. Computers may constitute the process of analysis or can be ignored from the process. Data science is mainly related to the field of 'Statistics'. It includes the steps of data collection, organization of data, analysis of data, and representation of data. The huge amounts of data in the organizations they have resorted to data science for survival and sustenance. Data science has become an integral part of Information Technology. As the technological advancements enable and provide an edge to data science. For example, a company that has huge amount of data can use data science for collecting, storing, managing and also analyzing that data effectively. This data is then run with many tests in a scientific method to extract results. Provost, F. et al. (2013).

Data Science is a bigger picture of data analytics which includes not only simple statistical modeling but also mathematics and calculations. Data science is an umbrella under which all the three parts come up, with subject expertise. Data science mostly tackles big data, data cleansing, preparations and analysis. Data science is used to generate deep insights from the collected data set. It uses the data mining concepts, tools of data mining, predictive analytics and machine learning to generate critical information. Domain knowledge is very important when it comes to data science even more than from data analytics. In the current times Machine learning algorithms also play a pivotal role.