The Role of Data Science in Big Data Analytics - Overview

S. Balakrishnan
Professor and Head, Dept. of Computer Science and Business Systems, Sri Krishna College of Engg. and Technology, Coimbatore, Tamilnadu, India.

Data is made ceaselessly, and at an ever-growing rate. Phones, Online social networking, imaging advances to choose a therapeutic assurance all these and more make new data, and that must be taken care of some spot for reasons unknown. Gadgets and sensors therefore produce explanatory information that ought to be taken care of and arranged persistently. Big Data is making critical new open doors for associations to infer new esteem and make upper hand from their most important resource: information. The field of data science is "rising at the crossing point of the fields of social science and statistics, information and computer science and design".

1. Introduction

Around 100 hours of video are moved to YouTube reliably and it would take around 15 years to watch every video moved in one day. AT&T is thought to "hold the world's greatest volume of data in one remarkable database – its phone records database is 312 terabytes in size, and contains pretty much 2 trillion lines". Consistently we send "204,000,000 messages, create 1,800,000 Facebook likes, send 278,000 Tweets, and up-load 200,000 photographs to Facebook 570 new sites spring into reality each moment of consistently".

Data science is the "investigation of where data originates from, what it speaks to and how it tends to be transformed into a significant asset in the production of business and IT systems". A key part of information science is the use of the logical strategy to shape and move theories to approve decisions about fundamental examples in information.

The general goal of data science may appear to be direct; however execution is an intricate procedure and includes various strides before the estimation of an information science item can be watched. This is what that resembles:

1. "Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. DS team evaluation
6. Stakeholder evaluation
7. Deployment"

The data and analytics landscape is evolving. The business around enormous information and information science is one aftereffect of this advancement/ transformation. In spite of the fact that the market frequently utilizes the terms enormous information and information science reciprocally, they are actually very unique. Big data alludes to the capacity to oversee huge volumes of divergent information at the correct speed and inside the perfect time span to empower examination and activity. Large information is about the three v's for example volume, variety, and velocity – and some would include esteem. Associations are pushing toward progressively cross breed conditions to deal with this enormous and multistructured information. This regularly incorporates the cloud, Hadoop, and information lakes just as NoSQL databases and different stages. Enormous information examination habitually requires the utilization of MPP (massively parallel processing engines), in-memory processing, and different advancements that can deal with huge amounts of information.

2. Data Science Landscape

Data science landscape can be divided into the following categories: (i) Fields (ii) Objects (iii) Techniques and (iv) Approaches.

2.1 Data Science Fields

Data Science has different fields such as "Nanotechnologies, Physics, Robotics, Mathematics, Statistics, Information theory, Information technology and Artificial Intelligence".

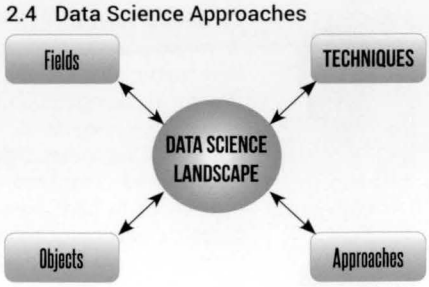
2.2 Data Science Objects

Methods that "scale to Big Data are of particular interest in data science, although the discipline is not generally considered to be restricted to such data".

2.3 Data Science Techniques

Data science techniques may be in the form of "Signal processing, Probability models, Machine learning, Statistical learning, Data mining, Database, Data engineering, Pattern recognition, Visualization, Predictive analytics, Uncertainty modeling, Data warehousing, Data compression, Computer programming and High Performance Computing".

2.4 Data Science Approaches



The development of "machine learning, a branch of artificial intelligence used to uncover patterns in data from which

predictive models can be developed, has enhanced the growth and importance of data science".

3. Big Data Landscape

So as to design big data architecture it is essential to get a handle on the information on the current large information scene and fuse it into existing foundation. In conventional information the board structures, the organized data or information was nourished into the endeavor combination device which moved the gathered organized information into information stockrooms or operational units. At that point diverse scientific capacities were utilized to uncover the information, however the new type of information the executives structures that acquire large information scene are intended to meet the "velocity, volume, value and variety of requirements". To deal with these enormous informational indexes, new designs have been framed that fuse multi hub equal handling strategies. Big data landscape has a further characterization dependent on preparing prerequisites and various techniques are proposed for group handling and continuous preparing.

A few innovations through which we can outfit big data are:

1. Massively Parallel Processing
2. MapReduce
3. NoSQL
4. Hadoop

3.1 Massively Parallel Processing

The data is circulated among various hubs for quicker handling. The procedure is done resemble on each machine and the yield is gathered to reason the necessary outcome. This innovation requires information on SQL and costly equipment to chip away at.

3.2 MapReduce

Mapreduce also use the concept of "multi nodes and parallel processing". It consists of two functions:

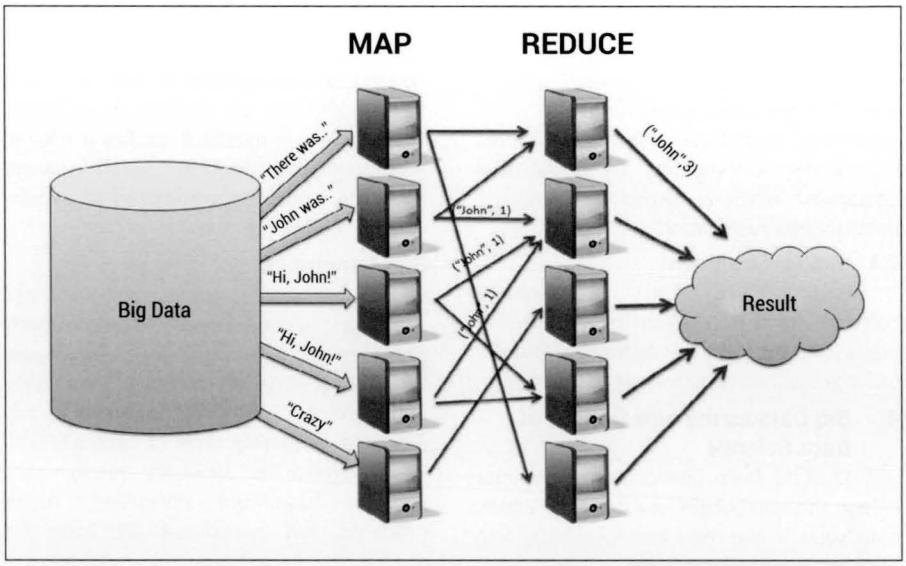
(i) Map - It isolates data over various hubs which are then prepared in parallel.

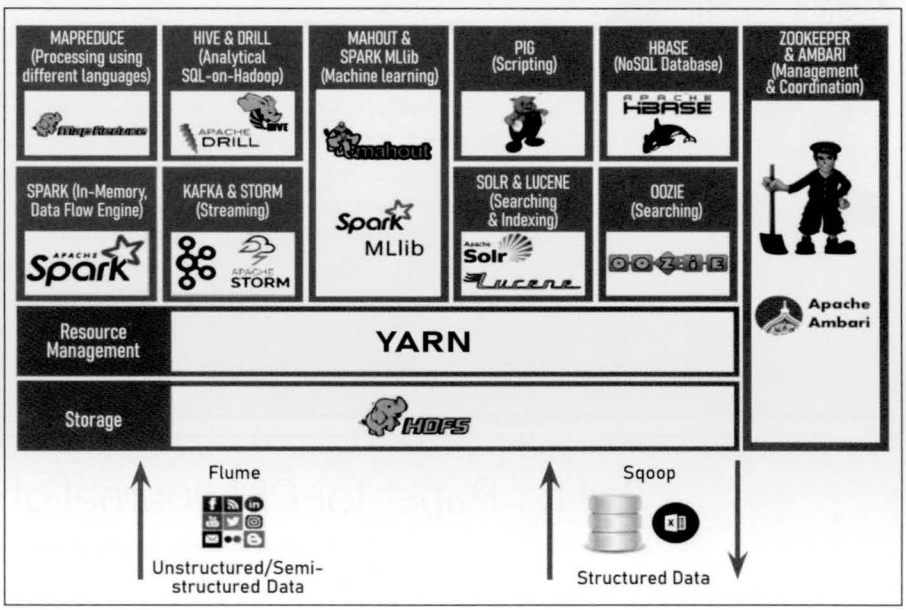
(ii) Reduce - This capacity consolidates the outcome sets into a last reaction.

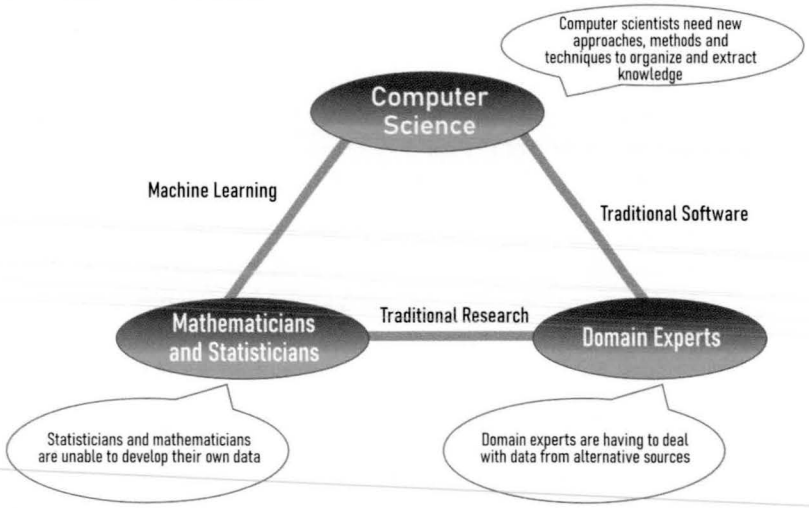
Massively parallel processing uses "SQL queries however Map Reduce uses java and doesn't require exorbitant submitted stages".

3.3 NoSQL

NoSQL database-management systems are "unlike relational database-







management systems, in that they don't utilize SQL as their query language". They forgo the overhead of "indexing, schema and ACID transactional properties to make enormous, reproduced information stores for running investigation on economical equipment, which is helpful for managing unstructured information".

3.4 Hadoop Ecosystem

Hadoop is an "open-source software framework used for storing and processing Big Data in a distributed manner on large clusters of commodity hardware".

4. Big Data as the new frontier of Data Science

Starting from phenomena such data deluge, the existence of new and alternatives data sources like the Internet, sensors and images, the availability of data not ad-hoc collected but automatically generated, it is understood that relations between scientific fields could not be confined to a binary interdisciplinary relationships, but it needed a triangulation and a transdisciplinary approach, and the identification of a data-driven scientific method.

5. Conclusion

Data science, big data, and advanced analytics have been "progressively perceived as significant main impetuses for cutting edge advancement, economy, and instruction". Despite the fact that they are at a beginning time of improvement, vital conversations about the "master plan, patterns, significant difficulties, future headings, and possibilities are basic for the sound advancement of the field and the network". The up and coming age of information science, including a wide scope of orders, science, and economy, depends intensely on the key arranging and visionary activities that will be embraced in organized information look into territories and new companies. Most assuredly, the present inquiries, for example, "for what reason do we need information science" will be supplanted by a group of logical speculations and devices to address the obvious stupendous difficulties and huge issues confronting tomorrow's huge information, science, business, society, and the economy. We will be significantly astonished by the astounding advancements and potential changes that will occur in the following 50 years.

About the Author

Dr. S. Balakrishnan (CSI Membership No. 2060000034) is a Professor at Sri Krishna College of Engineering and Technology, Coimbatore, Tamilnadu, India. He has 17 years of experience in teaching, research and administration. He has published over 15 books, 3 Book Chapters, 11 Technical articles in CSI Communications Magazine, 1 article in Electronics for You (EFY) magazine, 3 articles in Open Source for You Magazine and over 100 publications in highly cited Journals and Conferences. Some of his professional awards include: 100 Inspiring Authors of India, Deloitte Innovation Award, Cash Prize ₹ 10,000/-, from Deloittee for Smart India Hackathon 2018, Patent Published Award, Impactful Author of the Year 2017-18. His research interests are Artificial Intelligence, Cloud Computing and IoT. He has delivered several guest lectures, seminars and chaired a session for various Conferences. He is serving as a Reviewer and Editorial Board Member of many reputed Journals and acted as Session chair and Technical Program Committee member of National conferences and International Conferences at Vietnam, China, America and Bangkok. He has published more than 6 Patents on IoT Applications.