Statistical Learning Issues: Challenges in Convergence of Data , Cloud Computing & Data Science

Yash Kulkarni[1] Arya Salokhe[2]
[1]B.E. [2]B. Tech Student
[1]Department of Computer Engineering [2]Department of Information Technology
[1]Keystone School of Engineering Pune, India [2]DR. J.J. Magdum College of Engineering Jaysingpur, India.

Abstract

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. cloud computing is the delivery of computing services including servers, storage, databases, networking, software, analytics, and intelligence over the Internet to offer faster innovation, flexible resources, and economies of scale. Data science is a field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Big data is the subset of data analysis. This paper discusses the convergence of big data, cloud and data science. Also identifies various issues faced during big data, cloud, data science and their convergence.

Keywords – Big data, Cloud computing, Data science, Issues

I)Introduction

Today more than 2.9 million emails are sent across the internet every second. 375 megabytes of data is consumed by households each day. Data is being created every minute of every day without us even noticing it. Data is being created every minute of every day without us even noticing it.

Big data may be defined as data sets whose size is beyond the ability of typical database software tools to capture, create, manage and process data. The definition can differ by sector, depending on what kinds of software tools are commonly available and what sizes of data sets are common in a particular industry. Cloud computing is Internet-based utility computing, basically shared resources, software and information that are used by end-users hosted on virtual servers.

Data Science was created to understand the data and their relationships, analyse them, but above all to extract value and ensure that, properly interrogated and correlated, they generate information that is useful not only to understand the phenomena but above all to orient them.

The paper analyses the convergence of these three terms. How they are impacting on the IT industry. The paper discusses the issues in big data analysis through cloud computing. The issues are i) Issues in big data ii) Issues in cloud computing iii) Issues unsolved after convergence. These challenges needs to be addressed and resolved.

The rest of the paper is organized as follows. Section II consists relationship between big data analysis and cloud computing . Section III describes difference . Section IV describes convergence. Section V states issues and lastly section VI provides the summery of the work.

II)Relationship between Big data, cloud computing and data science.

There are infinite possibilities when we combine Big Data and Cloud Computing! If we simply had Big Data alone, we would have huge data sets that have a huge amount of potential value just sitting there. Using our computers to analyse them would be either impossible or impractical due to the amount of time it would take. Cloud Computing services largely exist because of Big Data. Likewise, the only reason that we collect Big Data is because we have services that are capable of taking it in and deciphering it, often in a matter of seconds. The two are a perfect match, since neither would exist without the other.

both Big Data and Cloud Computing play a huge role in our digital society. The two linked together allow people with great ideas but limited resources a chance at business success. They also allow established businesses to utilize data that they collect but previously had no way of analysing.

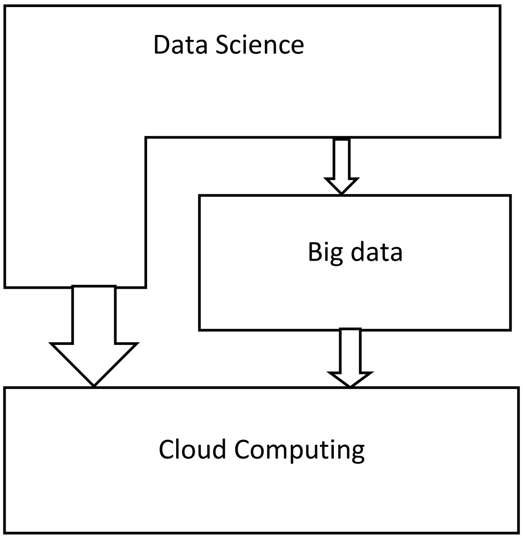

Fig: Relation between Data Science ,Big Data And Cloud Computing

III)Difference between big data science and cloud computing

Big data is a terminology used to describe huge volume of data and information. It refers to structured, semi-structured, or unstructured data that can be further processed for analysis. Computers are used to unlock patterns from the data sets that are further analysed to provide business insights. It includes all kinds of data in many different formats. Big data can exist without cloud computing.

Cloud computing a technology used to store data and information on a remote server rather than on a physical hard drive. Cloud refers to the internet which in this case, acts as an infrastructure as a service. It utilizes a vast network of cloud servers over the internet to analyse data and information, instead of using a personal computer or local server. It's a new paradigm to computing resources. Cloud requires big data for computing resources.

Key Differences are the cloud computing is the computing service delivered on demand by using computing resources dispersed over the internet. On the other hand, the big data is a massive set of computer data, including structured, unstructured, semi-structured data which cannot be processed by the traditional algorithms and techniques. The cloud computing provides a platform to the users to avail services such as SaaS, PaaS, and IaaS, on demand and it also charges for the service according to use. In contrast, the primary objective of big data is to extract the hidden knowledge and patterns from a humongous collection of the data. High-speed internet connection is the essential requirement for the cloud computing. As against, big data uses distributed computing in order to analyse and mine the data.

Both Big data and Cloud computing are the two most trending terms in the ever-growing IT (information technology) world nowadays. Big data is kind of a buzzword used among the marketers to represent large volume of data so huge that is virtually impossible to process by just one machine whether structured or unstructured. Cloud computing is like an application that systematically stores data and programs using a network of remote servers over the internet. Cloud is just a metaphor representing internet. For example, if big data is content, cloud computing is infrastructure.

IV)Convergence Big Data, Cloud Computing and Data Science

Cloud computing virtualizes computer resources as a resource pool to provide computing resources over the network by optimizing resource usage in terms of the CPU, RAM, network, and storage. The main advantage of big data comes through big data analytics. By using big data analytics in cloud, businesses are able to derive better analysis from the large amounts of structured and unstructured data in their possession. The flexibility of the cloud makes it ideal for big data analytics. Also, cloud computing is much cheaper for companies to use than the large-scale big data resources that organizations have used before. Moreover, the cloud also makes data integration from numerous sources easier for companies.

As organizations are shifting their operations and big data analytics into the cloud, this is offering major financial advantages to participating companies. Big data analytics places rigorous demands on networks, storage, and servers. This is why some businesses are outsourcing this hassle and expense to the cloud.

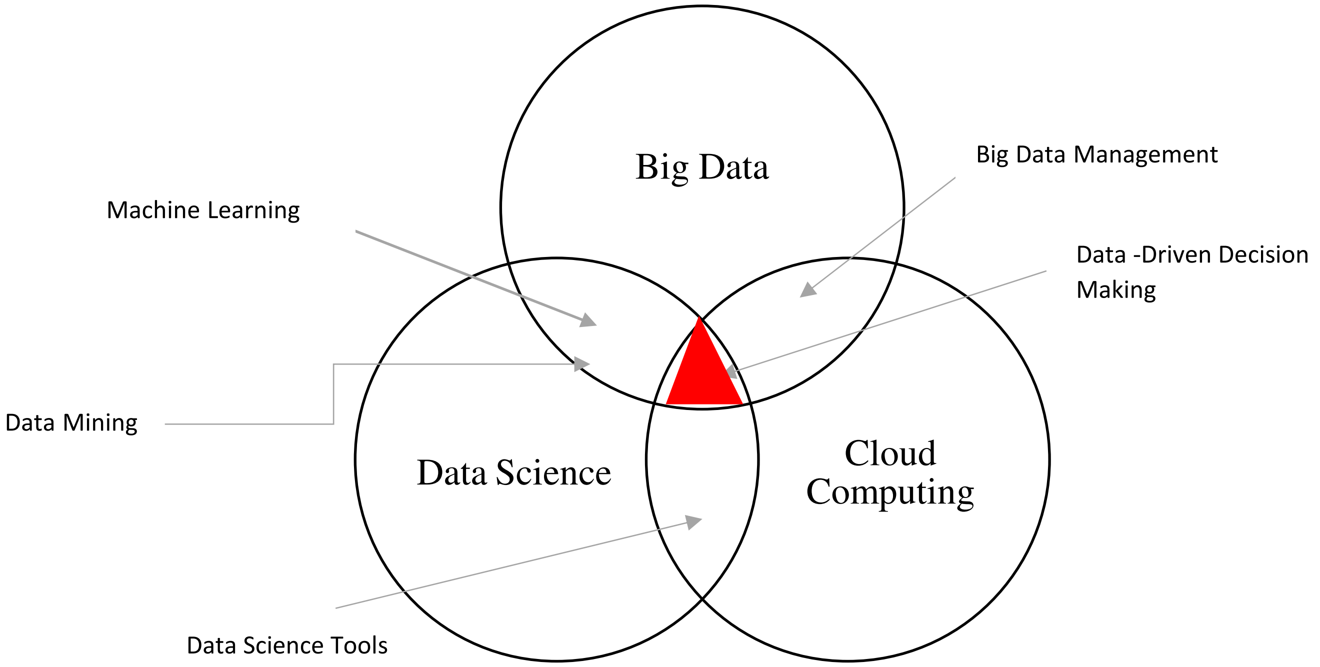

Fig: Convergence of big data, cloud computing and data science

Big data in the cloud is providing new business opportunities that support big data analysis and confront various, architectural hurdles. Big data involves manipulating petabytes (and perhaps soon, exabytes and zettabytes) of data, and the cloud's scalable environment makes it possible to deploy data-intensive applications that power business analytics. The cloud also simplifies connectivity and collaboration within an organization, which gives more employees access to relevant analytics and streamlines data sharing.

V)Issues in Big Data Analysis through Cloud Computing

Just as Big Data has provided organizations with terabytes of data, it has also presented an issue of managing this data under a traditional framework. How to analyse the large sum of data to take out only the most useful bits? Analysing these large volumes of data often becomes a difficult task as well. Security issues in the cloud are a major concern for businesses and cloud providers today. It seems like the attackers are relentless, and they keep inventing new ways to find entry points in a particular system**.** Other issues include ransomware, which deeply affects a company's resources, Denial of Service attacks, Phishing attacks and Cloud Abuse. Globally, 40% of businesses experienced a ransomware incident during the past year. Both clients and cloud providers have their own share of risks involved when making an agreement on cloud solutions. Insecure interfaces and weak API's can give away valuable information to hackers, and these hackers can misuse this information for the wrong reasons. Data replication must be done in such a way that it leaves zero room for error; otherwise it can affect the analysis stage largely. It is important to make the searching, sharing, storage, transfer, analysis, and visualization of this data as smoothly as possible.

The way to deal with these challenges is to implement next-generation technology which can predict an issue before it causes more damage. Fraud detection patterns, encryptions and smart solutions are immensely important to combat attackers. At the same time, it is our responsibility to own our data and keep it protected at our end while looking for business intelligent solutions. Data Acts is also a serious issue which requires data centre to be closer to a user than a provider.

VI) Summary

The paper described the convergence of big data, cloud computing and data science .A relationship between big data, cloud computing and data science established. This paper also identifies the difference between them. And states the issues between convergence of big data and cloud computing. The observations can be stummed up as follows. With data increasing on a daily base, big data systems and in particular, analytic tools, have become a major force of innovation that provides a way to store, process and get information over petabyte datasets. Cloud environments strongly leverage big data solutions by providing fault-tolerant, scalable and available environments to big data systems.

Thus clod and big data become a competitive advantage. More data results into more accurate analysis. And accurate analysis results into better decision making. To realise the full potential of big data, the challenges posed by it must be addressed. For successful big data analysis on cloud privacy, data integrity, data quality these kind of issues must be resolved. And data driven decisions can be made by businesses who manage to handle these issues.

References

[1] Agarwal, D., Das, S. and Abbadi, A. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM 978-1-4503-0528-0/11/0003.

[2] Fan J., Han F., Liu H.
Challenges of Big Data Analysis
ResearchGate, 1 (1) (2013), pp. 1-38
CrossRefView Record in ScopusGoogle Scholar

[3] Chang, V., 2015. Towards a big data system disaster recovery in a Private cloud. Ad Hoc Networks, 000, pp.1–18.

[4] Gens, F. “New IDC IT Cloud Services Survey: Top Benefits And Challenges”. IDC eXchange., 2009. [online].