Chapter 9

Big Data: Techniques, Tools, and Technologies – NoSQL Database

Vinod Kumar
Maulana Azad National Institute of Technology, India

Ramjeevan Singh Thakur
Maulana Azad National Institute of Technology, India

ABSTRACT

With every passing day, data generation is increasing exponentially, its volume, variety, velocity are making it quite challenging to analyze, interpret, visualize for gaining the greater insights from the available data. Billions of networked sensors are being embedded in devices such as smart phones, automobiles, social media sites, laptop, PC’s and industrial machines etc. that operates, generate and communicate data. Thus, the data obtained from various resources exists in structured, semi-structured and unstructured form. The traditional database system is not suitable to handle these data formats. Therefore, new tools and techniques are developed to work with these data. NoSQL is one of them. Currently, many NoSQL database are available in the market, each one of them specially designed to solve specific type of data handling problems, most of the NoSQL databases are developed with special attention to problem of business organizations and enterprises. The chapter focuses various aspects of NoSQL as tool for handling the big data.

INTRODUCTION

Big Data analysis involves making “sense” (Oracle White Paper, 2013) out of huge amount of varied data that are in its raw format. Today, technologies of storage devices have attained a great height in storage capacity. The organizations are capable to store the data of big data Category. But only storage of big data is not enough for any organization. The benefit lies in getting the valuable insights from the stored data to help in making planning, decisions and other organization’s strategies. The valuable information can only be obtained by analysis of big data (Katal A., et al, 2013). Moreover, the data comes from variety

DOI: 10.4018/978-1-5225-0536-5.ch009
of resources such as smart phones, sensor embedded machines, social media sights, IT log Files, Web servers, email servers etc. Due to the large volume, Variety, Velocity, Variability and Complexity (Katal A., et al, 2013) of Big Data, the analysis task becomes very challenging and difficult. Thus, it requires advanced tools, techniques and specialized data analysis skills over the traditional tools and methods.

It has been obviously noticed that Relational Data Base Management Systems (RDBMS) can no longer handle all the data management issues posed by many currently running application software. It is mostly because of:

1. The large, and constantly increasing, amount of data needed to be stored by many companies/enterprises/diverse organizations.
2. The tremendous query workload needed to access and analyze these data.
3. The need of flexibility in the database schema.

With the arrival of Web 2.0 service, the amount of data managed by large-scale web services has grown exponentially, posing new challenges and infrastructure requirements. This has led to new programming paradigms and architectural choices, such as map-reduce and NoSQL databases/data stores, which make up two of the key peculiarities of the specialized extremely, distributed systems well-known as Big Data architectures. The basic computer infrastructures generally encounter complexity requirements, resulting from the need for efficiency and speed in processing over vast surfacing data sets. This is made possible by taking benefits from the features of new technologies, such as the automatic scaling and replica provisioning of Cloud environments. Although performance is the main issue for the considered applications. Precisely, NoSQL (Olivier Cure, et al. 2011) databases are the next generation databases mostly dealt with some of the points: being non-relational, distributed, open-source and horizontally scalable.

HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

- **Hadoop**: Hadoop (White, 2010) is an Apache Open Source software framework for working with big data. It was created by Doug cutting and Michael J. Cafarella. Doug who was working at yahoo at that time, it named after his son’s toy elephant. It was derived from Google technology and put to practice by yahoo and others. But big data is too varied and complex for a one-size-fits-all solution. While Hadoop has surely captured the greatest name, recognition, it is just one of three classes of technologies well suited to storing and managing big data. Hadoop is software framework which means it includes a number of components that were specifically designed to solve large-scale distributed data storage, analysis and retrieval tasks.

Basic Key and Important Terms Related to the Hadoop

- **Open-Source Software**: Open source software is dissimilar from commercial and proprietary software due to the wide and open network of developers that form and handle the programs. Usually, it’s free to download, use and contribute to, though more and more commercial versions of Hadoop are making available for use.

- **Framework**: Framework provides users everything needed to develop and run software applications – programs, tool sets, connections, etc.