Chapter 5
Role of Open Source Software in Big Data Storage

Rupali Ahuja
University of Delhi, India

Jigyasa Malik
University of Delhi, India

Ronak Tyagi
University of Delhi, India

R. Brinda
University of Delhi, India

ABSTRACT

Today, the world is revolving around Big Data. Each organization is trying hard to explore ways for deriving value out of huge pile of data we are generating each moment. Open Source Software are widely being adopted by most academicians, researchers and industrialists to handle various Big Data needs because of their easy availability, flexibility, affordability and interoperability. As a result, several open source Big Data tools have been developed. This chapter discusses the role of Open Source Software in Big Data Storage and how various organizations have benefitted from its use. It provides an overview of popular Open Source Big Data Storage technologies existing today. Distributed File Systems and NoSQL databases meant for storing Big Data have been discussed with their features, applications and comparison.

INTRODUCTION

The emergence of data from new data sources such as the Internet of Things, Sensor Networks, Open Data on the Web, Data from Mobile Applications, Social Network data have made traditional database management systems inadequate to handle large volumes of data sets. Moreover, the size of data being generated is expanding at speed of light increasing the need of new tools and technologies for handling Big Data.

DOI: 10.4018/978-1-5225-3142-5.ch005

Big Data technology provides the capability of capturing, analyzing and processing huge volumes of disparate data at the right speed and within the right time frame which allows real time analysis. For instance, BMW uses sensor data to inform its customers when their cars need to be serviced (Patten, 2015). The interest in Big Data in every field has risen across academicians, researchers and industry alike because of the value that may be generated by it.

Open Source Software is any software whose source code is publicly available for use, modification, sharing and re-distribution under a licensing policy (Wikipedia, 2016c). Open Source Software is the driving force behind the success of many Big Data applications because of their collaborative and knowledge sharing aspects. Many Open Source Software have been developed to cater to the needs of Big Data Storage, Processing, Handling, Analysis, Management and Visualization. Organizations are either using these software directly or customizing them according to their needs adding to the number of Open Source Software available today.

NoSQL databases and File Systems form the core component of Big Data Storage competency. A File System is a component of the Big Data stack which administers the Distributed Storage nodes for storing data into databases/data stores efficiently. Since data is distributed across networks, a file system is responsible for communication with requisite nodes and aggregating data from vast nodes to perform analysis and thereby generate the result. It also deals with Organization, Storage, Naming, Sharing and Protection of files. (Kune, Konugurthi, Agarwal, Chillarige & Buyya, 2016)

Big Data demands massive and specialized storage infrastructure due to its characteristics of high velocity, wide data variety and huge data volume. NoSQL databases are the physical storage house of large amounts of varied data coming from wide variety of sources and generated at a high velocity. NoSQL databases rein the Big Data Storage world. Depending on the type of data and velocity of data being generated by various types of sources, different types of NoSQL open source databases are available today.

This chapter focuses on the role of Open Source Software in Big Data Storage and various Open Source tools available for storing Big Data. Also, it lists some popular companies who have successfully exploited Open Source Big Data tools to establish, enhance and improve profitability of their business.

BACKGROUND

The amount of data generated each second is continuously growing at an exponential rate. Facebook, a social networking website, is home to 40 billion photos and more than 100 hours of videos are uploaded to YouTube every minute and these statistics are burgeoning at speed of light in almost every field increasing the interest and demand for Big Data Storage and management technologies. A new forecast from International Data Corporation (IDC) sees the Big Data technology and services market growing at a Compound Annual Growth Rate (CAGR) of 23.1% over the 2014-2019 forecast periods with annual spending reaching $48.6 billion in 2019 (IDC, 2016).

Open Source tools are playing prominent role in managing Big Data Storage issues. The most dominant technologies used in Big Data world, Hadoop and Apache Spark are Open Source tools. The most popular Big Data software distribution companies like Cloudera and HortonWorks have based their business around open source technologies. Open Source is the platform best suited for Big Data solutions. Almost all Big Data solutions work on top of UNIX Operating System which is open source. Without