Trending Big Data Tools for Industrial Data Analytics

T

A. Bazila Banu
Bannari Amman Institute of Technology, India

V. S. Nivedita
Bannari Amman Institute of Technology, India

INTRODUCTION

The Big Data in its natural form is of no usage. Therefore, now let us realize Big Data Analytics. It encapsulates past data into a form that individuals can easily read. This helps in generating reports, such as a company's profits and sales comparison based on quarterly/monthly and annually. It Is further classified into descriptive, predictive, and diagnostic analytics (Chunquan Li et al.,2021). Big data analytics tools are used in various stages of data processing such as Hadoop to collect and evaluate data, MongoDB to handle data that gets updated frequently, Talend for data incorporation and administration, Cassandra to handle aggregates of data, Spark for real-time administering to handle large volume of data in a distributed environment, STORM for time computational approach and Kafka for streaming and handling fault-tolerant storage. It's essential for any organization to find the best way to deal with the varied data sources and still meet the aims of the analytical process (Chunquan Li et al.,2021). This takes a shrewdness approach that integrates hardware, software, and procedures into a wieldy process that delivers results within an adequate time frame Storage is another critical element for Big Data. The data have to be stored in some repositories which can be readily available and secured. This has proved to be an exclusive challenge for many administrations, since network-based storage, such as SANS and NAS are very expensive to purchase and succeed. Storage has evolved to become one of the more unimaginative elements in the typical data center. However, today's enterprises are met with budding needs that can put the strain on storage technologies. But investing in open-source technologies are required to expand the business on a large scale. A case in point is the push for Big Data analytics, a concept that brings BI abilities to large data sets (F. Xu et al., (2019). For the decision maker looking to power Big Data, Hadoop solves the most communal problems linked with Big Data: storing and accessing large amounts of data in an efficient fashion. Big data Analytics is a collection of process that are associated with the industry. Despite of handling various characteristics of Big Data such as volume, velocity, variety, variability, veracity, value, visualization, validity, vulnerability, volatility etc. the Big Data tools are used to generate business outcomes benefitting the organization in various productive means.

DOI: 10.4018/978-1-7998-9220-5.ch032

BACKGROUND

Industrial Big Data

In the era of digital economic globalization, intelligent decision making has attracted a lot of attention from the digital industry market. One prime technology in artificial intelligence is big data driven analysis. This enhances the productivity and helps in making wise decisions by mining the hidden knowledge and the potential ability of the Big Data (M. Ghasemaghaei & G. Cali, 2019). Many real-time large-scale data are applied to the industrial process. Mostly the real-time data are streamed from noisy environment Also among the acquired data certain data will be labelled and few may not. Such kind of substantial amount of data with various challenges within are processed and expected to produce an optimized intelligent output without compromising the time and space dimensions. Hence the Big Data processing requires extensible methods to distribute and store real-time data, to suggest and dynamically adapt with the changes made in the process to provide automatic decisions (Ritu Ratra & Preeti Gulia, 2019). Thus, the End-to End Big Data process is expected to integrate, adapt and generalize the data in all stages within to create intelligent decisions with respect to the process.

Therefore, non-traditional techniques and strategies are required to store, organize and process the big data sets. There are several big data tools available, the following are the important big data analytics tools are highly recommended and applied in industry servicing various needs such as data collection, data cleaning, data filtering and extraction, data validation, and data storage.

Hadoop -- To collect and evaluate data.

MongoDB -- To handle data that gets updated frequently.

Talend -- To provide data incorporation and administration.

Cassandra -- To handle aggregates of data.

Spark -- To provide real-time administration while handling large volume of data in the distributed environment.

STORM -- To process high velocity data in distributed real-time computational environment.

The following context discusses the aforementioned tools in detail with working strategies, application possibilities, its merits and exceptions in order to point out how effective the tools are applied in Big Data Analytics.

APACHE HADOOP

Hadoop is an open-source framework developed by Apache Foundation. Apache Hadoop is adopted by various organizations for storing and processing humongous datasets. Hadoop outsmarted supercomputers and turned out to be the fastest system in sorting a terabyte of data in 2008.The journey of Hadoop started from the year 2002 with Apache Nutch Project (see Apache.org,2021). The Nutch's developers got inspired by Google's Google DFS & MapReduce and started implementing the open source framework, the Nutch Distributed File System (NDFS) and map reduce in the middle of 2004. Soon after Doug Cutting and Mike Cafarella developed an independent licensed sub project Hadoop. Hadoop initially contributed only 20 to 40 clusters later in 2006, Yahoo scaled the Hadoop project to more clusters. In 2007, Yahoo started using Hadoop on 1000 node clusters. Later Hadoop confirmed its success and became the most popular. Various releases on Hadoop evolved later. Hadoop 3.1.3 is the latest version of