Chapter 82
Need of Hadoop and Map Reduce for Processing and Managing Big Data

Manjunath Thimmasandra Narayanapppa
BMS Institute of Technology, India

A. Channabasamma
Acharya Institute of Technology, India

Ravindra S. Hegadi
Solapur University, India

ABSTRACT

The amount of data around us in three sixty degrees getting increased second on second and the world is exploding as a result the size of the database used in today's enterprises, which is growing at an exponential rate day by day. At the same time, the need to process and analyze the bulky data for business decision making has also increased. Several business and scientific applications generate terabytes of data which have to be processed in efficient manner on daily bases. Data gets collected and stored at unprecedented rates. Moreover the challenge here is not only to store and manage the huge amount of data, but even to analyze and extract meaningful values from it. This has contributed to the problem of big data faced by the industry due to the inability of usual software tools and database systems to manage and process the big data sets within reasonable time limits. The main focus of the chapter is on unstructured data analysis.

INTRODUCTION

Big data is the collection of datasets that are huge in size and difficult to handle by commonly used data processing tools and its applications. These datasets are unstructured and usually originated from various sources such as social media, scientific applications, social sensors, surveillance cameras, electronic health records, web documents, archives, web logs and business applications. They are larger in the size

DOI: 10.4018/978-1-5225-7501-6.ch082

with fast data in/out. Organizations would be interested in capturing and analyzing these datasets because they can add considerable value to the decision making process. However, such processing may involve complex workloads, which move the boundaries of what are possible using traditional data management and data warehousing techniques and technologies. Further, big data must have high value and ensure trust for decision making process. These data come from diverse sources and heterogeneity is one more important property besides volume, variety, velocity, value and veracity. Data gets collected and stored at unprecedented rates. Moreover the challenge is not only to store and manage the large amount of data, but even to analyze and extract meaningful values from it. This has contributed to the problem of big data faced by the industry due to the inability of usual database systems and software tools to manage and process the big data sets within reasonable time limits. Processing of Big data can consist of various operations depending on usage like culling, classification, indexing, highlighting, searching, faceting, etc.

Two significant data management trends for processing the big data are relational DBMS products meant for analytical workloads (also called analytic RDBMSs, or ADBMSs) and the non-relational systems (sometimes called NoSQL systems) meant for processing multi-structured data. A non-relational system can be used to generate analytics from big data or to pre-process big data before consolidated into a data warehouse.

Analytic RDBMS - ADBMS

An analytic RDBMS is an integrated solution for managing the data and generating analytics that offers better price/performance, simplified management and administration. The performance improvements are achieved by making use of massively parallel processing architectures, data compression, enhanced data structures and the capability to push analytical processing into DBMS.

Non-Relational Systems

Non-relational systems are useful for processing the big data where most of data is multi-structured. These are particularly popular with the developers who prefer to use procedural programming language, rather than a structured language such as SQL, to process the data. These systems support different types of data structures including document data, key-value pairs and graphical information.

One of the most important non-relational systems is the Hadoop distributed processing system introduced by open source Apache Software Foundation. Apache defines the Hadoop as, a framework for running applications on large hardware cluster built of commodity hardware. It includes a distributed file system (HDFS) which can distribute and manage bulky data across the nodes of a hardware cluster to offer high throughput. Hadoop makes use of the MapReduce programming model to divide the application processing into small fragments that can be executed on multiple nodes of same cluster to provide massively parallel processing. The Hadoop also includes Pig and Hive languages for generating and developing MapReduce programs. Hive includes Hive-QL, which provides subset of SQL.

The main focus of the chapter is on unstructured data analysis. The unstructured data is the information that either does not fit well into relational tables or does not have pre-defined data model. As compared to others the fastest growing type of data is the unstructured data. Some of examples are imagery, sensors, telemetry, video, log files, documents and email data files. Big data is a collection of techniques and technologies that involve new forms of integration to uncover hidden values from large datasets that are complex, diverse and of a massive scale. There are several techniques for gathering, storing, process-