Chapter 9
Programming and Pre-Processing Systems for Big Data Storage and Visualization

Hidayat Ur Rahman
University of Swat, Pakistan

Rehan Ullah Khan
Al Qassim University, Saudi Araba

Amjad Ali
University of Swat, Pakistan

ABSTRACT

This chapter of the book chapter provides detailed overview of the major concept used in Big Data. In order to process the huge volume of data, the first step is the pre-processing which is required to anomalies such as, missing values by applying various transformations. This chapter provides a detail overview of preprocessing tools used for Big Data such as, R, Yahoo! Pipes, Mechanical Turk, Elasticsearch etc. Beside preprocessing tools, the chapter provides detailed overview of storage tools, programming tools, data visualization, log processing tools and caching tools used for Big Data analytics. In other words, this chapter is the core of the book and provides the overview of the major technologies discussed later in the book.

INTRODUCTION

Big Data can be completely characterized by using three metrics i.e. volume, variety and velocity ((Dumbill, 2011); (Russom, 2011)). Volume refers to the amount of data which is beyond the capacity of traditional database management systems. Volume introduces the challenge of scalable storage and distributed computation, processing this large amount of data two approaches are used such as massively parallel approach (data warehouse or database) and distributed batch processing approach (Apache Hadoop). The database approaches are used for storing structure data while the distributed batch pro-

DOI: 10.4018/978-1-5225-3142-5.ch009

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

cessing approaches such as Apache Hadoop has no such restriction and it can be used for unstructured data. Due to the fast-moving data, real time analytics are required to analyze this streaming data which cannot be stored in databases. Various tools are used to handle the velocity of big data. Tools such as Twitter Storm and Yahoo S4 are widely used to handle the velocity problems.

Another challenge that arises in Big Data is to handle the variety of Big Data. Traditional databases such as Relational Database Management Systems (RDBMS) are used to store structural data while the data available on the internet are mostly unstructured or semi-structured. A mechanism is required to store the irregular data and extract useful information from these unstructured data. The NoSQL databases and RDF (Rich Data Format) are used to tackle this problem of data representation. This chapter is an important chapter of the book as it covers the overall process of Big Data i.e. from data acquisition to data visualization. The chapter provides overview of the various processing tools such as R, Yahoo Pipes, Datameer, Big Sheets, etc. Besides programming tools, it also covers the storage and visualization tools used for Big Data systems.

BACKGROUND

The term Big Data is commonly used for huge volumes of data which cannot be operated using traditional databases, they are beyond the capabilities of commonly used software tools to store, manipulate and process data within limited time (Garcia, 2015). There are often terabytes (Tb) or petabytes (Pb) of information stored in a single dataset. Some of the problems related with big data are capturing steaming data, storage, indexing, sharing and visualization. Enterprises use this high volume of data to extract useful knowledge using various software tools. However, as the size of the dataset increase, the difficulty in management also increases. In order to manage this huge amount of data, advanced tools and techniques are used. Since traditional data analysis and management tools are unable to exploit the data, it requires more sophisticated and specialized tools to store and manipulate data. Big Data analytics comprise of tools and techniques which helps in decision making process (Russom, 2011). Big Data analytics comprises of three main areas, the storage and management tools used for data, processing tools used for extracting useful information from the data and visualization. These three areas form different phases of a decision-making process in Big Data (Russom, 2011).

The first problem faced by the organization in managing Big Data is in regard to the storage of Big Data when the data is acquired. Traditional tools used for storing these data are data warehouses and data marts, data from various operational data sources are uploaded using Extract, Transform and Load tools (ETL), these ETL tools are used for extraction of information from external sources and then it applies certain transformation in order to remove the heterogeneity in data and make a common data layer and finally the load tools are used to load the data into the data warehouse. In order to analyze this data, various analytic tools are used but before applying the data to analytic tools it has to be cleaned and transformed (Chang & Dean, 2011). Several solutions to handling Big Data storage problems exist which ranges from Distributed Database Systems (DDBMS) to Massively Parallel Databases (MPDBMS), both DDBMS and MPDBMS ensures data and platform scalability and provides high performance query operation such as the Not only Structure Query Language (NoSQL) database (Chang & Dean, 2011).

NoSQL also known as non-relational database management system provides massive data scalability and is aimed to store unstructured and non-relational data. NoSQL databases separate data storage and data management tasks by providing high performance data storage and data management (Chang &

24 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/programming-and-pre-processing-systems-for-big-data-storage-and-visualization/198765

Related Content

Web Services, Service-Oriented Computing, and Service-Oriented Architecture: Separating Hype from Reality
John Ericksonand Keng Siau (2008). *Journal of Database Management (pp. 42-54).*
www.irma-international.org/article/web-services-service-oriented-computing/3390

Modeling Design Patterns for Semi-Automatic Reuse in System Design
Galia Shlezinger, Iris Reinhartz-Bergerand Dov Dori (2012). *Cross-Disciplinary Models and Applications of Database Management: Advancing Approaches (pp. 29-56).*
www.irma-international.org/chapter/modeling-design-patterns-semi-automatic/63661

Cross-Correlation Measure for Mining Spatio-Temporal Patterns
James Ma, Daniel Zeng, Huimin Zhaoand Chunyang Liu (2013). *Journal of Database Management (pp. 13-34).*
www.irma-international.org/article/cross-correlation-measure-for-mining-spatio-temporal-patterns/86282

A Model of Error Propagation in Conjunctive Decisions and its Application to Database Quality Management
Irit Askira Gelman (2012). *Journal of Database Management (pp. 103-126).*
www.irma-international.org/article/model-error-propagation-conjunctive-decisions/62034

Fabric Database and Fuzzy Logic Models for Evaluating Fabric Performance
Yan Chen, Graham H. Rongand Jianhua Chen (2008). *Handbook of Research on Fuzzy Information Processing in Databases (pp. 538-562).*
www.irma-international.org/chapter/fabric-database-fuzzy-logic-models/20367