Chapter 5
Handling Critical Issues of Big Data on Cloud

Madhavi Vaidya
VES College, India

ABSTRACT

Big Data is driving radical changes in traditional data analysis platforms. To perform any kind of analysis on such voluminous and complex data, scaling up the hardware platforms becomes impending. With the entire buzz surrounding Big Data; it is being collected at an unprecedented scale. Big Data has potential to revolutionize much more than just research. Loading large data-sets is often a challenge. Another shift of this Big Data processing is the move towards cloud computing. As many communities begin to rely on cloud based data management, large shared data goes up extensively. Analysis of such large data on distributed processing system or cloud is a bit difficult task to handle. The aim of this chapter is to provide a better understanding of the design challenges of cloud computing and analytics of big data on it. The challenge is related to how a large extent of data is being harnessed, and the opportunity is related to how effectively it is used for analyzing the information from it.

INTRODUCTION

Data is the information that has been translated into a form that is more convenient to move or process; this is the definition given on whatis.com. Wikipedia says it is a set of values of qualitative and quantitative variables; pieces of data are individual pieces of information. Generally, data is said as distinct pieces of information, usually formatted in a special way. The value of Big Data can only be extracted by data analytics. Although many different data analytics algorithms and techniques including statistical analysis, data mining, and machine learning can be on Big Data, they all rely on extremely intensive computations. Big data is a phenomenon that is characterized by the rapid expansion of raw data. The challenge is related to how a large extent of data is being exploited, and the opportunity is related to how effectively it is used for analyzing the information from it on cloud. The most useful approaches and categories of data tools which are to be chosen have been discussed in this chapter.

DOI: 10.4018/978-1-4666-9834-5.ch005

The Need of Data Processing

Firstly, we have to study why the data has to be processed. There are certain reasons for which the data is being processed. The data can be:

- **Incomplete:** Lacking attribute values, containing attribute data.
- **Noisy:** Containing errors or outliers.
- **Inconsistent:** Containing discrepancies in code or names.
- The quality data should be available.

To obtain the required information from huge, incomplete, noisy and inconsistent set of data is the need of data processing.

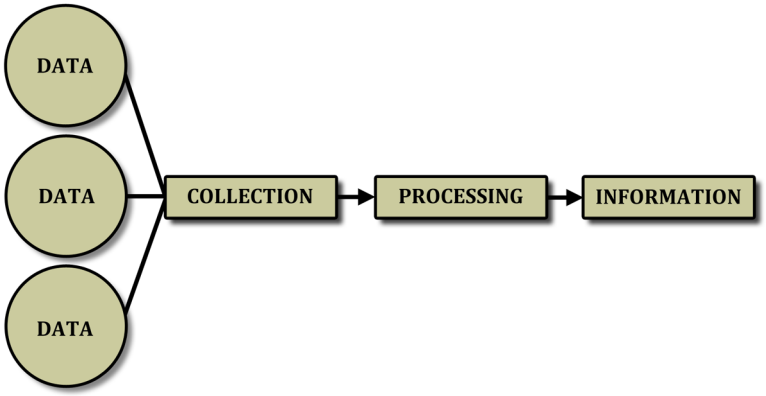
The steps of Data Processing:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Summarization

Let's study the above steps of Data Processing one by one which are depicted in Figure 1.

- **Data Cleaning:** Data cleaning is especially required when integrating heterogeneous data sources. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data.
- **Data Integration:** It combines data from multiple sources into a coherent data store, as in data warehousing. In short it is an integration of multiple databases, data cubes, or files.

Figure 1. Data Processing Flowchart



A Study on the Performance and Scalability of Apache Flink Over Hadoop MapReduce
Pankaj Latharand K. G. Srinivasa (2019). *International Journal of Fog Computing (pp. 61-73).*
www.irma-international.org/article/a-study-on-the-performance-and-scalability-of-apache-flink-over-hadoop-mapreduce/219361