Chapter 44
Efficient Big Data-Based Storage and Processing Model in Internet of Things for Improving Accuracy Fault Detection in Industrial Processes

Mamoon Rashid
https://orcid.org/0000-0002-8302-4571
School of Computer Science and Engineering, Lovely Professional University, India

Harjeet Singh
https://orcid.org/0000-0003-3575-4673
Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, India

Vishal Goyal
Department of Computer Science, Punjabi University, India

Nazir Ahmad
Department of Information Systems, Community College, King Khalid University, Saudi Arabia

Neeraj Mogla
Nike Inc. Global Headquarters, USA

ABSTRACT

As the lot of data is getting generated and captured in Internet of Things (IoT)—based industrial devices which is real time and unstructured in nature. The IoT technology—based sensors are the effective solution for monitoring these industrial processes in an efficient way. However, the real—time data storage and its processing in IoT applications is still a big challenge. This chapter proposes a new big data pipeline solution for storing and processing IoT sensor data. The proposed big data processing platform uses Apache Flume for efficiently collecting and transferring large amounts of IoT data from Cloud—based server into Hadoop Distributed File System for storage of IoT—based sensor data. Apache Storm is to be used for processing this real—time data. Next, the authors propose the use of hybrid prediction model of Density-based spatial clustering of applications with noise (DBSCAN) to remove sensor data outliers and provide better accuracy fault detection in IoT Industrial processes by using Support Vector Machine (SVM) machine learning classification technique.

DOI: 10.4018/978-1-6684-3662-2.ch044

1. CHAPTER OUTLINE

This chapter is structured around the concepts of efficient storage of sensor IoT based data and its processing in Big Data pipeline. Further the inclusion of novel prediction model will help to improve fault detection in IoT Industrial processes. In today's scenario we have to consider the big source of data generation as well as the plausible suitable platform of such huge data analysis. Therefore, the associated challenges are also included in this chapter.

1.1 Introduction to Big Data and IoT

Big Data Analytics along with Internet of Things (IoT) finds its use in areas of Smart Cities, Healthcare, Agriculture and Industrial Automation Units. The challenge of large amount of data generation in IoT devices is fulfilled by Big Data technologies in terms of its storage and processing (Chen et al., 2016). The advanced IoT devices and their applications have given rise to voluminous data in different varieties (Mavromoustakis et al., 2016). On the other side, Big Data technologies have discovered new kind of opportunities for developing IoT based systems (Rashid et al., 2013). Therefore, IoT based Systems and Big Data technologies integration will create new challenges in terms of storage and processing which needs to be addressed by the researchers (Singh et al., 2015).

1.1.1 Types of Big Data

The classification of Big Data is mostly given in terms of structure of data. The structure of data depends usually on its organization. Based on this, Big Data is classified into structured, unstructured and semi-structured data (Oussous et al., 2018). These types are explained below:

- **Structured Data:** Structured data is having fixed format and is easily stored, processed and accessed. Structured data is always following particular order as in row and column format and always results into ordered output. This data is easy to process as the format of data is always known in advance. All traditional databases containing data in row column format belong to this category.
- **Unstructured Data:** Unstructured data is usually huge data which is not in organized manner. This kind of data remains usually unknown and poses numerous challenges while processing for valuable insights as output. Moreover, this data is not having any kind of order and is raw in nature. Data in the form of images, audio, video and sensor-based data belong to this category.
- **Semi-Structured Data:** This kind of data usually contains both the forms but remain undefined. Usually this kind of data is not organized inherently at the beginning, but it can be turned into structure form while taking its analysis. Representation of data in terms of XML files belong to this category.

The different types of Big Data are shown in Figure 1.