An Efficient Distributed Data Processing Method for Smart Environment

C. Hemanth Kumar and A. Siva Sangari
Department of Information Technology, Sathyabama University, Chennai-600119, India;
hemanthkumar031@gmail.com, sivasangarikavya@gmail.com

Abstract

Background/Objectives: In current times, huge volume of data at a very high velocity gets generated through social media and various sensors in embedded systems that are connected to the Internet which causes a very Big data problem. These challenging Big data's need to be processed and stored by traditional Relational Database Management Systems (RDBMS). Due to this reason the need for new software solutions has emerged for managing the Big data in an efficient, scalable and smart way. Methods/Statistical Analysis: In this study, an approach to combine the concept of batch processing and stream processing to an end where we can query the data set which also supports Adhoc Querying with less latency, that can be run on any Large scale Machine Learning Algorithms for recognizing any interest pattern in the streaming data set was employed. The functionalities of Hadoop ecosystem's tool HIVE can also be used to produce the results to Adhoc queries, User Defined Functions (UDF) similar to writing a SQL Stored Procedures in the Spark System. An interface with SerDes which is Serialization and De-serialization that helps us to talk to the standard stream where we can exactly query the dataset are employed. Findings: By proposing a new software solution AllJoyn Lambda, in which AllJoyn is integrated in the lambda architecture and the prototype implementation of the architecture is done using Apache Hadoop Yarn over Apache Spark Streaming are presented . This study illuminates the high velocity streaming data set on a database without losing any data from the streaming domain, to support Adhoc Querying from the data set and to provide a mechanism for fast data processing and analytics using Large Scale Machine Learning. This paper highlights the analysis of large scale dataset processing, handling challenges, and its comprehensive systematic review. Applications/Improvements: From this study, we conclude that, building a smart environment by using the big data setup platform improves and enhances the results for the smart environment.

Keywords: AllJoyn Lambda Architecture, Big Data Analytics, Internet of Things, Smart Environment, Spark Streaming

1. Introduction

In engineering sector, Database management is gaining interest and achieving importance day by day becoming extremely important. So many researchers have worked on this subject and published number of papers and reports. Therefore, benefits of this field need to be utilizing fully by knowing its efficiency and state-of-the-art. Relational Database Management Systems (RDBMS) which are ideally suited for storing structured data have been used predominantly in the past decade. However, in the present day world, users and devices often generate a lot of semi-structured and unstructured data in various formats, that too at a very fast rate. In this context, it is essential to provide a distributed fault tolerant data architecture that satisfies the following requirements: To store the high velocity streaming data set on a database without losing any data from the streaming domain, to support ad hoc querying from the data set and to provide a mechanism for fast data processing and analytics using large-scale machine learning.

In recent past many authors have studied and worked on big data analytics by using various technologies on Hadoop platform. For example, Losup, Ghit, and Epema

*Author for correspondence
et al. reported processing of big data more efficiently by optimizing the MapReduce processing further to process huge amount of data with MapReduce programming model. Suzumura studied and worked on studies related to processing big data with standard sets of data by various applications. Semi structured cloud storage, non-structural storage and distributed file system was employed for big data processing in cloud computing environments were explored by Ji et al. The authors of reported “Shared disk big data analytics with Apache Hadoop” in which they retrieved valid information for huge data and uncover the hidden patterns by MapReduce Framework. The authors of reported "Addressing Big Data Problem Using Hadoop and Map Reduce” in which optimal solutions using MapReduce programming architecture for processing volumes of data and Hadoop cluster for storage were experimentally worked. To meet the requirements like processing large scale dataset, dataset handling challenges, a new software solution, AllJoyn lambda, which is integration of AllJoyn in the lambda architecture, has emerged. AllJoyn technology being used for executing various applications across heterogenous embedded devices, providing fascinating solutions for mobility, networking, security and dynamic configuration issues, it is not sufficient to handle complex smart environment and also communication among devices belonging to different broadcast domains is not being supported and also it does not provide a feature for Bigdata analytics, and storage. In order to overcome the shortcomings of the AllJoyn, a new scalable software solution has emerged by integrating AllJoyn in the Lambda Architecture, which is presented in this study. Hence it enables Bigdata storage, processing and real time analytics.

This paper addresses the designed/proposed software architecture which can be adapted to the evolution of a smart environment by its activities. Also it can be well fitted into the multiple IoT smart environments enabling the production of different context-aware applications and services.

2. The Methodology

2.1 Relational Data Processing in Spark: Spark SQL

Spark SQL can act as a distributed SQL query engine and provides a programming abstraction called Dataframes.

A collection of distributed data which are organized into named columns is called Data frames. With higher optimizations it can be relevant to a data frame in R/Python or table in a relational database. Huge number of sources like structured data files hive tables, external databases, or existing RDDs are used to construct Data Frames. Optimization rules and data sources acts as a catalyst to speed up the evaluation of Data frame API performing relational operations on both built-in collections of spark and external data sources.

High memory-efficient can be achieved by Spark SQL providing a columnar store for many aggregates than naive Spark code in computations expressible in SQL. This architecture shows that the database can be queried with the functionalities of Hadoop ecosystems tool. Hive can also be used to produce the results to ad hoc queries, User Defined Functions (UDF) similar to writing a SQL stored procedures in the Spark System. The interface with SerDes which is Serialization and De-serialization helps to talk to the standard stream where the dataset can be exactly queried. Figure 1 shows the system architecture.

Figure 1. System Architecture

2.2 Large Scale Smart Environments Management by AllJoyn Lambda Architecture

AllJoyn framework plays a significant role in enhancing and developing IoT systems which aimed at the interoperability among heterogeneous devices, creation of dynamic proximal network and execution of distributed applications. AllJoyn is an open source project which enables the development of multiple applications like sharing media, proximal applications, chat rooms, multi-player games, domestic and social applications on looking forward at the evolution of IoT by providing a common interface towards smart devices.

Important functions offered by AllJoyn are listed below:
ability to adapt the framework to specific devices;
- transferring data between devices through Bluetooth, Wi-Fi, and other communication technologies;
- efficient and secure exchange of data through D-Bus: Interoperability between different operating systems.

AllJoyn Architecture is presented in Figure 2.

Figure 2. AllJoyn architecture

Though the physical layer of AllJoyn architecture shows that it is responsible to manage IP connections, using different communication technologies like Bluetooth, Wi-Fi, it is not enough to manage complex smart IoT environments. The number of devices and information acquisition frequency increase, data management becomes quite tedious because AllJoyn does not support communications among devices belonging to different broadcast domains and does not provide any feature for storage of large amount of data and real time analytics which is called as Big Data Problem. Due to these limitations of AllJoyn, it is not able to address two main challenges of the IoT which includes large-scale smart environment management and Big Data storage and analytics. To overcome the shortcomings of AllJoyn, this paper propose a new scalable solution by integrating AllJoyn system with storm for processing real time data. This Architecture supports MongoDB for bulk storage and processing of data by three different types of data patterns which are given below.

- **Regular patterns** – Many embedded devices running AJTC applications which needs to be repeated at intervals by set of actions comes under regular patterns.
- **Event based patterns** – it has a set of action performed by several devices according to particular events E.g., keep windows open on detection of smoke by sensors.
- **Automated patterns** – set of actions that are triggered by the complex algorithms and not configured by users. In this, priority is given to each pattern. E.g., movement of curtain when it is exposed to light.

2.3 Discretized Streams

Stream processing by structuring computations as a set of short, challenging tasks instead of continuous operators can be avoided by D-Streams. Later, the state in memory across tasks as fault-tolerant data structures that can be recomputed deterministically and stored by them. Degradation of computations into short tasks results in exposing possibilities at a fine granularity and allows
powerful recovery techniques like parallel recovery and speculation. Other than fall tolerance the D-stream model gives other benefits like powerful unification with batch processing\textsuperscript{13}.

3. Conclusion

In this paper, two main challenges are discussed namely management of large-scale smart environment and big data storage/ analytics by proposing a new software solution, AllJoyn Lambda, mainly to overcome the limitations of AllJoyn. Thus the data generated at a very large scale can be stored and processed using the machine learning approach to manage the large scale environment efficiently.

4. References