Raspberry Pi Hadoop Cluster based Data Processing

Hemangi Karchalkar
Student, SPPU
S.K.N College of Engineering,
Pune, India

Poonam Railkar
Professor, SPPU University
S.K.N College of Engineering,
Pune, India

ABSTRACT
Implementing Raspberry PiHadoop Clusters to process data collected by sensor nodes and make them available to users via internet for real time decision making .Hadoop cluster is installed on multiple Raspberry Pi computers. This comprises installing Hadoop on each Raspberry Pi node, and then configuring it thereby allowing the nodes to all communicate with each other properly .Using Hadoop Map-reduce framework and HDFS(Hadoop Distributed File System) for collecting ,managing data from sensor nodes by running map-reduce and accessing the processed ,aggregated data via internet (Internet Of Things).

Keywords
Raspberry Pi,Hadoop, Map-Reduce, HDFS (Hadoop Distributed File System), Sensornodes, Cluster, Internet of Things (IoT).

1. INTRODUCTION
In Ubiquitous environment efficient data management, monitoring is must for data analyses. Different sensors are deployed for various applications and the sensor data can be accessed via internet, the sensor networks need giant data collection, management, monitoring .Apache Hadoop Map reduce is extremely useful for big data processing and management .Since the basic function of map reduce is to split the input data into large independent chunks which are further processed parallely .Therefore it extends its capabilities to various cloud environments. HDFS: Distributed file system written in Java for storing large chunks of data. The Raspberry Pi is a low cost, mini-computer that plugs into a computer. Raspberry Pi's are cheap and are generally used for research and educational purpose. They have Ethernet ports which are used to connect to internet and other Raspberry Pi clusters. Increasing, connecting multiple Raspberry Pi Hadoop clusters adds to the processing speed and computational power. The information sensed by the sensors and processed by Hadoop Map-reduce can be used for providing services which will help users to make real-time decisions.

2. LITERATURE SURVEY
Jamie Whitehorn [1] configured and installed Hadoopon a group of Raspberry Pi computers (clusters). He explained the implementation at the Strata andHadoop World conference wherehe sharedhis Hadoop Raspberry Pi idea. He discussed the problems a student comes across while learning, implementingthe Hadoop which is a distributed architecture requiring several computers.

In-Yong Jung, Ki-Hyun Kim, Byong-John Han, and Chang-Sung Jeong [2] implemented distributed sensor node management system using Hadoop Map-reduce framework and distributed file system (DFS).The paper states efficient ways for gathering, managing and processing sensor data by implementing specific Map-reduce applications on the various sensor nodes which will upload and retrieve the sensor node data to the Hadoop DFS as and when required.

Shaun Franks and JohnathanYerby [3] implemented a low-cost supercomputer with Raspberry Pi .The paper covered in detail the design, challenges and advantages of constructing a low-cost supercomputer.

Vijaykumar S, Dr. M Balamurugan, Ranjani K [4] proposed an idea of fault –tolerant, reliable process on Big Data using ARM and Hadoop framework. This research work includes implementation of high level data management at low cost.

P. Turton, T. F. Turton [5] implemented an idea of using Raspberry Pi as a supercomputer without the use of simulation or visualization techniques.

Mahesh Maurya, SunitaMahajan [6] used various map-reduce techniques like wordcount, pi, grep and basically provided a research analysis of Map-reduce algorithms.

Mamoru Sekiyama, Bong Keun Kim, SeishoIrie, and TamioTanikawa [7] made a portable IoT (Internet of Things) Device based on Raspberry Pi, used Hadoop as a database, the positional information system is found using data log system and RT-Middleware.

3. PROPOSED IDEA
A network of sensor nodes on which the data aggregation techniques are to be applied thereby supplying the useful data to the users via internet .The Master node functions as a manager which controls all the operations of HDFS .The Hadoop slave comprises of the Task tracer and data node which interacts with the master Hadoop node. Accessing and flushing of the sensor data in HDFS is done by using Hadoop commands. The big volume of data can be processed since data is spread across all nodes in HDFS. The huge volume of sensor data accumulated is processed by using map-reduce function which will break the data into large independent chunks and process it further as per user requirement. Hadoophas a built-in scheduler which allocates different map-reduce tasks or other tasks to idle nodes. All map-reduce tasks execute in parallel thereby increasing the speed, efficiency of the entire system .The aggregated, processed data is made available to the users via internet. The aggregation results of each individual raspberry pi Hadoop node are then sent to a master raspberry pi Hadoop node.

Big data software perform heavy tasks on clusters of networked computers .While sending data across a network

All the CPU'S in a cluster compete for resources such as memory, and on a system with limited resources the Raspberry Pi is a huge benefit. IoTis basically the result of smart machines communicating and interacting with each other, objects, environment resulting in generation and processing of data that when put into use can make life easier. Accessing the sensor data can result in various services like smart cars, smart energy systems, smart homes, smart

healthcare hospice centers .The basic requirements of an IoT common to every scenario include Sensing, data collection, processing, connection (wired /wireless),cooperation and communication among various devices connected and security across the network. The type of sensor node to be deployed is dependent of the type of application for which it is to be deployed. Each sensing node must carry a unique ID so that is can be accessed even remotely. Sensors +Connectivity =Made life easier. The figure 1 illustrates the proposed architecture of one Raspberry Pi Hadoop cluster .The data from various sensor nodes is collected and made fed into Raspberry Pi Hadoop slave nodes .Each Cluster consists of multiple Slave nodes and one master node .The slave nodes process the data using map reduce operations .The master node controls and manages the working of slave nodes and HDFS (Hadoop Distributed File System).The aggregated, processed data is then made available to the users via internet for real-time decision making tasks.

.

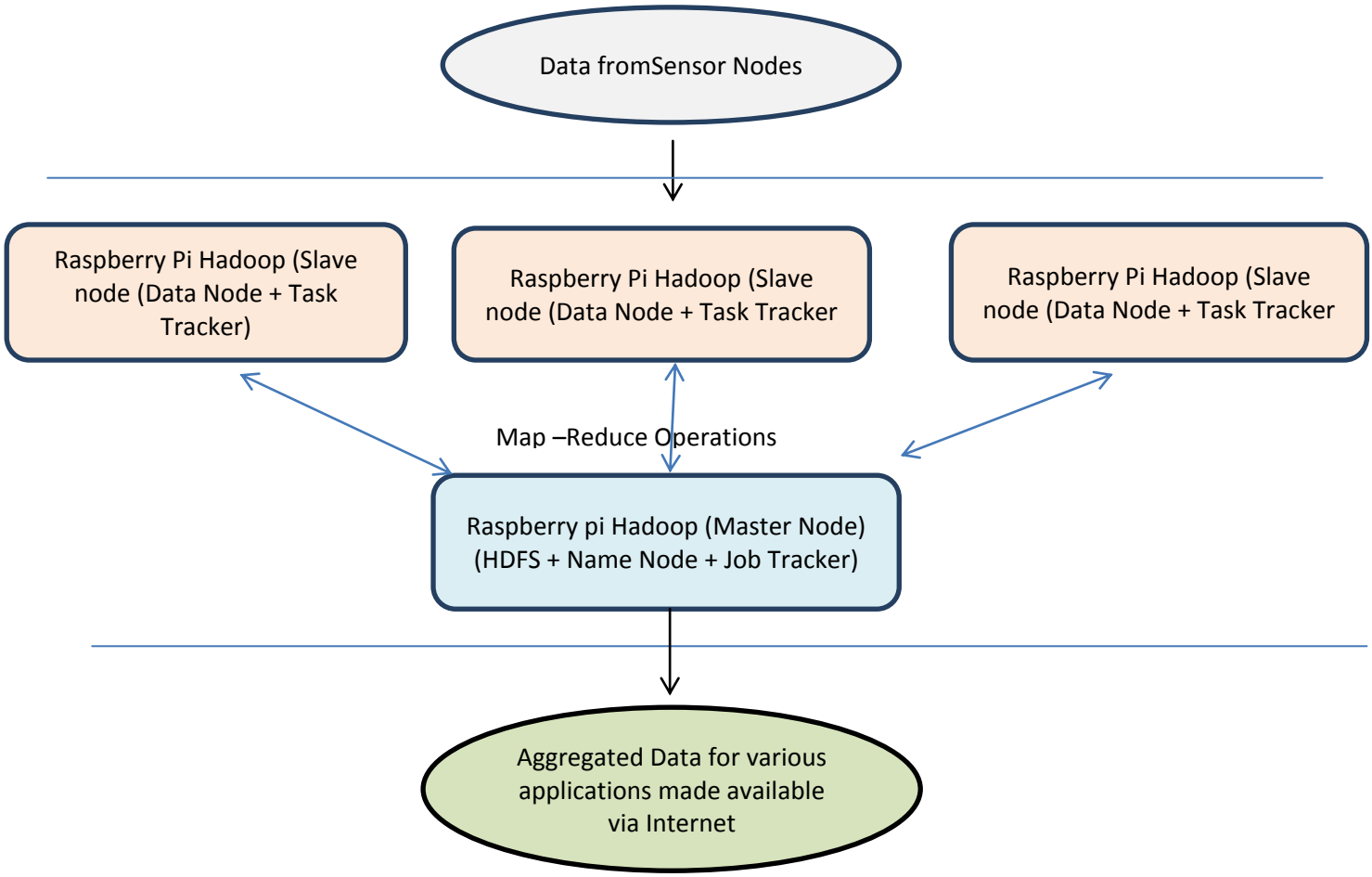

Fig 1: General Architecture of a single Raspberry Pi Hadoop Cluster

4. FEATURES

Due to small size, low cost of Raspberry-Pi and good performance,it isoften used to build a cluster. It has built-in Raspberry Pi Ethernet port, which allows you to connect it to a switch, router, or similar device. Multiple Raspberry Pi devices connected to a switch can then be formed into a cluster. Multiple Raspberry Pi combined adds to more CPU's processing to work with. The traditional systems consists of executing single instruction at a time using a single processor .Supercomputing using Raspberry Pi consists of breaking the large task into smaller task to process data simultaneously .By connecting multiple Raspberry Pi's together, the computational power is increased. High flexibility and fault-tolerant architecture. Scalable and exploiting Hadoop Map-reduce feature. By maintaining a cluster of Raspberry Pi Hadoop clusters if a single node fails the entire program doesn't fail.

5. APPLICATIONS

Adding more Raspberry Pi clusters, a comprehensive and robust cluster Supercomputer is formed.

Cognitive approach for data management in IoT.Forest Fire Detection. Monitoring of combustion of gases and fire conditions to alert vulnerable zones. Monitoring of CO_2 emissions by cars , pollution caused by factories and toxic gases. Snow level measurement in real time to assess the quality of ski tracks and take measures to prevent avalanche . Monitoring, prevention and evacuation in case of Landslides and Avalanches. Monitoring of soil moisture and patterndetection in land conditions. Liquid detection in warehouses, data centers andother sensitive areas to prevent corrosion and break downs. Detection, monitoring of radiation levels in nuclear power plants and their surroundings to generate leakage alerts.

Detection of leakages, gas levels in industrial areas, chemical

factories , inside mines and their neighboring areas .

Monitoring and evacuation during Earthquake and control in places of tremors.

6. CONCLUSION

The purpose of thispaper is to present sensor data management and accessing the required data via the internet by implementing Hadoop Map reduce across each Raspberry Pi Hadoop cluster .It offers efficient ways for managing sensor data by using map reduce applications and uploading, retrieving of data from HDFS (Hadoop Distributed File System).By connecting multiple Raspberry Pi Hadoop Clusters the computational power, efficiency is increased. The aggregation results of each individual raspberry pi Hadoop cluster are then sent to a master raspberry pi Hadoop cluster. The aggregated, useful data is then made available to the users via internet for decision making in real time activities.

7. ACKNOWLEDGEMENT

I would like to give my sincere thanks to my teachers and Mrs. PoonamRailkar Ma'am for her constant support.

8. REFERENCES

[1] Alex Woodie, George Liopold, Steve Conway, "Hadoop on a Raspberry Pi-Isaac Lopez" "http://www.datanami.com/2013/11/27/hadoop_on_a_raspberry_pi/"

[2] In-Yong Jung, Ki-Hyun Kim, Byong-John Han, and Chang-Sung Jeong "Hadoop-Based Distributed Sensor Node Management System"Volume 2014(2014), Article ID 601868

[3] Shaun Franks, JohnathanYerby "creating a lost-cost supercomputer with Raspebby Pi'

[4] Vijaykumar S, Dr. M Balamurugan, Ranjani K "Big Data: Hadoop Cluster Deployment on ARM Architecture" (IJARCCE) Vol 4, June 2015

[5] P. Turton, T. F. Turton, "PIBRAIN - A COST-EFFECTIVE SUPERCOMPUTER FOR EDUCATIONAL USE"

[6] Mahesh Maurya, SunitaMahajan, "Performance analysis of Map-Reduce Programs on Hadoop cluster" Year: 2012, IEEE Conference Publications

[7] Mamoru Sekiyama, Bong Keun Kim, SeishoIrie, and TamioTanikawa,"Sensor Data Processing Based on the Data Log System Using the Portable IoT Device and RT-Middleware" Year: 2015, IEEE Conference Publications