Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

In-Mapper combiner based MapReduce algorithm for processing of big climate data

Gunasekaran Manogaran [a], Daphne Lopez [b], Naveen Chilamkurti [c,*]

[a] University of California, Davis, USA
[b] School of Information Technology and Engineering, VIT University, Vellore, India
[c] Department of Computer Science and Computer Engineering, LaTrobe University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 31 October 2017
Received in revised form 13 January 2018
Accepted 25 February 2018
Available online xxxx

Keywords:
Big data
Internet of Things
Weather sensor devices
MapReduce programming
Model
Hadoop distributed file system

ABSTRACT

Big data refers to a collection of massive volume of data that cannot be processed by conventional data processing tools and technologies. In recent years, the data production sources are enlarged noticeably, such as high-end streaming devices, wireless sensor networks, satellite, wearable Internet of Things (IoT) devices. These data generation sources generate a massive volume of data in a continuous manner. The large volume of climate data is collected from the IoT weather sensor devices and NCEP. In this paper, the big data processing framework is proposed to integrate climate and health data and to find the correlation between the climate parameters and incidence of dengue. This framework is demonstrated with the help of MapReduce programming model, Hive, HBase and ArcGIS in a Hadoop Distributed File System (HDFS) environment. The following weather parameters such as minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity are collected for the study are Tamil Nadu with the help of IoT weather sensor devices and NCEP. Proposed framework focuses only on climate data for 32 districts of Tamil Nadu where each district contains 1,57,680 rows and so there are 50,45,760 rows in total. Batch view precomputation for the monthly mean of various climate parameters would require 50,45,760 rows. Hence, this would create more latency in query processing. In order to overcome this issue, batch views can precompute for a smaller number of records and involve more computation to be done at query time. The In-Mapper based MapReduce framework is used to compute the monthly mean of climate parameter for each latitude and longitude. The experimental results prove the effectiveness of the response time for the In-Mapper based combiner algorithm is less when compared with the existing MapReduce algorithm.

1. Introduction

Big data analytics have been playing a significant role in various fields with increased concentration from government sectors, business enterprises, and research centers. This section elaborates on how big data is expected to rise in the future. Recently, big data analytics is being used to gain more important hidden values from various environments that include healthcare analytics, environment and natural resource management, public sector units, business enterprises, government organizations, social networking and computational platforms [1]. Data generation sources in healthcare departments have been increasing dramatically.

These data generation sources generate a variety of data such as pharmaceutical data, electronic medical records, scanned images, data on individual food and dietary preferences, data on exercise patterns, financial details and so on [2–4]. A combination of all these data becomes big data. This combination helps taking good decisions in disease diagnosis, healthcare services, drug recommendation, healthcare delivery and drug interventions [5]. Fig. 1 represents the sources of big data.

Climate change plays a significant role in the day-to-day life of human beings. Data scientists highlight global climate changes and their implications [6]. Understanding of a global climate change requires researchers from multi-disciplinary domains [7]. Especially, individuals from climate modeling, data analytics, and database management system are required to process such huge volume of big data and analyze the global climate change [8]. In recent years, researchers and practitioners from multi-disciplinary domains have been focusing on potential climate change circumstances. Fig. 2 represents the relationship between IoT and big data analytics.

EnviroAtlas is a geospatial software that provides climate information relating to all over the world. Environmental management department from the United States has developed the EnviroAtlas [9–11]. This software provides information on future climate change scenario in a comprehensible manner. Pickard et al. have

* Corresponding author.
E-mail addresses: gmanogaran@ucdavis.edu (G. Manogaran), n.chilamkurti@latrobe.edu.au (N. Chilamkurti).

https://doi.org/10.1016/j.future.2018.02.048
0167-739X/

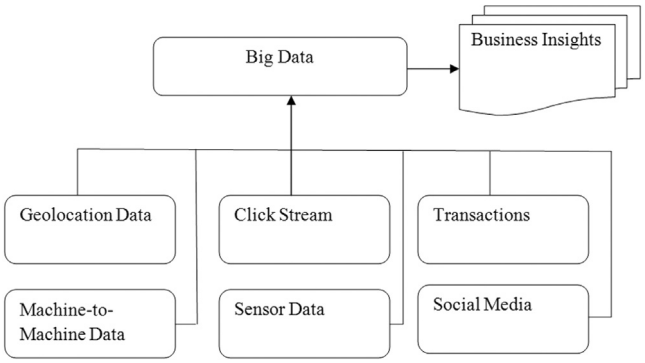

Fig. 1. Sources of big data.

discussed the impact of climate changes on clinical care, humanity and ecology [12]. Climate Analytics-as-a-Service is developed by NASA to provide scalable, high-performance data management services, software application virtualization and domain synchronized API for big climate data [13]. In addition, a team from EU-BrazilCC project has developed a PDAS platform to process the huge volume of biodiversity and climate change data [14].

The volume of the data generated from healthcare department has been increasing considerably. Hence, analytics on healthcare data is becoming more complex in everyone's life [15,16]. The more useful patterns are used in disease diagnosis such as classification of heart diseases, genome sequencing, microarray analysis, next generation sequencing, and processing of various medical scanned images such as MRI, CT, PET and Ultrasound [17,18]. Bates et al. have discussed the various uses of big data in healthcare such as reducing the cost to patients, re-admissions, unfavorable events, and clinical care optimization for diseases upsetting various organs [19]. Similarly, Raghupathi et al. have discussed the uses of big data in healthcare including hospital management and service delivery, patient activities, medical decision making, and sustainable services [20]. Jee and Kim have identified various approaches to reshape the traditional healthcare system. These include proper treatment course, reduced clinical care services, medical resource management, and development of healthcare systems [21,22].

The objective of this work is to develop a big data processing and prediction framework of dengue arising as a result of climate change. The framework is demonstrated with the help of MapReduce programming model, Hive, HBase and ArcGIS in a Hadoop Distributed File System (HDFS) environment. This framework is used for finding the correlation between the climate parameters and the incidence of dengue. This paper is structured as follows: Section 1 provides an overview of big data and its characteristics, followed by the big data challenges and potential solutions, big data applications, and big data analytics in climate change and healthcare. Section 2 provides an extensive review of literature relating to big data processing framework, big data predictive modeling approaches, and big data architectures for climate change. Section 3 proposes a big data processing framework for finding the correlation between the climate parameters and dengue incidence. In Section 4, the framework is demonstrated with the help of MapReduce programming model, Hive, HBase and ArcGIS in a Hadoop Distributed File System (HDFS) environment. In Section 5, the proposed framework is compared with various existing scalable architectures in the big data environment. Section 6 concludes the research work.

2. Review of literature

In this section, an exhaustive survey of research on big data processing and prediction framework is presented with an example. Yao et al. have proposed a novel data reduction algorithm based on MapReduce programming model [23]. A hierarchical encoded decision table is used in the proposed framework to identify the significant features from the big data. The result generated from this approach is supplied to the data reduction method for reduction of the size of the data. Schnase et al. have developed the big data based spatial indexing approach called MERRA to integrate weather data with various climate models to generate a Spatio-temporal mixture of 26 weather parameters [24]. The proposed big data based spatial indexing approach is widely used by researchers doing research in climate change analysis and health decision making systems. Li, Hu, Schnase, Duffy, Lee, Bowen and Yang have developed a scalable and distributed Spatio-temporal indexing approach on the basis of the MapReduce programming model to process the large size of climate data [24]. In this framework, the observational data is stored directly into the HDFS in original file format. Li, Huang, Carbone and Hu have proposed a scalable, high-performance query processing framework for the enhancement of integration of Hive and cloud computing technologies [25]. The proposed framework demonstrated as a climate analysis method, consists of a SQL-style query to process the terabytes of climate data. Zhu has used ArcGIS software to model the climate change and weather prediction [26]. In this framework, the Antarcticas land surface is taken to model the climate change. Initially, the Antarcticas land surface is divided into a number of polygons and counties, followed by the performance of regression analysis on the incidence data collected from British Antarctic Survey. Hajat et al. have used a storyboard method to develop a spatial database server to query, store and process the spatial locations [27].

Fig. 2. Relationship between IoT and big data analytics [1].

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

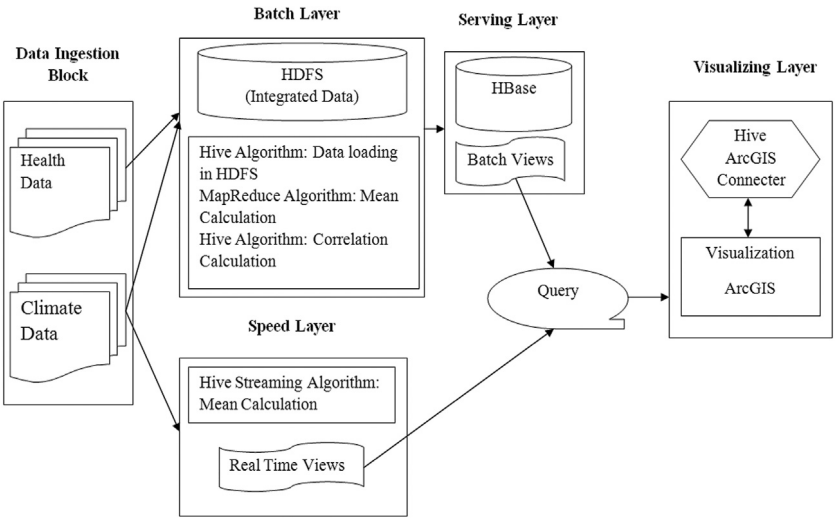

Fig. 3. The proposed big data processing framework.

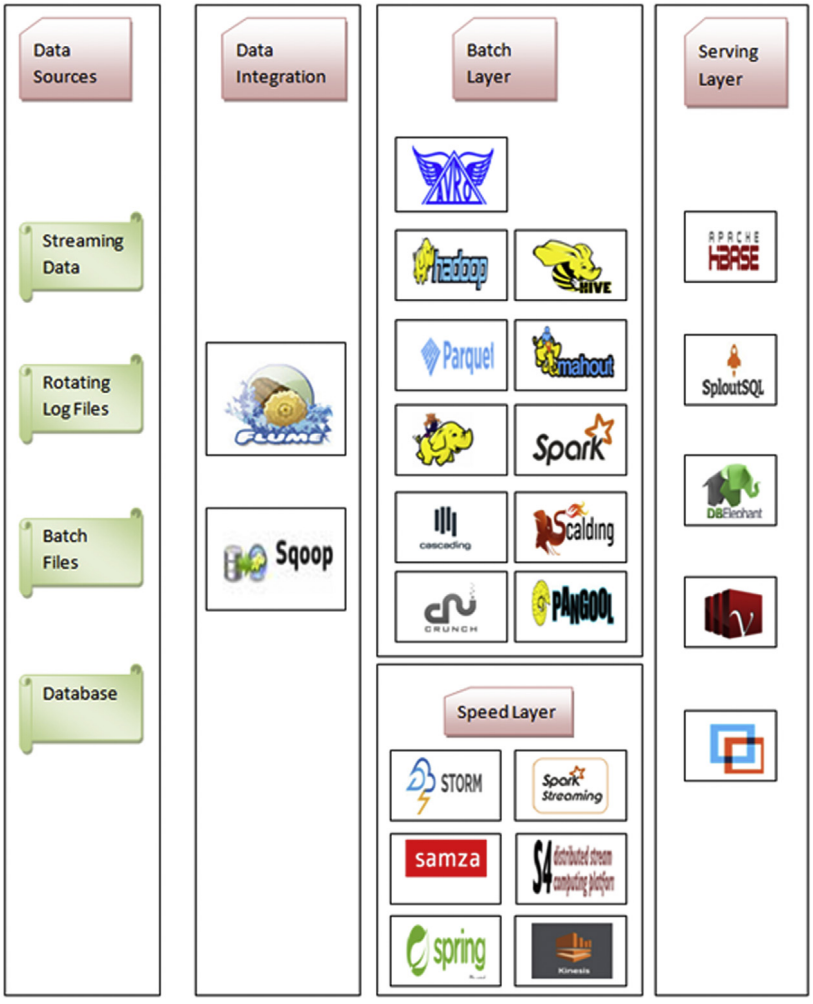

Fig. 4. Components of proposed framework.

The spatial database server is capable of handling user queries with a browser and GUI. This browser is also used for visualizing the regression results in a geospatial information system environment. Chasparis and Eldawy have studied the impact of climate change using big data analytics [28]. In their article, a case study was used to explain the global climate change and its implications. Ackermann et al. have proposed a scalable big data processing framework called Jet platform [29]. The primary goal of the Jet platform is to process batch data of a huge size in a distributed manner. Gao et al. have developed a high performance and scalable

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

```
Date,Longitude,Latitude,Elevation,MaxTemperature,MinTemperature,Precipitation,Wind,Relative
Humidity,Solar
1/1/1979,79.0625,13.8942003250122,409,24.001,17.365,0.3278742264,1.566199145656,0.8886190936,12.30403348,
1/2/1979,79.0625,13.8942003250122,409,25.817,14.111,0.0102996792,1.835927473671,0.873943504333,19.493721,
1/3/1979,79.0625,13.8942003250122,409,25.231,12.991,0,1.90932099044504,0.850141727895694,20.09739132,
1/4/1979,79.0625,13.8942003250122,409,25.162,12.907,0,1.86838639889342,0.835131190267102,20.104311348,
1/5/1979,79.0625,13.8942003250122,409,25.908,11.847,0.0781059168,1.6394340241707,0.802525084434,20.14002,
1/6/1979,79.0625,13.8942003250122,409,25.51,13.303,0.1441958544,1.87781918316264,0.834196067495133,18.93,
1/7/1979,79.0625,13.8942003250122,409,26.117,12.981,0,1.6757060896263,0.795431155405995,20.292809688,
1/8/1979,79.0625,13.8942003250122,409,26.701,14.226,0,1.49576174732698,0.766072592823415,20.412171432,
1/9/1979,79.0625,13.8942003250122,409,27.036,15.321,0,1.86507940597598,0.740040020348654,20.148522372,
1/10/1979,79.0625,13.8942003250122,409,25.813,13.84,0,2.09980346349538,0.80843842463949,20.31598458,
1/11/1979,79.0625,13.8942003250122,409,25.215,12.002,0,1.99474999541674,0.794919321398066,20.483435052,
1/12/1979,79.0625,13.8942003250122,409,25.506,10.974,0,1.594475807156,0.790534103060224,20.69099928,
1/13/1979,79.0625,13.8942003250122,409,26.329,14.381,0,1.71615448541087,0.761920371234461,20.681832168,
1/14/1979,79.0625,13.8942003250122,409,26.614,12.775,0,1.99117412616105,0.767678930018668,20.69740926,
1/15/1979,79.0625,13.8942003250122,409,26.903,13.428,0,2.10653682505959,0.723011561155546,20.887990308,
1/16/1979,79.0625,13.8942003250122,409,27.077,13.008,0,1.6733786779527,0.767045136865706,20.72435904,
1/17/1979,79.0625,13.8942003250122,409,27.086,13.333,0,1.82442404034068,0.808374437532221,20.569895568,
```

Fig. 5. Raw weather station data.

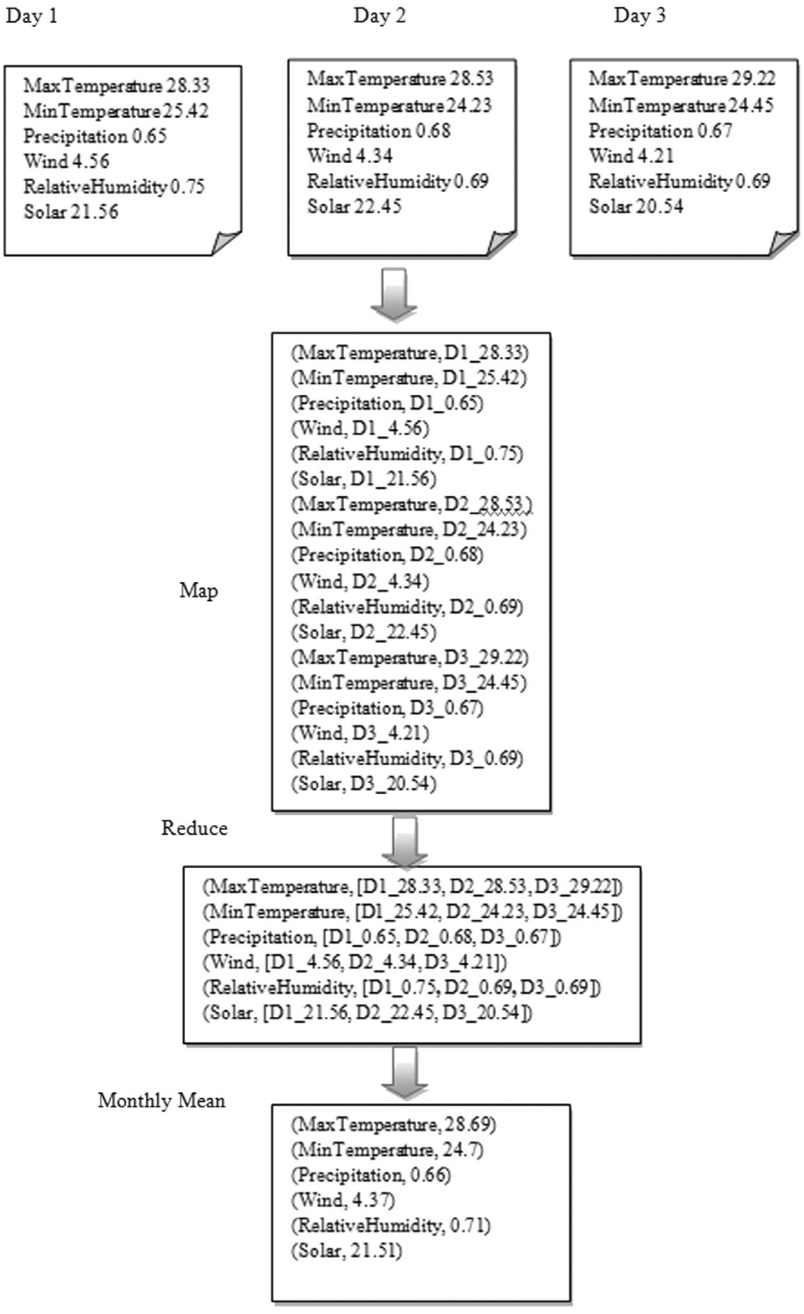

Fig. 6. Key and value pair for the batch view generation.

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

Table 1
Monthly average of climate data.

Month	Long	Lat	MaxT	MinT	Wind	Precip	Solar	Hum
Jan2000	79.14	12.93	34.5	28.5	15	0.4	22.3	0.45
Feb2000	79.14	12.93	33.4	27.5	16	0.3	22.5	0.49
Mar2000	79.14	12.93	23.2	30.1	17	0.4	22.7	0.53
Apr2000	79.14	12.93	35.6	31.2	14	0.5	23.6	0.56
May2000	79.14	12.93	36.4	32.5	13	0.6	23.8	0.58
Jun2000	79.14	12.93	37.3	33.6	12	0.7	23.9	0.62

Table 2
Monthly dengue cases.

Month	Long	Lat	Dengue
Jan2000	79.14	12.93	9
Feb2000	79.14	12.93	12
Mar2000	79.14	12.93	8
Apr2000	79.14	12.93	15
May2000	79.14	12.93	38
Jun2000	79.14	12.93	32

spatial computing platform based on the Hadoop distributed file system for processing the huge geospatial based workflows [30]. The essential goal of this framework is to collect crowd-sourced gazetteer entries with geo-locations. Zheng et al. have presented a big data architecture that could integrate and process streaming data with the help of a cloud computing environment [31,32].

Marz and Warren developed the Lambda Architecture for implementing big data systems [33]. Nathan Marz has implemented the Lambda Architecture on distributed data processing systems for Twitter data analytics. Martínez-Prieto et al. have identified a Service-On-Line-Index-Data (SOLID) architecture to manage big semantic data in real-time [34]. SOLID uses a huge size data storage block for storing big semantic data, which is indexed to allow high-speed querying. Krämer and Senner have identified the Modular Software Architecture for processing a large scale geospatial datasets in the cloud [35]. The proposed architecture is flexible and it supports variety of recent big data frameworks such as MapReduce, in-memory computing or agent-based programming. Gorton and Klein have developed the MongoDB-based Healthcare Data Management Architecture to increase availability and reduce latency for globally distributed users [36]. Vatsavai et al. have reviewed various spatial data mining algorithms for big data analytics such as Gaussian process learning, spatial autoregressive models, and Gaussian mixture models [37,38]. These algorithms are discussed in the light of computational and input and output requirements.

Moreover, The IoT devices sense the patients' health status and then transfer the clinical data to doctors and care holders [39]. This data is most often used for disease diagnosis and clinical care [40,41]. University of Virginia, has recently developed the Alarm-Net infrastructure to monitor patient health in an assisted-living environment [42–44]. UbiMon is the healthcare project originally developed for observing the patients' health condition in a continuous manner. MobiCare is another project originally developed on the basis UbiMon for monitoring the individuals' health condition on a continuous basis [16,45,46]. Recently, CodeBlue project has been developed by Harvard Sensor Network Lab to monitor individual health [47].

Table 4
Pearson correlation.

Correlation coefficient value r	Correlation type
Value is near ± 1	Ideal
Value lies between ± 0.50 and ± 1	Strong
Value lies between ± 0.30 and ± 0.49	Average
Value lies below ± 0.29	Minute
Value is zero	No correlation

3. Climate big data processing framework

Big data describes the enormous volume of structured, semi-structured and unstructured data that cannot be processed by traditional data processing tools and technologies [48]. Nowadays, the data generation sources generate the enormous amount of data. These data generation sources generate data which is not in the structured format; hence, difficulty, size, heterogeneity, appropriateness and privacy with big data cause issues in every stage of data processing [49,50]. Hence, there is a need to develop a big data architecture to process such voluminous data. It requires scalable parallel and distributed computing platforms such as Hadoop MapReduce, Hive, HBase, Strom, Spark, and Cassandra to process and gain more useful hidden information [51]. Climate simulation plays a vital role in public health [52]. Meteorological data collected from various weather stations and meteorological centers provide valuable universal information [53,54]. The data collected from the weather stations are often used to forecast the short-term weather. Processing of huge climate data provides even more valuable hidden information. However, the traditional data mining algorithms and statistical methods are not adequate to store and process the big climate data [55]. There is need for a scalable distributed framework to store and process the big climate data and attain more meaningful change information in the seasonal climate [56]. Moreover, Geographical Information System (GIS) is widely used to utilize the climate data in an efficient manner [57]. However, the weather stations and metrological centers are continually generating an enormous amount of data that cannot be visualized by traditional GIS platforms. Hence, there is need for a scalable distributed geospatial environment to utilize the climate data. Recently, Faghmous and Kumar have discussed the issues and challenges introduced in the processing of the big climate data [8,58]. In recent years, researchers from multidisciplinary environments have intended to use big climate data to predict the water and vector-borne diseases such as dengue, malaria, and chikungunya. For example, Hay et al. have discussed the potential opportunities to use big data analytical methods to model the infectious diseases [59].

Table 3
Integrated data.

Month	Long	Lat	MaxT	MinT	Wind	Precip	Solar	Hum	Dengue
Jan2000	79.14	12.93	34.5	28.5	15	0.4	22.3	0.45	9
Feb2000	79.14	12.93	33.4	27.5	16	0.3	22.5	0.49	12
Mar2000	79.14	12.93	23.2	30.1	17	0.4	22.7	0.53	8
Apr2000	79.14	12.93	35.6	31.2	14	0.5	23.6	0.56	15
May2000	79.14	12.93	36.4	32.5	13	0.6	23.8	0.58	38
Jun2000	79.14	12.93	37.3	33.6	12	0.7	23.9	0.62	32

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

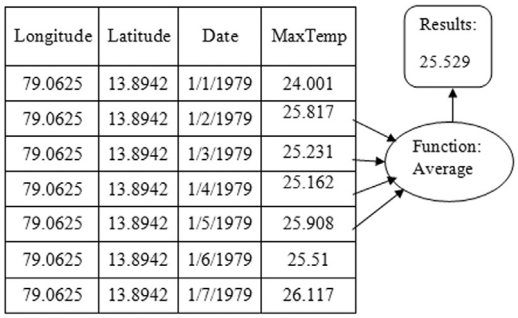

Fig. 7. Batch view precomputation.

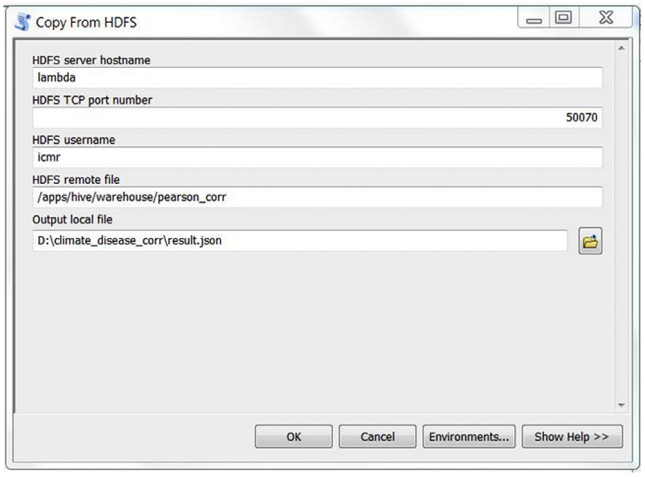

Fig. 8. Copy from HDFS tool.

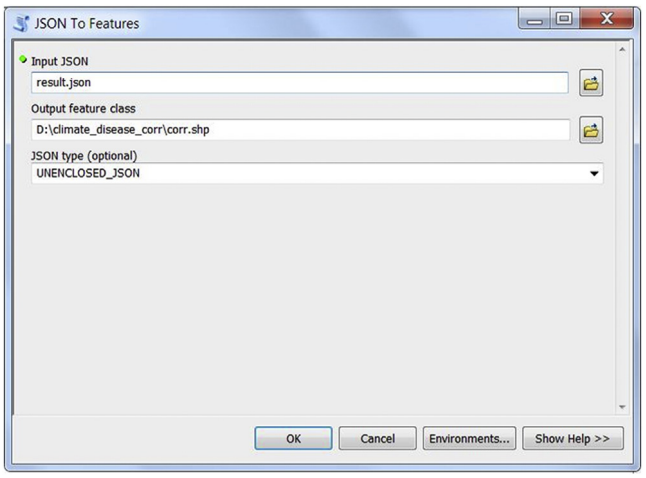

Fig. 9. JSON to features tool.

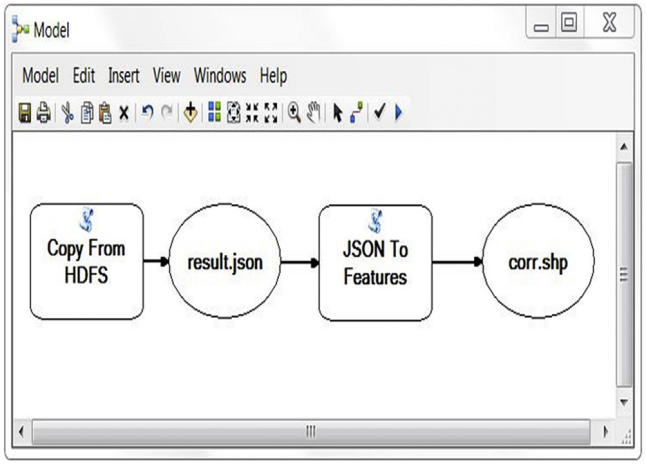

Fig. 10. ArcGIS model builder.

3.1. Data processing

The intention of the proposed framework is to model the correlation between the climate parameters and dengue incidence. Fig. 3 represents the climate data processing framework. The framework is demonstrated in a Hadoop distributed file system environment with the help of Hadoop MapReduce, Apache HBase, Apache Hive and Hive ArcGIS API. The framework consists of five different layers, namely, batch layer, speed layer, serving layer and visualization layer. The batch layer is implemented with the help of Apache Hadoop MapReduce and Hive. Apache HBase and Apache Hive streaming are used for implementing the serving layer and speed layer respectively. ArcGIS10.2 software is used for the implementation of the visualization layer with the help of Hive ArcGIS API. The framework is implemented in a Hadoop multi-node cluster environment.

3.2. Components of climate data processing framework

Fig. 4 represents the supporting programming tools for developing the batch layer, serving layer and speed layer of the framework.

3.3. Layer wise components

The namenode is used for controlling the namespace of the distributed file system and storing the details relating to the location and size of the data block. The tasks performed at the namenode include name changing, opening, and closing of directories and records. The essential role of datanode is to manage storage facilities, and execute read and write functions on the HDFS. The namenode instructions followed by the datanodes include creation, deletion, and replication of blocks. The datanode sends the heartbeat message to the namenode. The heartbeat message consists of block information relating to block size, the number of available blocks and the number of the task being done. The secondary namenode is used for updating the namespace information into the fsimage file. The goal of the secondary namenode is to increase the restarting speed of the namenode while the role of the jobtracker is to perform the job scheduling operations for the map and reduce the task on the datanodes. The jobtracker also monitors the task failure and reschedules those tasks on various available datanodes. The jobtracker runs over the namenode and determines the location of the data. The tasktracker is used for executing the tasks assigned by the namenode. The execution status of the tasktracker is periodically reported to the namenode. In other words, the essential role of the task tracker is to run the map or reduce tasks. Hive is also used for generating the batch views in the lambda architecture. A Structured Query Language (SqL) interface is used in Hive to analyze, query and summarize the large volume of data in the Hadoop environment. Hive uses the query language named HiveQL. The role of the HiveQL is to convert the SQL like queries

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

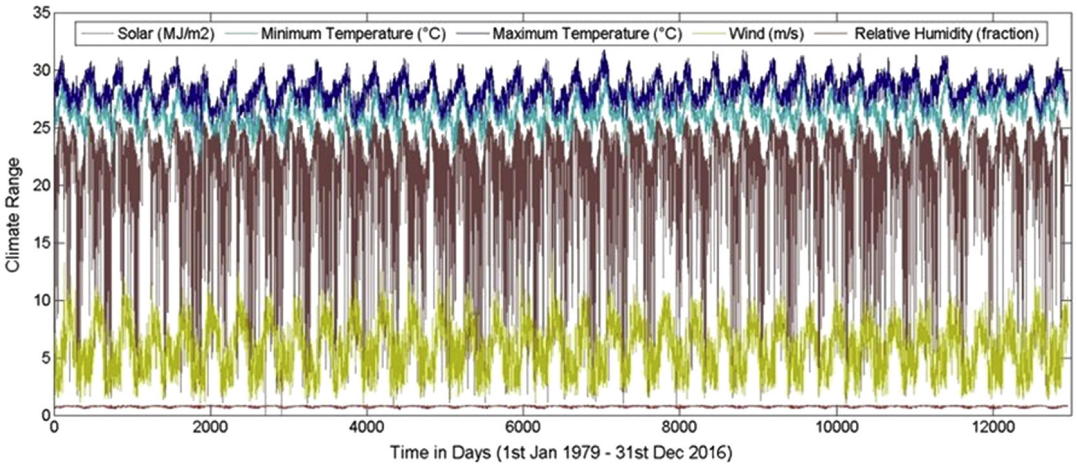

Fig. 11. Climate data (Jan 1979–Dec 2016).

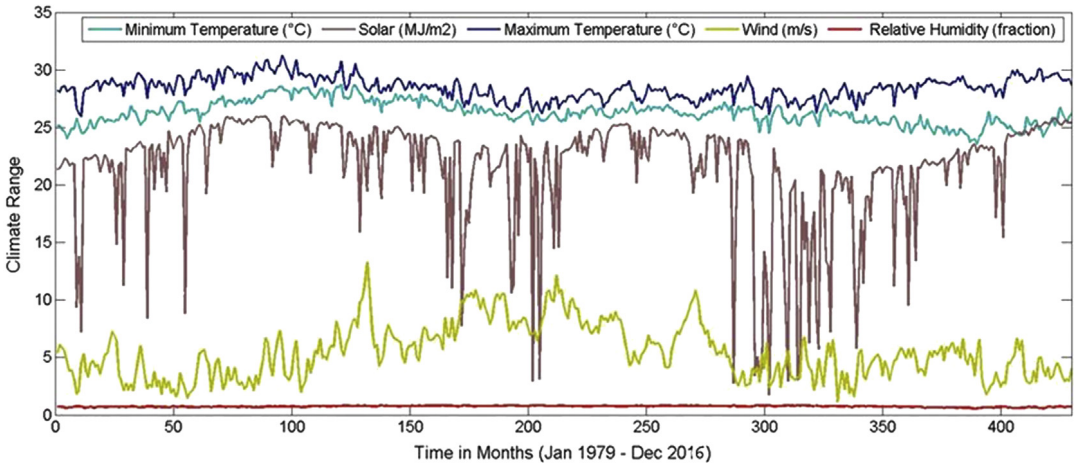

Fig. 12. Monthly mean climate range (Jan 1979–Dec 2016).

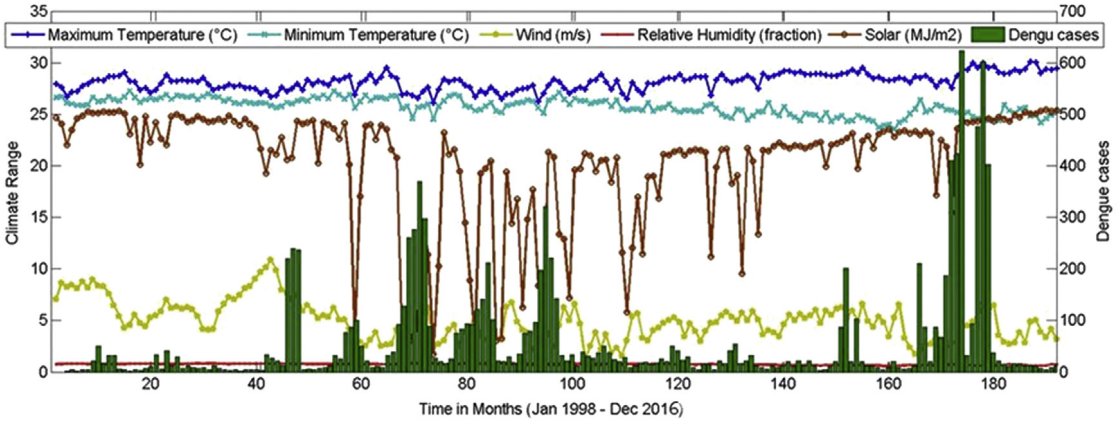

Fig. 13. Integration of monthly dengue cases and mean climate data (Jan 1998–Dec 2016).

into MapReduce jobs in the hadoop environment. Apache HBase is a database identified after the Googles BigTable was developed. Apache HBase can store millions of rows and columns in a scalable and distributed environment. Master/Slave architecture is also followed by Apache HBase to provide high scalability and manage the node failure. Fig. 4 represents the tools that are available to implement the climate data processing framework.

4. Implementation

4.1. Data ingestion block

National Centers for Environmental Prediction (NCEP) and IoT weather sensor devices have been collecting the weather data for the entire globe. As shown in Fig. 5, weather data consist of

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

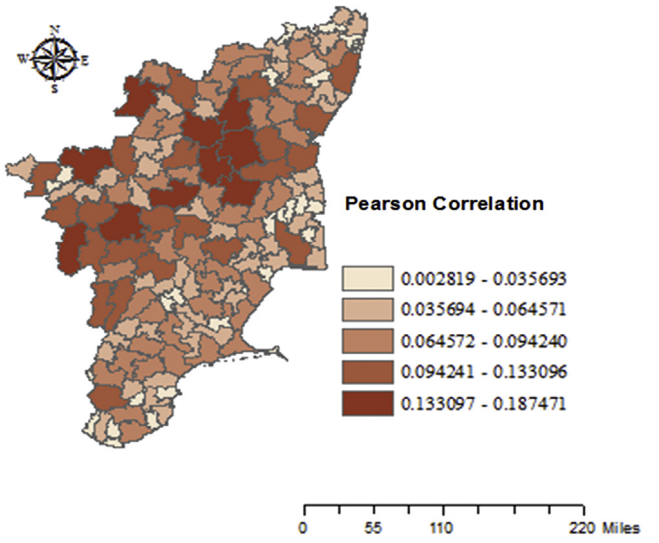

Fig. 14. Geo-spatial visualization of correlation results between maximum temperature and dengue for Tamil Nadu, India (Jan 1998–Dec 2016).

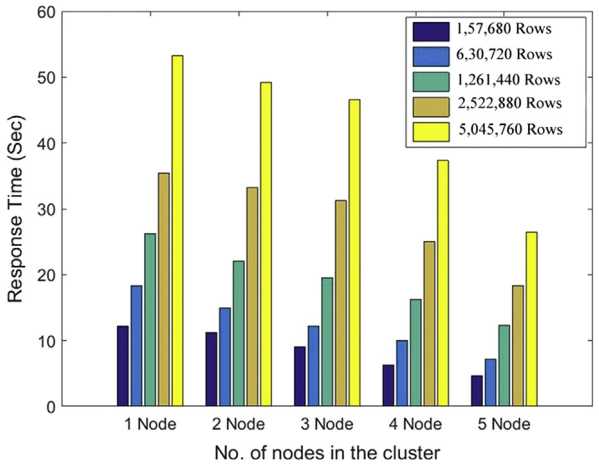

Fig. 16. Performance evaluation parameters for varying query sizes.

following parameters, namely, date, elevation, longitude, latitude, minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity. The NCEP has provided weather data for 36 year period from 1979 to 2016. It contains day wise climate data relating to about 432 years for the entire world. The proposed framework focuses only on climate data for 32 districts of Tamil Nadu where each district contains 1,57,680 rows and 50,45,760 rows in total.

4.2. Batch layer and batch views

The essential role of a batch layer is divided into two major tasks: (1) master dataset management and (2) batch views precomputation. Apache Hive is used for managing the master data in the batch layer whereas Hadoop MapReduce framework is used compute the batch views. A SQL interface is used in Hive for analysis, querying and summarizing the large volume of data in the Hadoop environment. Hive uses the query language named HiveQL. The role of the HiveQL is to convert the SQL-like queries into MapReduce jobs in the Hadoop environment. HiveQL is used in the framework for bulk loading of huge climate data into HDFS.

Batch view pre-computation for the average of various weather parameters would require 50,45,760 rows of weather data. Hence, this would create a high latency and reduce efficiency.

In order to overcome this issue, batch views can be precomputed for a smaller number of records and involve further computation to be done at query time. The MapReduce framework is used for computing the monthly mean of climate parameter for each latitude and longitude.

Fig. 6 represents the Key and value pair for the batch view generation. Fig. 7 represents the batch view precomputation for the average of four days.

4.3. Data integration

As shown in Tables 1–3, the monthly average of climate parameters and monthly dengue incidence are integrated on the

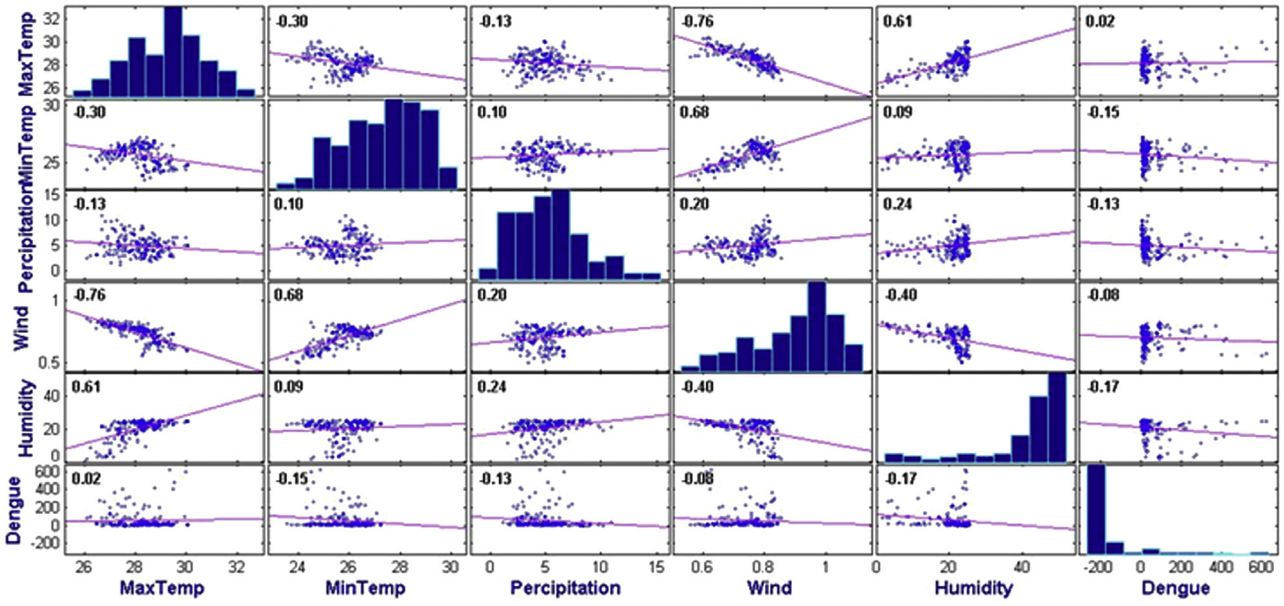

Fig. 15. Pearson correlation results.

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

Algorithm 1: Master data management using Apache Hive

Data: Day wise weather data
Input: Day wise weather data: minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity
Output: Day Wise Weather Table
Step 1: if (Table Name NULL) or (Table Field NULL) then
Step 2: for each column in Table Field Day Wise Weather Table do add field name, field data type
Step 2: for each row in Table Fields Day Wise Weather Table do Table Fields terminated by comma
Step-3: Return Day Wise Weather Table

Algorithm 2: Loading Day wise weather data into Weather Table in HDFS

Input: Day wise weather data: minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity
Output: HiveQL Load Table query
Step 1: if (Table Name NULL) then
Step 2: load data local inpath 'env:home/weatherdata.txt' into table Day Wise Weather Table;
Step 3: Return the Hive Load Table query

Algorithm 3: Batch view generation - MapReduce algorithm for monthly mean of climate parameters

```
class Mapper
        Wid=Weather parameter ID
        wpv=Weather parameter value(day wise)
method Map(integer Wid, double wpv)
        Emit(integer Wid, double wpv)
class Reducer
method Reduce(integer Wid, double [wpv 1 , wpv 2 , . . .])
        sum   0
        count   0
for all double wpv double [wpv 1 , wpv 2 , . . .] do
        sum   sum + wpv
        count   count + 1
        avgwpv sum/count
Emit(integer Wid, float avgwpv)
```

basis of geo-location (latitude and longitude). Integration of the epidemiological and climate datasets is handled by Apache Hive. The vital role of Apache Hive query is to join the uploaded dengue data with climate datasets on the basis of geo-location.

4.4. Pearson correlation coefficient

Pearson correlation coefficient is used for computing the linear relationship between the variables. Table 4 represents the degree of correlation between any two variables x and y, and Pearson correlation coefficient r is defined by,

$$r = \frac{\sum_i (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}} \quad (1)$$

where,
$\overline{x} =$ Mean of x
$\overline{y} =$ Mean of y
$r =$ Pearson correlation coefficient.

In this paper, x represents the dengue incidence and y represents the climate parameter.

4.5. Serving layer

As a batch layer does not store batch views (Pearson correlation between the monthly mean weather parameters and monthly dengue cases), there is a need to store the batch views in a scalable distributed environment. Hence, the serving layer is connected to the batch layer to store the batch views. Apache HBase is used in the proposed framework for storing the batch views. Hive–HBase integration API is used for storing the Apache Hive results into the Apache HBase.

The vital role of the serving layer is to update the batch views in a continuous manner, Due to the high latency of batch view computation, the batch views will always be out of date. As speed layer resolves the queries on the most recent data, the high latency

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

Algorithm 4: The proposed In-Mapper based MapReduce algorithm

```
Data: Weather data collected from multiple weather stations
Input: Weather data climatedata.txt from HDFS
Output: Seasonal Average Weather Parameters to HDFS
Function: Mapper
        method INITIALIZE
                Sum=new ASSOCIATIVE ARRAY
                Count=new ASSOCIATIVE ARRAY
        method MAP<string nwp,double wpv >
                nwp=Weather Parameter Name
                wpv=Weather Parameter Value(day wise)
                Sum nwp=Sum nwp+wpv
                Count nwp=Sum nwp+1
        method CLOSE
                for all term nwp 2 Sum do
                Emit Intermediate(string nwp,pair(Sum nwp,Count nwp) )
Function: Reducer
        method REDUCE(string nwp,pairs [(Sum1;Count1:::)]
                double Total Sum=0.0;
                int Total Count= 0;
        for all pair (Sum,Count) 2 pairs [(Sum1;Count1:::)] do
                Total Sum=Total Sum+Sum;
                Total Count=Total Count+Count
                Avg wpv=Total Sum / Total Count
return(string nwp,double Avg wpv);
```

Algorithm 5: Calculation of Pearson Correlation in HDFS

Input: Monthly mean climate parameters

Output: Pearson correlation between the variables

Step 1: if (Table Name = NULL) then

Step 2: for each (pair of Table Field Monthly Mean Climate Table) do select count(*), corr (field name, field name) from Monthly Mean Climate;

Step 3: Return the Hive correlation query

issue of the batch layer is overcome by speed layer. A few hours is taken by the batch layer to complete a single batch view, whereas the serving layer takes less than an hour to updates the batch view. In general, updates of the most recent batch views are stored in the serving layer database.

4.6. Speed layer and speed views

Batch view computation in big data requires a few hours to complete. Hence, batch view will become stale during the batch processing time. However, the data ingestion block continuously stores the data into the batch layer. The most recent incoming data has to be stored in the speed layer, for resolving the queries on the most recent data.

4.7. Visualizing layer

The Pearson correlation between the various weather parameters such as minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity and a number of dengue incidence is geographically visualized with help of Hive ArcGIS connector. This API is used for transferring of the files from ArcGIS to Apache Hadoop, and Apache Hadoop to ArcGIS. The GUI for the Hive ArcGIS connector is represented in Figs. 8–10.

5. Performance evaluation

Details of dengue incidence have been collected from Directorate of National Vector Borne Disease Control Programme (NVBDCP) and various recent research studies. The NCEP provides weather data for a 36 year period from 1979 to 2016. It contains day wise climate data of about 432 years for the entire world. The following weather parameters such as minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity are collected for the study are Tamil Nadu. Proposed framework focuses only on climate data for 32 districts of Tamil Nadu where each district contains 1,57,680 rows and so there are 50,45,760 rows in total. Batch view precomputation for the monthly mean of various climate parameters would require 50,45,760 rows. Hence, this would create more latency in query processing. In order to overcome this issue, batch views can precompute for a smaller number of records and involve more computation to be done at query time. The MapReduce framework is used to compute the monthly mean of climate parameter for each latitude and longitude. The day wise weather data and monthly average for the year Jan 1979 to Dec 2016 of the study area is represented in Figs. 11 and 12 respectively. The Pearson correlation between the monthly mean weather parameters and monthly dengue cases is calculated with the help of Hive aggregate built-in function corr ( ).

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

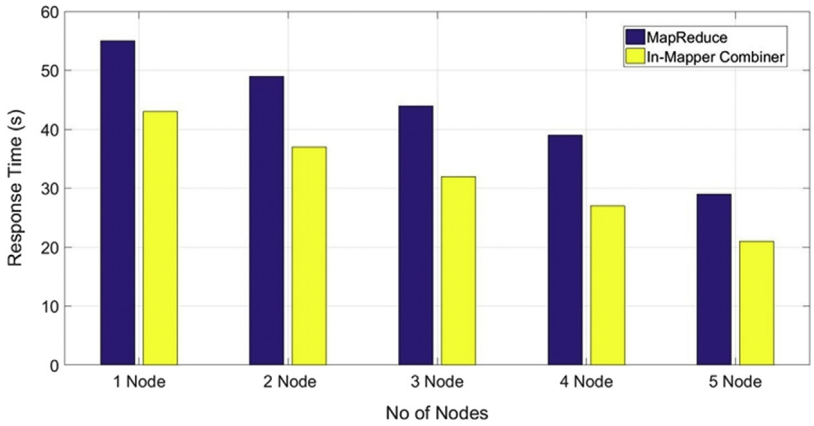

Fig. 17. Comparison between the traditional MapReduce algorithm and In-Mapper based combiner algorithm.

The correlation results are stored in the Apache HBase (serving layer) with the help of Hive–HBase integration. The Apache HBase updates the batch views in a continuous manner. Fig. 13 represents the integration of monthly dengue cases and mean climate data for the year Jan 1998 to Dec 2016. Fig. 15 represents the correlation results between maximum temperature and Dengue for Tamil Nadu, India (Jan 1998–Dec 2016). Similarly, Geospatial visualization of correlation results between maximum temperature and Dengue for Tamil Nadu, India (Jan 1998–Dec 2016) is represented in Fig. 14 with the help of Hive ArcGIS connector. It is observed from the results; the dengue incidence is positively correlated with maximum temperature and negatively correlated with the precipitation, wind and humidity. The large volume of climate data is reduced with the help of Hadoop MapReduce. The results are stored in the serving layer. The serving layer is implemented in a Hadoop distributed file system environment. The framework is used for finding the correlation between the dengue incidence and various weather parameters such as minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity. The results generated from the batch layer are stored across the hadoop cluster. The performance of the framework is evaluated with the help of various queries in a cluster of nodes. The performance evaluation parameters for varying query sizes is represented in Fig. 16.

Fig. 17 represents the comparison between the traditional MapReduce algorithm and In-Mapper based combiner algorithm. The experimental results prove the effectiveness of the In-Mapper based combiner algorithm. As shown in Fig. 5. the response time for the In-Mapper based combiner algorithm is less when compared with the existing MapReduce algorithm. The In-Mapper based combiner algorithm is effectively used in the proposed system to compute the seasonal averages of various climate parameters such as minimum temperature, maximum temperature, wind, precipitation, solar and relative humidity.

6. Conclusion

The objective of this chapter is to propose a big data processing framework to store and process the big climate data. The proposed framework is capable of monitoring the correlation between the climate parameters and dengue the incidence in a continuous manner. The framework is demonstrated in a HDFS environment with five layers, namely, data ingestion layer, batch layer, speed layer, serving layer and visualization layer. Hadoop MapReduce and Hive is used for implementation of the batch layer. Apache HBase and Apache Hive streaming are used for implementation of serving layer and speed layer respectively. ArcGIS10.2 software is used for implementation of the visualization layer with the help of Hive ArcGIS API. The proposed framework is implemented in a Hadoop cluster environment. The number of cluster nodes in the Hadoop environment varies on the basis of the measurement and optimization technique. The performance evaluation results proved the effectiveness of the proposed framework. The experimental results show the efficiency in storing the big climate data and predicting the correlation between the number of dengue cases and climate parameters.

This research study would be useful for developing a continuous monitoring system for disease surveillance. The goal of this disease surveillance system is to collect the steaming data from various climate sources and identify the change changes in the seasonal climate. The speed layer could be used to store and compute the real time views. Moreover, Internet of Things can be integrated with the disease surveillance system that combining, location-based tracking, smart sensors, wireless technology and cloud computing which can help optimizing the work of employees, patients, equipment and medical supplies. For example, wearable sensor devices are used for suggesting physiological exercises and food habits by a two or three day period of continuous physiological monitoring of patients. In this period, wearable sensors would continuously observe and store the patients health data into a data store. This would help doctors in efficient diagnosis of the patient's health condition. Reliable results are feasible through use of laboratory tests, and also patient's health data collected from wearable body sensors.

Hence, sensor data could be used for initiating appropriate action for ensuring improvement to the patient's health and early diagnosis. Traditional data storage techniques and platforms are not adequate to provide storage for above mentioned emerging sensor application domains where the volume, velocity and variety of the data grow by leaps and bounds. Solution of this problem requires the development of an efficient system for storage and processing of voluminous big data. A scalable IoT based architecture to be proposed to store the real time sensor (physiological) data and identify drastic changes using scalable big data based change detection algorithms.

References

[1] G.-H. Kim, S. Trimi, J.-H. Chung, Big-data applications in the government sector, Commun. ACM 57 (3) (2014) 78–85.
[2] D. Lopez, G. Manogaran, Big data architecture for climate change and disease dynamics, in: Geetam S. Tomar, et al. (Eds.), The Human Element of Big Data: Issues, Analytics, and Performance, CRC Press, 2016.
[3] G. Manogaran, D. Lopez, Disease surveillance system for big climate data processing and dengue transmission, Int. J. Ambient Comput. Intelli 8 (2) (2017) 1–25.

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

[4] G. Manogaran, D. Lopez, A Gaussian process based big data processing framework in cluster computing environment, Cluster Comput. (2017) 1–16.
[5] G. Manogaran, D. Lopez, Spatial cumulative sum algorithm with big data analytics for climate change detection, Comput. Electr. Eng. (2017).
[6] S. Nativi, P. Mazzetti, M. Santoro, F. Papeschi, M. Craglia, O. Ochiai, Big data challenges in building the global earth observation system of systems, Environ. Modell. Softw. 68 (2015) 1–26.
[7] J.-G. Lee, M. Kang, Geospatial big data: challenges and opportunities, Big Data Res. 2 (2) (2015) 74–81.
[8] J.H. Faghmous, V. Kumar, A big data guide to understanding climate change: The case for theory-guided data science, Big Data 2 (3) (2014) 155–163.
[9] R. Varatharajan, G. Manogaran, M.K. Priyan, V.E. Balaş, C. Barna, Visual analysis of geospatial habitat suitability model based on inverse distance weighting with paired comparison analysis, Multimedia Tools Appl. (2017) 1–21.
[10] R. Varatharajan, K. Vasanth, M. Gunasekaran, M. Priyan, X.Z. Gao, An adaptive decision based kriging interpolation algorithm for the removal of high density salt and pepper noise in images, Comput. Electr. Eng. (2017).
[11] D. Lopez, G. Manogaran, J. Jagan, Modelling the H1N1 influenza using mathematical and neural network approaches, Biomed. Res. 28 (8) (2017) 1–5.
[12] B.R. Pickard, J. Baynes, M. Mehaffey, A.C. Neale, Translating big data into big climate ideas, Solutions 6 (1) (2015) 64–73.
[13] J.L. Schnase, D.Q. Duffy, G.S. Tamkin, D. Nadeau, J.H. Thompson, C.M. Grieg, M.A. McInerney, W.P. Webster, Merra analytic services: meeting the big data challenges of climate science through cloud-enabled climate analytics-as-aservice, Comput. Environ. Urban Syst. 61 (2017) 198–211.
[14] S. Fiore, M. Mancini, D. Elia, P. Nassisi, F.V. Brasileiro, I. Blanquer, Big data analytics for climate change and biodiversity in the eubrazilcc federated cloud infrastructure, in: Proceedings of the 12th ACM International Conference on Computing Frontiers, ACM, 2015, p. 52.
[15] D. Lopez, M. Gunasekaran, B.S. Murugan, H. Kaur, K.M. Abbas, Spatial bigdata analytics of influenza epidemic in Vellore, India, in: Proc. 2014 IEEE International Conference on Big Data, IEEE, 2014 October, pp. 19–24.
[16] R. Varatharajan, G. Manogaran, M.K. Priyan, A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing, Multimedia Tools Appl. (2017) 1–21.
[17] D. Lopez, M. Gunasekaran, Assessment of vaccination strategies using fuzzy multicriteria decision making, in: Proc. Proceedings of the Fifth International Conference on Fuzzy and NeuroComputing, FANCCO-2015, Springer International, 2015, pp. 195–208.
[18] D. Lopez, G. Sekaran, Climate change and disease dynamics - A big data perspective, Int. J. Infect. Dis. 45 (2016) 23–24.
[19] D.W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, Big data in health care: using analytics to identify and manage high-risk and high-cost patients, Health Affairs 33 (7) (2014) 1123–1131.
[20] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health Inf. Sci. Syst. 2 (1) (2014) 3.
[21] K. Jee, G.-H. Kim, Potentiality of big data in the medical sector: focus on how to reshape the healthcare system, Healthcare Inf. Res. 19 (2) (2013) 79–85.
[22] G. Manogaran, V. Vijayakumar, R. Varatharajan, P.M. Kumar, R. Sundarasekar, C.H. Hsu, Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering, Wirel. Pers. Commun. (2017) 1–18.
[23] Q. Yao, Y. Tian, P.-F. Li, L.-L. Tian, Y.-M. Qian, J.-S. Li, Design and development of a medical big data processing system based on hadoop, J. Med. Syst. 39 (3) (2015) 23.
[24] J.L. Schnase, D.Q. Duffy, G.S. Tamkin, D. Nadeau, J.H. Thompson, C.M. Grieg, M.A. McInerney, W.P. Webster, Merra analytic services: meeting the big data challenges of climate science through cloud-enabled climate analytics-as-aservice, Comput. Environ. Urban Syst. 61 (2017) 198–211.
[25] Z. Li, Q. Huang, G.J. Carbone, F. Hu, A high performance query analytical framework for supporting data-intensive climate studies, Comput. Environ. Urban Syst. 62 (2017) 210–221.
[26] Y. Zhu, Global climate change studying based on big data analysis of antarctica, in: Proceedings of the Fourth International Forum on Decision Sciences, Springer, 2017, pp. 39–45.
[27] S. Hajat, C. Whitmore, C. Sarran, A. Haines, B. Golding, H. Gordon-Brown, A. Kessel, L.E. Fleming, Development of a browser application to foster research on linking climate and health datasets: challenges and opportunities, Sci. Total Environ. 575 (2017) 79–86.
[28] H. Chasparis, A. Eldawy, Experimental evaluation of selectivity estimation on big spatial data, in: Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data, ACM, 2017, p. 8.
[29] S. Ackermann, V. Jovanovic, T. Rompf, M. Odersky, Jet: An embedded dsl for high performance big data processing, in: International Workshop on End-to End Management of Big Data, BigData 2012, Number EPFL-CONF-181673, 2012.
[30] S. Gao, L. Li, W. Li, K. Janowicz, Y. Zhang, Constructing gazetteers from volunteered big geo-data based on hadoop, Comput. Environ. Urban Syst. 61 (2017) 172–186.
[31] Z. Zheng, P. Wang, J. Liu, S. Sun, Real-time big data processing framework: challenges and solutions, Appl. Math. Inf. Sci. 9 (6) (2015) 3169.
[32] Xiong Li, Jianwei Niu, Md Zakirul Alam Bhuiyan, Fan Wu, Marimuthu Karuppiah, Saru Kumari, A robust ECC based provable secure authentication protocol with privacy preserving for industrial internet of things, IEEE Trans. Ind. Inf. (2017). Online.
[33] N. Marz, J. Warren, Big Data: Principles and Best Practices of Scalable Realtime Data Systems, Manning Publications Co, 2015.
[34] M.A. Martínez-Prieto, C.E. Cuesta, M. Arias, J.D. Fernández, The solid architecture for real-time management of big semantic data, Future Gener. Comput. Syst. 47 (2015) 62–79.
[35] M. Krämer, I. Senner, A modular software architecture for processing of big geospatial data in the cloud, Comput. Graph. 49 (2015) 69–81.
[36] I. Gorton, J. Klein, Distribution, data, deployment: Software architecture convergence in big data systems, IEEE Softw. 32 (3) (2015) 78–85.
[37] R.R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, S. Shekhar, Spatiotemporal data mining in the era of big spatial data: algorithms and applications, in: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, ACM, 2012, pp. 1–10.
[38] Xiong Li, Jianwei Niu, Saru Kumari, Fan Wu, Arun Kumar Sangaiah, Kim-Kwang Raymond Choo, A three-factor anonymous authentication scheme for wireless sensor networks in internet of things environments, J. Netw. Comput. Appl. (2017). Online http://dx.doi.org/10.1016/j.jnca.2017.07.001.
[39] C. Thota, R. Sundarasekar, G. Manogaran, R. Varatharajan, M.K. Priyan, Centralized fog computing security platform for iot and cloud in healthcare system, in: Exploring the Convergence of Big Data and the Internet of Things, IGI Global, 2018, pp. 141–154.
[40] S. Moosavi, T. Gia, E. Nigussie, A. Rahmani, S. Virtanen, H. Tenhunen, J. Isoaho, End-to-end security scheme for mobility enabled healthcare internet of things, Future Gener. Comput. Syst. 64 (2016) 108–124.
[41] Xiong Li, Jianwei Niu, Saru Kumari, Fan Wu, Kim-Kwang Raymond Choo, A robust biometrics based three-factor authentication scheme for global mobility networks in smart city, Future Gener. Comput. Syst. (2017). Online http://dx.doi.org/10.1016/j.future.2017.04.012.
[42] A. Whitmore, A. Agarwal, The internet of things-a survey of topics and trends, Inf. Syst. Front. 17 (2) (2015) 261–274. http://dx.doi.org/10.1007/s10796-014-9489-2.
[43] M. Masdari, S. Ahmadzadeh, Comprehensive analysis of the authentication methods in wireless body area networks, Secur. Commun. Netw. 9 (17) (2016) 4777–4803.
[44] R. Chakravorty, A programmable service architecture for mobile medical care, in: Proc. Pervasive Computing and Communications Workshops, 2006. PerCom Workshops 2006. Fourth Annual IEEE International Conference on, IEEE, 2006, March, p. 5.
[45] S. Al-Janabi, I. Al-Shourbaji, M. Shojafar, S. Shamshirband, Survey of main challenges (security and privacy) in wireless body area networks for healthcare applications, Egypt. Inf. J. (2016).
[46] M.K. Priyan, G.U. Devi, Energy efficient node selection algorithm based on node performance index and random waypoint mobility model in internet of vehicles, Cluster Comput. (2017) 1–15.
[47] Khan, Medical applications of wireless body area networks, Int. J. Digit. Content Technol. Appl. 3 (3) (2009).
[48] M. Marjani, F. Nasaruddin, A. Gani, A. Karim, I.A.T. Hashem, A. Siddiqa, I. Yaqoob, Big iot data analytics: architecture, opportunities, and open research challenges, IEEE Access 5 (2017) 5247–5261.
[49] Xiong Li, Fan Wu, Muhammad Khurram Khan, Lili Xu, Jian Shen, Minho Jo, A secure chaotic map-based remote authentication scheme for telecare medicine information systems, Future Gener. Comput. Syst. (2017). Online http://dx.doi.org/10.1016/j.future.2017.08.029.
[50] Xiong Li, Maged Hamada Ibrahim, Saru Kumari, Arun Kumar Sangaiah, Vidushi Gupta, Kim-Kwang Raymond Choo, Anonymous mutual authentication and key agreement scheme for wearable sensors in wireless body area networks, Comput. Netw. (2017). Online http://dx.doi.org/10.1016/j.comnet.2017.03.013.
[51] P. Chandarana, M. Vijayalakshmi, Big data analytics frameworks, in: Circuits, Systems, Communication and Information Technology Applications, CSCITA, 2014 International Conference on, IEEE, 2014, pp. 430–434.
[52] R. Varatharajan, G. Manogaran, M.K. Priyan, R. Sundarasekar, Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm, Cluster Comput. (2017) 1–10.
[53] C. Thota, R. Sundarasekar, G. Manogaran, R. Varatharajan, M.K. Priyan, Centralized fog computing security platform for iot and cloud in healthcare system, in: Exploring the Convergence of Big Data and the Internet of Things, IGI Global, 2018, pp. 141–154.
[54] G. Manogaran, R. Varatharajan, D. Lopez, P.M. Kumar, R. Sundarasekar, C. Thota, A new architecture of internet of things and big data ecosystem for secured smart healthcare monitoring and alerting, Future Gener. Comput. Syst. (2017).
[55] G. Manogaran, D. Lopez, Health data analytics using scalable logistic regression with stochastic gradient descent, Int. J. Adv. Intell. Paradigms 9 (2016) 1–15.

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.

[56] S.E. Hampton, C.A. Strasser, J.J. Tewksbury, W.K. Gram, A.E. Budden, A.L. Batcheller, C.S. Duke, J.H. Porter, Big data and the future of ecology, Front. Ecol. Environ. 11 (3) (2013) 156–162.
[57] B.C. Pijanowski, A. Tayyebi, J. Doucette, B.K. Pekin, D. Braun, J. Plourde, A big data urban growth simulation at a national scale: configuring the gis and neural network based land transformation model to run in a high performance computing (hpc) environment, Environ. Modell. Softw. 51 (2014) 250–268.
[58] G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K.M. Abbas, R. Sundarsekar, Big data knowledge system in healthcare, in: Internet of Things and Big Data Technologies for Next Generation Healthcare, Springer International Publishing, 2017, pp. 133–157.
[59] S.I. Hay, D.B. George, C.L. Moyes, J.S. Brownstein, Big data opportunities for global infectious disease surveillance, PLoS Med. 10 (4) (2013) e1001413.

Gunasekaran Manogaran is currently working in University of California, Davis, USA. He has received his Ph.D. from the Vellore Institute of Technology University, India. He received his Bachelor of Engineering and Master of Technology from Anna University and Vellore Institute of Technology University respectively. He has worked as a Research Assistant for a project on spatial data mining funded by Indian Council of Medical Research, Government of India. His current research interests include data mining, big data analytics and soft computing. He is the author/co-author of papers in conferences, book chapters and journals. He got an award for young investigator from India and Southeast Asia by Bill and Melinda Gates Foundation. He is a member of International Society for Infectious Diseases and Machine Intelligence Research labs.

Daphne Lopez is a Professor in the School of Information Technology and Engineering, Vellore Institute of Technology University. Her research spans the fields of grid and cloud computing, spatial and temporal data mining and big data. She has a vast experience in teaching and industry. She is the author/co-author of papers in conferences, book chapters and journals. She serves as a reviewer in journals and conference proceedings. Prior to this, she has worked in the software industry as a consultant in data warehouse and business intelligence. She is a member of International Society for Infectious Diseases.

Naveen Chilamkurti is currently working as a Senior Lecturer at Department of Computer Science and Computer Engineering, La Trobe University, Australia. He received his Ph.D. from La Trobe University. He is also the Inaugural Editor-in-Chief for International Journal of Wireless Networks and Broadband Technologies launched in July 2011. He has published about 125 journal and conference papers. His current research areas include intelligent transport systems (ITS), wireless multimedia, wireless sensor networks, vehicle to infrastructure, vehicle to vehicle communications, health informatics, mobile communications, WiMAX, mobile security, mobile handover, and RFID. He currently serves on editorial boards of several international journals. He is a senior member of IEEE. He is also an Associate Editor for Wiley IJCS, SCN, Interscience JETWI, and IJIPT.

Please cite this article in press as: G. Manogaran, et al., In-Mapper combiner based MapReduce algorithm for processing of big climate data, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.02.048.