Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2020, Article ID 8081578, 12 pages
https://doi.org/10.1155/2020/8081578

Research Article

Semantic Smart World Framework

K. ElDahshan,[1] **E. K. Elsayed**,[2] **and H. Mancy**[2]

[1]*Department of Mathematics, Faculty of Science, Al-Azhar University, Cairo, Egypt*
[2]*Department of Mathematics, Faculty of Science (Girls), Al-Azhar University, Cairo, Egypt*

Correspondence should be addressed to H. Mancy; dr.hendfathi@azhar.edu.eg

Received 30 July 2019; Revised 5 October 2019; Accepted 16 October 2019; Published 20 January 2020

Academic Editor: Miin-Shen Yang

This paper presents a general Semantic Smart World framework (SSWF), to cover the Migratory birds' paths. This framework combines semantic and big data technologies to support meaning for big data. In order to build the proposed smart world framework, technologies such as cloud computing, semantic technology, big data, data visualization, and the Internet of Things are hybrid. We demonstrate the proposed framework through a case study of automatic prediction of air quality index and different weather phenomena in the different locations in the world. We discover the association between air pollution and increasing weather conditions. The experimental results indicate that the framework performance is suitable for heterogeneous big data.

1. Introduction

Migratory birds can move from one place to another without borders between countries; so, we need to use the concept of "Smart World". Big data can serve the world in "Smart World" challenges. Most of these challenges are related to data management. The most cited problems are privacy issues and dealing with the heterogeneity of world data. An important issue is how to build a generic smart world framework to support all dimensions of any city regardless of its size and characteristics.

The rapid evolution of Information and Communication Technologies (ICT) and the Internet of Things (IoT) has impacted cities in the physical infrastructure, buildings, transportation systems, governance, environmental monitoring, healthcare, etc. The integration of devices, platforms, and applications using ICT is of great significance to smart cities [1].

The expression "Smart City" has many different definitions. Some authors define a Smart City as the integration of social, physical, and IT infrastructure to improve the quality of city services. Other authors focus on a set of Information and Communication Technology (ICT) tools to integrate the Smart City environment [2].

City computing is a process of acquisition, integration, and analysis of a huge amount of heterogeneous data generated by diverse sources in city spaces, such as sensors, devices, vehicles, buildings, and humans. These sources are the aim of addressing the major issues cities face (e.g., air pollution, increased energy consumption, and traffic congestion) [3]. There are three main challenges in city computing: city sensing and data acquisition, computing with heterogeneous data, and hybrid systems combining the physical and virtual worlds.

Recently, many frameworks have been proposed in different dimensions of smart cities including transportation [4], environment [5], energy [6], social [7], economy [8], and public safety and security [9]. Most of these frameworks did not include semantic interpretation of the results and focused on specific domains. In general, the data generated from smart cities are usually not easy to understand by humans because it has the challenges related to big data. The concept of Big Data is clarified by considering five Vs [10]:

(1) Volume: refers to the size of data that has been generated by all different sources.
(2) Velocity: refers to the speed of data changes.
(3) Variety: refers to the different types of data being generated.
(4) Veracity: refers to the quality of the data.
(5) Value: refers to the value of the data.

Therefore, there are necessary needs to build a general framework to overcome the challenges posed by data of the city.

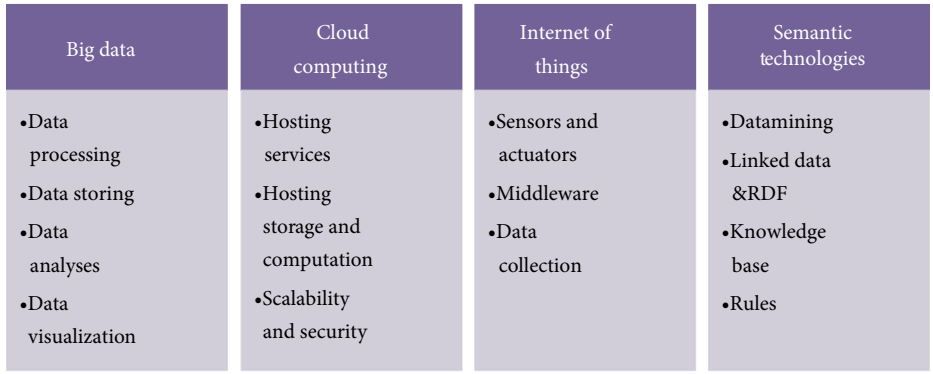

Figure 1: Smart City platforms technologies.

Table 1: A comparative study among Smart City frameworks and its technologies.

Smart City framework	Type	Technologies					
		IoT	Big data	Semantic	Cloud computing	API\services	Security
SCDAP	F	√	√				
CityPulse	F	√		√			
Zhang et al.	F	√		√		√	
Spitfire	F	√		√			
iCore	F	√		√			
CITIESData	F		√		√		
Krishna	F	√	√				
Simon	F	√	√		√		

To achieve the research objectives, this paper is structured as follows: Section 2 reviews background and previous related works. Section 3 illuminates the proposed framework architecture. Section 4 describes the implementation of our proposed framework SSWF. Section 5 provides a case study of the SSWF to analyze air pollution and weather on migratory birds' path. Section 6 explains the result of applying the proposed framework. Section 7 discusses the significant contribution and limitations of this research and concludes the paper.

2. Related Work

Different cities have already built IoT infrastructures and various sensor devices to collect the data needed. A huge number of research projects concentrates on the collection and economy of IoT data generated from smart cities.

Many Smart City frameworks can classify into three different classes [11]:

(1) Models: abstract frameworks for Smart City.
(2) Specific purpose model: framework and applications related to one domain of the Smart City.
(3) Multidomain models: framework and applications that describe the Smart City as a complex system and consider more than one domain.

Smart City frameworks have a major focus on existing Smart City platforms. The existing works are mainly in four key areas: (1) data acquisition, (2) semantic interoperability, (3) data analysis, and (4) Smart City application development support [12].

We divided the framework in the Smart City into four categories, according to technologies used. Almost all of the frameworks use at least one or more of the following technologies (Big Data, Cloud Computing, Internet of Things, and Semantic Technology). Figure 1 presents technologies and their functions.

Table 1 presents a comparative study among Smart City frameworks and the technologies used. Table 1 explores also security and API services. SCDAP "Smart City Data Analytics Panel" is a big data analytics framework for Smart City applications, the main feature of this architecture is limited to Apache Hadoop suite as an underlying data storage and management layer [13]. The "CityPulse" framework supports Smart City service creation by means of a distributed system for semantic discovery, data analytics, and interpretation of large-scale near the real-time Internet of Things data and social media data streams [12]. Zhang et al. [14] presented a semantic framework that integrates the IoT with machine learning for smart cities. This framework retrieves and models urban data for certain kinds of IoT applications based on semantic and machine-learning technologies. It is used to detect pollution from vehicles and to detect traffic patterns. Spitfire and iCore are frameworks that use semantic technologies for IoT data collection [15]. CITIESData is a Smart City data management framework that includes data collection, cleansing, and publishing [16]. It divides Smart City data

Figure 2: The concept of SSWF.

insensitive, quasi-sensitive, and open/public levels. Then it suggests different strategies to process and publish the data within these categories. Mohbey [17] presented a Smart City framework using different technologies of big data and the Internet of Things. It focuses on problems related to real-time decisions for a smart city. Bibri [18] proposed a framework for a Smart City based on big data and sensor data.

There are points still to be covered. On the one hand, Hybrid technologies such as Big Data, Semantic technology, cloud computing, the Internet of Things, and Data Vitalization are not integrated to support a more efficient smart city framework. On the other hand, big data frameworks did not support meaning to add value to the data. Semantic frameworks are so slow for data retrieving and processing. Security issues still prevail in previous frameworks.

3. The Proposed Semantic Smart World Framework (SSWF) Architecture

The world contains the set of Data Centres for Smart Cities which specializes in collecting and measuring big data for natural phenomena like bird migration, environmental pollution, and Climate Change. Many problems are there in these data centres; they are not connected together; they serve Smart Cities in the world; and, there is no open access to the data.

SSWF is a general semantic big data framework that combines semantic web and big data technologies to connect, predict, and discover the knowledge of world big data. That is without boundaries between cities.

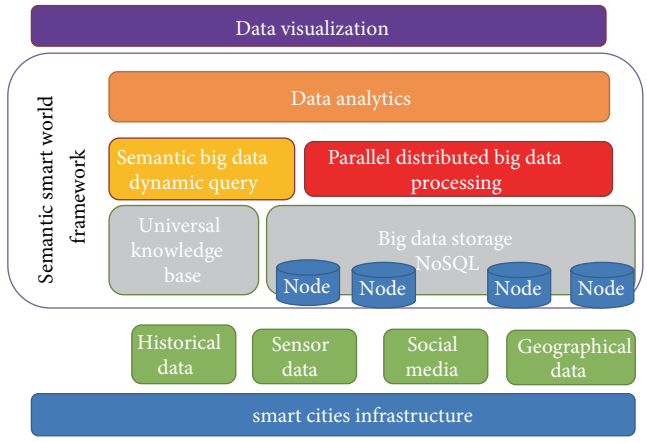

Figure 3: SSWF architecture.

Figure 2 shows the concept of semantic smart world framework.

Big data technologies are required for most data-related activities, such as storing, processing, analyzing, and sharing, while Semantic technologies are required for meaning-related activities, such as event detection, reasoning, and decision support. Thus, this research is aimed to build and develop a general framework for smart cities that utilizes a combination of data-related and meaning-related activities.

As shown in Figure 3, The Smart Cities' infrastructure generates the heterogeneous big data. The architecture of SSWF consists of the following phases over Smart Cities

```
SELECT ?x1,?x2,.....?xi
 WHERE {

        X ?x1 . Y ?x2 . Z?x3 .

        FILTER( condition)
}
MODIFIERS
```

Figure 4: General dynamic query.

infrastructure: (1) Big data storage; (2) universal knowledge base; (3) parallel distributed big data processing; (4) Semantic big data dynamic query (5); Data analytics; and (6) Data visualization.

In this section; we describe each component in SSWF architecture. The Smart Cities' infrastructure generates heterogeneous big data. The main challenge is the ability to collect and push timely data of city events from a huge number of heterogeneous sources such as sensors, servers, devices, vehicles, buildings, and human activities, and deal with both historical and real-time big data.

3.1. Big Data Storage. This phase is responsible for storing the data collected from the Smart cities. It used big data storage systems like Hadoop Distributed file system (HDFS) and NoSQL Database. Moreover, this component should be capable of performing useful preprocessing tasks, such as data filtering, normalization, and transformation.

3.2. Universal Knowledge Base. In this phase, we build a universal semantic data model as Ontology-based. This model should automatically classify the data, associate relationships, and find new relationships. This is done by using the OWL "Web Ontology Language" and Ontology re-engineering method as Merging.

3.3. Parallel and Distributed Big Data Processing. This phase is responsible for the processing of smart city data in distributed cluster nodes. There are two types of data processing: Stream processing, to perform real-time data flow; and Batch processing, to perform large historical data-sets. We should choose suitable distributed big data processing frameworks.

3.4. Semantic Big Data Dynamic Query. This phase integrates semantic dynamic query with big data distributed processing. We connect NoSQL with the universal knowledge base. Sematic dynamic queries can run directly on data stored in HDFS/NoSQL without requiring any data movement or transformation. There are two main steps for run query: (1) The RDF Loader converts an RDF dataset into the data layout using MapReduce. (2) The Query Compiler rewrites a given Sematic dynamic into the SQL on big data ecosystem based on the algebraic representation of SPARQL expressions.

General Dynamic SPARQL query as shown in Figure 4. The query consists of three parts: the SELECT clause identifies the variables $(x_1, x_2, \ldots, x_i)$ to appear in the query results, The WHERE clause provides the basic graph pattern (X, Y, Z) to match against the data graph, and filter which contains association rule or condition. The query can include modifiers like (Group BY, HAVING, ORDER BY, LIMIT, OFFSET, and VALUES).

We configure general dynamic SPARQL query in Data visualization phase.

3.5. Data Analytics. The processed data are further analyzed in this phase to utilize events and help decision-makers to take the correct decision. This phase supports semantic filtering, semantic monitoring, event detection, and knowledge discovery. Semantic filtering is s filtering based on expressions in the form of a conjunction of description logics atoms enriched with OWL data types and SWRL (Semantic Web Rule Language) built-ins. We build filter expression in the universal semantic data model. Then the filter expression is translated into a SPARQL query. Semantic monitoring allows event types taxonomy and event parameters to define in Ontology. We define them in the universal semantic data model. Event detection translates into finding a set of event types for which a given event occurrence belongs to their domains.

Knowledge discovery (or data mining techniques) must be adapted to be suitable for Big Data analysis. In this phase, we handled the Gamma Association Coefficient to be suitable to use.

The *gamma association coefficient* (also called the *gamma statistic)* shows us how closely sets of items in a series of data or transactions "match". Gamma can be calculated for ordinal (ordered) variables that are continuous variables (like temperature or humidity) or discrete variables (like "hot" or "cold"). The gamma estimator is based on the number of observations that are concordant and discordant. It ignores tied pairs (i.e., pairs of observations having the same X values or the same Y values). The Gamma coefficient ranges between −1 and 1. Value 1 means perfect positive association, while value −1 means perfect inverse association. If there is no association between the variables, the value will be zero [19].

We assume that cross tabulation or 2×2 cross table formula is shown as Equation (1).

$$\begin{array}{c|cc|c} & X^- & X^+ & \text{Total} \\ Y^- & a & b & e \\ Y^+ & c & d & f \\ \hline & g & h & n \end{array} \quad (1)$$

where a is the frequency of variable X^- against Y^-, b is the frequency of variable X^+ against Y^-, c is the frequency of variable X^- against Y^+, d is the frequency of variable X^+ against Y^+.

Equation (2) shows Gamma association coefficient γ

$$\gamma = \frac{(\mathrm{ad} - \mathrm{bc})}{(\mathrm{ad} + \mathrm{bc})}. \quad (2)$$

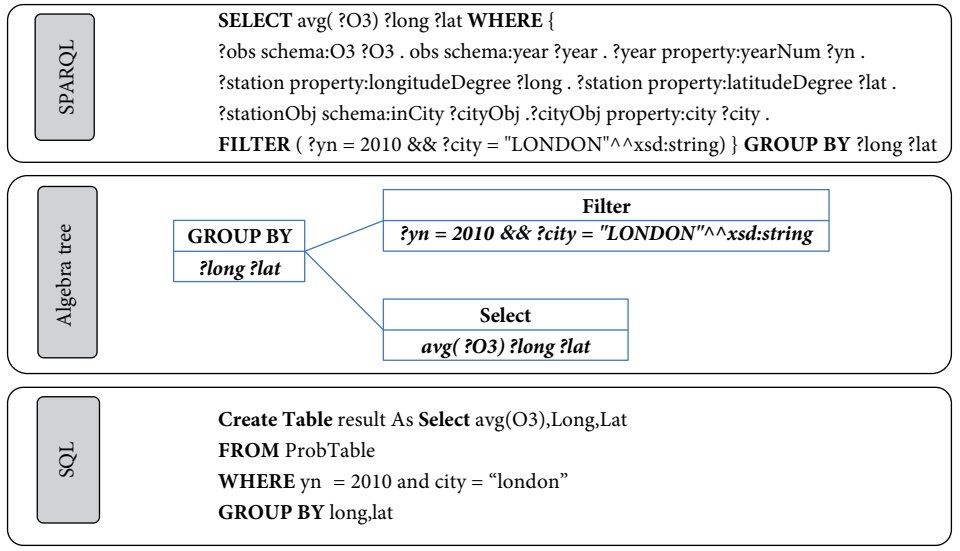

Figure 5: Dynamic query flow.

Table 2: The software packages used in the framework and its function.

Software	Function
Big queue [22]	Data transformation
HDFS [23], HBASE [24]	Data storage
Spark [25] and Hadoop [23]	Data processing
Spark [25]	Stream processing
Hadoop YARN [23]	Clustering management
REST API's	Data access
Spark MLib [25]	Machine learning
SPARQL [26]	Semantics logic
OWL [27] and RDF [28] Protégé	Knowledgebase
Sempala Alexander [20]	Interactive SPARQL query processing on Hadoop
Java [29]	Dashboard

Notice that γ compares the product of diagonal cells (ad) to a product of the off-diagonal cells (bc). The denominator is an adjustment that ensures that γ is always between +1 and −1.

We Generalized 2×2 tables for any category attributes in datasets in Equation (3):

Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ then the cross table will be

$$\begin{array}{c|ccc|c} & X_1 & \cdots\cdots & X_n & \\ Y_1 & a_{11} & & a_{1n} & \sum_{k=1}^{n} \alpha_{1n} \\ \cdot & \cdot & & \cdot & \\ \cdot & \cdot & & \cdot & \\ Y_m & a_{m1} & & a_{mn} & \sum \alpha_{mn} \\ & \sum_{r=1}^{m} \alpha_{r1} & & \sum_{r=1}^{m} \alpha\, r_n & T \end{array} \quad (3)$$

Then for each two categories (x_i, y_j), we have 2×2 cross table in Equation (4).

$$\begin{array}{c|ccc|} & X_i^{+} & \cdots\cdots & X_i^{-} & \\ Y_j^{+} & a_{ij} & & \sum_{k=1}^{n} \alpha_{ik} & t_{1k} \\ & \cdot & & \cdot & \\ & \cdot & & & \\ Y_j^{-} & \sum_{r=1}^{m} \alpha_{rj} & & \sum_{r,k=1}^{n,m} \alpha_{rk} & t_{mk} \\ & t_{j1} & & t_{jn} & T \end{array} \quad (4)$$

We can simplify the variables in 2×2 cross table to be

$a = a_{ij}$ it is number of records which satisfy x_i & y_j classes conditions,

$b = \sum_{k=1}^{n} a_{ik} = t_{1k} - a_{ij} = e - \boldsymbol{a}$ where t_{1k} total number of records which satisfy $y_j\, k \neq j$,

$c = \sum_{r=1}^{m} a_{rj} = t_{j1} - a_{ij} = g - \boldsymbol{a}$ where t_{j1} total number of records which satisfy $x_i\, r \neq i$.

$d = \sum_{r,k=1}^{n,m} a_{rk} = t_{mk} - \sum_{r=1}^{m} a_{rj} = t_{mk} - \left(t_{j1} - a_{ij}\right) = \left(T - t_{1k}\right) - \left(t_{j1} - a_{ij}\right) r, k \neq i, j.$

Then Equation (5) shows the general 2×2 cross table will

$$\begin{array}{c|ccc|} & X & \text{Not } X & \text{Total} & \\ Y & a & e-a & e & \\ \text{Not } Y & g-a & (T-g)-(e-a) & T-e & \\ & g & T-g & T & \end{array} \quad (5)$$

Then Equation (6) shows the simplest form of γ to be suitable for big data

$$\gamma = \frac{[a((T-g)-(e-a))-(e-a)(g-a)]}{[a((T-g)-(e-a))+(e-a)(g-a)]}. \quad (6)$$

The simplification in this way required just count (a, e, g, T) and this calculation must also count in any association rule discovery to calculate support and confidence.

3.6. Data Visualization. The previous component produces output as a series of values. To represent these values, it will be necessary to use visualization techniques. In this type of

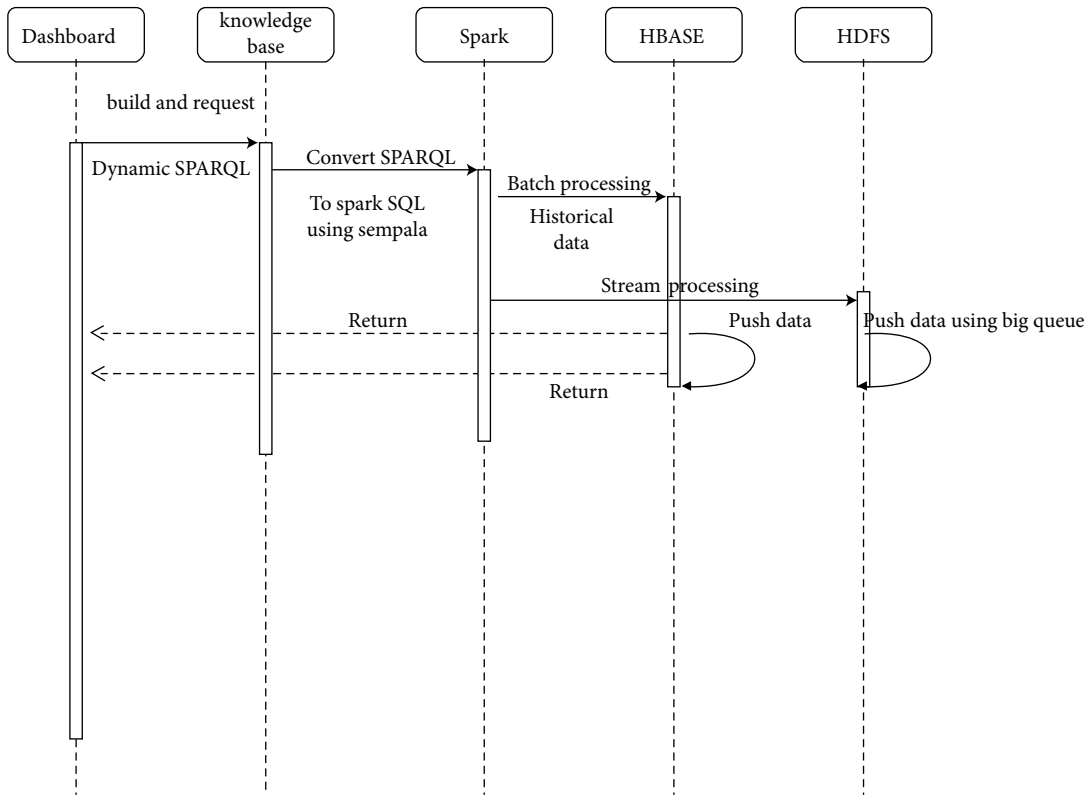

Figure 6: Data sequence diagram.

Table 3: Hbase table structure and design.

Cf: goe		CF: date time						Cf: airquality			
Longitude	Latitude	Day	Month	Year	Hour	CO	O_3	NO_2	SO_2	PM10	PM2.5
CF: weather											
Temperature		Humidity		Rainfall		Precipitation		Wind		Solar-irradiance	

Table 4: The air quality index.

Band	Index	Nitrogen dioxide 1-hour mean (μg m^{-3})		Ozone 8-hourly mean (μg m^{-3})		PM10 particles 24-hour mean (μg m^{-3})		PM2.5 particles 24-hour mean (μg m^{-3})		Sulphur dioxide 15-minute mean (μg m^{-3})	
		Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
	1	0	67	0	33	0	16	0	11	0	88
Low	2	68	134	34	66	17	33	12	23	89	177
	3	135	200	67	100	34	50	24	35	178	266
	4	201	267	101	120	51	58	36	41	267	354
Moderate	5	268	334	121	140	59	66	42	47	355	443
	6	335	400	141	160	67	75	48	53	444	532
	7	401	467	161	187	76	83	54	58	533	710
High	8	468	534	188	213	84	91	59	64	711	887
	9	535	600	214	240	92	100	65	70	888	1064
Very high	10	601		241		101		71		1065	

technique, we focus on how to make the representation of the knowledge which is minded more understandable. Some representation forms may be better suited than others for particular kinds of knowledge. Users can choose or maybe data-driven. Visualization techniques can be classified into (graphical, tabular, or using color only). Users can access the

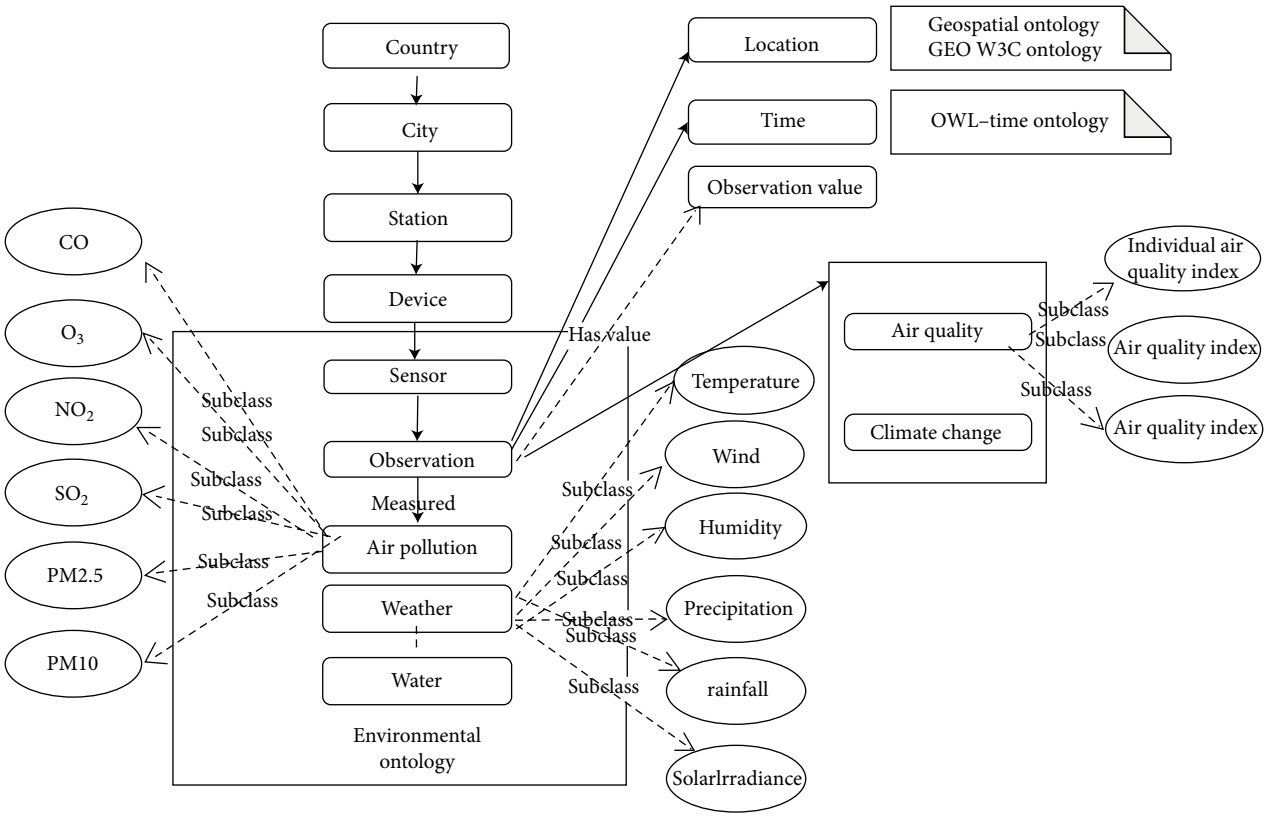

Figure 7: Smart City environmental ontology.

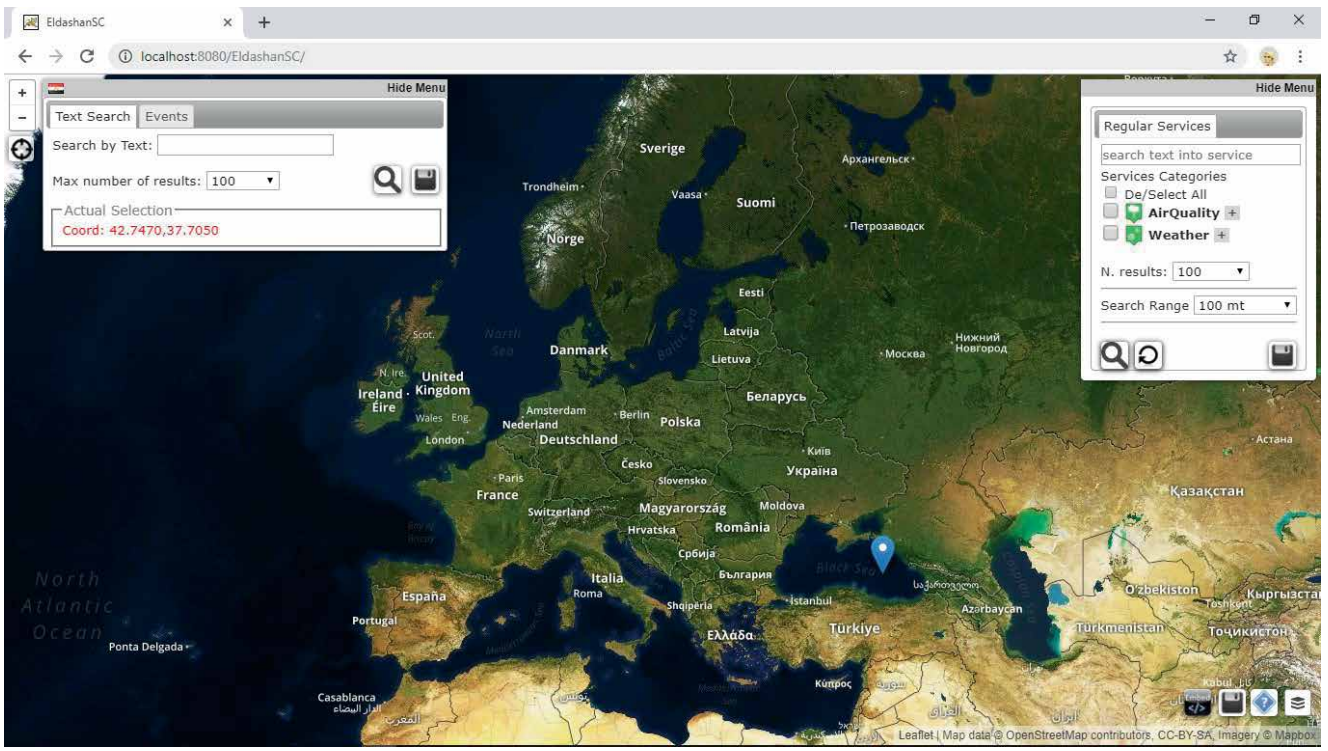

Figure 8: Dashboard for SSWF.

```
SELECT ?O3 ?NO2 ?CO ?PM25 ?PM10 ?SO2 ?Temp ?Hum ?Wind ?Rainfall ?pre
?solar WHERE {

        ?obs schema:O3 ?O3.
        ?obs schema:NO2 ?NO2.
        ?obs schema:CO ?CO.
        ?obs schema:PM2.5 ?PM25.
        ?obs schema:PM10 ?PM10.
        ?obs schema:SO2 ?SO2.
        ?obs schema:Temperature ?Temp.
        ?obs schema:Humidity ?Hum.
        ?obs schema:Wind ?Wind.
        ?obs schema:Rainfall ?Rainfall.
        ?obs schema:precipitation ?pre.
        ?obs schema:Solarirradiance  ?solar.
        ?obs schema:year ?year.
        ?year property:yearNum ?yn.
    ?obs schema:sensor ?sensor.
        FILTER( ?yn >= 1970 && ?yn <= 2010)
}
```

Figure 9: Dynamic query to retrieve all air quality and weather parameters filtered in different period of times.

proposed framework by a set of data visualization components like a dashboard, mobile application and APL's.

4. Implementation of the Proposed Framework SSWF

In this section, we illuminate hardware and software packages used in the proposed framework SSWF and explain data flow for the proposed framework SSWF.

For Implementing SSWF, we need to use suitable software in each layer. Where we use the HPC System of Bibliotheca Alexandrina which has a SUN cluster of peak performance of 11.8 Tflops, 130 eight-core compute nodes, 2 quad-core sockets per node, each is Intel Quad Xeon E5440 @ 2.83 GHz, 8 GB memory per node, Total memory 1.05 Tera Bytes, 36 TB shared scratch, Node-node interconnect, Ethernet & 4x SDR Infiniband network for MPI, 4x SDR Infiniband network for I/O to the global Lustre filesystems.

We develop a semantic dashboard using java JDK, then build a universal knowledge base using Protégé. The data are pushed from different data sources to NoSQL storage by Big Queue tool. Depending on the size of the data, the SSWF stored data in HDFS or Hbase. The SSWF processed data as batch processing in case of historical data or streaming processing in case of real-time data. The SSWF used Sempala Alexander [20] as interactive SPARQL query processing on SQL on Hadoop. The SSWF generates a dynamic SPARQL query over the universal semantic data model of the city.

A complete example of how a dynamic SPARQL query is translated to Spark SQL is illustrated in Figure 5, (1) the SPARQL query asks for an average of Q3 in "London" during 2010, the corresponding algebra tree is illustrated in (2) and the Spark SQL query is given in (3).

In the data processing phase; Alkatheri [21] built a comparative study among big data frameworks. In comparison with Spark, Apache Storm, Flink and Apache Hadoop frameworks for nonreal-time data, this comparison recognized Spark as a winner across various key performance indicators (KPI), while, for stream processing, Flink was the best. These KPIs are processing time, CPU consumption, Latency, Execution time, task performance, and Scalability. We compare Spark and Flink frameworks on high-performance computing (HPC). We found that the Spark Framework is the best framework against the pervious KPIs.

Spark is very fast and easy to collect a huge amount of data processing. Apache Spark is a distributed processing framework that works on the in-memory system. It is known for its high performance. It is easy to use and has flexibility with efficiency in handling huge data. Also, it supports application development in languages like Python and Java using Hadoop based storage system.

Table 2 presents the software packages used in the proposed framework SSWF and its functions.

A corresponding sequence diagram illustrating this data flow process is shown in Figure 6.

5. Case Study of the SSWF to Analysis Air Pollution and Weather on Migratory Birds' Path

World health organization (WHO) shows that 9 out of 10 people breathe air containing high levels of pollutants. It estimates

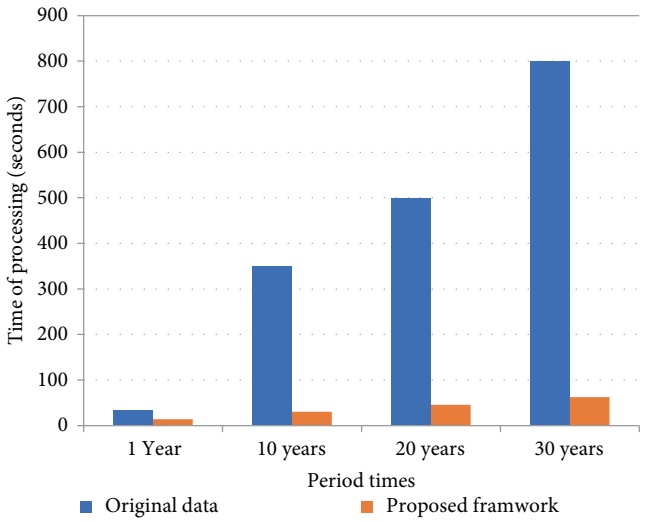

Figure 10: Bar chart for time processing for original data and proposed framework in different period of times.

Table 5: The comparison between time processing of the data (seconds) and a different period of times.

Time	Original data	Proposed framework
	Time of processing (seconds)	Time of processing (seconds)
1 year	34.3	14.1
10 years	350.1	30.5
20 years	500.4	45.3
30 years	800.6	62.4
40 years	1500.50	75.2

Table 6: Gamma association coefficient between the daily average of humidity and PM2.5.

No of event	Humidity	PM25	γ (Gamma association coefficient)
278	Low	Low	76.16%
62	Not low	Low	16.98%

```
SELECT
?station avg(?pm10)?long ?lat  WHERE {
?obs schema:PM10 ?pm10.
?obs schema:station ?station.
?station property:longitudeDegree ?long.
?station property:latitudeDegree ?lat.
?stationObj schema:inCity ?cityObj.
?obs schema:year ?year.
?year property:yearNum ?yn.
?cityObj property:city ?city.
?obs schema:sensor ?sensor.
FILTER(?city = "LONDON"^^xsd:string && ?yn >= 2000 && ?yn <= 2010)

} group by ?station  ?long ?lat
```

Figure 11: The monthly average air quality index for PM10 over London from 2008 to 2012.

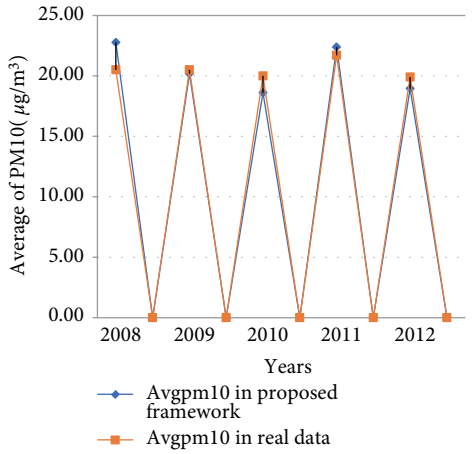

Figure 12: The monthly average air quality index for PM10 over London from 2008 to 2012.

that around 7 million people die every year from exposure to polluted air [30]. Most organizations deny access to their data by external researches due to privacy issues.

We study air quality [30] and weather forecasting [31] monitoring data for 40 European countries from 1969 to 2012. The size of the data per year is 1.5 GB. Multiple weather factors (temperature, wind speed, humidity, rainfall, etc,) are taken into consideration based on hourly monitoring. Air pollutants included particulate matter with an aerodynamic diameter $\leq 10\,\mu$m (PM10), PM2.5, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and Ozone (O_3). The European Environment Agency (EEA) launched the European Air Quality Index (AQI) to check the current air quality across Europe's cities and regions. We compare our results with the European Air Quality Index (AQI) (http://airindex.eea.europa.eu/) to check our predictions.

```
Events(?e) ^ E_PM25(?e,?m) ^ humidity  (?ws)^ swrlb:lessThan(?m, 70) ^
swrlb:greaterThan(?m, 10) ^ swrlb:lessThan(?ws, 20.0) –> swrlq:select(?e) ^
swrlq:orderBy(?m)
```

Figure 13: SWRL to count "*a*".

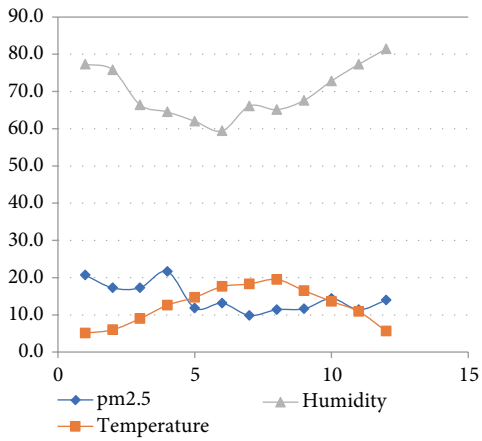

Figure 14: Curve relation among humidity, temperature, and PM2.5 during 2011 over London.

We discover knowledge by applying the Gamma association coefficient on certain classes in the universal knowledge base. So, we build association rules discovery between PM2.5, temperature, and humidity during 2011 over London.

We predict the value of any attribute in the universal knowledge base. The main challenge is the missing data in some places in the migratory birds' path. So, the SSWF search for the nearest area has data and predicts the missing data. In this case; we predict the annual average of NO_2 over Egypt in 2011.

Now, we apply the SSWF in the following sections.

5.1. Data Storage. At this stage, data can be collected from different sources. BigQueue pushes the data from all sources to the NoSQL storage and HDFS. It can open one connection to the database and synchronize others. In order to speed up the query location added in Geo column family, Date Time information domain is added in DateTime column family, Air pollution information domain is added in Air quality column family, and weather information domain is added in weather column family, respectively, with the logical view of the entire table shown in Table 3.

5.2. Universal Knowledge Base. We create a common knowledge base Smart City Environmental Ontology (SCEO) that can deal with static, semi-static and real-time data. SCEO can be used to make queries for predictions, suggestions, and deductions.

SCEO is a universal Ontology, which merges more than one ontology. This merge will extend the knowledge base. This merge is necessary for knowledge transfer among different knowledge bases. The common between ontologies is the location and time components.

The location and time components use the standard GEO W3C Ontology (http://www.w3.org/2003/01/geo/). We define a set of association rules for air quality index and weather. Table 4 shows the air quality index with air pollution parameters value.

Figure 7 shows a general picture of the ontology of the Smart City environmental structure. SCEO contains set of classes like (County, City, Station, Device, Sensor, Air pollution, Weather, Water, Air Quality, Climate Change). It is also contain set of subclasses like air pollution has (CO, O_3, NO_2, SO_2, PM2.5.PM10) and weather has (temperature, wind, humidity, rainfall, etc,).

5.3. Data Processing. In this phase, Dynamic SPARQL query can be processed over large distributed datasets in memory efficiently on top of the existing cluster HPC platform without data preparation overhead. Dynamic SPARQL query can be run over Sempala which converts SPARQL query to algebraic representation and then to Spark SQL.

5.4. Data Visualization. Geographical Dashboard has been implemented. Apache server tomcat was used to host Dashboard. Figure 8 shows the dashboard for the proposed framework—an easy to configure Dynamic SPARQL query using dashboard controls. The result can be filtered by a range or the number of output values. The dashboard provides an Animated Marker Clustering.

6. Analysis of Results

For evaluation purposes, we measure the time of processing a dynamic query to retrieve all air quality and weather parameters filtered in different periods of times as shown in Figure 9. Table 5 shows the comparison between normal RDF and the proposed framework in the processing time of query code 2. Figure 10 shows the bar chart for time processing (seconds) for original data and the proposed framework in different periods of time.

We measure monthly average Air quality index for PM10 over London from 2008 to 2012 https://data.london.gov.uk/ and our proposed framework as shown in Figure 11.

Figure 12 shows the comparison between the monthly average Air quality indexes for PM10 over London from 2008 to 2012 and our proposed framework. The matching ratio in the air quality index between the framework calculation and the real data is 98%.

Now, we can discover knowledge by applying Gamma association coefficient on any two classes.

We build association rules discovery between PM2.5 and temperature and humidity during 2011 over London.

Where *a* is the number of PM2.5 in each rule. i.e., satisfy the two classes at the same time.

e is the total number of PM2.5 in each class of the second class,

T is the total number of the sample.

For example, if we do not count a number of PM2.5 (events) with range low (10–20 (μg m^{-3})) and humidity low type (less than 70), then Figure 13 shows the Semantic Web Rule Language (SWRL) Rule.

Table 6 visualizes the outcome of the Gamma association coefficient and a relational association rules discovery for the daily average of humidity and PM2.5 during 2011 over London.

There is a strong relation between PM2.5 class low (10–20 (μg m^{-3})) and humidity low type (less than 70) and week relation between low PM2.5 and not low humidity.

Figure 14 shows the relation curve among humidity, temperature, and PM2.5 during 2011 over London that confirms the above knowledge discovery rule.

Finally, we can apply association rule discovery functions to predict the value of any attribute in the universal knowledge base. The main challenge is the data are not available in some countries or cities in migratory birds' path. So, SSWF searches for the nearest area that has available data and predicts new area data value. In this case, we predict the annual average of NO_2 over Egypt in 2011. Our case study data do not have any information about Egypt. The nearest area of Egypt is Cyprus that is far away from Egypt with 956 KM. According to the Egyptian Environmental Affairs Agency (EEAA) (http://www.eeaa.gov.eg), the annual average of NO_2 over Egypt in 2011 is 58 (μg m^{-3}). The framework predicts the annual average of NO_2 is 56.25 (μg m^{-3}) with an accuracy percentage of 96.9%.

7. Conclusions

The smart world is a dream, but we can do it, like the migratory birds around the world without a visa. This paper presents a framework aiming to build a general semantic big data framework for a smart world. The SSWF provides a universal knowledge base for data generated by different data sources. SSWF provides not only data but understandings of the meaning of data readings, context and relationships among data, facts, and events. SSWF has been adding millions of records in RDF triples by using air quality and weather monitoring data for 40 European countries. The advantages of this framework are (1) the increasing ability of end-users to self-manage data from different data sources. (2) Independent framework from the domain of services and environments. (3) Manages concepts and relationships from different data sources. The main characteristics of this framework are (1) build a universal semantic data model. (2) Define a set of association rule discovery for prediction, suggestions, and deductions. (3) Service-Oriented Architecture (SOA). (4) Use of semantic RDF standards to make the data "self-describing". (5) Management of big data. The matching ratio between framework calculation and real data in the Air quality index is 98%. The matching ratio between framework calculation and real data in prediction new values is up to 96.9%. Finally, in future work, we will add more analysis techniques. We will merge Smart City ontologies with different domains to increase knowledge and pattern detection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the HPC team who support our framework hardware and like to thank air quality and weather data providers.

References

[1] D. Evans, "The internet of things: how the next evolution of the internet is changing everything," 2011.

[2] E. E. Santana, "Software platforms for smart cities: concepts, requirements, challenges, and a unified reference architecture," *ACM Computing Surveys (CSUR)*, pp. 78–110, 2018.

[3] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, pp. 38–52, 2014.

[4] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A communications-oriented perspective on traffic management systems for smart cities: challenges and innovative approaches," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 125–151, 2015.

[5] Y. A.-Y. Zheng, "A cloud-based knowledge discovery system for monitoring fine-grained air quality," 2014, Microsoft Tech Report.

[6] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu, "Inferring gas consumption and pollution emission of vehicles throughout a city," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, NY, USA, 2014.

[7] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Beijing, China, 2012.

[8] Y. F. Xiong, "Modeling of geographic dependencies for real estate ranking," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, pp. 11–35, 2016.

[9] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, ACM, San Jose, CA, 2010.

[10] K. R. Jayashree, R. Abirami, and R. Babu, "A collaborative approach of IoT, big data, and smart city," *Big Data Analytics for Smart and Connected Cities*, IGI Global, pp. 25–37, 2019.

[11] G. D'Aniello, M. Gaeta, and F. Orciuoli, "An approach based on semantic stream reasoning to support decision processes in smart cities," *Telematics and Informatics*, vol. 35, no. 1, pp. 68–81, 2018.

[12] D. Puiu, P. Barnaghi, R. Tonjes et al., "CityPulse: large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.

[13] A. M. S. Osman, "A novel big data analytics framework for smart cities," *Future generations computer systems*, vol. 91, pp. 620–633, 2019.

[14] N. Zhang, H. Chen, X. Chen, and J. Chen, "Semantic framework of internet of things for smart cities: case studies," *Sensors*, vol. 16, no. 9, pp. 1501–15024, 2016.

[15] D. Pfisterer, K. Romer, D. Bimschas et al., "SPITFIRE: toward a semantic web of things," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 40–48, 2011.

[16] X. Liu, A. Heller, and P. S. Nielsen, "CITIESData: a smart city data management framework," *Knowledge and Information Systems*, vol. 53, no. 3, pp. 699–713, 2017.

[17] K. K. Mohbey, "An efficient framework for smart city using big data technologies and internet of things," *Advances in Intelligent Systems and Computing*, pp. 319–328, 2019.

[18] S. E. Bibri, "The IoT for smart sustainable cities of the future: an analytical framework for sensor-based big data applications for environmental sustainability," *Sustainable Cities and Society*, vol. 38, pp. 230–253, 2018.

[19] Coefficient Gamma, 2019, https://www.statisticshowto.datasciencecentral.com/gamma-coefficient-goodman-kruskal/.

[20] S. Alexander, P.-Z. Martin, N. Antony, and L. Georg, "Sempala: interactive SPARQL query processing on Hadoop," in *Proceedings of the 13th International Semantic Web Conference*, Springer, pp. 19–23, Italy, 2014.

[21] S. Alkatheri, S. Abbas, and M. Siddiqui, "A comparative study of big data frameworks," *International Journal of Computer Science and Information Security (IJCSIS)*, 2019.

[22] S. Zareian, M. Fokaefs, H. Khazaei, M. Litoiu, and X. Zhang, "A big data framework for cloud monitoring," in *Proceedings of the 2nd International Workshop on BIG Data Sofware Engineering - BIGDSE*, IEEE, pp. 58–64, Austin, TX, USA, 2016.

[23] Apache Hadoop, 2019, https://hadoop.apache.org/.

[24] Apache HBASE, 2019, https://hbase.apache.org/.

[25] Apache Spark, 2019, https://spark.apache.org/.

[26] SPARQL Query language, 2019, https://www.w3.org/TR/rdf-sparql-query/.

[27] Web ontology language, 2019, https://www.w3.org/OWL/.

[28] Resource Description Framework, 2019, https://www.w3.org/RDF/.

[29] Java Language, 2019, https://www.java.com/.

[30] European Environment Agency, https://www.eea.europa.eu.

[31] European Climate Assessment & Dataset, 2019, https://www.ecad.eu/