BIG DATA ANALYTICS AND SMART CITIES: A LOOSE OR TIGHT COUPLE?

Ahmed M. Shahat Osman
Ahmed Elragal
Birgitta Bergvall-Kåreborn
Luleå Technical University- Department of Computer Science, Electrical and Space Engineering
971 87 Luleå - SWEDEN

ABSTRACT

Smart City (SC) is an emerging concept aiming at mitigating the challenges raised due to the continuous urbanization development. To face these challenges, government decision makers sponsor SC projects targeting sustainable economic growth and better quality of life for inhabitants and visitors. Information and Communication Technologies (ICT) is the enabling technology for smartening. These technologies yield massive volumes of data known as Big Data (BD). If spawned BD are integrated and analyzed, both city decision makers and citizens can benefit from valuable insights and information services. The process of extracting information and insights from BD is known as Big Data Analytics (BDA). Although BDA involves non-trivial challenges, it attracted academician and industrialist. Surveying the literature reveals the novelty and increasing interest in addressing BD applications in SCs. Although literature is replete with abundant number of articles about SCs applications harnessing BD, comprehensive discussion on BDA frameworks fitting SCs requirements is still needed. This paper attempts to fill this gap. It is a systematic literature review on BDA frameworks in SCs. In this review, we will try to answer the following research questions: what are the big data analytics frameworks applied in smart cities? what are the functional gaps in the current available frameworks? what are the conceptual guidelines of designing integrated scalable big data analytics frameworks for smart cities purposes? The paper concludes with a proposal for a novel conceptual analytics framework to serve SCs requirements. Additionally, open issues and further research directions are presented.

KEYWORDS

Big data, Big data analytics Frameworks, Smart cities

1. INTRODUCTION

The concept of smart cities (SC) emerged as a strategy to mitigate unprecedented challenges of continuous urbanization, while at the same time provide better quality of life to the citizens (Hafedh Chourabi, et al., 2012). City smartness is realized by means of the advances of Information and Communication Technologies (ICT) (YIN ChuanTao, et al., 2015) and as a result SCs are usually characterized by an extensive use of digital technologies in various city domains in combination with a holistic view of the city where different domains should be closed integrated. Furthermore, the diffusion of digital technologies in people's daily life has boosted human-to-human, human-to-machine, and machine-to-machine interactions. These interactions yield massive volumes of data, commonly known as "Big Data" (BD) and characterized by large and fast growing volumes of complex datasets, which go beyond the abilities of commonly known data management systems to accommodate. By analyzing these data volumes valuable insights and correlations can be extracted (Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, & Athanasios V. Vasilakos, 2015). However, BD complexities result in non-trivial challenges when performing Big Data Analytics (BDA) processes (YIN ChuanTao, et al., 2015). Although literature is replete with articles addressing applications of BDA in different SC domains, detailed discussions on integrated BDA frameworks fitting SCs requirements are still needed. The lack of this type of articles is the primary motive for this research.

This article is a systematic literature review on BDA frameworks in SCs aiming at answering three basic research questions. RQ1: What types of BDA frameworks are available for the smart city context? RQ2:

What are the functional gaps in the current available frameworks? Finally, RQ3: What conceptual guidelines for designing integrated scalable BDA frameworks, relevant for smart city contexts, can be found in the literature? The literature review analyzed 30 articles addressing BD applications in SCs. The review process followed the widely known framework proposed by Vom Brocke.

This paper is organized as follows: Following this introduction a presentation of key the concepts of big data, data analytics, and smart cities are given. After this, the scope and method of the literature review is presented together with a discussion of the findings. Thereafter, the main contribution of this paper, i.e. a proposal of a novel conceptual BDA framework for SCs is described followed by recommendations for future research directions. The paper ends with a summary and conclusion.

2. TOPIC CONCEPTUALIZATION

In this section, the fundamental concepts of the study, “big data” and “smart cities”, are reviewed for better understanding for the challenges of harnessing big data analytics in various SC domains.

2.1 Big Data (BD)

Big data (BD) is a natural crop of the advanced digital artifacts and their applications. Sensors, mobiles and Social Media Networks are examples of modern digital technologies that have permeated our daily lives. Penetration of these technologies in our lives yields unprecedented massive volumes of data known as “Big Data”. The term “Big Data” refers to the continuously growing datasets to the extent it is difficult to manage using traditional relational database management systems. BD is commonly characterized by the four Vs: Volume, Velocity, Variety and Veracity (Figure 1).

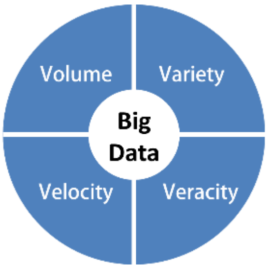

Figure 1 - Big data 4Vs

Volume of BD, as outlined before, goes far beyond the size of traditional operational databases or data warehouses. Traditional databases usually grow to the order of gigabytes or terabytes. Volumes of BD are big enough to the extent that new measuring units are required such as Petabyte (10^{15}) and Exabyte (10^{18}).

Velocity of BD refers to the high rate of data streaming into hosting platforms. In addition to the high rate of incoming data, velocity raises important question about data aging, i.e. for how long these data will be valuable. In some cases, real time analysis of streaming data is critical. For example, real time analysis for video streams captured by traffic surveillance cameras is critical to predict traffic jams and prevent bottlenecks.

Variety refers to the complexity of BD formats. BD is mostly composed of semi-structured and unstructured data, e.g. text data files and images, and stream data, e.g. geospatial data streams. This is in addition to traditional structured data. Usually, the ratio between structured data to other data types is estimated by 20% to 80%, respectively. We believe that the adjective "big" will fade over time. The self-evident meaning of “data” will intuitively be extended to include all data types mentioned above.

Veracity refers to the trustworthiness of the data. Incorrect data will definitely lead to misleading analytics. Therefore, there is a need to ensure that the data is correct as well as the analyses performed on the data are correct, especially in the case of automated decision-making, where no human is involved.

In fact, the above four V's are not the only characteristics of BD. Value, Variability and Visibility represent another set of dimensions. However, the first four are the most commonly known ones. These V characteristics bestow a wild nature to BD compared to the traditional structured data.

The chain of activities for extracting values out of BD is known as BD value chain (H. Gilbert Miller & Peter Mork, 2013; Curry, 2016) or BDA (Chun-Wei Tsai, et al., 2015). Although BDA involves non-trivial processes and many challenges due the wild nature of BD. However, BDA holds an unprecedented opportunity to shift the traditional methods of information extraction into new dimensions.

2.2 Big Data Analytics (BDA)

To deal with the challenges of BD, platform scalability is the intuitive solution. There are two commonly known scaling approaches: vertical and horizontal scaling or known as scale up and scale out, respectively.

Vertical scaling means empowering the processing platform with additional computing resources (memory, CPUs, disk space, etc.), to accommodate with the incremental volume of data. This approach usually involves single instance of an operating system.

Horizontal scaling, however, is a divide-and-conquer approach where the workload is distributed and processed in parallel across multiple independent computing machines. More machines can be added as much as needed to improve the overall system performance. This approach involves multiple instances of operating systems running on independent machines.

Of course, each approach has its advantages and disadvantages. For example, up or vertical scaling is restricted to the upper ceilings of the system upgrades. Nonetheless, out or vertical scaling shows more elasticity in this regard. In fact, scaling out is more complicated than scaling up since multiple instances of different operating systems have to be managed by a single mastering node. Table 1 shows a summarized comparison between vertical and horizontal scaling.

Table 1- Comparison between vertical and horizontal scaling

	Vertical Scaling	Horizontal Scaling
Scaling	Limited, compelled by technology manufactures.	Unlimited, unrestricted to specific technology
Management complexity	Easy, Most software applications can benefit additional resources easily.	Complex, management software has to handle the complexities of parallel processing on distributed data.
Financial investment	Considerable financial investment.	Affordable, added machinery is high-end servers.

In fact, the financial investment and scalability upper ceiling aspects are the major drawbacks of vertical scaling. These two considerable drawbacks come in favor of horizontal scaling when it comes to SCs. It is more reasonable to rely on horizontal scaling rather than vertical scaling in a multi-domain SC projects. This note interprets why most of the researches and designs of BDA frameworks and platforms are built using horizontally scalable platforms.

As for the data analysis, (Chun-Wei Tsai, et al., 2015) provided a thorough survey on the studies on BDA frameworks, platforms, machine learning and data mining algorithms. The survey demonstrated the dominant technologies in the area of BDA, such as parallel processing, cloud computing and Apache Hadoop. Apache Hadoop is a widely known open-source fault-tolerant software framework used for distributed storage and parallel processing of BD (Apache, n.d.).

(Chun-Wei Tsai, et al., 2015) adopted analogues approach using Knowledge Discovery in Databases (KDD) model to study the corresponding bottlenecks arise in the case of BD rather than data. In the KDD model, analytics is divided into three operators: input, analysis and output (Figure 2). In the case of BD, the processes of input operator affect the efficiency of the performance of data analysis execution. Although input operator processes are of high importance, Chun-Wei noted a significant observation that number of research articles and technical reports focusing on data analysis is significantly more than the number focusing on other operators.

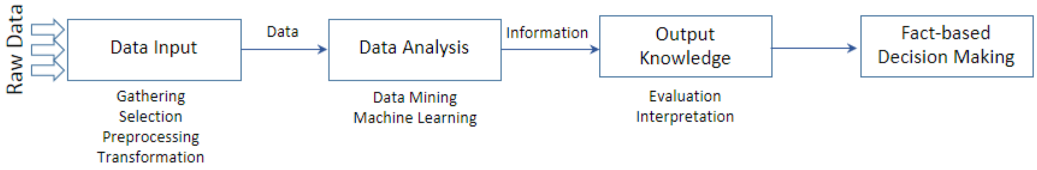

Figure 2 - Knowledge Discovery in Databases

2.3 Smart Cities (SCs)

Although the term “Smart City” seems to be simple and intuitively understandable, there is no global agreement on a unified definition of the term. There are various number of definitions in literature. Despite the variations of these definitions, there is a common consent on the centric role of ICT in realizing city smartness (Michael Kehoe, et al., 2011; YIN ChuanTao, et al., 2015). Within the context of SCs, the meaning of “smartness” has different connotations according to which perspective it is regarded; marketing, planning or technology (Nam, 2011). From technology perspective, a SC is composed of smart components such as smart buildings, smart farms and smart hospitals. “Smartness” of these components is realized through intensive use of digital artifacts such as sensors, actuators, mobiles and ICT applications. However, “smartening” of various city domains is not enough for a city to be smart. On the city level, a SC is viewed as a whole body of systems, *i.e. system of systems*, where the interrelationship between the underlying systems and subsystems are taken into account. This integrated vision for a SC implies cross domain sharing of information (Michael Kehoe, et al., 2011) and clarifies the difference between smartening of specific city domain and smartening a city as a whole. Therefore, the process of evaluating and modelling SCs encounters considerable complexities (Leonidas G. Anthopoulos, et al., 2015).

Researches pursue different approaches to establish assessment frameworks for SCs (Giffinger, 2010; Nam, 2011; Hafedh Chourabi, et al., 2012; Lazaroiu, 2012; Lombardi, 2012; Söderström, 2014; Neirotti, 2014; Felix Herrera Priano & Cristina fajardo Guerra, 2014; Lee, 2014). However, the pioneering work of Rudolf Giffinger, et al. is the most common in literature (Giffinger, 2010). Giffinger defined six main traits or domains that characterize the city smartness. These domains are: smart economy, smart people, smart governance, smart mobility, smart environment and smart living. These domains are further divided into factors and evaluate indicators.

2.4 Smart Cities (SCs) and Big Data (BD)

From an ICT perspective, industrialist and scholars adopt the layered (or tiered) approach to model smart cities. For example, IBM adopt a three-layer model for SCs including instrumented layer, interconnected layer, and intelligent layer (Michael Kehoe, et al., 2011). Similarly, YIN ChuanTao proposed a four-layer model for SCs: data acquisition and transmission layer; data vitalization layer; common data and service layer; finally applications layer (YIN ChuanTao, et al., 2015). Despite the number of layers, these models aim at projecting the journey of the data from its birth in raw form until valuable information is extracted benefiting end users, the citizens and decision makers. Recognizing the analogy (or even congruence) between SC models and BD value chain is straightforward. This analogy manifests the intersecting area between BDA and SCs.

In this context, an important question poses itself, what are the characteristics and features of the analytics platforms/frameworks to benefit BDA in SCs? The answer of this question led to the first research question of this article RQ1. In the literature, there are many articles about BD applications in SCs proposing solutions for specific domains such as energy and transportation using BDA frameworks. However, most these articles discuss functional requirements and architecture designs for specific domain. To develop an appropriate domain-independent and resilient analytic framework for SC purposes, there are two driving aspects that should be considered. Firstly, the framework should be capable, at least, of managing the 4V’s of BD. Secondly, the design of the framework should consider, at a high level, functional and non-functional requirements pertained to SC’s from a holistic perspective.

3. THE SCOPE AND SEARCH PROCESS OF THE REVIEW STUDY

This systematic literature review is based on the taxonomy proposed by (Cooper, 1988) and adapted by (Vom Brocke, 2009). In this literature review the focus is on research outcomes, practices and applications. The goal is to integrate finding about BDA frameworks in the development and evolution of smart cities and analyze the findings in order to identify research gaps. These gaps guide the design of integrated scalable BDA and future research directions and identify both conceptual and methodological structures. The perspective introduced leans to be neutral representation rather espousal of position and the coverage is exhaustive coverage with selected citations. Finally, the target audiences are specialized scholars, practitioners and SCs planners and decision makers. Literature search involves four steps: 1) identifying search databases; 2) search keywords; 3) forward and backward search and 4) evaluation of articles.

Step 1-Identifying search databases: To collect quality scholar articles and conference proceedings, the following internationally recognized online information systems databases were used for the search process: ACM DL, IEEE, SCOPUS, Springer Link, INSPEC and Web of Science.

Step 2-Search- keywords: As this paper comes at the intersection between the two subjects, the search keywords were “big data” and “smart city” and the search parameter was the document title. The search keywords did not include the two words “framework” or “analytics” to widen the scope of the search process as these keywords might not be used explicitly in the article title while they are used within the article contents. The search process was limited to English publications only. The total number of the retrieved publications was 247 distributed as shown in Table 2:

Table 2 - Search Databases and number of hits

Database	ACM DL	IEEE	SCOPUS	Springer Link	INSPEC	Web of Science	Total
No of hits	6	26	53	114	40	8	247

Step 3-Forward- and backward search: Returned articles were published between 2012 and 2017, which reflect the relative novelty of subject. Additionally, the 247 articles are considered enough pool for the analysis step, hence the authors decided not to apply forward nor backward search.

Step 4-Evaluation- of articles: Search database, article id, type, title, year, source title, authors, volume, abstract, URL and DOI attributes of the retrieved publications were collected in an Excel sheet. An elimination process was applied to exclude duplicates, work in progress and irrelevant publications from the analysis process, resulting exclusion of (85) articles. The remaining (162) articles are further filtered according to how likely they address the main objective of this review i.e. big data analytics in smart cities. This filtration process was performed by evaluating the abstracts of the remaining (162) articles, resulting exclusion of additional (132) articles do not fit with the objectives of this article. The remaining (30) articles were used for analysis. The sources of the final articles candidate for analysis are listed in Appendix A.

3.1 Literature analysis and synthesis

The final list of the thirty articles subjected to the analysis are classified according to their relevance to the three operators of BDA value chain: input, analysis, and output (Figure 2). Articles classification process involved reviewing each article’s abstract and conclusion to decide how likely the article belong to one or more of the three classification operators. In case of article abstract and conclusion does not lead to clear decision for classification, the article full body is reviewed for final decision. Results of the classification process are shown in Appendix B.

To answer research questions RQ1 and RQ2 the contribution(s) of the articles are evaluated to identify the uncovered requirements of the value chain operators/processes and SC domain, if any. Compiling the results of the two research questions will derive the answer for the third research question RQ3. The answer to the last research question will derive the design of the conceptual BDA framework to serve SCs requirements.

3.1.1 Data Input

In (Takuro Yonezawa, et al., 2016) authors introduced a scalable, distributed infrastructure city-wide sensor network for sharing *social big sensor data* in SCs called SOXFire. By social big sensor data authors mean the data which has usefulness and sociality to be shared. Through this access city employees, citizen and WEB developers can contribute through unified APIs. Design of *SOXFire* rely on already matured IoT communication protocols. The essential design goals of *SOXFire* are: coping with heterogeneous types of sensors; securing access, that enable different formats; easing management; and providing Scalability and Federation of multi-community system.

In (Bo Tang, et al., 2015), authors presented a prototype for hierarchical distributed architecture based on Fog Computing to integrate massive number of infrastructure components and services in SCs. Fog computing is a decentralized computing infrastructure in which data, compute, storage and applications are distributed in the most logical, efficient place between the data source and the cloud. The prototype was tested to evaluate event detection performance of the recognition of 12 distinct events. Results demonstrate the feasibility to implement the proposed hierarchical architecture in a city-wide implementation.

Civitas is a distributed object-oriented middleware designed for SCs (Villanueva, et al., 2013). It is used to facilitate the development and deployment of SC applications. Citizens connect to the middleware via a special device called the *Civitas Plug* to ensure privacy and security. To promote software consistency and reusability. The middleware design adopted two principles: Everything is a Software Object and City Layout should be Independence.

In (Kemp, G., et al., 2015) authors presented an approach for developing data storage, cleaning and integration services. Such integration aims at making an efficient decision support system based on cloud computing using a new programming paradigm called Service Oriented Computing. The objective of the presented approach is to manage and aggregate cloud services for managing BD and assist decision making for transport systems. The proposed decision support system relies on the following services: data acquisition service; information extraction and cleaning service; integration and aggregation services; BD analysis service; and decision support service. The proposed approach was applied to the transport industry, which can bring new understanding to town transport infrastructures.

3.1.2 Data Analysis

In (Debopriya Ghosh, et al., 2016) authors presented intelligent solution using machine learning to automate and support crime analysts. The proposed solution enables identifying connected entities and events by collecting, integrating and analyzing diverse data sources. It helps crime analysts identify dead end leads to avoid having the investigators spend time on these leads. A prototype for the proposed solution is built using R package for machine learning classification module and preprocessing with MySQL as a backend database.

A novel system for sentiment analysis using convolutional neural network is introduced in (Tang B, et al., 2015) for visual sentiment prediction. The research focuses on images and fine-tuned convolutional neural network initially trained on a large natural image recognition dataset (Imagenet ILSVRC2012) to learn feature representations for visual sentiment prediction. To evaluate the proposed method, real world dataset has been taken from Twitter. Authors used a benchmark including 603 tweets with photos. Experimental results led to two important conclusions: 1) Logistic regression model has more accuracy compared to J48 and Random Forest; and 2) Domain specific fine-tuning is effective in improving the performance of neural network.

(Ciprian Barbieru & Florin Pop, 2016) addressed important issue regarding scheduling of analysis tasks in a SC environment. In some cases, data must be analyzed on real-time bases while others can be analyzed as a batch process. To handle these cases, authors design a real-time and job scheduler using Hadoop for BD processing. Hadoop processing addresses both the problem of small tasks that need to be executed in real time, and in the same time, adjust long-running jobs. A case study is applied as a support of SC applications. Data gathered, routed, stored via mobile devices, while processed and diffused via cloud. Experimental results showed that the proposed scheduler behaved better than the standard one when fed with the same job distribution. Unfortunately, performance degraded when the distribution is the exact opposite. Authors expect that this can be solved by choosing other schemes of resource limits in real implementation.

Authors of (D. Singh, et al., 2016) proposed a BDA framework for video recorded data, through the analysis of images or videos to find semantic patterns that are useful for interpretation. The research aims at providing a solution for city traffic control for detection of bike-riders without helmet in city traffic. The proposed framework utilizes data recorded by video surveillance cameras for visual data analytics.

In (Tosi D & Marzorati S., 2016) authors describe a novel use of BD coming from the cellular network to extract mobility patterns for SCs. These mobility patterns are able to describe different mobility scenarios of the city, starting from how people move around Point of Interests (PoI) of the city in real-time. These mobility patterns can be exploited by policy makers to improve the mobility in a city or by navigation systems and journey planners to provide users with accurate travel plans. Several BD mining algorithms have been discussed to support the prediction and estimation of vehicular traffic conditions, speed profiles for roads, flows of people moving among subway stations and around PoI of the city.

3.1.3 Output

Few articles address the output operator. In (Rathore MM, et al., 2016) authors addressed the subject as a part of complete proposals for BDA system for SCs based on IoT. Results are displayed to end user in terms of predefined graphs to facilitate decision making process.

3.2 Discussion

From the above literature survey, we can draw the following observations about big data analytics frameworks from infrastructure, platform and functional aspects.

Since SCs rely on sensed data collection, IoT technology provides a viable means for this requirement via large number of data acquisition devices through the Internet. Also, cloud computing technology provides an elastic, robust and high available infrastructure technology capable for accommodation with the dynamic nature of SCs and increasing volumes of data. Marriage between the two technologies, i.e. IoT and cloud computing (sometimes referred to as ClouT),) represents a workable infrastructure platform for BDA. Additionally, Fog computing is an efficient infrastructure for raw data processing on the edge of the network, where raw data can be preprocessed close to data sources before analysis processing.

Hadoop/Spark provides an efficient scalable fault tolerant platform for developing efficient BDA frameworks of SCs.

Surveyed BDA frameworks dealt with both batch analysis of historical data and real-time analysis for stream data. Most of the surveyed articles introduced solutions for specific SC application domains (*e.g. mobility, sentiment analysis*…etc.). However,.) while few articles addressed BDA from the holistic perspective of SCs (Martin Strohbach, et al., 2015; Khan, Z., et al., 2015).

Despite of its importance for decision making processes, few articles addressed the output operator processes (evaluation and interpretation).

The aforementioned observations give an overview of the characteristics of BDA frameworks applied in SCs. They also provide an answer to the first research question RQ1. Furthermore, it indicates the tight relation between BDA and SCs development. However, with the re-emphasis on the holistic perspective of the SC as a multi-domain system of systems, we can identify the following missing features on the surveyed frameworks:

- How extracted knowledge for different city domains can be merged together for more detailed and complicated analytics (*i.e. ensemble modeling*)?
- Considering the computational complexities of BDA, including a repository for extracted knowledge retention, will enable saving extracted data models for future services without additional cost (*i.e. model persistence*).

The above two missing features answers the second research question RQ2 and motivated the proposal of a novel conceptual analytics framework.

4. PROPOSAL FOR NOVEL CONCEPTUAL FRAMEWORK

Based on the knowledge surveyed in the above section, we can summarize the design guidelines of integrated scalable BDA frameworks for SCs in the following points:

- Ability to analyze stream data for online and real time applications (D. Singh, et al., 2016; Debopriya Ghosh, et al., 2016),
- Ability to analyze batch data for applications that afford latency (Ciprian Barbieru & Florin Pop, 2016),
- Ability to retain extracted models for future processing,
- Ability to merge one or more extracted models for higher level of analytics.

Based on these guidelines, we present a novel conceptual framework for developing of BDA frameworks for SCs (Figure 3). The design of the proposed framework is domain independent and satisfies the above-mentioned features. The function of each component is described hereafter:

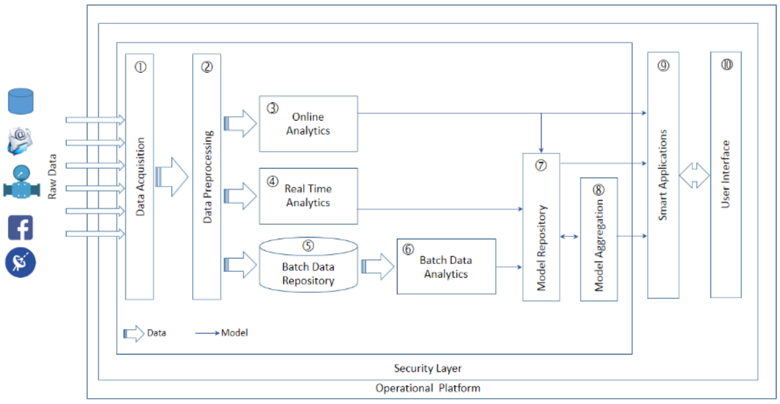

Figure 3 - Proposed Big Data Analytics Framework for Smart Cities

① **Data acquisition**: The main characteristics of this layer are scalability, interoperability and location awareness.

② **Data preprocessing**: The ability to clean input data and transform these data transformation into standard output messages.

③ **Online analytics**: The ability of performing stream data processing for applications that involve interactivity with acceptable limited latency.

④ **Real time analytics**: The ability of performing stream data processing for real time applications. Real time applications are applications that functions within time frames that the user senses as immediate or current. The latency must be less than a defined value.

⑤ **Batch data repository**: Data storage management system (e.g. Hadoop HDFS, NoSQL database management systems).

⑥ **Batch data analytics**: Batch data analytics for applications that have latency tolerance.

⑦ **Model repository**: Model storage management system, where resultant data analysis models can be persisted with relevant metadata for current and future inquiries.

⑧ **Model aggregation**: Ability to ensemble persisted resultant data analysis models for higher level and more complicated inquiries.

⑨ **Smart application**: SC applications built on resultant data analysis models.

⑩ **User interface**: End user interface that provides with efficient flexible tools allowing access, reporting and ad hoc inquiries for persisted and/or aggregated models.

Although a complete vision for the functionality of the security layer will be clearer during physical implementation, the following security measures should be adhered to on physical design especially for critical analytics:

- Restricted sign on access to the framework should be granted for critical and sensitive data;
- Multi levels user authentication;
- A complete audit log should be kept for important operations.

Apache Hadoop and Spark are viable platform options to realize the proposed framework. Hadoop has robust ecosystem that is well suited to meet the analytical needs of developers. Hadoop Ecosystem comes with a suite of tools and technologies making it a very much suitable to deliver to a variety of data processing needs. The distributed parallel processing feature of Hadoop platform enables handling massive volumes of unstructured and structured data efficiently. Also, as an open source software and horizontally scalable using commodity hardware, the initial cost saving is considerable while it can continue to grow as needed.

Technology companies provided various distributions of Hadoop bundled with other software tools to simplify Hadoop usage and extend its functionality e.g. Cloudera, Hortonworks, MapR, IBM InfoSphere, Amazon Elastic MapReduce and Windows Azure HDinsight.

5. FUTURE RESEARCH DIRECTIONS

Future research directions involve the following points:

- Realize the proposed conceptual framework using Hadoop/Spark ecosystems.
- Develop efficient model persistence and ensemble algorithms.
- Enable analytical model export/import functionality to/from other analytical frameworks.
- Develop efficient, powerful and friendly end-used interfaces. This involves multi model visualization.
- Test the framework in real case SC application.

Since the ultimate goal behind SCs is improving citizens' quality of life, elevation of the moral and cultural aspects of citizens' lives represents a significant challenge. Yet, it has been observed that this aspect is not given the necessary attention. More research dedicated to studying human privacy, behaviors, education, innovations, emotions and sentiments is required.

6. CONCLUSIONS

This paper is a literature survey on BDA frameworks in SCs. The paper aimed at answering the three research questions. RQ1: What are the big data analytics frameworks applied in smart cities? RQ2: What are the gap(s) in the current available frameworks? RQ3: What are the conceptual guidelines of designing integrated scalable big data analytics frameworks for smart cities purposes? To answer these questions, 30 articles have been reviewed and analyzed. The analysis process indicated that available proposed frameworks lack two important features, namely model persistence and ensemble. To fill in this gap, a novel conceptual framework is proposed to provide the ability to persist and ensemble extracted models. The proposed framework enables the ability to retain the extracted models, enabling future studies on these models without the need to re-analyze the data. This feature saves the considerable time for data re-analysis. Additionally, extracted model ensemble will enable the ability for more complicated cross-domain analytics.

REFERENCES

Anthopoulos, L. G., et al, 2015. Comparing Smart Cities with different modeling approaches. 24th International Conference on World Wide Web. Florence, Italy,pp. 525-528.

Apache, H., n.d. Hadoop Apache. [Online] Available at: http://hadoop.apache.org/

Barbieru, C. and Pop, F., 2016. Soft Real-Time Hadoop Scheduler for Big Data Processing in Smart City. IEEE 30th International Conference on Advanced Information Networking and Applications. Crans-Montana, Switzerland, , pp. 863-870.

Brocke, V. J. e. a., 2009. Reconstructing the giant: On the importance of rigour in documenting the literature search process. Verona, Italy, s.n., pp. 2206-2217.

Cheng, B. et al., 2015. Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander. IEEE International Congress on Big Data. New York, USA, pp. 592-599.

Chourabi, H. et al, 2012. Understanding Smart Cities: An integrative framework. Hawaii, 45th Hawaii International Conference on System Sciences (HICSS), pp. 2298-2297.

Chun-Wei Tsai, et al, 2015. Big data analytics: a survey. Journal of Big Data, 2(1), p. 20.

Cooper, H. M., 1988. Organizing knowledge syntheses: A taxonomy of literature reviews. Knowledge in Society, 1(1), pp. 104-126.

Corradi, A. et al., 2015. Automatic extraction of POIs in smart cities: Big data processing in ParticipAct. IFIP/IEEE International Symposium on Integrated Network Management (IM). Ottawa, Canda, pp. 1059-1064.

Costa, C. and Santos, M.Y., 2016. BASIS: A big data architecture for smart cities. In SAI Computing Conference (SAI). London, United Kingdom, pp. 1247-1256.

Curry, E., 2016. The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In: New Horizons for a Data-Driven Economy. Gewerbestrasse: Springer International Publishing, pp. 29-37.

Deng, L. and Zhou, C., 2015. Protype Framework of Smart City Base on Big Data and Smart Grid. 2015 International Conference on Computer Science and Mechanical Automation. Hangzhou, Zhejiang, China,pp. 102-107.

Felix Herrera Priano, F. H. and Guerra,C.F., 2014. A Framework for measuring smart cities. Proceedings of the 15th Annual International Conference on Digital Government Research. Aguascalientes, Mexico, pp. 44-54.

Ghosh, D.; Chun. S.A.; and Shafiq, B., 2016. Big Data-based Smart City Platform: Real-Time Crime Analysis. 17th International Digital Government Research Conference on Digital Government Research. Shanghai, China, pp. 58-66.

Giffinger, R. a. H. G., 2010. Smart cities ranking: an effective instrument for the positioning of the cities?. ACE: Architecture, City and Environment, 4(12), pp. 7-25.

Gilbert, M.H. and Mork, P., 2013. From Data to Decisions: A Value Chain for Big Data. IT Professional, 15(1), pp. 57-59.

Girtelschmid, S. et al, 2014. On the application of big data in future large-scale intelligent smart city installations. International Journal of Pervasive Computing and Communications, 10(2), pp. 168-182.

Hashem, I.A.T., et al, 2016. The role of big data in smart city. International Journal of Information Management, 36(5). pp. 748-758.

Horban, V., 2016. A Multifaceted Approach to Smart Energy City Concept through Using Big Data Analytics. IEEE First International Conference on Data Stream Mining & Processing. Lviv, Ukraine, pp. 23-27.

Huanan, Z.; Shijun, L. and Hong, J., 2015. Guangzhou smart city construction and big data research. International Conference on Behavioral, Economic and Socio-cultural Computing (BESC). Nanjing, China, pp. 143-149.

Iancu, V., et al, 2016. A Smart City Fighting Pollution, by Efficiently Managing and Processing Big Data from Sensor Networks. In: Resource Management for Big Data Platforms. Springer International Publishing AG., pp. 489-513.

Jara, A.J., et al, 2015. Big data for smart cities with KNIME a real experience in the SmartSantander testbed. Software - Practice and Experience, 45(18), pp. 1145-1160.

Jara, A. J.; Genoud, D.; and Bocchi,Y., 2014. Big Data in Smart Cities: From Poisson to Human Dynamics. 28th International Conference on Advanced Information Networking and Applications Workshops. Victoria, Canada,pp. 785-790.

Kemp, G., et al., 2015. Aggregating and managing realtime big data in the cloud: Application to intelligent transport for smart cities. 1st International Conference on Vehicle Technology and Intelligent Transport Systems, VEHITS 2015. Lisbon, Portugal, pp. 107-112.

Khan, Z. et al, 2015. Towards cloud based big data analytics for smart future cities. Journal of Cloud Computing, 4(1), p. 11.

Lazaroiu, G. C. a. M. R., 2012. Definition methodology for the smart cities mode. Energy - ElSevier, 47(1), pp. 326-332.

Lee, J. H. M. G. H. a. M.-C. H., 2014. Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco. Technological Forecasting and Social Change, pp. 80-99.

Lombardi, P. E. A., 2012. Modelling the smart city performance. Innovation: The European Journal of Social Science Research 25.2, 25(2), pp. 137-149.

Kehoe, M., et al., 2011. Smart Cities Series: A Foundation for Understanding IBM Smarter Cities. s.l.:IBM Redbooks.

Moreno-Cano, V. et al, 2015. Big Data for IoT Services in Smart Cities. 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT). Milan, Italy, pp. 418-423.

Nam, T. a. T. A. P., 2011. Conceptualizing smart city with dimensions of technology, people, and institutions. 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Time. Maryland, USA, pp. 282-291.

Nandury, S.V. and Begum, B.A. , 2016. Strategies to handle big data for traffic management in smart cities. 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Jaipur, India, pp. 356-364.

Neirotti, P. E. A., 2014. Current trends in Smart City initiatives: Some stylised facts. Cities, Volume 38, pp. 25-36.

Psomakelis, E., et al, 2016. Big IoT and Social Networking Data for Smart Cities Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications. s.l., s.n.

Rathore, M. M.; Ahmad, A.; and Paul A., 2016. IoT-based smart city development using big data analytical approach. IEEE International Conference on In Automatica (ICA-ACCA). Curicó, Chile,pp. 1-8.

Schatzinger, S. and Lim, C.Y.R., 2017. Taxi of the future: Big data analysis as a framework for future urban fleets in smart cities. In: Smart and Sustainable Planning for Cities and Regions. Switzerland: Springer International Publishing, pp. 83-98.

Singh, D. and Reddy, C.K. 2014. A survey on platforms for big data analytics. Journal of Big Data, 2(1), p. 8.

Singh, D.; Vishnu, C.; and Mohan, C. K., 2016. Visual Big Data Analytics for Traffic Monitoring in Smart City. 15th IEEE International Conference on Machine Learning and Applications (ICMLA). Los Angeles, California, USA, pp. 886-891.

Shuangmei Ma and Liang , Z., 2015. Design and Implementation of Smart City Big Data Processing Platform Based on Distributed Architecture. 2015 International Conference on Intelligent Systems and Knowledge Engineering, Taipei, Taiwan, pp. 428-433.

Söderström, O. T. P. a. F. K., 2014. Smart cities as corporate storytelling. City: Analysis of urban trends, culture, theory, policy, action., 18(3), pp. 307-320..

Souza, A., et al., 2016. Using Big Data and Real-Time Analytics to Support Smart City Initiatives. IFAC-PapersOnLine, 49(30), pp. 257-263.

Strohbach, M., et al, 2015. Towards a Big Data Analytics Framework for IoT and Smart City Applications. In: Modeling and Processing for Next-Generation Big-Data Technologies. Gewerbestrasse, Switzerland: Springer International Publishing, pp. 257-282.

Tang, B. et al., 2015a. A Hierarchical Distributed Fog Computing Architecture for Big Data Analysis in Smart Cities. ASE BigData & Social Informatics. Kaohsiung, Taiwan, p. 28.

Tang, B., et al., 2015b. A hierarchical distributed fog computing architecture for big data analysis in smart cities. Proceedings of the ASE BigData & SocialInformatics. Kaohsiung, Taiwan, p. 28.

Tosi, D. and Marzorati, S., 2016. Big Data from Cellular Networks: Real Mobility Scenarios for Future Smart Cities. IEEE Second International Conference on Big Data Computing Service and Applications. Oxford, UK, pp. 131-141.

UN - Department of Economic and Social Affairs, 2014. World Urbanization Prospects - The 2014 Revision, New York: United Nations.

Villanueva, F. J., et al, 2013. Civitas: The Smart City Middleware, from Sensors. 2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. Taichung, Taiwan, pp. 445-450.

Wang, X., 2015. Calibration of Big Traffic Data for a Transport Smart City. 15th COTA International Conference of Transportation Professionals: Efficient, Safe, and Green Multimodal Transportation, CICTP 2015. Beijing, China, pp. 387-397.

Xiong, L., et al, 2015. Multi-source macro data process based on the idea of sample=overall in big data: An Applicability Study on Influence Factors to Smart City. 2015 International Conference on Logistics, Informatics and Service Sciences (LISS). Barcelona, Spain,

YIN ChuanTao, et al., 2015. A literature survey on smart cities. Science China Information Sciences, 58(10), pp. 1-18.

Appendix A - Sources of Analysed Articles

Springer Book Chapter

Modeling and Processing for Next-Generation Big-Data Technologies 2015

Resource Management for Big Data Platforms 2016

Conference Proceedings

Database: ACM DL

2015 Proceedings of the 17th International Digital Government Research Conference on Digital Government Research

2015 Proceedings of the ASE Big Data & Social Informatics

2016 Proceedings of the 2Nd International Workshop on Smart

Database: IEEE

2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing

2014 28th International Conference on Advanced Information Networking and Applications Workshops

2015 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)

2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)

2015 IEEE International Congress on Big Data

2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)

2015 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)

2015 International Conference on Computer Science and Mechanical Automation (CSMA)

2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)

2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)

2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)

2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)

2016 IEEE International Conference on Automatica (ICA-ACCA)

2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)

2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

2016 SAI Computing Conference (SAI)

Database: INSPEC

CLOSER 2016. 6th International Conference on Cloud Computing and Services Science Proceedings

Database: SCOPUS

2015 International Conference on Logistics, Informatics and Service Science, LISS 2015

CICTP 2015 - Proceedings of the 15th COTA International Conference of Transportation Professionals

VEHITS 2015 - Proceedings of the 1st International Conference on Vehicle Technology and Intelligent Transport System

Journal Article

Database: SCOPUS

Green Energy and Technology

International Federation of Automatic Control (IFAC) Papers Online

International Journal of Pervasive Computing and Communications

Journal of Cloud Computing

Software - Practice and Experience

Total = 30

Appendix B - Value Chain Classification

Operator	Reference
Data Input	
Gathering	(Takuro Yonezawa, et al., 2016), (Bo Tang, et al., 2015), (Villanueva, et al., 2013), (Cheng, et al., 2015), (Costa C & Santos MY., 2016), (Rathore MM, et al., 2016), (Kemp, G., et al., 2015), (Girtelschmid, S. , et al., 2014)
Selection	(Rathore MM, et al., 2016), (Kemp, G., et al., 2015), (Girtelschmid, S. , et al., 2014)
Processing	(Rathore MM, et al., 2016), (Kemp, G., et al., 2015), (Li Xiong, et al., 2015), (Girtelschmid, S. , et al., 2014)
Transformation	(Rathore MM, et al., 2016), (Kemp, G., et al., 2015), (Girtelschmid, S. , et al., 2014)
Analysis	(Debopriya Ghosh, et al., 2016), (Tang B, et al., 2015), (Vasylyna Horban, 2016), (Ciprian Barbieru & Florin Pop, 2016), (Jara, et al., 2014), (D. Singh, et al., 2016), (Cheng, et al., 2015), (Shuangmei Ma & Zhengli Liang , 2015), (Costa C & Santos MY., 2016), (Rathore MM, et al., 2016), (Psomakelis E, et al., 2016), (Souza A, et al., 2016), (Voichita Iancu, et al., 2016), (Huanan Z, et al., 2015), (Tosi D & Marzorati S., 2016).
Output	(D. Singh, et al., 2016), (Rathore MM, et al., 2016)