Accepted Manuscript

A novel big data analytics framework for smart cities

Ahmed M. Shahat Osman

PII: S0167-739X(17)30744-6
DOI: https://doi.org/10.1016/j.future.2018.06.046
Reference: FUTURE 4308

To appear in: *Future Generation Computer Systems*

Received date : 27 April 2017
Revised date : 18 March 2018
Accepted date : 25 June 2018

Please cite this article as: A.M.S. Osman, A novel big data analytics framework for smart cities, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.06.046

A NOVEL BIG DATA ANALYTICS FRAMEWORK FOR SMART CITIES

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

Abstract

The emergence of smart cities aims at mitigating the challenges raised due to the continuous urbanization development and increasing population density in cities. To face these challenges, governments and decision makers undertake smart city projects targeting sustainable economic growth and better quality of life for both inhabitants and visitors. Information and Communication Technology (ICT) is a key enabling technology for city smartening. However, ICT artifacts and applications yield massive volumes of data known as big data. Extracting insights and hidden correlations from big data is a growing trend in information systems to provide better services to citizens and support the decision making processes. However, to extract valuable insights for developing city level smart information services, the generated datasets from various city domains need to be integrated and analyzed. This process usually referred to as big data analytics or big data value chain. Surveying the literature reveals an increasing interest in harnessing big data analytics applications in general and in the area of smart cities in particular. Yet, comprehensive discussions on the essential characteristics of big data analytics frameworks fitting smart cities requirements are still needed. This paper presents a novel big data analytics framework for smart cities called "Smart City Data Analytics Panel – SCDAP". The design of SCDAP is based on answering the following research questions: what are the characteristics of big data analytics frameworks applied in smart cities in literature and what are the essential design principles that should guide the design of big data analytics frameworks have to serve smart cities purposes? In answering these questions, we adopted a systematic literature review on big data analytics frameworks in smart cities. The proposed framework introduces new functionalities to big data analytics frameworks represented in data model management and aggregation. The value of the proposed framework is discussed in comparison to traditional knowledge discovery approaches.

Keywords:

Analytics framework; Apache Hadoop; Apache Spark; Big data; Smart cities

1. Introduction

The concept of smart cities emerged as a strategy to mitigate the unprecedented challenges of continuous urbanization, increasing population density and at the same time provide better quality of life to the citizens and visitors [1]. A smart city is composed of smart components such as smart buildings, smart farms and smart hospitals, which constitute various city domains where the meaning of the label "smart" has different connotations in each domain [2]. ICT applications and intensive use of digital artifacts such as sensors, actuators and mobiles are essential means for realizing smartness in any of smart city domains [3]. However, "smartening" of various city domains is not enough for a city to be smart, whereas the interrelationship between the underlying city domains should be taken into account to realize city smartness [3, 4]. As such, a smart city is viewed as a whole body of systems or *system of systems*. This integrated view for a smart city implies cross-domain sharing of information [5]. This holistic view for smart city characterizes the meaning of "smart" in the context of "smart city" compared to smartening of particular city domain.

On the other hand, the extensive use of digital technologies in various city domains and the diffusion of digital technologies in people's daily life have boosted human-to-human, human-to-machine, and machine-to-machine interactions which yield massive volumes of data, commonly known as *big data* which is a mixture of complex data characterized by large and fast growing volumes datasets which go beyond the abilities of commonly known data management systems to accommodate. By analyzing these big data volumes, valuable insights and correlations can be extracted [6]. The process of analyzing big data to extract useful information and insights is usually referred to as big data analytics or big data value chain [6], which is considered as one of the key enabling technologies of smart cities [7, 8, 9].

However, big data complexities comprise non-trivial challenges for the processes of big data analytics [3]. Although literature is replete with articles addressing big data analytics frameworks and their applications of in different smart domains, detailed discussions on the characteristics of big data analytics frameworks fitting smart city's requirements are still needed. The lack of this type of articles is the essential motive for this research. The main contribution of this paper is a proposal of a novel big data analytic framework for smart cities. To identify the necessary characteristics of big data analytics frameworks for smart cities, we adopted a systematic literature review approach on big data analytics frameworks in smart cities to answer two basic research questions. RQ1: what are the characteristics of big data analytics frameworks applied in smart cities in literature? And RQ2: what are the essential design principles that should characterize big data analytics frameworks to serve smart cities purposes? To achieve this objective, 30 articles addressing big data analytics frameworks and applications in smart cities are analyzed.

This paper is organized as follows: This section is an introductory section about the subjects and motive for this paper. The second section presents fundamental concepts about big data and smart cities and how the two subjects are related. The scope of the review is defined in the third section. In the fourth section, the 30 articles selected for review are analyzed with respect to the value chain operators and the functional requirements that fit smart cities. Findings are discussed in the fifth section. The sixth section presents the main contribution of this paper, proposal for a novel big data analytics framework for smart cities and its Hadoop-based prototype implementation. In the same section, SCDAP design principles are discussed. Also, the value of SCDAP approach is demonstrated in comparison to traditional knowledge discovery approaches. The seventh section is the conclusion section. Finally, in the eighth section SCDAP architecture limitation is discussed and list of recommended directions for future research is presented. This paper includes three appendices: Appendix A: Details of the search process, Appendix B: Results with respect to big data value chain operators and Appendix C: Results with respect to Functional Requirements. Appendices are available on the following URL: [Appendices]

2. Topic Conceptualization

In this section, fundamental concepts of the two main subjects of this article big data and smart cities are presented to uncover the challenges of harnessing big data analytics in smart cities. Uncovering these challenges help determining, at a high level, essential functional and non-functional requirements that should be considered in the design of big data analytics frameworks for smart city purposes.

2.1 Big data

Big data is a natural crop of the advanced digital artifacts and their applications. Mobiles, sensors and Social Media Networks are examples of modern digital technologies that have permeated our daily lives. Prevalence of these technologies in the everyday life boosted human-to-human, human-to-machine and machine-to-machine interaction into unprecedented levels yielding massive volumes of data known as big data. However, volume of data is not the only characteristic of big data. Big data is commonly characterized by four Vs characteristics: Volume, Velocity, Variety and Veracity (Figure 1) [10].

Figure (1) - Big data 4Vs

Volume, as the name indicates, big data volumes goes far beyond the size of traditional operational databases or data warehouses. Traditional databases usually grow to the order of gigabytes (10^9 byte) or even terabytes (10^{12} byte). Big data volumes are big enough to the extent that new measuring units are required such as Petabyte (10^{15} byte) and Exabyte (10^{18} byte).

Velocity refers to the high rate of data streaming into hosting platforms. For example, how many mouse clicks per second can be captured from Social Media Network applications such as Facebook or LinkedIn? In addition to the high rate of incoming data, velocity raises an important concern on data aging i.e. “for how long these data will be valuable?” In some cases, real-time\online analysis of streaming data is critical. For instance, real-time analysis for video streams captured by traffic surveillance cameras is critical to predict traffic jams and prevent bottlenecks within limited time brackets.

Variety refers to the complexity of big data formats. Big data is mostly composed of semi-structured (e.g. IoT sensed data files); unstructured data (e.g. text data files and images) and stream data (e.g. geospatial data streams) in addition to traditional structured data. It is usually estimated that the ratio between structured data to other data types is by %20 to %80, respectively.

Veracity refers to the trustworthiness of the data. False data will definitely lead to misleading results. Therefore, there is necessity to ensure that data sources are trustworthy and data are correct especially in case of automated decision-making where no human intervention is involved.

However, we expect the adjective *big* will fade over time and the self-evident meaning of *data* will intuitively be extended to include all types of data as mentioned above, i.e. semi-structured, unstructured and stream data in addition classically known structured data.

2.1.1 Big data analytics platforms

Big data analytics refers to the entire processes and tools required for knowledge discovery including data extraction, transformation, loading and analysis; specific tools, techniques, and methods; and how to successfully provide results to decision makers. Although developing big data analytics platforms encounter non-trivial challenges due to the complex nature of big data, big data analytics holds an unprecedented opportunity to shift the traditional methods of information extraction into new dimensions. This opportunity induced researchers and technology providers to develop sophisticated platforms, frameworks and algorithms to compete with the challenging of big data [6, 11, 12].

To deal with the challenging nature of big data, platform scalability represents the logical solution. There are two commonly scaling approaches: vertical and horizontal scaling which are known also as scale up and scale out respectively [6]. Vertical scaling means empowering the processing platform with additional computing power (memory, CPUs…etc.) to accommodate with the incremental volumes of data. This approach involves execution of a single operating system. On the other side, horizontal scaling is a divide-and-conquer approach. 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 execution of multiple instances of different autonomous operating systems running on independent machines, which as such a further complexity (Figure 2).

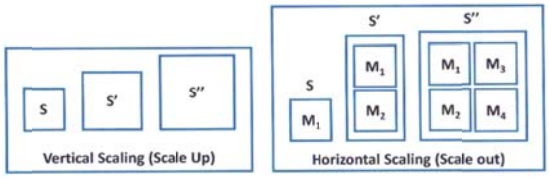

Figure (2) - Vertical vs Horizontal Scaling

Each approach has its advantages and disadvantages. Although vertical scaling shows resilience in platform upgrading, yet, it is restricted by the platform upper limits (e.g. maximum memory, number of CPUs…etc.). On the other side, in horizontal scaling more computing nodes can be added as much as needed. However, horizontal scaling involves maintaining multiple instances of different operating systems that is a complicated challenge. High Performance Computing (HPC) clusters and Apache Hadoop [13] are two examples of vertical and horizontal scaling platforms.

However, the cost of vertical scalability and upper ceiling limitations are major and considerable drawbacks of vertical scaling. These two drawbacks come in favor of horizontal scaling when it comes to smart city projects, considering the dynamics of multi-domain nature of smart cities projects and prospects for future expansion, it is more rational to rely on horizontal scaling rather than vertical scaling. This note interprets why most of the researches and designs of big data analytics frameworks and platforms are built using horizontally scalable platforms such as Apache Hadoop platform.

2.1.2 Big data value chain operators

In [6], authors adopted analogues approach used for Knowledge Discovery in Databases (KDD) model to study the corresponding bottlenecks in case of big data analytics. In the KDD model, analytics is divided to three operators: input, analysis and output (Figure 3). In the case of big data, data input operator ingest stream volumes of noisy incomplete raw data from heterogeneous sources. The processes of input operator have a crucial impact in mitigating the effect of data

voluminous on the overall analysis processes. Picking up relevant data, data cleansing and compression will influence the efficiency of the data analysis execution performance. Despite the importance of the input operator processes, the authors in [6] noted an important observation that the number of research articles and technical reports that focus on data analysis operator is typically more than the number focusing on other operators.

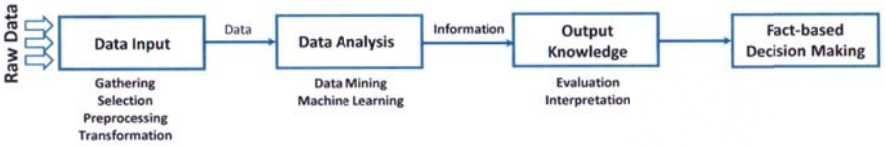

Figure (3) - Knowledge Discovery in Databases

As for the data analysis, the survey in [6] revealed that parallel computing and cloud computing technologies have a strong impact on the research on this area, where most of the big data analytics frameworks and platforms were developed based on these technologies. However, reliance on parallel distributed processing platforms compelled researchers to modify traditional machine learning and data mining algorithms to adapt with the new environment and programming paradigm (e.g. Map and Reduce) [13].

2.2 Smart cities

Although the term smart city seems to be simple and intuitively understandable, there is no globally recognized definition of the term; this is due to different perspectives about the how the label "smart" is loaded in the context of smart cities [1]. There are several surveys in literature addressing the definition of "smart city" from different perspectives [1, 2, 3, 5]. For example, IBM defines a smart city as: "the city that makes optimal use of all the interconnected information available today to better understand and control its operations and optimize the use of limited resources" [5]. Also, YIN ChuanTao, et al [3] defined a smart city as "a system of technological infrastructure that relies on advanced data processing with the goal of making city governance more efficient, citizens' happier, business more prosperous and the environment more sustainable". Authors in [2] analyzed and categorized several definitions of the term "smart city" in literature setting three fundamental factors that make a city smart. These factors are technology (infrastructures of hardware and software), people (creativity, diversity, and education) and institution (governance and policy), given the connection between these factors. However, from these definitions we can draw out three essential characteristics that characterize smart cities as:

- The vital role of ICT as key enabling technology in developing smart cities, where ICT represents the essential backbone for connecting city core systems together infusing data and information between different city systems. Digital artifacts and ICT applications generate raw data that can be blended and analyzed to extract useful information and insights about the city using artificial intelligence and data analysis applications.
- The integral view of a smart city, where the interrelationship between the core systems of the city should be considered. Scholars and relevant organizations define different domains that comprise a smart city, for example, authors in [4] defined six *smart* domains for a city to be smart. In this regard, a smart city is viewed as a whole body of systems (i.e. system of systems) where no system in the city domains functions in isolation [14, 15, 16].
- Sustainability, which means in its broad meaning the ability to continue and grow without significant deterioration. In smart city context, the concept of sustainability applies into different aspects of life in a smart city such as economics and environment.

These characteristics are not exclusive to smart cities, whereas there are many other relative labels in literature such as “digital city”, “intelligent city”, “virtual city” and “ubiquitous city” which address different aspects that characterize modern cities and share some of these characteristics with smart cities with different degrees of deepness.

2.2.1 Modeling of smart cities - an ICT perspective

From an ICT perspective, industrialist and scholars adopted the layered approach to model smart cities. For example, IBM adopted a three-layer model including instrumented layer, interconnected layer and intelligent layer [5]. Similarly, YIN ChuanTao, et al. proposed a four-layer model including: data acquisition and transmission layer; data vitalization layer; common data and service layer; and the applications layer [3]. Regardless of the number of layers, these models project the journey of the big data from its birth in its raw form until valuable information and insights are extracted for the benefit of decision makers and citizens (Figure 4). Although there is no one-to-one correspondence between big data value chain operators (Figure 3) and the layers of smart city models, they share a common objective of projecting the processes of extracting insights and useful information out of raw big data to support smart services and decision-making processes [17].

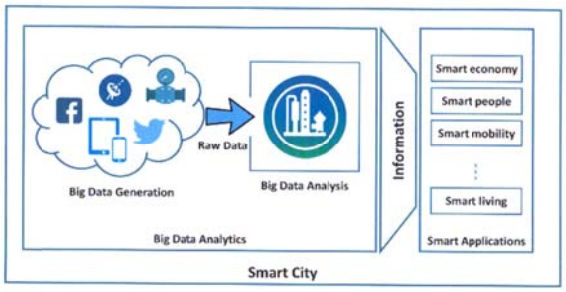

Figure (4) - Big Data in Smart Cities

2.3 Big data analytics in smart cities

The role of big data in developing smart cities is undeniable [7]. There are many applications of big data analytics in different domains of smart cities, such as planning [8], traffic control and transportation [18, 19], crime analysis [20], energy [21] and environment [22]. To design software platforms and architectures fitting smart city purposes, specific non-functional and functional requirements linked to smart city’s nature of data sources and applications should be clearly identified first. For example, the heterogeneity of data sources (e.g. sensed IoT data, Social Media Network and Electronic Medical Records) entails considerable design factors such as data integration, system scalability, privacy and security [9, 23]. Also, the dynamic nature of smart city life necessitates considerable attention for stream data analytics to enable online and real-time services [8, 9]. Additionally, “historical data” or “batch data” analytics is an essential requirement for planning (short and long term) and decision-making purposes in smart cities.

In [17], authors compiled a comprehensive list of eleven essential requirements that should be considered in designing smart cities software architectures. Typically, these requirements could be classified into two categories. Firstly, the functional requirements: object interoperability, real time monitoring, historical data, mobility; service composition and integrated urban management. The second category is the non-functional requirements: sustainability, availability, privacy, social aspects and flexibility/extensibility (scalability).

Similarly, authors in [24] introduced a comprehensive study based on examining 23 smart city platforms to specify the fundamental functional and non-functional requirements to develop a

software platform to enable construction of scalable integrated smart city applications. The study concluded with eight functional requirements: data management, application run-time, Wireless Sensor Network WSN management, data processing, external data access, service management, software engineering tools and definition of city model. This is in addition to eight non-functional requirements: interoperability, scalability, security, privacy, context awareness, adaption, extensibility and configurability.

Authors in [23] specified six characteristics a big data analytics platform should maintain to accommodate with the V challenges of big data, specifically: scalability, I/O performance, fault tolerance, real-time processing, data size and support for iterative tasks.

In summary, we can recap the functional requirements into interoperability, real time analysis, historical data analysis, mobility, iterative processing, data integration and model aggregation. While the non-functional requirements are scalability, security, privacy, context awareness, adaption, extensibility, sustainability, availability, and configurability.

3. The Scope of Literature Review

In order to define the scope of this literature review, authors followed a widely known taxonomy scheme proposed by [25] and adapted by [26] that includes six characteristics for literature review: (1) focus, (2) goal, (3) organization, (4) perspective, (5) audience and (6) coverage.

3.1 Focus

Focus is the central area of interest to the review process. According to [25], it could be research outcomes, research methods, theories, practices or applications. The candidate articles for review are analyzed in two dimensions. The first is the analysis with respect to big data value chain operators presented in subsection 2.1.2 (appendix B). The second one is the analysis with respect to the functional requirements mentioned in section 2.3 (appendix C). The reason for choosing these two dimensions for analysis is that they complement each other. This approach gives a broader picture about the traits of big data analytics frameworks in smart cities, which in return gives a pointer to answer the first research question. In this regard, it is worth mentioning that non-functional will be not be addressed in this article as it is considered general requirements applicable for any analytics system which also required in smart cities with different levels of complexity.

3.2 Goal

Goal refers to the objectives to be fulfilled by the review that could be integration, criticism, central issue. As the objective of this research is to study the traits of the available big data analytics frameworks applied in smart cities and identify the essential functionalities that should characterize big data analytics frameworks to serve smart cities purposes, the goal of this research is to integrate and criticize the finding of the past literature.

3.3 Organization

Organization refers how the literature review is organized. The literature could be organized as chronological order, conceptual order (sharing of the same ideas) or methodological order (sharing of the same methods of work). In this article, literature is organized and discussed in a conceptual order.

3.4 Perspective

Perspective refers to the reviewer's point of view in discussing the literature. Perspective could be: a neutral position (impartial role as an honest "judge") or an espousal position (advocate to certain

idea(s) or methodology). In this research, the author adopted a neutral position search perspective since there is no need to foster a specific position.

3.5 Audience

Audience refers to beneficiaries whom the review addresses (specialized researchers, general researchers, practitioners, policy makers). Since the second research question is identifying essential functionalities that should big data analytics frameworks have to serve smart cities purposes, the audience of this literature review are specialized scholars, practitioners and smart cities planners.

3.6 Coverage

Coverage refers to how the reviewer searches the literature and how he makes decisions about the suitability and quality of documents. According to Cooper, there are four categories of coverage: exhaustive (including the entirety of literature on a topic or at least most of it), exhaustive with selective citation (considering all the relevant sources, but describing only a sample), representative (including only a sample that typifies larger groups of articles), and central (reviewing the literature pivotal to a topic). In this literature review, authors adopted exhaustive with selected citations coverage since it is not realistic to claim exhaustive coverage. Additionally, adoption of representation or central coverage does not serve the objectives of this review.

Table (1) summarizes the choices made by the author, regarding the Cooper's taxonomy about the review scope.

Table (1) Taxonomy of literature review

Characteristics	Categories			
Focus	Research outcomes	Research methods	Theories	Applications
Goal	Integration	Criticism		Central Issue
Organization	Historical	Conceptual		Methodological
Perspective	Neutral representation		Espousal of position	
Audience	Specialized scholars	General scholars	Practitioners\ Policy-makers	General public
Coverage	Exhaustive	Exhaustive with selective citation	Representative	Central or pivotal

4. Literature analysis and synthesis

According to the reference review scheme [26], the search process involves four steps: (1) identifying search databases; (2) search keywords; (3) forward and backward search; and (4) evaluation of articles.

To collect quality scholar articles, six information systems online databases were searched using two search keywords "big data" and "smart city", total of 247 articles were retrieved. After filtration and evaluation, process only 30 articles were shortlisted for analysis. Details of the search process and sources of shortlisted articles are listed in appendix A.

The process of evaluating the articles in terms of addressing the two dimensions of analysis in section 3.1 involved reading and evaluating the article's abstract and conclusion sections to decide which of the focus points (section 3.1) are addressed. If article abstract and conclusion does not

lead into clear decision, the article's full body is reviewed. In the following two subsections, the results of analysis with respect to big data value chain operators and functional requirements are demonstrated. For a more detailed analysis of some proposed end-to-end architectural frameworks are reviewed in details, namely: BASIS [27], SWIFT [28] and RADICAL [16] are reviewed in details in subsections 4.3, 4.4 and 4.5 respectively. Points of strength and weakness of each architectural framework are listed after each review.

4.1 Analysis with respect to value chain operators

Results of the evaluation process with respect to big data value chain operators are shown in the table presented in appendix B. From these, we could recognize that most of the analyzed articles focused on data gathering (21 out of 30). This ratio reflects the high interest of researchers in finding efficient solutions for data gathering from different sources, while relatively, less number of researches addressed the rest of data input functionalities (selection, preparation, transformation) although its significant impact on the efficiency of the analysis processes. In the following three subsections, we will review how each of these operators is addressed.

4.1.1 Data Input

The central challenge in this operator is related to the ability to acquire timely raw data about city events from a large number of heterogeneous sources. Also, in through this operator data are prepared to the following operator for analysis, where data are either analyzed online for real time (or near real time analysis) or stored for later analysis. In this context, researches have been directed towards dealing with the technologies enabling city smartness such as IoT, Smart Sensor Networks, Social Networks and mobile applications. Researchers adopted several techniques to deal with this challenge such as developing an autonomous middleware layer with unified data access through WEB API services e.g. [27, 29, 30]. Authors in [31] introduced a location-aware architecture using the merge between IoT and Cloud technologies (referred to as FOG computing) where local computations and analytics can be performed on the edge of network where machine learning and other artificial intelligence techniques are used in. This approach supports quick response at neighborhood-wide providing high computing and network traffic performance. Relatively little number of articles addressed other data input functionalities (e.g. selection, filtering, transformations) in addition to data gathering [32, 19].

4.1.2 Analysis

There are two main types of data analytics in general, online and historical. The first one serves real-time or near real-time applications while the second one is appropriate for applications that can afford late results. In the context of smart cities, both types of analysis is required. For example, online analytics is required traffic surveillance while historical data analytics is required in the initial phases of city planning in addition to online analytics. Surveyed articles have addressed many areas related to smart cities domains such as traffic control [18, 15, 19], pollution monitoring [22], crime analysis [20] and planning [33]. To mitigate the V challenges of big data, Hadoop and Cloud technology are reliable platforms to host and process data [34, 32, 35, 36].

4.1.3 Output

Within the scope of reviewed articles, it is noted that output results are presented to the end user in the form of graphs and charts that serve the results of the addressed domain of applications.

4.2 Analysis with respect to functional requirement

Results of the evaluation process with respect to functional requirements are shown in the table presented in appendix C. Results in appendix C show that number of articles targeted real-time analysis (25 out of 30) is three times the number of articles targeted historical data analysis (8 out of 30) while only (8 out of 30) addressed the two types of analysis. In addition, only nine articles address integration of data from multiple sources. In the following subsections, we will review how each of these functional requirements is addressed.

4.2.1 Interoperability

In a smart city environment, there are various types of data generating devices. For example, sensors from multiple vendors, systems implemented with different languages and standards (refer to subsection 4.1.1). Interoperability challenge is to make all these devices operate in integrated software platforms. In this context, researches adopted layered design where separate layer is dedicated to interact with our world through standard interfacing gateways [27, 29, 30].

4.2.2 Real-time and Historical data Analysis

Choice between real-time and\or historical data analysis depends on the nature of the target analysis-based applications. As mentioned in section 4.1.2, surveyed articles have addressed the two types of analysis. Traffic control [18, 15, 19] and pollution monitoring [22] are examples of real-time analysis. Crime analysis [20] and planning [33] are examples of historical data analysis.

4.2.3 Mobility

Mobility is one of the key characteristics of smart cities [4]. Linking of traffic, communication and analytics is becoming increasingly important. Transportation infrastructures are pushed to their limits demanding for smart and adaptive means of transporting and routing policies to optimize existing systems grows. In return, mobility requires data analysis systems to be able to deal with mobile data sources. In [37, 38] authors introduced a real-time mobility patterns detection system able to describe how people move around Point of Interests (POI). Policy makers and Journey Planners to provide final users with accurate travel plans can exploit these mobility patterns.

4.2.4 Iterative processing

The demand for iterative processing in a big data analytics system emerged because the supported programming paradigm supported by Hadoop, MapReduce, in primarily sequential data processing model which is not efficient for fast data processing requirements (e.g. real time analysis). That was the reason for introducing Spark for in-memory iterative data processing. In [39, 32, 40] authors introduced various design scheme based on Spark over Hadoop for real-time data analytics. The test results show significant performance efficiency of big data processing platform using distributed architecture with iterative processing. In this context, it is worth mentioning that iterative processing is not an alternative for batch processing. However, both processing schemes are required in smart cities applications.

4.2.5 Data Integration

The challenge of data integration involves combining data from several disparate sources, which are generated and\or, stored using various technologies to provide a unified view of the data. Data integration becomes increasingly important in case of smart cities being a complex system of systems. In [20] authors introduced an effective real time crime analysis system based on integrating data from different data sources. The systems applies machine learning approach for

developing ‘tactical information profiles’ timely manner to support and optimize investigators decision making actions. Also, in [32] proposed a smart city system based on data generated from smart home sensors, vehicular networking, weather and water sensors, smart parking sensors, surveillance objects, etc. The system implementation includes various data preparation steps starting from data generation and collecting, aggregating, filtration, classification, preprocessing.

4.3 BASIS

BASIS is a three-layer big data architecture for smart cities [27]. The architecture of BASIS is built upon a fundamental design principle of strict separation between abstraction layers. The three layers are the conceptual layer, the technological layer and the infrastructure layer.

The conceptual layer encapsulates the internal and external functionalities of BASIS including data capturing, integration, storage and APIs for web streams and data analysis. The technological layer is realized using Hadoop platform. The infrastructural layer addresses the physical hardware design aspects to interact with external world. In other words, it is the interfacing part of the technological layer to interact with external requests. Also, the architectural design of BASIS considered also the following design principles:

- Makes available Open Data, facilitating the appearance of new services developed by Smart Cities governments, organizations or citizens,
- Multiple abstraction layers, from the most conceptual to the most technological,
- Distributed data storage and processing,
- Data security, privacy and thrust concerns,
- Incorporates ways of managing data’s lifecycle, using inherent concepts about data partitioning,
- Establishes cooperation strategies between entities involved in the development of smart city services,
- Service-oriented and client-independent,
- Use of open source technologies, except in cases where that choice is not cost effective.

BASIS was validated using real case study to find the delay profile in flights in several cities at the USA. The profile was extracted using K-means algorithm running in multiple clusters (13 cluster) using dataset containing 3.5 million records about flight performance. Table (2.a) shows the strengths and weaknesses of BASIS architecture.

Table (2.a) Strengths and weaknesses of BASIS architecture

Strengths	Weaknesses
- Adoption of layered design principle (three separate layers), - Domain independent design, - Support of both batch and stream data analytics, - Service-oriented analytics (analytics as-a-service), - Data capturing and data analysis services are provided through APIs, - Implementation using Hadoop, a commonly	- Distinction between the technological and infrastructure layers are not clearly demonstrated. As the infrastructure layer is concerned the physical hardware design aspects to interact with external world which could be considered as an extension functionality of the technological layer, - Although BASIS is comprehensive big data analytics architecture, it did not show the particular design considerations pertaining to

Strengths	Weaknesses
known horizontally scalable platform.	smart cities, - Relevance of the validation case (flight delay profile) to smart cities is questionable, it would much better if the authors considered extracting insights related to one or more smart city domains (transportation, energy, traffic…etc.). Also, the validation case dealt with batch data analysis, - Batch data analysis for offline applications is not clearly compared with stream data analysis.

4.4 SWIFT

SWIFT stands for: "Smart wireless sensor network (WSN)-based Infrastructural Framework for smart Transactions". It is a three-tiered architectural framework that supports integration of heterogeneous devices. The base layer is Smart Wireless Sensor Network (S-WSN). The second layer is the "Smart Wireless-based Pervasive Edifice" (SWIPE) which resides on the S-WSN layer. The third layer is the "Smart Decision & Control Enabler" (SDCE).

The S-WSN layer acts as the sensory probes of SWIFT where several hundreds of physically dispersed wireless sensor nodes that sense a phenomenon of interest and report the data for further analysis. Sensors are grouped in clusters and headed by a Smart Cluster Heads (SCHs) deployed at various locations in the city. SCHs collect and aggregate data from nearby sensor nodes. In addition to sensed data, the nodes transmit their node identification and battery status to the SCHs. The SCHs are capable of taking low-end, but important decisions like raising an alarm, generating emergency actuation signals etc. After aggregating and processing the data, the SCHs transmit data to nearby Smart Fusion Nodes (SFN) in the following layer SWIPE.

SWIPE is the heart of SWIFT architecture, which comprises several Smart Fusion Nodes (SFN) that act as the edifice for SWIFT architecture. SFNs act as data classifiers and perform data fusion to draw meaningful interpretation of the sensed data for query processing and other related services. They collect data from S-WSN layer to facilitate ubiquitous computing. Smart Decision & Control Enabler (SDCE) is the core layer that provides a host of services (cloud) to all smart objects in the city based on data provided by SWIPE. Table (2.b) shows the strengths and weaknesses of SWIFT architecture.

Table (2.b) Strengths and weaknesses of SWIFT architecture

Strengths	Weaknesses
- Relevance to smart city domains is emphasized through the proposed application domain (traffic monitoring and control), - Scalability to accommodate with heterogeneous large number of sensors, - Adoption of the hierarchical uncomplicated layered approach, - Ability to build location-aware data,	- Only sensor-captured data is considered, data from other data sources are not considered, - Batch data analysis for offline applications are not applicable, - Performance of SWIFT framework is not tested.

Strengths	Weaknesses
- Online or near online application geared architecture, - Although SWIFT is proposed for traffic monitoring and control domain, it can be adapted for other online analytics domains.	

4.5 RADICAL

"RApid Deployment on Intelligent Cities And Living", RADICAL is a Service Oriented Architecture -based (SOA) platform that enables retrieval and analysis of IoT (Internet of Things) sensed and Social Networks (SN) data to offer variety of added-value services for smart cities.

IoT data are pushed into RADICAL repository (MySQL database) through the respective Application Programming Interface (API). Device related data are saved in the form of observations and measurements. Observations correspond to general reported IoT events while measurements to more specific metrics included in an observation (e.g. CO_2 measurements). On the other side, SN data are accessed in real time from underlying SN adaptors by communicating with the respective Networks' APIs.

On top of the main platform, RADICAL provides a set of application management tools that allow end users to make better use of the platform, such as configuring the registered IoT devices or extracting general activity statistics. Table (2.c) shows the strengths and weaknesses of RADICAL architecture.

Table (2.c) Strengths and weaknesses of RADICAL architecture

Strengths	Weaknesses
- Combination of retrieval and analysis of data sourced from heterogeneous sources (IoT and SN). As for smart cities, this is a key feature for considering the interrelationship of between the city underlying domains, - Interoperability to accommodate with heterogeneous large number of sensors (IoT and SN), - Support for of SOA to provide smart city services, - Domain independent architecture.	- Reliance on MySQL database as a main repository is a restrictive factor when it comes to big data considerations, - Inability to retain analysis results for future analyses or reference.

5. Findings and discussion

In this section, we analyze and discuss the findings of the reviews in section 4. Findings of these reviews are compiled from three different perspectives: design principles, enabling technologies and application domains to identify the characteristics of the big data analytics frameworks in smart cities hence answer the first research questions.

Design principles

- Adoption of layered design principle. Where each layer is designated a specific functionality with well-defined interfacing with preceding and following layers,
- Standardization of the data acquisition from external world (interoperability and mobility). Similarly and output data access through unified access (e.g. API or standard message formats) to enable fact-based applications [29],
- Enabling of both real-time and historical data analytics are necessary,
- Support of both iterative and sequential data processing to accommodate with real-time and historical data analytics requirements,
- Scalability to adapt with potentials of increasing number of data capturing devices and the volumes of data generated from these devices.

Enabling technologies

In this subsection, we review key enabling technologies for harnessing big data analytics in smart cities. However, we will try to focus only on the aspects pertained specifically to smart city's applications to avoid repeating what is already addressed in literature.

- Horizontally scalable platforms: Hadoop seems to be feasible to accommodate with scalability requirements of big data analytics frameworks in smart cities. They support cost effective resilient platforms able to host and process increasing volumes of big data [27, 34].
- Fog computing (also known as Edge computing) a technology which provides computing and storage between end devices and cloud data centers are used for data preprocessing to detect anomalous and hazardous events [31].
- Cloud systems provides scalable, robust and highly available hosting environment for both data storage and computation fitting with the volume characteristic of big data and complex dynamic nature of smart cities [19, 40].
- Data mining (DM) and machine learning (ML): DM and ML are crucial technologies for data-centric smart cities. DM is used to extract hidden, unknown and potentially valuable information from big data. DM is a broad field that includes many algorithms and techniques from statistics, ML and information theory to extract information from data. Additionally, ML is an application of artificial intelligence (AI) that provides computers the ability to learn and improve from experience from data without being programmed. ML algorithms are usually classified into two categories, namely supervised and unsupervised learning. Algorithms belong to the first category learn from past data with labeled examples (training dataset). After sufficient training iterations, the system will be able to predict future events. In contrast, unsupervised machine learning algorithms are used when the data used for training are not labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data.

  There are many applications of DM and ML for smart cities in literature **[41]**. The main challenge in this domain is finding appropriate data sets from massive city data to fit with data mining requirements. Solving the problem of dataset selection and data combination is very important for smart city data mining.
- In-memory databases for high performance analytics: relational databases systems (RDBMS) were not designed to cope with big data scalability and agility challenges, nor were they built to

take advantage of the commodity storage and processing power available today. In contrary, not only SQL databases (NoSQL) are designed to fit with big data requirements from that aspect. There are four basic types of NoSQL databases: key-value, document-based, column-based and graph-based. Apache Hadoop suite includes HBase. HBase is a NoSQL database that runs on top of HDFS which supports both key-value and column-based store. HBase provides real-time read/write access to those large datasets. The main advantage of HBase is the high linear scalability of handling huge data sets with large number of rows and columns. In addition, HBase can easily combines data sources that use a wide variety of different structures and schemas.

- Data visualization: data visualization is the presentation of data and\or information in a friendly pictorial format. This technique of presenting data helps users and decision makers to quickly identify interesting patterns from data visually. When it comes to big data, visualization becomes more challenging because of its V characteristics. Applications of big data in smart cities, especially those targeting decision makers, add more complexity for data visualization since data is sourced from different domains [42]. However, the emergence of new technologies such as augmented reality (AR), virtual reality (VR), mixed reality (MR) and Google Maps are paving the way for the development of practical efficient smart city applications. Traffic management [18] and points of interest (POI) [38] are examples data visualization applications in smart cities.

Application domains

Within the scope of reviewed articles, application domains are approached in two ways. Some research articles proposed analytics solutions driven by specific application domain (e.g. environment, traffic…etc.) while other articles were driven by finding technical solutions for analysis functionalities (e.g. interoperability, mobility…etc.). In the latter case, domain of application is used as a proof of concept to the proposed idea.

Based on article analysis and review in the previous section and the above findings, we can realize that factors influencing the design of the proposed frameworks are driven by either finding solutions for specific smart city domain(s) or finding solutions for some technical challenges. Another important observation, the basic design concept of the reviewed architectures lack addressing the holistic view of smart cities being complex system of systems where the interrelationship between different smart city domains should be taken into consideration. Consideration of the interrelationship between different smart city domains necessitates comprehensive analytics based on datasets generated from different domains which might be reiterated for a wider scope of extracted information and insights (*e.g.* during the smart city planning phase were the strategic objectives are set) [17, 43, 44]. There are two approaches to meet this demanding requirement. The first one is to persist datasets generated from different domains, integrate selected datasets for analysis then start the whole analytics processes. The second approach is ability to persist the resultant information form the analytics processes for later analysis usages, where the extracted data models can be retrieved and combined together for comprehensive analytics. Of course, this approach is conditioned by the applicably of merging these models with each other. However, the second approach will optimize execution time compared to the first approach in the sense that no need to re-analyze the same dataset many times. Of course, the second approach does not mean dispensing with the first approach completely. However, both approaches are complementary to each other to optimize execution time.

In the context of big data analytics for smart cities, we recommend inclusion of two significant design aspects, namely: *model management* and *model aggregation*. The first one enables persisting extracted information an insights (data models) for future analysis without reiterating the

analysis process. In other words, it will help building a some sort of library for the extracted models. The second one will enable model ensemble for more comprehensive multi-domain analytics.

6. Proposal for a Novel Framework

In light of the findings and discussion in section 5 and understanding the holistic view of a smart city, we propose a "Smart City Data Analytics Panel (SCDAP)". The schematic architecture of SCDAP is shown in figure (5). SCDAP is a 3-layer architecture including: a) Platform layer, b) Security layer and c) Data processing layer.

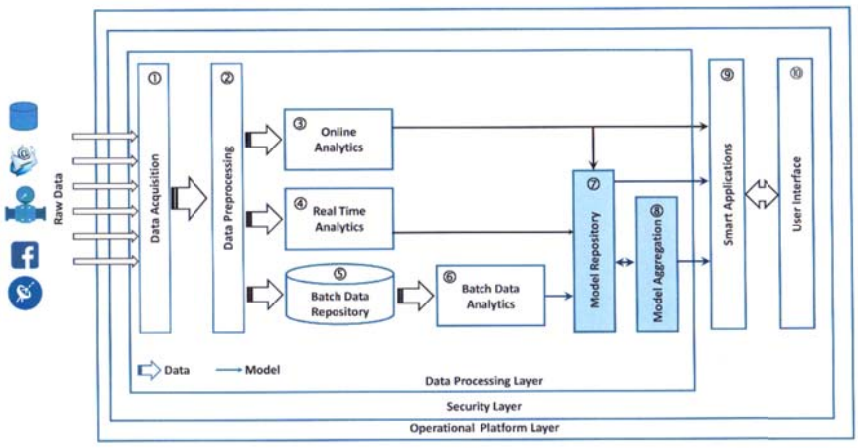

Figure (5) - Smart City Data Analytics Panel SCDAP Architecture

- Platform Layer:

  The platform layer is a horizontally scalable platform including hardware clusters, operating systems and communication protocols. In horizontally scalable platforms, additional computing nodes can be added as needed.

- Security Layer:

  Although a complete vision for the functionality of the security layer will be clearer during physical implementation of SCDAP, 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.

- Data Processing Layer:

  This is the core data processing engine that provides all the data processing functionalities from data acquisition to knowledge extraction. This layer supports both online and batch data processing for real-time and historical data analytics respectively. Additionally, this layer provides two important functionalities that distinguish SCDAP namely: *model manager* and *model aggregation* where extracted data models can be managed (*i.e.* persisted, retrieved and deleted) and aggregated, respectively. Broadly, components composing this layer are summarized as follows:

  ① **Data acquisition**: The main component for data capturing from the external world. It is characterized by scalability, interoperability. Scalability is the ability to accommodate

dynamically with the increasing number of data generating artifacts. Interoperability is the ability to interact with heterogeneous types of generating artifacts (Wireless Sensors Networks (WSN), IoT, Social Media Networks (SMN), etc.).

② **Data preprocessing**: Provides input data cleansing, transformation and integration functionalities. This component is responsible for transforming input data into analysis-ready format(s).

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

④ **Real time analytics**: Ability of performing stream data processing for real time applications. Real time applications are applications that function 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 afford latency.

⑦ **Model management**: Extracted data model management system, where resultant data analysis models can be persisted, retrieved (or deleted) with relevant metadata for future inquiries or reuse. Additionally, static (semi-static data) of the city is persisted in this repository.

⑧ **Model aggregation**: Extracted data model ensemble functionality for higher level and more complicated analytics and inquiries.

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

⑩ **User interface**: End user interface to provide efficient flexible tools allowing access, reporting and ad hoc inquiries for persisted and\or aggregated models.

6.1 Hadoop Based Prototype Implementation

In this section, we discuss the implementation of SCDAP using Apache Hadoop and Spark suites. Since Hadoop is a popular horizontally scalable platform for storing, managing and processing big data it fits smart city's big data analytic applications. It constitutes the base platform in realizing SCDAP. For this reason, we opted to introduce the initial structure of SCDAP using Hadoop and Spark stacks. A high-level schematic diagram for the functional architecture of SCDAP is shown in figure 6.a. The architecture is composed of eight components. The function of each component is described hereafter:

① **Resource Management Layer (YARN)**: YARN (Yet, Another Resource Negotiator) is Hadoop 2.0's resource management framework. YARN isolates resource management and scheduling from the data processing components. Also, it enables Hadoop to support diverse of processing approaches and a broader range of applications. For example, YARN enables Hadoop to run interactive queries, real-time applications and batch jobs simultaneously on one shared dataset.

② **Data Storage Layer:** data storage layer is the data stowing layer. In Hadoop platform, available types of data storage are:

HDFS (Hadoop Distributed File System): It is the native Hadoop data management system. It is a high performance scalable, distributed, fault-tolerant and reliable data storage. HDFS is designed to span large clusters of commodity servers and manage large volumes data files. HDFS allows file creation, write once, read many and remove operations. It does not allow update operations.

NoSQL (Not Only SQL) Databases: The demand for NoSQL databases technology evolved to handle huge volumes of semi-structured and unstructured data properly at which traditional relational databases are not designed for these types of data. NoSQL databases have no declarative query language and no predefined schema. There are four types of NoSQL

databases. Each of these categories has its own specific attributes and limitations (Key-value, Column-oriented, Graph, Document oriented). There is no single solution that is better than all others are; however, there are some databases that are better to solve specific problems. HBase and Kudu are examples of column-oriented database where every column (or family of columns) is treated individually. Column-oriented databases are appropriate for aggregation queries, data warehouses and business intelligence application. However, Apache Kudu provides some additional features similar to traditional relational databases.

③ **Data Integration Layer:** This is the first layer where data is collected for the external world to the proposed framework. The proposed design of this layer enables the ability to collect stream data (e.g. twitter and sensed data) and batch data from traditional enterprise databases and data warehouses. Apache Hadoop ecosystem provides efficient application to serve data integration purposes:

Apache Flume: Flume is an efficient distributed service for collecting, aggregating, and moving large volumes of log data for streaming into Hadoop. Flume's main use-case is to ingest data into Hadoop.

Apache Kafka: Kafka is a distributed publish-subscribe message streaming platform. It is used to build real-time streaming data pipelines that reliably get data between systems or applications (in which Hadoop is one of them). Also, it allows processing of data stream records as they occur.

Apache Sqoop: Sqoop is a tool designed for transferring bulk data between relational databases (e.g. MySQL and Oracle) into the Hadoop Distributed File System (HDFS) and vice versa.

④ **Data Analytics Layer:** This is the main layer for Analyzing incoming data to extract knowledge and insights. This layer consists of two components: Stream data analyzer (e.g. machine learning) and batch data analyzer (e.g. data mining). Stream data analyzer capture data from the data integration layer (stream data). Apache Hadoop ecosystem provides applications appropriate for this objective, namely Apache Storm and Apache Spark. Batch data analyze access data persisted in any data storage system (e.g. HDFS, HBase…etc.).

Apache Storm: Storm is a powerful distributed real time computation framework for processing streaming data and can perform micro-batch processing. It can process huge number of records per second per node on a cluster of modest size. Storm supports many programming languages such as Java, Scala and Python.

Apache Spark: Spark is a batch in-memory computing framework and can perform micro-batch procession via Spark-streaming. Spark is an efficient alternative to Hadoop MapReduce programming framework as it offers faster performance given memory requirements considerations. It provides APIs and supports Java development, Scala and Python programming languages. Also, Spark supports MLlib (Machine Learning Library), an algorithm library about big data machine learning algorithms such as classification, clustering and regression.

Extracted models are dispatched to the model management layer to be persisted for future reference or transferred directly to the smart applications layer for real-time and online applications.

⑤ **Model Management Layer:** This layer provides two modules:

Model management: This module is responsible for managing extracted models (stowing, discarding) with relevant metadata for future inquiries and applications.

Aggregation manager: This module provides the ability to ensemble persisted models for higher level inquiries. This functionality enables more complicated smart city cross-domain analytics.

⑥ **Smart Applications Layer:** This is the main layer for developing smart applications based on the extracted analytical models. This layer can provide applications based on either real time based analytics or batch analytics.

⑦ **Security Layer:** Apache Hadoop provides Sentry as a system for enforcing role based authorization to data and metadata stored on a Hadoop clusters.

⑧ **End User Presentation Layer:** This is the end user interface where extracted knowledge and models can visualized with simple, interactive and effective GUI interfaces. The more simplicity and interactivity of this layer is, the more complexity of the underneath application.

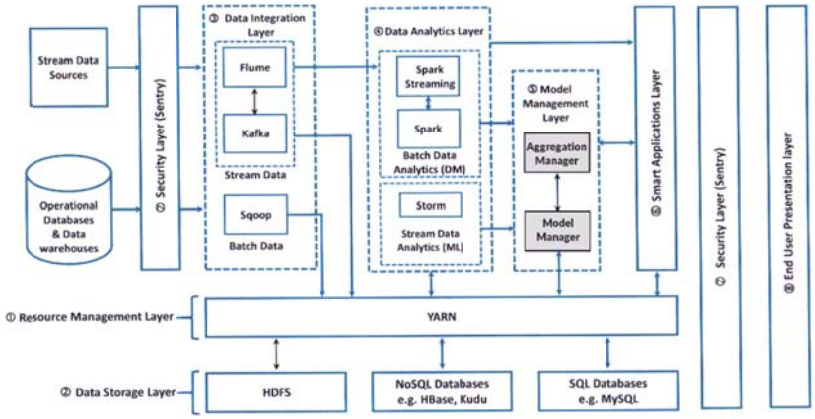

Figure (6.a) - SCDAP Prototype Implementation using Apache Hadoop

It is worth mentioning that there are many efficient open source and commercial software platforms suitable for realizing SCDAP. Anaconda (https://www.anaconda.com) and Google TensorFlow (https://www.tensorflow.org) are examples of these open source platforms. SAS (https://www.sas.com/en_us/home.html) and RapidMiner (https://rapidminer.com/) are examples of commercial software platforms. These platforms offer rich easy-to-use components that facilitate the development of the required functionalities that dramatically minimize the development time without (or with minimal) programming.

Anaconda is a distribution for Python and R languages powered by hundreds of data mining, data visualization and numerical analysis libraries (*e.g.* Python NumPy and pandas). TensorFlow is an open source Python library released in 2015 by Google to enable developers to design, build and train flexible deep learning models easily. Similarly, SAS and RapidMiner include large powerful libraries for data mining, machine learning and visualization. These platforms adopt the approach of data flow graphs where nodes in this graphs represent mathematical operations and edges represent data that is communicated from one node to another.

Since SCDAP is in its initial phase, more iterations are required to study the design details of each component. For example, inclusion of the above software packages (*e.g.* TensorFlow and\or Anaconda) in the design of SCDAP will affect the ingredients of the data analytics layer (component④) as shown in figure 6.b. Although these software packages can be integrated with Hadoop platform to reduce the development cycle of SCDAP, yet it is technically challenging task that is out of scope of this paper.

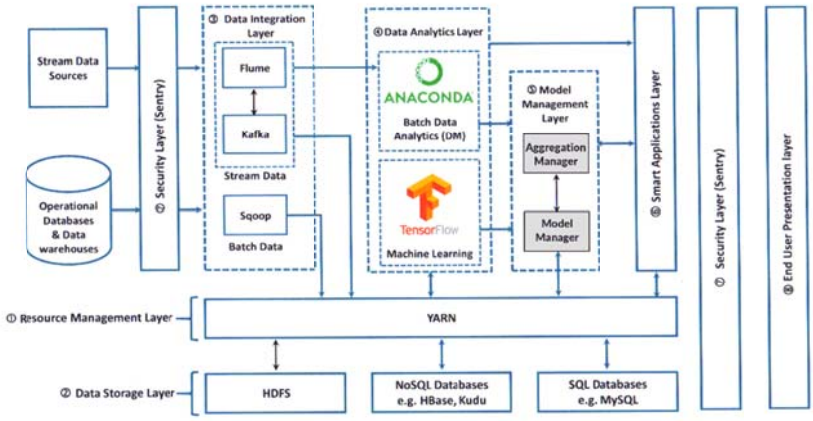

Figure (6.b) - SCDAP Prototype Implementation using Hadoop, TF and Anaconda

6.2 SCDAP Design Principles

In this subsection, we present the design principles for SCDAP grounded in the findings and conclusions reached in findings and discussion section and the main objective of SCDAP as a big data analytics panel providing powerful analytical functionalities is providing to stakeholders including model management and aggregation. The principles of SCDAP design fall into category of *materiality oriented design principles* (category 2) [45]. This category prescribe how an artifact should be built, comprise and state the system features. There are a number of design science research (DSR) methodologies in literature. However, we opted to adopt the six-step design science research methodology (DSRM) proposed by [46]. In summary, we can formulate the following six design principles of SCDAP as follows:

(DP1) Principle of layered design approach
The layered design approach is a most common and solid architecture pattern. Each layer in the architecture forms an abstraction level in the overall system functionality where each layer has a specific role and responsibility. For example, a security layer would be responsible for handling all user accessibility independently of the hosting platform layer.

(DP2) Principle of standardized data acquisition\access
Standardization of data acquisition and access is a key design principle in leveraging big data analytics in the context of smart cities. As a smart city consists of many smart components each has its own standards for the generated data, the proposed architecture should be able to deal with this various standards (*interoperability*). Also, the data and information generated from SCDAP should be accessed through standard API.

(DP3) Principle of enabling of both real-time and historical data analytics
Both real-time and historical data analytics are essential functionalities to satisfy various the analytical requirements of different smart city domains.

(DP4) Principle of enabling of both iterative and sequential data processing
Iterative and sequential data processing are complementary to each other whereas there is no preference for one technique over the other. Choice between the two techniques depends on the size of data and nature of the functional requirements.

(DP5) Principle of model management
This principle is pertained to the system ability to manage resultant data analysis models for future inquiries and analytics. This principle is supported with the ability to manage city reference data (*e.g.* maps).

(DP6) Principle of model aggregation
This principle enables model ensemble for more comprehensive multi-domain analytics.

6.3 Comparison to traditional approaches

In this section, the value of SCDAP approach is demonstrated in comparison to traditional knowledge discovery and management approaches. The commonly known approach in this context is the six-step Cross-industry standard process for data mining (CRISP-DM) [47]. This approach involves specific successive and iterative steps until an optimal prediction model is reached satisfying the business objectives (Figure 7).

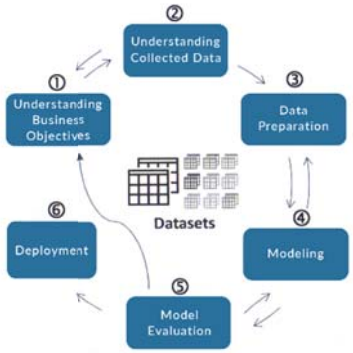

Figure (7) - Cross-industry standard process for data mining (CRISP-DM)

As in a typical data analytics case, the first step is understanding the business objectives of the analytics case. The following step is understanding the collected data *i.e.* the coding scheme, normalized data and missing data. The following three steps: data preparation, modeling and model evaluation are reiterated several times until an acceptable model satisfying the case objectives is reached (Figure 8), where various modeling techniques are applicable, *e.g.* regression models, decision tree, machine learning. In each iteration, model parameters are calibrated and various attributes are selected to generate temporary prediction model as a candidate model for selection as an optimum model. If we assume that the optimal model is generated after n iteration and the average time required to generate each temporary model is t, then the total required time T to reach the optimal model is calculated from:

$$T = n \times t \qquad (1)$$

In some cases, model ensemble techniques are applied to improve the accuracy of predictive model where several modeling techniques are applied before the final ensemble algorithm is applied. Assuming that the total number of the generated candidate prediction models is p then the total required time T to reach the optimal model in equation (1) is calculated from:

$$T = n \times t \times p \qquad (2)$$

In this context, it is worth noting that the process of reaching an optimal model involves generation of several temporary predictive models for comparisons and evaluations at which the same temporary models might be regenerated many times in different iterations. This process will be

significantly challenging and demanding requirement in case of complex prediction models involving large number of attributes and massive volumes of data.

To avoid regeneration of temporary models unnecessarily, SCDAP provides two powerful functions to manage and reuse of the temporary generated models. These functions are *model manager* and *model aggregation manager*. The first function enables storing, retrieval and deletion of temporary models. This functionality allows better controls on generating temporary models unnecessarily where stored model can be retrieved for reuse whenever need. The other function *model aggregation manager* enables applying model ensemble techniques using stored temporary model (Figure 8). In this case, if we assume that number of reused prediction models is p_r then the total required time to reach the optimal model is reduced by $(t \times p_r)$, then T in equation (2) will modified to:

$$T = n \times t \times p - t \times p_r \qquad (3)$$

Or

$$T = t\,(\,n \times p - p_r\,) \qquad (4)$$

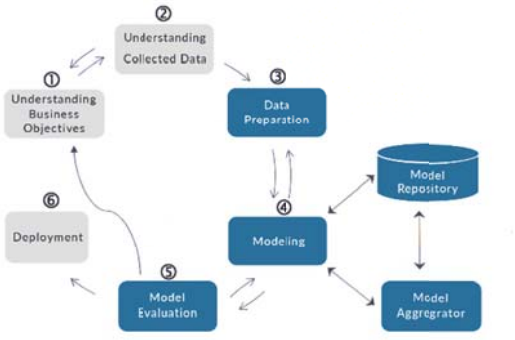

Figure (8) - SCDAP model manager and aggregation functionalities

These two functions maximize the usability of temporally generated models and control the model generation processes to the needed cases only. From a broader perspective within the context of smart cities, over time accumulation of analytical models in addition to it descriptive meta data enables creation of a *library of models* that can be referenced for any future urban studies.

7. Conclusions

Smart City is an emerging concept aiming at mitigating the challenges raised due to the continuous urbanization development and increasing population density in cities where ICT plays a crucial role for city smartening. Big data, as natural crop of extensive use of ICT artifacts and applications in smart cities, have gained increasing attention from academics, industrials and governments for extracting hidden insights and correlations. Although literature survey reveals the increasing interest in big data analytics, comprehensive discussions on the essential characteristics of big data analytics frameworks fitting smart cities requirements are still in demand.

In this paper, we reviewed 30 papers addressing big data analytics in smart cities from two perspectives, namely: big data analytics value chain and functional requirements. The review process aimed at answering two research questions: what are the characteristics of big data analytics frameworks applied in smart cities in literature? And, what are the essential design principles that should guide the design of big data analytics frameworks have to serve smart cities purposes? Based on the answers of these two questions, we proposed a novel framework titled "Smart City Data Analytics Panel – SCDAP"

The answer of the first research question showed that adoption of layered design approach, standardized data acquisition\access, enabling of both real-time and historical data analytics, support of both iterative and sequential data processing and scalability are the commonly followed design principles. In addition, horizontally scalable platforms, cloud systems, data mining, machine learning, in-memory databases and data visualization are the enabling technology. On the other side, the answer of the second research question showed the necessity of including two additional functions to the analytics frameworks for smart cities, namely: model management and model aggregation, that is the main contribution of this paper. The proposed framework SCDAP features the compiled functionalities of the surveyed architectures and suggested ones on the answers of second research question. Additionally, a schematic architectural design for SCDAP and it implementation using Hadoop platform were provided. Finally, the value of SCDAP approach is discussed through comparison with traditional knowledge extraction approaches.

8. SCDAP Limitation and Future Research Directions

Although the proposed framework SCDAP presents new functionalities to big data analytics frameworks for smart city applications, the main feature of this architecture is limited to Apache Hadoop suite as an underlying data storage and management layer. Separation between SCDAP functionalities and the underlying data storage and management layer will add enhance the generality of SCDAP and its ability to deal with many other platforms. However, recommended future research directions involve the following points:

- Develop efficient model persistence, retrieval and ensemble algorithms.
- Develop efficient, powerful and friendly end-used interfaces involving multi-model visualization and OLAP style tools.
- Enable export/import of smart city analytical data models functionality to/from other analytical frameworks (e.g. knowledge exchange).
- Defining smart city performance measures (Key Performance Indicators - KPIs) in which analytics frameworks can be evaluated.

REFERENCES

[1] Hafedh Chourabi, J. Ramon Gil-Garcia, Theresa A. pardo, Taewoo Nam, Sehl Mellouli, Hans Jochen Scholl, Shawn Walker and Karine Nahon, "Understanding Smart Cities: An integrative framework," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, Hawaii, 2012.

[2] T. a. T. A. P. Nam, "Conceptualizing smart city with dimensions of technology, people, and institutions.," in *Proceedings of the 12th Annual International Digital Government Research*

Conference: Digital Government Innovation in Challenging Time, ACM., Maryland, USA, 2011.

[3] YIN ChuanTao, XIONG Zhang, CHEN Hui, WANG JingYuan, COOPER Daven and DAVID Bertrand, "A literature survey on smart cities," *Science China Information Sciences,* vol. 58, no. 10, pp. 1-18, 2015.

[4] R. a. H. G. Giffinger, "Smart cities ranking: an effective instrument for the positioning of the cities?," *ACE: Architecture, City and Environment,* vol. 4, no. 12, pp. 7-25, 2010.

[5] Michael Kehoe, Michael Cosgrove, Steven De gennaro, Colin Harrison, Wim harthoorn, John Hogan, John Meegan, Pam Nesbitt and Christina Peters, Smart Cities Series: A Foundation for Understanding IBM Smarter Cities, IBM Redbooks, 2011.

[6] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey," *Journal of Big Data,* vol. 2, no. 1, p. 20, 2015.

[7] Ibrahim Abaker Targio Hashem, Victor Chang and Nor Badrul Anuar, "The role of big data in smart city," *International Journal of Information Management,* vol. 36, no. 5, pp. 748-758, 2016.

[8] Souza A, Figueredo M, Cacho N, Araújo D and Prolo CA, "Using Big Data and Real-Time Analytics to Support Smart City Initiatives," *IFAC-PapersOnLine,* vol. 49, no. 30, pp. 257-263, 2016.

[9] E. Al Nuaimi, H. Al Neyadi, N. Mohamed and J. Al Jaroodi, "Applications of big data to smart cities," *Journal of Internet Services and Applications,* vol. 6, no. 15, p. 15, 2016.

[10] Min Chen, Shiwen Mao and Yunhao Liu, "Big Data: A Survey," *Mobile Networks and Applications,* vol. 19, no. 2, p. 171–209, April 2014.

[11] Nada Elgendy and Ahmed Elragal, "Big Data Analytics: A Literature Review Paper," in *Advances in Data Mining. Applications and Theoretical Aspects*, New York, NY, USA, 2014.

[12] J. P. Shim, Aaron M. French, Chengqi Guo and Joey Jablonski, "Big Data and Analytics: Issues, Solutions, and ROI," *Communications of the Association for Information Systems,* vol. 37, no. 1, p. 797 – 810, 2015.

[13] H. Apache, "Hadoop Apache," [Online]. Available: http://hadoop.apache.org/.

[14] Jara, A.J., Genoud, D. and Bocchi, Y., "Big data for smart cities with KNIME a real experience in the SmartSantander testbed," *Software - Practice and Experience,* vol. 45, no. 18, pp. 1145-1160, 2015.

[15] A. J. Jara, Dominique Genoud and Yann Bocchi, "Big Data in Smart Cities: From Poisson to Human Dynamics," in *28th International Conference on Advanced Information Networking and Applications Workshops*, Victoria, Canada, 2014.

[16] Psomakelis E, Aisopos F, Litke A and Tserpes K, Karda, "Big IoT and Social Networking Data for Smart Cities Algorithmic Improvements on Big Data Analysis in the Context of RADICAL City Applications," in *arXiv preprint arXiv:1607.00509* , 2016.

[17] Welington M. da Silva, Alexandre Alvaro, Gustavo H. R. P. Tomas, Ricardo A. Afonso, Kelvin L. Dias and Vinicius C. Garcia, "Smart cities software architectures: a survey," in

Proceedings of the 28th Annual ACM Symposium on Applied Computing, Coimbra, Portugal, 2013.

[18] D. Singh, C. Vishnu and C. K. Mohan, "Visual Big Data Analytics for Traffic Monitoring in Smart City," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Los Angeles, California, USA., 2016.

[19] Kemp, G., Vargas-Solar, G., Da Silva, C.F., Ghodous, P. and Collet, C., "Aggregating and managing realtime big data in the cloud: Application to intelligent transport for smart cities," in *1st International Conference on Vehicle Technology and Intelligent Transport Systems, VEHITS 2015;*, Lisbon, Portugal, 2015.

[20] Debopriya Ghosh, Soon Ae Chun and Basit Shafiq, "Big Data-based Smart City Platform: Real-Time Crime Analysis," in *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, Shanghai, China., 2016.

[21] Vasylyna Horban, "A Multifaceted Approach to Smart Energy City Concept through Using Big Data Analytics," in *IEEE First International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2016.

[22] Voichita Iancu, Silvia Cristina Stegaru and Dan Stefa, "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., 2016, pp. 489-513.

[23] Dilpreet Singh and Chandan K Reddy, "A survey on platforms for big data analytics," *Journal of Big Data,* vol. 2, no. 1, p. 8, 2014.

[24] Eduardo Felipe Zambom Santana, Ana Paula Chaves, Marco Aurelio Gerosa, Fabio Kon and Dejan S Milojičić, "Software Platforms for Smart Cities: Concepts, Requirements, Challenges, and a Unified Reference Architecture," *ACM Computing Surveys (CSUR),* vol. 50, no. 6, January 2018.

[25] H. M. Cooper, "Organizing knowledge syntheses: A taxonomy of literature reviews," *Knowledge in Society,* vol. 1, no. 1, pp. 104-126, 1988.

[26] J. e. a. Vom Brocke, "Reconstructing the giant: On the importance of rigour in documenting the literature search process," in *European Conference on Information Systems (ECIS)*, Verona, Italy, 2009.

[27] Costa C and Santos MY., "BASIS: A big data architecture for smart cities," in *In SAI Computing Conference (SAI)*, London, United Kingdom, 2016.

[28] Nandury SV and Begum BA. , "Strategies to handle big data for traffic management in smart cities," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, India., 2016.

[29] Takuro Yonezawa, Tomotaka Ito and Jin Nakazawa, "SOXFire: A Universal Sensor Network System for Sharing Social Big Sensor Data in Smart Cities," in *16 Proceedings of the 2nd International Workshop on Smart*, Trento, Italy, 2016.

[30] B. Cheng, Salvatore Longo, Flavio Cirillo, Martin Bauer and Ern, "Building a Big Data Platform for Smart Cities: Experience and Lessons from Santander," New York, USA, 2015.

[31] Bo Tang, Zhen Chen, Gerald Hefferman , Tao Wei, Haibo He and Qing Yang, "A Hierarchical Distributed Fog Computing Architecture for Big Data Analysis in Smart Cities," in *Proceedings of the ASE BigData & Social Informatics 2015*, Kaohsiung, Taiwan, 2015.

[32] Rathore MM, Ahmad A and Paul A., "IoT-based smart city development using big data analytical approach," in *IEEE International Conference on In Automatica (ICA-ACCA)*, Curicó, Chile, 2016.

[33] Schatzinger, S. and Lim, C.Y.R., "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, 2017, pp. 83-98.

[34] Ciprian Barbieru and Florin Pop, "Soft Real-Time Hadoop Scheduler for Big Data Processing in Smart City," Crans-Montana, Switzerland, 2016.

[35] Martin Strohbach, Holger Ziekow , Vangelis Gazis and Navot Akiva, "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, 2015, pp. 257-282.

[36] Khan, Z., Anjum, A., Soomro, K. and Tahir, M., "Journal of Cloud Computing," *Towards cloud based big data analytics for smart future cities,* vol. 4, no. 1, p. 11, 2015.

[37] Tosi D and Marzorati S., "Big Data from Cellular Networks: Real Mobility Scenarios for Future Smart Cities," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications*, Oxford, UK, 2016.

[38] Corradi A, Curatola G, Foschini L, Ianniello R and De Rolt CR, "Automatic extraction of POIs in smart cities: Big data processing in ParticipAct," in *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Ottawa, Canda, 2015.

[39] Shuangmei Ma and Zhengli Liang , "Design and Implementation of Smart City Big Data Processing Platform Based on Distributed Architecture," in *2015 International Conference on Intelligent Systems and Knowledge Engineering*, Taipei, Taiwan, 2015.

[40] Khan, Z., Anjum, A., Soomro, K. and Tahir, M., "Towards cloud based big data analytics for smart future cities," *Journal of Cloud Computing,* vol. 4, no. 1, p. 11, 2015.

[41] Raja A. Alshawish, Salma A. M. Alfagih and Mohamed S. Musbah, "Big data applications in smart cities," in *International Conference on Engineering & MIS (ICEMIS)*, Agadir, Morocco, 2016.

[42] Sinung Suakanto, Suhono H. Supangkat, Suhardi and Roberd Saragih, "Smart city dashboard for integrating various data of sensor networks," in *2013 International Conference on ICT for Smart Society (ICISS)*, Jakarta, Indonesia, 2013.

[43] Thiago Poleto, Victor Diogho Heuer de Carvalho and Ana Paula Cabral Seixas Costa, "The Roles of Big Data in the Decision-Support Process: An Empirical Investigation," in *International Conference on Decision Support System Technology*, Belgrade, Serbia, 2015.

[44] Guido Perboli, Alberto De Marco, Francesca Perfetti and Matteo Maroned, "A New Taxonomy of Smart City Projects," *Transportation Research Procedia,* vol. 3, pp. 470-478, 2014.

[45] Leona Chandra, Stefan Seidel and Shirley Gregor, "Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions," in *2015 48th Hawaii International Conference on System Sciences*, Kauai, 2015.

[46] Ken Peffers, Tuure Tuunanen, Marcus A. Rothenberge and Samir Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems,* vol. 24, no. 3, pp. 45-77, 2007.

[47] Rüdiger Wirth and Jochen Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in *4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000.

[48] Kaoutar Ben Ahmed, Mohammed Bouhorma, Mohamed Ben Ahmed and Atanas Radenski, "Visual Sentiment Prediction with Transfer Learning and big data analytics for smart cities," in *4th IEEE International Colloquium on Information Science and Technology (CiSt)*, Tangier, Morocco, 2016.

[49] F. J. Villanueva, Maria J. Santofimia , David Villa and Jes´us Barba, "Civitas: The Smart City Middleware, from Sensors," in *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, Taichung, Taiwan, 2013.

[50] Li Xiong, Shan Xue, Shufen Yang and Changling Han, "Multi-source macro data process based on the idea of sample=overall in big data: An Applicability Study on Influence Factors to Smart City," in *2015 International Conference on Logistics, Informatics and Service Sciences (LISS)*, Barcelona, Spain, 2015.

[51] Girtelschmid, S. , Steinbauer, M. and Kumar, V., "On the application of big data in future large-scale intelligent smart city installations," *International Journal of Pervasive Computing and Communications,* vol. 10, no. 2, pp. 168-182, 2014.

[52] Huanan Z, Shijun L and Hong J., "Guangzhou smart city construction and big data research," in *2015 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, Nanjing, China, 2015.

[53] Liguo DENG and Chuansheng ZHOU, "Protype Framework of Smart City Base on Big Data and Smart Grid," in *2015 International Conference on Computer Science and Mechanical Automation*, Hangzhou, Zhejiang, China , 2015.

[54] Moreno-Cano V, Terroso-Saenz F and Skarmeta-Gómez AF, "Big Data for IoT Services in Smart Cities," in *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT))*, Milan, Italy., 2015.

[55] X. Wang, "Calibration of Big Traffic Data for a Transport Smart City," in *15th COTA International Conference of Transportation Professionals: Efficient, Safe, and Green Multimodal Transportation, CICTP 2015*, Beijing; China, 2015.

Ahmed Shahat received born in Alexandria, Egypt. He received his B.Sc. and M.Sc. degrees from the department of automatic control and computer engineering, Alexandria University in 1982 and 1986 respectively. Currently, Ahmed is a partner and senior manager at Amiral Management Corp., an Egyptian ICT corporation.

Ahmed is a Ph.D. student at Department of Computer Science, Electrical and Space Engineering – Lulea Technical University. His research focus is in the field of big data applications in smart cities. His research aim at investigating the opportunities and challenges of harnessing big data analytics in enhancing decision making process in smart cities.