NoSQL method for the metric analysis of Smart Cities

Elsa Estrada-Guzman, Rocio Maciel, *Leopoldo Gómez, IEEE Members*

IEEE Guadalajara Metrics for Smart Cities Working Group

Abstract— The objectives of Smart Cities are to ensure the sustainability of the cities and to improve quality of life. The city's performance indicators offer insight into the conditions and factors that influence the improvement of the well-being of the population of the city, and, consequently, these indicators support the metrics of Smart Cities. The source of information of the metrics can be found in the relevant Big Data, which requires extraction and processing for future analysis. Because the relational database models are not large enough to handle the extensive variety of formats and volume of the consultations, this paper demonstrates the design of a NoSQL schema with storage in the columns to facilitate the scalability and speed of searches.

Index Terms—Smart city metrics, Schema NoSQL, Smart Cities.

1. Introduction

THE goals of this document are to present the main plan of Guadalajara's Smart City Metrics Structure and to provide a brief description of the Guadalajara Smart City Framework that will process the information of the city. The term Smart Cities includes economic, political, social and cultural aspects. The Smart City concept began out of the need to improve the conditions and the lifestyle of the cities that tend to develop in a disorganized manner. The current literature on Smart Cities was intended to establish objectives to achieve the well-being of the population, as well as to identify its elements, as the main actors (the citizens, the government and the private sector) must collaborate to remove the barriers of division and segmentation that obstruct growth (development). With these objectives, it is possible to promote the role of the government in these tasks, the participation of the citizens in decision-making, and the contribution of the private sector to the stimulus of the economy and innovation.
Another important element is the Information and Communication Technology, which is expected to reduce the digital divide and "promote the information society", which can be monitored through the metrics of the cities that seek the establishment of sustainable cities.

The first part of this work describes the proposed non-relational database schema to facilitate the analysis of the prevalent metrics of cities on the Internet for the consumption of the relevant actors. The second part of this work presents a panorama of the metric cities, expressing the importance of measuring the appropriate parameters, recollecting information for the identification of the metrics and the search for the sources of information.

In the third part, the metrics model is analyzed, with the objective to understand the complexity of its requirements, within which can be found the sources of the metrics as well as the size of the audience or users (actors) who use the internet as an ordinary means of communication; the internet acts as a repository of information on the metrics. In the fourth part, the storage models of NoSQL are studied, and some of the characteristics are described, with the objective of identifying the model that best fits to the solution of the proposed/suggested problems. In the fifth part, the required transformative process for the analysis of necessary data is described, and the NoSQL schema is established.

2. Context

2.1 Metrics of Smart Cities

The word "metric" is often used to refer to a "measurement" activity; despite both terms overlapping in meaning, they have their differences. Defining a metric is something abstract; it is also a subjective attribute, intangible and qualitative, that it is used as a tool to facilitate the decision-making. A measurement helps to approximate metrics because a measurement provides quantifiable and observable results, and the measurements are the objective.

The quality of life, which is a metric of a Smart City, is considered not quantifiable and abstract, although it helps to understand the state of a city. With this metric, it is possible to evaluate the current conditions and to guide the progress. However, measurements must be

- *Elsa Estrada-Guzman is with the ITPhD CUCEA UDG, CO 45100. E-mail: elsa.estrada@ieee.org*
- *Rocio Maciel is with the Coordinator of the ITPhD CUCEA UDG, E-Mail: r.maciel.mx@ieee.org*
- *Leopoldo Gómez is a Researcher at Smart Cities Innovation Center CUCEA UDG, E-Mail: dr.leopoldo.gomez@ieee.org*

found that support this metric: performance indicators of the city that should be shared with the population for its analysis in order to build an information society.

2.2 The search of the key indicators of performance

The key performance indicators (KPIs) help to define the objectives of an organization. KPIs are measurements used in the process of quantification of the efficiency of the internal functions. In this study, we refer to the performance of a city, to the search of its parameters, and to the factors that influence its conditions.
Any city possesses multiple indicators, many of which have been developed by public and private institutions, and the work on them continues and tends to increase. As a consequence, the minimum and maximum of the cluster is unknown because each indicator can multiply when it is deconstructed into other, more specific, indicators.

The citizens must know how to improve their quality of life and how to measure it, which implies time, recollection effort, analysis and fixation of the associated indices. This activity is dynamic because of the continuous changes in the social paradigms, and, most of all, because of the evolution that is noted in the paradigm of well-being that directly influences the feelings and priorities of the citizens, impacting the appreciation of new KPIs.

Currently, sustainability and equity receive a major proportion of the attention. This appreciation has served as a foundation for the compilation of KPIs, from which different criteria have been established, with the aim of classifying them and directing them to different focuses. For example, the Global Power City [1] index identifies the actors (researcher, manager, visitor, resident, and artist) as reference points to classify and evaluate the aspects of a city. City Vitals [2] groups the indicators around four intangible values: talent, innovation, connection and distinction. The authors are convinced that human attitudes have the power to transform the future. Global Ranking for Smart Cities [3] contributed the model that shows the structure of the KPIs around six city areas: **Smart Government, Smart Economy, Smart Mobility, Smart Environment, Smart People and Smart Living,** which are referred to as smart areas. This model serves as a reference for this study.

As was previously mentioned, previously established metrics exist that are multiplying in different sources of information, such as the internet, the government, industry, and nonprofit organizations. These metrics participate by sharing their data. The challenge that is anticipated is the recollection and analysis due to the multiplicity of formats that make its treatment complex.

2.3 The tendency towards the open data framework with a participative audience

The participative producers of open data audiences are the government, citizens, and the public and private sectors. These producers have special interest in the extraction of the information generated from the cities' KPIs to build new knowledge or to make decisions based on the analysis (Table 1). At the same time, each of the participating audience members is a producer of information of different situations and events. The government propagates information from the census that is a good approximation of the population, social behavior, happiness parameters, quality of life, employment and education situation, income and poverty. The public courts or public sector share information related to the indices of contamination, environment, economy, etc.

Actor	Type of required data	Type of visualization	Type of access
Government	Influence factors the quality of life, demand of resources and scarcity	• Sorted query by priority • Statistic correlations between factors • Decision making models	• On line consultation • Evaluation to sectors • Recommendations • Direct download
Private Sector	Areas of opportunity, economic development factors, education pro-grams, gross production.	• Statistic correlations • Metadata	• On line consultation • Evaluation and voting • Direct download
Citizens	Budget use Resources distribution, Social and cultural projects, Factors that influence public security	• Interface for an: interactive participation	• On line consultation • Evaluation and voting • Direct download

TABLE 1
USE OF DATE EXPECTATION, CLASSIFIED BY RECIPIENT

The participating audiences are the government, citizens, public and private sectors. These audiences have a special interest in the extraction of the information generated from the cities' KPIs to build new knowledge or to make decisions based on the analysis. At the same time, each member of the participating audience is a producer of information of different situations and events. The private sector participates with indices of business economics, finances, investments, market tendencies, etc. The citizens contribute by providing of information through the social network, the use of electronic mail and forums to display preferences, beliefs and perceptions of their life conditions. In such a way, the data generated by some participants are of value to others. These metrics are localized in several sites, covering large numbers of groups of registers and files. The benefit of their analysis

would be the revelation of patterns that could be significant to the actors; this idea is already considered by the "big data" concept, by which a revision of the characteristics that outline this group of metrics must be performed to obtain an effective analysis.

3. Data model requirements for smart city metrics

With the Internet, the volume of information increases dramatically: "almost every company is procuring the digital representations of its existing data, which results in high increase of digital data growth" [4]; as a result, the personal computer processing capacity is insufficient, and there is a demand of high efficiency in the availability and computer performance for analysis under heavy loads and with less cost and time. Therefore, the problems encountered with this expected amount of information are the three Vs of volume, velocity and variety, which are commonly used to characterize different aspects of Big Data [5], as explained from the perspective of recovering metrics in the following paragraphs.

3.1 Volume management requirement

The metric's sources are found in the internet throughout several storage locations; for example, transparency in government generates repositories distributed in an open style environment, reaching terabytes of population and household census storage; public security, estate and annual municipal justice census is aggregated, as well as polls, studies and registries with different focuses. In the same way, the Private Sector share open access, such as the economic and financial indicators related to Merchandise trade.

The next trend that is envisioned will cause a growth of the amount of data is the development of the phenomenon called "The Internet of Things" [6]. With the proliferation of sensors in objects, events that are stored in the systems that flow in the network are captured. These mechanisms or devices write simultaneously in the designated storage, and there are indicators in them that must be filtered for the analysis. As a result, petabytes of data are expected to be reviewed.

3.2 Velocity requirement

The actors' common activities are online consultation and direct upload, which demand rapid response times. The consultations involve readings that simultaneously require extraction and analysis. The uploading function warns the writing function about the storage that will be available to the user, with actualization speed according to the actualization speed of Big Data sources.

3.3 Management requirement of the variety of formats

The variety that is considered in the metrics of Big Data includes different types of documents that have different internal structures, depending on each system or place of origin. Table 2 presents some sources of KPIs that are classified by smart areas, according to Cohen's model.

TABLE 2
FORMATTING OF METRIC SOURCES F, CLASSIFIED BY AREAS OF A SMART CITY

KPA	Metric	Source	Format
ENVI-RONMENT	Temperature, humidity, temperature, lighting, carbon monoxide, nitrogen dioxide	Sensors	JSON
PEOPLE	Life expectancy Education level	Census	XLS,XLSX,CSV, PDF
GOVERN-MENT	Usage of public expenses Public budget	OpenData	CSV, XML, ZIP, KMZ, SOAP+XML,VND., CSV, JSON,RAR PNG
MOBILITY	Public transportation services. Connectivity	Sensors, urban applications	HTML,XLS,PDF,JASON
ECONOMY	Gross Domestic Product Inflation	Census Business applications	PDF, XLS, Multiple database formats
LIVING	Housing quality Parks	Housing and urban development department	PDF, XLS, Multiple database formats

3.4 Data analysis requirement

Currently, the field of Big Data is changing the way in which investigation is conducted, e.g., it is possible to adopt skills to solve complex problems associated with scientific discoveries that surround multiple knowledge areas, such as health, security, education, etc. In this case, patterns must be found, starting with the city's KPI analysis. Scientific analysis is one of the tools required to assist in this process, together with the algorithms to classify the information, such as the Bayesian model, especially the useful kernel for classification tasks, regression analysis, and cluster analysis of spatial data. Performing the correlations between the KPIs leads to the establishment of forecasting via the observation of patterns.

4. NoSQL storage models

4.1 Antecedents

NoSQL databases were developed after 2009; the main idea was to give more storage power to the web sites whose user size is constantly increasing, while taking into account the scalability [7]. In other words, the storage size increases are set to maintain high availability. Another important characteristic is that these databases do not maintain an internal organization of tables with elements of the same kind, as in the relational model, which, at the same time, generates advantages in some aspects and disadvantages in others.

Relational databases (RDB), originating with the work of Codd [8], were and still are an important model; normalization is a powerful function that is used to decrease the redundancy and to reduce storage costs. The model was organized in tables, containing tuples of the same kind

based in the established relationship. In this system, an attribute of each table can be designated as a primary key, and its value is used to extract a unique registry or to make inquiries in one or several tables. The query language (or consultation language) for the RDB is Structure Query Language (SQL), which facilitates the construction of queries incorporating restrictions, groupings and recursive functions [9]. These activities are units of work, called transactions. A transaction consists of one or more statements in SQL, and each transaction must have the following properties: atomicity, consistency, isolation, and durability (ACID) [10]. The atomicity means that a transaction (create, update, select, delete, etc.) must be processed entirely, and these are indivisible. The consistency is the capacity to write valid pieces of information according to rules and restrictions imposed. The isolation refers to the concurrency control and the durability of the permanency of storage.

The priorities of the relational database systems (RDBSs) lean more towards consistency than towards availability because, usually, whoever implemented them at the in the first place maintained little information of their businesses and enterprises.

The challenge starts with the net traffic and the huge information volumes and the ability to handle different formats that block/stop the table design with a unique key structure for only one registry. The distributed database systems for the RDB attempt to resolve the huge volume of problems, but do not support the segmentation for the parallel processing. In this way, the NoSQL models must provide an answer to these concerns/worries and, with it, the possibility for other solutions is available.

4.2 The CAP Theorem (Consistency, Availability, and Partition Tolerance)

Brewer [11] proposed a model for distributed bases systems in which he starts with definitions describing three characteristics: consistency, availability, and partition tolerance. The consistency is the capacity of the machines to obtain access to the same data at the same time and at any time. The availability is considered to be the capacity to respond to petitions. The partition tolerance is the capacity to operate any of the faulted servers. The theorem established that it is impossible to have the three functions at the same time for which the RDB are deemed appropriate for applications where consistency is absolutely required.

If the partition tolerance must be maintained, which is the case for the storage that secures the process even when a server fails, then the NoSQL storage models are considered.

Figure 1 shows the Relational DataBase Management System (RDBMS) with availability attributes and consistency. The NoSQL storage for databases (DBMS) maintains only the consistency and the partition tolerance. The NoSQL is based only in the availability and the partition tolerance. Still, there are no systems with all three attributes. However, companies and users are increasingly choosing NoSQL databases. The three most popular types are **Key - value stores, Document-based stores** and **Column-oriented databases** [12].

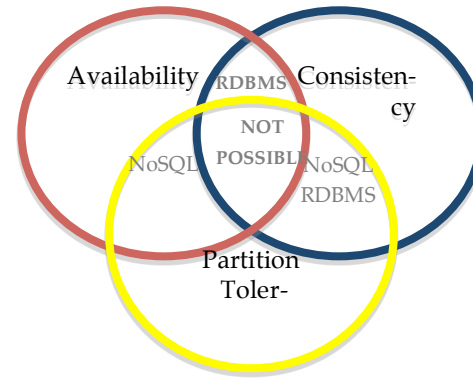

Figure 1. Classic model of the CAP theorem: A distributed storage system can only guarantee two functions simultaneously

4.3 Key - value stores

This type of storage is very similar to the organization of hash tables, in which the key is the position of the arrangement of the storage. The value that is directed by the key can be any type of data, such as a chain, JSON or XML documents, and a binary blob. The disadvantage of this structure is that the access throughout the values is impossible and deriving consultations or queries from the values is difficult; however, it is easy to escalate these types of databases due to the simplicity of their table distribution in different computers [13].

4.4 Document – based stores

This is a kind of database that allows for the organization of a collection of documents, which can be of any sort. One difference with respect to the key-value is that is possible to store an entire page in only one registry. This model allows for the consultation and location of data through the values without requiring the key. The registries are kept independently, and if a change occurs, it does not affect the rest of the registrations. One of the advantages is the complete recovery of a set of data with one simple consultation. Registrations are divided among many machines when the traffic is increased, speeding the consultations/queries of any content because it is enroute to one computer alone. This model is useful for web sites where there are more readings than content writings.

4.5 Column-oriented databases

The main support is to address the huge number of columns, the sparse nature of the data and frequent changes in schema. Column values are stored contiguously, with the aim to improve performance for aggregations and dynamic queries. In addition, columns of those rows that satisfies the

where clause of the query are retrieved, which causes unnecessary disk input/output [14].

5. NOSQL OUTLINE DESIGN FOR METRIC DATABASES

The design of an outline "database schema" for the metrics must be adjusted to the characteristics and requirements that were analyzed previously; furthermore, the number of users that would have access to the same values at the same time must be considered. This quantity of users is the Mexican population of more than 112 336 538 million people registered in 2010 [15]. The method of access is the reading of the uploading of the enquiry.
There will be demands of consistency because the trust level depends on the accuracy level, not only for the enquiries but also for the synchronization among the simultaneous requests of the users. Speed plays a substantial roll because the volume of data that will be uploaded will be very high.
The data must be processed in advance; the primary tasks are the recollection, screening, classification and analysis. The purpose of this procedure is to find meaningful information for the users. This process is visualized in Figure 2; this flow is composed of blocks of programs, algorithms, and formats that represent the ins and outs of the blocks.

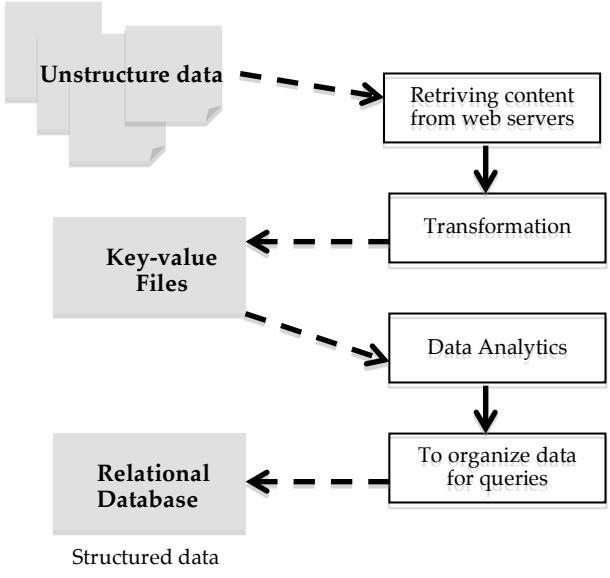

Figure 2 Metrics of Smart Cities, process of treatment inside Big Data

The process starts with the extraction of the data without structure and with a variety of formats that prevail in the sites that contain indicators of city from Big Data. Next, the transformation NoSQL is produced to facilitate the speed of the analysis; specifically, the model key-value as there will be no internal searches to the process. Only finding correlations and classifying are required to identify cluster(s) over data in spaces of statistics methods.
The last block organizes the registers obtained by the analysis in a structured format to accelerate the free access on line enquiries of the users; in that way, a transformation to the format requested by the user can be performed.
The outline that is proposed here is for the archives that are produced by the transformation. The outline is inspired by the Dremel processing style, a system of enquiries for nested data analysis [16]. Dremel has the advantage of the disposition in columns to complete the distribution of parallel processing, thereby performing in seconds enquiries regarding trillions of tables. Dremel processing is used by Google and allows for scaling into thousands of CPUs and petabytes and it is a complement of Map Reduce. Map Reduce is an algorithm that transforms the data in key - value registries and utilizes scalability [17].
The reason why the model in the columns is appropriate for the structure of the metrics' content is because each document includes indicators of different areas of a smart city and it is necessary to group them for their analysis. The format is supported by different technologies, such as MR, Sawzall, and FlumeJava.

Figure 3 shows the outline for the nested indicators of People and Living that correspond to two areas of Smart Cities in Cohen's model. The messages definition "Search request" was used for its creation.

```
message Document
    {
    required int64 DocId;
    repeated group Inidcator
       {
        repeated group People
          {
          required string Education;
          required string Security;
          }
        repeated group Living
          {
          required string Housing;
          }
       }
    }
```

Figure 3 NoSQL schema for extracting metrics in Big Data

Figure 4 shows two examples of content processed by the previous outline, with an exit to the 4 tables that are below Figure 5. Each table contains values of only one attribute and a column; for example: Indicador.People.Education found schools with two repetitions, Indicador.People.Security shows kidnapping 21 times and theft nine times, and so on.

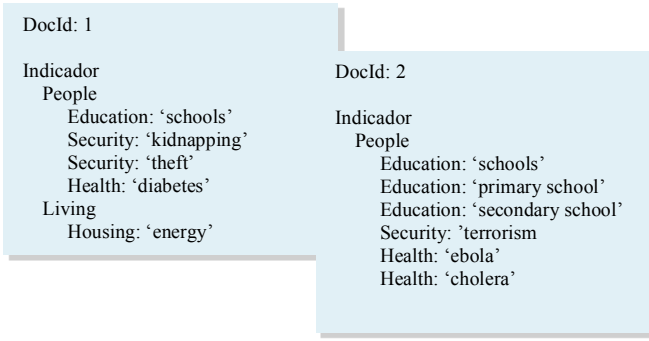

Figure 4 Two examples produced by the extraction using the proposed NoSQL schema

Indicador.People.Education	
value	r
schools	6

Indicador.People.Security	
value	r
kidnapping	21
theft	9

Indicador.People.Health	
value	r
diabetes	34

Indicador.Living.Housing	
value	r
energy	12

Figure 5 Four tables with results after of the processing was applied

One of the advantages of this format is the possibility to group the indicators together that belong to only one area, for example, Education and Security are linked to People.

6. Concluding remarks

A variety of tools exist that enable the processing of large volumes of information, but these still depend on the design of the outline of the databases, as one must specify the desired attributes that must be escalated to add others. The existing outlines are limited to the extraction of chains, which prevents the analysis of the factors associated to other types of data and could surround an event that is presented in the content of the document.
To evaluate the proposed outline, it is necessary to integrate a platform that allows for the monitoring of the consistency and tolerance to the participation; however, this platform is outside the scope of this paper.

7. References

[1] The Mori Memorial Foundation, “Global Power city 2013”, Octubre 2013; [Online] http://www.mori-m- fountion.or.jp/english/research/project/6/pdf/GPCI2011_English.pdf

[2] J. Cortright, Impresa Consulting, CEOS for cities, “City Vitals”, 2014; [Online] http://www.ceosforcities.org/city-vitals/

[3] B. Cohen, “What exactly is a smart city”, Septiembre 2012. [Online]. http://www.fastcoexist.com/1680538/what- exactly-is-a-smart-city. [Last accessed: 11 09 2014].

[4] P.Bedi, V. Jindal, A. Gautam, "Beginning with Big Data simplified," IEEE, Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on, vol., no., pp.1,7, 5-6 Sept. 2014 DOI: 10.1109/ICDMIC.2014.6954229

[5] “Big Data Now: 2012 Edition”, O’Reilly Media, Inc., October 2012, p. 3-7. ISBN: 978-1-449-35671-2

[6] J. Wielki, “Implementation of the Big Data concept in organizations – possibilities, impediments and challenges”, IEEE, Proceedings of the 2013 Federated Conference on Computer Science and Information Systems pp. 985–989

[7] S. Edlich, “NoSQL Databases”, Berlin, Germany, http://nosql-database.org/, [Last accessed: 25 02 2015]

[8] E. F. Codd, “A Relational Model of Data for Large Shared Data Banks”, IBM Research Laboratory, Communication of the ACM, San Jose, California, P. BAXENDALE, Editor, 1970, Volume 13, Number 3.

[9] L. Davidson, K. Kline & K. Windisch, “Pro SQL Server 2005 Database Design and Optimization”, 2006, p.47.

[10] G. Bell, L. Lamport & B. W. Lampson, “James N. Gray 1944-2012 Biographical Memoirs”,2013, National Academic on Sciences, [On line]
http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/gray-james.pdf

[11] E. Brewer, “CAP Twelve Years Later: How the Rules Have Changed”, IEEE, Computer Society, 2012, Volume 43, DOI: 10.1109/MC.2012.37

[12] N. Leavitt, “Will NoSQL Databases Live Up to Their Promise?”, IEEE, 2010, Volume 43, DOI:10.1109/MC.2010.58

[13] M. Manoochehri, “Data Just Right”, Addison Wesley Data & Analytics Series, 2014, p. 32

[14] K. Kaur, R. Rani, “Modeling and querying data in NoSQL databases”, Big Data, 2013 IEEE, International Conference on, DOI: 10.1109/BigData.2013.6691765

[15] INEGI, “Instituto Nacional de Estadísticas y Geografía”, “Población de México”, 2014, New Delhi,
http://cuentame.inegi.org.mx/poblacion

[16] Google, Inc. “Dremel: Interactive Analysis of WebScale Datasets”, The 36th International Conference on Very Large Data Bases, September 2010, Singapore.

[17] Google developers, “Protocol buffers, lenguaje guide”, 2014; [Online].
https://developers.google.com/protocol-buffers/docs/proto