Study Case: The Database Selection Process for the Big Data-based System to Reduce Health Effects of Air Pollution in Ciudad Juárez, Mexico.

Caso de estudio: El proceso de selección de la base de datos para el gran sistema basado en datos para reducir los efectos en la salud de la contaminación del aire en la Ciudad de Juárez, México

[a]Adrián Vásquez, [b]Fernando Estrada, [c]Alicia Jiménez, [d]Angel Nieves, [e]Nabile Rodríguez, [f]Israel Hernández

[a]Laboratory of Climatology and Air Quality, al197503@alumnos.uacj.mx,Orcid:0000-0003-3807-565X, Universidad Autónoma de Ciudad de juárez, Ciudad de Juárez, México.

[b]National Laboratory of Information Technology, al181765@alumnos.uacj.mx,Orcid:0000-0003-3807-565X, Universidad Autónoma de Ciudad de juárez, Ciudad de Juárez, México.

[c]Institute of Engineering and Technology,al116437@alumnos.uacj.mx,Orcid:0000-00033807-565X, Universidad Autónoma de Ciudad de juárez, Ciudad de Juárez, México.

[d]Institute of Engineering and Technology,al171496@alumnos.uacj.mx,Orcid:0000-00033807-565X, Universidad Autónoma de Ciudad de juárez, Ciudad de Juárez, México.

[e]Institute of Engineering and Technology,al171503@alumnos.uacj.mx,Orcid:0000-00033807-565X, Universidad Autónoma de Ciudad de juárez, Ciudad de Juárez, México.

[f]Institute of Engineering and Technology,israel.hernandez@uacj.mx,Orcid:0000-00033807-565X, Universidad Autónoma de Ciudad de juárez, Ciudad de Juárez, México.

Recibido: 7 de Enero de 2018 Aceptado: 15 de Mayo de 2018

Forma de citar: A. Vásquez, F. Estrada, A. Jiménez,A. Nieves,N. Rodríguez, I. Hernández, "Study Case: The Database Selection Process for the Big Data-based System to Reduce Health Effects of Air Pollution in Ciudad Juárez, Mexico.", *Mundo Fesc,* vol. 9, no. 17, pp. 23-30, 2019.

Abstract:

The Border 2020 is a U.S.-Mexico effort to address binational environmental problems along the border. This project involved the city of El Paso, Texas and Ciudad Juárez, México to improve the transboundary air quality. A large portion of the Ciudad Juarez population resides in areas with very few or none air quality monitoring stations and also people is not educated on the health effects of exposure to air pollution. This motivated an innovative community-based climate monitoring scheme to increase the awareness among people on the effects of air pollutions. The idea was to manufacture a large amount of low-cost air quality sensors, located at different strategic sites to cover a major portion of the city and then to measure and analyze meteorological variables and alert people about outdoor activities when their health is at risk. To achieve this, it was considered a big data-based system to collect, store, analyze and visualize a large amount of data. Selecting the appropriate database software to store large volumes of data is a key element in these projects. Recent advances in storage technology show two main approaches of databases: SQL relational and NoSQL non-relational databases. This paper discusses important factors to consider when selecting the database software for climate data and presents a performance comparison between SQL and NoSQL databases in specific scenarios involving operations such as inserting, deleting and updating a massive volume of both structured and unstructured data.

Keywords: Community Monitoring, Air Pollution, Environmental Quality Index, Databases, Big data.

Autor para correspondencia:
*Correo electrónico: israel.hernandez@uacj.mx

Introduction

The Border 2020 program is a joint effort by the U.S. Environmental Protection Agency (EPA), Mexico's Secretariat of the Environment and Natural Resources (SEMARNAT), and the Border Environment Cooperation Commission (BECC). It is focused on the environmental and public health challenges through joint-governance, key partnerships, and projects implemented by federal, state, tribal and local partner [1]. The U.S.-México border extends from the Gulf of México to the east to the Pacific Ocean to the west. Along the border there are border cities with a considerable amount of population (See Figure 1). The 2020 program also considers mutual collaboration in research projects that generate tangible benefits to the U.S.-Mexico border communities such as 1) improve air quality; 2) provide clean and safe drinking water; 3) revitalize land and prevent contamination; and 4) enhance emergency preparedness and response [1-4].

Figure 1. U.S. – México Border Region [1].

This project addresses the Border 2020 Program's Region-wide Priority Area "Strengthen existing air quality monitoring network, to include the development of/or reinforcement of existing strategies for establishing air quality monitoring between the border cities of El Paso, Texas and Ciudad Juárez, Mexico. Ciudad Juarez is the most populated city in the state of Chihuahua, Mexico with a population of 1,391,180 inhabitants and an extension of 3,560 Kms. [5]. A large portion of the Ciudad Juarez population resides in areas with few or none air quality monitoring stations and also people is not educated on the health effects of exposure to air pollution.

This motivated an innovative community-based monitoring scheme to increase the awareness among people on the effects of air pollutions. Thus, it was created the Ciudad Juárez Climate Network, composed by air quality monitoring station located at different site Most of these stations were based upon the development of a large amount of low-cost sensors which allow to cover a major portion of the city. Each station has been attached with a wireless device to send data to a remote database system. Thus, the emission of the main air pollutants (PM2.5 or higher, Ozone and Carbon monoxide) across the region is directly measured in real time. This network allows to effectively measure the impact of air pollution on human health of the area residents and to alert people about outdoor activities when their health is at risk. The main objectives of this project are: 1). Strength the existing air-monitoring network of Ciudad Juarez by adding air quality data on portions of the city under represented by the existing network; 2). Inform and educate the community on the health effects of exposure to air born contaminants; 3). Create awareness on the combined effects of meteorology with waste incineration on the quality of the air.

To achieve these objectives, it was considered the development of a big data-based system to collect, store, share, analyze, visualize a large amount of data. The term big data was first introduced by Roger Magoulas from O`Reilly media and associated with a large volume of data that existing data tools cannot manage and process due to the complexity and volume of this data [6]. Big data is characterized by three V's: Volume, Velocity and Variety. Volume denotes a large amount of data (in order petabytes or more). Velocity refers to the rate at which data is collected from a considerable high number of sources and made available for its use. Variety refers not only both structured and unstructured data, but the heterogeneous nature of data. A common mistake is to think that big data is just about data when what it really matters, is not the data itself but what we can do with the data. Note that, in this case, it is intended to benefit the community with a big-data based system focused

on reducing health effects of air pollution in Juárez City. A key component of the big data is the database software to store data generated by the climate monitoring stations. Recent advances in database technology show two main approaches of databases: SQL and NoSQL. The selection of the appropriate database software represents an important decision that must be guided by the consideration of specific factors which are described above.

The Community-based Climate Monitoring scheme.

The monitoring of air quality is the fundamental pillar of this project because it is directly related to the public health of the inhabitants of the different regions of Ciudad Juárez. To address this problem, the creation of an Environmental Quality Index (EQI) is proposed, which aims to incorporate the main atmospheric irritants, observed in the Ciudad Juárez -El Paso region. The calculation of the EQI incorporates empirical weights to account for the synergies between stressors such as temperature and humidity with atmospheric irritants such as ozone, fine particles and carbon monoxide. Stressors are monitored through community stations designed and manufactured ad-hoc for each region of the city. Their low cost foresees that they will be able to be deployed in great numbers in all the colonies of the city. The EQI allows determining the risk related to daily activities of people exposed to a neighborhood scale. Through the use of Information Technology, the aim is to disseminate the values of the EQI through citizen alerts that empower the inhabitants of the city to take preventive actions to avoid respiratory diseases. To achieve the above, the following steps are followed: 1) Divide the city by regions (see Figure 2). 2) Install sensors in strategic places in each region. 3) Store the data in a database. 4) Create an application to periodically compute the EQI. 5) Use mobile technologies to disseminate and monitor the EQI among the inhabitants of the city. Just as important as the technological challenge, it is the challenge of educating and raising awareness among people about their

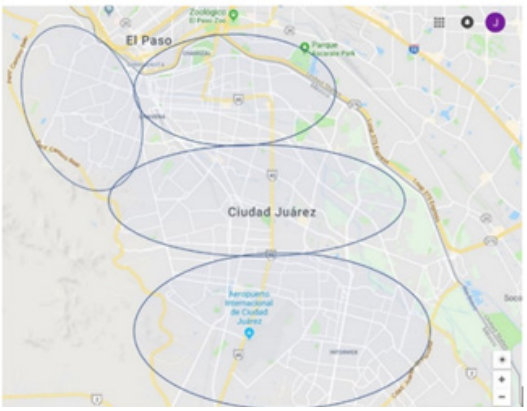

Figure 2. Regions of Ciudad Juárez.

The SQL and NoSQL approach

In the context of database technology, there exists two types of database models: SQL and NoSQL. While both approaches are viable alternatives, there are certain differences that users must consider when selecting the best alternative for the application.

The SQL approach.

The SQL database is also known as relational database, which is a structured method of storing data in the form of tables with relationships. A relational database is based on a rigid storage schema composed by tables, fields, relations among tables and a set of rules to keep the data integrity in the database. A schema must be previously well designed before storing data, because a change in the structure, even a simple change like adding a column, would be difficult to made. The Database normalization is the process of restructuring a relational database in accordance with a set of rules contained in a sequence of normal forms, in order to reduce data redundancy and improve data integrity. All the transactions made in relational databases follow the ACID properties [7-8]: 1. Atomicity, which ensures that either all the operations of a particular transaction are reflected in database or none. 2. Consistency, which preserve the consistency of database, the execution of a particular transaction should take place in isolation (it means that no other transaction should run concurrently

when there is a transaction already in progress). 3. Isolation, which state that for every pair of transactions, one transaction cannot start execution unless the other transaction finished execution. 4. Durability, which consider that once a transaction completes successfully, all the data updates made into the database must remain in time, even in the presence of a system failure.

The NoSQL approach.

NoSQL means Not-Only-SQL, in other words, another alternative to store data. Generally, NoSQL are aligned to distributed systems, allowing horizontal scalability, which means that it is possibly to add server easily to the system. Generally, NoSQL maintains copies of the same data across the different servers that compose the system. There are four types of NoSQL Databases [9-10]: key-value databases, document-oriented databases, graph databases and column-family databases. All NonSQL databases claim to be schema-less, which means that do not follow a rigid structure to store data. It means that it is possible to store data with different attributes in the same column. All the transactions made in NoSQL databases follow the BASE properties: 1. Basically Available, which indicates that the system does guarantee availability by maintaining data stored in distributed servers. 2. Soft State, indicates that the state of the system may change over time, even without input. This is because of the eventual consistency model. 3. Eventually consistent, indicates that at some point in the future, data will become consistent. No guarantees are made, however, about when this will occur.

The Database Selection Process.

In this section it is discussed the database selection process to store the data generated by climate stations deployed in the different regions of the city. For this, three key aspects were considered: the nature of data, the database functionality and a performance evaluation in different scenarios.

The nature of data.

Before considering any technical details and perhaps the most important factor when selected the database software, was to understand the nature of data flowing to and through the big data system This project considers the use of large number of sensors to measure different climatic variables.

In many occasions the sensors can be provided by different vendors. This causes the generation of the data to be heterogeneous, which means that each vendor defines the manner in which the data record is generated in each sensing. Heterogeneous data record may vary in the number of fields or even the case in which two sensors from different vendors, which measure the same climatic variable, can generate the same amount of data but with different format (i.e., one sensor can report the value of the climatic variable as a floating value and the other sensor as a double value). Homogeneous data means that the records generated by all the sensor are always the same in number of fields and data type.

The database functionalities.

The database functionalities that this project pursued includes: 1) Extract and filter data. 2) Script Autonomation. 3) Data exchange format. 4) Volume of Data. 5) Data Storage. 6) Scalability.

Functionality	SQL Databases	NoSQL Databases
Extract and filter data	Relational databases allow the extraction and filtering of data from databases through the use of the SQL language, which can be used by any relational database of the literature.	Unlike the relational databases, the NoSQL found in the literature have their own query language. This certainly can cause confusion when deciding which NoSQL database to be used.
Script Autonomation	There are operations with the data that are carried out in a certain order every certain time. It is required that these operations can be integrated into a script and executed automatically. The SQL language allows the creation of scripts that any database can execute and schedule its automatic execution.	In general, the query languages of the NoSQL databases allow the creation and automatic execution of scripts.
Data Exchange Format	The project requires to convert a set of data stored with a certain structure to a data file with a format recognized in the literature. Relational databases provide the means to perform these operations.	NoSQL databases, like SQL, allow these operations to be carried out with most of the known formats.
Volume of Data	The volume of data stored in a database has a direct impact on its performance. It is usually accepted that a query takes more time when the volume of data is greater. SQL databases tend to be the most affected by the volume of data, partly because of the rigid scheme that determines the use of several tables to store the data.	The flexible scheme of NoSQL databases, which among other things considers the least number of tables to store information, allows queries to the database to perform better than relational databases.
Data Storage	The SQL databases have the characteristic that when a field has the value of NULL in a data record, when the record is stored in the database, space is reserved for the field, as if it had some value. The above causes the hard drive to increase its storage.	In the case of NoSQL, when a field has a value of NULL, it is simply not included in the registry and no disk space is occupied. This causes a better use of the hard disk.
Scalability	The SQL databases tend to scale vertically, which means that once a data server is reaching its storage capacity limit, it is necessary to acquire a new server with more storage capacity.	The NoSQL databases are prepared to scale vertically, which means that once a data server is reaching its storage capacity limit, it is possible to add another server with the same characteristics and connect it in a network.

Table 1. Description of the database functionalities.

Performance Evaluation.

This section shows a performance comparison between SQL and NoSQL databases. It is divided in two parts: the first part describes the methodology used and the second part shows the results and analysis of the evaluation.

Methodology.

I.The NoSQL database used in this evaluation was MongoDB.

II. The SQL database used was SQL Server.

III. The scenarios were planned to consider 10,000, 30,000, 50,000, 100,000 and 1,000,000 data records to which specific operations were applied.

IV. Definition of the metrics: Inserting new records, Disk Space Usage, Inserting new records with image of 2MB, Inserting new records with image of 20MB, Updating the 50% of records.

V. The characteristics of the equipment where the tests were carried out: Lenovo ThinkPad L440 Laptop, Processor: Core i5, 2.60Hz and 8GB of RAM.

Results and Analysis.

The results obtained to measure the time when inserting data records in both databases are shown in Figure 3, where it is observed that for all the scenarios, the SQL Server takes more time to perform this operation. In the scenario of 1,000,000 records, SQL Server needs up to 25% more time than MongoDB.

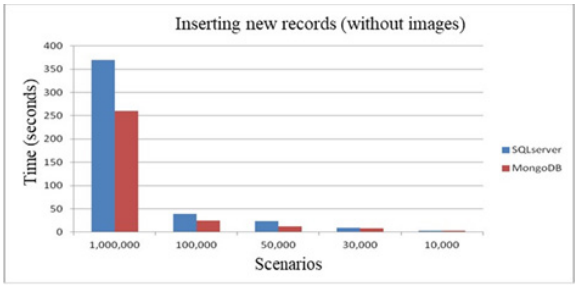

Figure 5. The database Insertion Operation (without images)

Next, the disk space utilization for both databases was measured and the results are shown in Figure 6. The graphic shows that MongoDB occupies more disk space than SQL Server, perhaps because MongoDB used only a big table to store data, this may increase the data redundancy. The results show that, in practically all the scenarios, MongoDB requires up to 2.2 times more disk space than the SQL Server.

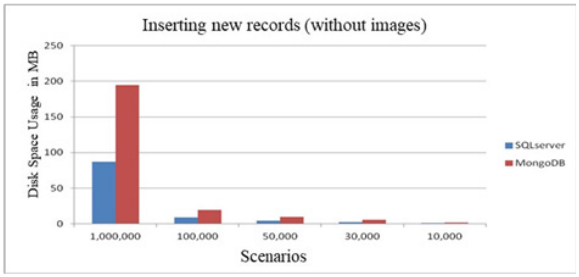

Figure 6. Disk Space Utilization.

The database updating operations results are shown in Figure 7, where for all the scenarios SQL Server requires more time to perform this operation than MongoDB. It is important to mention that this experiment considered the update of the 50% of the records stored. In the scenario with 1,000,000 records, SQL Server requires up to 4 times more time than MongoDB to update half of the records.

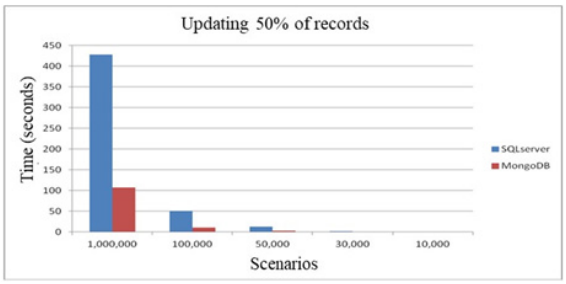

Figure 7. The database Update Operation.

Now, in order to stress the data load, it was considered a specific scenario where every 10 records inserted, an image of size 2 MB was attached to the 11th record and inserted in the database. The results obtained in both databases are shown in Figure 8, where it is observed that for all the scenarios SQL Server requires more time to perform this operation. In the case of 100,000 inserted records, SQL Server aborted the process while MongoDB finished the process properly. In the scenario of 1,000,000 records, SQL Server simply could complete the case.

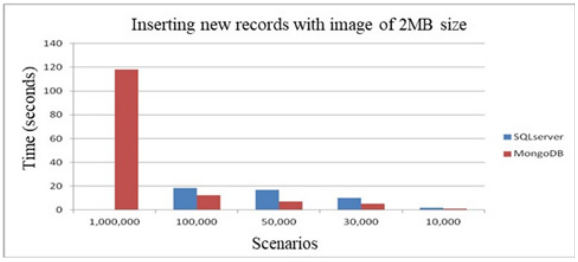

Figure 8. The database Insertion Operation with image of 2MB size.

Now, an extreme scenario was carried out where for each 10 records inserted, an image of size 20 MB was attached to the 11th record and inserted in the database. The purpose of this last two scenarios was to observe the performance of both databases in extreme conditions of data load. Figure 9 shows the results and it is observed that for the case of 10,000 records, MongoDB finished properly, but SQL Server simply could complete the case and therefore it was not necessary to continue with the remaining cases.

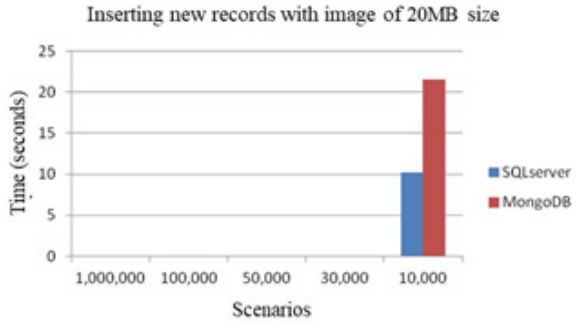

Figure 9. The database Insertion Operation with image of 20MB size.

Conclusions.

This paper described the study case of the database selection process for the big-data system to reduce health effects of air pollution in Ciudad Juárez, México. The project considers the use of a large number of inexpensive and strategically located sensors that measure different climatic irritants scattered in the environment. For the selection of the database, three key elements were considered: the nature of the data, the functionality of the database and a performance evaluation between the SQL and NoSQL databases, where several metrics were considered. After analyzing the results obtained, it was concluded that the database that best aligns with the project's objectives is the NoSQL MongoDB database.

References

[1] EPA (United States Environmental Protection Agency, 2017. [En línea]. Available: https://www.epa.gov/border2020. [Último acceso: 2018 09 28].

[2] EPA (United States Environmental Protection Agency), 2018. [En línea]. Available: https://www.epa.gov/newsreleases/us-epa-awards-84500-childrens-health-protection-along-us-mexico-border. [Último acceso: 28 09 2018].

[3] United States Environmental Protection Agency, 2018. [En línea]. Available: https://www.epa.gov/newsreleases/us-epa-anuncia-389000-para-proyectos-ambientales-lo-largo-de-la-frontera-arizona-sonora. [Último acceso: 15 09 2018].

[4] EPA (United States Environmental Protection Agency), 2018. [En línea]. Available: https://www.epa.gov/newsreleases/us-epa-announces-389000-environmental-projects-along-arizonasonora-border. [Último acceso: 10 09 2018].

[5] INEGI [En línea]. Available: http://cuentame.inegi.org.mx/monografias/informacion/chih/poblacion/. [Último acceso: 2018 08 10].

[6] H.J. Watson, «Tutorial: Big Data Analytics: Concepts, Technologies and Applications.,» Communications of the Association for Information Systems:, vol. 34, nº 65, 2014.

[7] T.G., "Usage-Driven Database Design: From Logical Data Modeling through Physical Schema Definition", New Jersey, Apress, 2017, p. 374.

[8] N. a. S.R. Umanath, Data Modeling and Database Design, Boston: Cengage Learning, 2014.

[9] MongoDB, "Top 5 Considerations When Evaluating", A MongoDB White Paper , New York, 2015.

[10] N.B.F.D.D. T. a. T.R. Schulz W., "Evaluation of relational and NoSQL database architectures to manage genomic annotations", Journal of Biomediacal Informatics, vol. 64, pp. Pages 288-295, 2016.