Application of Big Data tools in the analysis of visitors to museums and archaeological sites in the State of Oaxaca

Aplicación de herramientas de Big Data en el análisis de visitantes a museos y zonas arqueológicas del Estado de Oaxaca

MORALES-HERNÁNDEZ, Maricela†*, RAFAEL-PÉREZ, Eva, DIAZ-SARMIENTO, Bibiana and ALTAMIRANO-CABRERA, Marisol

Tecnológico Nacional de México/Instituto Tecnológico de Oaxaca. Avenida Ing. Víctor Bravo Ahuja No. 125 Esquina Calzada Tecnológico, C.P. 68030; Oaxaca de Juárez, Oaxaca. México

ID 1st Author: *Maricela, Morales-Hernández* / **ORC ID**: 0000-0002-3521-2041, **CVU CONACYT ID**: 731036

ID 1st Co-author: *Eva, Rafael-Pérez* / **ORC ID**: 0000-0003-2793-1254, **CVU CONACYT ID**: 905268

ID 2nd Co-author: *Bibina, Díaz-Sarmiento* / **ORC ID**: 0000-0003-4350-6311, **CVU CONACYT ID**: 820776

ID 3rd Co-author: *Marisol, Altamirano-Cabrera* / **ORC ID**: 0000-0001-5800-9655, **CVU CONACYT ID**: 657390

DOI: 10.35429/EJRC.2022.14.8.26.33 Received March 27, 2022; Accepted June 20, 2022

Abstract

This article is the result of the analysis of visitors to the archaeological zones of the state of Oaxaca, in the year 2021 applying Big Data tools. The original data was taken from the INAH (2021). The CSV (Comma Separated Values) file contains the columns: State, SIINAH Code, Acronyms, Work Center, Year, Month, Type of visitors, Number of visits, and Nationality. The objective is to illustrate the application of big data tools in the analysis of large volumes of data, in this case study, only a fragment of the data has been used, which corresponds to the Oaxaca State. The analysis was carried out with its own methodology based on the MAMBO methodology (Muñoz and Sánchez, 2019), in the applied methodology the following phases are identified: data acquisition, their management, the search for information in the data and their order and display. One contribution is the generation of a guide in which the reader will be able to identify the process of applying big data tools.

Big data, Analysis, Visualization

Resumen

Este artículo es el resultado del análisis de visitantes a las zonas arqueológicas del estado de Oaxaca, en el año 2021 aplicando herramientas de Big Data. Los datos originales se tomaron de la página del INAH(2021). El archivo CSV (Comma Separated Values), contiene las columnas: Estado, Clave SIINAH, Siglas, Centro de trabajo, Año, Mes, Tipo de visitantes, Número de visitas, y Nacionalidad. El objetivo es ilustrar la aplicación de herramientas de big data en el análisis de grandes volúmenes de datos, en este caso de estudio, se ha usado solo un fragmento de los datos, los cuales corresponden al estado de Oaxaca. El análisis se realizó con una metodología propia basada en la metodología MAMBO (Muñoz y Sánchez, 2019), en la metodología aplicada se identifican las siguientes fases: adquisición de los datos, el manejo de los mismos, la búsqueda de información en los datos y el orden y visualización de los mismos. Una contribución, es la generación de una guía en la que el lector podrá identificar el proceso de aplicación de herramientas de big data.

Big data, Análisis, Visualización

Citation: MORALES-HERNÁNDEZ, Maricela, RAFAEL-PÉREZ, Eva, DIAZ-SARMIENTO, Bibiana and ALTAMIRANO-CABRERA, Marisol. Application of Big Data tools in the analysis of visitors to museums and archaeological sites in the State of Oaxaca. ECORFAN Journal-Republic of Cameroon. 2022, 8-14: 26-33

* Correspondence to Author (e-mail: maricela.morales@itoaxaca.edu.mx)
† Researcher contributing first Author.

Introduction

In this article, a guide to the application of Big Data analysis tools for data analysis is presented, in this case, a public access dataset (INAH, 2021) has been selected, which stores information on visitor records to archaeological sites in Mexico during the year 2021. The CSV format file was fragmented, using in the analysis only the records corresponding to the state of Oaxaca.

For Skiena (2017) Big Data consists of massive amounts of rows (records) in a relatively small amount of columns (features). For this author, Big Data is often excessive to accurately fit a single model for a given problem.

So, a customized model can be trained to fit each user of the data. For Sathi (2012), there are two sources of Big Data, the first has to do with data that is generated within an organization, these may include: emails, PDF documents, events, blogs, and in general, any structured, unstructured or semi-structured data available in the organization. The second source of data is found outside the organization, in this group are free public data such as the one used in this research work, others are available for a fee and still others are available to business partners or specific customers. They can be found in social networks, information from competitors, governmental or non-profit organizations, among others.

According to Dietrich, Heller and Yang (2015) three attributes stand out that define the characteristics of Big Data: (1) large volume of data, (2) complexity of data types and structures, and (3) speed of creation and growth of new data. When a very large store of data is available, the data can be analyzed to gain knowledge from it. Understanding that knowledge discovery according to Singhal and Himanshu (2022) is defined as the method used to discover interesting, previously unknown and potentially useful patterns from a large amount of data.

For Big Data analysis, there are different tools in the market, most of them have a cost. For this case, and in order to exemplify the use of such tools, the services that AWS (Amazon Web Services) offers for the treatment and analysis of Big Data were specifically selected; those that were used are briefly explained below:

Redshift is a petabyte-scale data storage service fully managed in the AWS cloud (AWS, 2022). Candela *et al* (2011) define a petabyte as 1015 bytes, which implies a very large amount of data storage space. The Amazon Redshift data store is a collection of compute resources called nodes, which are organized into a group called a cluster. So each cluster runs an Amazon Redshift engine and can store one or more databases (AWS, 2022). So you are talking about a data warehouse that can seamlessly manage large volumes of data.

The first step in creating a data warehouse is to launch a set of nodes, called an Amazon Redshift cluster. After configuring the cluster, a dataset is loaded, and then data analysis queries are performed. Regardless of the size of the dataset, Amazon Redshift provides query performance.

For the present work, Redshift has been used to clean, prepare and load the data to be analyzed (creating the Data Warehouse), analyzing relevant information through SQL queries (Structured Query Language), which can give rise to interpretations of this information and thus be able to make decisions regarding what is found. Piñeiro (2015) defines SQL as a language that is used as a standard and includes instructions for data definition, manipulation and control. It is an official standard in the United States by ANSI (American National Standards Institute) and as an international standard by ISO (International Standards Organization).

On the other hand, Fernandez (2022) explains that S3 is the Amazon Web Services (AWS) object storage service of type PaaS (Platform as a Service), being a common solution to store data in the cloud in a secure, efficient and scalable way. Data in S3 is stored as objects within Buckets. An object is the basic unit of storage in S3, consisting of a file with an identifier and associated metadata. While a Bucket in Amazon S3 is nothing more than a high-level logical directory in which objects are located, each of them identified with a key.

For the present work, S3 has been used to store the CSV and JSON files with which the data analysis is performed.

Jensen (2020), describes QuickSight as a business analytics service with tailored analytics to derive business insights from data.

AWS (2022) describes it as a cloud-scale business intelligence service that can be used to deliver easy-to-understand insights to key people, wherever they are. QuickSight connects to data in the cloud and combines data from disparate sources. As a fully managed cloud-based service, it also provides enterprise-grade security, global availability and built-in redundancy.

For the case study, QuickSight is used to display an easy-to-understand dashboard with summarized and concrete information to support organizational decision making.

The article is composed of seven sections, which are described in the next paragraphs.

Methodology to be developed

For Big Data analysis there are different methodologies, although all of them agree on general steps and some in particular add or omit some. For example, Lin *et al* (2008) describe the CRISP-DM methodology (Cross Industry Standard Process for Data Mining), which consists of six phases as shown in Figure 1.

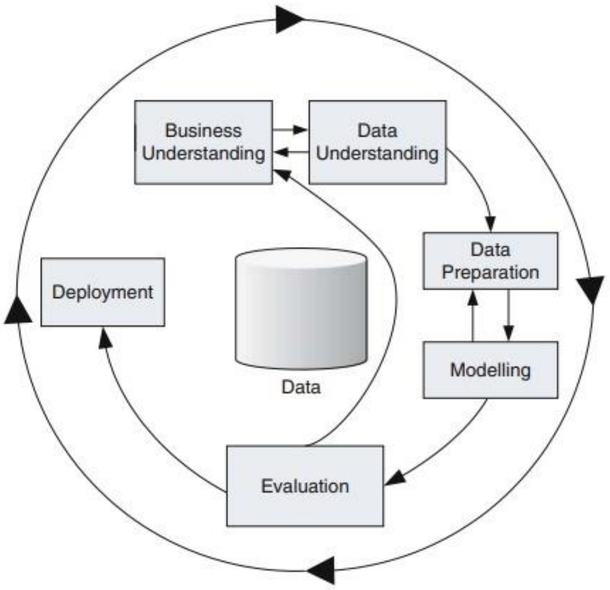


Figure 1 Cross-industry standard process for data mining
Lin et al (2008)

On the other hand, Muñoz and Sanchez (2019) propose a method they have called MAMBO (Meditate on the business, Acquire the data, Manage the data, Search the data, and finally, Sort and visualize), as can be seen in Figure 2.

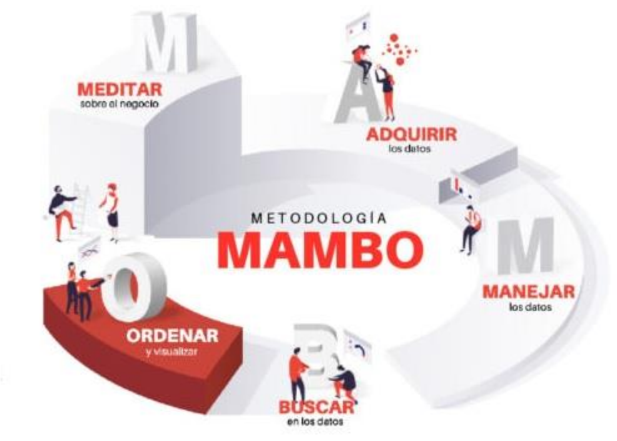


Figure 2 MAMBO Methodology
The art of measuring (2022)

Both methodologies have similar phases or stages, and for the present work only four steps were proposed based on the MAMBO methodology:

1. Acquire data
2. Manage the data
3. Searching the data
4. Sort and visualize

Development

Based on the MAMBO methodology, the steps followed for the analysis of visitors to the archaeological sites of the state of Oaxaca in the year 2021 are those described in the previous section. The following paragraphs describe the activities carried out in each of these steps.

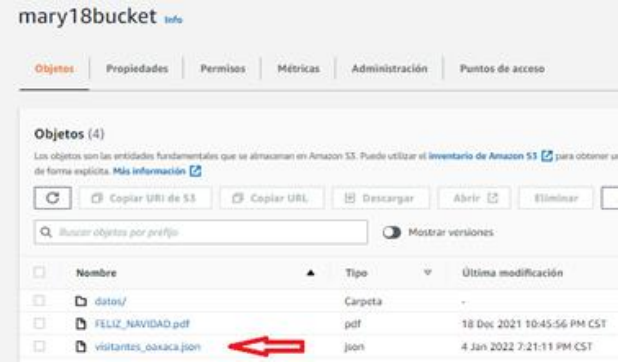
Acquire the data

The acquisition of data was done directly from a public data source maintained by INAH (2021), this data source contains information from all the archaeological sites administered by INAH throughout the Mexican Republic. Therefore, there was no need to compile data from the original sources.

Data management

As mentioned above, data in csv format were downloaded from the web page https://datos.gob.mx/busca/dataset/visitantes-a-museos-y-zonas-arqueologicas-abiertas-al-publico, concentrated in a general report of national and foreign visitors to Mexico's cultural sites open to the public during the year 2021.

From the acquired file, only the records for the state of Oaxaca were selected, converting the file to a "JSON" (JavaScript Object Notation) file. This first file was uploaded to an S3 bucket in AWS, as shown in Figure 3.



mary18bucket Info

Objetos (4)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el inventario de Amazon S3 para obtener u de forma explícita. Más información

Nombre	Tipo	Última modificación
datos/	Carpeta	-
FELIZ_NAVIDAD.pdf	pdf	18 Dec 2021 10:45:56 PM CST
visitantes_oaxaca.json	json	4 Jan 2022 7:21:11 PM CST

Figure 3 Files stored in the bucket created in S3. AWS Amazon

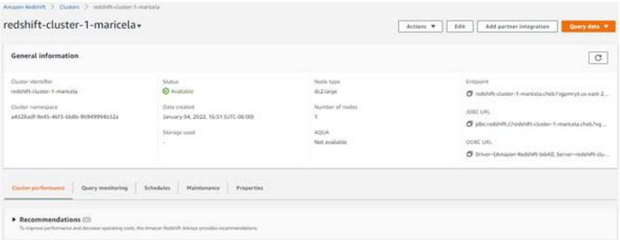
The original data file contains the columns shown in Table 1.

State
Key
SIINAH
Acronym
Work center
Year
Month
Type of visitors
Number of visits
Nationality

Table 1 Columns of the CSV file
INAH (2021)

Searching the data

In order to start the data analysis, the Redshift tool is used, so a cluster is created. For this work, the cluster redshift-cluster-1-maricela was created with the characteristics presented in Figure 4.



Amazon Redshift > Clusters > redshift-cluster-1-maricela

redshift-cluster-1-maricela

General information

Cluster identifier: redshift-cluster-1-maricela
Cluster namespace: a4328adf-9e45-46f3-bbdb-9b949994b32a
Status: Available
Date created: January 04, 2022, 16:51 (UTC-06:00)
Storage used: -
Node type: dc2.large
Number of nodes: 1
AQUA: Not available
Endpoint: redshift-cluster-1-maricela.cheb7vgpmryt.us-east-2...
JDBC URL: jdbc:redshift://redshift-cluster-1-maricela.cheb7vg...
ODBC URL: Driver={Amazon Redshift (x64)}; Server=redshift-clu...

Recommendations (0)
To improve performance and decrease operating costs, the Amazon Redshift Advisor provides recommendations.

Figure 4 Summary of the redshift-cluster-cluster-1-maricela cluster
AWS Amazon

Once the cluster is created, Redshift is accessed to import the JSON files to this service, for this, the table vistantes_oaxaca is created, to which the records that are in the JSON file are injected, the SQL instructions are detailed below:

```
create table visitantes_oaxaca
(
        estado varchar(max),
        clave varchar (max),
        siglas varchar(max),
        centro_t varchar (max),
        anio varchar (max),
        mes varchar (max),
        tipo_vis varchar(max),
        num_visitas varchar(max),
        nacionalidad varchar(max)
);
```

Once the table has been created, data injection is performed with the following instructions:

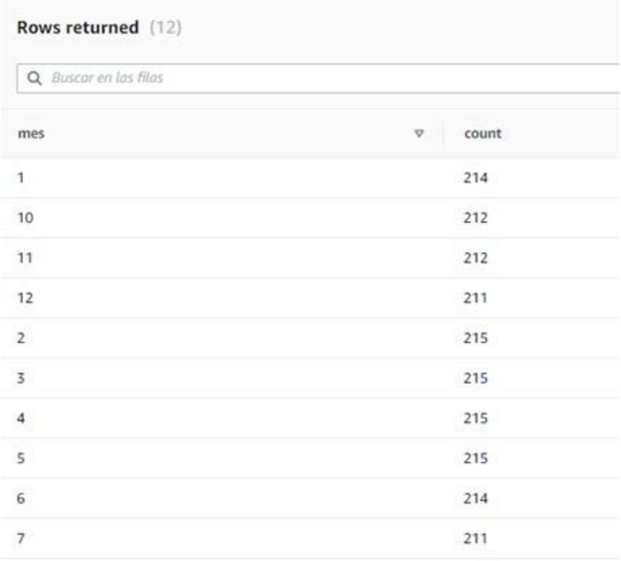
```
copy visitantes_oaxaca
from
's3://mary18bucket/visitantes_oaxaca.json'
iam_role
'arn:aws:iam::198272932368:role/redshift1'
format as json 'auto'
region 'us-east-2';
```

As a result of the data injection, a total of 2563 records were loaded into the vistantes_oaxaca table. It should be noted that it is required to have the necessary permissions to be able to write on the cluster that has been previously created. As well as the "iam" role is created in advance.

Once the data is loaded, different simple queries are made on the data and as an example some queries such as: the number of visitors per month, for this the following SQL instructions are used:

```
select mes,count(centro_t)
from visitantes_oaxaca
group by mes
order by mes;
```

The result of the consultation can be seen in Figure 5. This result shows that the flow of visitors is very similar in the different months of the year.



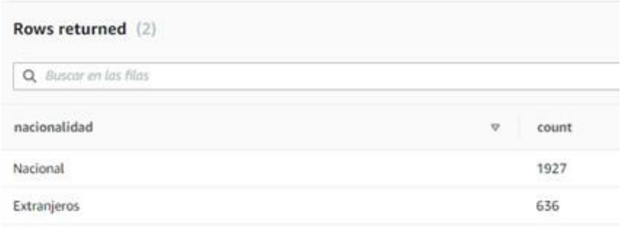
Rows returned (12)

Buscar en las filas

mes	count
1	214
10	212
11	212
12	211
2	215
3	215
4	215
5	215
6	214
7	211

Figure 5 Visitor query results by month for Oaxaca
AWS Redshift

Another simple query would be to know how many national and how many foreign visitors visited the archaeological zones of Oaxaca, the result can be seen in Figure 6.



Rows returned (2)

Buscar en las filas

nacionalidad	count
Nacional	1927
Extranjeros	636

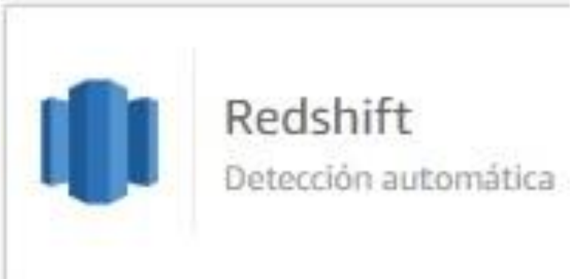
Figure 6 Visitors by nationality for Oaxaca.
AWS Redshift

Here it can be seen that national visitors predominate, so it can be assumed that in the archaeological zones of Oaxaca, the flow of visitors is mostly national visitors.

Sort and visualize

Once the Data Warehouse is generated in Redshift, it is exported to Quick Sight in order to present the results of the data analysis in a Control Panel using graphs that are easy to interpret by the user of the information.

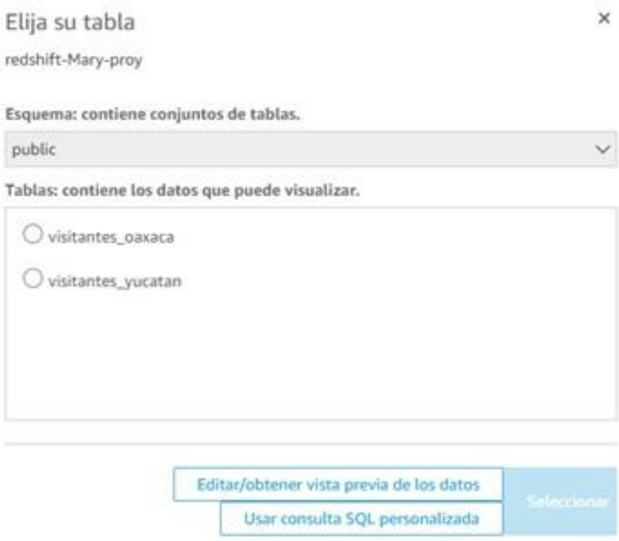
In order to use QuickSight it is required to have a user created specifically for the use of this tool, for the exercise a user was created previously. So the next step is to create a new analysis in QuickSight, once this new analysis is requested, a new data source must be created, so that the data loaded in Redshift can be used, Redshift (automatic detection) is selected, as can be seen in figure 7.



Redshift
Detección automática

Figure 7 Data source selection
QuickSight Amazon

In order to import the data previously loaded in Redshift, it is required to perform the proper configuration of the data source, done this, the available tables are shown as illustrated in Figure 8.



Elija su tabla
redshift-Mary-proy

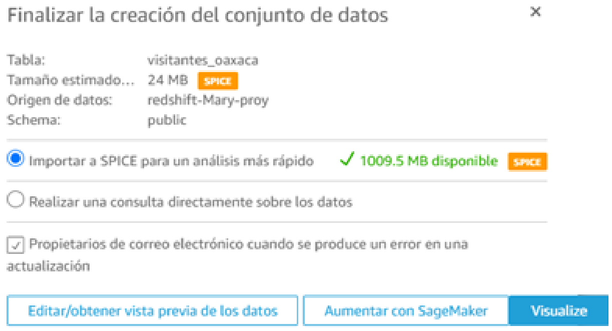
Esquema: contiene conjuntos de tablas.
public

Tablas: contiene los datos que puede visualizar.
○ visitantes_oaxaca
○ visitantes_yucatan

×
Editar/obtener vista previa de los datos
Usar consulta SQL personalizada
Seleccionar

Figure 8 Data source table selection
QuickSight Amazon

When the table is selected, the data is loaded into QuickSight and can now be visualized by clicking on the "Visualize" button, as shown in Figure 9.



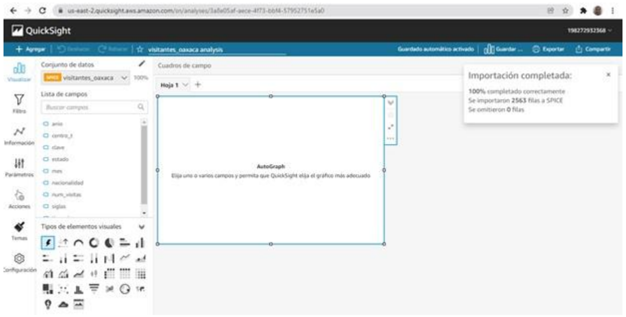
Finalizar la creación del conjunto de datos

Tabla:	visitantes_oaxaca
Tamaño estimado…	24 MB SPICE
Origen de datos:	redshift-Mary-proy
Schema:	public

● Importar a SPICE para un análisis más rápido ✓ 1009.5 MB disponible SPICE
○ Realizar una consulta directamente sobre los datos
☑ Propietarios de correo electrónico cuando se produce un error en una actualización

Figure 9 Successful data upload
QuickSight Amazon

The main panel for the analysis is shown in Figure 10, as can be seen, it notifies that the data import has been completed and shows the number of imported rows.



Importación completada:

100% completado correctamente
Se importaron **2563** filas a SPICE
Se omitieron **0** filas

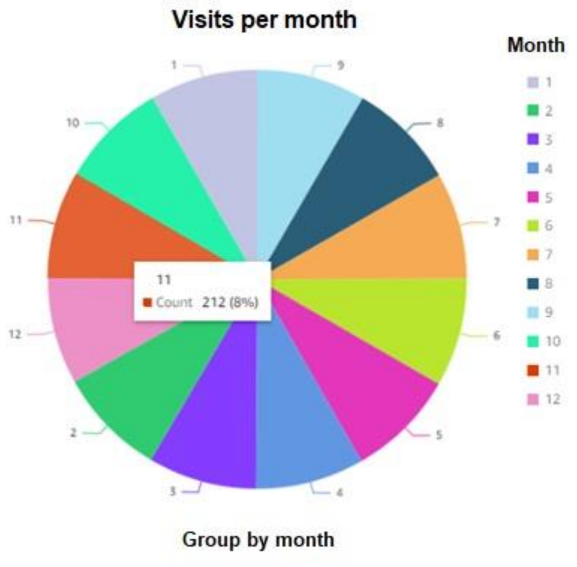
AutoGraph
Elija uno o varios campos y permita que QuickSight elija el gráfico más adecuado

Figure 10 Main panel of an analysis in *QuickSight Amazon*

From this point on, you can start visualizing the data through graphs that facilitate its analysis and allow the data user to make decisions about the information found. The results section shows the graphs obtained, as well as a brief interpretation of them.

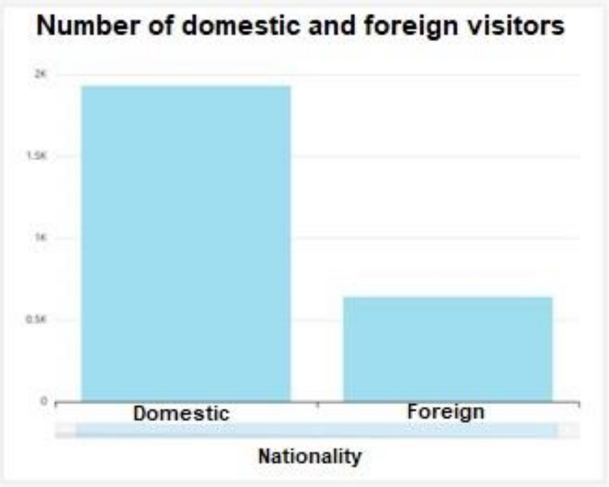
Results

As a result of the application of the AWS data analysis tools, the following graphs are presented. Graph 1 shows the number of visits per month, for example, the value shown means that in November there were 212 visits to the archaeological sites and museums in the state of Oaxaca. This could be an opportunity for INAH to look for strategies to increase the interest of the population in the knowledge of the pre-Hispanic cultures in the state.



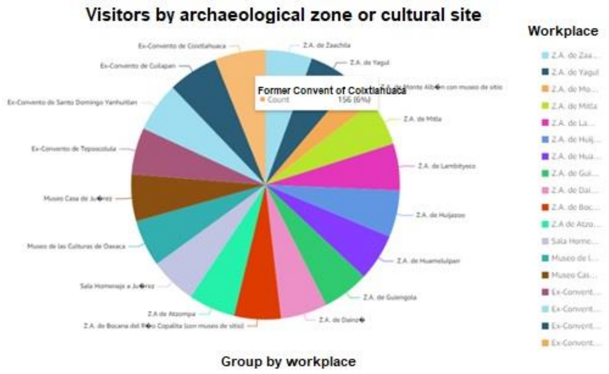
Graph 1 Visits per month
Own Elaboration

An analysis of the types of visitors was also made, so that, in Graph 2, a summary of how many visitors are nationals and how many are foreigners is presented, showing that nationals predominate.



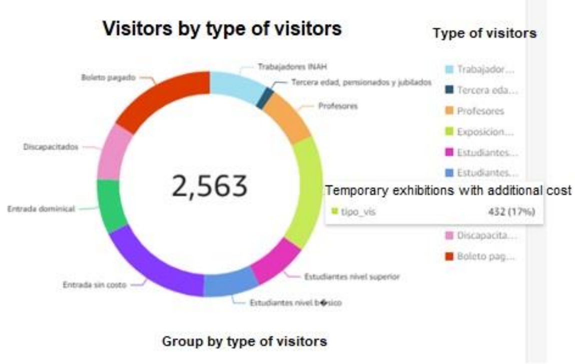
Graph 2 Number of domestic and foreign visitors
Own Elaboration

Graph 3 shows the number of visits per archaeological zone or museum or cultural space registered. In QuickSight, by sliding the pointer over the graph, the quantity for that category can be visualized, as can be seen in the same graph.



Graph 3 Visitors by archaeological zone or cultural site
Own Elaboration

Graph 4 shows the number of visits by type of visitor, and shows that there are people who visit cultural spaces and are willing to pay for their admission, as shown in the category of temporary exhibitions with additional cost and the paid ticket category, since these are the records that appear with the highest percentage in the graph.



Graph 4 Visits by type of visitor
Own Elaboration

Acknowledgements

This work is the result of the collaboration of the authors of this work, and we thank them for their support in their contributions to achieve this result; as well as the payment of the service required for the use of AWS Amazon tools, which as mentioned in previous paragraphs is not free, but it is important to demonstrate its advantages for those who find in them a possible solution to their approaches to information and knowledge from large volumes of data that their organizations have accumulated over time.

Conclusions

The use of Big Data analysis tools is becoming an imperative need for organizations if they need to gain knowledge from the data they have accumulated over time. The data itself has a value, as it can be consulted by the personnel who manage it; however, in order to make decisions it is important to treat it through a methodology. The problems regularly faced by the team analyzing the data are that they come from different sources and may have different formats. So, the first challenge is to homogenize them, the second challenge will be the data cleaning process, based on the experience of the data analyst, this process will be more or less complicated. And, once an adequate data warehouse is in place, the next step is to apply the tools for its analysis, management and visualization.

At the beginning of this paper, the objective was to illustrate the application of Big Data tools in the analysis of the data recorded by INAH of the visitors who come to the archaeological sites of the state of Oaxaca.

Therefore, it is hoped that this work can offer readers a guide to the transition in the use of data analysis tools; without forgetting that a methodology is required to obtain useful results, in this case a methodology based on the MAMBO methodology was proposed.

The experience obtained in this exercise is that the tools are useful to be able to deploy the knowledge found in the Big Data of the organizations. But, looking to the future, it is important that the data from its origin be as consistent as possible so that the result of the analysis can really provide organizations or companies with the certainty to make decisions based on their data.

In the future, comparative analyses with other tools can be established in order to analyze the convenience of using one or the other.

References

Amazon (19 de julio de 2022). Getting started with Amazon Redshift. https://docs.aws.amazon.com/redshift/latest/gsg/getting-started.html.

Candela, S., Castrillón, M., Domínguez, A., Doreste, L., Freire, D., Hernández, J., Lalchand, S. y Salgado, A. (2011). *Fundamentos de informática y programación para ingeniería.* Ediciones Paraninfo, S.A.

Dietrich, D., Heller, B. y Yang, B. (2015). Data Science & Big Data Analytics. Discovering, Analyzing, Visualizing an Presenting Data. John Wiley & Sons, Inc.

Fernández, O. (28 de noviembre de 2022). *Introducción a Amazon S3.* https://aprenderbigdata.com/amazon-s3/.

INAH. (2021). Visitantes a Museos y Zonas Arqueológicas abiertas al público. Datos y Recursos. https://datos.gob.mx/busca/dataset/visitantes-a-museos-y-zonas-arqueologicas-abiertas-al-publico.

Jensen, L. (2020). *Usability evaluation of AWS QuickSight for Real-time IOT data*. Royal Institute of Technology. Stockholm, Sweden

Lin, T. Y., Xie, Y., Wasilewska, A. y Liau, C.J. (2008). *Data Mining: Foundations and Practice Volumen 118 de Studies in Computational Intelligence.* Springer Science & Business Media.

Muñoz, G. y Sánchez, E. (2019). *Big Data como activo de negocio*. Ediciones Anaya Multimedia.

Piñeiro, J.M. (2015). *Lenguajes de definición y modificación de datos SQL*. Ediciones Paraninfo, S.A.

Sathi, A. (2012). *Big Data Analytics. Disruptive Technologies for Changing the Game*. IBM Corporation.

Singhal, N. y Himanshu (2022). *Electronic Systems and Intelligent Computing. A Review on Knowledge Discovery from Databases*. Pradeep Kumar Mallick, Akash Kumar Bhoi, Alfonso González-Briones, Prasant Kumar Pattnaik Editors.

Skiena, S. (2017). *Data Science Design Manual*. Springer.