Big Data and Data Analysis differences based on the different NoSQL Databases.

Hillary Noguera Sanchez.  
Universidad Fidelitas.  
Estudiante.San jose.  
e-mail: [hsanchez40651@ufide.ac.cr](mailto:hsanchez40651@ufide.ac.cr)

. . . . Mateo Bolaños Maya.  
Universidad Fidelitas.  
Estudiante.San Jose, Curridabat.  
e-mail: [teobm3677@gmail.com](mailto:teobm3677@gmail.com)

Big Data and difference Data Analysis based on the different NoSQL Databases.

# What is big Data?

Finally, we are going to define what Big Data is that with its name we can give us an idea of ​​what it means, they are quantities of massive data that when its size increases, its manipulation, management, and processing becomes difficult with the explanation above Defined, large companies take combinations of SQL and NoSQL environments to combat or resolve the large gaps they encounter when their data grows exponentially.

Is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

The definition of big data as the three V’s:

**Volume**: Organizations collect data from a variety of sources, including business transactions, smart (IoT) devices, industrial equipment, videos, social media and more. In the past, storing it would have been a problem – but cheaper storage on platforms like data lakes and Hadoop have eased the burden.

**Velocity**: With the growth in the Internet of Things, data streams in to businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

**Variety**: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions.

If we consider two additional dimensions when it comes to big data:

Variability: In addition to the increasing velocities and varieties of data, data flows are unpredictable – changing often and varying greatly. It’s challenging, but businesses need to know when something is trending in social media, and how to manage daily, seasonal and event-triggered peak data loads.

Veracity: Veracity refers to the quality of data. Because data comes from so many different sources, it’s difficult to link, match, cleanse and transform data across systems. Businesses need to connect and correlate relationships, hierarchies and multiple data linkages. Otherwise, their data can quickly spiral out of control.

The most importance of big data doesn’t revolve around how much data you have, but what you do with it. You can take data from any source and analyze it to find answers that enable.

* Cost reductions
* Time reduction
* New product development and optimized offerings.
* Smart decision making.
* Accomplish business.

Types of big Data.

* Structured
* Unstructured
* Semi-structured

Structured

The term *structured data* generally refers to data that has a defined length and format for big data. Examples of structured data include numbers, dates, and groups of words and numbers called *strings*. Most experts agree that this kind of data accounts for about 20 percent of the data that is out there. Structured data is the data you’re probably used to dealing with. It’s usually stored in a database.

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data.

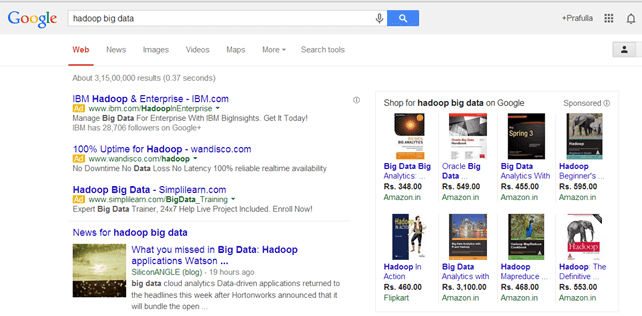
Examples of structured data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Employee\_ID** | **Employee\_Name** | **Gender** | **Department** | **Salary\_In\_lacs** |
| 2365 | Rajesh Kulkarni | Male | Finance | 650000 |
| 3398 | Pratibha Joshi | Female | Admin | 650000 |
| 7465 | Shushil Roy | Male | Admin | 500000 |
| 7500 | Shubhojit Das | Male | Finance | 500000 |
| 7699 | Priya Sane | Female | Finance | 550000 |

## Unstructured

Any data with unknown orm or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.

Example of unstructured



Ilustración

## Semi-structured

A simple definition is data that can’t be organized in relational databases or doesn’t have a strict structural framework, yet does have some structural properties or loose organizational framework. Semi-structured data includes text that is organized by subject or topic or fit into a hierarchical programming language, yet the text within is open-ended, having no structure itself.

Emails, for example, are semi-structured by *Sender, Recipient, Subject, Date,* etc., or with the help of [machine learning](http://monkeylearn.com/machine-learning/), are automatically categorized into folders, like *Inbox, Spam, Promotions,* etc.

<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>

<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>

<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>

<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>

<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>

A comparison of NoSql database management system and models

Gráfico

Descripción generada automáticamente con confianza media

Ilustración

When most people think of a database, they often envision the traditional relational database model that involves tables made up of rows and columns. While relational database management systems still handle the lion’s share of data on the internet, alternative data models have become more common in recent years as developers have sought workarounds to the relational model’s limitations. These non-relational database models, each with their own unique advantages, disadvantages, and use cases, have come to be categorized as *NoSQL databases*.

# Limitation of relational databases.

*Databases* are logically modeled clusters of information, or *data*. A *database management system* (DBMS), meanwhile, is a computer program that interacts with a database. A DBMS allows you to control access to a database, write data, run queries, and perform any other tasks related to database management. Although database management systems are often referred to as “databases,” the two terms are not exactly interchangeable.

Historically, the relational model has been the most widely used approach for managing data, and to this day [many of the most popular database management systems implement the relational model](https://db-engines.com/en/ranking). However, the relational model presents several limitations that can be problematic in certain use cases.

Another limitation presented by RDBMSs is that the relational model was designed to manage *structured data*, or data that aligns with a predefined data type or is at least organized in some predetermined way, making it easily sortable and searchable.

# NoSQL

Despite these different underlying data models, most NoSQL databases share several characteristics. For one, NoSQL databases are typically designed to maximize availability at the expense of consistency. In this sense, consistency refers to the idea that any read operation will return the most recent data written to the database. In a distributed database designed for strong consistency, any data written to one node will be immediately available on all other nodes; otherwise, an error will occur.

Conversely, NoSQL databases oftentimes aim for *eventual consistency*. This means that newly written data is made available on other nodes in the database eventually (usually in a matter of a few milliseconds), though not necessarily immediately. This has the benefit of improving the availability of one’s data: even though you may not see the very latest data written, you can still view an earlier version of it instead of receiving an error.

In the beginning, we will take a few simple definitions to better understand the final concept of NoSQL databases based on Big Data analysis.

We are going to review what a SQL database is. An SQL database is a store where information is systematically stored, commonly in the form of tables, all these tables are created with a specific language for their manipulation, definition and execution that is called SQL (Structured Query Language) these data by Having the quality of being structured, a data analyst can easily with a little study of the data controller, manipulate and execute entire relational databases.

Like its counterpart (NoSQL) we are going to highlight a few advantages of this type of database.

Management: This type of database has been on the market for a long time; therefore, it has greater support and greater tools for its management.

Integrity: the data in these databases must follow certain rules for their formalization since as they are structured, they must eliminate certain concepts such as redundancy of the data, this is an advantage since when making the tables the data is taken for granted to be found, making them easier to analyze

Now, a NoSQL database does not have structured data, which means that it is a cluster of information that does not have a very logical structure. Some data that can be saved are the following:

* Word documents
* Pdf document
* Emails: these are defined as semi-structured since the message text is unstructured and the analysis tools cannot categorize or identify them (human intervention is required)
* videos
* audios
* social media posts

All this information regardless of the format that is sent has certain advantages, which are what differentiates them from SQL databases, it should be emphasized that this is a different model that only seeks to solve problems that relational databases have.

* Performance
* Scalability
* Changes in the schemes: changes in the data can be easily made, since they are not structured, the image of the base itself will not change much, this is a double-edged sword, then, you must be very careful since a person who does not knowing the purpose of the database could change your focus.
* They run on low resource machines.
* Availability: with the replication between several servers, reliable and secure access can be guaranteed since if any server fails, another will be backing it up

[1] [2]

## Databases Key value

This type of database is based on two columns differentiated one by a key that must be unique and, on another column, where a value or values ​​can be found, both normal and embedded, these data after their classification are stored in a type of dictionary which makes them easier to read and write, this helps them to be more horizontally scalable and better in real-time operations thus facilitating the query of a user in a cluster of large amounts of data. It should be noted that there are few functionalities and that if we are not very careful, we can have a lack of consistency since the database does not contemplate the structure of the stored data.

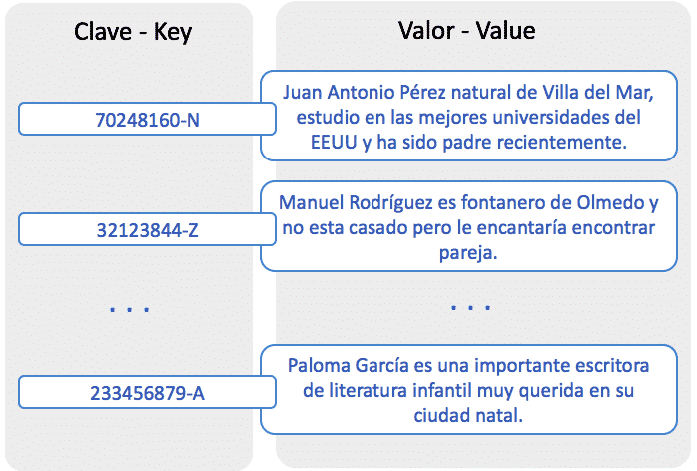


Ilustración 3 Base de datos clave - valor

## Column databases

As its name indicates, the data is entered in columns, its main objective is focused on business intelligence, that is, it is taken for granted that they are large companies that seek to manipulate their data for a specific purpose either for internal measurements or for simple marketing to know how your product behaves in a specific market before customers, so it seeks to perform efficiency in the writing and processing of data to give a quick response to a query.

As an advantage

* Facilitates the calculation of aggregated data

As Disadvantage

* Has poor write performance
* It does not offer JOINS due to its read-only approach.



Ilustración Base de datos columnares

## Documentary databases

This type of database is handled with types of XML, JSON, BSON documents and they sit in a specific key as well as being very versatile since they easily adapt to any data extension.

They are like the Key - value databases have the same speed, but they have greater flexibility thanks to the fact that they contain semi-structured data, their access is through already defined indexes.

Certain advantages include

* Easily transformable from the relational...
* It also allows complex searches and operations within documents.
* They are easy to develop and program.
* like relational databases that handle objects like Oracle, PostgreSQL, DB2.



Ilustración 5 Bases de datos documentales

[2] [3]

## Graph-oriented databases

In case it is not known that it is a graph, it is a collection of points that are related through lines, these have a small explanation of the relationship that leads (Juan ------- friend of ------ --- Carlos). This allows users to model data more intuitively since the shape of the graph is shown and thus it can be easily deduced if what is being sought is being done well or not. It is used a lot in the connections of patterns for commercial uses, problems with the products, common patterns, anomalies, communities, the customer's view corresponding to a point of view, etc ... These graphs can process the data and thus know what element is with more connections than other shipments in the same graph, it is used too much in massive companies that focus on social networks, technological implementations and both normal purchases and daily consumption (eBay, Walmart, Telenor, UBS, Cisco, Hewlett-Packard, or Lufthansa)

An example from the real world is to use a graph model to be able to go from point A to B so the intersections are nodes, and each street is a line that leads to the destination (Point B)

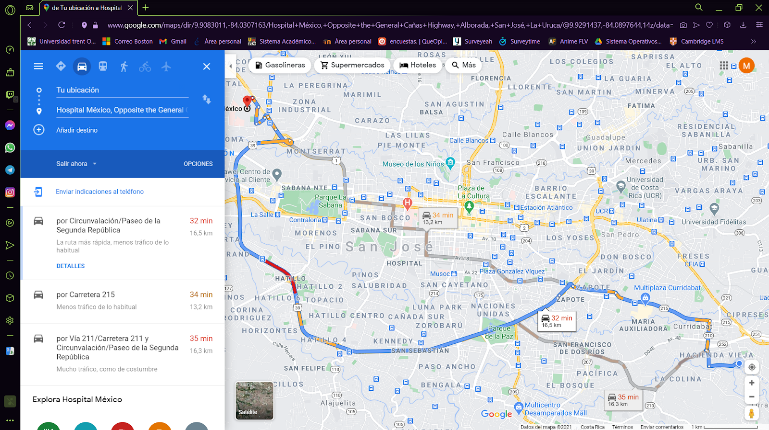


Ilustración 6 Ejemplo de base de datos grafos

Now the advantage that we can get from this database of graphs based on the opinion of many sites on the internet is that this is the perfect combination between programming and the database, in addition to having the visualization of how the graph.

## Multidimensional databases

are databases that are assimilated to a hypercube where each angle serves to access information that with OLAP implementations allow discovering new ways of using their data, even, applying data mining, they can be in a single table (they can even be databases contained in a single table) and extract several dimensions to filter or retrieve data in huge amounts of data to generate reports depending on what is needed. They are normally used in data warehouses to facilitate the analysis and abstraction of data in the record time of a given context.

The most notable advantage is the instant analysis of massive amounts of data.

[4] [5]

# bibliografias

[[1].R. España, "Diferencias entre datos estructurados y datos no estructurados", *Agenciab12.com*, 2020. [Online]. Available: https://agenciab12.com/noticia/diferencia-datos-estructurados-y-datos-no-estructurados. [Accessed: 14- Mar- 2021].](#Referencia_1)

[[2].R. España, "¿Qué es NoSQL y qué relación tiene con el Big Data?", *Agenciab12.com*, 2021. [Online]. Available: https://agenciab12.com/noticia/que-es-nosql-relacion-con-big-data. [Accessed: 14- Mar- 2021].](file:///C:\Users\teobm\OneDrive\Escritorio\Bases%20de%20datos%20no%20sql\investigacion\%5b2%5d.R.%20España,%20%22¿Qué%20es%20NoSQL%20y%20qué%20relación%20tiene%20con%20el%20Big%20Data%3f%22,%20Agenciab12.com,%202021.%20%5bOnline%5d.%20Available:%20https:\\agenciab12.com\noticia\que-es-nosql-relacion-con-big-data.%20%5bAccessed:%2014-%20Mar-%202021%5d.)

[[2].R. España, "¿Qué es NoSQL y qué relación tiene con el Big Data?", Agenciab12.com, 2021. [Online]. Available: https://agenciab12.com/noticia/que-es-nosql-relacion-con-big-data. [Accessed: 14- Mar- 2021].](file:///C:\Users\teobm\OneDrive\Escritorio\Bases%20de%20datos%20no%20sql\investigacion\%5b2%5d.R.%20España,%20%22¿Qué%20es%20NoSQL%20y%20qué%20relación%20tiene%20con%20el%20Big%20Data%3f%22,%20Agenciab12.com,%202021.%20%5bOnline%5d.%20Available:%20https:\\agenciab12.com\noticia\que-es-nosql-relacion-con-big-data.%20%5bAccessed:%2014-%20Mar-%202021%5d.)

[[4]."¿Qué es una base de datos orientada a grafos? | Oracle España", *Oracle.com*, 2021. [Online]. Available: https://www.oracle.com/es/big-data/what-is-graph-database/#:~:text=orientadas%20a%20grafos-,Bases%20de%20datos%20orientadas%20a%20grafos,las%20bases%20de%20datos%20relacionales. [Accessed: 16- Mar- 2021].](file:///C:\Users\teobm\OneDrive\Escritorio\Bases%20de%20datos%20no%20sql\investigacion\%5b4%5d.%22¿Qué%20es%20una%20base%20de%20datos%20orientada%20a%20grafos%3f%20|%20Oracle%20España%22,%20Oracle.com,%202021.%20%5bOnline%5d.%20Available:%20https:\www.oracle.com\es\big-data\what-is-graph-database\)

[[4]."¿Qué es una base de datos orientada a grafos? | Oracle España", Oracle.com, 2021. [Online]. Available: https://www.oracle.com/es/big-data/what-is-graph-database/#:~:text=orientadas%20a%20grafos-,Bases%20de%20datos%20orientadas%20a%20grafos,las%20bases%20de%20datos%20relacionales. [Accessed: 16- Mar- 2021].](#Referencia_5)

[6].USA, Big Data [https://www.sas.com/en\_us/insights/big-data/what-is-big-data.html. [Acces-17](https://www.sas.com/en_us/insights/big-data/what-is-big-data.html.%20%5bAcces-17)-march-2021]

[7]. USA, What is BIG DATA? <https://www.guru99.com/what-is-big-data.html#2> [Access-18-march-2021]

[8]. Mark Drake on 9 agust,2019, <https://www.digitalocean.com/community/tutorials/a-comparison-of-nosql-database-management-systems-and-models> [Access-16-march-2021]

[9]. Judith S. Hurwitz, Structured data in a big data environment, <https://www.dummies.com/programming/big-data/engineering/structured-data-in-a-big-data-environment/#:~:text=The%20term%20structured%20data%20generally,words%20and%20numbers%20called%20strings.&text=Structured%20data%20is%20the%20data,usually%20stored%20in%20a%20database>. [Access-18-march-2021]

[10]. USA, What is Semi-Structured Data & How to Analyze it, <https://monkeylearn.com/blog/semi-structured-data/> , [Access-18-march-2021]

# Imágenes

[1]D. Calvo, *Key-value database*. 2017.

[2]j. moreno,Columnar databases. 2017.

[3]H. Hernández, *Columnar databases*. 2019.

[4]Creative Commons, *dimensional cube*. 2021.