New Trends in Databases to NonSQL Databases

6

Antonio Sarasa-Cabezuelo
https://orcid.org/0000-0003-3698-7954
Facultad de Informática, Universidad Complutense de Madrid, Spain

INTRODUCTION

In recent decades, the relational model has been the mechanism of persistence of information most used in the computer applications. It provides a mature and stable model to represent the data, and also offers a widely used standard query and manipulation (SQL) language.

1. However, the appearance of the "Big Data" phenomenon has meant a change in the storage and information processing needs. This new context is characterized by: a) enormous amounts of information are available in heterogeneous formats and types, b) information must be processed almost in real time, and c) data models evolve periodically.
2. This requires a greater processing capacity, greater flexibility in terms of the types of information data that can be stored and greater flexibility in the data models that facilitate their adaptation.
3. Relational databases have limitations to respond to these needs in an optimal way given that:
    - The solution to achieve more processing capacity is to use more powerful machines (vertical scaling). It is not economically advantageous.
    - The evolution of the data models implies a modification of the database schemas. This causes problems of maintenance and consistency of previously stored information.
    - The storage of heterogeneous and complex data requires the transformation of information in order to store complex data in basic data types (the only ones that manage relational databases).

For these reasons, some companies such as Google or Amazon decided to create new database models (different from the relational model) that solve the needs raised in the context of Big Data without the limitations of relational databases. These new models are the origin of the so-called NonSQL Databases.

Currently, NonSQL databases have been constituted as an alternative mechanism to the relational model and its use is widely extended.

The main objective of this chapter's proposal is to introduce the NonSQL databases. Specifically:

- To show the processing and persistence needs of information in the Big Data.
- To describe the main characteristics and the conceptual basis of the NoSQL databases.
- To show how the NoSQL databases solve the needs raised in the Big Data
- To describe the main families of NoSQL databases and the data models on which they are based.

The structure of the chapter is as follows. Section 1 describes the storage needs that appear in the Big Data. Section 2 presents the limitations of the databases to cover the needs of Big Data. Next, the

DOI: 10.4018/978-1-7998-3479-3.ch054

NoSQL databases and their main characteristics are introduced. In the following section, persistence and distribution models are discussed. Next, the CAP theorem is presented. Finally, the trends in the NoSQL world are discussed, and a set of conclusions is established.

BACKGROUND

Big Data is a technological challenge that arises due to the confluence of two phenomena (Chen et al, 2014). On the one hand, there is a rapid increase in the storage capacity of the hardware and a reduction in its cost (in the 1980s it was possible to handle data sizes of 0.02 EB, in 2013 they handled sizes of 4.4 ZB, and in 2020 it is estimated that it will be possible to handle sizes of 44 ZB). And on the other hand, some companies and institutions begin to generate large amounts of data in a very fast way (for example, Google generates around 20 PB of data daily and Facebook 500 TB). The possibility of storing this data by having storage with sufficient capacity makes companies start to exploit the stored data for economic or strategic purposes, and to obtain benefits from the results of the analyzes. Thus, an infinite loop begins in which more and more data are generated, which forces us to develop media with greater storage capacity. When initial storage needs have been met, they are no longer able to meet current needs (since the amounts of data generated increase exponentially), which makes it necessary to create new storage media with greater capacity.

Big Data affects different technical aspects related to the storage and processing of information (Khan et al, 2014):

- The types of data that need to be managed are very different in nature, they require the storage of data that can be structured, semi-structured or simply without structure.
- The quantities of data that need to be processed are increasing, so that the necessary processing capacity also increases very rapidly. In this sense, observe that another requirement is the need for the processing to be done almost in real time.
- The data sources that need to be processed can change at any time, which is why we need sufficiently flexible information organization schemes that adapt to different information structures.
- The new context in terms of data types and number of data, makes it necessary to develop new techniques for analyzing information and interpreting the results thereof.

Thus, the technological areas affected by these new needs are varied. However, this chapter will focus on changes in information storage systems. Next, it will be analyzed the limitations of the relationship databases to cover the new needs of Big Data in terms of storage of information.

During the last decades, the relational model has been the most widespread way of carrying out the persistence of information. However, it presents important limitations to cover the main needs arising in the Big Data (Li et al, 2013). Then, each of them will be analyzed (Jacobs, 2009):

- The problem of the data schema (Zhao et al, 2014). In the relational model, a task prior to the storage of the data consists of creating a schema of the type of information to be stored. In this way, they describe the types of data are admitted. If it neccesary to store other types of data then it will have to make changes to the scheme. However, these changes often introduce anomalies in the stored data (for example, in terms of the relational model there will probably be rows with many columns with null values). In this sense, it is said that the information stored in a relational