NoSQL Database Systems for Big Data Scalability and Replication

Posted on September 19, 2019

View this essay online:
https://academic-master.com/nosql-database-systems-for-big-data-scalability-and-replication/

Database

Data refers to a set of values of both quantitative and qualitative variables. While information and data are used interchangeably, the extent to which a set of data is detailed and informative to a person depends on the degree to which it is unexpected by that individual. Shannon entropy can be used to characterize the volume of information contained in a data stream. In most cases, the data concept is mostly associated with scientific research.

However, data is obtained by a wide range of institutions and organizations, including businesses, such as revenue, sales data, stock prices and profits, among others; governments, for instance, unemployment rates, crime rates, literacy rates and levels; and also non-governmental organizations, i.e., censuses on the number of homeless populations by non-profit institutions.

Typically, data is usually collected, measured, assessed and analyzed. Consequently, after these steps, the data can be visualized using images, graphs, and other analytical tools. In general, concepts, data refers to the reality that some of the information or knowledge in existence is coded or represented in some form that is appropriate for better application of processing. Unprocessed or raw data is a collection of information in the way of numbers or characters before it has undergone cleansing and correction by researchers (“DATABASE Editorial Board 2009”, 2016). This type of data requires rectifications, which ensure the removal of outliers or visible parts or data entry mistakes, as is the case, for instance, with a thermometer recording from an outdoor Arctic place reading a tropical temperature.

Data processing is done in stages, and the processed resultant data from step one can be regarded as raw data when presented in the next stage. Field information is the unprocessed data obtained in an uncontrolled or unaltered “in situ” environment, while experimental data is the information picked within the concepts of scientific research and study through observation and recording. Data, in most instances, is termed the new oil in the digital economy. Consequently, data is information, while its collection is described as a database. This is a well-organized collection of data in various forms, such as tables, schemas, reports, queries, and views, among other objects (Mohanty, Jagadeesh & Srivatsa, 2013).

While the multi-media capacities are increasingly becoming a standard phenomenon for computers of varying sizes, the databases are shifting to multi-media as well. The advanced DBMS allows the keeping and retrieval of types of data such as spreadsheets, business graphs, clip art or art clippings, sound, pictures, and videos, among others. These new types of data are referred to as objects. Database management systems, abbreviated as DBMS, are computer software applications that interact with other implementations, users, and the database itself in capturing and analyzing data. The primary objective of a DBMS is to develop it in a way that facilitates the creation, definition, querying, updating, and administration of databases (Krishnan, n.d.). DBMSs are, at times, loosely termed databases.

Knowledge of databases is dependent on the individual's experience and may date back a couple of years or even back to the 1980s and 1990s. Some people, therefore, may not be aware of the old timers and types of databases. Most of the major database systems have their roots dating back to the mid and late 1970s. This, therefore, emphasizes the fact that the database market is yet to be fully mature. The databases available in the market also compete stiffly. Most enterprise database systems offer a variety of features, such as precise and effective backups and high availability of extended cloud services ("NoSQL Database: Cassandra is a Better Option to Handle Big Data", 2016). These database systems vary in price. In this essay, we will focus majorly on the NoSQL Database.

Digital innovations are advancing tremendously due to new technologies. This results in a constant rise in customers' expectations. For almost every company, this means that customer experiences and behaviour have rapidly become some of the most significant competitive advantages ("NoSQL in Action – A New Pathway to Database", 2016). In response to this, most institutions are initiating and implementing the use of NoSQL databases with the aim of delivering extraordinary experiences to consumer demands.

While Altoros remains the company behind some of the world's significant NoSQL innovations, it is known from experience that NoSQL was created differently, hence differing in quality. The company points out the differences between Cassandra or DataStax Enterprise, Couchbase Server, and MongoDB and further offers explanations on how each of them is likely to impact the success or failure of any business project. In its report, the critical areas that the company focuses on include the most significant architectural variations, typical application instances, and functional problems, primary evaluation criteria across the operation, availability, as well as ease of use ("NoSQL in Action – A New Pathway to Database", 2016). The report finally addresses three database functions: reading, writing, and querying data. It also discusses data consistency, deployments, and the support of numerous data centres.

In its report, Altoros evaluates the three responses from various angles of view, including availability, performance, installation and maintenance ease, fault tolerance, recovery, replication, and scalability, among many more. This is a different approach from the NoSQL comparisons, which mainly focus on just a single or two dimensions.

There are different types of NoSQL Databases. These include:

- Document databases tend to pair each key with a complicated data structure referred to as a document. Differing high-value pairs, nested certificates or major-array pairs are usually contained in documents.
- Graph databases are used to store information on data networks, such as social connections. Giraph and Neo4J are examples of graph stores.
- The most straightforward NoSQL databases are the key-value stores. Each of the items in the database is kept as an attribute title or key together with its value. The various examples of key-value stores include Berkeley and Riak DB. Some of the key-value stores allow each value to have types, such as integers, which enhance functionality. An example of these types of key-value stores includes Redis.
- Wide-column stores such as HBase and Cassandra have optimizations for queries over large datasets. They store data together in the form of columns rather than rows.

There are numerous advantages associated with NoSQL. When the comparison is conducted between NoSQL and relational databases, it is notable that NoSQL databases offer superior performance and are more scalable. Consequently, their data models account for several concerns which are, by design, not addressed by the relational model. These concerns include the significant volumes of tremendously changing unstructured, semi-structured and structured data, quick schema iteration, agile sprints, and frequent code pushes (Rodriguez, Malgapo, Quick & Huang, 2017). There are concerns about object-oriented programming, which is flexible and easy to use, and the geographically distributed scale-out architecture rather than the costly monolithic architecture.

Dynamic Schemas

In relational databases, schemas are required to be defined before data is added. For instance, one may wish to store information regarding their customers, such as first and last names, phone numbers, city, and state, as well as the address—a SQL database requires that it is aware of what is to be stored in advance. This is poorly incorporated with the agile development concepts since every time one finalizes new features, the schema of the database will often require changes (Krishnan, n.d.). This means that in case one decides, a couple of iterations into advancement, that they desire to store their customers' favourite items together with their phone numbers and addresses, they will be required to add that column to the database and later transfer the whole database to the newly created schema.

In instances where the database is extensive, the process can be slow and time-consuming since it entails significant downtime. In cases where data changes frequently occur in the application store, downtime may also be constant due to rapid iteration. Relational databases also restrict any attempts to efficiently address entirely unstructured data or that which is unknown in advance.

NoSQL, on the other hand, is structured in a way that facilitates the insertion of data without the need for a predefined schema. This makes it easy to address significant application changes in actual time without constant service interruptions or downtime. This means that the development is fast and reliable regarding code integration and requires minimal database administrator time. To emphasize data quality controls such as the presence of data types, specified fields, and permissible values, the developers have typically had to include application-side code (Feinleib, 2014). More advanced NoSQL databases facilitate the validation of rules applicable within the database, thus permitting the users to implement governance across data while at the same time maintaining the agility advantages of the dynamic schema.

Auto-Sharding

Due to their structural characteristic, relational databases typically scale vertically. An individual server is required to host the whole database to enable acceptable functionality for transactions and cross-table joins. This quickly becomes expensive, placing limits on the scale and also establishing a relatively small number of failure areas for database infrastructure. The response to support the tremendously advancing applications is to scale horizontally through the addition of servers rather than concentrating more capacity on an individual server.

On the other hand, NoSQL databases usually facilitate auto-sharding. This means that they are automatically and natively spread across an arbitrary network of servers without the need for the system to have prior knowledge of the content of the server pool. There is an automatic balancing of the data query and loading across the servers (“DATABASE Editorial Board 2009”, 2016). In instances where the server experiences downtime, it can be replaced quickly and transparently without disruptions to the application.

This is significantly eased by cloud computing, with suppliers such as Amazon Web Services offering virtually unlimited capacities upon demand while still taking into account all the appropriate administration roles. Developers are no longer required to create sophisticated, costly platforms to enhance their applications and can concentrate on formulating the application code (Krishnan, n.d.). Commodity servers are capable of providing processing and storage capabilities similar to those of individual high-end servers at a subsidized price.

Replication

Most of the NoSQL databases also facilitate automated replication of databases with the aim of maintaining availability in case of planned maintenance or outages. More advanced NoSQL databases are entirely self-healing, providing automatic failover and recovery as well as the capacity to share the database across various geographic places to cope with local failures and enable data localization (Neeraj, n.d.). NoSQL databases, as compared to relational databases, lack any requirements for

different applications or costly add-ons to apply the replication.

Integrated Caching

SQL database systems obtain a caching tier from some commodities. The methods can substantially enhance read performance but have no ability to improve the write performance. Consequently, they result in the addition of operational complexity in the system deployments. In an application where reads are dominant, a distributed cache can be considered, but in instances where the app has a modest write volume, the overall experience of the end users may not be enhanced by a distributed cache. Furthermore, cache invalidation management will be involved. Most of the NoSQL databases have exhibited excellent capabilities in integrated caching, thus eliminating the need for a different caching layer while maintaining frequently retrieved data in the system's memory as much as possible. Entirely managed, integrated in-memory database management layers are provided by some NoSQL databases for workloads in need of high throughput and lowest latency (Deshpande, 2014).

Whenever we talk about the major types of databases, we cannot fail to mention big data. So, what is significant data? For enterprises of all sizes, there has been a shift in data management from the necessary competencies to the critical differentiators with the capability of determining the market winners as well as the has-beens. Fortune 1000 corporations and government institutions have begun reaping benefits from the innovations and advancements of web pioneers. These institutions are formulating new programs while still reevaluating and reassessing the already established strategies and plans with the aim of evaluating the ways through which they can transform their enterprises using Big Data (Roza, 2014).

While implementing this initiative, it is notable that Big Data does not entail a single technique, technology or program. On the contrary, this is a trend or pattern across numerous areas of business and technological innovations. In other words, Big Data refers to the innovations and initiatives involving diversified, fast-developing and changing as well as massive data for current technological developments, techniques/skills, as well as the infrastructure to address them effectively and efficiently (Neeraj, n.d.). From a different perspective, data is of a significant volume, velocity, or variety.

In the contemporary world, new advanced technologies facilitate the realization of the value associated with big data. For instance, retailers can track the user's web clicks and views, thus identifying the behavioural patterns and trends that enhance the advertisements, pricing, and stocking. Household energy usage rates and levels can be captured using utilities, thus making efficient predictions on outages and incentivizing proper energy use. Google and even governments can detect and keep track of disease outbreaks through social media channels and signals. The gas and oil industries can take into account the output made by sensors attached to their drilling machines, thus creating more effective, efficient, and safe drilling plans and decisions. It is impractical

and impossible to manage big data, which is described as very large and complex, using traditional software tools (Feinleib, 2014). In particular, Big Data tends to relate to the creation, storage, retrieval, as well as analysis of data, which are typically remarkable regarding velocity, volume, and variety:

- Velocity. Ad impressions and clickstreams collect the user character and behaviour of millions of activities every second; high-frequency trading algorithms in the stock market reflect the changes in the market within microseconds; also, computer-to-computer processes share data across numerous devices; sensors and infrastructure generate considerable log information in actual time; online gaming tools facilitate the applications by millions of concurrent gamers, with each generating multiple and differing inputs each second.
- Volume. A typical computer had the possibility of having 10 gigabytes of storage back in 2000. Currently, the Facebook website collects 500 terabytes of new information daily; the proliferation of digital phones, the information they generate and use; a Boeing 737 is capable of producing 240 terabytes of flight information while on a flight across America; sensors attached to everyday objects have the potential to generate billions of new, continuously updated data feeds comprising location, environmental, and other types of information, including videos.
- Variety. Big Data does not only entail dates, numbers, and strings. It includes geospatial data, audio and video, 3D data, and unstructured texts such as log files and social media. Older database systems are structured in a manner that they address relatively smaller volumes of structured information, limited updates, or predictable and consistent data. Consequently, the outdated database systems are also designed in a way that they function on a single server, thus making increased capacity not only expensive but also finite. While the applications have consistently evolved to serve higher volumes of users, and application development activities have become agile, the traditional use of the relational database is apparently becoming a liability to many companies rather than a facilitating factor in their operations. Databases such as MongoDB can solve issues while offering various means to companies to establish remarkable business value.

Other forms of database systems that support Big Data include MongoDB, MySQL, Microsoft SQL Server, Oracle Database and SAP Sybase ASE.

Most corporations always start with the implementation of a small-scale NoSQL trial version. Through this move, they are able to develop a proper understanding of the advancement at low stakes. One of the prominent features of NoSQL is that it is open-source, which means that it is downloadable and then implemented and scaled at a minimal cost. Due to the rapid development cycles, institutions can also advance at a fast rate and ensure the delivery of a more profound customer experience at lowered prices. As organizations take into consideration the alternatives to the legacy developments, they may have a variety of motivations: identification of viable alternatives to costly proprietary applications and software, performing or scaling beyond the capabilities of their already existing systems or increasing the agility and speed of advancement.

Conclusively, NoSQL databases are among the significant databases available in the market. They facilitate the handling of large volumes of data while also ensuring easy transfer of data across some servers. Additionally, the databases are reliable since there is no downtime. This makes them efficient and reliable. The databases also facilitate various activities, such as sharding when changes are needed in the system.

References

DATABASE Editorial Board 2009. (2016). *Database*, *2009*(0), bap023-bap023. https://dx.doi.org/10.1093/database/bap023

Deshpande, T. (2014). *Mastering DynamoDB*. Birmingham: Packt Pub.

Feinleib, D. (2014). *Big Data Bootcamp*. Berkeley, CA: Apress.

Krishnan, K. *Data warehousing in the age of big data*.

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big Data Imperatives*. Berkeley, CA: Apress.

Neeraj, N. *Mastering Apache Cassandra*.

NoSQL Database: Cassandra is a Better Option to Handle Big Data. (2016). *International Journal Of Science And Research (IJSR)*, *5*(1), 24-26. https://dx.doi.org/10.21275/v5i1.nov152557

NoSQL in Action – A New Pathway to Database. (2016). *International Journal Of Science And Research (IJSR)*, *5*(6), 872-877. https://dx.doi.org/10.21275/v5i6.nov164282

Roza, G. (2014). *Databases*. New York: Rosen Publishing.