Paper on
NoSQL
(Not only SQL)

Created by

Shekhar Tyagi

Sr. Functional Consultant

Abstract

Using RDBMS databases in the past was the best option for gathering and accessing of data but due to advancement in technology the number means to capture data are increased. Data is coming from various spots and in different formats like from mobile devices, tablets and other means. There are tons of data that is coming everyday and is increasing with jaw breaking speed. Hence using the conventional relational databases in today's world is not a good option. Many technologies have been developed to counter this situation like Hadoop, Map reduce, Big data and NoSQL. These technologies are being effectively used in present and are proving to be a breakthrough in database technology.

In this paper, we will be discussing about NoSQL. Its basic concepts, types of databases, advantages/disadvantages and followed by a conclusion.

Keywords: NoSQL, types of databases, CAP, Key value stores, BASE.

Paper Type: Literature review.

Table of Contents

List of Figures

Introduction:

The scenario of using interactive applications have changed dramatically over the last decade. People are using more and more applications that are generating large volumes of data everyday to be processed. As these applications are accessible through internet and mobile devices the number of users are skyrocketed. Hence there is large amount of unstructured or semi structured data that is being stored. Using the Relational database to solve this situation is difficult cause of their inability to deal with scalability and agile challenges that are now faced in the modern applications. The reason being that it was architected to run a single machine, use a strict and schema based approach for modeling of data (Couchbase's NoSQL whitepaper 2013) .

There are 3 major factors that are driving the database industry to adopt an innovative technology. These are Big Data, increasing number of Users, and Cloud computing. Thinking about the past when 1000 users online was considered a lot and 10000 an extreme case and comparing it to now it doesn't feel much as many companies now provide services 24x7 a year. More than 2 billion people are active on the internet and the numbers are still steadily growing. In present, it's no wonder if an application has million of users daily.

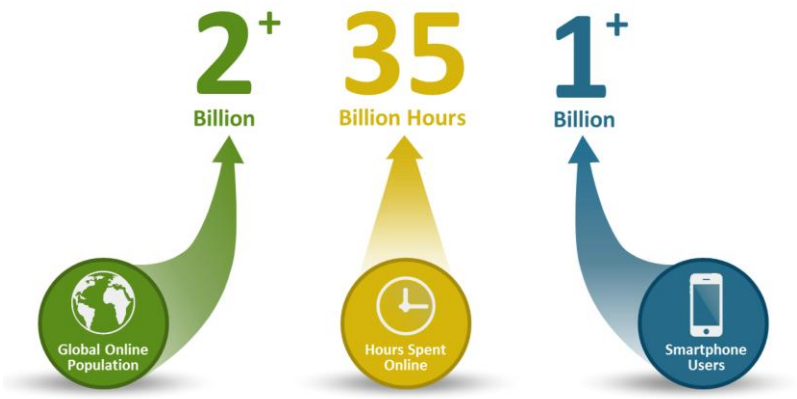

fig 1: Increasing number of users-global internet use, active users online and Smartphone users

In current environment collecting data has become very easy and apps like Facebook and twitter are storing huge sizes of data daily in different formats like user information, his DOB, work profile, multimedia data like photographs or videos, geographical location data, logs are just a few types of data being stored every day. The problem that arises with capturing this amount of data is that most of it is non-structured or semi structured and relational databases can't deal with them. Therefore, their arises a need for a technology that can handle that kind of data and NoSQL comes as the cure for this problem. NoSQL provides several types of databases to handle this kind of data.

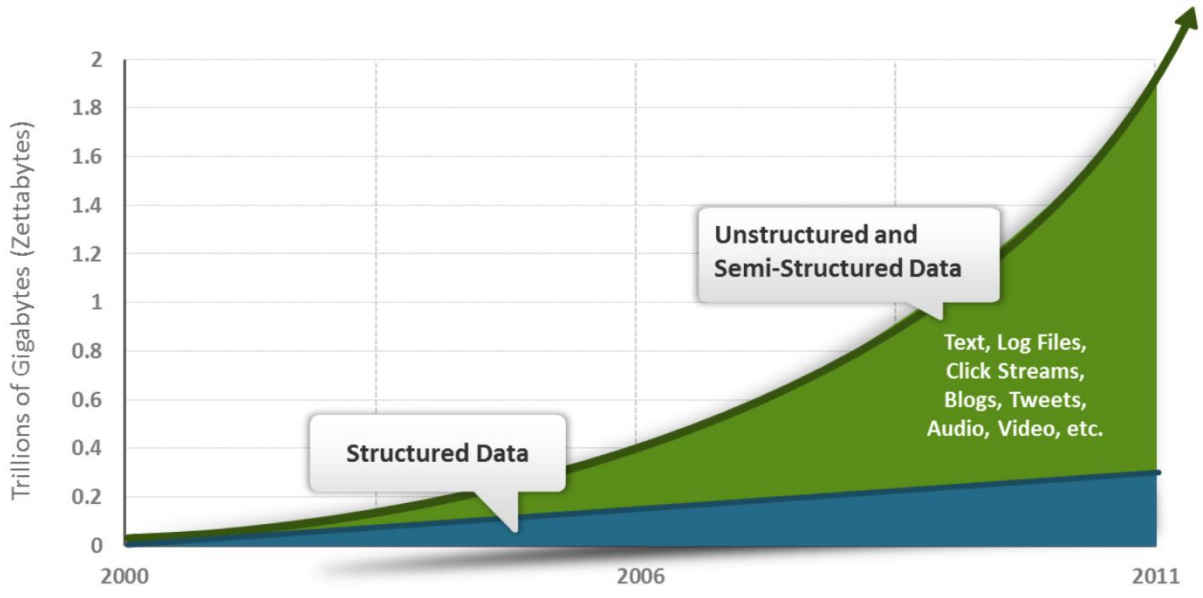

fig 2: Big data- more than 80% of data is unstructured or semi structured

If we look in the past we can see that most of the software and apps were made to run on single platform and using RDBMS databases on them was very efficient but as time changes

many applications today are being developed on three tier architecture which happens to support many users.

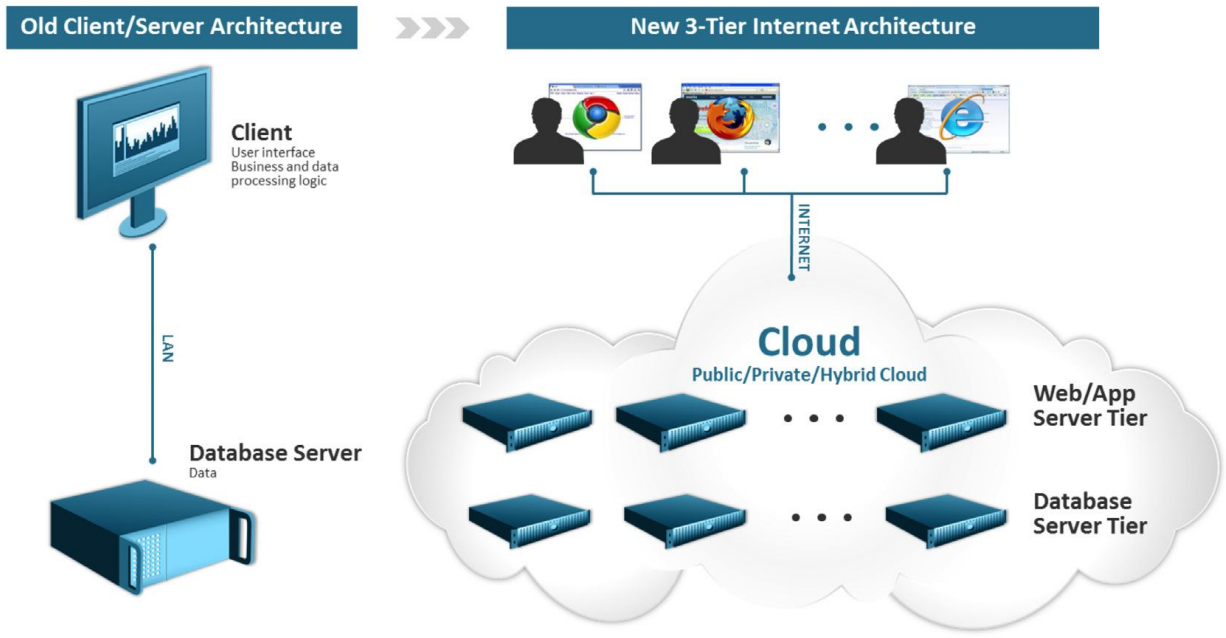

fig 3: Applications today are increasingly developed using a three-tier internet architecture.

The major factor about NoSQL technology that why it is efficient in handling distributed databases is that it's made from the beginning and follow a scale out approach that fits properly to the three-tier architecture. As the time passes more and more developers are adopting NoSQL.

In the next section this paper will give an overview of NoSQL, discuss the key concepts, its advantages and disadvantages followed by conclusion.

An overview of NoSQL

There's a confusion between how to pronounce NoSQL. Is it 'No' SQL that is it doesn't use SQL or it's something else. To justify that it's not 'No' SQL is that you can use SQL in NoSQL databases it's the choice of the developer what to opt for because NoSQL offers XML to query which is relatively easy than SQL. So, a major portion of the developers are calling it 'Not only SQL'. The reason behind its emergence is the ever-increasing data. As discussed above that the data scenario in current world is changed a lot as compared to 90's. Users, enterprises and business associates are storing, accessing terabytes of data daily and that too in different formats about their products or some other objects. NoSQL provides the technique of scaling up and scaling out the databases depending on the size of data which is a useful technique as it will be easier to handle big data.

The following are some feature of NoSQL (NoSQL in enterprise whitepaper 2103):

1. It's perfect for cloud computing.
2. Can handle unstructured, semi-structured, structured data.
3. Modifications can be done in the schema design without any downtime.
4. Can handle big data use cases like velocity, variety and others
5. It uses data in the system for real time, group analytic and activity search operations.

Basic concepts:

The CAP Theorem

Eric Brewer at the 2000 Symposium on Principles of Distributed Computing (PODC) stated the theorem as "that it is impossible for a distributed system to provide all 3 guarantees simultaneously".

In 2002 Seth Gilbert and Nancy Lynch discussed this and since then this is called CAP theorem. (Lynch and Gilbert 2002)

The acronym stands for:

Consistency: Different users can see the same data at the same time even if some other user is performing tasks on the database.

Availability: System is up all time. There is no downtime i.e. every request gets a response it doesn't matter whether it was success or a failure.

Partition tolerance: In the event of a failure system continues to operate. To be able to keep working after a server crash a server is divided into many parts keeping in mind that they don’t communicate with each other and doesn't affect each other.

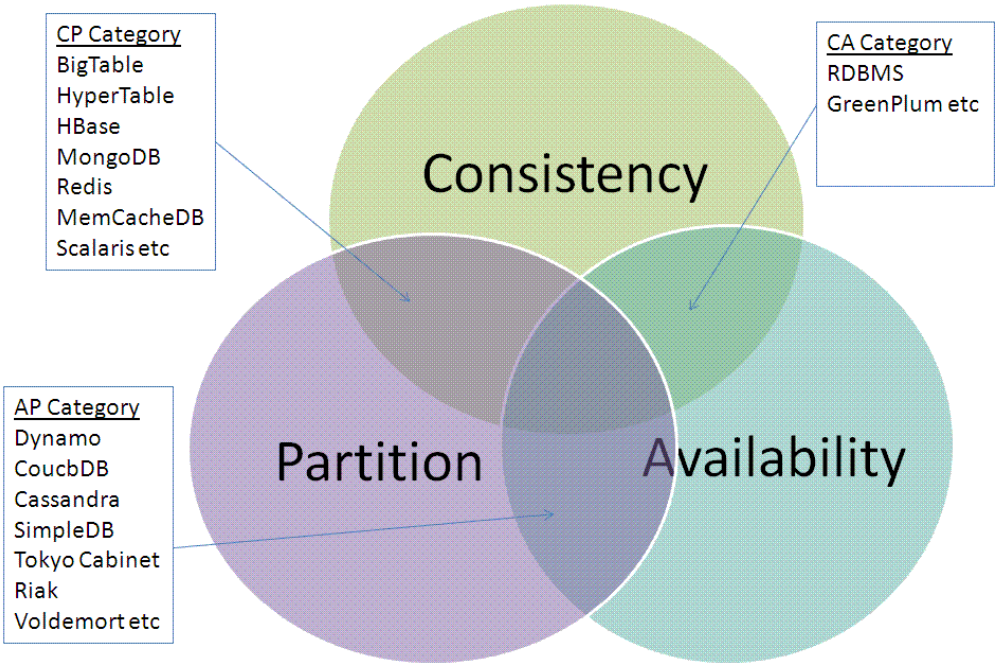

fig 4: description of 3 combination CA,CP,AP

CA - Consistency and availability are the main aspects here. All databases are in contact with each other as a result of the single cluster and no partition of the system is allowed.

CP - Consistency and partition tolerance are the main aspect of this kind of the system, it can happen that some data might not be available but consistency and partition tolerance are kept intact. Hence their arises no need for having a distributed parallel control.

AP - With the availability and partition approach, the resultant data may not be correct but system will be available despite of any division in the database. This approach is best where duplicate data is needed.

ACID vs BASE:

As discussed above, the CAP theorem says that we can only choose two options out of consistency, availability and partition tolerance. As there is vast number of applications in the market, properties like availability and partition tolerance are given more importance than consistency (Christof Strauch). To survive in the today's environment, applications have to decide between availability and redundancy to be a successful application. To tackle these problems, we opted to implement approaches like BASE. The BASE approach relinquishes the ACID properties in favor of availability, graceful degradation, and performance.

The acronym BASE is stands for the following:

- Basically Available
- Soft-state
- Eventual consistency

Types of Databases in NoSQL

NoSQL provides four basic types of databases (Girish Kumar & Rahul Checker):

Key-Value Store:

The idea behind the key value store databases is that they use hash table in combination with a unique key and a pointer which pinpoints a specific data in a database. They use a cache mechanism which increases their performance greatly. The only downside to it is as the unique keys increases it becomes tough to handle them.

example of Key Value Store: Riak, Amazon Dynamo.

Document-based Store:

As the name says it stores the data in a document. We can say that they are like Key Value store in regard that also use a unique key to identify a data. But as key value is inefficient for querying document store is exactly opposite. We can query the database with ease and efficiently.

example: CouchDB, MongoDB.

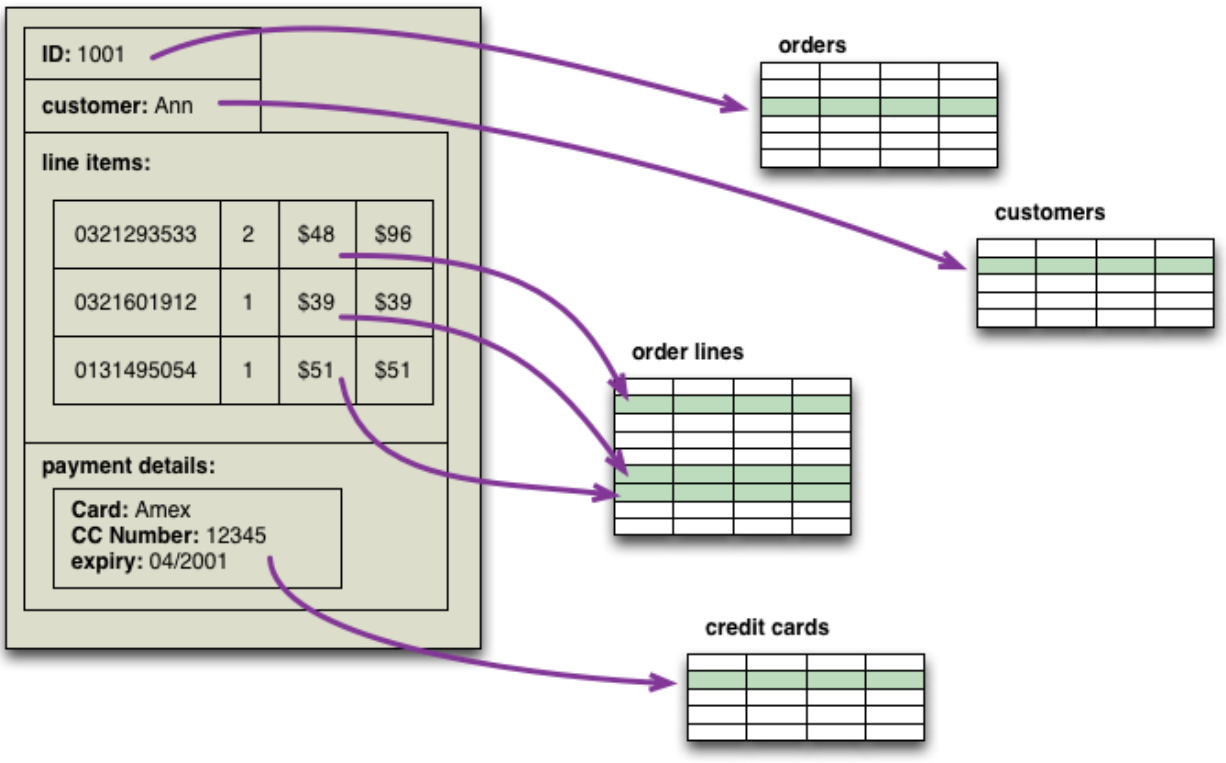

fig 5: diagram representing how data is stored in document store.

Column-based Store:

In column based store data is stored in columns rather than row. Their use lies in querying for huge databases.

example: Cassandra, Hbase

Graph-based:

Graph based databases consists of nodes, edges and set properties tie with them. Depicting the scalability concern is the best way to be represented by this database.

example: Neo4j, HypergraphDB.

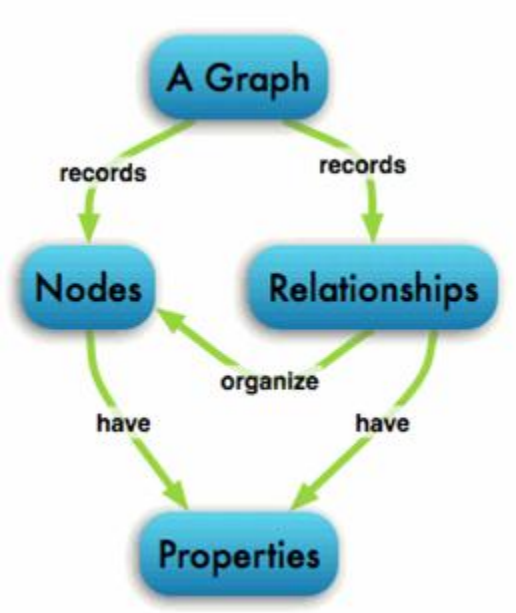

fig 6: a representation of graph based data

Advantages of NoSQL Databases (Nicholas green 2013)**:**

1) It provides less restriction as compared to SQL database. The data coming from the applications are stored virtually in any format making the modification in data easy.

2) They are easy to work with. NoSQL is less difficult and notably simple to set up than other relational database technologies.

3) The database, tables and datatypes are automatically generated when data is added.

4) Clustering and stack balancing is relatively easy.

5) NoSQL is an open source technology.

6) NoSQL is easily synchs with cloud computing.

Disadvantages of NoSQL Databases:

1) There are no connection in between the tables.

2) Lack of Stored procedures.

3) Managing NoSQL is tough as management mostly depends on scripting.

4) Gear to the GUI mode are not easily available.

5) NoSQL is only created for creation of data not for backing up data which is a negative point.

Conclusion:

So far, we have discussed about NoSQL. It's capabilities, features, advantages and disadvantages. Judging by this NoSQL is one of the best and efficient technology to face the distributed data challenges and to deal with humongous sizes of data. It's ability to work with cloud computing and being an open source makes it even more lustrous to work with it. Hence, we can say that NoSQL technology and hadoop are competing shoulder to shoulder.

A thing to keep in mind while choosing NoSQL is to analyze which type of database is best for an organization developing an application cause a wrong choice could have terrible effects.

References

Girish kumar and rahul checker different types of databases

from: http://blog.3pillarglobal.com/exploring-different-types-nosql-databases

Nancy Lynch and Seth Gilbert, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services", *ACM SIGACT News*, Volume 33 Issue 2 (2002), pg. 51-59.

from: http://lpd.epfl.ch/sgilbert/pubs/BrewersConjecture-SigAct.pdf

Nicholas green five key advantages of NoSQL

from: http://greendatacenterconference.com/blog/the-five-key-advantages-and-disadvantages-of-nosql/

NoSQL in enterprise

from: http://www.datastax.com/wp-content/uploads/2011/09/WP-DataStax-NoSQL.pdf

NoSQL Market Forecast 2013-2018
from: http://www.marketresearchmedia.com/?p=568

Why NoSQL whitepaper

from: http://www.couchbase.com/sites/default/files/uploads/all/whitepapers/NoSQL-Whitepaper.pdf