NoSQL Databases

B

Manoj Manuja
Education and Research Department, Infosys Ltd, India

Neeraj Garg
Education and Research Department, Infosys Ltd, India

INTRODUCTION

Organizations across the globe have been actively working on drawing the business values out of the huge data generated through various systems on a daily basis which is on the tune of petabytes and zeta bytes. This data is generally non-traditional, non-relational, less structured and termed as big data having three characteristics i.e. volume, variety and velocity. The sources of big data are Weblogs, emails, social media files, sensing devices outputs and audio/video bitmaps that can be mined for useful information. Till date, most of the data handling systems are relational in nature and do not handle unstructured data in a mature way. In this article, we have compared relational and non-relational databases in the context of big data. Three main characteristics of these databases namely performance, replication and sharding are evaluated and compared while managing big data. Parallel databases which are relational in nature can help in achieving the performance and replication in certain scenarios. These are difficult to scale and are not suitable for unstructured data. Another kind of non-relational databases are multimedia databases which are suitable to store multimedia objects like images, audio, videos and graphic objects. It is/may not be suitable to satisfy the three characteristics (volume, velocity and variety) of Big Data. There would be challenges like data modeling, storage requirements, indexing and accuracy of queries/retrieval to name a few in case of multimedia databases. Comparison reflects the advantages of non-relational database systems like NoSQL (Not Only SQL like MongoDB) over relational database systems like MySQL and shows that this is the way forward to handle big data.

BACKGROUND

Over the years, companies have been making critical business and financial decisions based on the transactional data stored in their relational databases. Beyond this business critical data, there is a treasure trove of other non-traditional, non-relational, less structured data in the form of Weblogs, emails, social media files, sensing devices outputs, bitmaps in the form of audio and video files that can be mined for useful information. Over the last few years, companies have started storing this data treasure because of the decreasing cost of both data storage and computing power. This data has primarily three characteristics namely

- Data is generated in huge **V**olumes.
- Data is available in a **V**ariety of data formats.
- Data is gathered with high **V**elocity.

3 V's make this data termed as Big Data which is huge in volume, coming in variety and with high velocity. There are broadly two major ways to store and manage big data i.e. using relational databases and non-relational database. Most of the existing DBMS (Database Management Systems) available across the globe are primarily relational in nature, can be expanded vertically and have predefined-schemas. Most importantly, they do not have matured way to analyze unstructured data and recognize data patterns at run time. Therefore, organizations are exploring the need of a DBMS which can handle structured as well as unstructured data and can be expanded horizontally as well as vertically. It should have the characteristic of accommodating changes in the database schemas during project life cycle and must provide the flexibil-

DOI: 10.4018/978-1-4666-5888-2.ch037

ity to update it as and when required. Non-relational databases like NoSQL have all these characteristics and properties to support 3Vs (Volume, Variety and Velocity) of big data along with the normal operational functionalities of SQL (Structured Query Language) (Lai, 2010; LaValle, et al., 2011).

MAIN FOCUS OF THE ARTICLE

In this article, we explore and examine the characteristics of relational and non-relational databases in the context of managing big data. Below section provides a brief of relational and non-relational databases with MySQL as an example for RDBMS (Relational Database Management System) and MongoDB as non-RDBMS (NoSQL) example. Subsequent three sections provide a detailed comparison of these two databases on the basis of three critical characteristics namely "High Performance and Flexible Schema," "Replication" and "Sharding." The section "Analysis and Discussion" provides a consolidated feature by feature analysis. Last two sections provide conclusion and references being used.

During the article, naming conventions for non-relational DBMS like NoSQL and MongoDB have been used interchangeably.

RELATIONAL AND NON-RELATIONAL DATABASES

Relational DBMS

RDBMS is a database management system that is based on relational model defined by E. F. Codd (2001). The data in RDBMS is stored in database objects called tables. The table is a collection of related data entries and it consists of columns and rows. Every table is broken up into smaller entities called fields. A field is a column in a table that is designed to maintain specific information about every record in the table. A record, also called a row of data, is each individual entry that exists in a table. A record is a horizontal entity in a table. Main features of RDBMS constitute storage of data in tables with row and column structure, unique identification of rows using primary and foreign keys, indexed tables for quicker data retrieval and multi user accessibility.

RDBMS is the basis for SQL. Some of the most popular RDBMS are Microsoft SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access. Let us now explore MySQL database as an example of RDBMS.

MySQL

MySQL is an open source RDBMS, which is developed by Swedish company MySQL AB. It can run on many different platforms including Microsoft Windows, the major Linux distributions, UNIX, and Mac OS X. It has free and paid versions, depending on its usage (non-commercial/commercial) and features. MySQL comes with a very fast, multi-threaded, multi-user, and robust SQL database server. It is widely used for Web based applications (DuBois, 2008; MySQL, A. B., 2001).

Non-Relational DBMS (NoSQL)

Non-Relational DBMS differ from Relational-DBMS in many ways as they don't use SQL as their query language, ACID properties (Atomicity, Consistency, Isolation, Durability) are not guaranteed, join operations are not performed but horizontal scale-up is possible along with vertical scaling. Non-relational Databases are sometimes also known as NoSQL databases. "NoSQL" is also termed as "Not only SQL" as it handles structured as well as unstructured data also. This point highlights a very critical characteristic of NoSQL in terms of its ability to vertically as well as horizontally scale-up. NoSQL is primarily useful when we are working on big data storage on the tunes of petabytes and zeta bytes. Some of the main characteristics of NoSQL databases are "flexible schema" approach, sharding, reduced latency and distributed query support to name a few. Data is saved with a unique key and value pair which can be scaled to handle large dataset very quickly. Some of the widely used NoSQL databases are Google's BigTable, MongoDB's MongoDB, Apache's Cassandra and Amazon's DynamoDB. Let us now get an insight about MongoDB (Pokorny, 2013).

MongoDB

MongoDB is an open source document oriented database system which is a part of NoSQL family of database systems. "Mongo" name has been borrowed from the word "hu**MONGO**us" that means huge or extremely large. MongoDB is an open source NoSQL