w4112p1

Cody De La Vara, Evan Drewry

March 27, 2013

Abstract

Introduction

The relational model, initially proposed by Edgar F. Codd in 1969, has dominated the world of databases ever since.[2]

Main Memory Databases

Column-store Databases

NoSQL and MapReduce

NoSQL and MapReduce are technologies that have emerged in the past decade to answer to the explosive increase in data processing demands that has resulted from the emergence of Web 2.0 and large, data-centric internet companies like Google, Amazon, and Facebook.[4] The goal of these technologies, therefore, is availability and horizontal scalability beyond what is achievable with traditional relational database systems.

Because of this, both MapReduce and the vast majority of NoSQL database systems add a layer of abstraction above cluster parallelization that provides seamless, automatic scaling and fault tolerance.

NoSQL

NoSQL is a blanket term (and maybe even a misnomer, depending on who you ask) used to classify database systems that do not conform to the traditional relational model. It is not even fully agreed upon what the abbreviation even stands for, though the most common interpretations are "Not only SQL" or simply "Not relational."[1] Sometimes the label is done away with altogether, and replaced by the slightly more accurate "Structured Storage." The term NoSQL was coined because SQL ("Structured Query Language") is tightly coupled with the relational model, and the NoSQL trend represents a departure from the "one-size-fits-all" spirit of SQL and relational databases. It is important to note that this terminology does not prescribe any specific data model, nor even a total rejection of SQL and joins; in fact, there exist many databases that fall under the NoSQL umbrella and also have a SQL-like query language associated with them.

Though the term NoSQL describes what a data store is *not* rather than what it *is*, there are several prevailing characteristics that are core to these so-called "NoSQL" data stores that differentiate them from other non-traditional database solutions like the main memory databases and column-stores described above (though there do exist both in-memory and column-oriented NoSQL data stores). Because the main motivation behind the NoSQL movement is the lack of scalability present in traditional relational databases, these data stores are most strongly characterized by their ability to horizontally scale simple database operations to millions of users, with application loads distributed across many servers.[5] Most of the other characteristics that have come to define NoSQL data stores are simply consequences of this primary goal.

Properties of a general NoSQL system

A simple data model One of the many reasons traditional relational databases have trouble scaling is the rigid, structured data model that defines them. Because of the one-size-fits-all

spirit of the relational model, there is lots of unnecessary overhead introduces A simple, non relational, model that stores STRUCTURED data rather than providing structure for the data like a Lightweight and scalable avoid unnecessary complexity introduced by one-size-fits-all mentality RDBMS's are feature-rich and rigid Three main types: document stores (Mongo, ...), key-value stores (memcached, dynamo, ...), and Extensible Record Stores (or wide column stores, or column-oriented) (BigTable, Cassandra, ...)

Loss of ACID compliance Sacrifice consistency for scalability

CAP-theorem

Consistency c

Availability foo

Partition Tolerance blah

Document Stores

MongoDB

CouchDB

Key-Value Stores

Extensible Record Stores

Cassandra

BigTable

MapReduce

MapReduce is a simple high-level programming model for processing huge quantities of data in parallel on a cluster. It is powerful because it provides a layer of abstraction over all the complexities of parallelization on a large number of nodes–including execution scheduling, handling of disk and machine failures, communication between machines, and all partitioning of data among the cluster–while still providing a simple and flexible programming model.[3]

The

MapReduce is also the name Google gave to their widely mimicked implementation of the MapReduce model. The most popular open source implementation is Apache's Hadoop.

Hive

Compared to traditional relational databases

Bibliography

[1] Rick Cattell. Scalable sql and nosql data stores. *ACM SIGMOD Record*, 39(4):12–27, 2011.

[2] Edgar F. Codd. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, June 1970.

[3] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[4] Neal Leavitt. Will nosql databases live up to their promise? *Computer*, 43(2):12–14, 2010.

[5] Christof Strauch and Walter Kriha. Nosql databases, 2011.