w4112p1

Cody De La Vara, Evan Drewry

March 27, 2013

Abstract

Introduction

Main Memory Databases

Column-store Databases

NoSQL and MapReduce

NoSQL and MapReduce are technologies that have emerged in the past decade as an answer to the explosive increase in data processing demands that has resulted from the emergence of large data-centric internet companies like Google, Amazon, and Facebook. The goal of these two technologies, therefore, is scalability and availability beyond that achievable by 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

- NoSQL = "Not only" SQL
- Departure from relational model
- Lightweight and scalable
- Sacrifice consistency for scalability
- Two main types, document stores (Mongo, ...) and key-value stores (BigTable, Cassandra, ...)

Document Stores

MongoDB

CouchDB

Key-Value 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.[1]

The

MapReduce is also the name of Google's widely mimicked implementation; however the most popular implementation is Apache's open source Hadoop.

Hive

Compared to traditional relational databases

Bibliography

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