A. Silva

Master in Informatics and Computing Engineering
Data Modeling with NoSQL: How, When and Why
Carlos André Reis Fernandes Oliveira da Silva

—

Abstract

As the Web continues to grow in size, more and more services are being created that require data persistence. The amount of data that these services need to archive is growing at an exponential rate and so is the amount of accesses that these services have to serve. Additionally, the relationships between data are also increasing. In the past, problems involving data persistence have been consistently solved by relying on relational databases. Still, the requirements of these new services and the needs for scalability have led to a depletion of relational database technologies.

New approaches to deal with these problems have been developed and the NoSQL movement was formed. This movement fosters the creation of new non-relational databases, specialized for different problem domains, with the intent of using the “right tool for the job”. Besides being nonrelational, these databases also have other characteristics in common such as: being distributed, trading consistency for availability, providing easy ways to scale horizontally, etc.

As new technologies flourish, there is a perceived knowledge impedance that stems from the paradigm shift introduced by these technologies, which doesn’t allow developers to leverage the existing mass of knowledge associated with the traditional relational approach.

This work aims to fill this knowledge gap by studying the available non-relational databases in order to develop a systematic approach for solving problems of data persistence using these technologies.

The state of the art of non-relational databases was researched and several NoSQL databases were categorized regarding their: consistency, data model, replication and querying capabilities.

A benchmarking framework was introduced in order to address the performance of NoSQL databases as well as their scalability and elasticity properties. A core set of benchmarks was defined and results are reported for three widely used systems: Cassandra, Riak and a simple sharded MySQL implementation which serves as a baseline.

Data modeling with NoSQL was further researched and this study provides a simple methodology for modeling data in a non-relational database, as well as a set of common design patterns. This study was mainly focused on both Cassandra and Riak.

Additionally, two prototypes using both Riak and Cassandra were implemented, which model a small chunk of a telecommunications operator’s business. These prototypes relied on the methodology and design patterns described earlier and were used as a proof of concept. Their performance was put to test by benchmarking a set of common (and usually expensive) operations against a traditional relational implementation.

Both Cassandra and Riak were able to yield good results when compared to the relational implementation used as a baseline. They also proved to be easily scalable and elastic. Cassandra, specifically, achieved significantly better results for write operations than the other systems. The developed design patterns proved themselves useful when implementing the prototypes and it is expected that given this work it will be easier to adopt a NoSQL database.

Resumo

n/a

Jury

- Chair: Maria Eduarda Silva Mendes Rodrigues
- External Examiner: Rui Oliveira, U.Minho
- Supervisor: João Correia Lopes
- Date: 13/7/2011