Automatic Query Driven Data Modelling in Cassandra

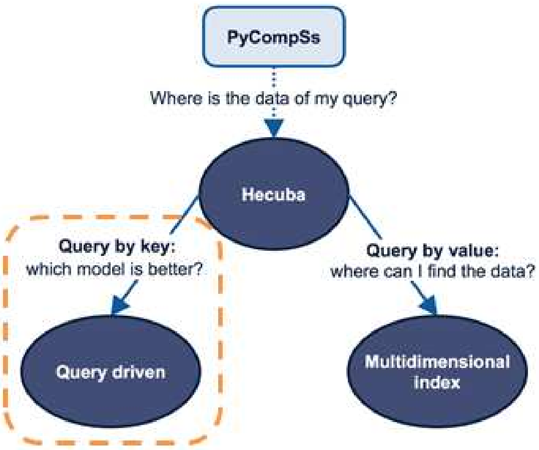
Roger Hernandez, Yolanda Becerra, Jordi Torres, Eduard Ayguadé
Barcelona Supercomputing Center
roger.hernandez@bsc.es

***Abstract**-Non-relational databases have recently been the preferred choice when it comes to dealing with Big Data challenges, but their performance is very sensitive to the chosen data organisations. We have seen differences of over 70 times in response time for the same query on different models. This brings users the need to be fully conscious of the queries they intend to serve in order to design their data model. The common practice then, is to replicate data into different models designed to fit different query requirements. In this scenario, the user is in charge of the code implementation required to keep consistency between the different data replicas. Manually replicating data in such high layers of the database results in a lot of squandered storage due to the underlying system replication mechanisms that are formerly designed for availability and reliability ends. We propose and design a mechanism and a prototype to provide users with transparent management, where queries are matched with a well-performing model option. Additionally, we propose to do so by transforming the replication mechanism into a heterogeneous replication one, in order to avoid squandering disk space while keeping the availability and reliability features. The result is a system where, regardless of the query or model the user specifies, response time will always be that of an affine query.*

Description and Previous Results

The leap into the Big Data world can be too challenging our out of the field for some information technologies users. This, among other concerns motivated the apparition of Hecuba, a set of tools and interfaces developed in our research group, which aims to facilitate programmers an efficient and easy interaction with non-relational databases. Currently all implementations are made on Apache Cassandra as an example database although it is easy to port this implementation to any non-relational key-value data store.

In the work presented here, we focus on the Query Driven Data Modelling part of Hecuba, a concept for which we have also developed a prototype which automatically decides which data model should be queried by taking advantage of the differences in the topology of the data stored in the Cassandra models and replicas, and overall increase performance by aiming at the affine query-model relationships. Currently, all this management falls on the user responsibility; therefore he would manually require building all the different replicas. This implies further hassle such as assuming the extra redundancy, maintaining the consistency between all the models that share common data when it had to be inserted or changed, or being mindful about performance issues or extra load time when having to modify database data since it would involve to do one write per data replica instead of a single write to the database. With our suggestion, all this hassle would disappear becoming yet another part of Hecuba that could benefit end-users by removing their need to specialise on the details of the database behaviour.



In our previous publication [2] we could confirm that there existed affinities between queries and models, with great performance differences when combining their possibilities, and the results were so motivating this gave an obvious track opening towards the development of a Query Driven Data Modelling prototype, which seeks to transparently apply the query assignations of the client towards the database aiming to get the best performance it can offer between the available models.

Since we already presented performance results in our previous work, this publication focuses on the design, development and functional analysis of the prototype, finally confirming that the possibilities on performance increases we saw were possible, transparently become a reality with the Query Driven Data Modelling prototype.

Past Publications

[1] C. Cugnasco, R. Hernandez, Y. Becerra, J. Torres and E. Ayguadé, “Aeneas: a tool to enable applications to effectively use non-relational databases” *Procedia Computer Science ICCS 2013*

[2] R. Hernandez, C. Cugnasco , Y. Becerra, J. Torres and E. Ayguadé, “Experiences of using Cassandra for molecular dynamics simulations” *Euromicro International Conference on Parallel Distributed and Network-based Processing PDP 2015*

[3] R. Hernandez, C. Cugnasco , Y. Becerra, J. Torres and E. Ayguadé “Automatic Query Driven Data Modelling in Cassandra” *Procedia Computer Science ICCS 2015*