José Magalhães

Master in Informatics and Computing Engineering
Suporte para Séries Temporais em Plataforma *e-Science*
José Tiago Paiva Antunes Magalhães
—

Abstract

In the last few years database management systems solutions, that do not follow the traditional object-relational approach, have gained popularity in specific cases where it is not necessary to maintain ACID properties or use SQL to query the database. The NoSQL databases differ from the relational approach because they use different kinds of structures to store the data like key-value data structures, columns, graphs or documents and they are being increasingly used in applications that deal with the so called Big data.

In the scope of Earth Sciences, including the use of laser sensors (LiDAR) for the analysis of wind conditions on wind towers facilities for electricity production, time series are produced and later used by researchers in their work research. Due to the huge amount of information that is produced by these sensors, it is inefficient to use relational database management systems to store time series.

The objective of this dissertation is to develop an innovative e-Science platform focused on the storage and availability of all meta-information and information necessary for correct interpretation of the scientific results, including repeatability and reproducibility. This platform is supported by two approaches: one using a relational database and another one using a NoSQL database. For its implementation it was studied, in detail, the impact that the two approaches have in the performance of the solution, especially in the storage and availability of data.

The tests performed showed that the approach which uses a NoSQL database management system provides a far superior performance compared with the first approach. The study also reveals that the limitations discovered in the first approach makes it unfeasible for a large amount of data. Finally, the second approach proves to be more scalable than the first which is an important property to be able to maintain the platform availability as the number of users and amount of data increase.

Resumo

Nos últimos anos têm ganho popularidade soluções de gestão de dados que não seguem a abordagem objeto-relacional tradicional, nos casos em que não é necessário manter as propriedades de Atomicidade, Consistência, Isolamento e Durabilidade (ACID) nem é necessária a utilização de SQL para a interrogação às bases de dados onde são guardados esses dados. As bases de dados NoSQL diferem da abordagem relacional por usarem estruturas de dados chave-valor, coluna, grafo ou documento, e estão a ser cada vez mais usadas em aplicações que tratam a chamada Big data.

No domínio de aplicação das Ciências da Terra, nomeadamente a utilização de sensores laser (LiDAR) para a análise das condições de vento em instalações de torres eólicas de produção de energia elétrica, são produzidas séries temporais usadas posteriormente por investigadores da área nos seus trabalhos de investigação. Devido à enorme quantidade de informação que é produzida por estes sensores, torna-se ineficiente a utilização de base de dados relacionais para o armazenamento das séries temporais produzidas.

O objetivo deste trabalho consiste na produção de uma plataforma e-Science, inovadora para investigadores, focada no armazenamento e disponibilização de toda a meta-informação e informação necessária à correta interpretação dos resultados científicos obtidos, nomeadamente a rastreabilidade e a reprodutibilidade dos

dados. Esta plataforma será suportada por duas abordagens distintas: uma com recurso a uma base de dados relacional e uma outra com recurso a uma base de dados NoSQL. Para a sua implementação foi estudado com detalhe o impacto que as duas abordagens terão no desempenho da solução, nomeadamente no armazenamento e disponibilização dos dados para os investigadores.

Os testes efetuados permitem concluir que a abordagem que faz uso de um sistema de base de dados NoSQL apresenta um desempenho muito superior na generalidade dos casos, quando comparada com a primeira abordagem. O estudo realizado revela ainda que as limitações descobertas na primeira abordagem inviabilizam a sua utilização para grandes quantidades de dados. Por fim, a segunda abordagem revela-se mais escalável que a primeira, propriedade importante para se conseguir manter a disponibilidade da plataforma, à medida que o número de utilizadores e a quantidade de dados aumentam.

Jury

- Chair: Maria Cristina Ribeiro
- External Examiner: Maria Benedita Malheiro
- Supervisor: João Correia Lopes
- Date: 16/07/2015