Clustering and Ranked Search for Enterprise Content Management

Juris Rats, Rix Technologies, Riga, Latvia
Gints Ernestsons, Clusterpoint, Riga, Latvia

ABSTRACT
The aim of this work is to understand more closely where the border lies between relational and Not Only Structured Query Language (NoSQL) platform as concerns Enterprise Content Management (ECM) area. Another objective (closely related to the first one) is to specify the conceptual architecture of the distributed ECM system. The authors specify the model of the prototype ECM system and compare two platforms for this model – MS SQL based for the relational platform and Clusterpoint for the NoSQL platform. The results of performance measurements of SQL and NoSQL technologies for Enterprise Content Management specific tasks are presented and analyzed. The viability of NoSQL Document-oriented database solution based on clustering and ranked search is demonstrated. The ways to leverage the improved performance and scalability of the software to better serve the business needs of the Enterprise are discussed. The conceptual architecture of the prototype system is outlined.

Keywords: Big Data, Clustering, Document-Oriented Databases, Enterprise Content Management (ECM), Not Only Structured Query Language (NoSQL), Ranked Search, Ranking

INTRODUCTION
Every day 2.5 quintillion bytes of new data are created – the result of this extremely rapid growth of data volume is that 90% of the data in the world today has been created in the last two years alone (Frank, 2012). In response to ever-increasing volume and variety of data software industry is moving towards distributed systems. Vast territories are discovered in the world of distributed systems where the relational data model and “one-size-fits-all” strategy is not a viable solution. A number of specialized solutions (so called NoSQL systems) are created in response, dedicated for use in separate areas – transaction processing, data warehouses, stream procession, data aggregation etc. The NoSQL technologies are important players in a rapidly growing Big Data market, expected to reach $16.9 billion in 2015 (Bradshaw, 2013), and the market share of those technologies is expanding.

DOI: 10.4018/ijeei.2013100102
This should not necessarily mean a death of the relational data model but rather that we have discovered its territory has borders. Still this frontier between the particular NoSQL and SQL solutions are vague.

The aim of this work is to understand more closely where the border lies between relational and NoSQL platform as concerns Enterprise Content Management (ECM) area. Another objective is to specify the conceptual architecture of the distributed ECM system. Both objectives are closely related because one cannot compare two platforms not understanding the nuts and bolts of each of them.

Our research concentrates on Enterprise Content Management (ECM) and, in particular, Document Management, which is influenced heavily by the rapid growth of data amount as well. The research is funded by European Regional Development Fund (ERDF) and its objective is to evaluate and compare the performance on large document databases of our in-house ECM system Namejs and the NoSQL database system Clusterpoint, and to design the architecture of the new NoSQL based ECM system.

The viability and advantages of the NoSQL based solution are demonstrated by our research and the conceptual architecture of the prototype Document management system is outlined.

ENTERPRISE CONTENT MANAGEMENT

Concept of Enterprise Content Management (ECM) has emerged during the last 20 years and is still evolving (see one of the most cited definitions in Blair, 2004). The term nowadays is much wider in scope than its origins. We will follow the more recent and comprehensive definition of Grafthmann, Lehnard and Slatter (2011, p. 5):

*Enterprise content management comprises the strategies, processes, methods, systems, and technologies that are necessary for capturing, creating, managing, using, publishing, storing, preserving, and disposing content within and between organizations.*

The most important enhancement of the latter is acknowledging the significance of the communication between the organizations. This is the important area for improvements because a lot of information that should be shared is duplicated between enterprises. It is an illusion that a correspondence saved in repositories of two organizations is safer in comparison with one copy in a decently protected shared storage. The shared solution has obvious advantages though – allowing for search in context of two enterprises (if the searcher has necessary authority).

ECM should be regarded as integrative middleware that uses internet technologies for in-house information provision (Kampffmeyer, 2007). It covers a wide area of functionality (Bantin, 2008; Nilsen, 2012; Korb & Strodl, 2010; Kampffmeyer, 2007) that surrounds five functional areas - Document Management, Collaboration of supporting systems (or groupware), Web Content Management (this includes portals), Records Management (archive and file management systems) and Workflow and Business Process Management.

Kampffmeyer (2007) characterizes the scope of the ECM systems by five C’s – Content, Collaboration, Compliance, Continuity and Cost. *Content* here stands for any structured or unstructured, dynamic or static contents, including records, documents, web pages, metadata etc. *Collaboration* includes direct communication with chats, instant messaging, forums etc., knowledge management and project work support. *Compliance* addresses fulfillment of legal and regulatory requirements with process documentation, secure storage and traceability of transactions. *Continuity* demands to support information availability and correctness. *Cost* designates the cost savings enabled through the support of collaboration, simplifying of administration, improving of working methods etc.
Entrepreneurship in Small-Scale Farms: From Production to Commercialization
www.igi-global.com/chapter/entrepreneurship-in-small-scale-farms/179701?camid=4v1a