"Alexandru Ioan Cuza" Iaşi University

Faculty of Economics and Business Administration

Doctoral School of Economics and Business Administration

Field of study : Business Information

DATA MANAGEMENT MODELS IN CLOUD

PHD Thesis - Summary

Scientific coordinator:

PHD Professor Marin Fotache

Author:

PHD student Dragoș Iulian Cogean

Iaşi, 2013

The author aims, in this paper, to address a new area of the informational technologies, cloud databases and their management. The perspective is a critical one, the paper will present both the advantages and disadvantages of using these technologies, together with the case studies that will confirm or infirm the author's hypothesis.

Along with the technical view, dominant in the paper, the author tries to bring aspects that have a role in the using of could technologies in organisations. Therefore, beginning with a literature overview, legal aspects of cloud computing, databases and the economical justification of these technologies are described. The author tries to give an overview of the legal aspects specific in the EU area as well as in the USA, taking into account that the USA is the international pole of the information technologies and, in many cases, the center from where the ideas and new trends in the field spread towards the rest of the world.

Although, right from the title, the reader receives information regarding the approach of storing and using the cloud technologies, a complete image of the current situation cannot be completed without a comparison of the two database models, taking into account the last years. The thesis includes relational and NoSQL databases analysis, considering aspects as performance, scalability, accessibility and functionality. More, in the last three chapters of the paper, a number of case studies are taken into account in order to bring new information to the readers together with statistical interpretations of the phenomena. The author take real situations as starting points, from the Faculty of Economics and Business Administration regarding the logging of the incoming data coming from various applications, as well as the need to create an application that manages the admission process of the new students generations. The conclusions of these studies are generalized into a larger category of applications. The last chapter contains a rigorous analysis of the cloud databases management. The topic of the tests are two database services, one is relational, the other one is a NoSQL one. The comparison of performances is possible with the use of a set of statistical tools, descriptive, as well as inferential. This paper is not aiming to name a winner of declare which technologies that are superior to others, but only to describe the current situation of databases, encouraging the reader to use the proposed testing methodology, in order to determine the optimal data storing and management

Right from the design phase of an application (desktop, web or mobile), among the biggest interests are the data storage methods. The databases are, by far, the best solution, but in the last years, choosing the storage system has been one of the most difficult tasks. Having as a starting point the theoretic background of E.F. Codd’s paper, „A Relational Model of Data for Large Shared Data Banks” and the theorems from the relational algebra, the relational databases have dominated the market starting in the 80’s and still are the choice of most companies. During the last years, the post-relational databases become more popular, but, at the same time, more controversial.

The present look of the storage and retrieving of the data is influenced by elements like the fact that people produce more and more data, but data is not only the result of the human activity. More than this, 90% of the data we are storing at this moment is produced during the last two years only. The sources of data are not only the human activities. Sciences, the internet, business, software, sensors, social networks, internet browsing, military operations, medical activities, all are data generators that need to be analyzed, stored, reused at a given point. The speed of data generators is different from the last years. The data is produced so fast, that it determines the impossibility of analyzing. In fact, it determines giving up the transactional states and the appearance of streaming database systems, or even online databases. The data types are different according to their types. Aside from the normal tabular data we are also considering multimedia, images, video, audio, software generated logs and, in general unstructured data. Coming to the size of the data storage, the units or measurement start from terabytes. The main storage options are the new technologies (mostly NoSQL) or optimized relational database engines. Each company wants to lower the storage costs of a data unit.

The purpose of the current paper is the analysis of the current situation of the databases, starting from relational technologies, and continuing with the more recent NoSQL technologies. By reading this paper, the reader, more or less technical, can have a better understanding of their options concerning the storage technologies, whether they are in cloud or on premise. The main methods considered in the research process are the literature review and the case study.

Relational databases have been for over 30 years the main choice of those who develop software or data analysis processing (OLTP or OLAP) due to their generic nature and adaptability. Shifting the focus to other types of databases as a source has three current problems for software developers: increasingly large data volumes, scalability requirements and adaptation to load peaks, multiplying the types of data stored. The attempts to adapt the current systems to relational databases to these requirements, often lead to an undesirable decrease in performance. On the other side, a process of understanding the current RDBMS problems must, however, be accompanied by awareness of the advantages they have. The generality of the problems we can solve, using a language that became the lingua franca, SQL, and the ACID benefit of transactions properties, the maturity of the data model, are just some of these benefits. The emergence of NoSQL databases, although a welcome change in the world of data storage and manipulation, must be seen and treated with caution and openness at the same time. Searching for a replacement to a solution that cannot meet the information requirements is a good step, but rushing into a decision and the lack of assessing it, may be a direction to lead the organization to massive losses. They can be caused by interruption of service delivery to customers, data loss, excessive amounts invested in the development of new components.

We notice that scalability is indeed a nerve of relational databases. Solutions exist from memory databases (Oracle Berkeley DB type) systems using procedures such as map-reduce (Hadoop, Dynamo), graph databases, and more. There is no general solution available, but only families of problems that can be solved with the same technology. Giants like Google and Facebook have managed to achieve scalability through the use of alternative technologies of relational databases, other companies have started from the premise that a good application architecture, caching technologies and a good understanding of how clusters can be created relational database will allow optimal required data processing. In the cloud, the situation is the same: there are post-relational NoSQL database-as-a-service solutions, but also instances of relational databases available to customers (Amazon offers instances of Oracle 11g and MySQL, Microsoft offers Azure). Another phenomenon observed and documented by programmers and designers is impedance mismatch. Impedance mismatch or object-relational impedance mismatch can be defined as the sum of the technical or modelling difficulties for encountered during the development of applications that support persistence using a relational database.

As we have seen, when we manage large volumes of data, scalability is an essential element of an organization's information system, relational database engines show their limitations. Here interfere the systems that are part of the great family of NoSQL. They can eliminate the problem of scalability, even if it forces the user to make other compromises. Specialty papers classify the NoSQL databases in the following families: document, graph, key-value, multivalve, object, column.

The database systems management market of post-relational type is still a place of apparitions accompanied by aggressive advertising, promises to solve all the problems of persistence and data manipulation, presentation antithetical to "the elephants" of the relational world. By adding these promises to the cloud computing hype, we obtain a competitive environment, a good playground for any ambitious marketer. However, there are examples of software products that have managed to escape from the insecure area. The big players in the IT field offer their own NoSQL products, software solutions presented as complementary to existing relational engines. Amazon (through Dynamo and SampleDB), Oracle (NoSQL Database) understand the need for different storage applications. At the same time, there are companies that, in addition to a moderate tone in advertising, succeeded to provide consistency from one software version to another, good documentation, tutorials, multiple specific driver for more programming languages, technical support, examples of use cases, all these for NoSQL database engine. These include 10Gen, manufacturer for MongoDB database or Neo4j, a company that provides a graph DBMS having the same name. In chapter 4 of the paper, the author describes the use of a document-oriented DBMS through a practical study, the chosen DBMS being the most popular among the NoSQL systems, MongoDB. As secondary objectives of this part of the work, we find the development of an application that has the MongoDB persistence layer and a comparison in terms of performance between two database services in the cloud. Studying the literature and case studies are the predominant research methods in this chapter. However, we try to combine qualitative and interpretative analysis, with a quantitative analysis based on statistical methods.

The development of MongoDB by 10Gen, began in 2007 and the project is distributed as open-source software since 2009. The current version is 2.4, an the company supports it along with version 2.2. Over time, MongoDB developers have introduced new components in the default installation, like utilities for import and export of data, performance monitoring, backup, or restore. A good example of evolution dictated by the user community is giving up the general locking mechanism when executing read and write operations. This change occurred with version 2.2. Currently, MongoDB provides a mechanism to ensure the competition for using the reads-write principle¸ more precisely, by reading on multiple threads in parallel, but writing data blocks, so that at any given time, only one thread can modify the data in the same database. MongoDB provides both primary means for performing CRUD (create, read, update, delete), an advanced framework for data aggregation (aggregation framework) and the option of running the sequences of map-reduce code written in Java Script.

In order to discover and better assess the MongoDB capabilities and how NoSQL databases can be used to solve real problems in the IT environment, the author describes the implementation of a management system for data logging using MongoDB as data persistence layer at the Faculty of Economics and Business Administration. The application developed in the Faculty of Economics and Business Administration aims to partially manage the outputs generated in the institution by the data logging systems. Among them is the data generated by the Blackboard e-learning platform, the usage data of the workstations in the laboratories, details about electronic books and articles accessed in the library or laboratories, information about sources of multimedia educational content (Scribed, Slide share, YouTube), navigation and data access on faculty's portal site. The choice for the development environment and runtime was Python. For hosting the web application, the chosen server was Tornado, known as able to manage client requests in a non-blocking manner. Both the database and web server are installed on Windows machines that are considered to have an average performance (commodity hardware), comparing them with the technologies of 2013. The implementation process of this application, even if in the prototype stage, can be considered a success, both in terms of use for the entity, for testing and validation of storage technologies of NoSQL data types. The DBMS of our choice appears to be mature and ready to handle massive amounts of data, but also provides useful tools to perform queries and data management tasks for the users.

In chapter 5, cloud databases are introduced. They represent one of the new challenges in the field of information systems. Aggressive advertising, specific to all products presented as cloud services, aims to induce the idea that most of the problems posed by the storage of data can be resolved by implementing such a storage service. However, price, safety and privacy, laws, contract compliance, performance and availability of services, are just a few things intensely debated by specialists. The author aims to address some issues related to cloud databases, as their development can influence the acceptance and popularity of several types of databases, thus making polyglot persistence a possible scenario for many organizations. Among the advantages of using cloud computing, by studying the relevant literature, the author identifies: reducing investment in hardware, reducing maintenance staff dedicated to this equipment, rational acquisition of licenses to use a software product, eliminating partial integration problems by choosing already tested and widely used solutions, easy management and access security policies (without forgetting to mention the risks), scaling of computing power. The analysis of cloud technologies continues with legal issues related to their adoption in organizations. Trying to summarize the results, we can say that the legal environment is an important factor in the decisions to adopt cloud services. Both the European Union and the United States, have laws which are not always correlated with technological progress and can be a barrier in many situations. In Europe, there are very strict laws regarding the protection of personal data and their transmission outside the community. In the United States, the obsession for terrorism leads to invasive acts from the authorities, making possible the risk of exposure of sensitive information such as financial reports, lists of customers and transactions, the result of research and development or trade secrets being a delicate issue. Companies that have adopted, or still deliberating on the adoption of cloud computing, need to consider the implications which appear in the relations with the authorities and with their business partners. For the second category, related contracts should contain clear sections on the obligations and rights of both parties involved and a clear definition of how responsibility is divided when stored or processed data that do not comply with copyright and privacy laws and rights, have an obscene content, or are an affront to human dignity. Also, the quality parameters of service providing into the cloud, such as performance and availability, should be clearly described, along with how to measure in a form such as SLA - Service Level Agreement. The advantages of cloud computing are often clearly defined for organizations, but legal risks and, particularly, the effect it may produce in the organization, should be thoroughly studied by mixed teams of technical people and law professionals.

The economic component is the next analysis with which the author continues the paper. Concepts, such as opportunity cost, capital investment and total cost of ownership are presented and explained. The human capital is also not ignored within the documentary approach. The author notes an increased interest from economists and specialists from other fields in understanding the business model adopted by service providers in the cloud, with economic performance that those who adopt these technologies could benefit from. Any decision to migrate to the cloud should be accompanied by a well-defined budget and a detailed plan of the costs related to this change that must be provided. The paper believes that, in order to avoid another crisis like the one in the early 2000s dot-com enterprises, the cloud technologies should be viewed cautiously and, in addition, to thorough technical evaluation; the economic efficiency must be calculated and communicated to decision makers.

The paper continues by presenting the characteristics of the databases in the cloud. Some of these features often correspond to the decisive preliminary selection criteria. For example, depending on the application that will use the persistent layer, only those services for relational or NoSQL can be selected. Other features are quantifiable and can participate (after applying a weighting factor) in adopting a decision. Into this category might come available resources, price, performance use (latency, number of operations per second supported). A third set of features can be observed, as the service is used and can be the basis for keeping a supplier or change it. Listed in this category are how and when comes the response from the provider to requests for assistance, administration and management tools, how they are maintained and documented, the number of service disruptions to the time the problem was fixed. In order to choose between different suppliers database in the cloud, organizations need to consider some of the above criteria and other specific circumstances. How these criteria are considered in the decision process must be one comparable to all services that are considered of potential use. The author proposes a methodology for evaluation of these services into the cloud database. Each stage is defined by a specific task and deliverable. The methodology described is reproduced in full in this summary, the main part of her thesis.

No.	**Activity**	**Deliverable**
1	Selection of the domain and the given situation for which the alternative of using cloud data storage is considered	List of the entities to be stored in the database and possibly a preliminary schema for the database (even a soft-schema)
2	Identifying of the mandatory features for the storage system: data model (rational, object, document, graph), instruments for service administration, data import and export utilities	List of technical features not subject to performance benchmarking
3	Comparison between the list obtained at previous step with all the characteristics already known for the cloud database service considered as being candidates for	List of cloud database providers able to fulfil technical demands

	the study.	
4	Identification of the users for the application and the way they access it	Clear report of the number of estimated operations (in normal usage and in high usage periods), together with details about the operation types: write, read, update, delete. Also, knowing data size and data types is important.
6	Estimation of the operational costs for the cloud storage service in an already defined time interval, knowing the prices for each provider and the data storage needs for the organization	List of cloud database providers which are accessible for the organization from the costs point of view
7	Identifying the infrastructure elements needed for running the performance tests	List of servers, network equipments and configuration options for them.
8	Identifying appropriate test tooling elements for the cloud database services involved in the decision process. These test applications have to be able to provide a unitary test approach and must be as less as possible influenced by the host	List of the test applications and their corresponding documentation

	operation system or connection type to the database.	
9	Test definition and their running sequence	Detailed description of the configuration parameters and code (scripts) needed for running the tests
10	Running of benchmarking tests in multiple rounds (for getting relevant data)	Results of the performance benchmarking tests
11	Interpretation of the results for the ran tests considering parameters like system capacity, average latency, maximum latency, number of obtained	Graphical representation of the results for a better understanding of their significance or statistical analysis
12	Creation of a decisional model to include selection criteria together with their weight	Decision table completed with all available data, able to give enough information for database service selection

The next part presents readers with a new case study, this time dedicated to the use of databases in the cloud, and their evaluation methodology presented above. This study has as a starting point an already existing problem of the belonging organization of the author of this thesis, the Faculty of Economics and Business Administration. The admission periods of enrolment and publication of the selection results generates data traffic increased significantly within the outcome periods. Even though we are only talking about a short period of 2-3 weeks, the administration efforts of the data sources are significant. Also, the possibility of disruption or dysfunctions in the software that makes the data available for the possible future students is not easily accepted. The large number of requests, phones, visits to the faculty could seriously disrupt teaching and secretarial activities. The image of the institution and the stress of students (taking into account the fact that each year more than 2000 students are accepted, but many more submit their applications) are factors that may lead to a decision to adopt a data storage service in the cloud. This paper presents three types of tests designed to simulate user behaviour in various situations such as the registration period, the querying and modification of data, and finally, the period of maximum traffic, corresponding to the admission results announcement. For the tests there are two databases in the cloud services used, each of them having one shared storage type instance. The tool used to generate the test and to measure the response time (latency) is a modified version of YCSB (originally developed by Brian Cooper in Yahoo, afterwards becoming an open-source project). The test environment is available by purchasing a virtual machine in the Rackspace cloud. The test results provide important information regarding the service performance. First of all, localization is an important factor. For the average values, a multiplier of 2.5 to 3 applies on the latency when the client and the service are not on the same continent (America and Europe in our case). Then, the author identifies the differences in the stability of the services (the deviation in time from the average value of the latency). The calculation is done using standard relative deviation, the result showing a significantly increased stability (multiplier factor is 2) for one of the two chosen services. If for the initially defined test scenarios, the author observed a slight advantage in terms of average latency of the services, when the number of requests in the test number 3 is increased (simulating a growing number of users), the ranking is reversed. This exercise has the purpose to confirm the possibility of extending the admission application, storage capacity and data processing, in case the number of candidates increases over time, or if it appears that the number of transactions estimated at the beginning of the current case study is not realistic. The new performance tests show that the service initially declared the winner, significantly decreases comparing to the number of operations, reaching that for 1000 operations per second, it takes over 3 times longer than the average service latency recorded concurrently. At the end of the case study, the author gives the example of a decision table to help readers to complete their approach for choosing a database service in the cloud based on the previously introduced methodology. Thus, in addition to performance, several categories with a large degree of subjectivity, like the confidence in the company, the functionality, the user assistance and acquisition costs are taken into account.

In the last chapter of the thesis, the author tries to highlight the different behaviours in various usage situations of relational and NoSQL (MongoDB and MySQL) database systems, continuing the approach begun in the previous chapter, namely running performance tests with the benchmarking tool YSSB on some database services in the cloud. After presenting the raw results obtained from running the tests, but also suggestive of their graphical representation, several indicators such as those of central tendency (mean, median) types, extreme values, standard deviation, variance types, coefficient of variation, and indicators such as arching shape or symmetry distribution are analysed. Towards the end of the study, the author tries to explain the relationship between variables using simple or multiple regression analysis.

The instrument used in the representation and processing of data is Microsoft Excel, having the data analysis module activated.

The used test scenarios are:

- Data insert (A test type) for 100 seconds
- Data insert for a longer period of time
- Data insert on the allocated memory space for the service (insert 9 GB data)
- Data read for 100 seconds (C test type)
- Data read for a long time
- Reading of stored data over the memory allocated size (9 GB)

The author draws the following conclusions through statistical analysis of test results rolled through YCSB:

- In case of data writing operations for 100 seconds, MongoDB keeps a lower average latency than MySQL
- The situation is reversed for data read operations, MySQL is apparently the better solution (limiting the target number of operations to 1000)
- It shows the influence of the target variable percentage of the number of operations per second and the total number of operations on the average latency in all test situations. Influence values are still quite small, being located between 9 and 17 %
- Since the two studied variables do not affect in a majority share the average latency, the identification of other factors could be the subject of a further study
- As in the case of read and write operation of the MySQL service, there is a negative coefficient for the total number of operations. In this

way, the greater the number of operations, the average latency should be low, probably by reducing the number of extreme values. In the case of data reading, one can advance the idea that the caching mechanism of MySQL has reduced the latency high values.

- It is confirmed the general expectation on NoSQL systems, that it has a better performance than the relational databases in the case of writing data (on the assumption of default distributions without other changes in performance)

The author then presents his own contributions to the field of study in the thesis:

- Chapter 3:
    - Making an extended description of the current situation of relational databases, including explanations of impedance mismatch problems and scalability
    - Make a list of NoSQL databases together with their main features identified in the relevant literature and the official documentation of the software systems that make the object of the study
- Chapter 4:
    - A detailed description of the MongoDB NoSQL database system together with the identification of recommended use cases
    - Demonstrating the ease to use MongoDB through the case study on the logging of data application management at the Faculty of Economics and Business Administration. In this experiment, were involved other technologies such as the web

server Tornado and Python running environment, client language JavaScript, or HTML5

- Chapter 5:
    - Presentation in general terms of cloud computing and cloud databases. The description and interpretation of economic, legal and technical aspects involved the adoption of technologies offered as services
    - Introducing a methodology for evaluating a service in the cloud database and its validation using a case study on the management of the admission process of students from the Faculty of Economics and Business Administration
- Chapter 6:
    - A realization of a practical study comparing the performance of for two systems in the cloud databases (SQL and NoSQL) with interpretation of the results using statistical tools

Finally, some of the possible research directions are listed. The author considers important issues such as the definition of generic tests for persistence in cloud services, which are easily customized by inserting the concerned database schema (even if we speak of a so-called soft-schema in the case of NoSQL) as ER diagrams or XML representations. It is also considered as a future development direction of the work, creating templates of virtual machines in the cloud for testing DBaaS, to make them available to the community of software developers.