Performance Evaluation of Non-relational databases (Cassandra, MongoDB, CouchDb and Riak): A Thesis Proposal
GROUP 31

Siddhartha Srinadhuni
9403119002
Sisr16@student.bth.se

Sai Pratheek Vasireddy
9401155446
sava16@student.bth.se

GROUP MEMBERS' PARTICIPATION

The contribution of individual group members involved in conducting the research project and reporting this document is shown in Table 1.

Group Member	Literature Review	Identify the Research Gap	Research Design	Report Writing
Sai Pratheek Vasireddy	**50%**	**50%**	**50%**	**50%**
Siddhartha Srinadhuni	**50%**	**50%**	**50%**	**50%**

Table 1. Group Member Contribution

ABSTRACT

In this article, we present the NoSQL databases and their dynamics and thus resulting in their performance evaluation. Non-relational databases have been leaving their marks of impact since the inception of web2.0. During the course of time, various non-relational databases have evolved. We propose to evaluate Cassandra [1], MongoDB [2], CouchDB [3] and Riak [4] for their performances and the effects. We propose a case study for answering the research questions that are formulated and conduct a grounded theory analysis for data analysis. Potential threats for our research that can be a part of our research have also been reflected. The introductions and their literature reviews are as follows.

Author Keywords
NoSQL Databases, Performance, Evaluation and Comparison

ACM Classification Keywords:
Information Systems: Database design and models
Information Systems: Data model extensions
Information Systems: Database administration

I. INTRODUCTION AND MOTIVATION

The revolution of NoSQL databases has been emerged from the era of Web2.0. The two major consents of web 2.0 have been micro content and social media. To enhance these two distinguished features, NoSQL databases have come to play. NoSQL are typically next generation databases that are mostly open sourced, scalable, distributed and being non-relational [5]. Performances of those with comparing different parameters are carried out as a part of this assignment. Cassandra, Riak, CouchDB and MongoDB have been compared to assess the performances and putting forth the best database among the mentioned ones.

Goal of NoSQL Databases: In terms of efficiency and managing data, relational databases have left a negative impact thus creating a void for much more efficient databases. This void is filled by "Not Only SQL" [6]. Amazon's dynamo and Google Big Table has influenced the new Information Technological enterprises to construct their own data stores and retrieve the data. This has been done by not using Structured Query Language. To give an insight, Cassandra, a NoSQL database works approximately 2500 times faster than MySQL when 50 gigabytes of data is involved.

This goal has motivated us to take up the research on this particular field of NoSQL databases and carry out evaluations to portray the efficient ones when certain parameters are involved. Given below are the background works on the databases and the results yielded on the performance criterion. The results are derived through the literature study that we have carried out and the expected outcomes for the proposal are also estimated through the best of our abilities.

Increasingly, over the last decade, software companies are moving towards being global conglomerates by adopting NoSQL databases integrated with the projects. Though there a slight question of data integrity, it is claimed to be the future of databases[7].Many software firms collect large amount of scientific, customer, sales and other data for future analysis with the help of Non-relational databases. Through this, handling unstructured data such as word processing files and emails becomes easy when it comes to handling [8]

Most generally, the NoSQL databases are divided into the following types. Wide Column store, Document store, Key stores, Eventual and consistent Key Value stores, Graph Databases. The relevant background for the databases that we have selected for the research proposal is given below.

Background and motivation

The databases that we have selected for the comparisons and evaluations are listed and the rationales for the selection is also depicted. The following are the databases that we have selected.

- Cassandra
- MongoDB
- CouchDB
- Riak

A brief descriptions of the databases used and the motivations behind us performing the evaluations are presented below.

Cassandra: According to [1] This NoSQL database is used for the management of very large amount of structured data through distributed storage system. The main features od Cassandra are providing scalability, high performance and wide applicability. The impact that Cassandra has made is phenomenal and we intended to perform an analysis through validating the results with the other three databases will give a knack to the field of Cassandra operations. Further, our research method which is a case study provides wider horizons in different scenarios as to how dynamics of Cassandra have been applied. This has been our rationale in selecting Cassandra as our subject of performing evaluations.

MongoDB:] The developer-friendly and administrator-friendly data model with its respective configurations, MongoDB is one of the most powerful and scalable data store. Major features such as indexing, stored JavaScript, aggregation and files storage constitute the MongoDB [9]. Our motivation behind selecting this database is the features such as *internal replication log, the oplog* which are one of a kind and we intended to check the parameters that are affected by these features and validate them against the rest of the subjected evaluations.

CouchDB: CouchDB has had the fame of being the new breed of database management systems. CouchDB is specifically designed to handle huge amount of traffics effortlessly. The striking features of this database can be explained through **C**onsistency **A**vailability **P**artition tolerance (CAP) theorem [10]. Also, the parameters influencing the performance are wrapped with appealing similarities with the other two databases. Hence, the quest for efficient databases with regards to performance has to be putforth and thus it became our motivation.

RiakDB: On the basis of [4], this powerful database has been promising when it comes to delivering maximum data availability by distributing data across servers. The features include ensuring low latency and robustness. [11] Riak has been supportive when considering fault tolerance and simplicity in operations. One engaging fact about Riak is that the it has its own role in BigData, Internet Of Things and Hybrid Cloud. This supportive nature compelled us to take up research on this subject.

For evaluations of databases a set of programs to assess the performance which are typically known as benchmarks [12] are used. In this scenario, for the literature, that we are evaluating, a common benchmark is used and that common workbench is Yahoo! Cloud serving Benchmark (YCSB).

Yahoo! Cloud Serving Benchmark The testing which yields the performance, displays the analysis in terms of speed of reading from and writing the data from the database. YCSB client is used for generating the operations which make up the workload. [13]. The prime feature of YCSB is that it is extensible as tailored workloads can be defined by the user through java code. The user can be placed in control when defining the number of records and also when YCSB core layer generates workload patterns. The tests and metrics generated by the YCSB are database-independent. The results are shown in the form of system read and write latencies under various aspects such as minimum to maximum latencies. Analyzations are shown in the terms of scalable and throughput.

- The only limitation of YCSB is that it does not provide a mean to evaluate one of the indispensable exchanges of modern storage systems.

Since the requirements of our literature selection has been satisfied, the below infographic explains how NoSql is an unmatchable entity while using the resources.

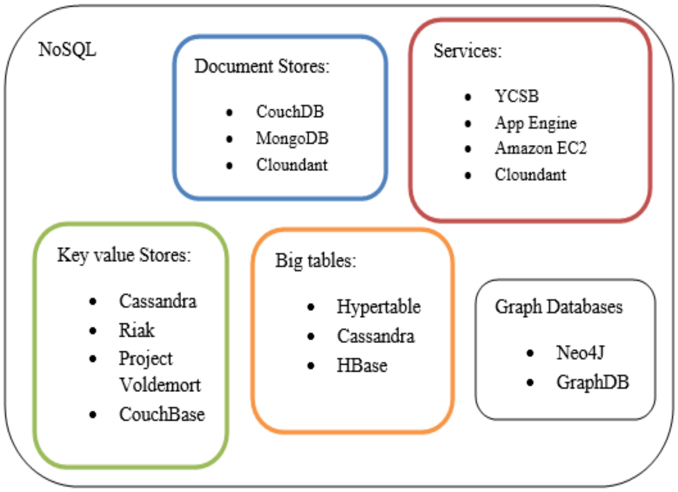


Fig1: NoSQL and their classification schematic

II. Literature review:-

The literature review was conducted by searching Scopus and IEEE Xplore database using the search string (Performance* OR Evaluation* OR Comparison*) AND (NoSQL Databases) which resulted in 96 and 70 documents and articles respectively. Out of the resulted 96 and 70 documents from the chosen databases 15 relevant articles were chosen. These articles were chosen by reviewing the abstracts. The pertinence of the literature was examined by studying the entire document which assisted us to judge the relevance of the article for our research.

Inclusion and Exclusion criteria: -

The articles were selected and shortlisted on the basis of the inclusion and exclusion criterion mentioned below:

1. The articles published from the years 2005-2016 were selected, as the concept of NoSQL was put-forth and was implemented in this period.
2. The research articles preferred for this research were conference papers, review papers, and journal articles which are available in full text.
3. The articles which were to be purchased and ordered were excluded keeping time constraint in mind and the available resources.

The rest of this section de-briefs the 15 articles bases on the following questions: -

a. The main point of the article?
b. How is the author arguing for this point based on the literature review?
c. What is the line of argumentation?
d. In what ways does the article relate to other research?
e. How does the author argue for the relevance of the research?
f. The relevance of the literature for our research?

Review of Paper 1

NOSQL EVALUATION, A use case survey: This international conference article [14] defines the evaluation of Nosql databases such as Voldemort, Redit, Riak, Membase, Mongo DB and other significant databases that have made their impacts in the era of web 2.0. These concepts and the requirements of web 2.0 motivated the authors to evaluate the performance of different NoSQL databases. This has been explicitly mentioned in the section which can be inferred as problem domain. This study has been evaluated against the results of a survey. The survey results are to be taken as a case for us, as a research method, since our method of research is a case study. The author has also mentioned the gap in research as different performance evaluations have been done in the past but there has not been a use case survey that has been performed. Hence to fulfil the gap, a use case survey has been carried out. To reflect this on his document, the author segregated different evaluation parameters as sections such as *Data Model, Query Possibilities, Concurrency Control, Partitioning and Replication and Consistency*. This particular literature is in line to our goals of finding that are determining the parameters and their effects on NoSql databases. Hence, evaluation of the databases has become a major criterion for us to choose this literature. Evaluations have been conducted on different parameters in line to answer our research questions. Cassandra, after testing proved to have an exemplar performance when compared to other databases with respect to parameters such as consistency, handling complex data structures and retrieval times. Hence, this is inline of our research to investigate different parameters that affect the performance of different NoSQL databases

Review of Paper 2

NoSQL Databases: MongoDB vs. Cassandra: Contrasts between relational databases and Nosql databases have been mentioned and thus addressing the transitions [15]. After addressing the limitations and the advantages in the line of those limitations, the author moves forward in comparing the two significant databases, MongoDB and Cassandra for their evaluations. The research method used in this research article is experimentation. The research of ours thoroughly gets supported by this literature as our findings are in line with affects that contribute to affect the performances. Several other related works have been mentioned to address the authors that have worked in the same field of NoSql databases. The line of argumentation is also mentioned to measure the execution parameters to the number of records used on each execution time. Different environments have also been selected to measure the processing time with lots of memory clusters. The goal of the paper is to increase the number of analysis and studies that already exist. This paper has also provided a reference to Web2.0 like [14] and has explained various concepts of NoSql databases. The question credible for this research is to point out an efficient database when MongoDB and Cassandra are compared. This has been a major rationale behind selecting this article as our research boils to affects in performance of NoSql. The comparisons are shown after depicting elaborately what MongoDB and Cassandra actually are. Conclusions favoured Cassandra as it performed better as it showed lower execution time irrespective of the database size used for evaluation when compared to MongoDB.

Review of Paper 3

Performance Evaluation of NoSQL Databases: In [16] The authors have collaborated to evaluate the databases and have come up with the parameters that show the significant effects on the performances of the NoSql databases. This scientific article provides paramount comparisons of the databases on a different workbench. Unlike [14] and [15], this article has taken Amazon EC2 services to evaluate the performances. The prime motivation for us to select this article is that it explored different NoSQL databases and yielded results on different parameters. We wanted to extend our research to know the parameters that affect the databases.

From this article, our findings extended to parameters such as Number of cores in a virtual machine upon which a DB is stress tested, Number of nodes and replications. Now that we have found out our parameters for evaluations, we moved forward to their affects on the databases. This paper has put forth the performances in the form of throughput time and update /read latencies for the respective parameters. Another rationale for behind choosing this article is that it deals with Cassandra and MongoDB to characterize the performance behaviour which is inline to our research.

Review of Paper 4

Application-Specific Evaluation of NoSQL Databases: In order to triangulate our research, we extended our search to application specific evaluations [17]. This international conference article was the most relevant in terms of this context as it talks about evaluations of different NoSQL databases on a particular application. An electronic healthcare record system is taken into consideration in this scenario. MongoDB, Cassandra and Riak have been used as test subjects which also happen to be our research areas. That being our major criterion for choosing this conference article, the evaluations for the article like [14] [15] [16] is based on Yahoo! Cloud Serving Benchmark. Riak and Cassandra proved to be better performing than MongoDB and the latter proved to better when compared with former but only under few conditions. Further this article has not measured the performance rigorously based on prototyping and has been limited to product specific evaluation. Inclusion and exclusion criterias have also been mentioned in the sections below to support the election of this scientific article.

Review of Paper 5

NoSQL Database: New Era of Databases for Big Data Analytics - Classification, Characteristics and Comparison: The review of [6] To further analyse the performances the databases towards entities that takes investment to manage and maintain data applications is the main line of argumentation. As our research method is a case study we wanted to investigate this scenario of big data as well and the relevancy in the line of our research, with this paper, is high. This is our primary motivation behind selecting this article for our proposal. Additionally, to answer our tentative research question of finding the parameters, this scientific articles has also been successful in comparing different articles in the terms of performance segregating them to groups like Hecht and Jablonski did in [14]**.** Different databases have been evaluated against different parameters. The research method used in this is a case study and different qualitative statistics have been derived to support the author's line of argumentation. Mentioning, the adoption of NoSQL databases from SQL databases, the author has also portrayed the limitations of NoSQL databases and this has been of paramount assistance in terms of answering our research questions.

Review of Paper 6

NoSQL databases: a step to database scalability in web environment: In [18] The author has addressed scaling of the NoSQL databases in the stream of Cloud based computing which can be inferred as the line of argumentation. In this, author has mentioned. From the inception of NoSQL to the architectures of them have been illustrated so as to implicitly address the performances in this context of cloud computing. This can also be served as a out of the box case for our case study. Additionally, we have not addressed the evaluations on a cloud based environment. Hence, this has been one such rationale to defend our selection of this article.

A set of cases have been taken upon which the scalability of NoSQL databases has been revealed and hence this article can be seen as a case study based article. This is also inline for our research as we can seek motivations from the method of study that the authors did. Also, performances have been evaluated and different approaches to NoSQL databases with more emphasis on cloud context has been carried out in this research paper. Hence this supported us to provide insights for our tentative research question in the proposal

The validity of this inclusion was further strengthened by identifying several other researchers specify communication challenges to being most critical as described below.

Review of Paper 7

A comparison between several NoSQL databases with comments and notes: Authors Bogdan George Tudorica, Cristian Bucur from Gas university of Ploiesti, Romania[19] have summed up the comparison between several NoSQL databases with various comments and notes. The comparison was done on the grounds of qualitative and quantitative point of views which constitutes various factors such as persistence, replication, Size, performance measurements and many other factors. This article focusses on the emergence of NoSQL databases in terms of handling heavy read/write workloads while replacing the ACID properties of the traditional databases with the BASE feature through comparison on various factors as mentioned above. This was identified as the research gap and motivation while conducting the literature review. Wide column store databases HBase and Cassandra were selected as the sample size from the wide range taxonomy of several NoSQL Databases and MySQL was also included in the point of comparison to test the relevance of the use of NoSQL databases over the traditional ones in this review article. The author claims that NoSQL databases are better in terms of handling huge data with complex reads/writes through which data management for large databases will be vivid and smooth. This review article presents an insight on the grounds of comparison of several evaluations for NoSQL Databases using different tools and various criterion which provides a benchmark for our research.

Review of Paper 8

A performance comparison of SQL and NoSQL databases : The article written by Yishan Li and

Sathiamoorthy Manoharan from University of Auckland, New Zealand[20] focuses on the comparison of key value stores implementation on NoSQL and SQL Databases. The main focus of this article is to potray the wide variation of performance levels of various NoSQL in comparison with SQL Databases. The evaluation of the performance is done on read, write, delete, and instantiate operations on the key-value storage. Author claims that through NoSQl databases are generally optimized for key value stores, not all of them perform better compared to the performance level of the SQl Databases which is considered to be the main line of argumentation. Finding the prime database for performing the read, write, delete, and instantiate operations with optimal performance level in the context of comparing NoSQL and SQl database is the aim of this article. Couchebase, mongoDB, Cassandra, RavenDB, MySQL, Hypertable, CouchDB were selected as the sample size from key-value stores for experimentation. On conducting the research with experimentation it was found out that Couchbase is the prime databases it is the fastest in performing read,write,delete operations and MongoDB is second in this context. This article provides insight on the drawbacks of NoSQl databases over SQL Databases and presents the optimal database for key-value stores in terms of performing various operations. It also reviews the grounds of evaluating the NoSQl databases.

Review of Paper 9

Survey on NoSQL database : Authors Jing Han, Haihong E, Guan Le and Jian Du[21] studied the basic characteristics of the NoSQL Databases and classified them in accordance to CAP theorem. The author discusses various aspects of the database technology which makes it more demanding and has an edge over selecting databases for performing heavy read/write operations. The features of NoSQL databases such as reading and writing data quickly, supporting mass storage and the ease to expand were mentioned by the author. This laid to the foundation of the classification of NoSQL Databases based on the Data model which includes key-value, column oriented and document store with the support of CAP theorem. This inference is the problem domain and the motivation behind this article. Considering the CAP theorem the author roots the classification of the NoSQL databases with the aid of a survey into the mainstream level which includes Redis, Tokyo-Cabinet and Flare into key-value store, Cassandra and Hypertable into Column-oriented database, MongoDB and CouchDB into Document database. The author concludes the article portraying various factors such as Data model, CAP support, MultiData-Centre support, capacity, performance, Query API, Reliability, Data Persistence, Rebalancing and business support involving in the selection of NoSQL databases. The implementation of NoSQL Database in cloud computing is considered as the future scope for this research. This article provides insight on the mode of classification of the NoSQL Databases on the basis of various aspects as mentioned above which provides a criterion for our research in assessing the parameters involved in our evaluation procedure.

Review of Paper 10

Performance Evaluation of NoSQL Databases: A Case Study : In paper[22], the authors John Klein, Ian Gorton, Neil Ernst, Patrick Donohoe, Kim Pham and Chrisjan Matser presented the evaluation of NoSQL databases on the basis of performance with the help of a case study. The authors moto was to replace the current RDBMS concept of the databases and adopt the NoSQL concept into EHR (Electronic health record) system. This was considered as the motivation of this research. The author proposes a method to evaluate the performance and scalability of the NoSQL databases in accordance to the need of the organisations. Choosing the prime database for an organisation in terms of performance evaluation is considered to be the problem domain of this research article. Riak, Cassandra and MongoDB were selected as the sample size, each from the taxonomy of key-value, column store and document store databases for evaluation. Evaluation criteria involves a setup of a test environment, mapping data model, Creating load test client, define and execution of test scripts. Performance and scalability criterion includes evaluation using strong consistency and eventual consistency. The results from this case study with the above criterion manifests that Cassandra provided the best overall performance in terms of latency, throughput and workload performance which suites the EHR system for better functionality. There were several evaluation methods proposed by Gorton and YCSB which were considered as the precursor for this research. Through this article we found out the depth in the evaluation criterion of the NoSQL databases. It also gives us insight on the parameters to be focussed which evaluating NoSQL databases.

Review of Paper 11

NoSQL Databases: A Software Engineering Perspective :Authors Joao Ricardo Lourenco, Veronika Abramova, Marco Viera, Bruno Cabral and Jorge Bernardino[23] have interpreted about NoSQL Databases in software engineering perspective. The author's main focus was to compare NoSQL databases in the context of real-world scenarios with real enterprise data as there were many articles stating the evaluation of these databases with pre-defined benchmarks in terms of performance and various factors. This inference is considered to be the problem domain and motivation of this article. The line of argumentation involves the inclusion of software quality attributes into the comparison criterion rather than the pre-defined benchmarks for evaluation of NoSQL databases. The software attributes chosen for the comparison criterion is as follows availability, scalability, durability, reliability, operational performance, recovery time and stabilization time. MongoDB, CouchDB, Cassandra, HBase, Voldemort, Aerospike, Couchbase were chosen for comparison from the wide range taxonomy of NoSQL databases. The author explicitly explained the characterises of the above mentioned NoSQL databases in accordance to the quality attributes. The results from this review shows that MongoDB is best suited for reliability and durability, Cassandra is a multipurpose database which mainly lacks in read performance and couch DB has the same characteristics as Mongo DB but it's availability factor is more compared to it. The future study to this research involves in evaluating these databases in the context of user experience with broader inputs. Through this article we found out that evaluation criteria for NoSQL databases has a broader perspective when we analyse them in real world contexts. This inference adds a new perspective and criterion in evaluating NoSQL Databases for our research.

Review of Paper 12

NoSQL databases: new millennium database for big data, big users, cloud computing and its security challenges: Author Asadulla Khan Zaki from BMS collge of engineering, Banglore, INDIA[24] gave a brief overview of NoSQL Databases , its characteristics and the security challenges. The author describes the need for the implementation the NoSQL databases in this modern world. As there is a huge surge of increase in the number of concurrent global users , it's difficult to manage the huge volume of data which is being collected and managed through traditional RDBMS. Thus author proposes the need of NoSQL databases which is considered as the problem domain and the motivation of this research. The author explains the transition of NoSQL databases from the traditional ones and introduces the concepts of Big data, Big users and the application of NoSQL in cloud computing. The characteristics of the NoSQL databases are explained which includes the introduction of CAP thereon where the taxonomy of the NoSQL databases id defined. The performance and scalability factors of the NoSQL databases are mentioned in this article. Author also explains the classification of NoSQL databases and defines key-value store, wide column store and document store databases respectively. The security challenges such as Transactional Integrity, Authentication Mechanisms, Susceptibility to Injection Attacks, lack of consistency and insider attacks are briefly explained by the author. The further study of this research involves in developing a security mechanism for the NoSQL database servers and scaling up the performance levels of databases in accordance to the rush in the increase of number of users. This article provides us an insight on the need of NoSQl databases in this millennium and urges to maintain stability in the overall performance of the NoSQL databases eliminating security threats.

Review of Paper 13

Which nosql database? a performance overview :In paper [25] Authors Veronika Abramova, Jorge Bernardino and Pedro Furtado gave a brief overview of the performance of NoSQL databases using Yahoo Cloud Servicing benchmark tool. The authors main focus was to propose a prime database which would satisfy the application needs and specific mechanisms to perform heavy workload operations in accordance to the user's perspective. The author's motto for this research was to compare wide range of NoSQL databases in the terms of operations as his previous work involved comparing only two databases. This inference is considered to be the motivation and problem domain of this research article. Cassandra, HBase, MongoDB, OrientDB and Redis were selected for comparison criterion in this research. The author briefly explained about the need for NoSQL databases and its classification. The research method chosen for this research was an experimental study. The experimental setup involved testing the databases in terms of workloads execution with A(50% read and 50% update), C(100% read), H(100% update)

Using YCSB benchmark tool. The results for each workload with respect to the databases and the overall execution time were mentioned explicitly by the authors. The results of this experiment portray that redis(key-value store) database has the best performance in terms of reads and updates and it is highly optimized due to mapping data into RAM. The precursor for this research was mentioned in the related works section which shows the previous studies involving the evaluation of NoSQL databases in various aspects. The future scope of this research includes further evaluation of their databases with exponential increase in the number of operations through multiple servers. This article provides an insight on the parameters involved in the performance criterion which would help us to compare the databases which aren't covered in this research.

Review of Paper 14

Comparative Study of Column Oriented NoSQL Databases on Characteristics :Authors Alireza Jomeiri, Mahboubeh Shamsi and Elham Kazemi[26] have done a comparative study on column oriented NoSQL databases on several factors and characteristics. The authors mainly focussed on explaining the properties and functionalities of the column oriented databases as it has an edge over key-value store and document databases in terms of various characteristics and advantages. Bigtable, Hypertable, HBase, Cassandra, SimpleDB and DynamoDB were selected as the sample size from the column oriented databases for this comparative study. The characteristics and dimensions chosen for the comparison criterion are consistency, availability, partition tolerance, persistence, concurrency control and replication. The databases were evaluated based on the comparison criterion mentioned above. This study will allow practitioners in choosing the best database for storage solutions. This is considered to be the motivation and problem domain for this article. The authors also mentioned various advantages of column-oriented databases over key-value stores and document databases to support the relevance of choosing column-oriented databases for this comparative study. The results of this article portray the nature of the above mentioned databases in various dimensions where each has its own relevance in various operations. This article provides a deeper insight on column oriented databases and introduces new dimensions for evaluating NoSQL databases.

Review of Paper 15

RDBMS to NoSQL: In [27] Reviewing Some Next-Generation Non-Relational Database's: Distributed storage mechanisms are a major factor for data storage for the new generation of software conglomerates. The web 2.0 reference has been taken to emphasize more on the era and the impacts of NoSQL databases. The reviews on the performances of the databases has been done and types of NoSQL databases has been presented. The author's line of argumentation lies in the implementation of NoSQL databases in grid and cloud computing. This has motivated us to take up this article to address the reviews on performances of Non-Relational databases. Additionally, this article relevancy with our research is high as the criterions of the evaluations are Cassandra, CouchDB and MongoDB. Since our method of research is a case study, this case will also serve for the study that we are going to perform for the proposal

III. PROBLEM DOMAIN :

On comprehensive analysis of the literature reviews mentioned in the above Section, it was evident that NOSQL databases play a major role in data storage and management in the current world scenario which is also termed as Web 2.0 environment. NoSQL databases are mostly classified into 3 categories namely key-value stores which includes Redis, Riak, DynamoDB, project Voldemort etc , column-oriented databases which includes Cassandra, HBase, Hypertable, Amazon SimpleDB etc and document databases which includes CouchDB, MongoDB, ElasticSearch, Orient DB etc. Each databases as mentioned above has its own importance and relevance for various instances. In order to determine the prime database from the wide range taxonomy of databases certain parameters were defined such as eventual consistency, strong consistency in accordance to latency, throughput and workload according to the article[22]. The evaluation criterion is also interpreted in the form of % of reads. Writes and updates in terms of execution time according to [25].In the article[23] evaluation was done in the perspective of software quality attributes which gave a broader overview in the selection of NoSQL databases. By taking all the criterion for evaluation into consideration we inferred that there wasn't complete coverage of all the NoSQL databases in terms of classification for comparison and the evaluation was mostly factor based which left out the scope for a wider evaluation basis consisting of real-world scenario perspective which includes many quality attributes. This inference is considered to be the problem domain for our research. To fill this gap we propose an evaluation criterion in our research which would cover wide range of factors on the basis of parameters in accordance to both real-time scenario which includes user experience and functionalities of the databases. We wish to compare Riak, MongoDB, CouchDB and Cassandra databases in our study where each one is chosen from key-value stores, column oriented stores and document stores. The motive behind choosing these databases was their prominence in the field of storage management in their own criteria. Papers [15],[23],[22]infer that the above mentioned databases perform better in their own kind and are currently implemented for data management by various MNC's such as Facebook, amazon in the current web 2.0 environment.

IV. RESEARCH QUESTIONS:

On identification of the gap in the research of NoSQL databases in the problem domain, the research question has been formulated as follows.

RQ1) What are the parameters effecting the performance of the NoSQL databases?

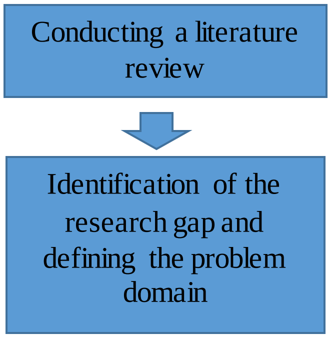
Motivations: The rationale behind choosing this question for research is to address the underlying parameters that have been effecting the performance of Non-relational databases. The extension of the answer which are the parameters points towards the affects of those parameters. Hence we formulated our second research question as follows.

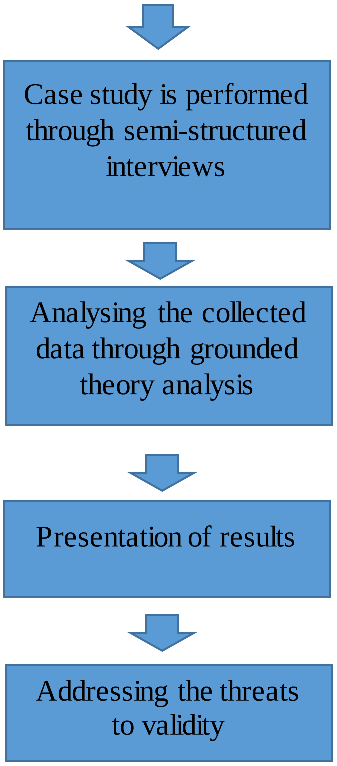
RQ2) What are the affects of these parameters on the NoSQL databases?

The research approach to be undertaken includes the research objectives and method, data collection method and data analysis method. This would lead to answering the research question and filling the research gap as identified from the considered literature review.

V. PILOT STUDY:

Our research was started initially by conducting a literature review through which we could identify the research gap and the problem domain which are mentioned in section iii and iv respectively. On defining the problem domain and identifying the research gap case study was chosen as our research method. Case study was performed in accordance to the selected cases as inputs for obtaining the results. Data was collected by interviewing practitioners which were selected as the sample size for our research. Convenience sampling method was selected for our study for precise analysis of the results. On the basis of the analysis method, results are presented explicitly. The inferences drawn from the results are mentioned in Expected outcomes section. Finally, limitations and threats to validity are mentioned in the research article.


Conducting a literature review
Identification of the research gap and defining the problem domain


Case study is performed through semi-structured interviews
Analysing the collected data through grounded theory analysis
Presentation of results
Addressing the threats to validity

VI. RESEARCH PROPOSAL:

Research Objectives :
We set our aim of the proposed research to put forth different parameters that affect the NoSQL databases and present those effects. Our aim has been successful as we have followed the objectives that we have set. They are as follows:

- Interpreting and understanding the concepts and dynamics of various NoSQL databases in terms of performance and features.
- Conducting a case study by taking inputs from multiple subjects and get qualitative data with reliability in realism.
- Source and observer triangulation are implemented to ensure triangulation
- Synthesizing and analysis of the collected data is performed.

The results would provide an analysis to portray the parameters that effects on the NoSQL databases and would set a path for the future scope.

Research Method: The goal as we have mentioned is to identify the research gap and determine the different parameters and have chosen to conduct an exploratory research as a part of this paper. Our research also works towards investigating so as to provide wider horizons of future scope and gain additional insights into the effects of the parameters that affect the performance of the NoSQL databases.
The crux of this research topic is NoSQL databases. The shift in the innovation from SQL implementing databases to NoSQL databases has to be understood from the properties that change from ACID to BASE. There is considerable amount of understanding the shift and the corresponding efforts that are involved and for this, we propose to conduct a qualitative study. We aimed to base our research method on the formulated research question[28]. Various empirical methods include survey, experimentation, case study and various other methods. We propose a literature review for the first research question and research method of case study for the second research question. The motivation behind this selection of literature review is that the selected literature has provided insights to the parameters that we are considering. The rationale behind selecting case study is that the maturity level is high, systematic approaches involved and there is a reliable way of realism [29] . Furthermore, the motivation can be supported by the fact that the questions that we have formulated can be triangulated as not many cases involve NoSQL databases and the parameters. We inferred this from the literature that we have explored.

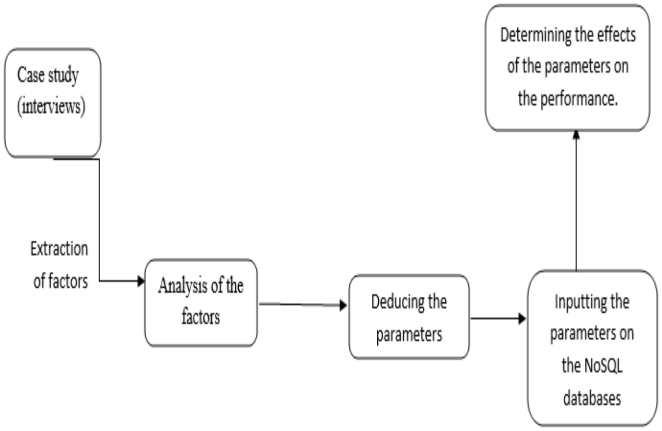

Case study (interviews)
Extraction of factors
Analysis of the factors
Deducing the parameters
Inputting the parameters on the NoSQL databases
Determining the effects of the parameters on the performance.

Fig2: Data Collection schematic

Units of Analysis: Cases are selected so as to address the factors which effect the performance of the considered NoSQL databases that are Cassandra, MongoDB, Riak and CouchDB. The threshold of the case study starts from the selection of the case study and is the sole base for the research. The considered case will therefore play a role of an input to put forth the factors affecting the case. The unit of analysis is a sample that can relevantly describe the perspective of the related population. The samples are of two types. Probabilistic and non-probabilistic sampling. We propose an alternative to the non-probabilistic sampling which is convenience sampling. Rejection of probabilistic sampling is due to lack of productivity and progress when it comes to qualitative research[30]. Here the entities under analysis are the professionals whom we interview for the cases that NoSQL and its performance is involved. These subjects are working professionals or the people with knowledge in the field of Cassandra, MongoDB, CouchDB and Riak. We propose to use experienced subjects for the extraction of case studies.

Data Collection Method: Real time systems are used as research tools for the data collection in the case study. We propose to observe the interpret the data required for the case study that is gathered from various sources from interviews and direct observations. We also would like to present our research to be triangulated and hence analysing different perspectives will enhance our knack. The data collection process will see its inception by analysing a documentations of Cassandra, MongoDB, Riak and CouchDB. The standard documentations of the databases and related authentic information are summarized for our understanding towards the knowledge.

Interviews design: When practitioner's opinion is required, interviews are taken. Here, interviews are taken to collect to information on factors affecting NoSQL databases and their parameters which effect. The motivation behind selecting Case study as interviews and not selecting survey is to get individual opinion on the performance analysis and additionally individual conversations are used to gather the information required. Semi structured interviews form the questions in the interview and are interviewed with confidentiality.

Data Analysis Method: Since our research is in an explorative structure, grounded theory is used. Grounded theory is a systematic research methodology using which a novel theory is derived from the analysis of the data. Grounded theory is said to have a superior approach for any data analysis of the qualitative data [31]. To build qualitative research on from the empirical data that is obtained, grounded theory is used [32]. Further, grounded theory provides a void for interpretation as case studies using interviews are being implemented. Given the fact that for amateur researchers, grounded theory provides supportive template for performing qualitative research [33]. Hence this is the motivation behind selecting grounded theory analysis.

This theory involves axial, selective and open coding to formulate the final and refined theory. In open codes, categorizing occurs through examination and detecting considerably high level concepts and further compared for analogous entities or differences. Thus open codes will be stated. Axial coding on the other hand restructures the data obtained from the open coding. The goals of axial coding are finding relationships and connecting the open to axial codes. The core category is where the refined theory is obtained.

VII. Discussions:-

Expected outcomes: -

On conducting the case study and comprehensive analysis of the acquired data, we desire to obtain the outcomes that would help us to identify the finest database in terms of strong consistency and eventual consistency with respect to latency and throughput, execution time for %reads and %writes, quality attributes such as operational performance, recovery time and stabilization time with respect to the operations performed and user experience. This study would help us to answer our second research question The above mentioned factors were derived from the literature review of articles [22], [25],[23] which answers our first research question These results would provide us an insight over a broader overview of evaluation of NoSQL databases and proposing a framework for its evaluation. By the end of this research we would urge a database which is consistent in nature corresponding to the factors mentioned above.

Threats to validity: -

Threats to validity are addressed by referring to the following article [34]

1. **Research Bias: -**

Though there were some initial assumptions and hypothesis about the results stated in the articles in determining the finest database, they were considered as the insights for our research in terms of selecting the parameters for the evaluation criterion. Hence research bias was mitigated.

2. **Confirmability: -**

As the sample size for our research constitutes the developers working on the databases for managing storage dependencies and different operations for various MNC's, this might be considered a threat for our research as the developers maybe biased on the determination of the finest database due to their lack of knowledge in implementing other databases.

3. **Internal validity threat: -**

The practice of convenience sampling method might be considered as a threat as people who weren't interested in the participation of the interviews were excluded as their opinion might draw a different consensus of the results. We ensured that the population of the sample was qualitative in nature so that explicit data can be collected.

Contributions and future scope: -

Our research would contribute to the field of database management as follows: -

1. We wish to propose a framework for performance evaluation criterion for NoSQL Databases.

2. Riak, MongoDB, CouchDB and Cassandra each chosen from the taxonomy from key-value stores, column oriented databases and document databases were evaluated to determine the finest database in terms of the parameters mentioned in the expected outcomes.
3. Through this research we urge to provide insights for various MNC's in choosing the NoSQL databases for data storage and management.

The future scope of this research involves in comparing each and every database from the taxonomy of NoSQL databases in terms of the factors mentioned above. This can be achieved by the repetition of this research with the involvement of multiple servers on a huge scale.

VIII. Conclusions:

In our research paper, through literature review we have answered the question of parameters that affect the NoSQL databases. The findings include Strong and eventual consistency in terms of latency and throughput. We tried the best of our abilities to triangulate the data obtained with literature review. We also proposed to conduct a case study backed up with semi structured interviews in order to obtain cases from the professionals. Our extended master thesis proposal is included in the Appendix-A.

IX. References:

[1] "Cassandra - A Decentralized Structured Storage System." [Online]. Available: https://www.cs.cornell.edu/projects/ladis2009/papers/lakshman-ladis2009.pdf. [Accessed: 22-May-2016].

[2] "MongoDB Documentation." [Online]. Available: https://docs.mongodb.com/. [Accessed: 24-May-2016].

[3] "Apache CouchDB." [Online]. Available: http://couchdb.apache.org/. [Accessed: 24-May-2016].

[4] "Riak KV." [Online]. Available: http://docs.basho.com/riak/kv/2.0.2/. [Accessed: 22-May-2016].

[5] "NOSQL Databases." [Online]. Available: http://nosql-database.org/. [Accessed: 29-Apr-2016].

[6] A. B. M. Moniruzzaman and S. A. Hossain, "Nosql database: New era of databases for big data analytics-classification, characteristics and comparison," *ArXiv Prepr. ArXiv13070191*, 2013.

[7] M. Indrawan-Santiago, "Database Research: Are We at a Crossroad? Reflection on NoSQL," in *2013 16th International Conference on Network-Based Information Systems*, Los Alamitos, CA, USA, 2012, vol. 0, pp. 45–51.

[8] N. Leavitt, "Will NoSQL databases live up to their promise?," *Computer*, vol. 43, no. 2, pp. 12–14, 2010.

[9] "MongoDB: The Definitive Guide." [Online]. Available: http://usuaris.tinet.cat/bertolin/pdfs/mongodb_the_definitive_guide.pdf. [Accessed: 22-May-2016].

[10] "CouchDB: The Definitive Guide." [Online]. Available: http://buhoz.net/public/libros/db/CouchDB-TheDefinitiveGuide.pdf. [Accessed: 22-May-2016].

[11] "Products," *Basho*. [Online]. Available: http://basho.com/products/. [Accessed: 22-May-2016].

[12] "EVALUATION OF DATABASE MANAGEMENT SYSTEMS." [Online]. Available: http://www.diva-portal.org/smash/get/diva2:367006/fulltext01.pdf. [Accessed: 21-May-2016].

[13] "Benchmarking Cloud Serving Systems with YCSB." [Online]. Available: https://www.cs.duke.edu/courses/fall13/cps296.4/838-CloudPapers/ycsb.pdf. [Accessed: 22-May-2016].

[14] R. Hecht and S. Jablonski, "NoSQL evaluation: A use case oriented survey," in *2011 International Conference on Cloud and Service Computing (CSC)*, 2011, pp. 336–341.

[15] V. Abramova and J. Bernardino, "NoSQL Databases: MongoDB vs Cassandra," in *Proceedings of the International C* Conference on Computer Science and Software Engineering*, New York, NY, USA, 2013, pp. 14–22.

[16] A. Gandini, M. Gribaudo, W. J. Knottenbelt, R. Osman, and P. Piazzolla, "Performance Evaluation of NoSQL Databases," in *Computer Performance Engineering*, A. Horváth and K. Wolter, Eds. Springer International Publishing, 2014, pp. 16–29.

[17] J. Klein, I. Gorton, N. Ernst, P. Donohoe, K. Pham, and C. Matser, "Application-Specific Evaluation of No SQL Databases," in *2015 IEEE International Congress on Big Data*, 2015, pp. 526–534.

[18] Jaroslav Pokorny, "NoSQL databases: a step to database scalability in web environment," *Int. J. Web Inf. Syst.*, vol. 9, no. 1, pp. 69–82, Mar. 2013.

[19] B. G. Tudorica and C. Bucur, "A comparison between several NoSQL databases with comments and notes," in *2011 RoEduNet International Conference 10th Edition: Networking in Education and Research*, 2011, pp. 1–5.

[20] Y. Li and S. Manoharan, "A performance comparison of SQL and NoSQL databases," in *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, 2013, pp. 15–19.

[21] J. Han, H. E, G. Le, and J. Du, "Survey on NoSQL database," in *2011 6th International Conference on Pervasive Computing and Applications (ICPCA)*, 2011, pp. 363–366.

[22] J. Klein, I. Gorton, N. Ernst, P. Donohoe, K. Pham, and C. Matser, "Performance Evaluation of NoSQL

Databases: A Case Study," in *Proceedings of the 1st Workshop on Performance Analysis of Big Data Systems*, New York, NY, USA, 2015, pp. 5–10.

[23] J. R. Lourenço, V. Abramova, M. Vieira, B. Cabral, and J. Bernardino, "NoSQL Databases: A Software Engineering Perspective," in *New Contributions in Information Systems and Technologies*, A. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, Eds. Springer International Publishing, 2015, pp. 741–750.

[24] A. K. Zaki, "NoSQL databases: new millennium database for big data, big users, cloud computing and its security challenges," *Int. J. Res. Eng. Technol. IJRET*, vol. 3, no. 15, pp. 403–409, 2014.

[25] V. Abramova, J. Bernardino, and P. Furtado, "Which nosql database? a performance overview," *Open J. Databases OJDB*, vol. 1, no. 2, pp. 17–24, 2014.

[26] A. Jomeiri, M. Shamsi, and E. Kazemi, "Comparative Study of Column Oriented NoSQL Databases on Characteristics."

[27] "RDBMS to NoSQL: Reviewing Some Next-Generation Non-Relational Database's." [Online]. Available: http://liacs.leidenuniv.nl/~stefanovtp/courses/StudentenSeminarium/Papers/DB/3.IJAEST-Vol-No-11-Issue-No-1-RDBMS-to-NoSQL-Reviewing-Some-Next-Generation-Non-Relational-Database's-015-030.pdf. [Accessed: 22-May-2016].

[28] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting Empirical Methods for Software Engineering Research," in *Guide to Advanced Empirical Software Engineering*, F. Shull, J. Singer, and D. I. K. Sjøberg, Eds. Springer London, 2008, pp. 285–311.

[29] D. Tofan, M. Galster, P. Avgeriou, and D. Weyns, "Software engineering researchers' attitudes on case studies and experiments: An exploratory survey," in *15th Annual Conference on Evaluation Assessment in Software Engineering (EASE 2011)*, 2011, pp. 91–95.

[30] M. N. Marshall, "Sampling for qualitative research," *Fam. Pract.*, vol. 13, no. 6, pp. 522–526, Jan. 1996.

[31] O. Badreddin, "Thematic review and analysis of grounded theory application in software engineering," *Adv. Softw. Eng.*, vol. 2013, p. 4, 2013.

[32] S. Looso, R. Börner, and M. Goeken, "Using grounded theory for method engineering," in *2011 Fifth International Conference on Research Challenges in Information Science (RCIS)*, 2011, pp. 1–9.

[33] J. Hughes and S. Jones, "Reflections on the use of grounded theory in interpretive information systems research," *ECIS 2003 Proc.*, p. 62, 2003.

[34] R. Feldt and A. Magazinius, "Validity Threats in Empirical Software Engineering Research-An Initial Survey.," in *SEKE*, 2010, pp. 374–379.

APPENDIX A :

We would like to propose an actual and realistic further study on this research as *‘Performance optimization of Cassandra on YCSB benchmark*, an extended master thesis proposal for future study.

Start Writing proposal	14 days
First draft submission	7 days
Second draft submission	7 days
Final draft submission	3 days
NoSQL databases training	36 days
Interview professionals for case studies	14 days
Collect data	20 days
Analyse the data	7 days
Evaluate the parameters	5 days
Evaluate the effects of the performance	6 days
Thesis documentation	21 days
Thesis documentation- second draft	12 days