DEGREE PROJECT

A method to evaluate database management systems for Big Data

focus on spatial data

Saleh Kanani

Information Security, master's level (120 credits)
2019

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering

LULEÅ UNIVERSITY OF TECHNOLOGY

Abstract

Big data of type spatial is growing exponentially with the highest rate due to extensive growth in usage of sensors, IoT and mobile devices' spatial data generation, therefore maintaining, processing and using such data efficiently, effectively with high performance has become one of the top priorities for Database management system providers, hence spatial database features and datatypes have become serious criteria in evaluating database management systems that are supposed to work as the back-end for spatial applications and services.

With exponential growth of data and introducing of new types of data, "Big Data" has become strongly focused area that has gained the attention of different sectors e.g. academia, industries and governments to other organizations and studies.

The rising trend in high resolution and large-scale geographical information systems have resulted in more companies providing location-based applications and services, therefore finding a proper database management system solution that can support spatial big data features, with multi-model big data support that is reliable and affordable has become a business need for many companies.

Concerning the fact that choosing proper solution for any software project can be crucial due to the total cost and desired functionalities that any product could possibly bring into the solution. Migration is also a very complicated and costly procedure that many companies should avoid, which justifies the criticality of choosing the right solution based on the specific needs of any organization.

Companies providing spatial applications and services are growing with the common concern of providing successful solutions and robust services. One of the most significant elements that ensures services' and hence the providers' reputation and positive depiction is services' high availability.

The possible future work for the thesis could be to develop the framework into a decision support solution for IT businesses with emphasize on spatial features. Another possibility for the future works would be to evaluate the framework by testing the evaluation framework on many other different DBMSs.

KEYWORDS: *Big data, spatial database, DBMS, GIS, selection method*

Acknowledgement

I would hereby appreciate all the kind support from my supervisor and fellow researchers at LTU and the authorities at Mobilaris AB and Simorgh Inc., who have given me the chance and courage to work on the subject that has always been of my interest.

I wish to express my utmost gratitude to Professor Ahmed Elragal for his inspirational comments and exceptional supervision, whose kind support has been given to me from all over the globe, regardless of his eventful schedule.

My special thanks goes to Mikael Nyström and Andreas Sikström at Mobilaris AB, Mohsen Rahmani at Simorgh Inc., Ali Padyab, Abdolrasoul Habibipour and Aya Rizk at LTU, whose kind support and annotations have driven me into the right perception of the business problem formulation and a proper artifact design that made the collaborative work feasible.

The joyful, collective work has led towards a method as design science artifact by which the chance to apply knowledge into a real-world business case problem solving became feasible.

Finally yet importantly, I would like to appreciate my dear wife, Shima and my dear parents for their unconditional love and support.

— Saleh Kanani

List of Abbreviations (Alphabetically Ordered)

- ***ADT* — A**bstract **D**ata **T**ype
- ***BD – B**ig **D**ata*
- ***DB* — D**atabase
- ***DBMS* — D**atabase **M**anagement **S**ystem
- ***DSR* — D**esign **S**cience **R**esearch
- ***FEDS* — F**ramework to **E**valuate **D**esign **S**cience Research
- ***GIS* — G**eographical **I**nformation **S**ystem
- ***ISO* — I**nternational **O**rganization for **S**tandardization
- ***NIC* — N**etwork **I**nterface **C**ontroller
- ***OGC* — O**pen **G**eo-spatial **C**onsortium
- ***OS* — O**perating **S**ystem
- ***OSGeo* — O**pen **S**ource **Geo**spatial Foundation
- ***SME* — S**mall and **M**edium size **E**nterprise
- ***RFID* — R**adio **F**requency **I**dentifier
- ***GNSS* — G**lobal **N**avigation **S**atellite **S**ystem

List of Figures and Tables

Table of Contents

CHAPTER ONE

Chapter 1 – Introduction

Within the following introduction, I would like to emphasize the significance of the studied field and further description of the motivation behind initiation of the research by industry as well as a brief overview of company's background and its contribution to the project.

1.1. Significance of Geospatial Information:

There are vast implementations and usages of geospatial data, ranging from geocoding, reverse geocoding, sensor data, map browsing, map search, Geotechnical usages Implications for environmental studies, such as flood risk analysis, avalanche forecast and alarm, toxic spill analysis and land information management, etc. which indicates the significance of the study.

Even though the concept of geospatial data has been around and implemented for a reasonably considerable time, but still there are many aspects, usages and benefits of such data and its derivatives which have been underestimated and hence there is a shallow knowledge among many producers and programmers of geo-spatial applications worldwide.

Thanks to the Open Geo-Spatial Consortium (OGC), comprehensive standards for utilization of spatial datatypes, functions, queries, relationship operators and indexing have been conducted in order to unify spatial features of spatial database management systems.

There are various implications and usages to spatial data, ranging from crime mapping and forecasting (Vijayakumar et al., 2014) to geological uses such as flood risk analysis (Suprio et al, 2011), Snow level monitoring and avalanche forecasting for transportation organizations and Toxic Spill (Suprio et al, 2011) Analysis for toxic spread among surface water to sending alarming notifications to mobile devices passing through dangerous firing area of a burning forest and even further utilizations such as forest and land use analysis (Trubins, 2013).

1.2. Big data Era:

The notion of "data are the new oil" has concentrated attention towards the value that data brings to the companies, governmental agencies and to the society (Couldry & Mejias, 2018). The collection and further processing of information brings strategic advantage to the organizations that leads to better efficiency, profit and support decision making processes. The term big data has defined by various authors, however, a single definition not agreed upon. In this thesis, I refer to

the definition noted by Kubick (2012, p. 26) as "data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time". Elgendy and Elragal (2014) enumerate the characteristics of big data as having 3 important aspects that needs to be considered. First is the volume, which refers to the data size and how enormous it is. Second, is the velocity, meaning that the data is changing at times, and this rate of changing is a critical factor. Last, variety of data, such as raw data, textual data, and other kinds of unstructured data, without interfering with critical production databases (EMC).

Spatial information has emerged as a prominent role player in many businesses and for the development of location-based services which has application in, but not limited to, smart cities, big data analytics, emergency response, space, health care, medical and many more yet to come (Birkin, 2019; Evans et al., 2019; Manogaran & Lopez, 2018). Lee & Kang (2015, p. 74) refer to spatial big data as "spatial data sets exceeding capacity of current computing systems". A significant portion of big data is spatial data (2.5 quintillion bytes of data is being generated every day (Carpenter & Snell (2013)), with size growing rapidly at least by 20% every year (Lee & Kang, 2015). Geospatial big data are collected from various devices such as RFID, mobile devices, unmanned aerial vehicles, cars, wearables, IoT and services such as social media, e-health and web mapping services, among others. Spatial big data has opened up a window of opportunity to the business and has created strategic advantages for many in recent years (Eldawy & Mokbel, 2015).

The cumulative volume of spatial big data, however, postures many new problems for the practitioners. For example, Ji et al. (2012), points to the issues of efficient index and query processing on largescale spatial data as a result of mismatch between DBMSs with the company's legacy systems. Moreover, Eldawy & Mokbel (2015) call for systems that are more specialized, techniques, and algorithms to support spatial big data. United nations initiative on Global Geospatial information management (UN-GGIM) predicts that the challenges of the management and integration of big spatial data will be significant (Carpenter & Snell, 2013). In this regard, the authors note that "techniques such as graphical processing units (GPUs), NoSQL and powerful in-memory SQL databases are becoming available, which will meet the demand for integrated spatial and non-spatial analytics in orders of magnitude less elapsed time" (Carpenter & Snell,

2013, p. 12). Hence, use of right information regarding handling of big spatial data at the right time, will have a significant effect for a well-informed decision-making for the developers and practitioners. Despite the advances of research in this field, there is a lack of support for the developers to be able to choose a suitable DBMS that matches their needs with the features provided for spatial big data analysis.

1.3. Promise of High-Availability:

Concerning systems demanding critical levels of availability, downtime in both measures of resistance and frequency might be a matter of seconds or even fractions of a second, hence the provider's fail to maintain high availability could possibly lead to its loss of revenue and users' and employees' dissatisfaction and more significantly, is able to cause regulatory issues to the service provider in critical and severe manner. (Schwartz, 2012) Since Confidentiality, Integrity and *Availability* are considered as the key components of Information Security (Whitman et al., 2011), hence, maintaining mission critical systems' high availability is of high significance.

There are a number of factors contributed to cause downtime and threaten information systems' high availability consequently. With respect to literature review that has been done so far on the theme, different insights in categorization among contributing elements have been found. Below are some examples of the above-mentioned insights into the issue:

According to Gartner report Scott (2001), the contributing factors are categorized into two major sectors of Human/Process failure and technology failure in which the first factor contributes 80% to the total downtime scenarios and the rest 20% goes to what is called technology failure, environmental failure or disaster.

MySQL (Schwartz, 2012) claims that the main causes of downtime accounts for: *System Failures, Physical Disasters, Scheduled maintenance* and *Operator or user errors* that are contributed to cause downtime as shown in percentages below:

- Failure and/or disaster events – 50%
- Scheduled maintenance operations – 30%
- Operator or user errors – 20%

Looking at the problem with through more holistic viewpoint (Baker et al., 2011), we can categorize the causes of downtime, technologically into two main categories:

- Planned downtime
- Unplanned downtime

Each category has a number of subclasses of failure as bellows (Baker et al., 2011):

i. Causes of Unplanned Downtime:

- Site Failure
- Cluster-wide failure
- Computer failure
- Storage failure
- Data corruption
- Human error
- Lost writes
- Hang or shutdown

ii. Causes of Planned Downtime:

- System and database changes
- Data changes
- Application changes

Since Spatial queries are considered to be among the highest resource demanding and time consuming queries (Suprio et al, 2011), optimization and conducting queries in such a way that consumes reasonable amount of resources is of high value and significance, as improperly designed, intensive queries could possibly overload the DB server and cause unavailability (Unplanned Downtime) as a consequence.

1.4. Company's Background and Contribution:

Mobilaris AB. is a mobile positioning solution provider, which handles positioning solutions in four major fields of *'Industrial Location Based Services', 'Public Safety Services', 'National Security and Law Enforcement'* and *'Commercial Location Based Services'*. The details of products and services could be found on company's website.

The company has designed and implemented two main middleware solutions, namely:

'Pacific Ocean' and *'EUNOMIA'* that could be tailored and configured to satisfy a number of purposes based on different intent and usages of positioning data.

Company's main concern is of course about providing robust and affordable solution as a package, in order to gain their provided systems' liability.

In order to approach this aim, the main objective has been defined as system's *'high availability'* since the provided system meant to handle critical positioning tasks, which in many cases may be a matter of humans' life or National Security. As explained above, maintaining high availability of the services is of significantly high value while using geospatial database therefore convinces the project's main intention.

The company would provide with two supervisors and server hardware as testing tool, which specifications would be included in the test report.

CHAPTER TWO

Chapter 2 – Research Motivation / Research Question

2.1. Business Need and Motivation:

Among several factors to be considered in the process of choosing an appropriate database type and the underlying methodology which determines the way of behavior with the spatial data. This research is an attempt to help the company filling the knowledge gap with respect to its applications' database needs, with main aim of providing a decision support mechanism that could help company's solution architects to decide upon the pros and cons of multiple available database management systems in the market. The main aim of the method is to provide with a platform in which the organization could measure the state-of-the-art features of big data provided by spatial database systems with respect to their interest and relevance of the features.

Costly licensing and skills requirements of the underlying database management systems in use by company's solutions have imposed rather high annual cost for maintenance to the clients of the company. The suggested research topic was to investigate if there is a more affordable alternative than the ones used by the company that could be utilized by their solution and at the same time satisfy the needs of the organization while reducing the annual licensing and maintenance costs.

2.2. Research Objective:

The thesis main objective is to design a method by which the system/solution architects could decide upon the DBMS solution, which suits the best for their very specific needs of their big data – Spatial application.

The method is comprised of a set of selection criteria based on the state-of-the-art features that enable database management system to handle spatial big data efficiently. The criteria has been chosen as the outcome of the literature search in the field as well as a number of business reports along with the researcher's own expectation of database systems from years of experience in the field from industry.

The ultimate use of the method would be in hands of software/system architects who are responsible for system design and making decisions about the development language for the solution as well as database systems among other factors to be decided in the process of design.

According the fact that technology for handling big data and spatial data are changing rather frequently in database management systems and therefore, there are many novice solution being introduced to the market that make the process of choosing the proper solution more complicated than ever.

The method is utilizing recent features provided by different DBMSs' and even enables the user to add their own selection criteria to the method that brings a level of flexibility to the usability of the method.

2.3. Research Delimitation:

This research is considering the database management systems' features that mostly are involved in handling big data with special focus in spatial data, therefore any other DBMS features are outside the scope of the study.

Other features provided in database systems are considered out-of-scope for this particular study and can be studied as future research possibility that will generalize the usability of the method.

2.4. Research Question:

Following is a list of research questions that will be addressed through literature review that forms the grounds for the design artifact through DSR methodology.

- What are the criteria to evaluate state-of-the-art DBMS for big data, with focus on spatial type?
- How could a software/solution architect decide upon a suitable solution for their DBMS?
- How to evaluate and compare different DBMS solutions based on a specific application needs?

CHAPTER THREE

Chapter 3 – Literature Review

3.1. Literature Review Framework

For reviewing the relevant literature, I have adopted to the framework for literature search process as suggested by Vom Brocke et al. (2009)

According Vom Brocke et al. (2009), the process of literature review should be well documented and rigorous, which is achievable through the framework they provided as follows.

The process as shown in figure5 consists of five identical steps that form the framework.

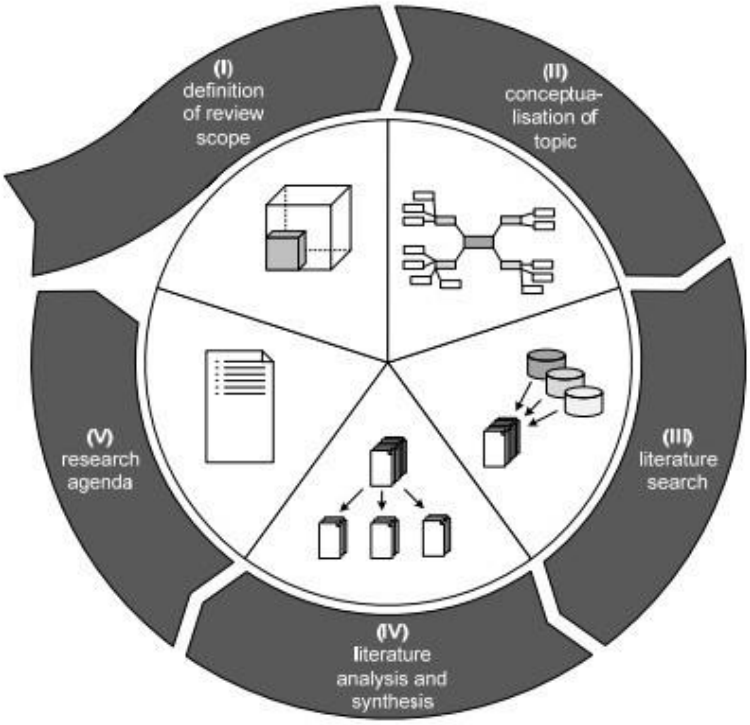

Figure3.1. *Framework for literature reviewing* (Vom Brocke et al., 2009)

3.1.1. Definition of review scope:

The first step in the LR work would be to define the scope of the review that is actually defined by what is the most important criteria and scope in the search and review process of the research.

The other factor to be criteria in this stage of the literature review is to set the period through which the results of the search would be narrowed down chronologically within the years 2015 to 2019 for the majority of the literature material.

Technology related literature are chosen to be among the most recent papers and the theoretical research papers have a broader time frame since the technology related material are generally changing with a higher rate than more theoretical material for the theoretical background frameworks used for this research.

3.1.2. Conceptualization of topic:

A major challenge in literature review is actually to conceptualize and define the scope and the essence of the research. In this research, the focus is to formulate the state-of-the-art in the field of big data and spatial type as well as the new features provided by DBMSs.

The main research topic would be the state-of-the-art spatial big data, features of database management systems for handling big data of type spatial for serving as database backend for geographical information systems. The result would then support the development of the evaluation method by identifying the new features of DBMSs as evaluation/selection criteria by adding a relevance factor that needs to be set by the person using the method in order to set the level of relevance of a certain feature to the very specific needs of their organization.

3.1.3. Literature Search:

Literature search process takes place online through Elsevier's abstract and citation database "Scopus" search engine.

The main topic for the search are as follows, therefore keywords used for the literature search initially are:

- *"Big Data"*
- *"Spatial"*
- *"Database"*

3.2. Literature Search Strategy:

Considering the fact that the research topic suggests multi-discipline search areas, the chosen keywords have been used in different forms of terminology that are commonly used in the academy papers i.e. combining the keywords big data, spatial data and database.

There are a number of different terminologies used in the research papers, therefore the keywords are used in combination so it covers up the whole spectrum e.g. "Spatial big data", "big spatial data" or "database" in parallel with "DBMS".

In accordance to the above-mentioned strategy to combine the "similarly-used" terms in literature, the following search queries are generated in order to narrow down the scope of the literature search and not missing relevant research papers due to use of different forms of terminology in different research papers due to not using similar words.

3.3. Literature Search Queries:

The First search query consists of keywords "big data" and "database" which gives 7,090 hits. The query is limited to return the papers that have both phrases "big data" and "database" among either their title, their abstract or keywords. The query is as follows:

Q1:

```
TITLE-ABS-KEY ( "big data" AND "database" )
```

The second search query would include "database" keyword so that the results would have considered the concept of databases in their abstract, title or keywords. The second query hit 390 results.

Q2:

```
TITLE-ABS-KEY ( "big data" AND " database " AND "spatial" )
```

Next, in order to broaden the chance of finding relevant literature, the keywords would be expanded to similarly-used keywords in the literature i.e. putting keywords "database" along with "DBMS" connected with OR operation by which different terminologies of the same concept would be covered in the literature search. Another example of such

generalization in literature search is using "GIS" and "geospatial" together with "spatial" keyword for a higher chance of relevant literature findings. This query gives 504 hits.

Q3:

TITLE-ABS-KEY ("big data" AND ("database" OR "DBMS") AND ("spatial" OR "geographic" OR "GIS" OR "geospatial"))

Knowledge Database	Queries/Number of hits		
	Q1	Q2	Q3
Scopus	7,090	390	504

Table 3.1. *No. of hits per query - before delimitations*

3.4. Literature Search Delimitations:

The process of delimitation of literature search results are performed through means of two major criteria of recent publish time, which enables the researcher to perform the search and therefore gather research data that is state-of-the-art.

3.4.1. Delimitating by Time:

Performing a research in such a novel, vibrant and changing field demands the information to be updated and recent, therefore the time scope for the literature search have been set to the recent five years, being from 2015 to 2019.

The two above-mentioned criteria would helped me to find proper literature that is recent and relevant to the research area and research question.

The result of such delimitation have resulted in an updated literature search query as follows:

Q4:

TITLE-ABS-KEY ("big data" AND ("database" OR "DBMS") AND ("spatial" OR "geographic" OR "GIS" OR "geospatial")) AND (LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015))

The above delimitation factor have reduced the search hits to 426.

3.4.2. Delimitating by Relevance:

Considering delimitation by relevance, I have put two rounds of delimitation, one through updating the search query by adding a further search delimiter by narrowing down the subject area to "Computer Science" and "Decision Science". This would narrow down the search-hit number even further to 323.

Q5:

```
TITLE-ABS-KEY ( "big data" AND ( "database" OR "DBMS" ) AND ( "spatial" OR
"geographic" OR "GIS" OR "geospatial" ) ) AND ( LIMIT-TO ( PUBYEAR , 2019 )
OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-
TO ( PUBYEAR , 2016 ) OR LIMIT-TO ( PUBYEAR , 2015 ) ) AND ( LIMIT-
TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "SOCI" ) )
```

Knowledge Database	Queries/Number of hits	
	Q4	Q5
Scopus	426	323

Table 3.2. *No. of hits per query - after delimitations*

The output of the above process has led to a number of 323 papers containing the search keywords and all published between 2015 and 2019. The next step in delimiting found papers will be by excluding the irrelevant papers yet by skim reading the papers abstracts and likewise selecting the most relevant pieces of work by findings throughout the study of their abstracts.

The eventual result of the literature search process is demonstrated in *table3.*

In addition to these research papers, I have employed backward/forwards search strategies in the areas that appeared significant to be followed up in the source papers.

In addition to the above-mentioned resources, a number of business reports, whitepapers, video presentations, lectures and seminar videos have been used as other types of resource.

3.5. Selected Research Papers:

The following table includes papers from the above literature search process as well as forward/backward search through their references as well as methodological papers used for the research.

The Last delimitation in the process of delimitating the research papers to include in the research

Is by ready through the remaining 323 research papers from Q5 and narrowing down the literature search results even further by studying the abstracts of the papers. In case the findings from abstract appeared to be relevant to the research are, then I would go through the body of the paper to ensure the

3.5.1. Further exclusion criteria:

After implementing all previously mentioned exclusion criteria of **keywords**, **ageing**, subject area of the research papers, the following criteria has been applied to the remaining papers, and the result has formed as shown in *table 3.*

1. Only documents in English language has been chosen
2. Project Reports: The project reports have been excluded from the literature search
3. Pure mathematical rather than socio-technical: in further exclusion of irrelevant papers, the papers concentrating in mathematical and rather engineering papers have been excluded and socio-technical papers and the ones emphasizing on socio-technical aspects of the field remained in the list.

No.	Research paper title	Author(s)	Journal/Conference
1	SQL versus NoSQL Databases for Geospatial Applications	Baralis and Rossi (2017)	IEEE International Conference on Big Data (BIGDATA)
2	Interactive and Scalable Exploration of Big Spatial Data - A Data Management Perspective	Sarwat (2015)	16th IEEE International Conference on Mobile Data Management
3	NoSQL Database: New Era of Databases for Big data Analytics -Classification, Characteristics and Comparison	Moniruzzaman and Hossain (2013)	International Journal of Database Theory and Application
4	Data-intensive applications, challenges, techniques and technologies: A survey on Big Data	Chen & Zhang (2014)	Information Sciences
5	Performance Evaluation of Querying Point Clouds in RDBMS	Ikawa et al (2019)	IEEE International Conference on Big Data and Smart Computing, BigComp
6	Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data	Thatcher (2014)	International Journal of Communication

7	Multi-model Data Management: What's New and What's Next?	Lu & Holubova (2017)	20th International Conference on extended Databases
8	A Study on Data Input and Output Performance Comparison of MongoDB and PostgreSQL in the Big Data Environment	Juang et al (2015)	International Conference on Database Theory and Application
9	Geospatial big data handling theory and methods: A review and research challenges	Li et al (2016)	ISPRS Journal of Photogrammetry and Remote Sensing
10	A BRIEF REVIEW ON LEADING BIG DATA MODELS	Sharma et al (2014)	Advance Publication, Data Science Journal
11	Indexing techniques for advanced database systems	Bertino et al. (2012)	Springer Science and Business media
12	Performance Evaluation of MongoDB and PostgreSQL for spatio-temporal data	Makris et al (2019)	EDBT/ICDT 2019 Joint Conference
13	Big data DBMS assessment: A systematic mapping study	Ortega et al (2017)	International Conference on Model and Data Engineering
14	Geographical information system parallelization for Spatial big data processing: a review	Zhao et al (2016)	Cluster Computing Journal
15	Evaluating query performance on object-relational spatial databases	Zhou et al (2009)	ISPRS Journal of Photogrammetry and Remote Sensing
16	GeoMesa: a distributed architecture for spatio-temporal fusion	Hughes et al (2007)	SPIE Defense and Security
17	Future trends in geospatial information management: the five to ten year vision	Carpenter & Snell (2013)	UN - GGIM
18	GeoSpark: a cluster computing framework for processing large-scale spatial data	Yu et al (2015)	International Conference on Advances in Geographic Information Systems
19	Survey on NoSQL Databases	Han et al (2011)	IEEE - International Conference on Pervasive Computing and Applications
20	Comparison of Relational Database with Document-Oriented Database (MongoDB) for Big Data Applications	Chickerur et al (2015)	International conference on advanced software and its applications

Table 3.3. *Literature search - selected papers*

3.6. Literature Review Synthesis:

3.6.1. Literature search findings:

The most significant technologies and feature involved in the state-of-the-art big data of spatial data nature are discussed in a large number of the literature search findings that are described as below:

As Lu & Holubova (2017) argue the classification of database systems into multi-databases and single-databases based on multi-model capabilities of DBMSs' and emphasizes on multi-model capabilities of database systems as a critical support.

Therefore, Multi-Model support has been chosen as one major criteria for the selection method.

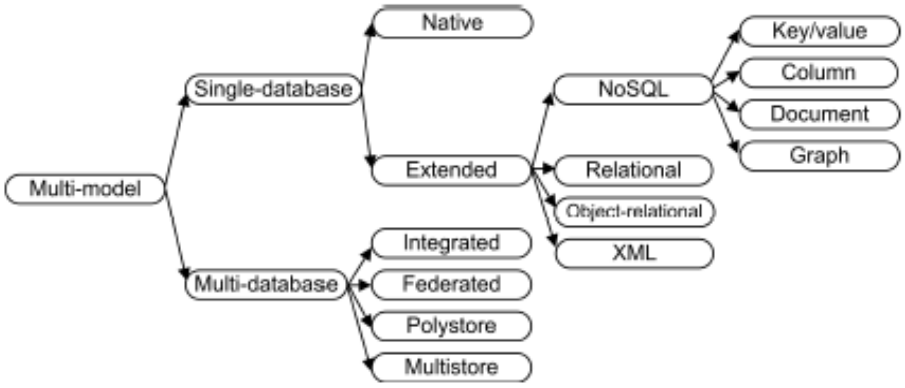

Figure 3.1. *Classification of multi-model data management (Lu & Holubova, 2017)*

Feinberg et al. (2015) has also emphasize the market urge for DBMS solutions to provide multi-modeling to a level that they estimate that by the year 2017 all leading operational database management systems will support some degree of multi-model support.

Hughes et al (2019) demonstrates the significance of key-value pair stores in big data and endorses the applicability and implication to spatio-temporal big data through a distributed architecture.

The importance of NoSQL applicability, in big data analytics is very well emphasized by Moniruzzaman & Hossain (2013), as the suitable database solution structure as opposed to classical relational database management systems (RDBMS). Baralis et al. (2017) have also emphasized the usability and significance of NoSQL and Cloud support for big data specifically for spatial data nature. Moniruzzaman & Hossain (2013) emphasize on graph database as being categorized in NoSQL databases to be similar to object-oriented databases as they are represented as a network of nodes. Graph databases are especially useful when the relationships between the data is more important than the data itself.

Chen & Zhang (2014) bring an example of NoSQL (Not-Only SQL) employment in apache Hadoop's HBase and argues that NoSQL employs a number of approaches that facilitate big data. Firstly, by separating the data storage and management into two separate parts in contrast to traditional relational database management, which combines the two. The segregation of data storage and management would facilitate scalability and high performance of the database system, therefore key-value pair and NoSQL support are key factors according to literature.

Sarwat (2015) addresses the scalability of spatial data systems and performance as major challenges and claims that cluster computation paradigm is a key to interactive spatial data exploration. Juang et al. (2015) and Makris et al. (2019) have also done a comparative benchmarking on MongoDB as the representative of a NoSQL database and PostgreSQL as a representative of a relational database management system. The study yet signifies the level of attention for application of NoSQL and its role in state-of-the-art BD.

In another feature comparison attempt Ikawa et al (2019) has made a benchmarking attempt on a sample data set with main aim of making a comparison between a B-Tree traditional three dimensional relational database systems opposed to spatial features of PostGIS add-on of PostgreSQL using R-Tree indexing method for spatial data type. The Research would further compare response time of different setups of the same data set and it claims that for queries returning 23 Million rows, the spatial R-Tree is performing faster than the traditional RDBMS implementation setting. The above-mentioned finding would then justify the spatial types and features for large volumes of data; hence, the significance of the spatial types, indexing and functions seems inevitable for big data of type spatial for GIS.

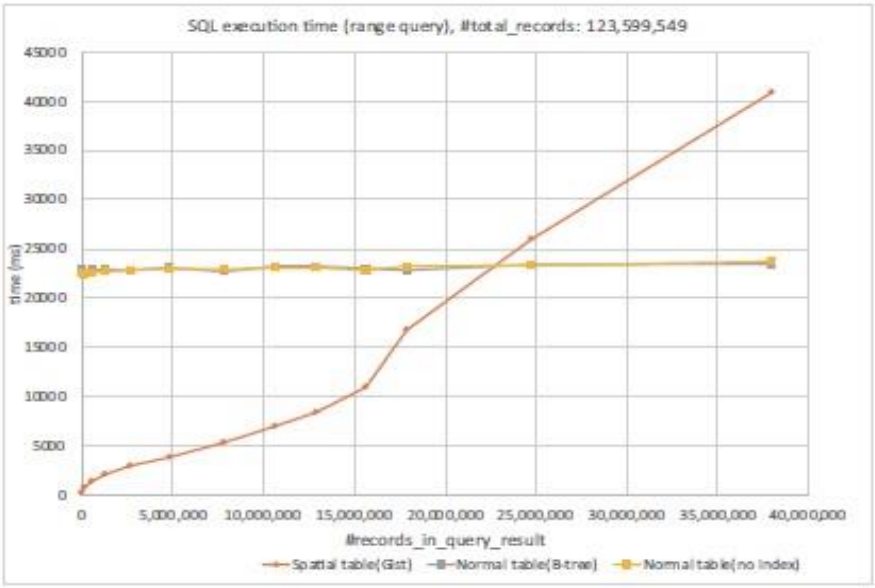

Figure 3.2. *Execution time of range query with the large set (Ikawa et al., 2019)*

Indexing is considered as a serious challenge in big data as argued by Sharma et al. (2014) yet indexing significance has been proven in databases particularly in dealing with extremely large

volume of data (Bertino et al., 2012), (Thatcher, 2014). The significance of spatial type indexing (R-Tree) comes into the light of importance.

Li et al. (2016) argues that new spatial indexing and algorithms are the new challenge for real time analytics of big data of spatial nature as further development in the field.

The systematic mapping study of Ortega et al. (2017) shows the significance of different aspects of big data capacity that impose new needs for handling big data in database management systems. The study also pinpoints the significance for industries to have quality DBMSs that are capable of handling big data in such growing volumes and in real time.

Yu et al. (2015) and Zhao et al. (2016) both argue that parallelization and distributed cloud parallel computing are highly crucial features of database systems dealing with big data and large scale spatial data.

Zhou et al. (2009) have performed a benchmarking of spatial functions through equal data set imported to databases MySQL, PostGIS, DB2 and Oracle Spatial in order to measure the query processing time of equal queries running on the above-mentioned DBMSs that would indicate the necessity of spatial functions in spatial databases.

Carpenter & Snell (2013) have addressed the challenges introduced by new trends and usages in today's geographical information systems. Among the discussed topics, the challenges of managing, integrating and analyzing such an exponentially growing volume of data in a timely manner and technologies related to it e.g. NoSQL databases and Cloud computing. The research also discusses the licensing and maintenance expenses as well as challenges due to policies and standards of managing geospatial big data.

In comparative studies e.g. Han et al. (2011) and Chickerur et al. (2015) that have aimed to compare two major categories of relational databases and document-based databases, the importance and its capabilities of document databases handling semi-structured databases are examined in comparison to the relational traditional databases through benchmarking that would emphasize the importance of such genre of database management systems.

3.6.2. Knowledge gap in literature:

Due to the findings of the literature search and synthesis of the research papers in the field of big data with focus on geospatial type, it has become inevitable that each research attempt has addressed a certain challenge and accordingly provided with discussing certain features or technologies connected to it. (*Figure 3.4*)

What is missing in the literature is a comprehensive solution that centralizes and gathers all the possible challenges in applicability of database systems with respect to usage of storing and retrieval of big data of geospatial nature, therefore the need to develop a research by which the industry and academy could benefit is expected.

The need of systematic evaluation method that centralizes the challenges and features provided by database systems of the state-of-the-art would be beneficial to both industry and academia.

Selection Criteria / Literature	Spatial Data Types, Indexes and Functions	Polyglot vs. Multi-model Databases	Parallel partitioning	Columnar/In-Memory	Graph database	Document Database	Key-value Pair Store	NoSQL Support	Cloud Support	Integrated Querying	Licensing Costs	Maintenance/Skill Requirements
Baralis & Rossi (2017)								✓	✓			
Sarwat (2015)			✓						✓			
Moniruzzaman & Hossain (2013)					✓			✓				
Chen & Zhang (2014)							✓	✓				
Ikawa et al (2019)	✓											
Thatcher (2014)	✓											
Lu & Holubova (2017)		✓										
Juang et al (2015)								✓				
Li et al (2016)	✓											
Sharma et al (2014)	✓											
Bertino et al. (2012)	✓	✓										
Makris et al (2019)								✓				
Ortega et al (2017)												

Zhao et al (2016)			✓									
Zhou et al (2009)	✓								✓			
Hughes et al (2007)							✓					
Carpenter & Snell (2013)	✓		✓	✓				✓	✓	✓	✓	✓
Yu et al (2015)			✓	✓					✓			
Han et al (2011)						✓						
Chickerur et al (2015)						✓						

Table 3.4. *Literature search – selection criteria correlation*

3.7. State-of-the-art GIS:

Geographical Information Systems (GIS) have made a vast progress during the recent ten years and new features and usages are yet to come by individual initiatives and organizations. Spatial data types are consisted of two major categories of data, namely Raster and Vector.

Raster data are discrete, continuous data that is usually presented as colorful pictures taken from satellites to work as the map background of today's geographical information systems e.g. Google® maps or Google® earth as well-known online services, which are proprietary and OpenStreetMap® as open source online map services. Following are examples of raster data:

Figure 3.3. *Basins Raster Map* (PostgreSQL Manual)

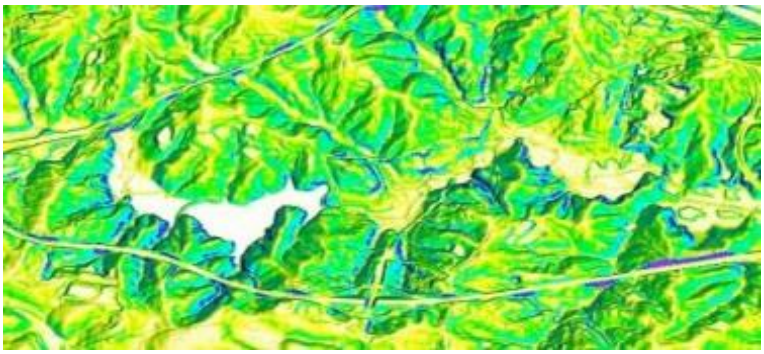

Figure 3.4. *Slope Raster Map* (PostgreSQL Manual)

Vector data are additional layers of data that may come in shapes of polygons, lines, curves or points with attributes as benefit to the plain background picture of the map. The data such as streets and street names, areas such as cities, regions, counties or countries presented as named polygons or multi-polygons, country boards as an example of curves etc. Following are samples of vector data represented on a map:

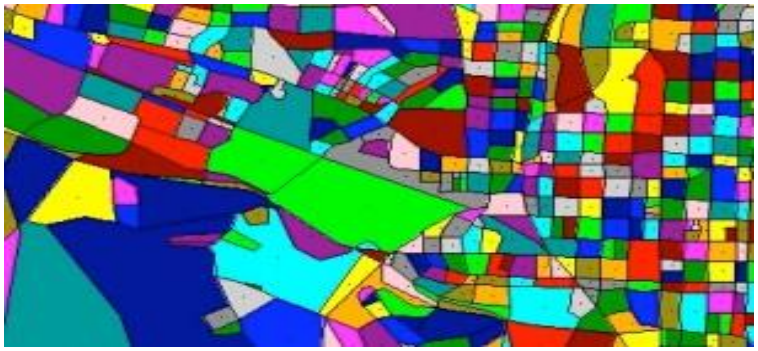

Figure 3.5. *Census Vector Layer* (PostgreSQL Manual)

Figure 3.6. *Lines, Curves and Polygons as Vector Layers* (PostgreSQL Manual)

The following example also shows a layer of vector data on top of a Raster image as a combinational representation of both data types viewed in a map:

Figure 3.7. *Combination of Vector and Raster Layers* (PostgreSQL Manual)

3.7.1. Evolution of GIS:

The three generations of GIS is shown in *figure 4.6* and described accordingly as argued by Davis et al (1998).

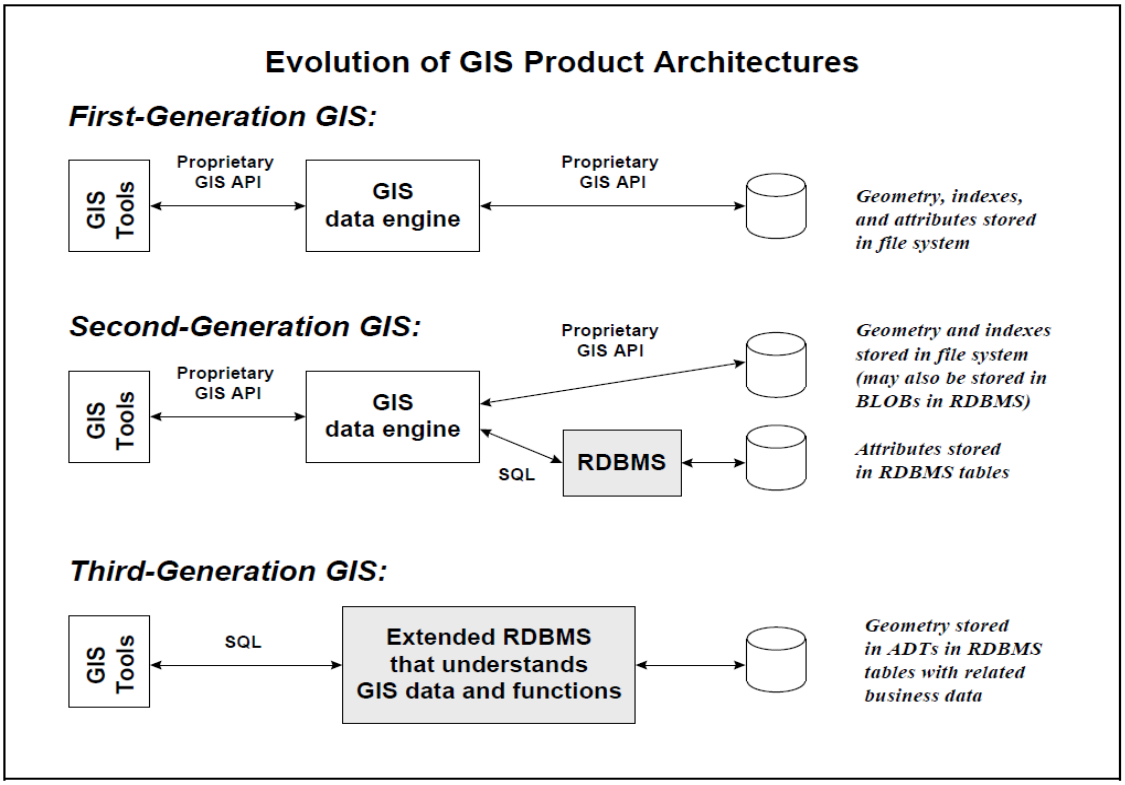

Figure 3.8: *Evolution of GIS Architectures* (Davis et al., 1998)

GIS Systems' Backend Components Evolution:

1. GIS Data Engine + File System.
2. RDBMS alongside with GIS Data Engine + File System.
3. Stand-alone Spatial RDBMS.

As shown in the above picture (*figure 4.6*), in the third generation of GIS, Spatial Database serves as the backend for *Geospatial* applications and geographical data is stored as *Abstract Data types* (ADT) in spatial database tables, in line with other additional business data.

3.7.2. OGC Compliance and SQL/MM Spatial Standard:

The OGC compliance open standard assures that our new system will be platform- and hence vendor-independent which gives the possibility to switch over to other platforms and/or vendors if needed over time without the need to rebuild the whole system from scratch. The international open standards are mainly concerned with providing a unique and open standard for solutions worldwide to ensure transition between the vendors that implicate the standard, with minimum amendments.

Open Geospatial Consortium (OGC) is the only recognized open international organization for geospatial information standardization and choosing vendors that follow the standard will have a considerable impact on minimizing the cost overhead enforced by platform change in case of need.

PostgreSQL and its geospatial extension package (PostGIS) follow the OGC open standard and hence they provide with further flexibility for the company's solution to be vendor-independent which is considered as one major advantage in the process of choosing the right open-source option available.

The other key factor in this chapter is ISO/IEC 13249-3 standard compliance to be considered. Another advantage of using a solution e.g. PostgreSQL's PostGIS is that it follows the SQL/MM or ISO/IEC 13249-3 standard – Part 3 as their standard for the SQL commands in use for spatial data. This feature enables the back-end programmer to choose their standard SQL commands, data types, procedures etc. for their solution with peace of mind that they can use the same queries in any other database management system that follows the same standard without any major compatibility concerns.

ISO/IEC 13246-3 SQL/MM Part 3: Spatial (Ashworth, 1999) is an international standard that includes the definitions of storing, retrieving and processing spatial data by the Structured Query Language (SQL) and also defines the structure of storing and representing spatial data as well as the available functions for conversion, comparison and processing of spatial data in a number of different ways. (Stolze, 2003)

3.7.3. Spatial vs. Two-Column Model

There are two main models to implement geo-spatial data in a database: (Suprio et al, 2011)

- Two-Column Model
- Spatial functions/data types indexing features

I. Two-Column Model:

The two-column model is a simple and efficient model, which stores geocoded location of a zero-dimensional geographic object (points) that has its own benefits, limitations and drawbacks.

The major benefit of the two-column model is its simplicity and hence efficiency and its major drawback is its limited nature in storing 1-dimensional and two-dimensional geographical objects position (Points, Lines/Curves and Surfaces rather than rectangular shapes). (Suprio et al, 2011) , (Aitchison, 2009)

In other words, using spatial data types, our system would be capable of store, process and query Lines and curves (1-Dimensional), Surfaces e.g. polygons (2-Dimensional) as well as bulk objects e.g. 3-D or "XYZ" mode shapes (3-Dimensional), rather than limiting our ability to save only points (zero-Dimensional) objects in 2-column model.

The company is already working on a project that includes CAD 3D maps and the above-mentioned features e.g. spatial-specific indexing features, could be a remarkable alternative to using BLOB data in a data engine.

Below is a simple (common implementation) table, designed in two-dimensional model, capable of storing location (Latitude and Longitude) of geocoded objects as coordinate value along with its corresponding name and address (Aitchison, 2009):

```
CREATE TABLLE Customers (
Name varchar (32),
Address varchar(255),
Lat decimal(18,9),
Long decimal(18,9)
);
```

Drawbacks to the two-column model are as follows: Firstly that the only implementation of surface would be rectangular areas which would in turn make it impossible to imply a search or any other query within any other shape of area such as polygons or multi polygons or any other complex shapes.

Secondly, the other possibility while using the two-column model is to calculate distance between two objects (points), by the *'spherical laws of cosines'*. Meanwhile there is a drawback of measuring distance between two points using this method is that the output would be an approximation rather than exact numbers, because of the shape of the earth not being a perfect sphere.

Below is a common implementation of spherical law of cosines for measuring the distance between two points of: *a* and *b*, holding position values of (Lat1, Lon1) and (Lat2, Lon2).

First and Second generation of GIS used the following method which lacks accuracy.

```
SELECT 3963.2 * ACOS (
SIN(Lat1) * SIN(Lat2) +
COS(Lat1) * COS(Lat2) * COS(Lon2 – Lon1)
);
```

Note: *3963.2* is the radius of earth in miles and it could be 6378.1 in scale of kilometers.

II. Spatial Model:

Using Spatial data types, functions and indexes would enable applications in order to store and proceed multiple functions and comparisons between multiple types of complex objects which brings outnumbered benefits and additional value to spatial application which in return, makes the query value much higher that the functions categorized in this class of queries, are classified among

the highest resource (memory) intensive queries, hence the query optimization and indexing become of high significance in overall performance.

Memory intensity and time consumption is going to be our main concern for the design of benchmarking in order to examine the amount of effect the platform change would affect the demand for resources in the system servers.

III. Spatial Data and Methods:

Spatial data in the context of this study refers to a variety of different geometries e.g. points, lines, curves or polygons and/or their composites such as multi polygons multi curves etc. as per SQL/MM standard specifications and hence applicable to a variety of data spaces.

One abstract definition of term *geometry* (Stolze, 2003) used by cartography is referred to as a point or a collection of points which has a reference on earth as their representation, therefore a geometry is often a representation of a geographic feature.

The following figure shows the hierarchy of a two dimensional geometry type based on the ISO/IEC SQL/MM standard.

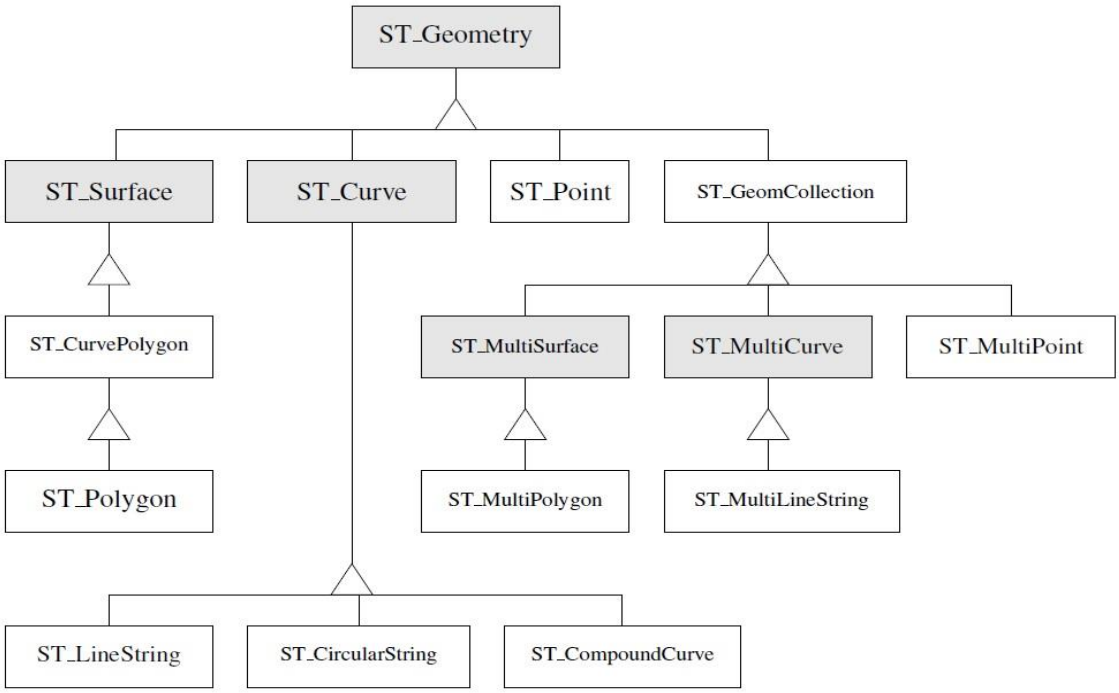

Figure 3.9. *SQL/MM Geometry type hierarchy* (Knut Stolze, 2003)

As shown in the above hierarchy figure, there are a numerous expansions with respect to the available application and representation different geographical references as compared to the two-column model in the first generation GIS.

Spatial Data can be used for representation of zero, one, two or three-dimensional geometrical objects. The zero-dimensional objects being points are able to be constructed using two-column model, by using latitude and longitude of the object point. The one-dimensional objects such as lines or curves could be easily implemented through geometry data types whereas it is close to impossible with the two-column model since you need to store indefinite number of point that is the only representation with use of coordinates, in order to be able to represent a line string or a curve. The same disadvantage comes into place while implementing a two- and three-dimensional object such as polygons that are easily accessible in the spatial model yet again close to impossible when it comes to two-column model.

In addition to the lack of presentation tools for one- and two-dimensional geometrical objects, e.g. roads, rivers, countries, etc. there are obviously numerous deficiency of applicable methods for manipulating and/or querying the spatial data in two-column model whereas a wide range of methods are available in spatial data for this reason.

CHAPTER FOUR

Chapter 4 – Methodology

4.1. DRS: Two Schools – One Choice

Based on Alan Hevner's three-cycle view of design science research (DSR), the methodology consists of three closely related cycles of activity (Hevner, 2007).

Since the research question suggests a method to evaluate DBMSs with respect to spatial nature of data, the method itself is considered the designed artifact in the thesis.

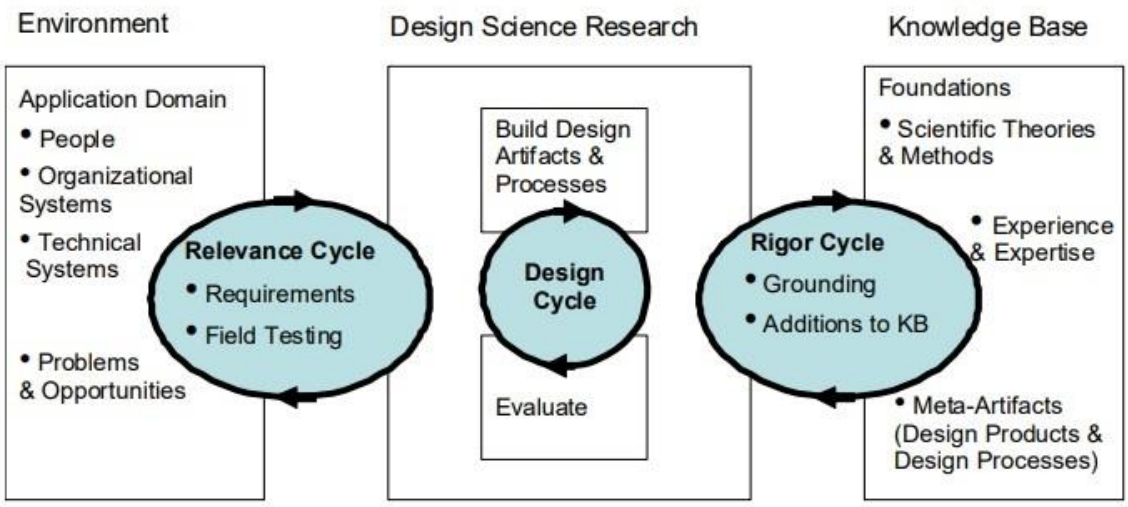

Figure 4.1. *Design science Research Cycles* (Hevner, 2007)

A proper DSR is supposed to begin by identifying the problems and opportunities in the application environment of the research area.

The problem was initially proposed in the form of comparison between two DBMSs that the company initially facilitated, in which case I had to make a lab experiment due to their needs of spatial data and functions.

The experiment could be performed on the same platform with the same test scenario, but different DMBSs in order to evaluate and compare efficiency and performance of each product.

In order to achieve a level of generalizability in the outcome of the research, I decided to take design science and design a framework by which any other DBMS could be evaluated based on specific needs of any other application.

The research process of design science will take place based on the stepwise methodology introduced by Peffers et al. (2007).

As stated in figure2, Peffers et al. (2007) defines the nominal process sequence of DSR process model, which goes through a process iteration.

Each process sequence will be discussed and explained as follows:

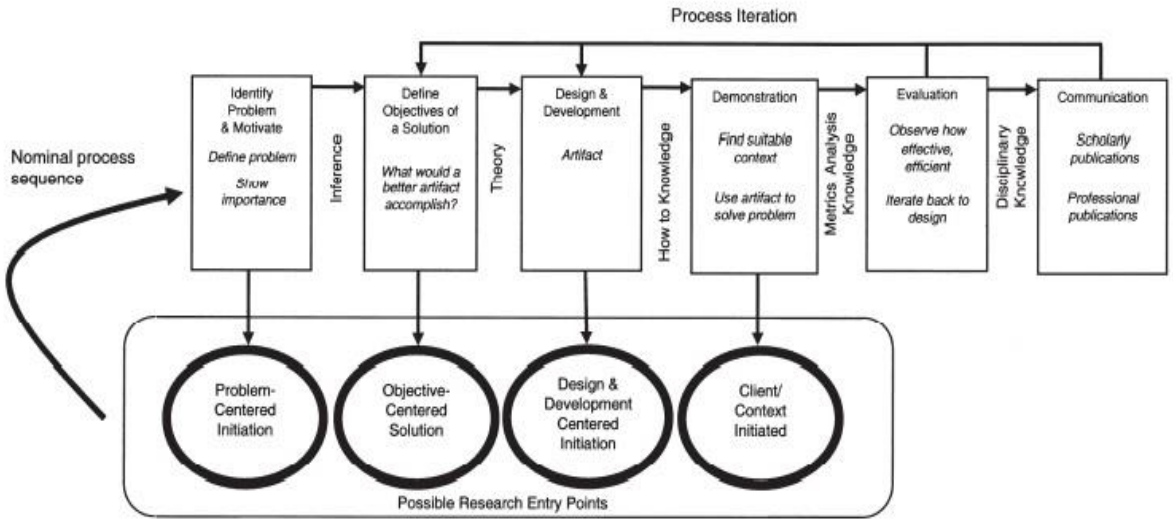

Figure 4.2. *DSRM Process Model* (Peffers et al., 2007)

As the first step in the model suggests, one needs to initiate and identify the problem and to motivate the research question by showing the significance of the study. The entry point to this step would be a problem-centered initiation.

The research question initially arose from business with initial intention of overcoming a commercial situation that the company have been facing due to additional annual licensing fees their products imposed to clients by using a certain DBMS solution.

The external supervisor initially aimed the research to pinpoint whether a specific DBMS solution would fit their needs or not. This has lead me to a more generalized research question formulation being How would a solution architect could be able to evaluate if a certain DBMS would best fit the needs of their application both technically and economically.

Through the following step in the iteration, the main objective(s) of such research needs to be formulated by showing e.g., what the outcome of the research would accomplish and come up with an objective-centered solution.

The accomplishment of such method as described in the previous step would be an enabler to those software system architects to decide upon to build their solutions based on a certain DBMS by being able to compare through different database systems facilitating the method's evaluation criteria. Another possible outcome would be in case the architect would need to migrate their data from a certain DBMS to another; they will be able to compare the other solutions possibilities, openings and pros/cons through instantiating the method between several database management systems.

The following step in the DSR would be to design and develop the method as artifact, which is the main outcome of the research study.

The possible artefact in this case would be an evaluation method that is comprised of a set of criteria and grading mechanism that would act as a decision support tool for software solution/system architect, enabling them to compare and decide upon the suitable DBMS among the possible options for the specific application and the specific needs of their organization. In this iteration, the actual method or framework needs to be designed and developed.

The method is basically comprised of a list of evaluation criteria which best describe the state-of-the-art features of today's big data environments

Following to the design and development phase, the artifact should be placed into the target context in order to achieve result. This is the actual use of the artifact in order to solve the problem by which the research has initiated.

CHAPTER FIVE

Chapter 5 – Artifact Development and Evaluation

The designed method will contain attributes as criteria to evaluate a DBMS for spatial big data as well as a weight for each criteria for indicating the relevance of each criteria to the specific application that makes it specific for any business implementation. For instance, if a company is interested in some specific features of a DBMS (e.g. in scale of 1 to 5), the weight of presence or lack of significance in such criteria for that specific application would be different to another business/app setting.

5.1. Selection Criteria Explained:

The method is comprised of a set of features of today's most advanced database management systems, accompanied by a weight factor that defines the level of relevance of the criteria to a specific application.

Following is a list of selection criteria categorized in two categories of technical and commercial, based on the findings from literature search in the field of big data and spatial database systems and Gartner business analysis reports, as discussed in the literature review synthesis, as well as author's own experience from business during the years of working as database specialist and system administration.

Technical Criteria:

1. Spatial Data Types, Indexes and Functions
   As discussed by: (Zhou et al., 2009), (Carpenter & Snell, 2013), (Ikawa, 2019), (Thatcher, 2014), (Li et al., 2016), (Sharma et al, 2014), (Bertino et al., 2012)
2. Polyglot vs. Multi-Model database
   As discussed by: (Lu & Holubova, 2017), (Bertino et al., 2012)
3. Parallel Partitioning
   As discussed in: (Zhao et al., 2016), (Yu et al., 2015), (Carpenter & Snell, 2013), (Sarwat, 2015)
4. Columnar/In-Memory
   As discussed by: (Yu et al., 2015), (Carpenter & Snell, 2013)
5. Graph Database
   As discussed by: (Moniruzzaman & Hossain, 2013)

6. Document Database (JSON, XML)
   As discussed by: Han et al (2011), Chickerur et al (2015)
7. Key-value pair stores
   As discussed by: (Hughes et al., 2019), (Chen & Zhang, 2014)
8. NoSQL support
   As discussed in: (Carpenter & Snell, 2013), (Baralis and Rossi, 2017), (Moniruzzaman & Hossain, 2013), (Chen & Zhang, 2014), (Juang et al, 2015), (Makris et al., 2019)
9. Cloud Support
   As discussed by: (Zhao et al., 2016), (Yu et al., 2015), (Carpenter & Snell, 2013), (Baralis and Rossi, 2017), (Sarwat, 2015)
10. Integrated Querying
    As discussed by: (Carpenter & Snell, 2013)

Commercial criteria:

1. Licensing (Carpenter & Snell, 2013)
2. Maintenance and Skill Requirements (Carpenter & Snell, 2013)

5.2. Method Evaluation Strategy:

For the evaluation phase of the DSR process iteration, I have decided to perform the evaluation phase using strategy of "Quick and Simple" that is conducted in form of naturalistic evaluation as suggested by Venable et al. (2016)

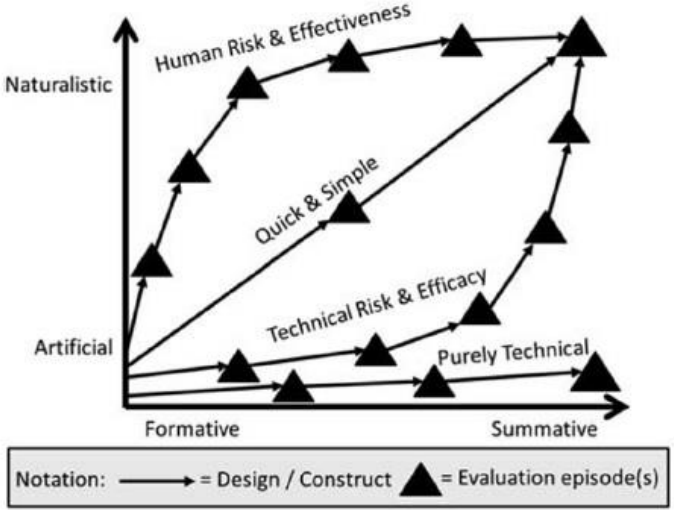

Figure 5.1. *FEDS Framework to Evaluate Design Science with evaluation strategies* (Venable et al., 2016)

The evaluation process as such consists of four steps as follows:

Step1-explicating the goals:

In the effort to design a method for evaluation of DBMS features, I have tried to design a method by which organizations' architects could be able to decide upon the best fit for their database management system that suits their specific needs for big data of spatial type.

As of introducing the new demands for data types and data process means and methods, the number of database solutions are rising with a high rate nowadays that makes the process of capability evaluation and the level of relevance of each product to a specific application complicated.

The developed method would serve as a decision support tool, hence facilitating the decision-making authorities in choosing the appropriate solution to their specific needs.

Step2-choosing the strategy for evaluation:

The next step into the evaluation phase of DSR is to choose a strategy for evaluation.

There are four heuristics to be address before we can decide upon the suitable strategy for evaluation. As the last point in the four-stage heuristic questions suggests, the level of complexity of the artefact should be examined if it is large and complex or small and simple.

As the artifact contains only certain evaluation criteria as well as a weight factor due to the state-of-the-art DBMS capabilities to handle big data, the structure of the artifact here would be considered simple and small and therefore one should go directly to "Quick and Simple" strategy.

Step3-determining the properties to evaluate:

This step includes deciding upon what to evaluate. Since the study's concentration is on designing a method by which the process of decision making for the relevant database system for specific application needs could become an easier and more facilitated process, the main property to evaluate for in method would be how the domain experts could facilitate and benefit from the method to make decision. Nevertheless, the domain experts could best evaluate the "Quick and Simple" evaluation of the method.

Step4-designing the evaluation episode:

The following steps should be taken in order to design the evaluation episode:

i. Identification and analyzing the environmental constraints. This means that one should figure out about what environmental resources are available and to what extent for the evaluation and which ones are not.
ii. Based on the above analysis, the aspects in evaluation should be prioritized so that we know which ones are essential, nice-to-have or very irrelevant to the evaluation process.
iii. Finally, a plan should consist of the number of evaluation iterations, timing of the evaluation and the actor(s).

5.3. Phase ONE: Ex Ante Concept Evaluation:

Since the method covers both technical aspects of the selection decision support as well as socio-technical parts, multiple sessions of naturalistic and artificial evaluation has been performed in order to guarantee the method's utility and ease of use from socio-technical point of view as well as artificial evaluations to ensure its technical usefulness.

The two rounds of Ex Ante evaluations at LTU in context of one-to-one feedback and brainstorming.

As Venable et al. (2012) categorizes the evaluation of DSR into two genres of Ex-Ante and Ex-Post as well as Naturalistic and Artificial, the evaluation has been performed in both forms of Ex-Ante and Ex-post. The evaluation rounds would hold naturalistic as well as artificial grounds. After the two rounds of feedback in the evaluation prior to the design (Ex Ante/Naturalistic), the researcher had a chance to develop the selection criteria into a method in which the selection criteria would form up a method that contains the selected criteria alongside grading scales which enables the user to emphasize and reflect upon the level of relevance for each criteria.

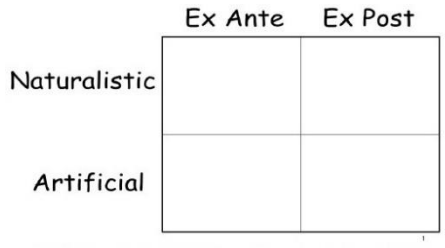

Figure5.2. *Strategic DSR Evaluation Framework (Venable et al., 2012)*

5.3.1. Evaluation - Round One:

First round of evaluation is performed as concept evaluation in forms of formative naturalistic brainstorming. The researcher has brought up the idea of addressing the knowledge gap by providing a decision support tool as a method that enables the system architects to decide upon their specific need of big data with focus on spatial data types and functions.

The method is targeted to support the decision of choosing among the available solutions in the market for database management systems. The initial idea has derived from business need into a literature search and finding the gap in research justifying the need for such method.

The researcher has introduced a list of selection criteria as result of the literature search as well as findings from business reports in the field of databases, powered by the researcher's own expectations and evaluation based on years of experience as database administrator and analyst in the industry.

The evaluation took place as a one-to-one brainstorming by which the evaluator, PhD Candidate, has approved the idea of listing the selection criteria, based on the state-of-the-art database solutions for big data and adding attributes that the user of the method is available upon their need for that specific attribute or feature.

The result of such ex ante evaluation is demonstrated as below (*figure 5.2*).

Item No.	Selection Criteria	Level of Relevance/Interest
1	Spatial Data Types	0-5
2	Advanced Spatial Functions	0-5
3	Spatial Indexing (R-tree)	0-5
4	NoSQL Support	0-5
5	Key-Value pair stores	0-5
6	Multi-Model Support	0-5
7	Document DB Support (JSON, XML)	0-5
8	Columnar/Wide-row data store	0-5
9	Cloud Support (DBaaS)	0-5
10	Cypher query language support	0-5
11	Integrated querying	0-5
12	Licensing Cost**	0-5
13	Maintenance Skills demand**	0-5

Table 5.1. *Design artifact – Initial Round*

5.3.2. Evaluation - Round Two:

In the Second round of evaluation, the method draft was discussed with supervisor and based on the findings and the initial design from the previous attempt in phase one.

The evaluator then commented as splitting the relevance factor into to attributes, being one as a YES/NO answer to the question if the corresponding feature in the method is applicable or interesting to the user's application or not and then having the weight criteria as of a scale of one to five instead. The results are show in (*figure 5.3).*

Another comment evaluator made was to add a method manual, which I later on created such manual, which could be found in the Appendix of the thesis.

Item No.	Selection Criteria	Relevant?	Level of Relevance/Interest
1	Spatial Data Types	0/1	1-5
2	Advanced Spatial Functions	0/1	1-5
3	Spatial Indexing (R-tree)	0/1	1-5
4	NoSQL Support	0/1	1-5
5	Key-Value pair stores	0/1	1-5
6	Multi-Model Support	0/1	1-5
7	Document DB Support (JSON, XML)	0/1	1-5
8	Columnar/Wide-row data store	0/1	1-5
9	Cloud Support (DBaaS)	0/1	1-5
10	Cypher query language support	0/1	1-5
11	Integrated querying	0/1	1-5
12	Licensing Cost**	0/1	1-5
13	Maintenance Skills demand**	0/1	1-5

Table 5.2. *Design artifact – Revision Two*

After the second round of feedback from Professor, a number of changes were in place with the design. The researcher decided to add two more column to the method in order to enable the user to reflect upon the available features in the target DBMS.

The user of the method could reflect upon each feature by checking if the feature is present in the target system or not and secondly by reflecting upon the level of correlation of each feature to the application or their specific need. In other words, the user of the method should study the target DBMS to see if the feature exists or not and accordingly reflect upon that in the third column, next comes the level they think that the feature is advanced and that it could serve their need.

The result of adding such columns to the method will bring more subjectivity to the method that is shown in (*figure 5.4).*

Item No.	Selection Criteria	Relevant?	Level of Relevance/Interest	Availability*	Feature Relevance *	Feature Band Score
1	Spatial Data Types	1	4	1	3	12
2	Advanced Spatial Functions	0	0	1	2	0
3	Spatial Indexing (R-tree)	1	2	1	5	10
4	NoSQL Support	1	5	1	4	20
5	Key-Value pair stores	1	4	1	3	12
6	Multi-Model Support	1	3	0	1	0
7	Document DB Support (JSON, XML)	1	5	1	3	15
8	Columnar/Wide-row data store	1	2	0	0	0
9	Cloud Support (DBaaS)	1	1	1	4	4
10	Cypher query language support	0	0	1	5	0
11	Integrated querying	0	0	0	0	0
12	Licensing Cost**	1	5	1	2	10
13	Maintenance Skills demand**	1	4	1	3	12
					Total Score	51

Table 5.3. *Design artifact – Revision Three*

The very last element of each row is the column called "Feature Band Score" that shows the product of the previous four values in the previous columns and multiplies them to calculate the-so-called Band Score which indicated the level of relevance and interest in that specific feature for the specific need of the method user.

There are two categories of selection criteria presented in the method. The first eleven criteria are added values, while the last two are costs. This will impose different values to calculation of the selection method "Total Score", meaning that the first eleven factors represent positive values and the latter two represent negative value in the total sum.

5.4. Phase TWO: Ex Post Evaluations:

In this chapter, two rounds of evaluation took place, which have lead the method towards its solid final state and has developed its capabilities in both usability and technical aspects. The process is described as following two rounds of evaluation.

5.4.1. Evaluation Round Three:

The third evaluation took place as an online meeting between the researcher and one person from industry in which the presenter described an overview of the research an demonstrated the usability of the method in order to motivate the participant to get engaged sharing his thoughts in technical efficacy perspective thoughts and feedback.

The participant is a senior database administrator and solution architect from Iran Insurance Company with more than seventeen years of academy and industry experience in field of software engineering and database administration being involved and playing crucial role in a considerable number of the corporate enterprise solutions.

After the session, the evaluator stated his comments as follows:

The biggest missing part from technical point of view is that the method fails to consider compatibility issues involved in the process of choosing a proper DBMS for any software solution.

Reflecting upon such statement and discussion, three more attributes have been added to the method in order to broaden its technical capabilities. The first criteria is called level of compatibility to the programming language of the application and the second criteria is addressing the level of compatibility of the target DBMS with the legacy system of the corporate. Ultimately, the participant addressed cypher query language support as a critical feature due to his field experience and he believed the above-mentioned are three rather crucial criteria for any software architect once making decision upon the right choice for database management systems.

Figure 5.5 shows the addition of the above-mentioned two selection criteria to the method and adding their total band score to the formula of "Total Score" accordingly.

Item No.	Selection Criteria	Relevant?	Level of Relevance/Interest	Availability*	Feature Relevance *	Feature Band Score
1	Spatial Data Types	1	4	1	3	12
2	Advanced Spatial Functions	0	0	1	2	0
3	Spatial Indexing (R-tree)	1	2	1	5	10
4	NoSQL Support	1	5	1	4	20
5	Key-Value pair stores	1	4	1	3	12
6	Multi-Model Support	1	3	0	1	0
7	Document DB Support (JSON, XML)	1	5	1	3	15
8	Columnar/Wide-row data store	1	2	0	0	0
9	Cloud Support (DBaaS)	1	1	1	4	4
10	Cypher query language support	0	0	1	5	0
11	Integrated querying	0	0	0	0	0
12	Compatibility with programming language	1	5	1	4	20
13	Compatibility with organisation's legacy systems	1	4	0	0	0
14	Licensing Cost**	1	5	1	2	10
15	Maintenance Skills demand**	1	4	1	3	12
					Total Score	71

Table 5.4. *Design artifact – Revision Four*

5.4.2. Evaluation Round Four: Focus group discussion as summative ex post evaluation

In this round of evaluation, the aim is to encourage users to engage with their thoughts and attitude towards their own expectations of functionality and usability of the design artifact. The attempt is combined with the target to recognize any unexplored needs users might have or needs that should be adjusted in some way. Summative evaluation should be reiterated until the concepts answer to relevant user needs substantially and no new insights about users' needs are recognized. The target of the summative evaluation is to identify how the selection support tool as a method should be connected and refined to satisfy the needs that have been identified from business in initial inquiries from Mobilaris. Dealing with innovative methods, it is essential to have in mind that it can often take years for the artifact to have its actual impact. A key concept here is to learn from previous failures to ensure that similar mistake will not happen again. During the process, it is vital to keep the key values in mind and to study how value can be created for the end users of the product and how the users can influence the evaluation process.

Following issues have been discussed among participants of the evaluation:

- What is the approach and purpose for the evaluation?
- What results can be expected?
- What is the main question that needs to be answered?
- How are the identified needs and/or requirements reflected in the concept?

A formative evaluation approach has been employed to evaluate the method as the artifact, trying to understand how well and easy the artifact works and what factors are affecting usability of the method. This approach adds worth to the method's life cycle by considering the qualitative method, with exploration of how well the method works and what elements from experience are associated with it to address those issues in the next round of design revision of the tool.

Summative Evaluation Design:

One focus group interview was conducted in English at May 22nd, 2019. The focus group was formed in Luleå, LTU's main campus with participants who were researchers or PhD candidates in Information Systems division. In total three people participated in the focus group along with me performing as the moderators of the discussion.

To recruit respondents, I emailed potential participation at LTU, inviting them to the evaluation meeting. I deemed not to limit the evaluators of the method solely to industry field-expert participants. In order to provide diversity in categories of people contributing to this study, I made a balance between the participants by engaging the researchers not solely related to industry. With respect to privacy concerns, the participants have been informed about that they would be anonymized in case they need to be quoted. The duration of a session was 60 minutes and took place in one afternoon at LTU.

The focus group interview was organized as follows:

1. Short introduction of the researcher, practicalities of the research, obtaining agreement for recording the session and the general plan of the meeting
2. A space for the contributors to introduce themselves and to explain their motivations behind the participation
3. Introduction to the project and types of feedback I expected from the evaluation round
4. General discussion about the research topic and challenges of picking the proper solution as database management systems for corporations
5. Presenting an overview of the previous rounds of evaluation and the feedback from each previous session
6. Description of how the previous evaluations have evolved the design of the artifact
7. Closing discussion dedicated to reflections and discussions upon the possible values the artifact could bring to the industry as a decision support tool for solution architects

Reflections/Feedback:

Participant A discussed that regarding the two evaluation factors of "Level of Relevance" and "Feature Relevance", the need for a textual description associated in each level ranging (from one to five) seems inevitable and adding a textual explanation for each number as in the range of answer could help clarifying the grading mechanism. In response to their feedback, I have added a textual description to the manual corresponding to each number in the evaluation process.

Participant A also believed that adding another selective attribute to the artifact would add up a level of subjectivity, by which the user of the method could add his/her own selection criteria to the method. In response to the above-mentioned comment, another row was added to the method in which the selection criteria is user-defined and can be added to the method at any point of time.

Participant B proposed to change the title for those selection criteria, which impose negative value to the calculation of total amount e.g. "licensing cost" to "cost efficiency".

According to the fact that the term "cost efficiency" could be related to other costs as well and due to the fact that the term would not reflect upon other selection criteria such as maintenance costs by skilled workforce. The term would be the underestimating a very significant range of selection criteria, so no actions taken due to the comment.

Participant C reflected upon the understandability of each selection criteria. They believe that adding textual description to each criteria would help the user to have a better understanding of the corresponding selection criteria, therefore could reflect upon the evaluation much easier. The description will be added as response to this comment in order to enrich the user experience by facilitating easier understanding of the meaning of each criteria.

Participant C also believes that adding comments to each column header in the method provides better readability and knowledge on how to use the evaluation method and correlate to the range of one to five. According the feedback, a comment added to each column header, which provides the corresponding additional comments to each column that needs user engagement.

Participant C also brings up the discussion of having all the streamlined database management systems available in the market and having the feature of auto-filling the attributes related to each DBMS for ease of use. This feedback is deemed as a very constructive yet time-consuming feature for the artifact. The feature would enable the user to have pre-filled attributes as per each DBMS that obviously adds a high level of usability and by making the evaluation process much easier by facilitating with pre-filled attributes for availability of each feature in the target database system.

The summative evaluation showed that the tool can provide benefits to its users through gathering the state-of-the-art features of database systems with respect to big-data and spatial data features of streamline database technology in one place. The method is also beneficial in raising awareness for decision makers about the factors crucial for decision-making in the fields. The method itself

could also bring ease in correlating the specific needs of an application of a certain use in GIS to the critical features available in the products in the database management systems market.

The ultimate effect of the four-round evaluation is reflected in the artifact's "final Version" (*figure 5.6*) as well as in the method manual (Appendix A)

Item No.	Selection Criteria	Relevant?	Level of Relevance/Interest	Av		ture Band Score
1	Spatial Data Types	1	4			12
2	Advanced Spatial Functions	0	0			0
3	Spatial Indexing (R-tree)	1	2			10
4	NoSQL Support	1	5			20
5	Key-Value pair stores	1	4			12
6	Multi-Model Support	1	3			0
7	Document DB Support (JSON, XML)	1	5	1	3	15
8	Columnar/Wide-row data store	1	2	0	0	0
9	Cloud Support (DBaaS)	1	1	1	4	4
10	Cypher query language support	0	0	1	5	0
11	Integrated querying	0	0	0	0	0
12	Compatibility with programming language	1	5	1	4	20
13	Compatibility with organisation's legacy systems	1	4	0	0	0
14	Licensing Cost**	1	5	1	2	10
15	Maintenance Skills demand**	1	4	1	3	12
16	User-specific criteria	1	4	1	3	12
					Total Score	83

Saleh Kanani:
To what extent is this feature important or relevant for your organization/application?
Answer in Scale of 1 to 5
where 1 represents least relevance and 5 represents highest relevance/significance to you

* In Target DBMS

** Licensing and Maintenance costs will impose negative value in the calculation of Total Score

Table 5.5. *Design artifact – Last Revision*

As Shown in the last revision of method (figure 5.6), revisions have been made on the method as follows:

1. User-specific selection criteria is added to the list of selection criteria
2. Comments added into the column headers for more readability
3. Two footers added as textual descriptions emphasizing on the last two columns being the factors which needs considering the grading with respect to the target DBMS
4. Last footer designates the fact that the items '14' and '15' impose negative value to the "Total Score"

Other reflections discussed above would affect the manual in (Appendix A). According to feedback from participants, a textual description is added to correlate each number from1 to 5 to textual descriptions and it contains textual explanations describing each selection criteria as reflection to the feedback.

CHAPTER SIX

Chapter 6 – Discussion and Future Research

6.1. Discussion

Motivated by business, bringing up the need for having a decision support mechanism that is capable of providing the state-of-the-art features of big data handling methods provided in today's database solutions with special emphasis on spatial nature of data, designing of such method initiated. Providing user engagement and subjectivity of the user of the method by adding elements that need user awareness and engagement in the decision support mechanism got in agenda.

As shown in literature search, the research papers that have taken comparative perspective of database management mechanisms and features have mostly comparing either a number of products with each other with respect to a specific feature in big data database management of by trying to implement a comparison between different methodologies in database systems for big data of geospatial type.

As stated in the synthesis, previous studies are either comparing specific features with each other through certain database systems or comparing specific database features together that are very limited in numbers, e.g. As Lu & Holubova (2017) have discussed the importance of Multi-model databases in big data application.

Another example of single criteria evaluation is Hughes et al (2019) work on key-value pair applicability in spatial large volume data, which again lacks the systematic centralized method for decision support by providing a holistic view of the available features.

Muniruzzaman & Hossain (2013) and Chen & Zhang (2014) have also emphasized on NoSQL applicability in big data, which also addresses a single criteria in evaluating functionality of a certain technology in big data application that also lacks the collectiveness of a systematic selection criteria for underlying database system.

Ikawa et al (2019) has also emphasized the significance of R-Tree indexing feature of spatial database management systems and how well they could improve the performance of GIS systems and pros and cons contributed to their usage, that has become another selection criteria in the systematic approach to database evaluation that this research is targeting.

The contribution of this research is to bring a systematic evaluation method for databases that are ought to be handling big data of spatial nature that was missing in the research literature.

Instead of comparing specific database system, the research aimed to develop a method through which any database management system of any kind could be evaluated subjectively as response of the knowledge gap. The method filled the gap by providing a comprehensive framework that contains significant criteria in the fields of big data of spatial nature to empower the system architects in corporates to make a comparison based on their own perception of big data features' usability in their own specific applications. The method can also as be used by researchers to provide with the state-of-the-art database features for big data spatial functions.

Centralizing all the features found in the literature as significant factors in such selection, combining them alongside the feedback through evaluation rounds both from academicians and field experts, have enriched the method by expanding the audience and evaluators.

6.2. Limitations and Future Research

As the evaluation had benefit from field experts' feedback in a rather limited evaluation phase, a limitation to the method is the low impact from practice on the method. The possibility to expand and enrich the method by getting further rounds of different practitioners in the industry would form a rather preferable future research possibility.

The results of such further evaluations would expand the value and functionality of the method.

Other limitation to this piece of work is that due to the limited amount of resources, the researcher did not have a chance to consider demonstrating the method through every possible streamline product available in the database market. Based on the above-mentioned limitation, another possible future research possibility would be to consider studying the streamline and pioneer products in the market share, and adding pre-filled values in the relevant attributes in the method once selecting from a list of pre-defined products.

Another possibility for further research would be to generalize the method so that it could be used for applications other that geo-spatial big data.

References:

Aitchison, A. (2009a). *Beginning spatial with SQL server 2008* Apress.

Ashworth, M. (1999). Information technology, database languages, SQL multimedia, and application packages, part 3: Spatial. *Iso/Iec,* , 13249-13243.

Baker, J., Bond, C., Corbett, J. C., Furman, J. J., Khorlin, A., Larson, J., . . . Yushprakh, V. (2011). Megastore: Providing scalable, highly available storage for interactive services.

Baralis, E., Dalla Valle, A., Garza, P., Rossi, C., & Scullino, F. (2017). SQL versus NoSQL databases for geospatial applications. Paper presented at the *2017 IEEE International Conference on Big Data (Big Data)*, 3388-3397.

Belanger, F., Hiller, J. S., & Smith, W. J. (2002). Trustworthiness in electronic commerce: The role of privacy, security, and site attributes. *The Journal of Strategic Information Systems, 11*(3-4), 245-270.

Bertino, E., Ooi, B. C., Sacks-Davis, R., Tan, K., Zobel, J., Shidlovsky, B., . . . Andronico, D. (2012). *Indexing techniques for advanced database systems* Springer Science & Business Media.

Birkin, M. (2019). Spatial data analytics of mobility with consumer data. *Journal of Transport Geography, 76*, 245-253.

Carpenter, J., & Snell, J. (2013). *Future trends in geospatial information management: The five to ten year vision* Ordnance Survey at the request of the Secretariat.

Cecchet, E., Candea, G., & Ailamaki, A. (2008). Middleware-based database replication: The gaps between theory and practice. Paper presented at the *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 739-752.

Çetintemel, U., Zimmermann, J., Ulusoy, Ö, & Buchmann, A. (1999). OBJECTIVE: A benchmark for object-oriented active database systems. *Journal of Systems and Software, 45*(1), 31-43.

Chen, C. P., & Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data.*Information Sciences, 275*, 314-347.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* Sage publications.

Davis, J. R. (1998a). IBM's DB2 spatial extender: Managing geo-spatial information within the DBMS. *IBM Corporation, May*,

Davis, J. R. (1998b). IBM's DB2 spatial extender: Managing geo-spatial information within the DBMS. *IBM Corporation, May*,

Eldawy, A., & Mokbel, M. F. (2015). The era of big spatial data. Paper presented at the *2015 31st IEEE International Conference on Data Engineering Workshops*, 42-49.

Evans, M. R., Oliver, D., Yang, K., Zhou, X., Ali, R. Y., & Shekhar, S. (2019). Enabling spatial big data via CyberGIS: Challenges and opportunities. *CyberGIS for geospatial discovery and innovation* (pp. 143-170) Springer.

Feinberg, D., Adrian, M., Heudecker, N., Ronthal, A. M., & Palanca, T. (2015). Magic quadrant for operational database management systems. *Gartner, ID G, 271405*

Franke, U., Johnson, P., König, J., & von Würtemberg, L. M. (2012). Availability of enterprise IT systems: An expert-based bayesian framework. *Software Quality Journal, 20*(2), 369-394.

Gray, J. (1992). *Benchmark handbook: For database and transaction processing systems* Morgan Kaufmann Publishers Inc.

Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. Paper presented at the *2011 6th International Conference on Pervasive Computing and Applications,* 363-366.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems, 19*(2), 4.

Hughes, J. N., Annex, A., Eichelberger, C. N., Fox, A., Hulbert, A., & Ronquest, M. (2015). Geomesa: A distributed architecture for spatio-temporal fusion. Paper presented at the *Geospatial Informatics, Fusion, and Motion Video Analytics V, , 947*394730F.

Ihm, J., Lopez, X., & Ravada, S. (2010). Advanced spatial data management for enterprise applications. *Oracle White Paper,*

Ikawa, G., Watanabe, Y., Yamada, S., & Takada, H. (2019). Performance evaluation of querying point clouds in RDBMS. Paper presented at the *2019 IEEE International Conference on Big Data and Smart Computing (BigComp),* 1-4.

Ji, C., Dong, T., Li, Y., Shen, Y., Li, K., Qiu, W., . . . Guo, M. (2012). Inverted grid-based knn query processing with mapreduce. Paper presented at the *2012 Seventh chinaGrid Annual Conference,* 25-32.

Jung, M., Youn, S., Bae, J., & Choi, Y. (2015a). A study on data input and output performance comparison of MongoDB and PostgreSQL in the big data environment. Paper presented at the *2015 8th International Conference on Database Theory and Application (DTA),* 14-17.

Jung, M., Youn, S., Bae, J., & Choi, Y. (2015b). A study on data input and output performance comparison of MongoDB and PostgreSQL in the big data environment. Paper presented at the *2015 8th International Conference on Database Theory and Application (DTA),* 14-17.

Jung, M., Youn, S., Bae, J., & Choi, Y. (2015c). A study on data input and output performance comparison of MongoDB and PostgreSQL in the big data environment. Paper presented at the *2015 8th International Conference on Database Theory and Application (DTA),* 14-17.

Keenan, P. B., & Jankowski, P. (2019). Spatial decision support systems: Three decades on. *Decision Support Systems, 116,* 64-76.

Lee, J., & Kang, M. (2015). Geospatial big data: Challenges and opportunities. *Big Data Research, 2*(2), 74-81.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., . . . Stein, A. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 119-133.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., . . . Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 119-133. doi:10.1016/j.isprsjprs.2015.10.012

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 134-142.

Lu, J., & Holubová, I. (2017). Multi-model data management: What's new and what's next? Paper presented at the *Edbt*, 602-605.

Makris, A., Tserpes, K., Spiliopoulos, G., & Anagnostopoulos, D. (2019a). Performance evaluation of MongoDB and PostgreSQL for spatio-temporal data.

Makris, A., Tserpes, K., Spiliopoulos, G., & Anagnostopoulos, D. (2019b). Performance evaluation of MongoDB and PostgreSQL for spatio-temporal data.

Manogaran, G., & Lopez, D. (2018). Spatial cumulative sum algorithm with big data analytics for climate change detection.*Computers & Electrical Engineering, 65*, 207-221.

Moniruzzaman, A., & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv Preprint arXiv:1307.0191*,

Ortega, M. I., Genero, M., & Piattini, M. (2017a). Big data DBMS assessment: A systematic mapping study. Paper presented at the *International Conference on Model and Data Engineering*, 96-110.

Ortega, M. I., Genero, M., & Piattini, M. (2017b). Big data DBMS assessment: A systematic mapping study. Paper presented at the *International Conference on Model and Data Engineering*, 96-110.

Osman, R., Awan, I., & Woodward, M. E. (2011). QuePED: Revisiting queueing networks for the performance evaluation of database designs. *Simulation Modelling Practice and Theory, 19*(1), 251-270.

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). Design science research evaluation. Paper presented at the *International Conference on Design Science Research in Information Systems*, 398-410.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

PostgreSQL 9.2 official online documentation. (2019). Chapter 25: High availability, load balancing and replication. Retrieved from http://www.postgresql.org/docs/9.2/static/different-replication-solutions.html

Ray, S., Simion, B., & Brown, A. D. (2011a). Jackpine: A benchmark to evaluate spatial database performance. Paper presented at the *2011 IEEE 27th International Conference on Data Engineering,* 1139-1150.

Ray, S., Simion, B., & Brown, A. D. (2011b). Jackpine: A benchmark to evaluate spatial database performance. Paper presented at the *2011 IEEE 27th International Conference on Data Engineering,* 1139-1150.

Sarwat, M. (2015). Interactive and scalable exploration of big spatial data--A data management perspective. Paper presented at the *2015 16th IEEE International Conference on Mobile Data Management, , 1* 263-270.

Schwartz, B., Zaitsev, P., & Tkachenko, V. (2012). *High performance MySQL: Optimization, backups, and replication* " O'Reilly Media, Inc.".

Scott, D. (2001a). Nsm: Often the weakest link in business availability. *Gartner Group AV-13-9472,*

Scott, D. (2001b). Nsm: Often the weakest link in business availability. *Gartner Group AV-13-9472,*

Sharma, S., Tim, U. S., Wong, J., Gadia, S., & Sharma, S. (2014). A brief review on leading big data models. *Data Science Journal, ,* 14-041.

Stolze, K. (2003). SQL/MM spatial-the standard to manage spatial data in a relational database system. Paper presented at the *Btw, , 2003* 247-264.

Thatcher, J. (2014). Big data, big questions| living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication, 8,* 19.

Trubins, R. (2013). Land-use change in southern sweden: Before and after decoupling. *Land use Policy, 33,* 161-169.

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. Paper presented at the *International Conference on Design Science Research in Information Systems,* 423-438.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77-89.

Vijayakumar, M., Karthick, S., & Prakash, N. (2013a). The day-to-day crime forecasting analysis of using spatial-temporal clustering simulation. *International Journal of Scientific & Engineering Research, 4*(1), 1-6.

Vijayakumar, M., Karthick, S., & Prakash, N. (2013b). The day-to-day crime forecasting analysis of using spatial-temporal clustering simulation. *International Journal of Scientific & Engineering Research, 4*(1), 1-6.

Vijayakumar, M., Karthick, S., & Prakash, N. (2013c). The day-to-day crime forecasting analysis of using spatial-temporal clustering simulation. *International Journal of Scientific & Engineering Research, 4*(1), 1-6.

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. Paper presented at the *Ecis, , 9* 2206-2217.

Whitman, M. E., & Mattord, H. J. (2011a). *Principles of information security* Cengage Learning.

Whitman, M. E., & Mattord, H. J. (2011b). *Principles of information security* Cengage Learning.

Yu, J., Wu, J., & Sarwat, M. (2015). Geospark: A cluster computing framework for processing large-scale spatial data. Paper presented at the *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems,* 70.

Zhao, L., Chen, L., Ranjan, R., Choo, K. R., & He, J. (2016a). Geographical information system parallelization for atial big data processing: A review. *Cluster Computing, 19*(1), 139-152.

Zhao, L., Chen, L., Ranjan, R., Choo, K. R., & He, J. (2016b). Geographical information system parallelization for spatial big data processing: A review. *Cluster Computing, 19*(1), 139-152.

Zhou, Z., Zhou, B., Li, W., Griglak, B., Caiseda, C., & Huang, Q. (2009a). Evaluating query performance on object-relational spatial databases. Paper presented at the *2009 2nd IEEE International Conference on Computer Science and Information Technology*, 489-492.

Zhou, Z., Zhou, B., Li, W., Griglak, B., Caiseda, C., & Huang, Q. (2009b). Evaluating query performance on object-relational spatial databases. Paper presented at the *2009 2nd IEEE International Conference on Computer Science and Information Technology,* 489-492.

Appendix A: Method Manual

Following is a stepwise user guide that users should consider following, in order to achieve best results from the method:

1- The user of the method should supposedly be aware of the following technologies, as they are the selection criteria used by the method to evaluate each database management system.
   - Spatial data types e.g. Geography and Geometry, Lines, Polygons, etc.
   - Spatial Functions e.g. ST_AREA, ST_COVERS, etc. from PostGIS
   - Spatial indexing i.e. specific type of indexing for spatial big data e.g. R-Tree
   - NoSQL (Not-Only SQL) language support
   - Key-Value pair stores as a NoSQL data store technology
   - Database multi-model support i.e. if the DBMS supports NoSQL, Document (XML, JSON) or Object-relational data types alongside with relational data
   - Columnar data store technology, which is very critical feature dealing with large volume of data
   - If the DBMS provides cloud support as a service or the feature as being served though private on-premises cloud infrastructure
   - Consider the cypher query language support that is vital feature for graph databases
   - Awareness of licensing and skill requirements for maintenance and administration
   - Level of compatibility of the DBMS with programming language and legacy systems in use by the organization

2- The first step in using the method is to study and understand all the features in the above list.

3- After studying the above features, the user needs to fill in the column "Relevant" with values "0" as "NO" or "1" as "YES" if they feel the each specific feature is relevant to their own specific use or not.

4- In the next column "Level of Relevance / Interest", the user should reflect upon the degree to which that certain feature is relevant or interesting to their own specific application.
   - Reflection upon the level of relevance / interest should be in scale of 1 to 5 where value 1 reflects the least relevance and 5 represents the highest level of relevance and/or interest.

5- After filling the "Level of Relevance / Interest", the user should fill in the next two columns by each, they reflect upon the availability and relevance of the technology that is facilitated in the target DBMS.

- The column "Available" should be filled in with values "0" as "NO" or "1" as "YES" if the feature id provided by the target DBMS or not, accordingly.
- The column "Feature Relevance" should be filled with values in scale of 1 to 5. Value "1" indicates that the feature is available but very minimalistic or merely relevant to their needs and "5" would indicate that the feature is very well developed and available in the target DBMS and it would serve the needs of the application the most convenient manner.

6- After filling the values for all four above-explained attributes for each selection criteria, a "Feature Band Score" would be generated as result of multiplication of the given values.

7- Finally, the "Total Score" will be generated as the sum of all the above "Band Scores".

Note 1: Please note that Items No.1 to 14 will be added to generate the "Total Score" and items No.15 and 16 will be subtracted, since they impose cost.

Note 2: Item No.14 is added for extended user engagement and flexibility and could be filled with any business/user-specific selection criteria that user feels necessary yet are not listed.