



Helwan University  
Faculty of Computers and Information  
Information Systems Department

---

---

# "Securing Big Data using Negative Databases"

---

---

Thesis Submitted For Partial Fulfillment of the Requirements for  
the Master Degree in Computers and Information, Information  
Systems Department, Faculty of Computers and Information,  
Helwan University

Submitted by

**Azza Ahmed Mohammed Ahmed**

Under Supervision of

**Prof. Sayed Abdel Gaber**

Information Systems Department,  
Faculty of Computers and Information,  
Helwan University

**Dr. Hanan Fahmy**

Information Systems Department,  
Faculty of Computers and Information  
Helwan University

**Dr. Mohammed Hassan**

Faculty of Computers and Information,  
Fayoum University

**2018**

---

# Acknowledgement

---

I can not express my thanks to ALLAH for every thing ALLAH gives to me beginning from being a Muslim person.

I am grateful to a number of people, which supported me to carry out this work. I would like to thank my thesis co-advisers; Prof. Dr. Sayed Abdel Gaber, Dr. Hanan Fahmy for her guidance, patience, and encouragement, Dr. Mohammed Hassan and Eng. Hussien Shahata they have given me during the work in this thesis.

Finally, a heart- felt “Thanks” to my wonderful Family for their pray, unconditional support and understanding, and for always being there for me, my dearest friend Asmaa for always being there for me.

---

# Abstract

---

Security related issues have many threats to the original data in any organization. There are users who always try to get into the internal information system and the data systems. There are organizations such as credit card companies, government agencies and security agencies that need their data secured to the highest extent. Hence, such organizations want their applications to provide high security. There are numerous ways to keep data secure but there is no way to secure it 100%. Anywhere, because of the different kinds of attacks that could be vulnerable at any time and the consequences that may be worse than expected. Therefore, no one knows the types and the severity of such attacks.

This thesis presents a new framework that addresses the problem of big data security. The proposed framework consists of two main phases, which are: NDG (Negative Data Generator) and NQC (Negative Query Converter). The first phase explains how to generate a Negative Data; this phase consists of big dataset, mapper, mapper result, NDG, and Negative Data. The second phase shows how to get a positive query on a Negative Data and retrieve the positive result of the query as the Positive Data; this phase consists of authority model and NQC.

The experiments that were conducted on the proposed framework are shown that the generator is capable of generating automatic data named ND from the original data (big data) in a low time and high processing and the malicious users can't get any useful information from ND. The increasing in the volume of big data leads to a decreasing in the volume of ND. Therefore, applying the concept of negative databases to the distributed

environment rather than just use it in the traditional environment to overcome the shortcomings of the traditional environment. This thesis consider the first research that presents a NQC that helps in dealing with negative data such as positive data, the experiments were provided the correctness of the NQC , and the importance of NQC that helps the user to extract data efficiently and easily from negative database.

---

# List of Abbreviations

---

<b>3Vs</b>	Volume, ,Velocity, Variety
<b>4Vs</b>	Volume, ,Velocity, Variety, Value
<b>ABAC</b>	Attribute Based Access Control
<b>ACLs</b>	Access Control Lists
<b>AM</b>	Authority Model
<b>ANN</b>	Artificial Neural Network
<b>CNF-SAT</b>	Conjunctive Normal Form- Satisfiability Problem
<b>CRM</b>	Customer Relationship Management
<b>DB</b>	Data Base
<b>DPLL</b>	Davis-Putnam-Logemann-Loveland
<b>ERM</b>	Environmental Resources Management
<b>ETL</b>	Extract, Transform, and Load
<b>FHE</b>	Fully Homomorphic Encryption
<b>GFS</b>	Google File System
<b>HDFS</b>	Hadoop Distributed File System

<b>HE</b>	Homomorphic Encryption
<b>IDC</b>	International Data Corporation
<b>IDS</b>	Intrusion Detection Systems
<b>IoT</b>	Internet of things
<b>IPS</b>	intrusion protection systems
<b>MPC</b>	Multi-Party Computation
<b>MR</b>	MapReduce
<b>ND</b>	Negative Data
<b>NDBs</b>	Negative Databases
<b>NDG</b>	Negative Data Generator
<b>NoSQL</b>	Not Only Structure Query Language
<b>NQC</b>	Negative Query Converter
<b>OLAP</b>	Online Analytical Processing
<b>PD</b>	Positive Data
<b>RDBMS</b>	Relational Database Management Systems
<b>RNDB</b>	Randomize Negative Database
<b>SAT</b>	Satisfiability Problem
<b>SNA</b>	Social Network Analysis
<b>SQL</b>	Structure Query Language

<b>SRP</b>	Secure Remote Password Protocol
<b>U</b>	Universal
<b>USB</b>	Universal Serial Bus
<b>VC</b>	Verifiable Computation

---

# Table of Contents

<b>Acknowledgement</b>	<b>II</b>
<b>Abstract</b>	<b>III</b>
<b>List of Abbreviations</b>	<b>V</b>
<b>List of Figures</b>	<b>VII</b>
<b>List of Tables</b>	<b>I</b>
<b>List of Tables</b>	<b>VIII</b>
<b>CHAPTER 1: Introduction</b>	
<b>1.1 Introduction</b>	<b>1</b>
<b>1.2 Security Definition</b>	<b>1</b>
<b>1.2.1 Security Foundation</b>	<b>2</b>
<b>1.2.1.1 Authentication</b>	<b>2</b>
<b>1.2.1.2 Authorization</b>	<b>3</b>
<b>1.2.1.3 Non-repudiation</b>	<b>3</b>
<b>1.2.1.4 Confidentiality</b>	<b>3</b>
<b>1.2.1.5 Integrity</b>	<b>4</b>
<b>1.2.1.6 Availability</b>	<b>4</b>
<b>1.2.1.7 Auditing</b>	<b>4</b>
<b>1.3 Big Data Overview</b>	<b>4</b>
<b>1.4 Negative Databases (NDBs)</b>	<b>5</b>
<b>1.5 Thesis Objectives</b>	<b>5</b>
<b>1.6 Thesis Outlines</b>	<b>6</b>
<b>CHAPTER 2: Preliminaries and Literature Survey</b>	
<b>2.1 Introduction</b>	<b>7</b>



<b>2.2 Big Data Definition</b>	<b>8</b>
<b>2.3 The Characteristics of Big Data</b>	<b>9</b>
<b>2.4 Big Data Techniques</b>	<b>10</b>
<b>2.5 The Issues of Big Data</b>	<b>12</b>
<b>2.6 Big Data Security Aspects</b>	<b>14</b>
<b>2.7 Big Data Security Technologies</b>	<b>16</b>
<b>2.8 Hadoop Architecture</b>	<b>19</b>
<b>2.8.1 MapReduce</b>	<b>21</b>
<b>2.8.1.1 MapReduce Advantages</b>	<b>22</b>
<b>2.8.1.2 MapReduce Disadvantages</b>	<b>23</b>
<b>2.9 Relational Algebra with MapReduce</b>	<b>24</b>
<b>2.10 Negative Databases Representation</b>	<b>25</b>
<b>2.11 Techniques for Generating Negative Databases</b>	<b>28</b>
<b>2.12 NDB Database-Related Logical Operations</b>	<b>42</b>
<b>2.13 Negative Databases Applications</b>	<b>44</b>
<b>CHAPTER 3: The Proposed Framework</b>	
<b>3.1 Introduction</b>	<b>47</b>
<b>3.2 The Proposed Framework</b>	<b>48</b>
<b>3.2.1 Negative Data Generation</b>	<b>50</b>
<b>3.2.1.1 NDG Time Calculations</b>	<b>53</b>
<b>3.2.2 Negative Data Query</b>	<b>54</b>
<b>3.2.2.1 The Negative Query Converter Algorithm Steps</b>	<b>56</b>
<b>CHAPTER 4: Testing &amp; Implementation</b>	
<b>4.1 Introduction</b>	<b>59</b>
<b>4.2 The Proposed Framework Prerequisites</b>	<b>59</b>
<b>4.3 The Proposed Framework Flowchart</b>	<b>60</b>
<b>4.2 NDB Generation Algorithm Implementation</b>	<b>63</b>
<b>4.2.1 Testing of Negative Databases</b>	<b>73</b>

<b>4.3 Implementation to Prove the Correctness of the Query Converter Algorithm</b>	<b>74</b>
<b>4.4 Results</b>	<b>98</b>
<b>CHAPTER 5: Summary, Conclusion &amp; Future Work</b>	
<b>5.1 Introduction</b>	<b>106</b>
<b>5.2 Summary of the Thesis</b>	<b>106</b>
<b>5.3 Conclusion</b>	<b>107</b>
<b>5.4 Future Work</b>	<b>108</b>
<b>REFERENCES</b>	<b>109</b>

---

# List of Figures

---

2.1	<b>Layered Architecture of Big Data System.</b>	<b>9</b>
2.2	<b>Hadoop Framework.</b>	<b>20</b>
2.3	<b>Hadoop Architecture.</b>	<b>20</b>
2.4	<b>MapReduce Architecture.</b>	<b>22</b>
2.5	<b>Negative Database Algorithms.</b>	<b>28</b>
2.6	<b>Prefix Algorithm Implementation Steps.</b>	<b>30</b>
2.7	<b>Randomize Algorithm Implementation Steps.</b>	<b>31</b>
2.8	<b>The <math>Q</math>-hidden-<math>NDBs</math> Algorithm Implementation Steps.</b>	<b>33</b>
2.9	<b>Implementation Steps for Generating the Hybrid-<math>NDBs</math> Algorithm.</b>	<b>34</b>
2.10	<b>GenComplete(s) Function to Generate Complete NDB</b>	<b>35</b>
2.11	<b>Makehardreverse (<math>NDB_C</math>, <math>S</math>) Function to Generate Hard-to-Reverse NDB Same as <math>Q</math>-Hidden Algorithm.</b>	<b>35</b>
2.12	<b>The Real-Valued Negative Databases Phases.</b>	<b>36</b>
2.13	<b>Dividing phase.</b>	<b>37</b>
2.14	<b>Encoding phase.</b>	<b>37</b>
2.15	<b>Generating the binary code from an integer.</b>	<b>38</b>
2.16	<b>Decoding Phase.</b>	<b>39</b>
2.17	<b>Algorithm for generating the <math>P</math>-hidden-<math>NDBs</math>.</b>	<b>40</b>
3.1	<b>The Proposed Framework of Big Data Security.</b>	<b>49</b>
3.2	<b>The Conceptual using Procedures of the proposed Algorithm.</b>	<b>56</b>
4.1 (A)	<b>The Negative Data Generator Algorithm</b>	<b>60</b>

	<b>Flowchart</b>	
<b>4.1 (B)</b>	<b>The Negative Data Generator Algorithm Flowchart</b>	<b>61</b>
<b>4.2</b>	<b>The Negative Query Converter Algorithm Flowchart</b>	<b>62</b>
<b>4.3</b>	<b>The Complement of Student _Name1.</b>	<b>99</b>
<b>4.4</b>	<b>The Compression of Student _Name1 Records before and after Applying NDBs.</b>	<b>99</b>
<b>4.5</b>	<b>The Complement of Student _Name2.</b>	<b>100</b>
<b>4.6</b>	<b>The Compression of Student _Name2 Records before and after Applying NDBs.</b>	<b>100</b>
<b>4.7</b>	<b>The Complement of Student _Name2 and Student_Grade.</b>	<b>101</b>
<b>4.8</b>	<b>The Compression of Student _Name2 and Student_Grade Records before and after Applying NDBs.</b>	<b>101</b>
<b>4.9</b>	<b>Measuring the CPU Time (second) for NDB on Single Node.</b>	<b>102</b>
<b>4.10</b>	<b>Measuring the CPU Time (second) for NDB on Cluster Node.</b>	<b>103</b>
<b>4.11</b>	<b>Measuring the Memory Size (MB) for NDB on Single Node.</b>	<b>103</b>
<b>4.12</b>	<b>Measuring the Memory Size (MB) for NDB on Cluster Node.</b>	<b>104</b>
<b>4.13</b>	<b>Measuring the CPU Time (second) for NDB between Single and Cluster Node</b>	<b>104</b>
<b>4.14</b>	<b>Measuring the Memory Size (MB) for NDB between Single and Cluster Node.</b>	<b>105</b>
<b>4.15</b>	<b>The Result of Negative And Positive Select Representation.</b>	<b>105</b>
<b>4.16</b>	<b>Reversing the Result of Data from Negative to Positive Projection.</b>	<b>106</b>

<b>4.17</b>	<b>The Result of Negative And Positive Intersection Representation.</b>	<b>106</b>
<b>4.18</b>	<b>The Result of Data Negative and Positive Union Representation.</b>	<b>107</b>
<b>4.19</b>	<b>The Result Of Data Negative And Positive Join Representation.</b>	<b>107</b>

---

# List of Tables

---

<b>2.1</b>	<b>Simple Negative Data Representation</b>	<b>26</b>
<b>2.2</b>	<b>Applying DB operation.</b>	<b>43</b>
<b>2.3</b>	<b>Applying DB Operations in Negative Database.</b>	<b>43</b>
<b>3.1</b>	<b>The Negative Data Generator Algorithm Pseudo Code</b>	<b>52</b>
<b>4.1</b>	<b>Positive Database (DB1) of Student_Name1.</b>	<b>63</b>
<b>4.2</b>	<b>Binary Representation of Student_Name1.</b>	<b>63</b>
<b>4.3</b>	<b>Decimal Format of Student_Name1.</b>	<b>63</b>
<b>4.4</b>	<b>The Complement of Student_Name1.</b>	<b>64</b>
<b>4.5</b>	<b>The Negative Database of Student_Name1</b>	<b>64</b>
<b>4.6</b>	<b>Positive Database (DB2) of Student_Name2.</b>	<b>67</b>
<b>4.7</b>	<b>Binary Format of Student_Name2.</b>	<b>67</b>
<b>4.8</b>	<b>Decimal Format of Student_Name2.</b>	<b>68</b>
<b>4.9</b>	<b>The Complement of Student_Name2.</b>	<b>68</b>
<b>4.10</b>	<b>The Negative Database of Student_Name2.</b>	<b>69</b>
<b>4.11</b>	<b>The Complement of Student_Name2 and Student_Grade</b>	<b>71</b>
<b>4.12</b>	<b>The Negative Database of Student_Name2 And Student_Grade</b>	<b>71</b>
<b>4.13</b>	<b>The NDB Output Comparison Table for Running on Single Node</b>	<b>73</b>
<b>4.14</b>	<b>The NDB Output Comparison Table for Running on Cluster Node</b>	<b>74</b>

<b>4.15</b>	<b>The Positive Database of Student _Name1 With 24- Bits Strings. And The Result of Applying Select Operation.</b>	<b>75</b>
<b>4.16</b>	<b>The Result of Negative And Positive Select Representation</b>	<b>75</b>
<b>4.17</b>	<b>The Positive Database of Student _Name1 with 24- Bits Strings. And The Result of Applying Projection Operation</b>	<b>76</b>
<b>4.18</b>	<b>Reversing the Result of Data from Negative to Positive Projection.</b>	<b>77</b>
<b>4.19</b>	<b>The Positive Database of Student _Name1 with 24- Bits Strings. And The Result of Applying Intersection Operation</b>	<b>80</b>
<b>4.20</b>	<b>The Result of Negative And Positive Intersection Representation.</b>	<b>80</b>
<b>4.21</b>	<b>The Positive Database of Student _Name1 with 24- Bits Strings. And The Result of Applying Union Operation.</b>	<b>85</b>
<b>4.22</b>	<b>The Result of Data Negative and Positive Union Representation</b>	<b>86</b>
<b>4.23</b>	<b>The Positive Database of Student _Name1 with 24- Bits Strings. And The Result of Applying Join Operation</b>	<b>91</b>
<b>4.24</b>	<b>The Result Of Data Negative And Positive Join Representation</b>	<b>91</b>