

**IMPROVING THE PERFORMANCE OF REAL-TIME
DATA ANALYTICS APPLICATIONS BY OPTIMISING
THE DATABASE AGGREGATION**

Thenamurage Dona Methma Pabasarie Samaranayake

199359L

MSc in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2022

**IMPROVING THE PERFORMANCE OF REAL-TIME
DATA ANALYTICS APPLICATIONS BY OPTIMISING
THE DATABASE AGGREGATIONS**

Thennamurage Dona Methma Pabasarie Samaranayake

199359L

This thesis was submitted in partial fulfilment of the requirements for the
Degree of MSc in Computer Science Specialising in Software
Architecture

MSc in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka
June 2022

DECLARATION

I declare that this is my own work and this has not incorporated without acknowledgement any material previously submitted for Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other media. I retain the right to use this content in whole or partially in future works.

.....
T. D. M. P. Samaranayake

.....
Date

I certify that the declaration above,

The candidate is true to the best of my knowledge and that this project report is acceptable for evaluation for the final research.

.....
Prof. Indika Perera

.....
Date

ABSTRACT

Organisations must make the best decision at the appropriate time to obtain a competitive advantage in a fast-changing market. To accomplish so, it's critical to make faster and more efficient judgments based on near-real-time data analysis. When it comes to these real-time streaming data analysis systems, the performance of the database is having a huge impact on such applications as it is required to achieve data availability and continuous processing for a large volume of data without a delay. When it comes to streaming data, data warehousing is more challenging. So, it is required to consider performance improvements in all the steps of the Extraction, Transformation, Load (ETL) process and the database architecture level. Therefore, the proposed approach is to improve the performance of the system by optimising the ETL process (Extraction, Transformation, and Load) and real-time data warehousing. In this approach, the optimised aggregation algorithm is introduced. Apart from that, the hardware, storage schemas, and query optimization of the data warehouse are also considered and this study is evaluating the performance of the centralised architecture for the real-time data warehouse.

ACKNOWLEDGEMENT

My profound gratitude goes to Prof. Indika Perera, my supervisor for the knowledge, supervision, advice, and guidance provided with his expertise, throughout in making the thesis a success. My appreciation goes to my family for the motivation and support provided throughout my life. I also would like to thank my colleagues in the MSc batch and at my workplace for their help and support in managing my research work.

THE TABLE OF CONTENT

DECLARATION	3
ABSTRACT	4
ACKNOWLEDGEMENT	5
THE TABLE OF CONTENT	6
LIST OF FIGURES	7
LIST OF APPENDICES	8
CHAPTER 1	
INTRODUCTION	1
1.1 Research Problem	3
1.2 Research Objectives	3
1.3 Scope	4
1.4 Outline	4
CHAPTER	2
LITERATURE REVIEW	5
2.1 ETL Process	6
2.2 Stream Data Warehousing	9
2.3 Summary	11
CHAPTER 3	
METHODOLOGY	13
3.1 ETL Module	14
3.1.2.1 Use Case	18
3.1.2.2 Aggregation Approach	19
3.1.2.3 Data Purging	22
3.2 Real-Time Data Warehouse	23

3.3 Statistics Retrieval	23
3.4 Performance Evaluation	25
3.5 Summary	26
CHAPTER 4	
SYSTEM ARCHITECTURE AND IMPLEMENTATION	27
4.1 Siddhi.io Stream Processing Architecture	28
4.2 Aggregation Algorithm	31
4.3 Centralised DB Architecture	33
4.4 Real-World Sales RTDW Schema	34
4.5 Summary	35
CHAPTER 5	
EVALUATION	36
5.1 Performance evaluation for the JVM factors.	38
5.1.1 Streaming Application	39
5.1.2 Data Warehouse	41
5.2 Performance evaluation for response time.	42
CHAPTER 6	
CONCLUSION	44
6.1 Research Contribution	45
6.2 Limitation of the Research Approach	45
6.3 Future Works	46
REFERENCES	47
APPENDIX A : CD/DVD	52

LIST OF FIGURES

Figure 2.1: ETL process can be applied for the streams from diverse sources	07
Figure 3.1: Overview of the ETA process	15
Figure 3.2: Using @store annotation for extracting data from a file	16
Figure 3.3: Pre-process null values	16
Figure 3.4: Persist data in RTDW	17
Figure 3.5: Sample table schema for the use case	18
Figure 3.6: Sample table schema for temporary tables	19
Figure 3.7: Sample aggregation table schema for the “SALES” table	21
Figure 3.8: Aggregation logic for the “Days” granularity tables	22
Figure 3.9: Sample purging SQL	22
Figure 3.10: Retrieve total revenue per store in last week	24
Figure 3.11: Retrieve daily revenue per store per day	24
Figure 3.12: Code snippet for the average response time calculator	26
Figure 4.1: SingleInputStream Query Runtime	29
Figure 4.2: Filter expression execution in Siddhi	30
Figure 4.3: Summarised aggregation flow	32
Figure 4.4: Centralised DB architecture	33
Figure 4.5: Sample schema used in store use case	34
Figure 5.1: Table schema of the evaluated system	37
Figure 5.2: Code snippet of the bash script used in capturing JVM stats	39
Figure 5.3: Heap utilisation against TPS in the Streaming application	40
Figure 5.4: CPU utilisation against TPS in the Streaming application.	40
Figure 5.5: Heap utilisation against TPS in the RTDW	41
Figure: 5.6: CPU utilisation against TPS in the RTDW	41
Figure 5.7: Average response time against simultaneous users for different TPS values	42