

**DRDP : DYNAMICALLY RE-CONFIGURABLE
DATA PIPELINE IN THE EDGE NETWORK**

M.G.I.M. Nuwanthilaka

219376N

Master of Science in Computer Science

Department of Computer Science Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

July

DRDP : DYNAMICALLY RE-CONFIGURABLE DATA PIPELINE IN THE EDGE NETWORK

M.G.I.M. Nuwanthilaka

219376N

Thesis submitted in partial fulfillment of the requirements for the degree Master
of Science in Computer Science

Department of Computer Science Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

July

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a 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. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date: 2023/07/1

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Master of Science in Computer Science Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. Indika Perera

Signature of the Supervisor:

Date: 13/07/2023

DEDICATION

Dedicated to all my past and present teachers.

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude towards my supervisors Prof. Indika Perera and Dr. Gayashan Amarasinghe at the Department of Computer Science and Engineering, University of Moratuwa, for their patience, dedication and guidance through out the research. Without their guidance, supervision and the expertise, this research could not have been completed.

Also I extend my gratitude to all the staff from the Department of Computer Science and Engineering who help me to improve my knowledge in various subject areas in multiple ways during this study.

I would like to express my heartfelt appreciation to my family members and friends in my study group. I am grateful to them for their understanding and encouragement throughout the course to make this research a success.

I wish to express my gratitude to all my colleagues at MillenniumIT ESP Pvt. Ltd and Cut+Dry, Inc for the support given to manage my MSc research work.

ABSTRACT

Pipelines are a highly discussed topic in today's technological world. There are different variations of pipelines; Data Science pipelines, DevOps pipelines, and DevSecOps pipelines, etc. A data science pipeline usually comes with a fixed architecture, which can be problematic in a fast-growing tech industry. Traditional data science pipelines may struggle to handle the volume, velocity, and variety of data at the edge, necessitating more dynamic and adaptable approaches. Many advancements are happening to bring the technology to the edge due to substantial data points generated in the sensor networks at the edge; from factory floors to log streams.

So, in this thesis we first discuss the existing literature in the data pipeline domain under three main topics; data pipeline challenges, data pipeline architectures, and data pipeline security. Then we propose a methodology for dynamically re-configurable data pipeline architecture in the edge network. This way we expect to achieve more efficiency, controllability, and scalability of the data across networks. The emerging field of edge architecture presents opportunities for innovative approaches to data pipelines, enabling organizations to harness the full potential of edge data for advanced analytics, machine learning, and real-time decision-making. Further, we propose a prototype with Raspberry Pi-based programs to discuss the effectiveness of this novel method. Using this proposed architecture we have evaluated the results and later discussed how this benefits the current and future data pipeline implementation. We hope this contributes to the emerging edge architecture subject area.

Keywords: data science, pipeline, architecture, edge

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor	i
Dedication	ii
Acknowledgement	iii
Abstract	iv
Table of Contents	v
List of Figures	ix
List of Tables	x
List of Appendices	xi
1 Introduction	1
1.1 Problem Statement	2
1.2 Research Objectives	2
2 Literature Review	4
2.1 Data Pipeline Challenges	4
2.1.1 Manufacturing process data analysis pipelines: a requirements analysis and survey [1]	4
2.1.2 Examining the Challenges in Development Data Pipeline [2]	5
2.1.3 Data Pipeline Selection and Optimization [3]	6
2.1.4 Data Life cycle Challenges in Production Machine Learning: A Survey [4]	6
2.2 Data Pipeline Architectures	7
2.2.1 Putting Data Science Pipelines on the Edge[5]	7
2.2.2 Modelling Data Pipelines[6]	8
2.2.3 Scalable data pipeline architecture to support the industrial internet of things[7]	8
2.2.4 Feedback Driven Improvement of Data Preparation Pipelines[8]	9
2.2.5 An Edge-Based Framework for Enabling Data-Driven Pipelines for IoT Systems[9]	9
2.2.6 On the Design and Architecture of Deployment Pipelines in	

Cloud- and Service-Based Computing – A Model-Based Qualitative Study[10]	10
2.2.7 Edge Based Data-Driven Pipelines (Technical Report) [11]	11
2.2.8 Review of social media analytical process and Big Data Pipeline [12]	11
2.2.9 Data Pipeline Architecture for Serverless Platform [3]	11
2.2.10 Pipeline architecture for mobile data analysis [13]	12
2.2.11 Service Analysis and Network Diagnosis[14]	12
2.2.12 JITA4DS: Disaggregated Execution of Data Science Pipelines Between the Edge and the Data Centre[15]	13
2.2.13 An Automated Software Pipeline for Quantitative Susceptibility Mapping (QSM) Data Analysis [16]	14
2.2.14 Efficient pipelined flow classification for intelligent data processing in IoT [17]	14
2.2.15 Pipemizer: An Optimizer for Analytics Data Pipelines [18]	15
2.2.16 Systematic and benchmarking studies of pipelines for mammal WGBS data in the novel NGS platform [19]	16
2.2.17 Tiny-HR: Towards an interpretative machine learning pipeline for heart rate estimation on edge devices [20]	16
2.2.18 CenFind: a deep-learning pipeline for efficient centriole detection in microscopy datasets [21]	17
2.2.19 CPSReliP: an integrated pipeline for analysis and visualization of population structure and relatedness based on genome-wide genetic variant data [22]	18
2.2.20 Development of Big Data-Analysis Pipeline for Mobile Phone Data with Mobipack and Spatial Enhancement [23]	18
2.2.21 A Parallel Fuzzy Load Balancing Algorithm for distributed Nodes over a Cloud System [24]	19
2.2.22 DataPipeline: Automated Importing and Fitting of Large Amounts of Biophysical Data [25]	19

2.2.23	FastFlow: Accelerating Deep Learning Model Training with Smart Offloading of Input Data Pipeline [26]	20
2.2.24	Resource-Saving Customizable Pipeline Network Architecture for Multi-Signal Processing in Edge Devices [27]	21
2.2.25	RESPECT: Reinforcement Learning based Edge Scheduling on Pipelined Coral Edge TPUs [28]	22
2.2.26	bulkAnalyseR: an accessible, interactive pipeline for analysing and sharing bulk multi-modal sequencing data [29]	22
2.2.27	QuantPipe: Applying adaptive post-training quantization for distributed transformer pipeline in dynamic edge environment [30]	23
2.2.28	Smart Data Placement Using Storage-as-a-Service Model for Big Data Pipelines [31]	23
2.2.29	Pipeline Parallelism for Inference on Heterogeneous Edge Computing [32]	24
2.2.30	Efficient Computer Vision on Edge Devices with Pipeline-Parallel Hierarchical Neural Networks [33]	24
2.2.31	Optimizing Pipelined Computation and Communication for Latency Constrained Edge Learning [34]	25
2.2.32	DORIAN in action: Assisted Design of Data Science Pipelines [35]	25
2.2.33	An Alternative to Cells for Selective Execution of Data Science Pipelines [36]	26
2.2.34	RCE-NN: A Five-Stage Pipeline to Execute Neural Networks (CNNs) on Resource Constrained IoT Edge Devices [37]	26
2.3	Data Pipeline Security	27
2.3.1	Integration Of Security Standards in DevOps Pipelines [38]	27
2.3.2	Security Support in Continuous Deployment Pipeline [39]	28
2.4	Summary	28
3	Methodology	30
3.1	Introduction	30

3.2	Proposed Components	33
3.2.1	Device resource parameter extraction module	33
3.2.2	MQTT server	33
3.2.3	Reconfiguration architecture module	33
3.2.4	Algorithms module	33
3.2.5	Cloud component	34
4	Experiments	35
4.1	Selection of Use case	35
4.2	Develop the experiment with proposed design	36
4.2.1	Data Generator/Simulator	36
4.2.2	Data subscriber to MQTT server	37
4.2.3	Device Parameters Identification	39
4.2.4	Algorithms Module	39
5	Results	41
5.1	Performance enhancement	41
5.2	Cost reduction	43
5.3	Conceptual architecture development	44
6	Discussion and Conclusion	45
	References	47
	Appendix A Implementation Code	52
A.1	simulator.py	52
A.2	subscriber.py	53
A.3	system-analyzer.py	55
A.4	pipeline.py	57
A.5	requirements.txt	58
A.6	data-enricher.py	58
A.7	data-cleaner.py	59
A.8	algorithms.py	61

LIST OF FIGURES

Figure	Description	Page
Figure 3.1	Raspberry PI device used to this experiments	30
Figure 3.2	Module configuration	31
Figure 3.3	Switch gates modeling	32
Figure 4.1	Azure IoT Edge design	35
Figure 4.2	DRDP design	36
Figure 5.1	Total processing time vs scenarios	42
Figure 5.2	Cost vs No. of data points	44

LIST OF TABLES

Table	Description	Page
Table 5.1	Time taken to process in edge device for 4 steps	41
Table 5.2	Performance matrix for 10 data point processing in the entire pipeline architecture	42
Table 5.3	AWS HTTP API request Cost	43
Table 5.4	Cost matrix for 10M data point processing	43

LIST OF APPENDICES

Appendix	Description	Page
Appendix -A	Implementation Code	52