**Comparative Evaluation Framework for Count-Min Sketch Variants in Streaming Data Environments**

by

Sergei Rogov

Final Year project submitted in partial fulfilment of the requirements for the Degree of

Bachelor of Science in Computer Science

Department of Computer Science

School of Sciences and Engineering

University of Nicosia

(June 2025)

***Comparative Evaluation Framework for Count-Min Sketch Variants in Streaming Data Environments***

***By***

***Sergei Rogov***

This Final Year Project has been accepted in partial fulfilment of the requirements for the Degree of

Bachelor of Science in Computer Science

Project Advisor:

/ /

(Name) (Signature) (Date)

Examiner:

/ /

(Name) (Signature) (Date)

Examiner:

/ /

(Name) (Signature) (Date)

Department of Computer Science

School of Sciences and Engineering

University of Nicosia

**Abstract**

This project proposes a comparative evaluation framework for Count-Min Sketch and its variants, addressing the challenges of gathering statistics in streaming data environments. The framework is designed with a modular architecture, and can be used by researchers to compare the performance of different Count-Min Sketch summarization algorithms or other summarization algorithms capable of answering frequency queries. The framework’s architecture includes multiple core modules such as input stream generation, algorithms implementation, performance metrics evaluation, and visualization component. It benchmarks six Count-Min Sketch (CMS) variants, including Count-Min Sketch, Count-Mean-Min Sketch, Conservative Count-Min Sketch, Count Sketch Hierarchical Count-Min Sketch, and Sliding Count-Min Sketch. For evaluation metrics, accuracy, memory usage, and average query time are monitored. Experiments are conducted on both real and synthetic datasets and real time visualisations are displayed on a dynamic dashboard as data stream processing goes.

**Acknowledgements**

I would like to deeply thank my advisor, Dr. Demetris Trihinas, for his valuable feedback, technical guidance, and encouragement throughout the course of this Final Year Project. I am also extremely grateful to my parents for their unwavering both moral and financial support during the entire period my studies, especially over the past year.

Table of Contents

[Chapter 1 Introduction 4](#_Toc200024669)

[1.1 Motivation 4](#_Toc200024670)

[1.2 Problem Statement 4](#_Toc200024671)

[1.3 Project Goals 5](#_Toc200024672)

[1.4 Scope and Limitations 5](#_Toc200024673)

[1.5 Structure of Report 5](#_Toc200024674)

[Chapter 2 Background 8](#_Toc200024675)

[2.1 Characteristics of Streaming Data 8](#_Toc200024676)

[2.2 Challenges in Real-Time Summarization 10](#_Toc200024677)

[2.3 Count-Min Sketch Approach 11](#_Toc200024678)

[2.4 Related Work 11](#_Toc200024679)

[Chapter 3 Framework Architecture 12](#_Toc200024680)

[3.1 Technology Stack 12](#_Toc200024681)

[3.2 Modular Design Overview 12](#_Toc200024682)

[3.3 Stream Generation Module 12](#_Toc200024683)

[3.3.1 Real Dataset Source 12](#_Toc200024684)

[3.3.2 Synthetic Dataset Source 12](#_Toc200024685)

[3.4 Algorithms Module 13](#_Toc200024686)

[3.4.1 Count-Min Sketch 13](#_Toc200024687)

[3.4.2 Conservative Count-Min Sketch 13](#_Toc200024688)

[3.4.3 Count-Mean-Min Sketch 13](#_Toc200024689)

[3.4.4 Count Sketch 13](#_Toc200024690)

[3.4.5 Hierarchical Count-Min Sketch 14](#_Toc200024691)

[3.4.6 Time Decay Count-Min Sketch 14](#_Toc200024692)

[3.5 Evaluation Module 14](#_Toc200024693)

[3.6 Visualization Module 14](#_Toc200024694)

[Chapter 4 Experiments and Results 15](#_Toc200024695)

[4.1 Optimal Parameters Selection 15](#_Toc200024696)

[4.2 Cross-Algorithm Comparisons 15](#_Toc200024697)

[4.3 Benchmark Summary Table 15](#_Toc200024698)

[4.4 Accuracy Trade-offs per Variant 15](#_Toc200024699)

[Chapter 5 Discussion 16](#_Toc200024700)

[5.1 Comparative Strengths 16](#_Toc200024701)

[5.2 Limitations of Variants 16](#_Toc200024702)

[5.3 Ideal Use Cases 16](#_Toc200024703)

[Chapter 6 Conclusion 17](#_Toc200024704)

[6.1 Summary of Contributions 17](#_Toc200024705)

[6.2 Major Findings 17](#_Toc200024706)

[Chapter 7 Future Work 18](#_Toc200024707)

[7.1 Automatic Algorithm Selection 18](#_Toc200024708)

[7.2 Adaptive Resizing 18](#_Toc200024709)

[7.3 Semantic Mapping 18](#_Toc200024710)

[References 19](#_Toc200024711)

[Appendices 20](#_Toc200024712)

**List of Tables**

**List of Illustrations**

# Introduction

## Motivation

In the era of big data, where data is generated in rapidly growing volumes, with high velocity, and in various shapes and forms, there is an obvious need in efficient real-time analysis of data streams, as it faces big challenges in terms of computation and storage. Traditional methods of data storage and processing are inefficient because of the memory and processing speed limitations. Approximate data stream summarization methods, however, offer a more feasible solution, trading an adequate amount of accuracy for a significant improvement in memory utilization.

Popular example of such approximate data stream summarization algorithm is Count-Min Sketch (CMS) – a technique for efficiently frequency queries, meaning estimating how often an element appears in a stream, using limited memory. However, despite a wide use of CMS, there is no standardized framework for evaluation and comparing it with either its own variants or with other algorithms aimed on frequency queries. The lack of such platform makes it difficult to determine the most suitable algorithm for a specific case.

## Problem Statement

There is a need in a standardized evaluation framework for detailed comparison and research of different variants of Count-Min Sketch summarization algorithm. Such platform will enable researchers and developers to better understand the strengths and weaknesses of different variants of approximate summarization techniques in different scenarios, as well as compare them in a clear visual manner, considering parameters like accuracy, memory usage, and average time taken to answer a query.

## Project Goals

The project aim is to develop a comparative evaluation tool with modular architecture for benchmarking multiple CMS variants. The framework should include:

* Generation of a data stream, both synthetically and from real sources
* Implementation of Count-Min Sketch algorithm and its variants
* Evaluation module, monitoring algorithm performance, tracking key metrics
* Real-time visualisation dashboard for easier interpretation and comparison

## Scope and Limitations

The framework only covers a limited number of selected Count-Min Sketch algorithm variations and does not explore any other summarization techniques with the similar functionality. Additionally, the work is focused on numeric and text data streams that do not allow deletions. This way, in regards to elements in the stream, the algorithm instance can only:

* Update its state with the processed item
* Query the item, providing its estimated frequency in the stream

## Structure of Report

The report consists of seven core chapters, each targeting a specific aspect of a Count-Min Sketch comparative framework.

* Chapter 1 – Introduction

This chapter provides a high-level overview of project’s context and motivation, stating the specific problem this work aims to address. It explains the pressing demand in efficient data stream processing algorithms as well as the need in a framework for their detailed comparison for the purpose of both research and decision-making regarding the selection of a better suiting algorithm variation for a certain scenario.

* Chapter 2 – Background

This section introduces the concept of streaming data and challenges associated with processing it. Besides, it provides an overview of the Count-Min Sketch core principles, as well as the explanation on how exactly this family of algorithms operate on a data structure level. Related work is also covered in this chapter.

* Chapter 3 – Framework Architecture

This chapter is dedicated to the explanation of the modular architecture of the framework and the technology stack used during the project implementation, including the programming languages and libraries used. The purpose and the structure of each module is discussed, as well as the communication between components.

* Chapter 4 – Experiments and Results

This section presents a series of experimental studies carried out under different conditions, conducted using the implemented framework. Various parameters values are tested and optimal ones are determined, and the behaviour of different Count-Min Sketch variations is demonstrated on multiple data streams of different modalities, both for streams generated from real and synthetic datasets.

* Chapter 5 – Discussion

The chapter is dedicated to the analysis and the evaluation of the results from the previous chapter. Strengths and limitations of different Count-Min Sketch variants are shown, and the ideal use cases for algorithm variations are identified.

* Chapter 6 – Conclusion

Major findings and key takeaways from the work carried out are discussed in this chapter. The project summary is presented, highlighting the main milestones and results.

* Chapter 7 – Future Work

The closing chapter proposes the future work that can potentially extend the functionality of the framework implemented, expanding its scope by adding more advanced techniques and capabilities to the platform.

# Background

## Characteristics of Streaming Data

In the modern day and age, the phenomenon of Big Data has fundamentally changed data-driven system across various domains, such as finance, healthcare, Internet of Things, E-commerce, and others. Big Data offers significant advantages by enabling specialists to extract meaningful insights from data and to leverage the gained knowledge for the purposes such as workflow optimisation, revenue growth, or predicting human behaviour based on the common patterns mined from vast datasets. However, alongside these benefits, Big Data introduces challenges, particularly in more difficult data processing and data storage tasks. Streaming data, as a dynamic and time-sensitive subset of Big Data, adds even more complexities because of its continuous and high-speed nature.

Streaming data is often characterized by the “V” model, each dimension of which starts with the letter “V”:

* Volume

The problem with the overall amount of data being produced. Streaming systems must handle extremely large volumes of data generated from wide range of sources, such as sensors, social media platforms, mobile devices. Without the optimized algorithms, specifically tailored high-throughput environments, it would not be feasible to deal with such data streams. In many real-world scenarios, storing raw streaming data is either impractical or cost-prohibitive, making the use approximation techniques necessary to reduce memory utilization with a sacrifice of accuracy.

* Velocity

The challenge that refers to the of speed at which streaming data arrives and should be processed. In many cases, such as traffic monitoring or fraud detection, data should be handled instantaneously, in a single pass, as any latency may lead to the loss of critical information and inability to access data again. Efficient and low latency algorithms are crucial when dealing with velocity problem.

* Variety

Data could appear in a structured (e.g., tables), semi-structured (e.g., JSON, XML), and unstructured formats (e.g., text, video, audio), creating a challenge of inconsistency – handling such heterogeneous data types require high level of flexibility in processing logic.

* Veracity

The problem of data quality, which refers to the data reliability and accuracy. Not all the sources can be trusted, because the information coming from the source that lacks credibility could lead to faulty conclusions and decisions. Besides, data streams could contain noise and inconsistencies in it, requiring data stream processors to include filtering, anomaly detection and validation techniques to mitigate erroneous outcomes from streaming data handling.

* Value

Some data is less useful than the other. Distinguishing between low-value data and high-value data in real time is crucial, because processing resources are usually limited. In order to effectively make use of data, streaming systems should identify which sections of data carry the most insights. Useful patterns should only be found and extracted from most informative parts of data streams. For example, irrelevant or outdated data should be ignored.

* Variability

Data streams may fluctuate because of seasonal effects, user behaviour or other external events. Adaptive streaming models are required for these scenarios for better accuracy and memory utilisation.

* Visualization

The problem of presenting the big data to the target audience in a clear and understandable way. This involves highlighting key trends and showcasing overall picture while omitting noise and irrelevant details that do not covey any insights.

* Vulnerability

Data streams are vulnerable to cyberattacks. It is essential not only to handle the data itself, but also to protect it in transit, applying security measures.

* Volatility

Streaming data has an important property – once missed, it cannot be restored. That is why it is crucial for stream handling systems to perform real-time low latency data processing.

## Challenges in Real-Time Summarization

Since the handling of streaming data enforces memory limitations and latency constrains, fast and lightweight approximations are required. Exact methods are often too slow and memory-intensive, which means they are not applicable in real-time data streaming environments. Therefore, approximation techniques like sketching or sampling, which trade off accuracy for speed and efficiency, must be used. Designing such algorithms introduces new challenges as it requires wise balancing between memory error bounds, memory consumption and computational overhead.

## Count-Min Sketch Approach

Basic Count-Min Sketch (CMS) summarization algorithm has a clever design, allowing it to answer frequency queries in an approximate manner, with a fixed memory utilization budget. It works with numeric and text data streams, or any data streams consisting of hashable items.

Basic CMS algorithm consists of the following attributes and methods:

* 2D array of counters

The main data structure of the algorithm, which is used to store the counters associated with items from the data stream. The matrix has dimensions *depth* × *width*, meaningit consists of *depth* rows and *width* columns. This 2D array is updated or queried by the algorithm methods to maintain frequency estimates.

* Hash functions

The CMS algorithm instance has *depth* different hash functions, each associated with each of the *depth* rows in a 2D matrix of counters. These hash functions are used to hash data items and map them to each of *depth* rows of length *width*, either to update CMS with the item being processed or to query the item.

* Add method

In order to update the CMS instance with the new item, the method is used.

Each hash function *hi* maps this item to a column index in the row , effectively allocating this object to one of the *width* buckets in that row, where . Once the insert positions are determined, the corresponding counters in the 2D array are incremented.

By the nature of the algorithm and due to the memory constraints, the width of the 2D counter array is limited, making hash collisions unavoidable.

* Query method

In order to query the item in the CMS instance, the *query* method is used.

To answer a frequency query, algorithm hashes an item using each of the *depth* hash functions, mapping it to *depth* different positions in the 2D array of counters. Then, it retrieves the counter values from these positions and returns the minimum value among them as an estimated frequency.

Due to the hash collisions, overestimations can occur because multiple objects might be hashed to the same cell, causing their frequency count values to be combined. However, returning the minimum across all hashed positions helps to mitigate this effect.

In a classic Count-Min Sketch approach, the algorithm never underestimate – it can only return frequency estimates that are equal or greater than the ground truth frequencies.

In different CMS variants, the implementations of individual components may vary, depending on the strategy of a particular variant.

## Related Work

Previous research on Count-Min Sketch benchmark comparisons.

# Framework Architecture

## Technology Stack

The framework’s architecture includes multiple core modules such as input stream generation, algorithms implementation, performance metrics evaluation, and visualization component

## Modular Design Overview

The framework’s architecture includes multiple core modules such as input stream generation, algorithms implementation, performance metrics evaluation, and visualization component

## Stream Generation Module

Supports both real and synthetic datasets for stream generation.

### Real Dataset Source

#### FIFA Tweets

#### Kosarak

### Synthetic Dataset Source

Generation of data stream with skewed data distribution.

## Algorithms Module

### Count-Min Sketch

#### Algorithm

#### Benefits

#### Limitations

### Conservative Count-Min Sketch

#### Algorithm

#### Benefits

#### Limitations

### Count-Mean-Min Sketch

#### Algorithm

#### Benefits

#### Limitations

### Count Sketch

#### Algorithm

#### Benefits

#### Limitations

### Hierarchical Count-Min Sketch

#### Algorithm

#### Benefits

#### Limitations

### Time Decay Count-Min Sketch

#### Algorithm

#### Benefits

#### Limitations

## Evaluation Module

#### Accuracy

Comparison of predicted item frequencies with ground truth.

#### Memory Usage

Total memory utilized to store a Count-Min Sketch instance.

#### Average Query Time

Time taken to answer a frequency query on average.

## Visualization Module

#### Graphs Generation

Multiple charts showing algorithm behaviour.

#### Dynamic Dashboard

Real-time dashboard displaying all metrics of two algorithms chosen.

# Experiments and Results

## Optimal Parameters Selection

Empirical tuning of CMS width and depth parameters.

#### Width

#### Depth

## Cross-Algorithm Comparisons

Side-by-side evaluation under identical data conditions.

## Benchmark Summary Table

Tabular overview of key metrics across all variants.

## Accuracy Trade-offs per Variant

Analysis of performance variations per algorithm.

# Discussion

## Comparative Strengths

Each algorithm’s niche strengths outlined based on empirical results.

## Limitations of Variants

Discussion of scenarios where certain CMS types underperform.

## Ideal Use Cases

Recommendations based on data type, skewness, and resource constraints.

# Conclusion

## Summary of Contributions

Outlines the framework’s capabilities and its role in CMS benchmarking.

## Major Findings

Highlights key differences observed among the variants.

# Future Work

## Automatic Algorithm Selection

Heuristics for picking the best CMS variant per data type or per dataset type.

## Adaptive Resizing

Dynamic adjustment of sketch dimensions.

## Semantic Mapping

Semantic item clustering instead of hashing.

# References

<http://dimacs.rutgers.edu/~graham/pubs/papers/timedecaypodc.pdf>

<https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/CharikarCF.pdf>

<https://aclanthology.org/D12-1100.pdf>

# Appendices