Cardinality Estimation of Distinct Items - An Overview

Sebastian Balsam

November 3, 2025

Abstract

In this paper I want to give an overview of different solutions for the Count Distinct Problem...

1 Introduction

The cardinality estimation problem or count-distinct problem is about finding the number of distinct elements in a data stream with repeated elements. These elements could be url's, unique users, sensor data or IP addresses passing through a data stream. A simple solution would be to use a list and add an item to this list each time we encounter an unseen item. With this approach we can give an exact answer to the query, how many distinct elements have been seen. Unfortunately if millions of distinct elements are present in a data stream, with this approach, we will soon hit storage boundaries and the performance will detoriate.

As we often do not need an exact answer, streaming algorithms have been developed, that give an approximation that is mostly good enough and bound to a fixed storage size. There different approaches to solve this problem. Sampling techniques are used as an estimate by sampling a subset of items in the set and use the subset as estimator, e.g the CVM algorithm by Chakraborty et al. [Chakraborty et al., 2023]. While sampling methods generally give a good estimate, they can fail in certain situations. If rare elements are not in the sampled set, they would not be counted and we would underestimate. For database optimization machine learning techniques have been used recently for cardinality estimation [Liu et al., 2015, Woltmann et al., 2019, Schwabe and Acosta, 2024].

The technique I want to focus in this paper are sketch techniques that implement a certain data structure and an accompanying algorithm to store information efficient for cardinality estimation.

2 Probabilistic Counting with Stochastic Averaging

- Overview of PCSA - Problem description

2.1 Flajolet & Martin

In the 80's streaming problems were not the problems, we have today, as there simply were not that many data streams. But databases existed and began to grow in size. As table sizes grew, evaluation strategies became important, how to handle such big data tables for join operations. Query optimizers were developed that could find the optimal strategy.

In this context Flajolet and Martin developed a method to estimate the number of distinct items in a set in a space efficient way[Flajolet and Martin, 1985].

Algorithm 1 The Flajolet Martin algorithm.

```
Require: $n \geq 0$
Ensure: $y = x^n$
$y \leftarrow 1$
$X \leftarrow x$
$N \leftarrow n$
while $N \neq 0$ do
    if $N$ is even then
        $X \leftarrow X \times X$
        $N \leftarrow \frac{N}{2}$                ▷ This is a comment
    else if $N$ is odd then
        $y \leftarrow y \times X$
        $N \leftarrow N - 1$
    end if
end while
```

- Compressed (Scheuermann)

3 HyperLogLog

4 Conclusion

References

[Chakraborty et al., 2023] Chakraborty, Vinodchandran, and Meel (2023). Distinct elements in streams: An algorithm for the (text) book.

[Flajolet and Martin, 1985] Flajolet, P. and Martin, G. N. (1985). Probabilistic counting algorithms for data base applications.

[Liu et al., 2015] Liu, Xu, Yu, Covinelli, and ZuZarte (2015). Cardinality estimation using neural networks.

[Schwabe and Acosta, 2024] Schwabe and Acosta (2024). Cardinality estimation over knowledge graphs with embeddings and graph neural networks.

[Woltmann et al., 2019] Woltmann, Hartmann, Thiele, Habich, and Lehner (2019). Cardinality estimation with local deep learning models.