2023-11-26

应用系统体系架构 — 作业9

学号：521030990006

姓名：VAHAGN GHAZARYAN

## **Install InfluxDB on your own machine , and monitor the status of your laptop as demonstrated in the course. Take a screenshot in Explore on the web interface and paste it in a Word document, and briefly describe your laptop based on the screenshot. Operating status.**

## I used influxdb and telegraf docker containers to setup my environment. They are connected with network to communicate with each other and set the url to influxdb (not localhost) in telegraf.conf.

## 

## I passed my cpu and memory information over time for test.

## 

## 

## For instance here you can see my free memory over time.

## 

## Here you can some of my cpu and total cpu usage over time.

## **B.Please use a Word document to answer the following questions:**

## ***i.Please explain what are the meanings of read amplification and write amplification in log-structured databases? ( 2 minutes)***

## **Read Amplification** is happening when more data than necessary is read from the storage .To retrieve a small amount of data, the system may need to read a larger set of log entries, which causes read amplification.

## In log-structured databases, read amplification is a significant issue due to the way data is stored across different layers, each potentially containing different versions of the data. To retrieve a specific piece of information, the system might need to access **multiple data files across these layers,** leading to a higher number of read operations than would be necessary in a more straightforward data storage layout. This not only increases the latency of read operations but also limits the performance of AP (Availability and Partition tolerance) queries in a distributed system.

## **Write Amplification** happens when the amount of data written to storage is more than the amount of data that the user intends to write. In log-structured databases, this occurs because every write operation involves writing a new log entry, even for small updates. This leads to more data being written than the original data size, especially due to metadata and additional information that needs to be logged.

## The write amplification issue in log-structured databases is characterized by several factors:

## Low-Latency Inserts

## Asynchronous Disk Writes

## Write Blocking:

## Severe Write Amplification Due to Asynchronous Writes

## *ii.Please explain the specific calculation methods used in more than two different similarity calculations in vector databases? (1 point)*

## Approaches:

## **Random Projection:** This method involves transforming high-dimensional vectors into a space of lower dimensions using a matrix of random projections.

## **Product Quantization:** This technique segments the original vector into smaller parts, condenses each part into a compact “code” that represents it, and then reassembles these parts.

## **Locality-Sensitive Hashing:** This is a strategy for indexing when conducting an approximate search for the nearest neighbors.

## **Hierarchical Navigable Small World (HNSW):** This approach constructs a layered, tree-like framework where each node symbolizes a collection of vectors. The connections between nodes indicate the resemblance between these vectors.

## Metrics for Similarity:

## **Cosine Similarity:**

## The scale here is from -1 to 1, with 1 indicating vectors that are exactly the same, 0 indicating vectors at right angles, and -1 indicating completely opposite vectors.

## **Euclidean Distance:**

## This metric varies from 0 to an undefined upper limit, where 0 means vectors are identical, and higher values indicate increasing dissimilarity.

## **Dot Product:**

## The range for this measure is from negative infinity to positive infinity. A positive value suggests vectors that are aligned in the same direction, 0 indicates vectors that are perpendicular, and a negative value indicates vectors that are in opposite directions

## 