CS 411 - Database Systems **Spring 2023** **PT2 - Research Proposal**

Submitted by -Payal Mantri (payaljm2)

Exploring the Use of Presto as a Query Engine for Vector Databases

Introduction

Vector databases are becoming increasingly important in a variety of applications, including machine learning, computer vision, natural language processing, and recommendation systems. Vector databases are scalable data platforms to store, index, and query across embedding vectors that are generated from unstructured data (images, text, etc.) using deep learning models. These databases are designed to efficiently store and retrieve high-dimensional vector.
However, querying large collections of vectors can be a challenging task, as traditional SQL query engines are not optimized for similarity search and other vector-specific operations.
Presto[4] is an open-source distributed SQL query engine that is designed to query data from a wide variety of data sources, including traditional relational databases, NoSQL databases, and data lakes. It is highly scalable and can be deployed on clusters of commodity hardware, making it well-suited for big data environments.
This research proposal aims to explore the use of Presto, an open-source distributed SQL query engine, for querying vector databases.

Motivation

Traditional SQL query engines are not well-suited for querying vector databases, as they are not optimized for similarity search and other vector-specific operations. As a result, there is a growing interest in distributed query engines such as Presto, which can be used to query vector databases in a scalable and efficient manner. However, there is currently a lack of research on the use of Presto as a query engine for vector databases, and there is a need to investigate the feasibility and performance of this approach. This research proposal aims to address this gap in the literature by exploring the use of Presto for querying vector databases, and developing best practices and guidelines for using Presto with different types of vector databases.

Related Work

Milvus[2] is an open-source vector database designed to store, manage, and search large-scale vector data. It is optimized for similarity search and other vector-based operations, and it provides a range of features to support machine learning applications.

Weaviate[3] is an open source vector database that stores both objects and vectors. This allows for combining vector search with structured filtering. Weaviate can be used stand-alone (aka bring your vectors) or with a variety of modules that can do the vectorization for you and extend the core capabilities.It has a GraphQL-API to access your data easily.

All these vector databases typically have their own query language or API, which can make it difficult to work with multiple databases in a unified way. By using a common query engine like Presto, it is possible to write queries that can be executed across multiple vector databases, even if they have different query languages or APIs.

Methodology/Approach

The proposed research will involve a series of experiments to investigate the feasibility and performance of using Presto as a query engine for vector databases. The methodology will include the following steps:

- **Selection of vector databases:** A range of vector databases will be selected for the experiments, including databases that are optimized for similarity search and other vector-specific operations.
- **Configuration of Presto:** Presto will be configured to connect to each of the selected vector databases using the appropriate driver or connector
- **Experiment design:** A series of experiments will be designed to evaluate the performance of Presto when used to query the selected vector databases. The experiments will vary the size and complexity of the vector data, the query workload, and the number of nodes in the Presto cluster.
- **Performance evaluation:** The performance of Presto will be evaluated in terms of query speed, accuracy, and scalability. The results will be compared to other query engines and tools that are commonly used with vector databases.
- **Development of best practices and guidelines:** Based on the results of the experiments, best practices and guidelines will be developed for using Presto with different types of vector databases.

Research Questions

The proposed research will aim to answer the following research questions:

1. Can Presto be used as an efficient and scalable query engine for vector databases?
2. How does the performance of Presto compare to other query engines and tools that are commonly used with vector databases?
3. What are the best practices and guidelines for using Presto with different types of vector databases?
4. How does the performance of Presto vary with different types of vector data and query workloads?
5. What are the key challenges and limitations of using Presto with vector databases, and how can they be addressed?

Answering these research questions will provide valuable insights into the feasibility and performance of using Presto as a query engine for vector databases, and will help to develop best practices and guidelines for using Presto with different types of vector databases.

References

1. Guo, R., Luan, X., Xiang, L., Yan, X., Yi, X., Luo, J., ... & Xie, C. (2022). Manu: a cloud native vector database management system. *arXiv preprint arXiv:2206.13843*.
2. Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., ... & Xie, C. (2021, June). Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data* (pp. 2614-2627).
3. https://weaviate.io/developers/weaviate
4. Sethi, R., Traverso, M., Sundstrom, D., Phillips, D., Xie, W., Sun, Y., ... & Berner, C. (2019, April). Presto: SQL on everything. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)* (pp. 1802-1813). IEEE.