Database Performance Comparison

Project Proposal

1. Executive Summary

Our project aims to compare the performance of three open-source database systems: PostgreSQL, MySQL, and MongoDB, for handling multimodal data from the YouTube-8M Dataset. We will evaluate these databases along the axes of query performance and consistency using a large, multimodal dataset on local machines. The impact of this project will be to provide insights into selecting the most appropriate database system for applications dealing with complex, multimodal data, potentially improving application performance and user experience in video analytics and recommendation systems.

2. Project Background

Modern applications, especially in the realm of video analytics and content recommendation, face challenges in managing large volumes of multimodal data while maintaining high performance. While there's research on individual database performance, there's a gap in understanding how different types of databases perform for specific multimodal data scenarios on commodity hardware.

Studies like "A Comparison of NoSQL Databases for Modern Applications" (Yassien et al., 2020) provide general comparisons, but lack the multimodal data focus our project proposes.

Our study aims to fill this gap by providing a targeted comparison for applications dealing with video, audio, text, and image data on locally available resources.

3. Solution Presentation

Vision

Create a comprehensive comparison of PostgreSQL, MySQL, and MongoDB for applications handling multimodal data, focusing on query performance and consistency using locally available computing resources.

Project Schedule (14-week semester)

- Week 1-2: Literature review and project setup
- Week 3-4: Dataset preparation and database setup
- Week 5-8: Conduct performance tests
- Week 9-11: Data analysis
- Week 12-14: Report writing and presentation preparation

Team Roles

- Project Manager & Database Specialist: Oversee project and manage databases
- Application Developer: Create testing scenarios for multimodal queries
- Data Analyst: Analyze test results
- Technical Writer: Compile findings into the report

Risk Register

- Risk: Limited computing power on local machines for large-scale multimodal data tests Mitigation: Use a subset of the YouTube-8M dataset and optimize queries for each data type
- Risk: Difficulty in creating equivalent scenarios across different database types for multimodal data Mitigation: Focus on equivalent functionality rather than identical queries, using database-specific optimizations where necessary

Deliverables

- Final report detailing performance comparisons
- Presentation of key findings
- GitHub repository with test scripts and results

Reporting

- Bi-weekly progress reports
- Final project report and presentation

4. Project Deliverables and Goals

Final Deliverables

- Comprehensive report (by Week 13)
- Presentation slides (by Week 14)
- GitHub repository (continuous updates, finalized by Week 14)

SMART Goals

1. Set up testing environments for PostgreSQL, MySQL, and MongoDB on local machines by the end of Week 4.
2. Develop and execute at least 5 multimodal data query scenarios by the end of Week 8.
3. Analyze performance data for all three databases across all scenarios by the end of Week 11.
4. Produce a final report with actionable insights for multimodal data applications by the end of Week 13.
5. Deliver a final presentation and publish the GitHub repository in Week 14.

5. Required Resources

Dataset

We will use the YouTube-8M Dataset available on Kaggle. (https://www.kaggle.com/datasets/youtube/youtube-8m) This dataset includes multimodal data from millions of YouTube videos, providing the necessary scale and complexity for our study.

Key components of the dataset:

1. Video data
2. Audio tracks from the videos
3. Text data (video titles, descriptions, and tags)
4. Extracted image frames from the videos

Hardware and Software

- Personal computers of team members for database hosting and testing
- Open-source database systems: PostgreSQL, MySQL, and MongoDB

- Open-source data analysis tools (e.g., Python with pandas, matplotlib)
- Git for version control and GitHub for project hosting

6. Conclusion

Our project addresses the need for a multimodal data-specific comparison of PostgreSQL, MySQL, and MongoDB on locally available resources. By conducting a focused study using the YouTube-8M dataset, we will provide valuable insights to help businesses select the most appropriate database system for their applications dealing with video, audio, text, and image data. This project has the potential to significantly impact database selection decisions for applications in video analytics, content recommendation, and other multimodal data processing fields, leading to improved application performance and user experience within the constraints of local computing resources.