Middleware for Supporting “Big Data” Analytics Across a Database Cluster
Daniel Salowe, Shayan Patel, Sahil Shah
Project Advisor: Zachary Ives, PhD

OVERVIEW
By leveraging a distributed infrastructure to optimize MySQL queries and a simple web UI, our software enables clients to easily run their own computations over their data.

MOTIVATION
- Need for efficient, big data analytics across a relational database cluster
- Lack of affordable and open-source support
- Scalability and reliability of data

KEY COMPONENTS
1. Distributed JOIN operations
2. Fault tolerance
3. Distributed infrastructure
4. Support for efficient computations before or after queries
5. Web UI

DESIGN/SYSTEM ARCHITECTURE

RESULTS
- Baseline JOIN (1 Node) was run after adding data from another node to compare to our use of replication
- With more nodes added to our system, we can expect speed increases
- This is due to the scalability of the system shown in the graph above

DISCUSSION
- Use a low-level language such as C to increase efficiency
- Increase concurrency

POSSIBLE OPTIMIZATIONS
- Improve fault tolerance for JOIN
- More MySQL query support

FUTURE DIRECTIONS