Introduction

Parallel data processing methods have allowed people of all professions to analyze data at an unprecedented scale, yielding key insights in myriad industries and academic disciplines. Contemporaneously (and not entirely independently), data privacy has become an important and immediate issue in our society. Organizations in both the public and private sectors constantly collect or access data on their users or constituents, often without explicit consent by the individual. As computing continues to shift into "the cloud" from users' local systems, user data is relinquished to, and becomes centralized in, the small number of prevalent cloud service providers. In distributed parallel data processing, where clusters of multiple physical or virtual machines execute jobs in parallel, this introduces a potential for privacy and confidentiality compromises of users' data, as well as results of their computations. Even an entirely privacy-respecting benign cluster provider, the contemporary model of centralized cluster control provides an attack focus for malicious hackers. Though there are many different architectures for parallel and distributed processing, we will focus exclusively on the MapReduce paradigm in this project.

Problem Statement

So, how can we protect the privacy of users' computations while still affording the user all (or at least most) of the major features of modern distributed data processing architectures? We need a system that:

1. is decentralized (peer-to-peer) in architecture with granular data access control, to prevent a malicious peer from being able to see too great a portion of the user's data.[1]
2. preserves the users' anonymity[2], keeping it hidden from other peers in the cluster.
3. maintains compatibility with existing MapReduce implementations without requiring the user to re-write large portions of their code base.

TorMR

Deliverables

1. code for simulations/proof of concept
2. code documentation

[1] this may be susceptible to sybil attacks

[2] or at least pseudonymity

3. formal document containing description of project, implementation details & design decisions, and analysis of experimental results.