



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Mining of Massive Datasets

Steven Lang

June 5, 2016

Motivation

Cluster Computing

Distributed File System (DFS)

MapReduce

Introduction

Terminology

Example: WordCount

Combiners

Further details on MapReduce

Coping with node failures

Matrix-Vector-Multiplication

- V Fits into main memory

- V does not fit into main memory

Relational Algebra Operations

Communication Cost

- Example: Natural Join

- Example: Cascaded Two-Way Joins

Complexity Theory for MapReduce

- Reducer Size and Replication Rate

- Example: Similarity Joins

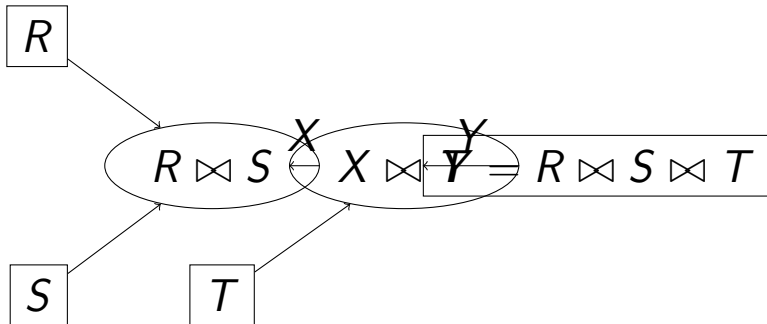


Figure: Schematic of a two two-way joins

