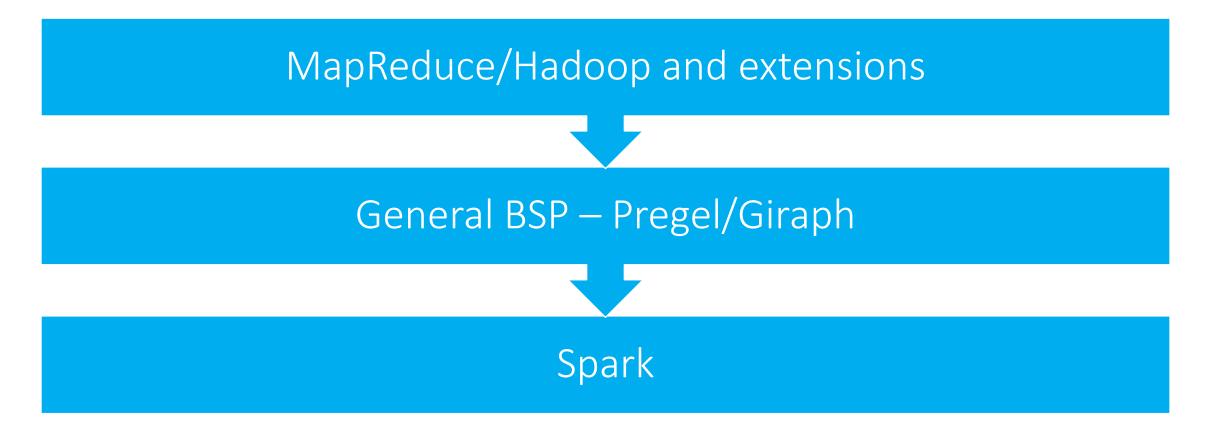
# Distributed Regression using Apache Spark

Neil Satra







#### ✓ Speed

- ✓ Ease of Use
- ✓ Generality
- ✓ Runs Everywhere

### MapReduce WordCount in Spark

wordCounts.collect()

#### The Problem to Solve



### How the application works

- 1. Write code in high-level scala, running transformations on the abstract collection (RDDs)
- 2. Package it into a jar file
- 3. Tell Spark the cluster to deploy the code on

#### Tasks

#### • Setup

- $\,\circ\,$  Get Scala installed and running
- o Get Spark installed and running
- Get a Mesos cluster running with Scala + Spark

#### • Implement regression

- $\circ$  Manually
- $\circ$  Using Mlib
- $\circ$  Using Mahout
- Benchmark performance
  - $\circ$  On a single node
  - $\circ$  On clusters of various sizes
  - $\circ~$  With each of the three implementations

## Why?

Testing assumption that distributed = better