Matrix Completion with ALS

Reza Zadeh

@Reza_Zadeh | http://reza-zadeh.com

Collaborative Filtering

Goal: predict users' movie ratings based on past ratings of other movies

$$R = \begin{pmatrix} 1 & ? & ? & 4 & 5 & ? & 3 \\ ? & ? & 3 & 5 & ? & ? & 3 \\ 5 & ? & 5 & ? & ? & ? & 1 \\ 4 & ? & ? & ? & ? & 2 & ? \end{pmatrix}$$

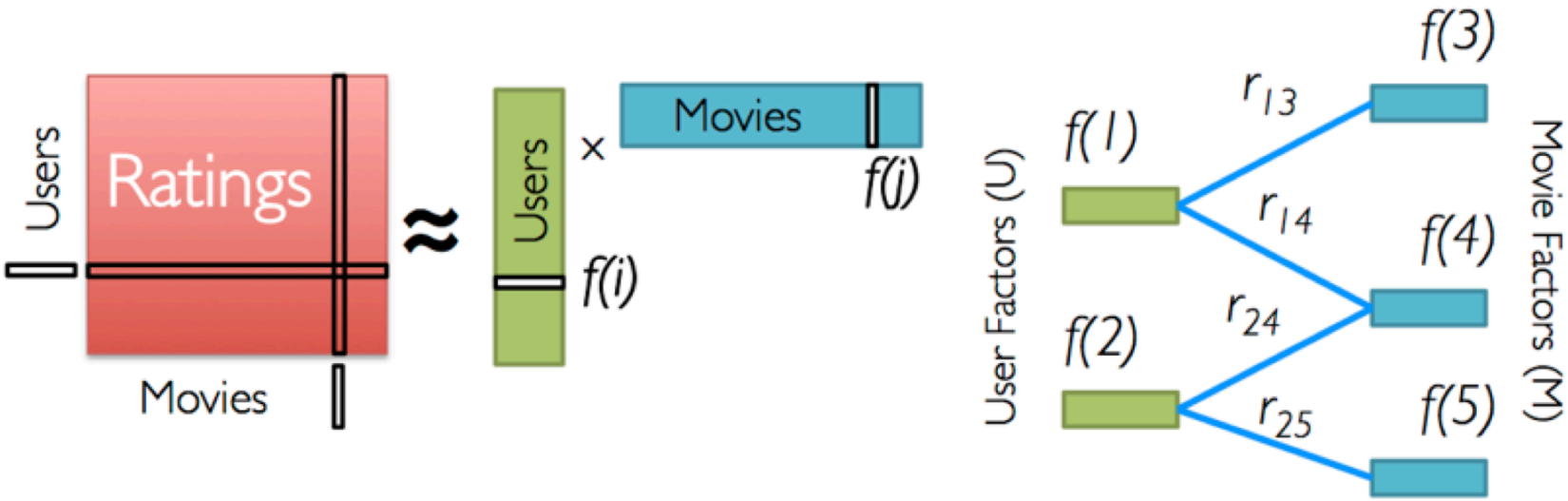
Users

Movies

Don’t mistake this with SVD.

Both are matrix factorizations, however SVD cannot handle missing entries.

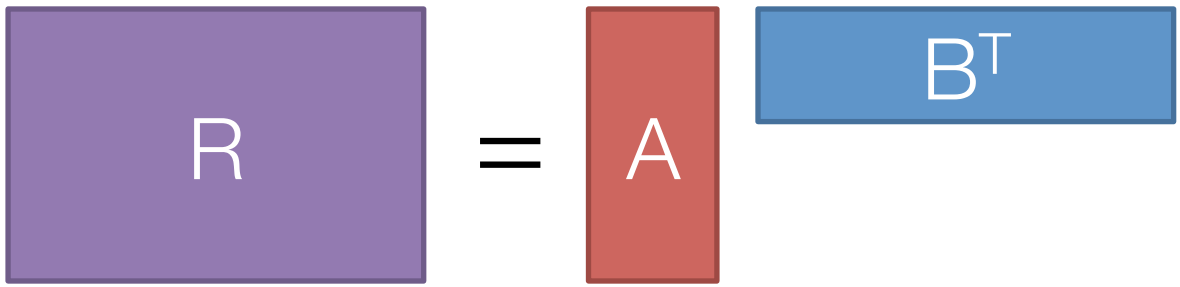
Optimization problem



Iterate:

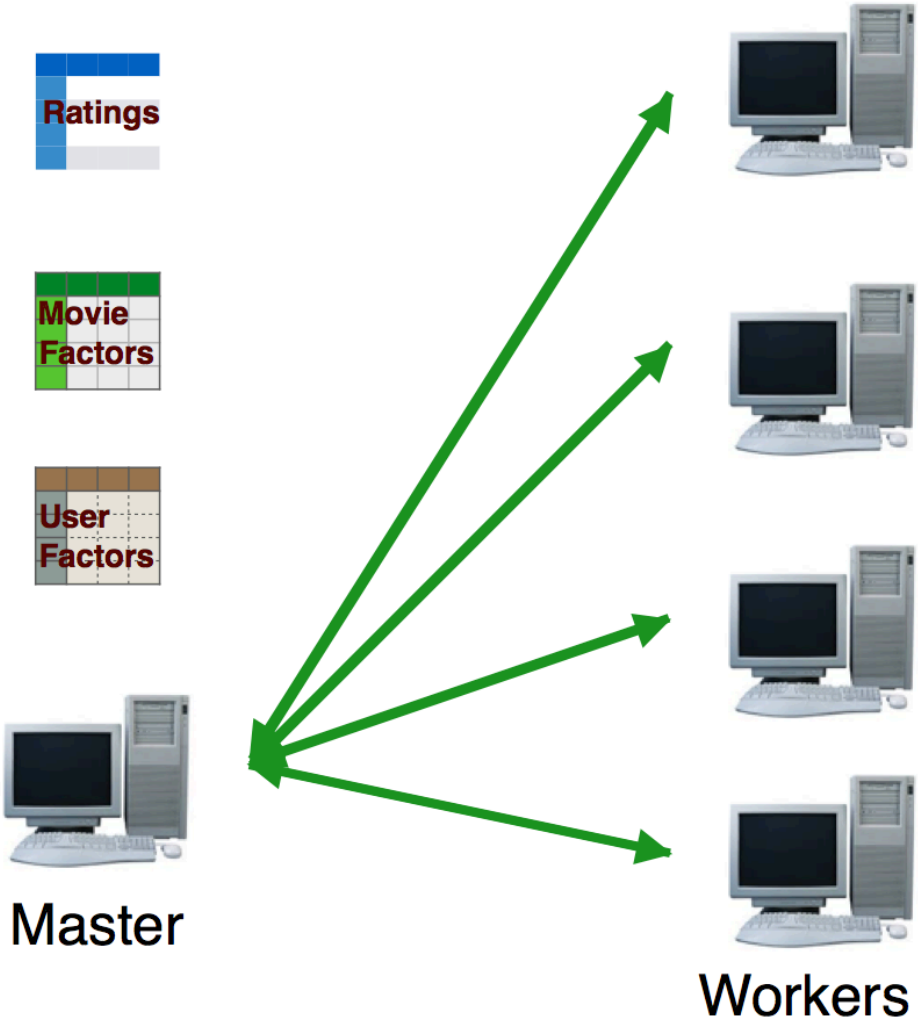
$$f[i] = \arg\min_{w \in \mathbb{R}^d} \sum_{j \in \mathrm{Nbrs}(i)} \left(r_{ij} - w^T f[j]\right)^2 + \lambda ||w||_2^2$$

Alternating Least Squares



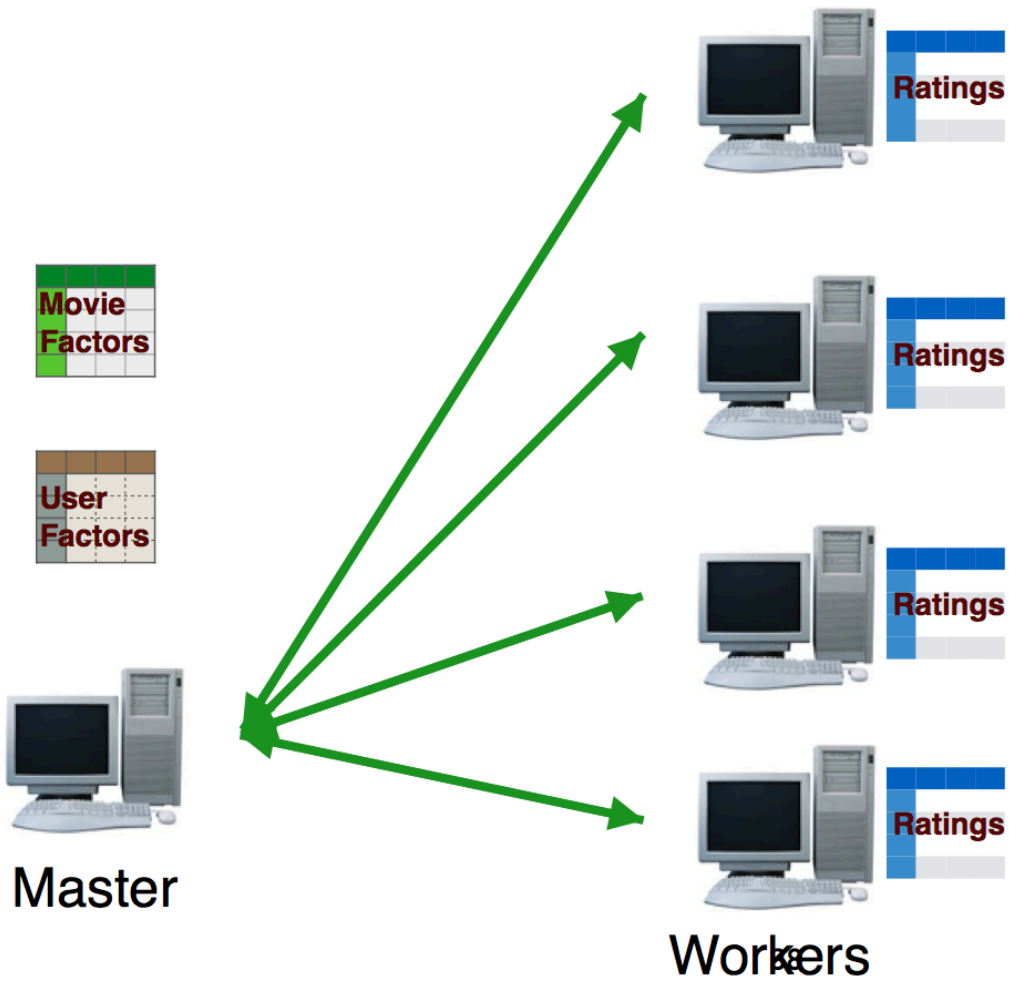
1. Start with random A_1, B_1
2. Solve for A_2 to minimize $||R - A_2 B_1^T||$
3. Solve for B_2 to minimize $||R - A_2 B_2^T||$
4. Repeat until convergence

Attempt 1: Broadcast All



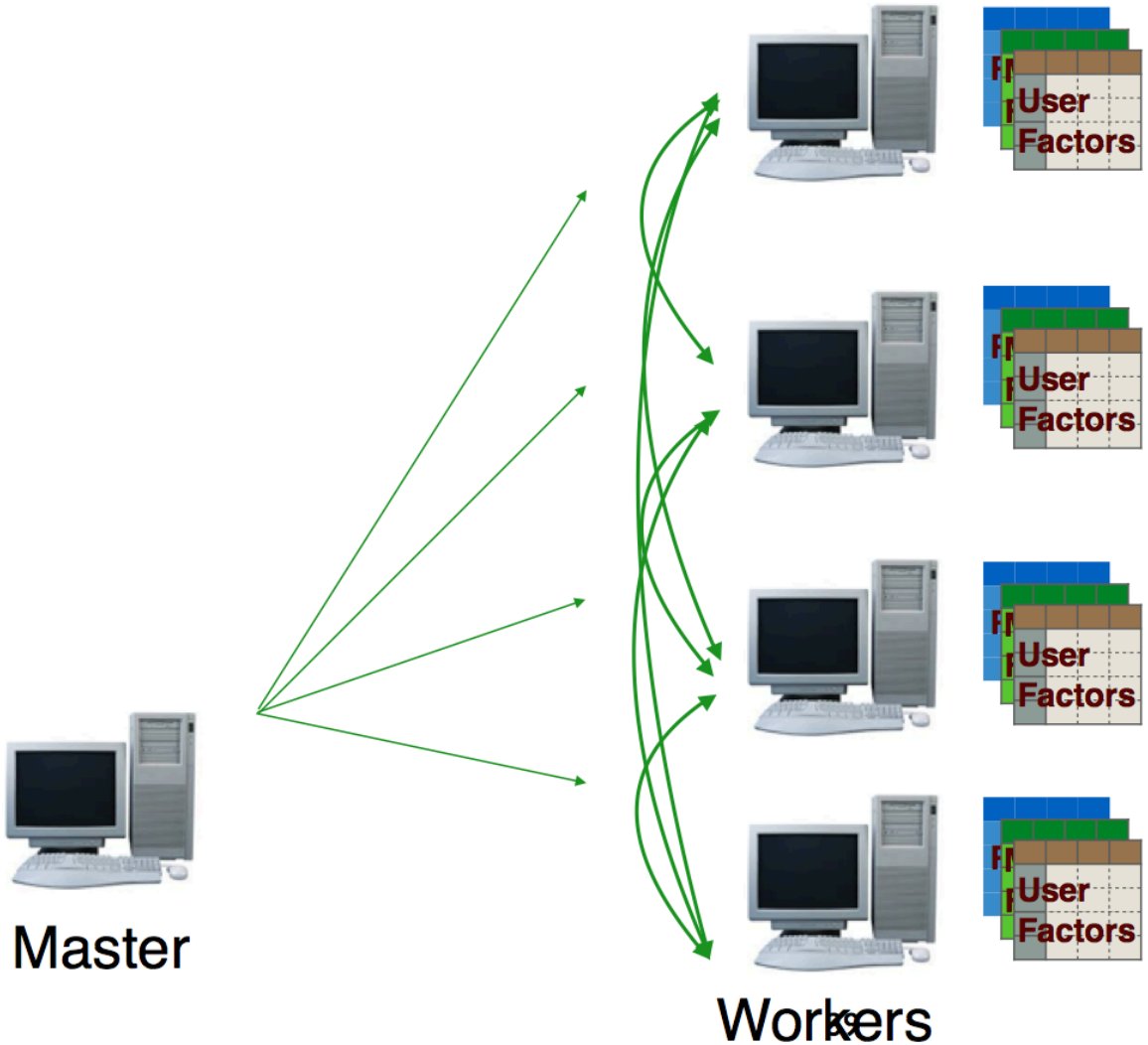
- Master loads (small) data file and initializes models.
- Master broadcasts data and initial models.
- At each iteration, updated models are broadcast again.
- Works OK for small data.
- Lots of communication overhead - doesn't scale well.

Attempt 2: Data Parallel



- Workers load data
- Master broadcasts initial models
- At each iteration, updated models are broadcast again
- Much better scaling
- Works on large datasets
- Works well for smaller models. (low K)

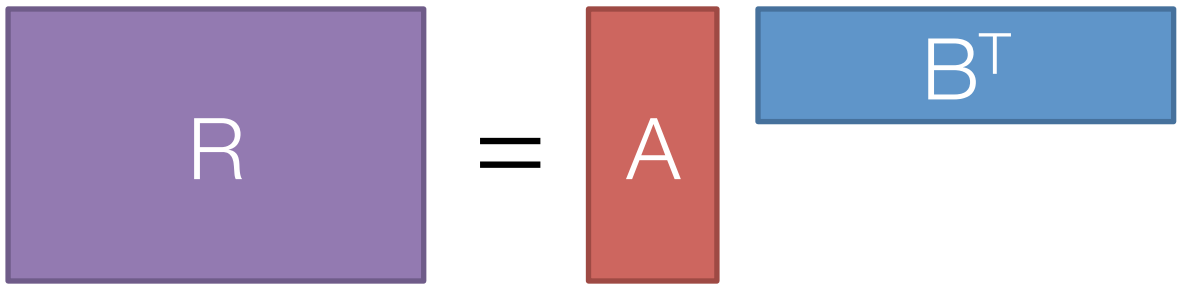
Attempt 3: Fully Parallel



- Workers load data
- Models are instantiated at workers.
- At each iteration, models are shared via join between workers.
- Much better scalability.
- Works on large datasets

ALS on Spark

Matei Zaharia,
Joey Gonzales,
Virginia Smith

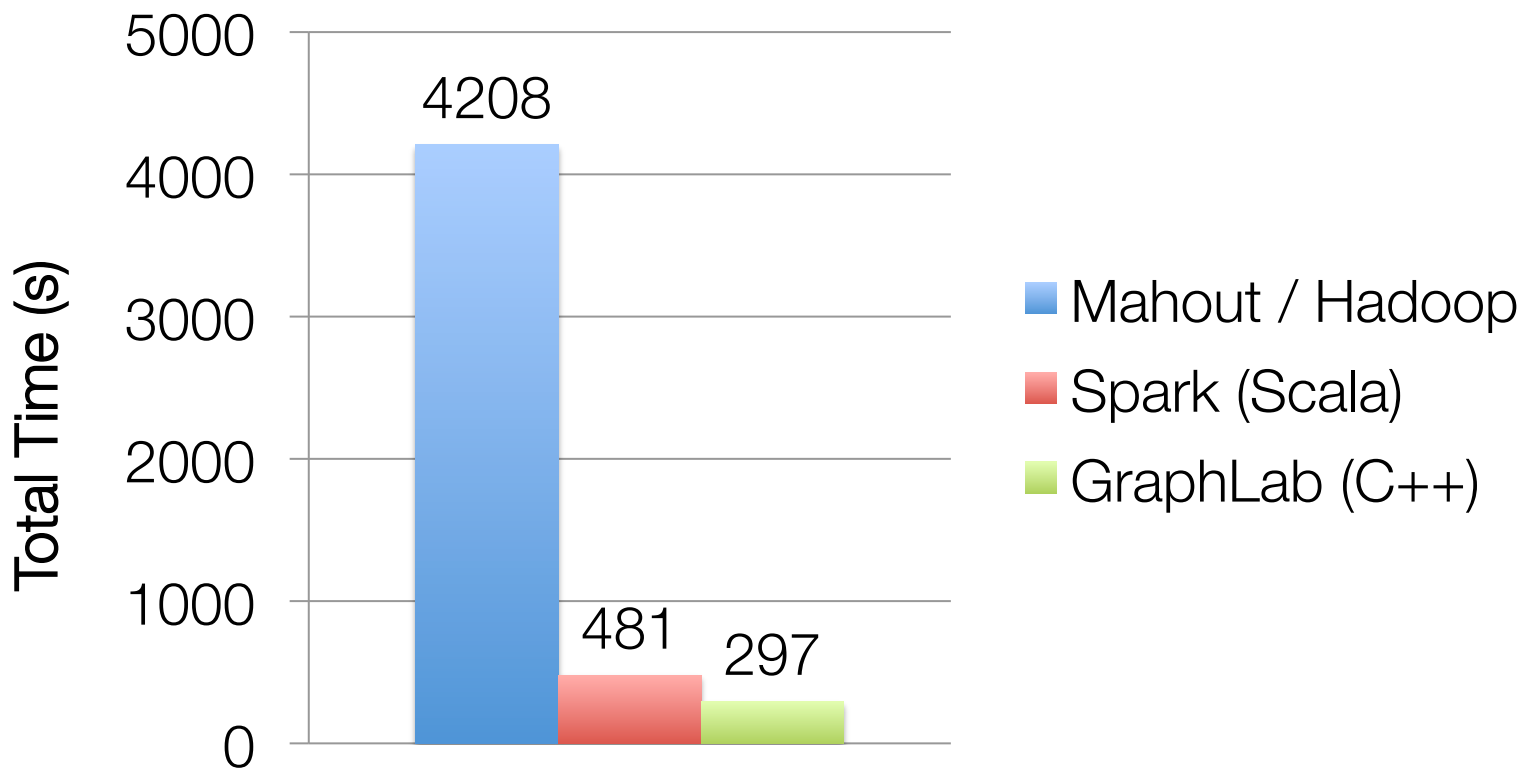


Cache 2 copies of R in memory, one partitioned by rows and one by columns

Keep A & B partitioned in corresponding way

Operate on blocks to lower communication

ALS Results

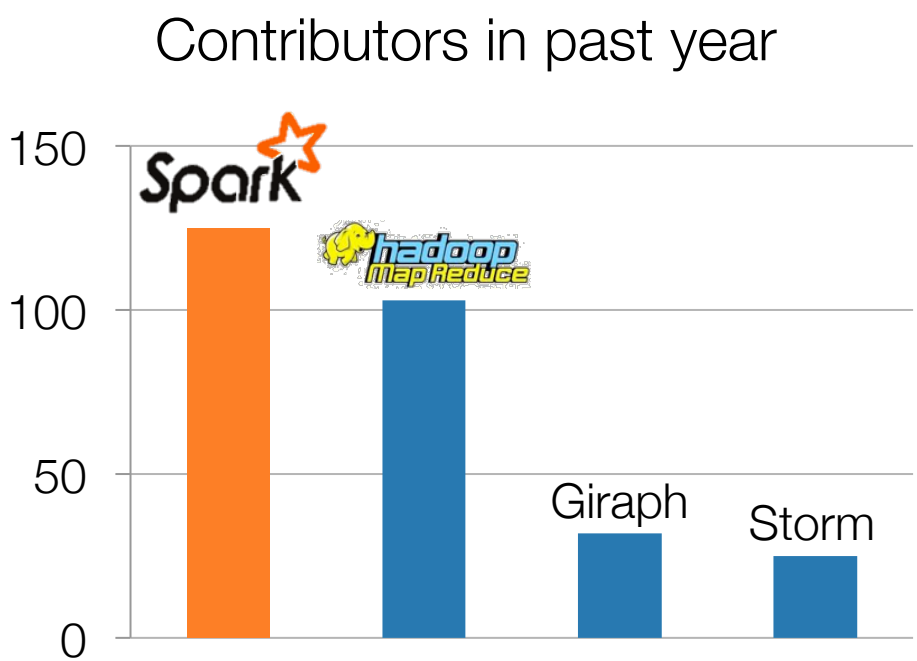

Total Time (s)
5000
4000
3000
2000
1000
0
4208
481
297
Mahout / Hadoop
Spark (Scala)
GraphLab (C++)

State of the Spark ecosystem

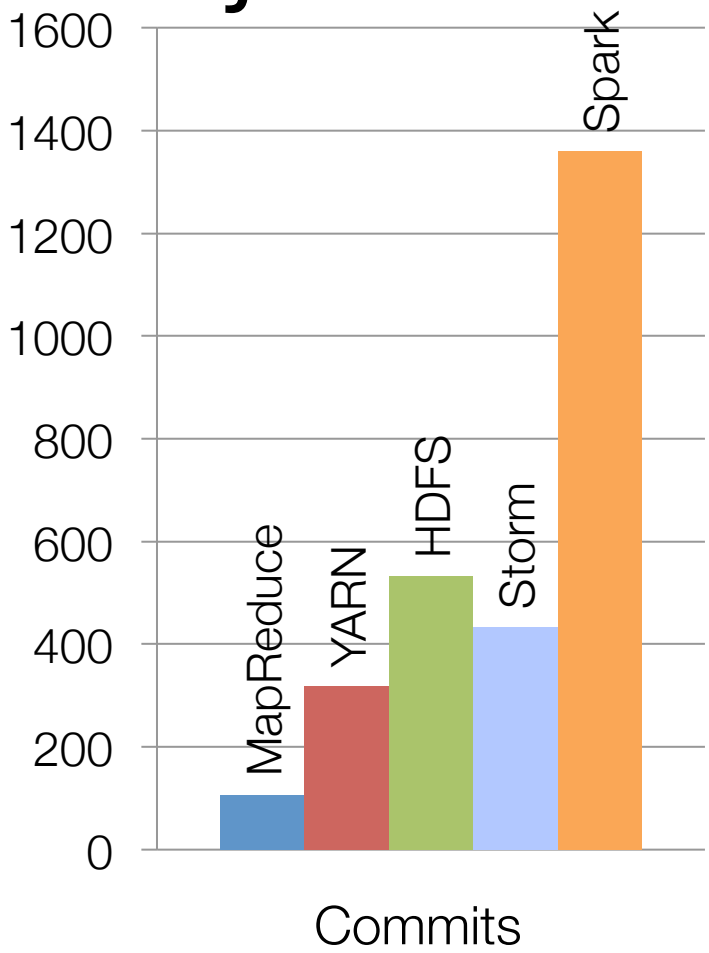
Spark Community

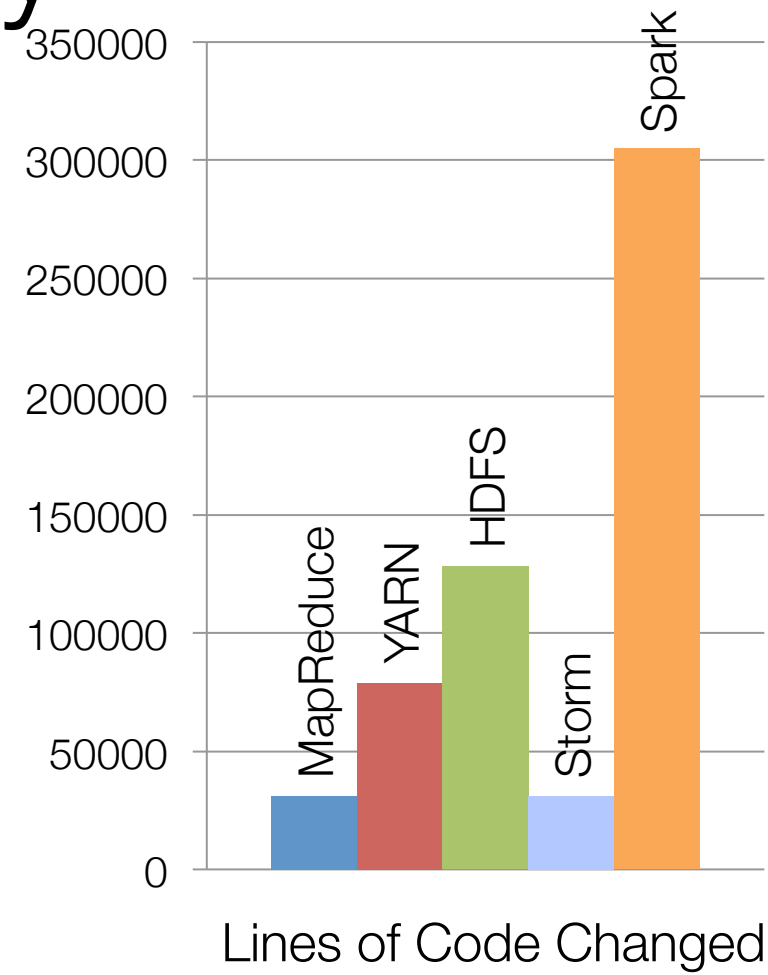
Most active open source community in big data

200+ developers, 50+ companies contributing



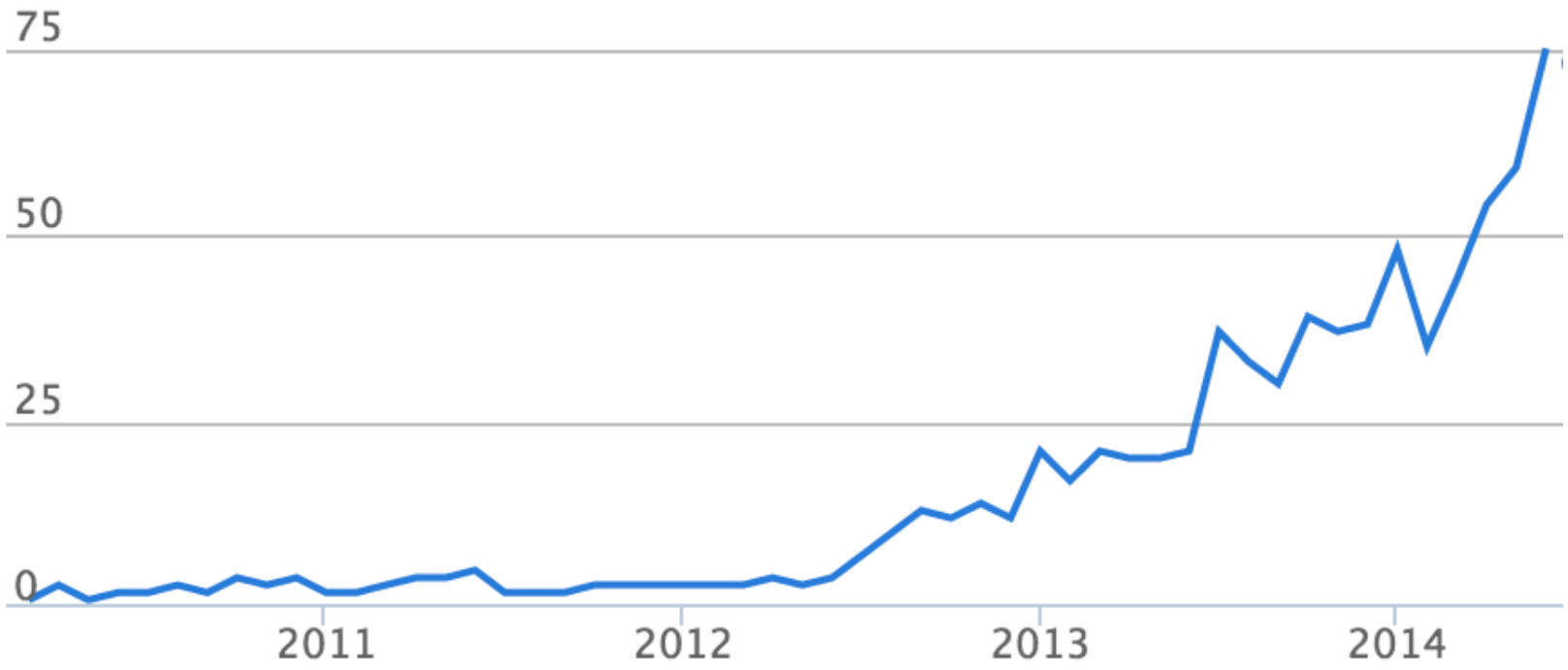
Project Activity





Activity in past 6 months

Continuing Growth



Contributors per month to Spark

source: ohloh.net

Conclusions

Spark and Research

Spark has all its roots in research, so we hope to keep incorporating new ideas!

Conclusion

Data flow engines are becoming an important platform for numerical algorithms

While early models like MapReduce were inefficient, new ones like Spark close this gap

More info: spark.apache.org