Econometrics at Scale: Spark Up Big Data in Economics

CENTER FOR FINANCIAL & BUSINESS ANALYTICS
Data Science Institute

Benjamin Bluhm
benjaminbluhm@gmx.de
(Data Science Freelancer)

Jannic Cutura
Janniccutura@gmail.com
(Goethe University)

Data Science Institute
COLUMBIA UNIVERSITY

Motivation

The amount of data available for research is ever increasing. New econometrics techniques borrowered from the machine learning literature gained considerable attention over the past years. The ability to handle and analyse datasets that are too large to fit in memory crucial to leverage 21 century opportunities. Yet there is as of now little guidance for economists on how to handle big data.

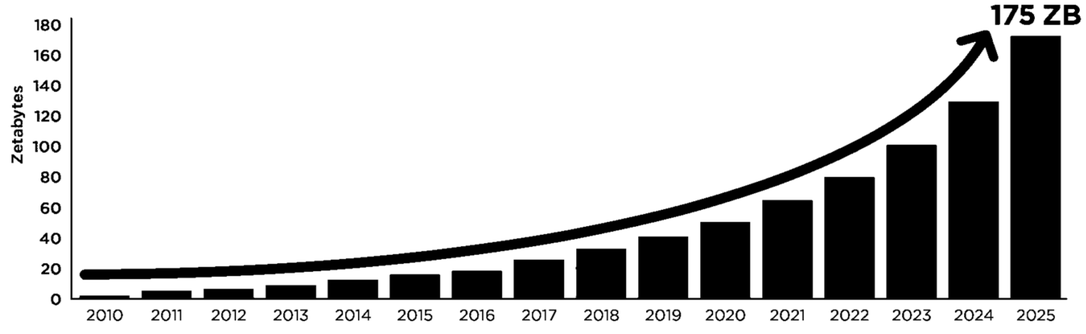

Figure 1. Anual Size of Global Datasphere

Spark Architecture

Spark solves memory problems by distributing the computation across clusters

- **Your computer: Take the data to the computation**
- **Spark: Bring the computation to the data**

This is based on efficient implementations of the map-reduce framework.

Food for thought: How do you compute the mean of large dataset?

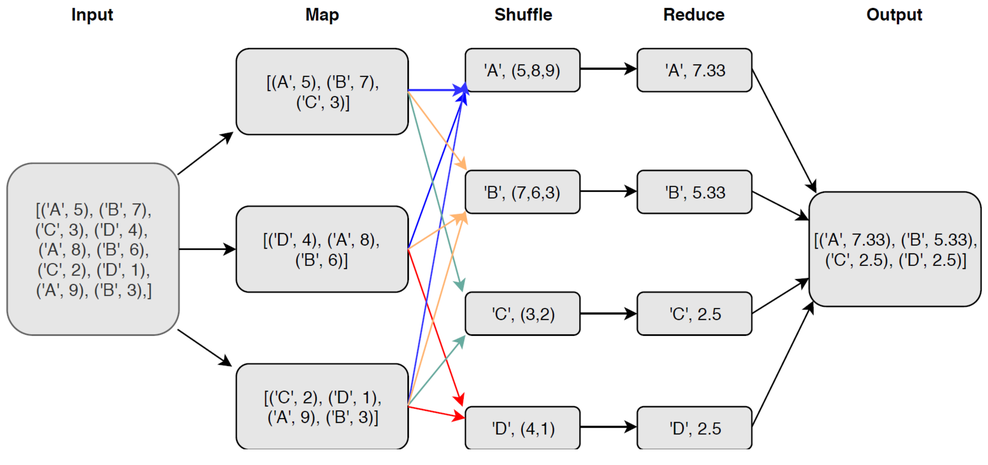

Figure 2. A simple map-reduce framework

Econometrics on Spark

Spark allows to use your existing data handling and analysis pipelines in a distributed fashion. Existing data handling pipelines in R/Python can seamlessly be transferred to a cloud based Spark system. We demonstrate ease of use for many standard empirical setups including:

- **Understanding data (parse, reshape, aggregate, filter, plot)**
- **Microeconometrics (OLS, Probit, Logit, Survival Analysis)**
- **Paneleconometrics (random and fixed effection regressions, error clustering)**
- **Time Series Econometrics (Monte Carlo Simulations, VARs)**

	OLS			Probit			Logit		
	Spark (local)	base R (local)	Spark (AWS)	Spark (local)	base R (local)	Spark (AWS)	Spark (local)	base R (local)	Spark (AWS)
Loan granted	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Income (000$)	0.0001***	0.0001***	0.0000061944***	0.0003***	0.0003***	0.0002***	0.0006***	0.0006***	0.0002***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Male	0.0250***	0.0250***	0.0298***	0.0639***	0.0639***	0.070***	0.0982***	0.0982***	0.07006***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Race									
White	0.01383***	0.01383***	0.00829***	0.0354***	0.0354***	0.0264***	0.0613***	0.0613***	0.02645***
	(0.0113)	(0.0113)	(0.0000)	(0.0109)	(0.0109)	(0.0000)	(0.0061)	(0.0061)	(0.0000)
Black	-0.1486***	-0.1486***	-0.1450***	-0.3823***	-0.3823***	-0.3629***	-0.6060***	-0.6060***	-0.3629***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
# Observations	147,329	147,329	137,819,151	147,329	147,329	137,819,151	147,329	147,329	137,819,151
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Loan Purpose FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Runtime (min)	**0.71**	**0.007**	**9.90**	**0.146**	**1.556**	**33.69**	**0.026**	**0.995**	**19.833**

Table 1. Some Microeconometrics on Spark

Conclusion

Cloud Computing is a useful tool for research in economics and social sciences more broadly. Cloud technologies enables usage of existing data handling pipelines to be executed on (i) arbitrary large datasets in (ii) a fraction of computing time. We provide intuitive explanations of Spark using econpseak alongside fully reproduciable minimal examples.

Acknowledgments

We would like to thank Sanjiv Das and Fraue Kreuter as well as members from the IMF Data@Fund group for helpful feedback and suggestions. We gratefully achknowledge a travel grant from the Bank of England

References

Einav, Liran and Jonathan Levin (2014). *Economics in the age of big data*. In: Science 346.6210
Foster, Ian et al. (2016). *Big data and social science: A practical guide to methods and tools* Chapman and Hall/CRC.