Making Big Data Analytics Interactive

Matei Zaharia, in collaboration with
Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Haoyuan Li, Justin Ma, Murphy McCauley, Joshua Rosen, Reynold Xin, Michael Franklin, Scott Shenker, Ion Stoica

AMP Lab, UC Berkeley

Motivation

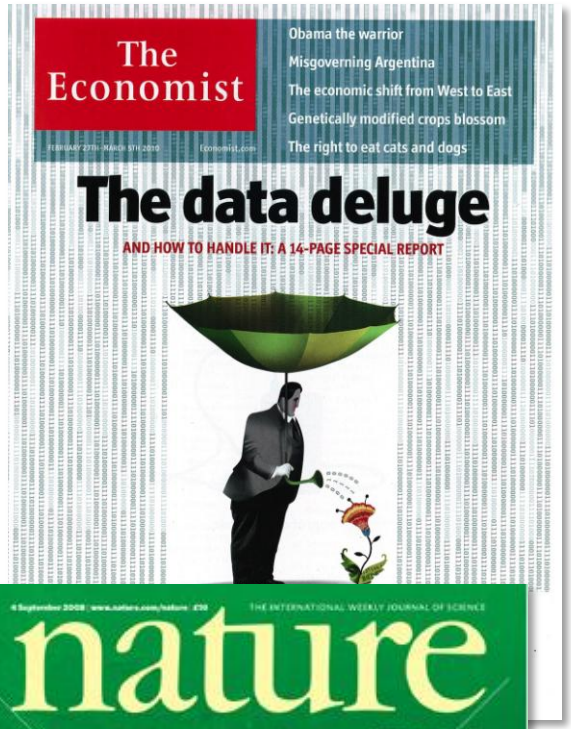
Data is growing faster than computation speeds

» Web apps, mobile, scientific, …

Requires large clusters to analyze

Programming clusters is *hard*

» Failures, placement, load balancing



Spark Programming Model

Manipulate distributed datasets as you would local collections

Concept: *resilient distributed datasets* (RDDs)

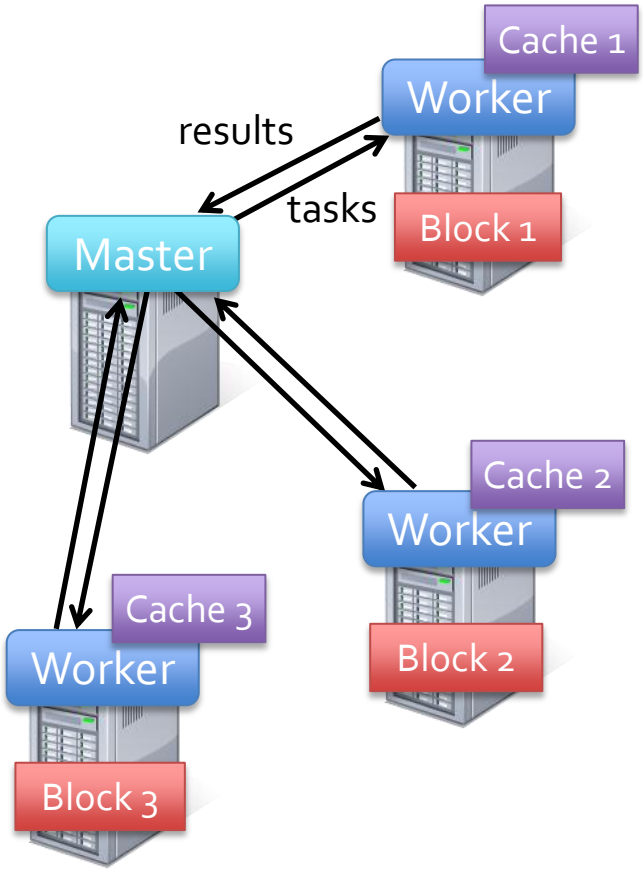
- » Distributed collections of objects that are automatically reconstructed on failure
- » Built through parallel *transformations* (map, filter, etc)
- » Can be *cached* in memory for fast reuse

Example: Log Mining

Load error messages from a log into memory, then interactively search for various patterns

```
lines = spark.textFile("hdfs://...")
errors = lines.filter(_.startsWith("ERROR"))
messages = errors.map(_.split('\t')(2))
messages.cache()

messages.filter(_.contains("Linux")).count()
messages.filter(_.contains("PHP")).count()
```



Result: scaled to **1 TB** data in **5 sec**
(vs 170 s for on-disk data)

Projects Built on Spark

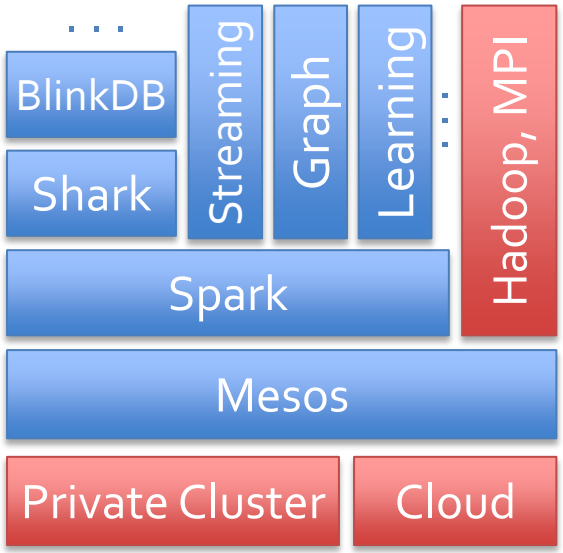
Spark available in Java, Scala, Python

Growing open source community

- » 500-member meetup
- » 14 companies contributing

Using it to develop a stack of fast analytics tools

- » Shark SQL, Spark Streaming, …



Berkeley Data Analytics Stack (BDAS)

Find Out More

Visit us at the **AMP Lab:** 4th floor Soda Hall

www.spark-project.org