Spark on Mesos

Who am I?

- Born in Rochester
- RIT Software Engineering alum
- Brand Networks for 3+ years
- http://geowa4.software

Topics

- Overview
- Use Cases
- Architecture
- Mesos
- Demo

Overview

> ...fast and general-purpose cluster computing system.
>
> — *http://spark.apache.org/docs/latest/index.html*

Language Support

- Scala
- Java
- Python
- R

Tools

- Spark SQL
- MLib
- GraphX
- Spark Streaming
- Spark Shell

Clustering Options

- Apache Mesos
- Standalone
    - Amazon EC2
- Hadoop YARN

Data Sources

- JDBC
- Cassandra
- Elasticsearch
- HDFS / S3

Streaming Sources

- Kafka
- AWS Kinesis
- Twitter
- HDFS / S3

Outputs

Pretty much the same as the inputs.

ETL

> Spark is my new go-to ETL tool.
>
> – *Brian Clapper*

Data Consumption

Pull data from external sources into local data store.

Custom Receivers

Machine Learning

- With static data set
- Iteratively with streaming

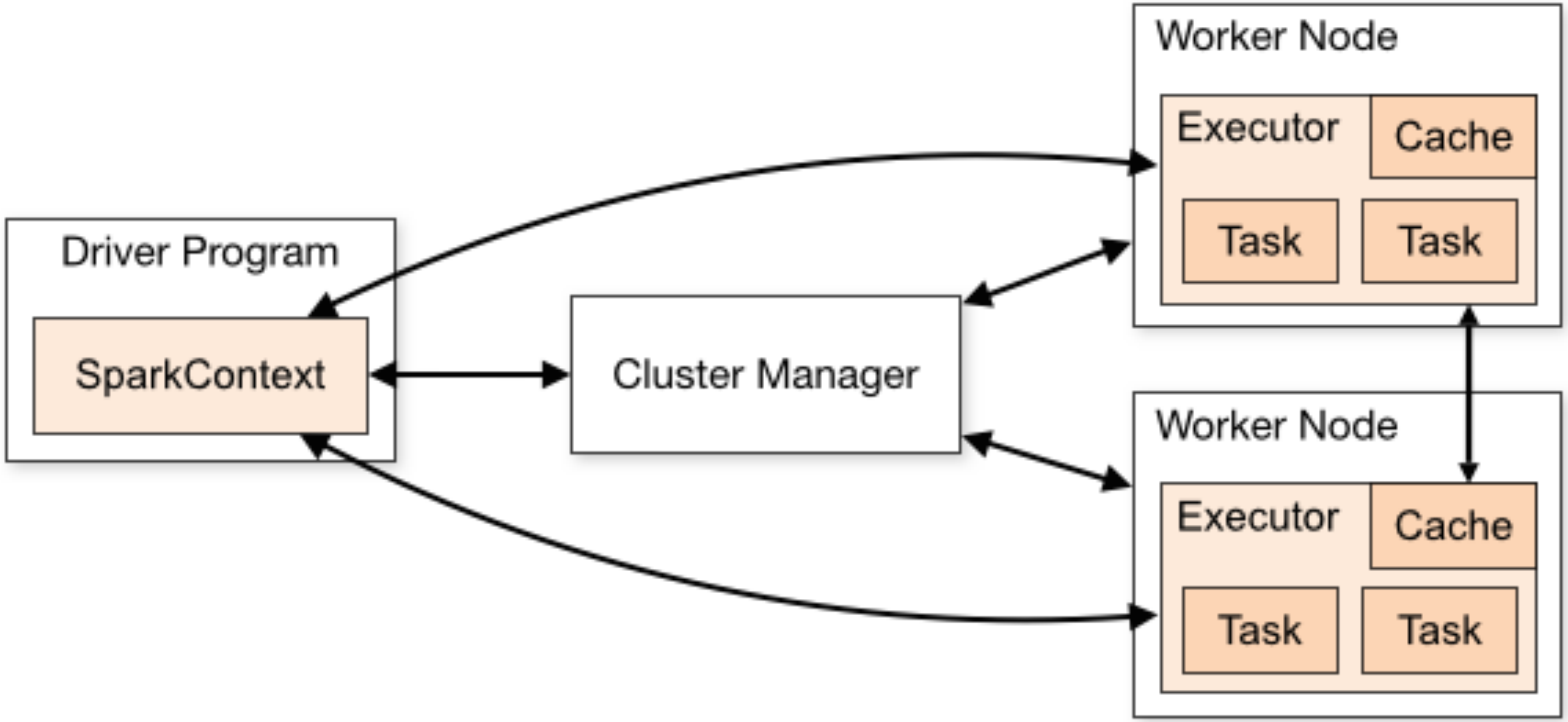
Scalable Code

Find the most common word in the lines containing "Spark".

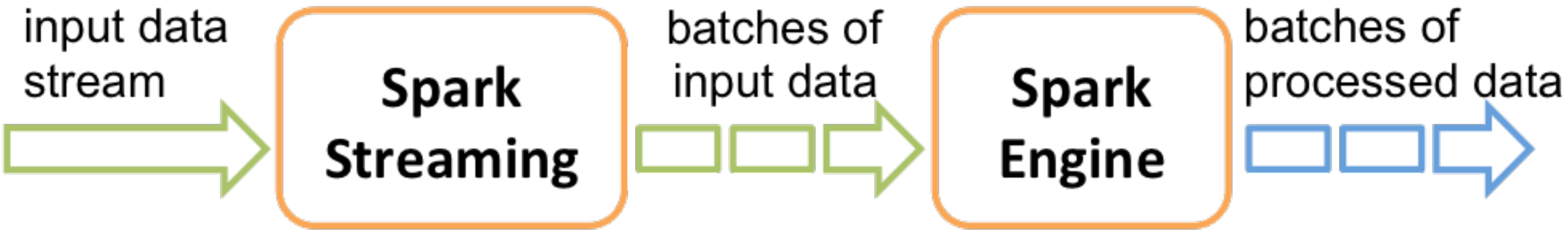
```
lines
    .filter(l => l.contains("Spark"))
    .flatMap(l => l.split(" "))
    .map(word => (word, 1))
    .reduceByKey((a, b) => a + b)
    .map(pair => pair.swap)
    .sortByKey(false)

(12, foo)
(3, bar)
(1, baz)
```

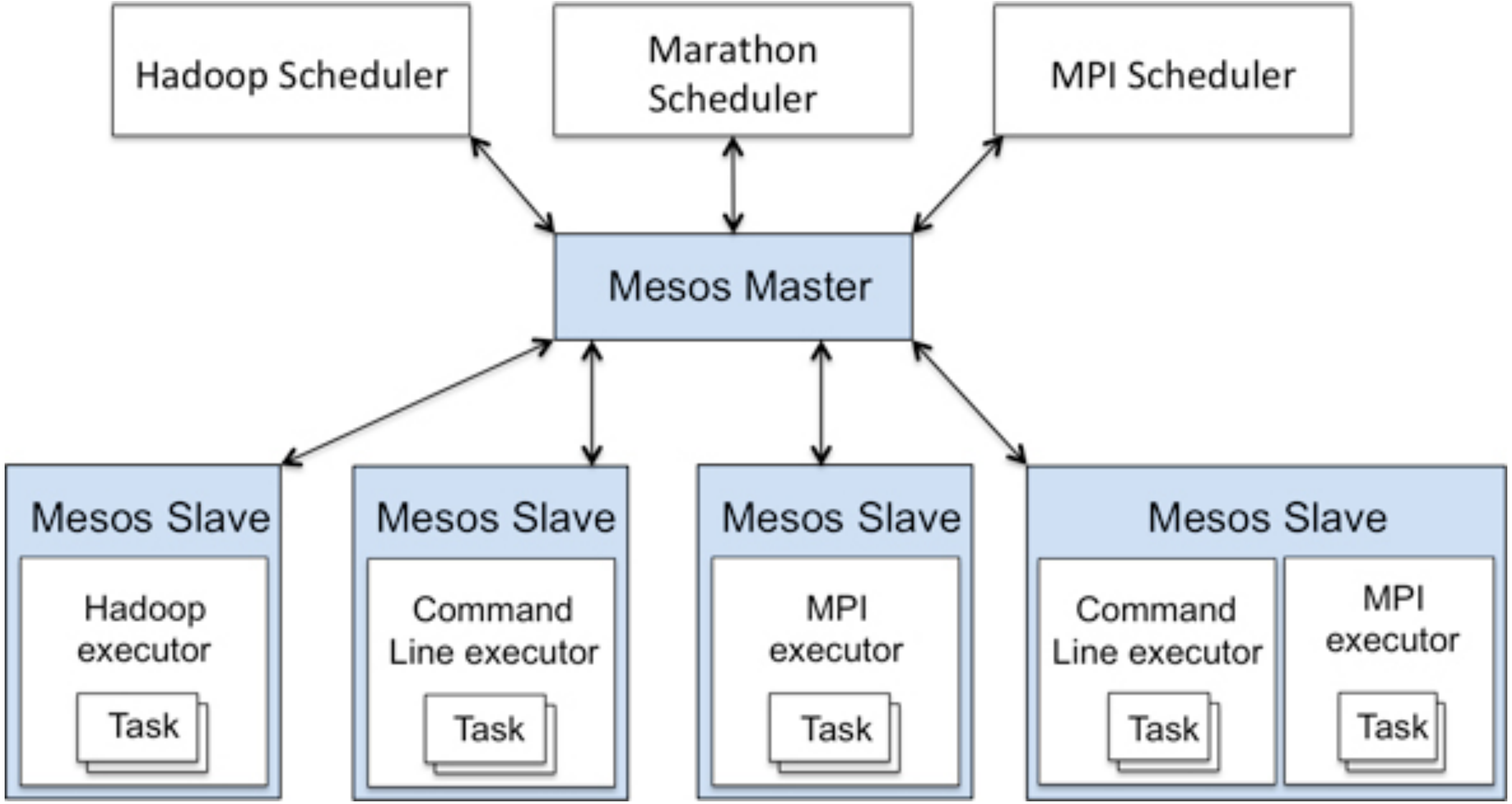
Architecture



Streaming Architecture



Mesos Achitecture



Let's run it