Spark

A Fast and General Engine
for Large-scale Data Processing

Shangyu Luo
Department of Computer Science
Rice University

Cluster Computing System

- Scalability
- Speed
- Usability
- Fault Tolerance


amplab
YAHOO!
ADATAO
BIG DATA INTELLIGENCE
Spark
tuplejump
databricks™
bizo

Outline

- What is Spark
- Why is Spark
- Experiments
- Discussion
- Conclusion and Future Work

What is Spark

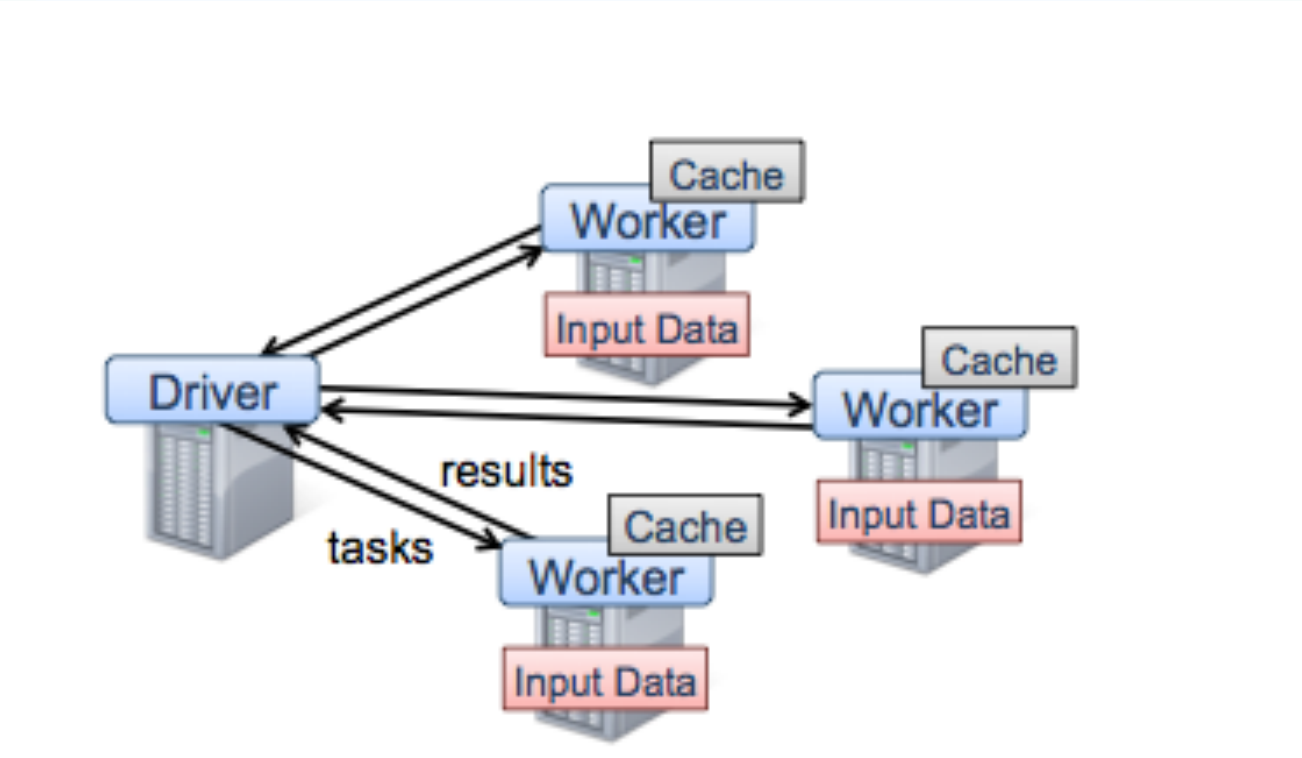
- An open source cluster computing system
- Fast data analysis — fast to run and fast to write.

  (Map-reduce like system)
- Up to 100X faster than Hadoop MapReduce
- Support Scala, Java and Python

Why is Spark

- Iterative algorithms (e.g., machine learning) and interactive applications
- Fast
  - Storing data into memory and then reuse it (cache)
  - E.g., Logistic Regression, 1 billion 9 dimension data points : 3s (normally 70s)
- Easy to use
- Fault tolerance
  - Lineage information

Spark Cluster Structure



Driver Node→ Partitioning → Tasking → Worker Nodes

Example

```
val file = spark.textFile("hdfs://...")

val counts= file.flatMap(line => line.split(" "))
                .map(word => (word, 1))
                .reduceByKey(_ + _)

counts.saveAsTextFile("hdfs://...")
```

Experiments

- Models
  - Gaussian mixture model (GMM)
  - Gaussian mixture model with imputation
  - The Bayesian Lasso
  - Hidden Markov model (HMM) for text
  - Latent Dirichlet allocation (LDA)
- Experimental Platform
  - Amazon EC2 m2.4xlarge machines (8 cores, 68 GB of RAM)
- Evaluated Systems
  - Spark, SimSQL, GraphLab and Giraph

Gaussian Mixture Model

	10 dimensions				100 dimensions
	Lines of code	5 machines	20 machines	100 machines	5 machines
Spark (python)	236	26:04	37:34	38:09	47:40
Spark (java)	737	12:30	12:25	18:11	6:25:04
SimSQL	197	27:55	28:55	35:54	1:51:12

Figure 1: GMM; Lines of code and average time per iteration. Format is HH:MM:SS or MM:SS. The size of the dataset is ten million data points per machine.

Bayesian Lasso

	Lines of code	5 machines	20 machines	100 machines
Spark (Python)	168	0:55	0:59	1:12
SimSQL	100	7:09	8:04	12:24

Figure 2: Bayesian; Lines of code and average time per iteration. Format is HH:MM:SS or MM:SS. The model has 1000 regressor dimensions and a one-dimensional response. The size of the dataset is 100,000 data points per machine.

Hidden Markov Model

	Word-based, 5 machines		Document-based, 5 machines	
	Lines of code	Running time	Lines of code	Running time
Spark	NA	Fail	214	4:21:36
SimSQL	131	8:17:07	123	3:42:40

Super Vertex Implementation				
	Lines of code	5 machines	20 machines	100 machines
Spark	215	3:45:58	4:01:02	Fail
SimSQL	136	2:05:12	2:05:31	2:19:10

Figure 3: HMM; Lines of code and average time per iteration. Format is HH:MM:SS or MM:SS. The size of the dictionary is 10,000 words. The number of topics is 20. The size of the dataset is 2.5 million documents per machine.

Discussion

- Fast: GMM, Bayesian Lasso
  - Reuse the same data in the cache in each iteration
- Slow: GMM with imputation, LDA, HMM for text
  - Data is changing in each iteration
  - Shuffling work is time-consuming
- Scalability: Bad for LDA and HMM, good for others.
- Ease-of-programming: Good. Short programs in most of experiments.

Discussion

- Python vs. Java
  - Python: Better support for mathematical calculation; Especially vector/matrix manipulation.
  - Java: Generally faster; More choices of operations on RDD.
- Spark vs. SimSQL
  - No overall winner in speed.
  - Spark: Easier for programming.
  - SimSQL: Better scalability. Can run almost all five experiments without problems.

Conclusion & Future Work

- Spark is a fast cluster computing platform, especially suitable for iterative and interactive applications.
- Still under development
  - Extensions (SQL, stream, machine learning libraries, graph, etc.)
  - Not so stable (May have unexpected problems)
- Contributors and users are increasing rapidly.
- SimSQL has the potential to improve its performance and usability.