DLI Accelerated Data Science Teaching Kit

Lecture 11.3 - Spark SQL and Other Spark Libraries

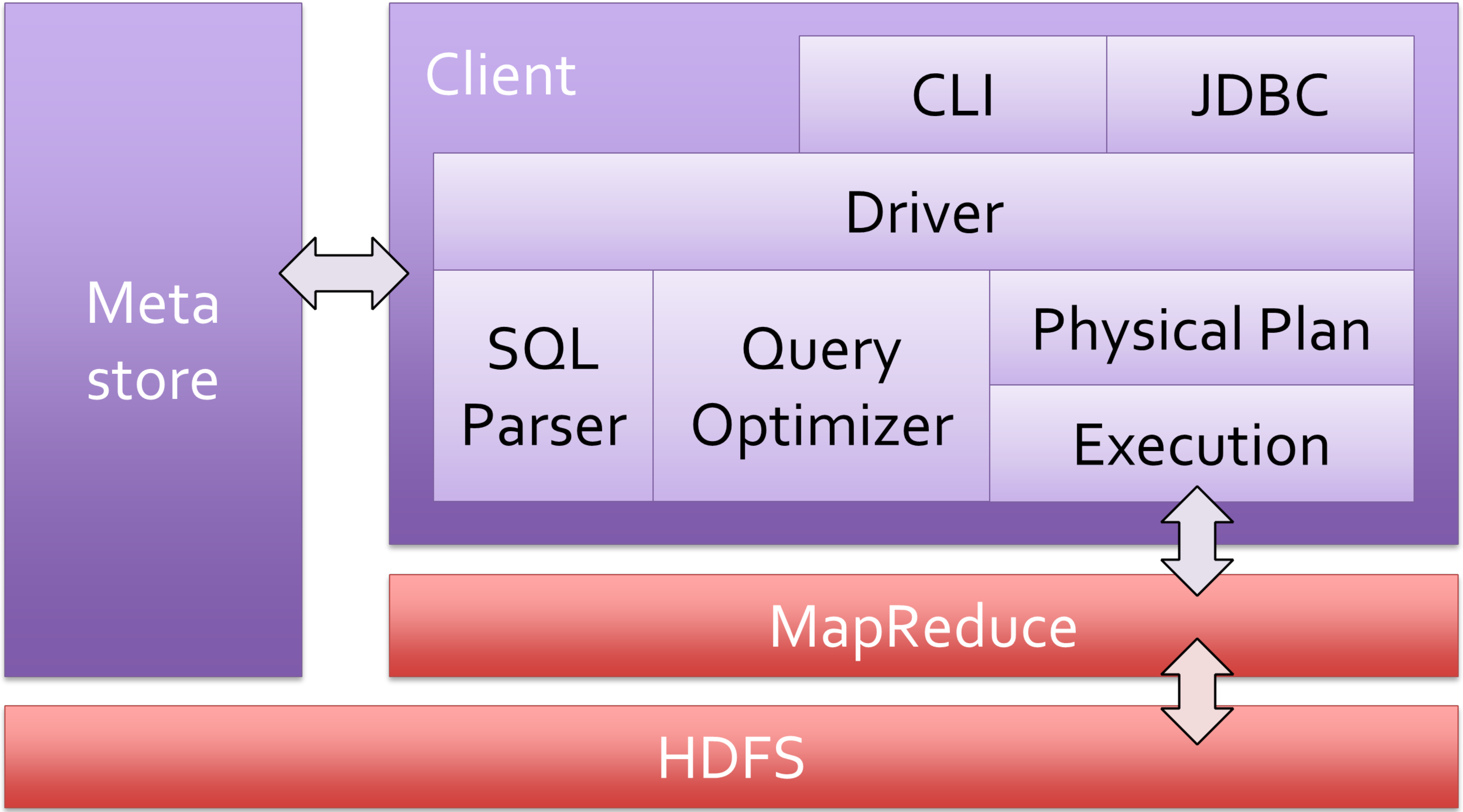
Motivation

Hive is great, but Hadoop’s disk-based engine can make even the smallest queries take minutes

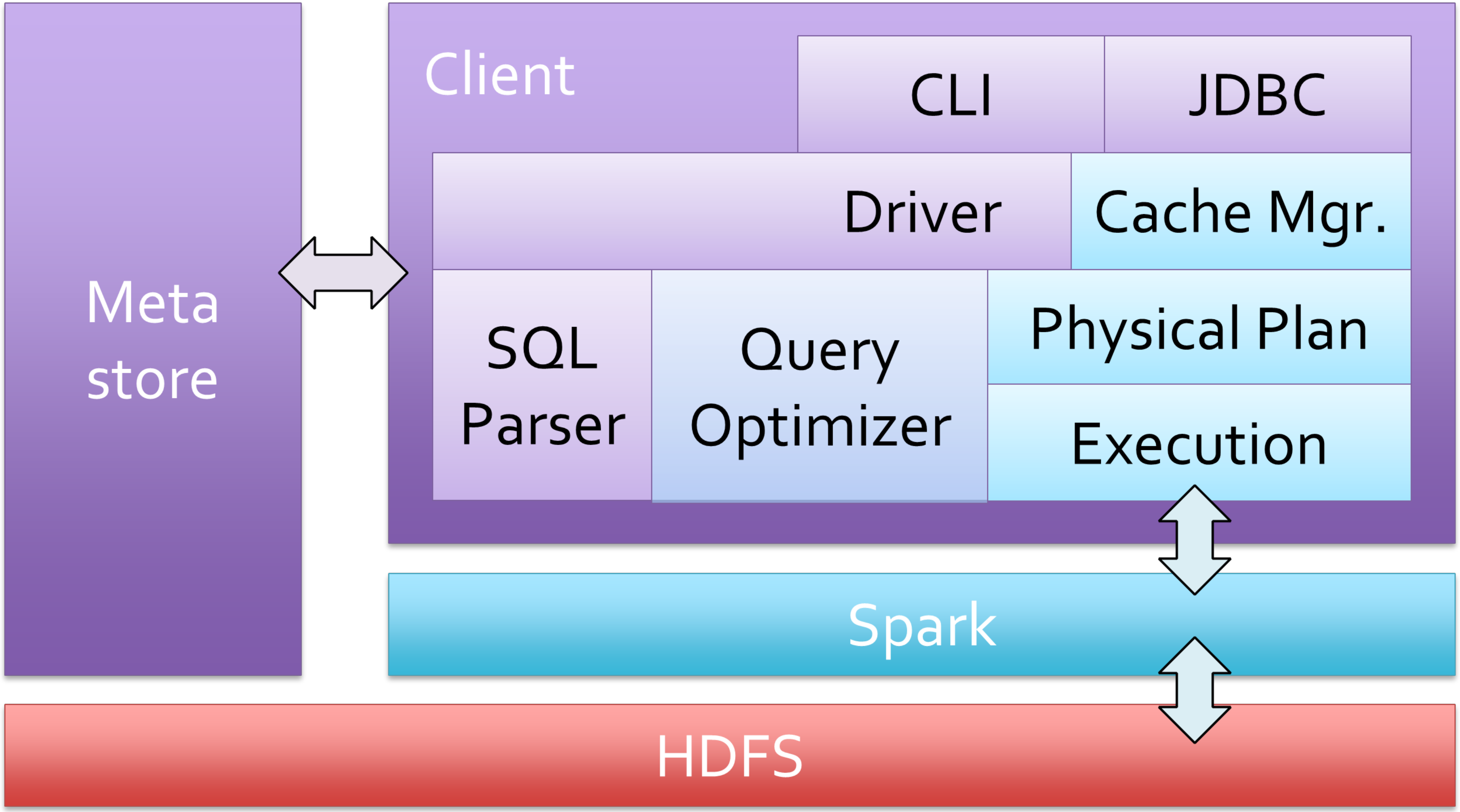
Can we extend Hive to run on Spark?

Yes! Spark SQL = Hive on Spark

Hive Architecture



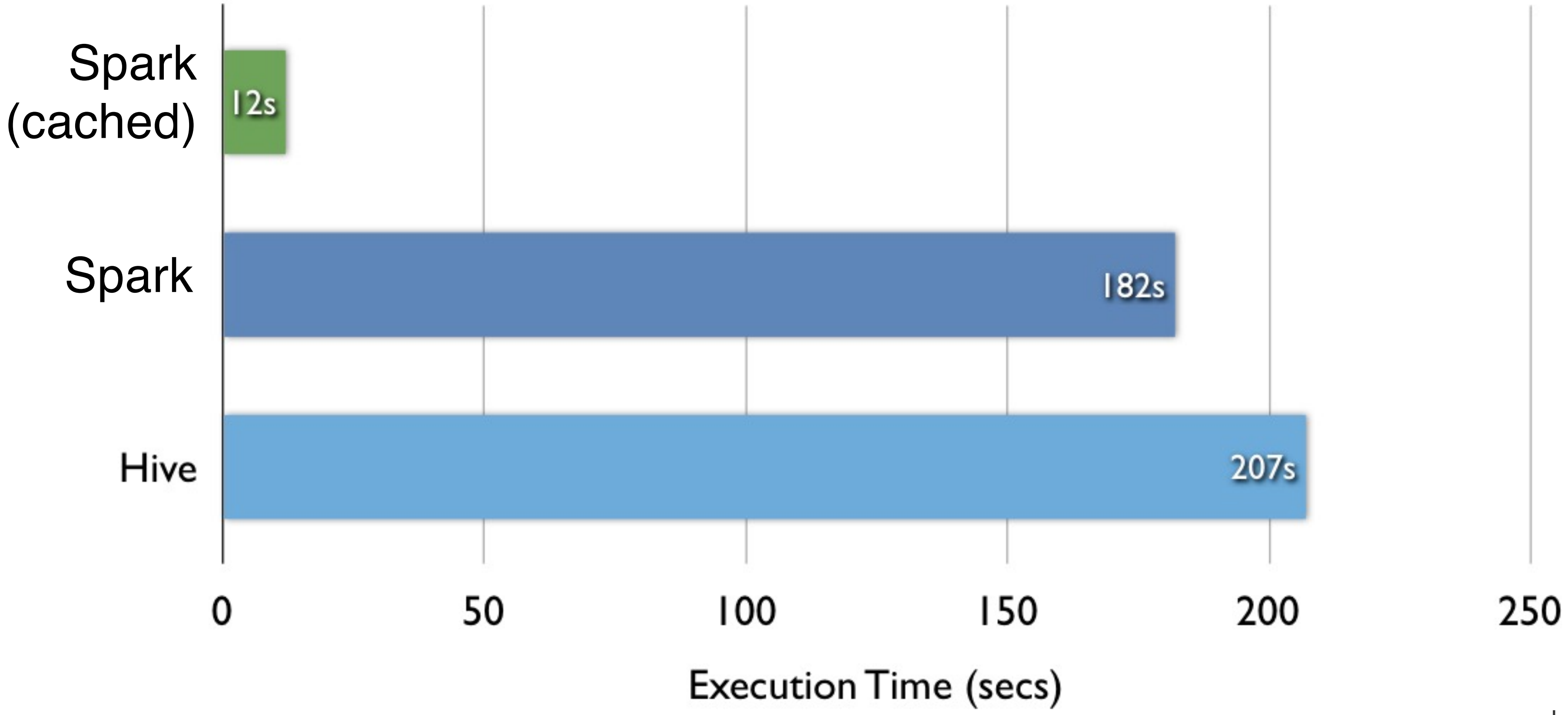
Spark SQL Architecture



[Engle et al, SIGMOD 2012]

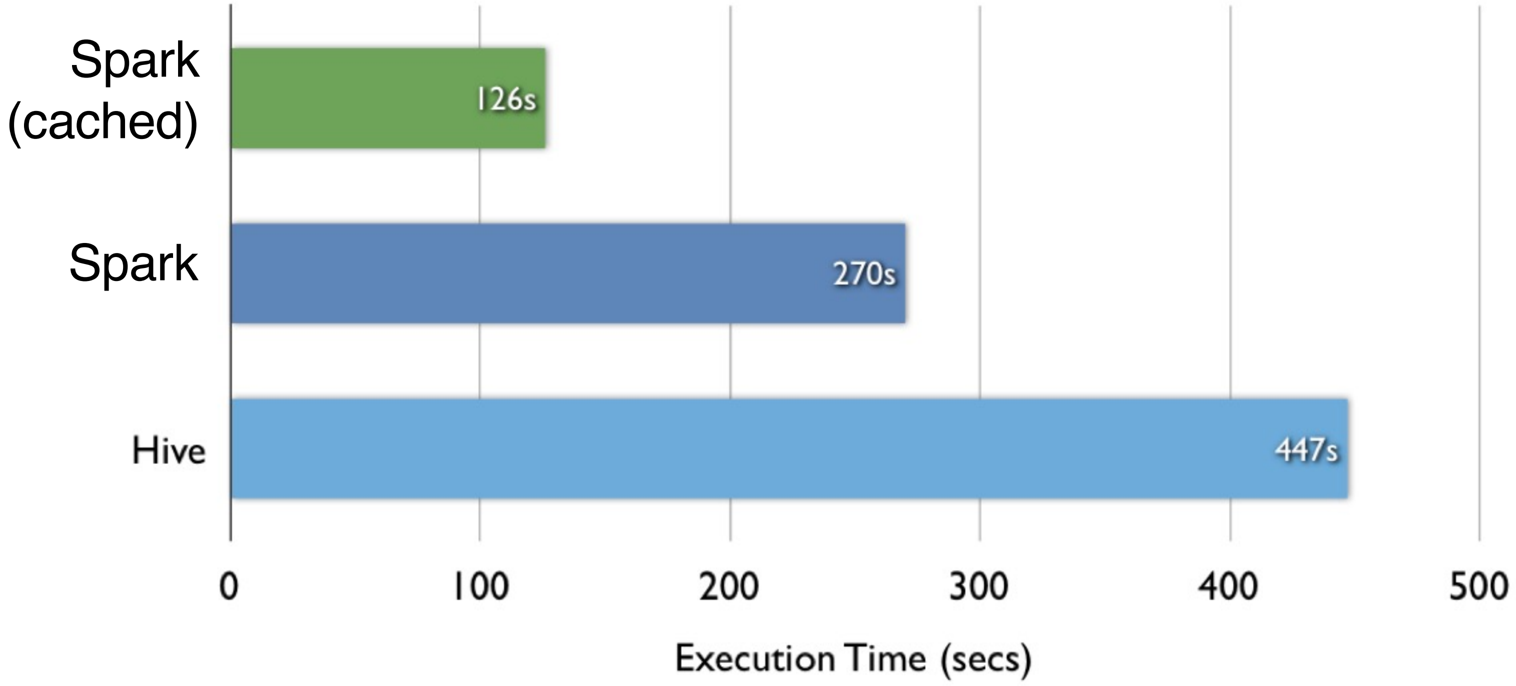
Benchmark Query 1

SELECT * FROM grep WHERE field LIKE ‘%XYZ%’;



Benchmark Query 2

```
SELECT sourceIP, AVG(pageRank), SUM(adRevenue) AS earnings
FROM rankings AS R, userVisits AS V ON R.pageURL =V.destURL
WHERE V.visitDate BETWEEN ‘1999-01-01’ AND ‘2000-01-01’
GROUP BY V.sourceIP
ORDER BY earnings DESC
LIMIT 1;
```



Spark Streaming

Recall that Spark’s model was motivated by two emerging uses (interactive and multi-stage apps)

Another emerging use case that needs fast data sharing is **stream processing**

- » Track and update state in memory as events arrive
- » Large-scale reporting, click analysis, spam filtering, etc.

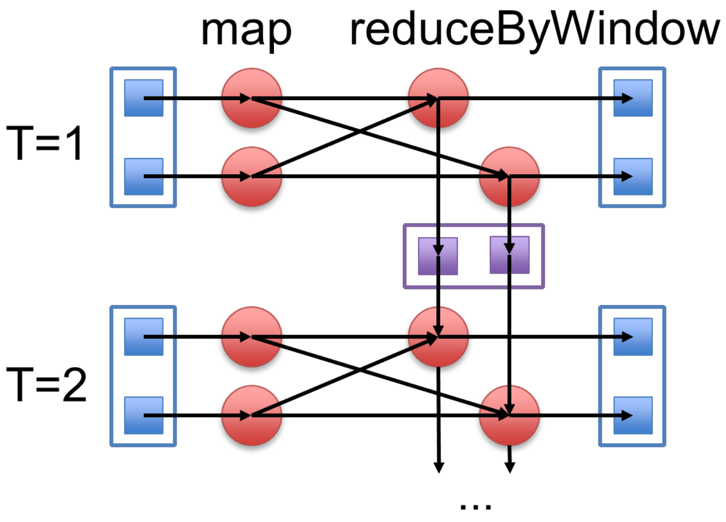
Spark Streaming

Extends Spark to perform streaming computations

Runs as a series of small (~1 s) batch jobs, keeping state in memory as fault-tolerant RDDs

Intermix seamlessly with batch and ad-hoc queries

```
tweetStream
 .flatMap(_.toLower.split)
 .map(word => (word, 1))
 .reduceByWindow(5, _ + _)
```



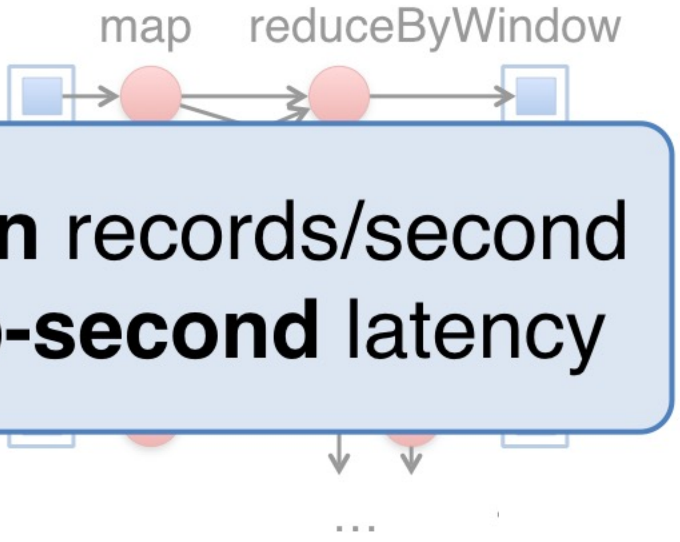
[Zaharia et al, HotCloud 2012]

Spark Streaming

Extends Spark to perform streaming computations

Runs as a series of small (~1 s) batch jobs, keeping state in memory as fault-tolerant RDDs

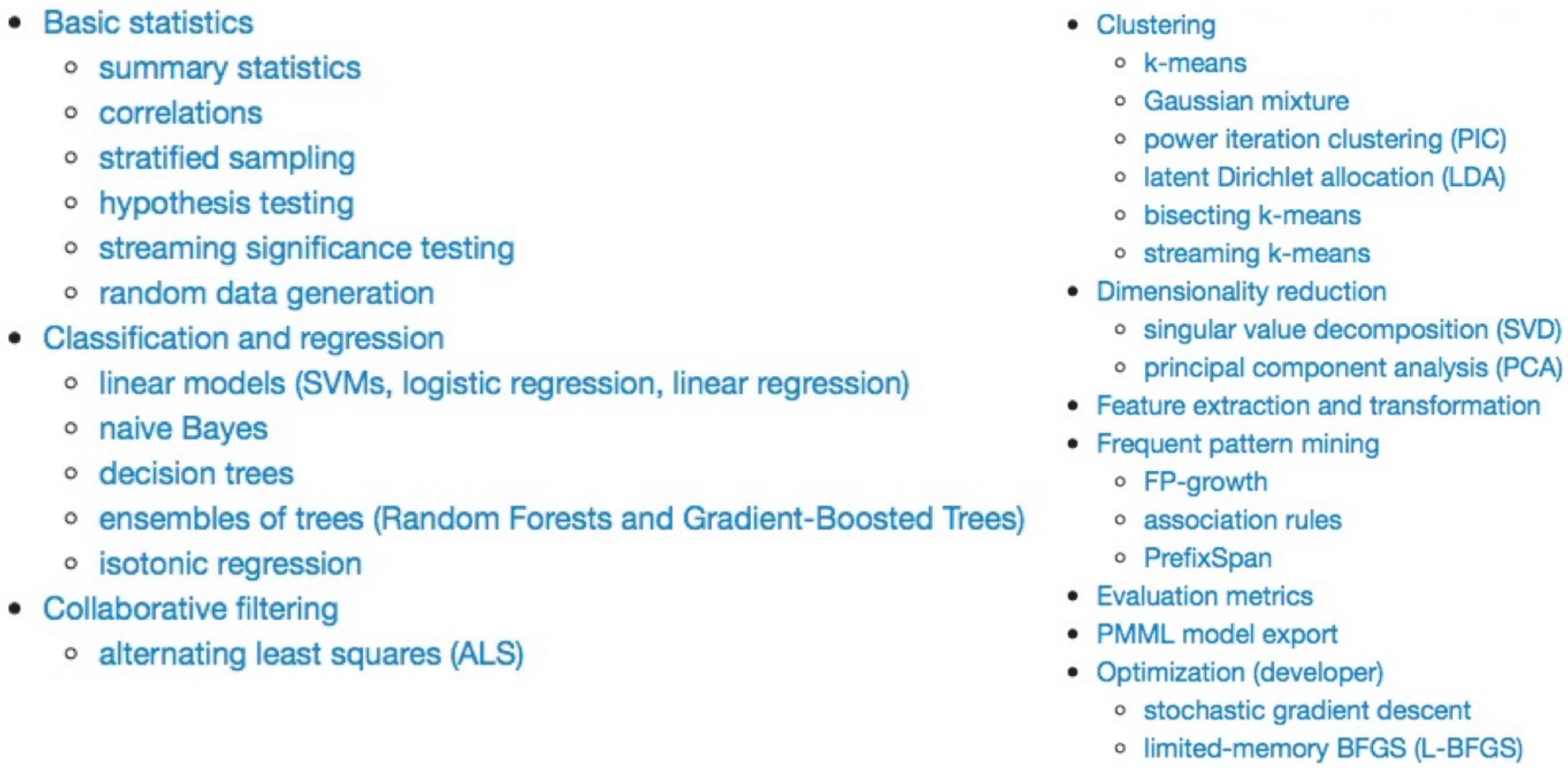
Intermix seamlessly with batch and ad-hoc queries



Result: can process **42 million** records/second (4 GB/s) on 100 nodes at **sub-second** latency

[Zaharia et al, HotCloud 2012]

MLlib



- Basic statistics
  - summary statistics
  - correlations
  - stratified sampling
  - hypothesis testing
  - streaming significance testing
  - random data generation
- Classification and regression
  - linear models (SVMs, logistic regression, linear regression)
  - naive Bayes
  - decision trees
  - ensembles of trees (Random Forests and Gradient-Boosted Trees)
  - isotonic regression
- Collaborative filtering
  - alternating least squares (ALS)
- Clustering
  - k-means
  - Gaussian mixture
  - power iteration clustering (PIC)
  - latent Dirichlet allocation (LDA)
  - bisecting k-means
  - streaming k-means
- Dimensionality reduction
  - singular value decomposition (SVD)
  - principal component analysis (PCA)
- Feature extraction and transformation
- Frequent pattern mining
  - FP-growth
  - association rules
  - PrefixSpan
- Evaluation metrics
- PMML model export
- Optimization (developer)
  - stochastic gradient descent
  - limited-memory BFGS (L-BFGS)

https://spark.apache.org/docs/latest/ml-guide.html

GraphX

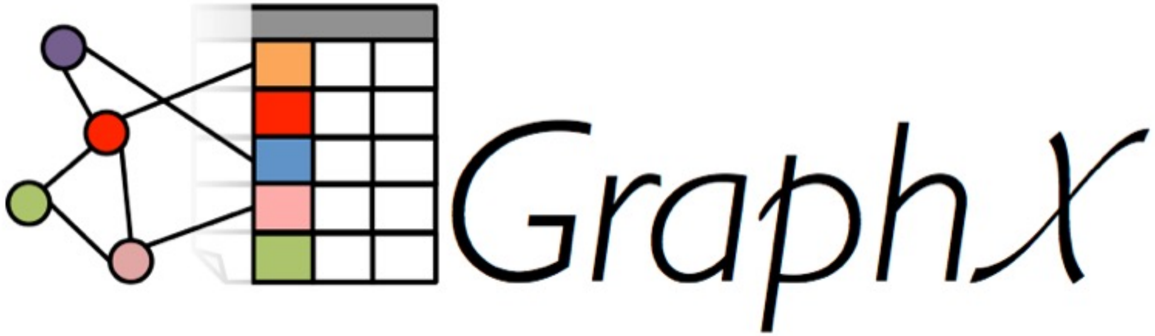
Parallel graph processing

Extends RDD -> Resilient Distributed Property Graph

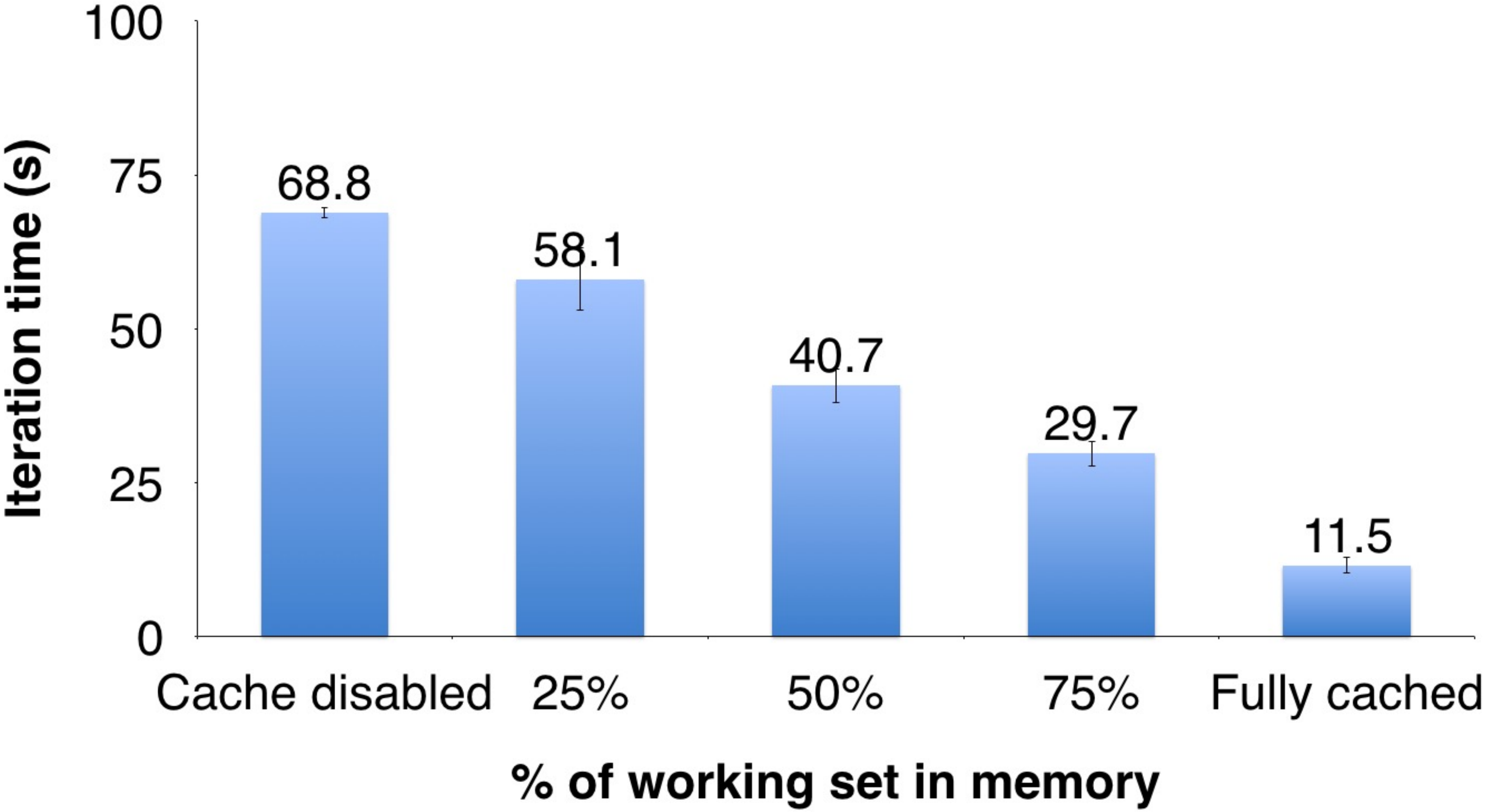
- Directed multigraph with properties attached to each vertex and edge

Limited algorithms

- PageRank
- Connected Components
- Triangle Counts



Behavior with Not Enough RAM



DLI Accelerated Data Science Teaching Kit

Thank You

We thank Dr. Matei Zaharia for sharing teaching materials for Spark.