Optimizations and the Future

Justin Pihony

@JustinPihony | justin-pihony.blogspot.com

Course Overview

- Basics of Spark
- Core API
- Cluster Managers
- Spark Maintenance
- Libraries
  - SQL
  - Streaming
  - MLlib/GraphX
- Troubleshooting / Optimization
- Future of Spark

Course Overview

Section Overview

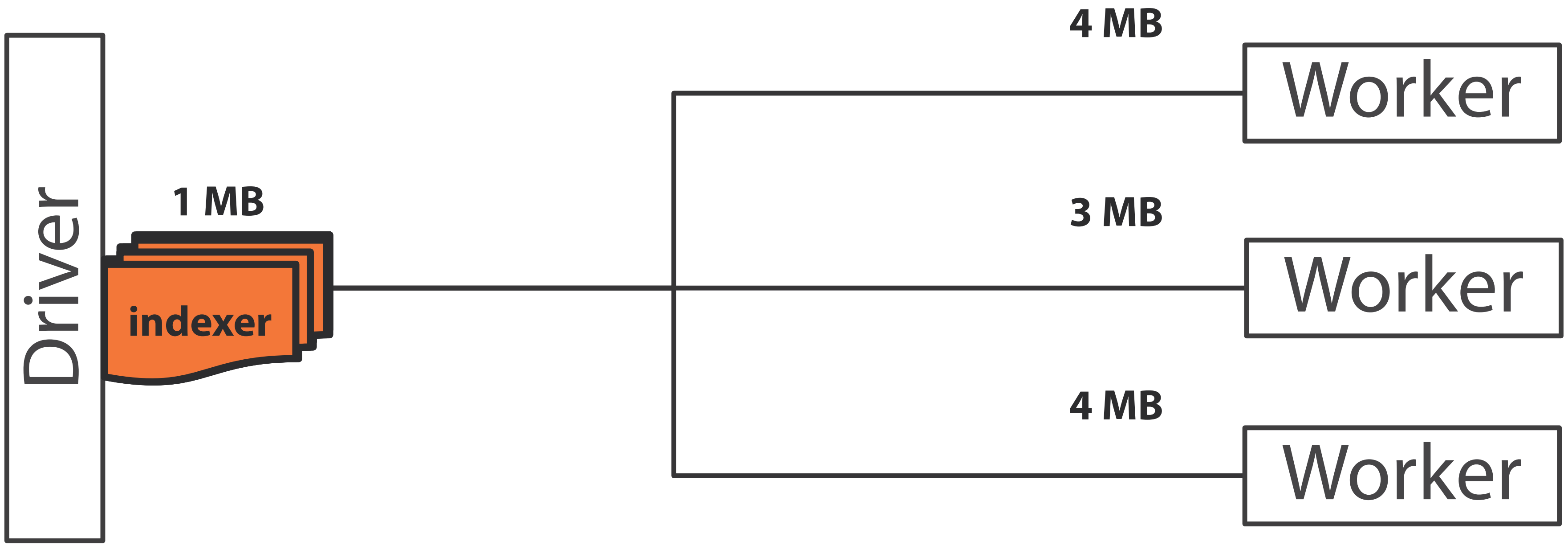
- Basics of Spark
- Core API
- Cluster Managers
- Spark Maintenance

- Libraries
  - SQL
  - Streaming
  - MLlib/GraphX
- Troubleshooting / Optimization
- Future of Spark

- **Troubleshooting / Optimization**
  - Closures
  - Broadcast
  - Partitioning
- **Future of Spark**

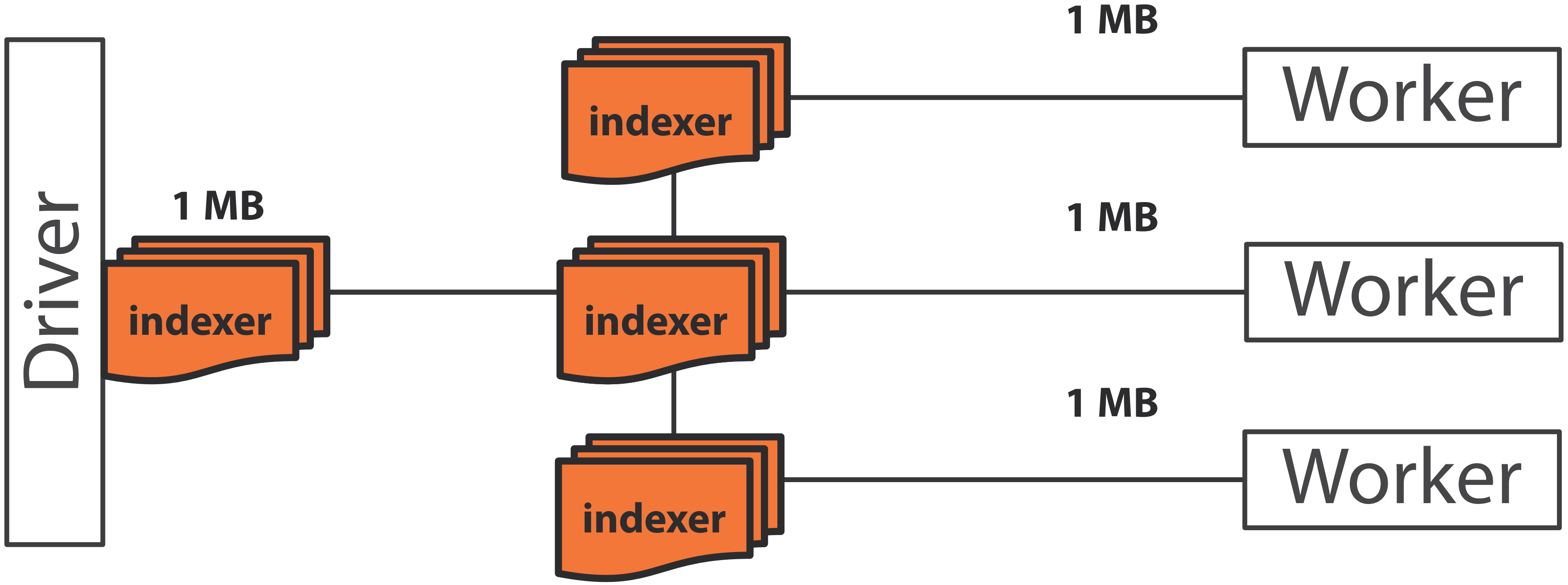
Broadcasting

```
val indexer = Map(…)//1MB

rdd.flatMap(rddVal => indexer.get(rddVal))
```



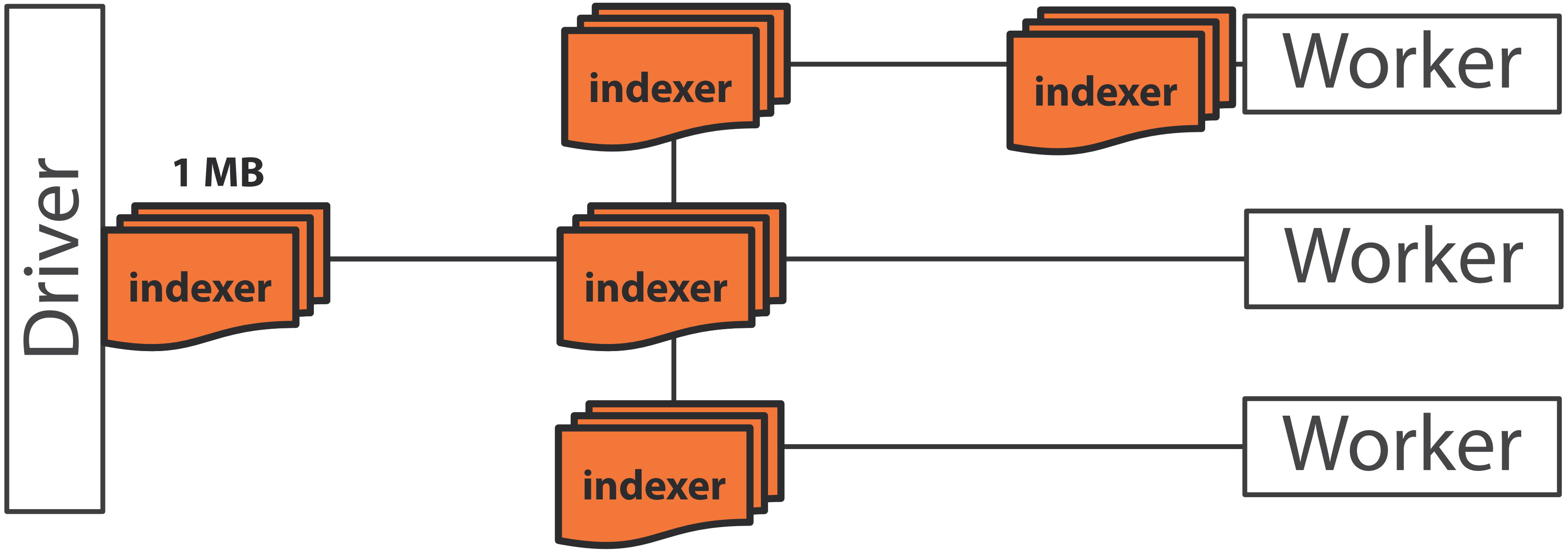
Broadcasting

```
val indexer = sc.broadcast(Map(…))//Map=1MB;indexer<1MB

rdd.flatMap(rddVal => indexer.value.get(rddVal))
```



Broadcasting

```
val indexer = sc.broadcast(Map(…))//Map=1MB;indexer<1MB

rdd.flatMap(rddVal => indexer.value.get(rddVal))
```



Future of Spark

> "zero-management cloud platform"
>
> \- Databricks FAQ

Future of Spark

R

TACHYON

Project Tungsten

Spark Packages

spark-jobserver

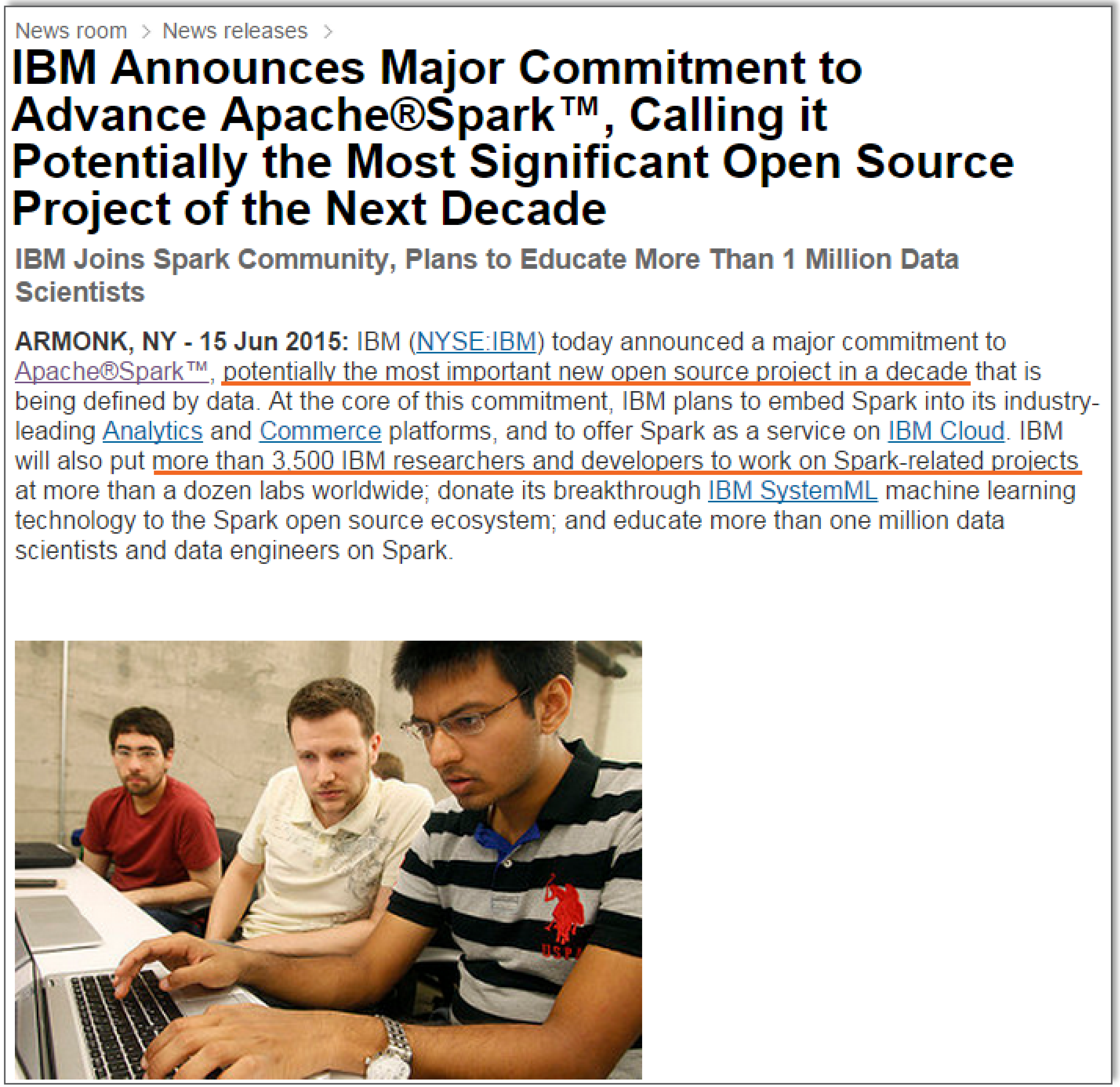
zeppelin

Data Sources

Streaming

Machine Learning

...



News room > News releases >

IBM Announces Major Commitment to Advance Apache®Spark™, Calling it Potentially the Most Significant Open Source Project of the Next Decade

IBM Joins Spark Community, Plans to Educate More Than 1 Million Data Scientists

ARMONK, NY - 15 Jun 2015: IBM (NYSE:IBM) today announced a major commitment to Apache®Spark™, potentially the most important new open source project in a decade that is being defined by data. At the core of this commitment, IBM plans to embed Spark into its industry-leading Analytics and Commerce platforms, and to offer Spark as a service on IBM Cloud. IBM will also put more than 3,500 IBM researchers and developers to work on Spark-related projects at more than a dozen labs worldwide; donate its breakthrough IBM SystemML machine learning technology to the Spark open source ecosystem; and educate more than one million data scientists and data engineers on Spark.

https://www-03.ibm.com/press/us/en/pressrelease/47107.wss

Resources

- Patrick Wendell – Spark Performance (2013)
  - https://www.youtube.com/watch?v=NXp3oJHNM7E
- Project Tungsten
  - https://databricks.com/blog/2015/04/28/project-tungsten-bringing-spark-closer-to-bare-metal.html
- DataStax – Common Spark Troubleshooting
  - http://www.datastax.com/dev/blog/common-spark-troubleshooting
- https://issues.apache.org/jira/browse/spark/
  - Labels = starter

Summary

- Beyond the basics
- What’s next?