

Spark

Overview



- Started as a research project at UC Berkeley in 2009
- Open Source License (Apache 2.0)
- Latest Stable Release: v2.0 (June 2016)
- 600,000 lines of code (75% Scala)
- Built by 800+ developers from 200+ companies

Spark

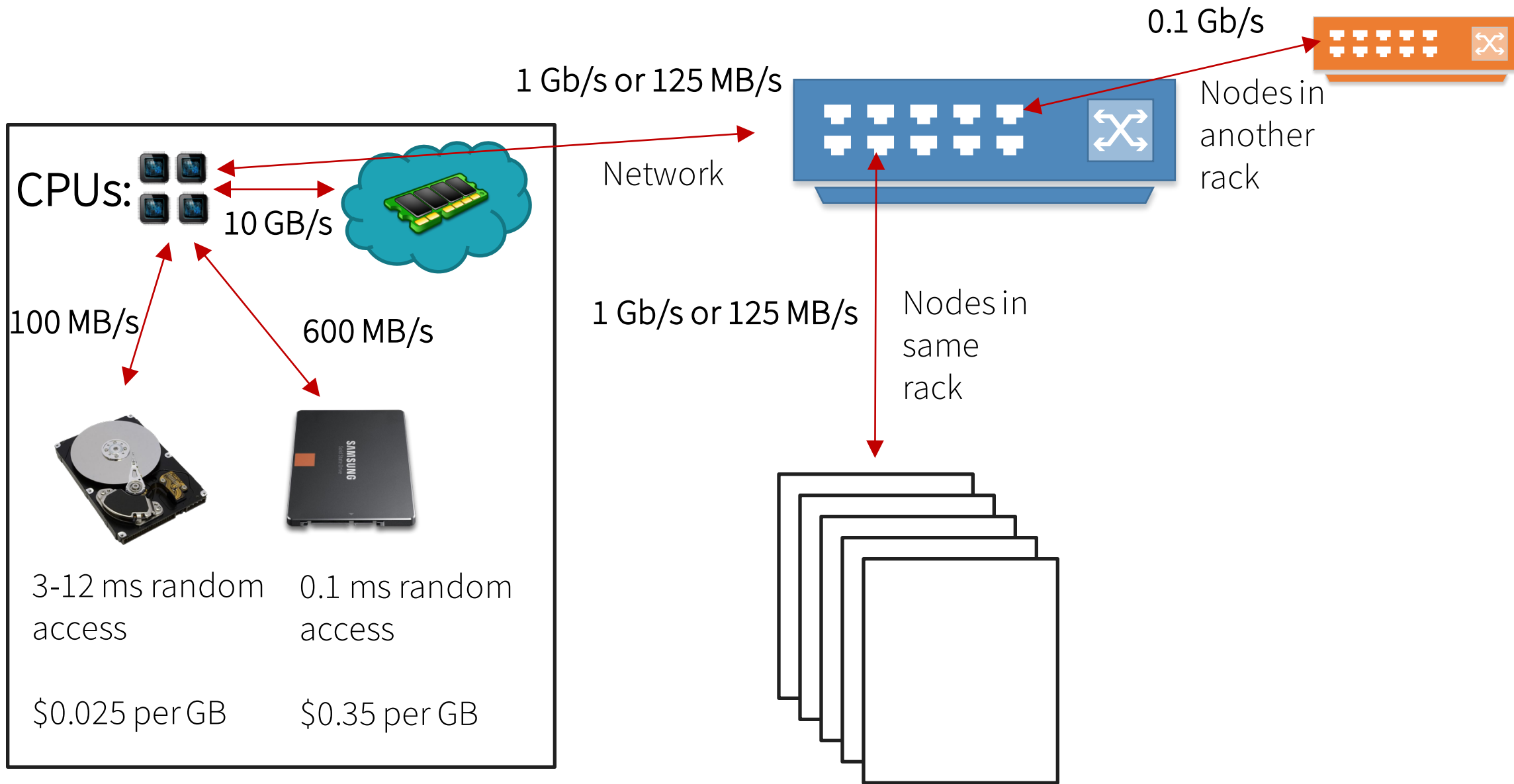
- 相比 Hadoop，更现代，更灵活
- 支持更通用的图形执行模型，允许 MapReduce 迭代以及更有效的数据重用
- 交互式操作界面
- 基于内存，比基于磁盘的 Hadoop 快得多
- 可以在 YARN 和 Mesos 上运行，也可以在笔记本电脑和 Docker 容器中运行

Spark

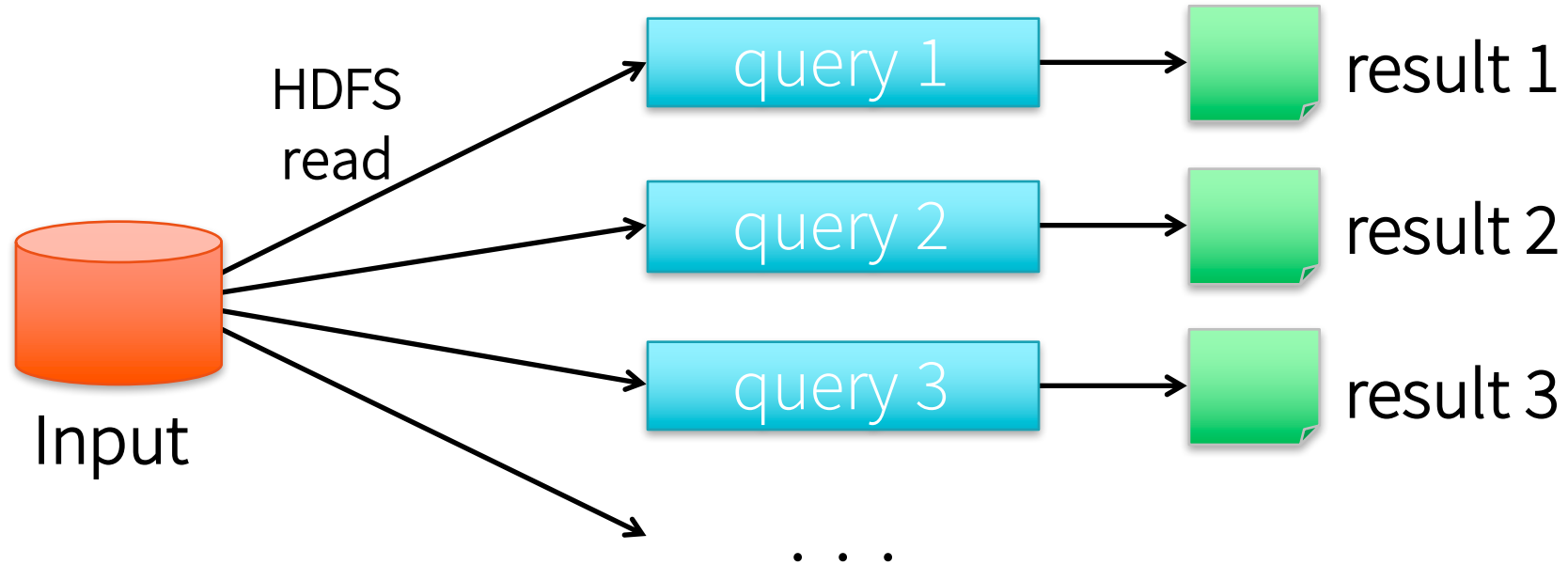
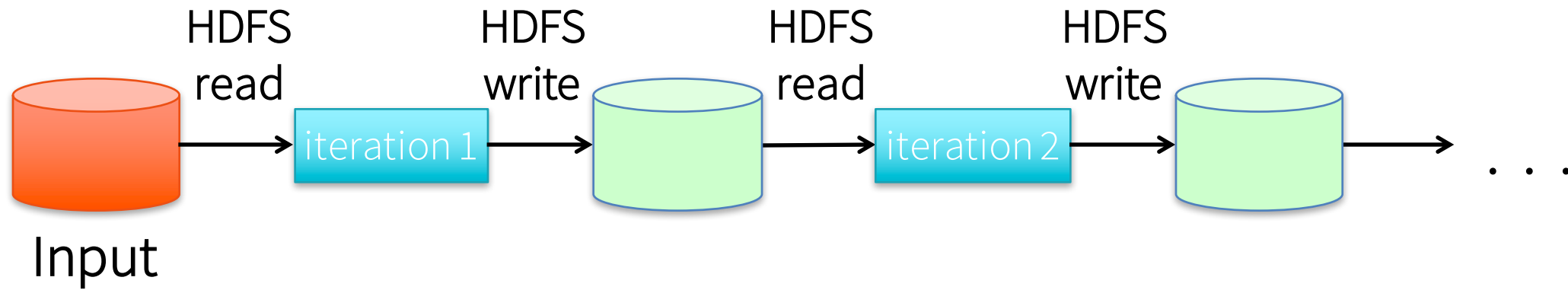
- 一个开放源代码群集计算框架
 - 最初在加利福尼亚大学伯克利分校的 AMPLab 开发
 - 与 Hadoop 一样，从批处理扩展而来
 - 提供内存计算范式
- 将线性的 MapReduce 计算范式扩展到 DAG 描述的任务的流水线处理
 - 将 Hadoop 扩展到流、迭代 MapReduce 和图形分析操作
 - 支持通用并行计算
 - Microsoft HDInsight, Amazon Elastic MapReduce 都支持 Spark

Spark Core

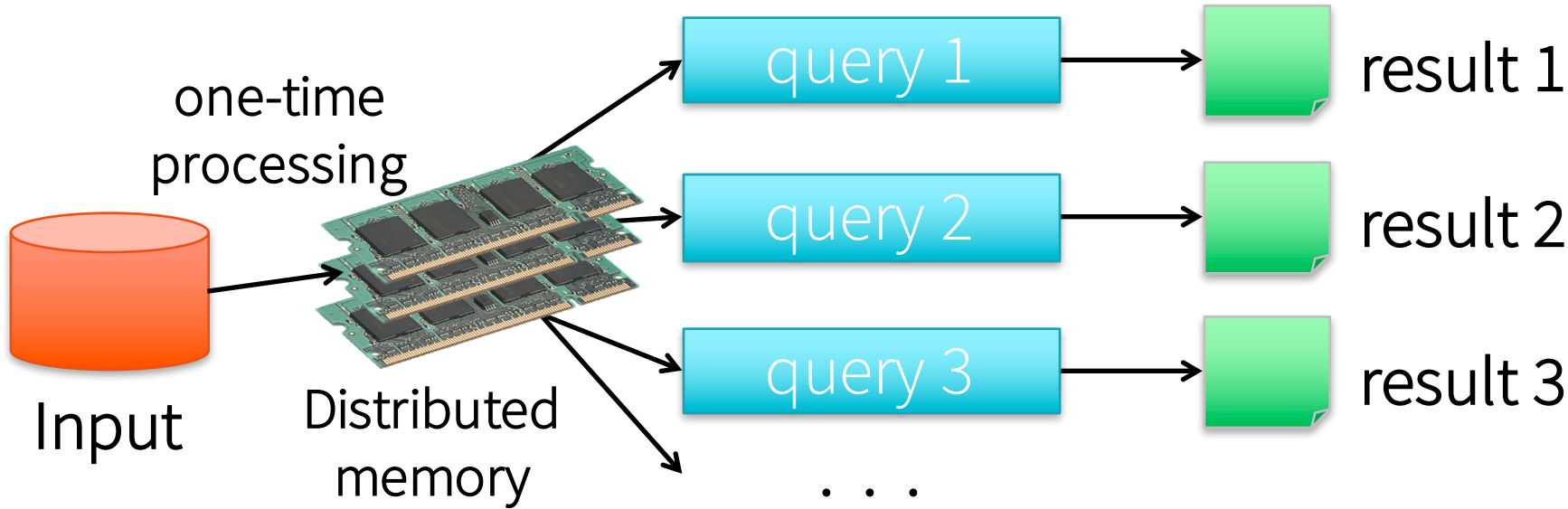
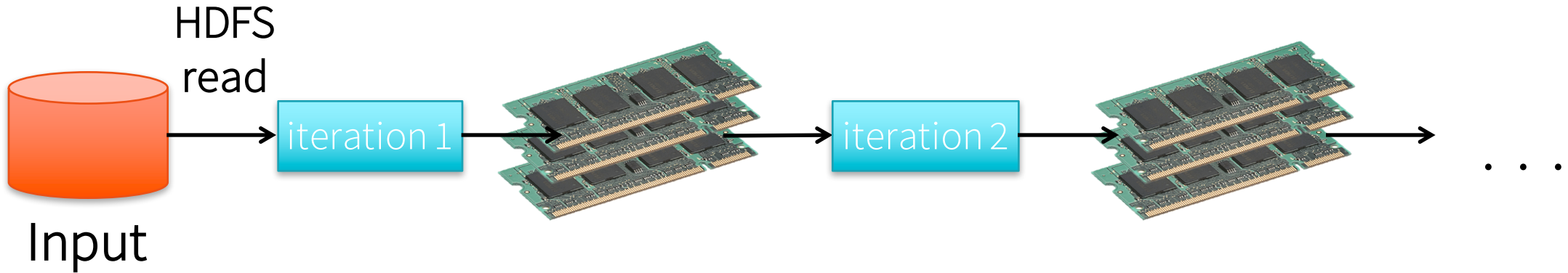
- Spark 核心提供分布式任务调度，调度和基本 I/O 功能
- 对于集群管理，Spark 支持其自己的独立调度程序，Hadoop YARN 或 Apache Mesos
- 对于分布式存储，Spark 可以与 HDFS，MapR 文件系统（MapR-FS），Cassandra，OpenStack Swift，Amazon S3 等接口



MapReduce: Use Disk



Spark: In-Memory Data Sharing



10-100x faster than network and disk

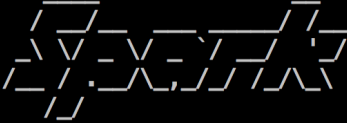
分布式执行模型

- 为了使用 Spark，开发人员编写一个驱动程序，该程序连接到一组 Worker
- 驱动程序定义一个或多个 RDD，并对它们调用操作
- 驱动程序上的 Spark 代码还会跟踪 RDD 的 lineage
- Worker 运行在集群服务器上，可将 RDD 分区存储在 RAM 中

Interactive Shell

```
1. pyspark (java)
$ pyspark
Python 2.7.9 (default, Jan 7 2015, 11:49:12)
Type "copyright", "credits" or "license" for more information.

IPython 2.4.0 -- An enhanced Interactive Python.
?          -> Introduction and overview of IPython's features.
%quickref  -> Quick reference.
help       -> Python's own help system.
object?    -> Details about 'object', use 'object??' for extra details.
Welcome to

           version 1.6.0

Using Python version 2.7.9 (default, Jan 7 2015 11:49:12)
SparkContext available as sc, HiveContext available as sqlContext.

In [1]: █
```

(Scala, Python and R only)

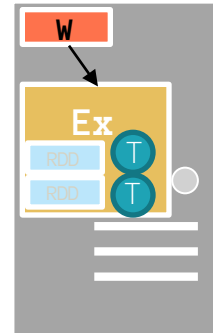
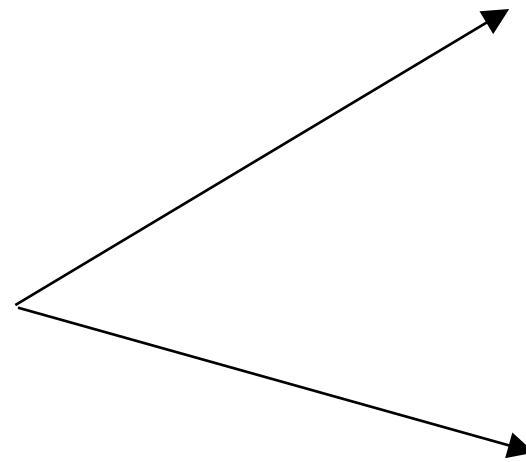
Driver Program

```
Python 2.7.9 (default, Jan 7 2015, 11:40:12)
Type "copyright" for more information.

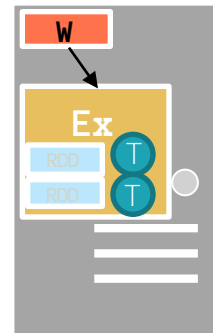
Python 2.4.0 - An enhanced Interactive Python.
Help >>> Introduction and overview of Python's features.
quit() >>> Quit the interpreter.
help() >>> Python's help system.
object? >>> Details about 'object' or 'object()' for extra details.
Welcome to

Python version 1.6.9

Using Python version 2.7.9 (default, Jan 7 2015 11:40:12)
Use Ctrl-C to abort, Ctrl-D for EOF.
In [1]:
```



Worker Machine



Worker Machine

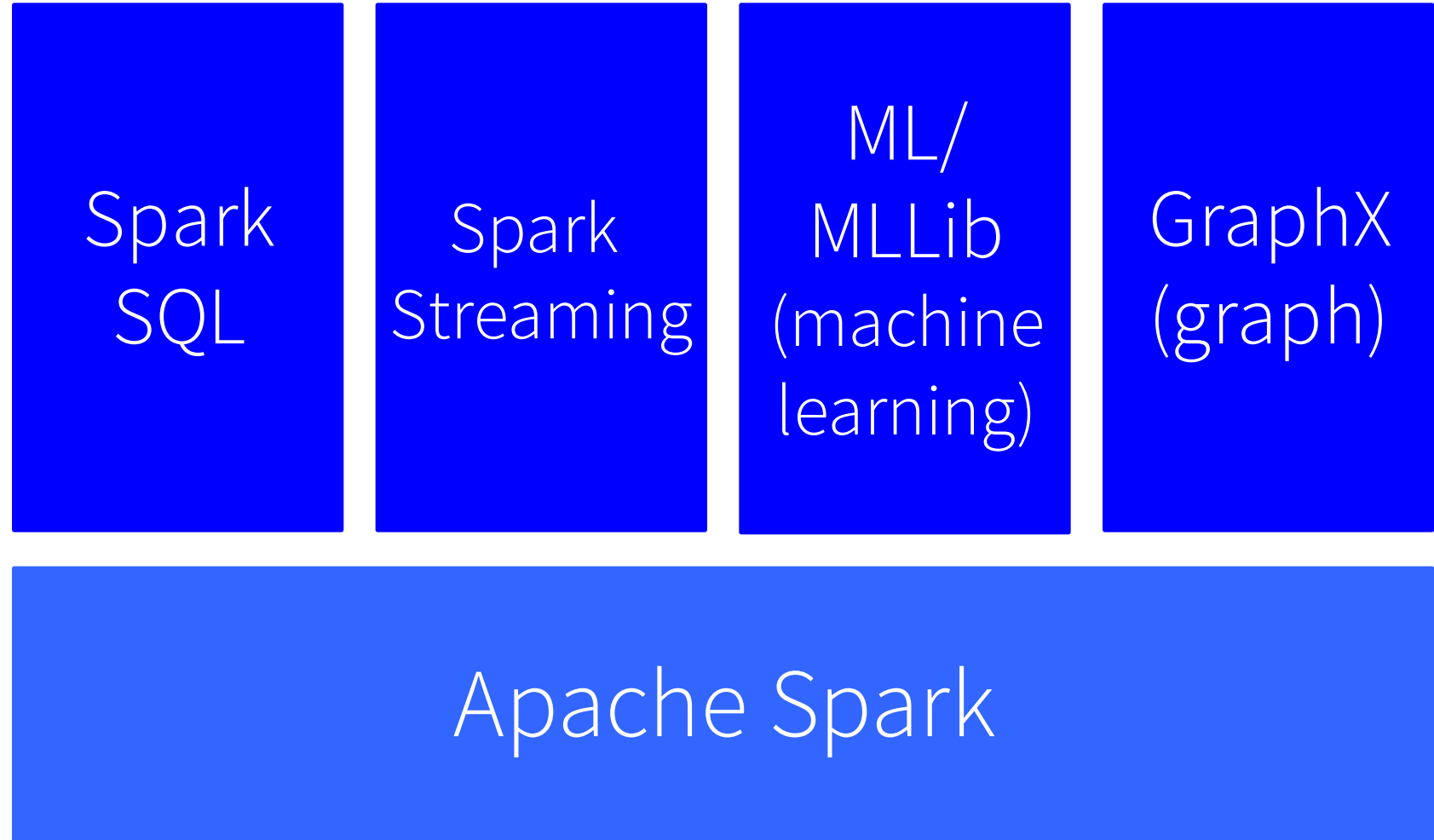
Spark 模块

- Spark SQL 处理结构化数据
- Spark Streaming 处理实时数据流
- MLlib 包含常见的机器学习功能
- GraphX 用于处理网络图

Opportunity



- Keep more data *in-memory*
- New distributed execution environment
- Bindings for:
 - Python, Java, Scala, R



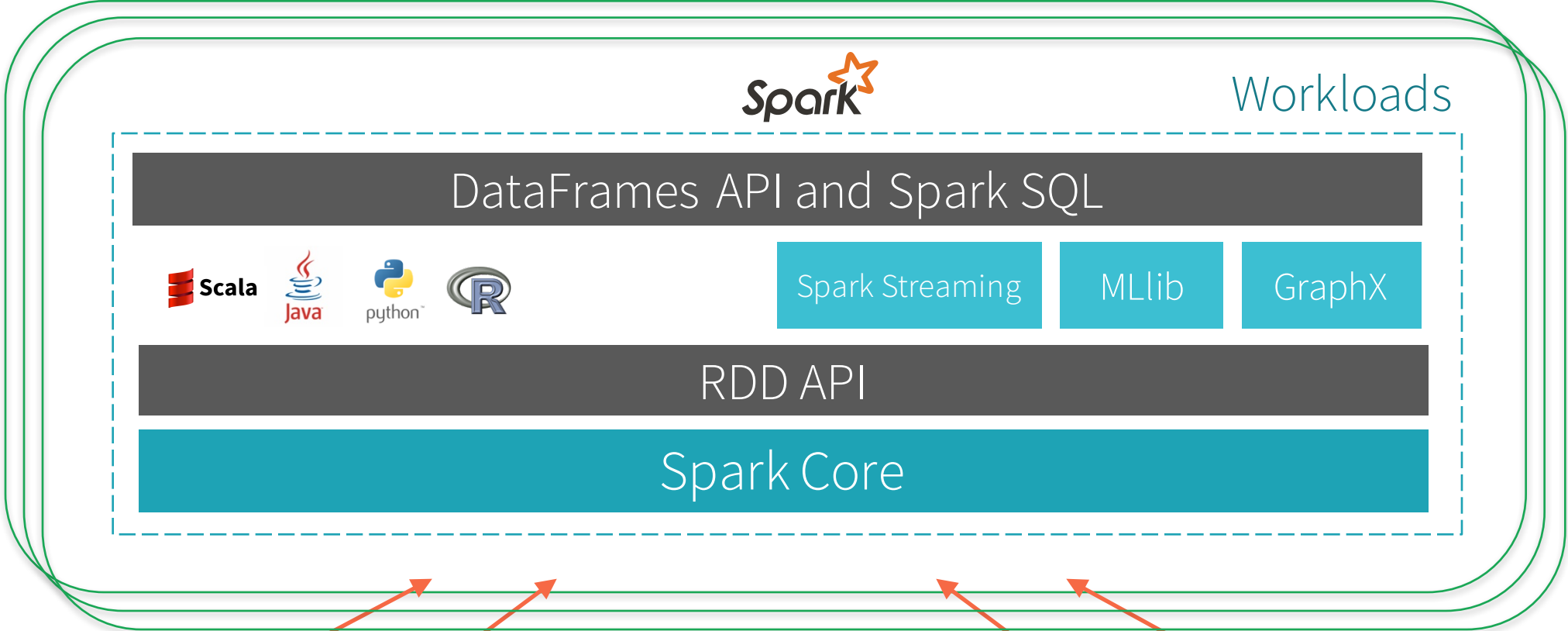
Environments

YARN

 docker

 EC2

 MESOS



A row of data source logos including **hadoop HDFS**, **cassandra**, **HIVE**, **APACHE HBASE**, **PostgreSQL**, **CSV**, **{JSON}**, **MySQL**, **elasticsearch.**, **Parquet**, and **Parquet**.

Data Sources

Additional Resources

Books

- *Learning Spark*. <http://shop.oreilly.com/product/0636920028512.do>
- *Advanced Analytics with Spark*.
<http://shop.oreilly.com/product/0636920035091.do>
- *Scala for the Impatient*. <http://www.informit.com/store/scala-for-the-impatient-9780321774095>

Web Sites

- Spark documentation: <http://spark.apache.org/docs/latest/>
- Databricks blog: <https://databricks.com/blog>