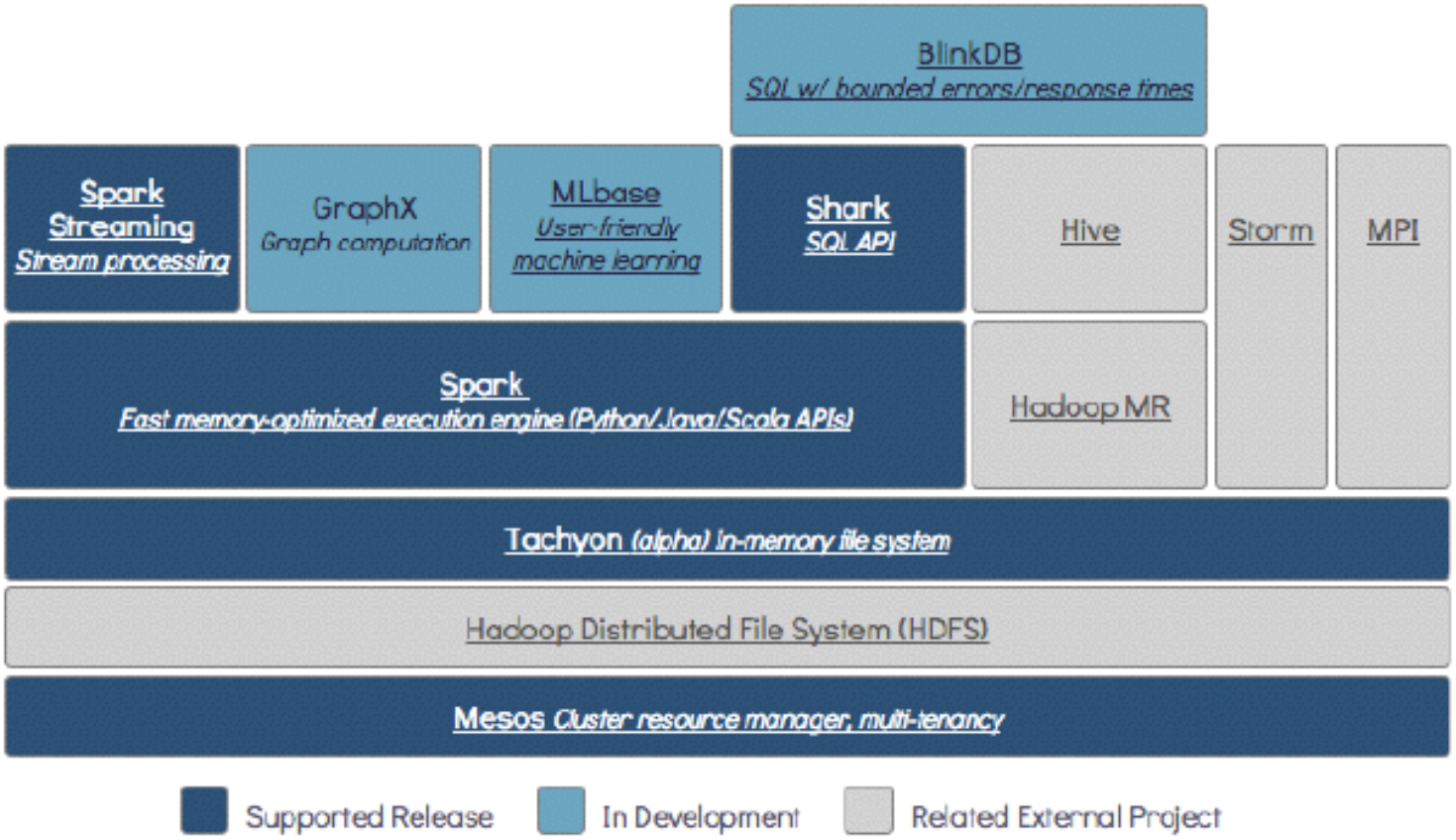
Spark 初始

讲师：陈博

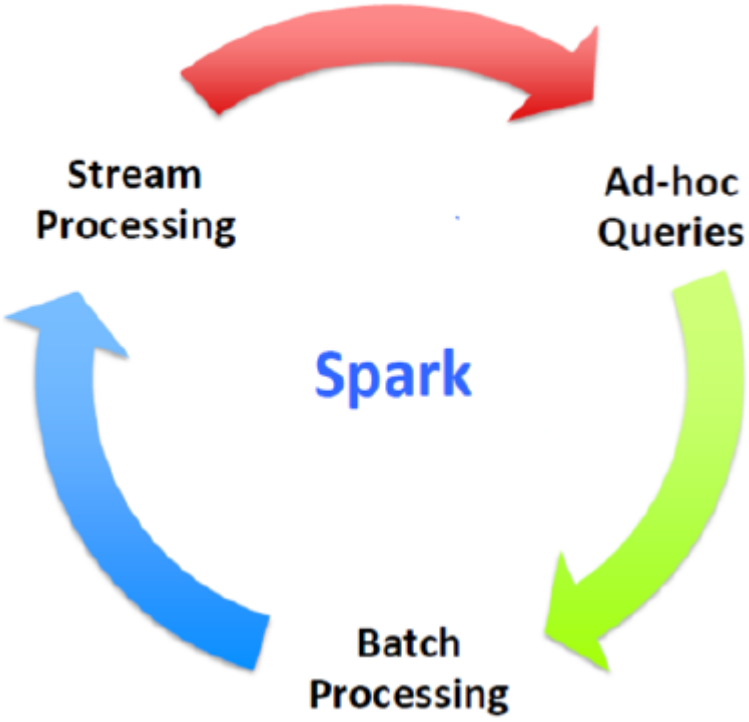
Spark是什么

- What’ s Spark?
- Apache Spark is an open source cluster computing system that aims to make data analytics fast
- both fast to run and fast to wrtie

- The Berkeley Data Analytics Stack



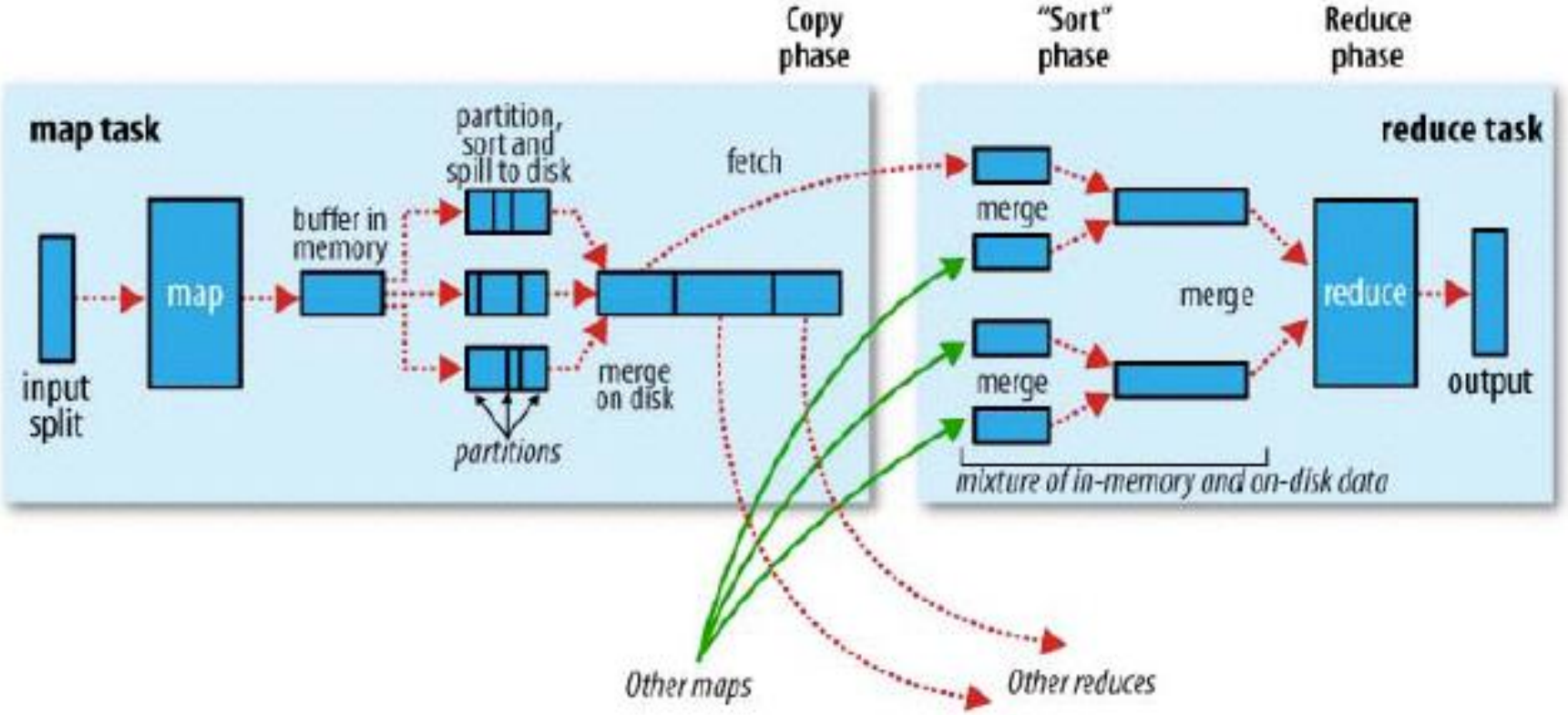
- One stack rule them all !!!



回顾hadoop

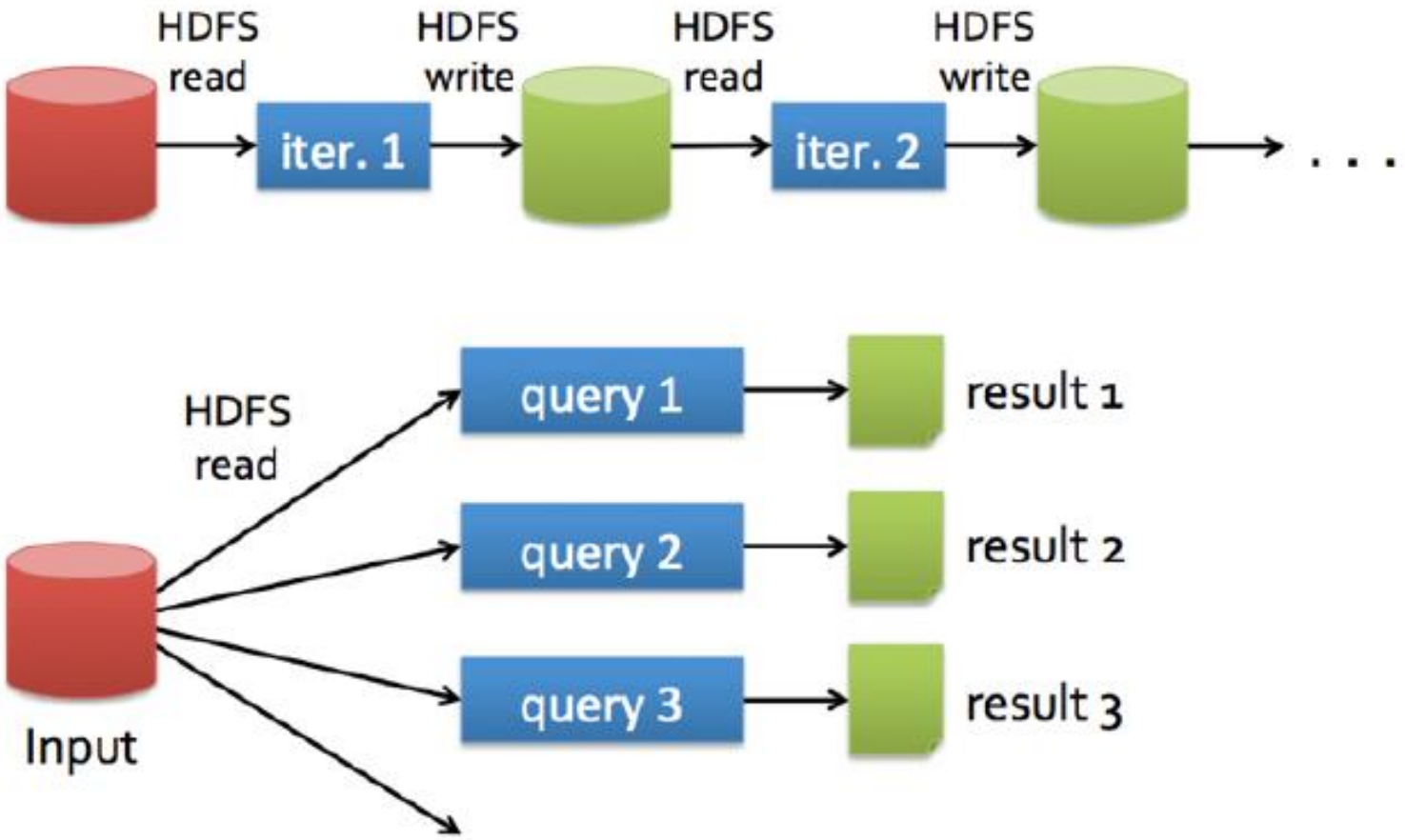
- Spark 相比 hadoop历史
  - 发展尤为迅速
  - Spark 5年时间
  - Hadoop历史 10年时间

回顾hadoop


Copy phase
"Sort" phase
Reduce phase
map task
reduce task
partition, sort and spill to disk
fetch
buffer in memory
map
input split
merge on disk
partitions
merge
merge
merge
reduce
output
mixture of in-memory and on-disk data
Other maps
Other reduces

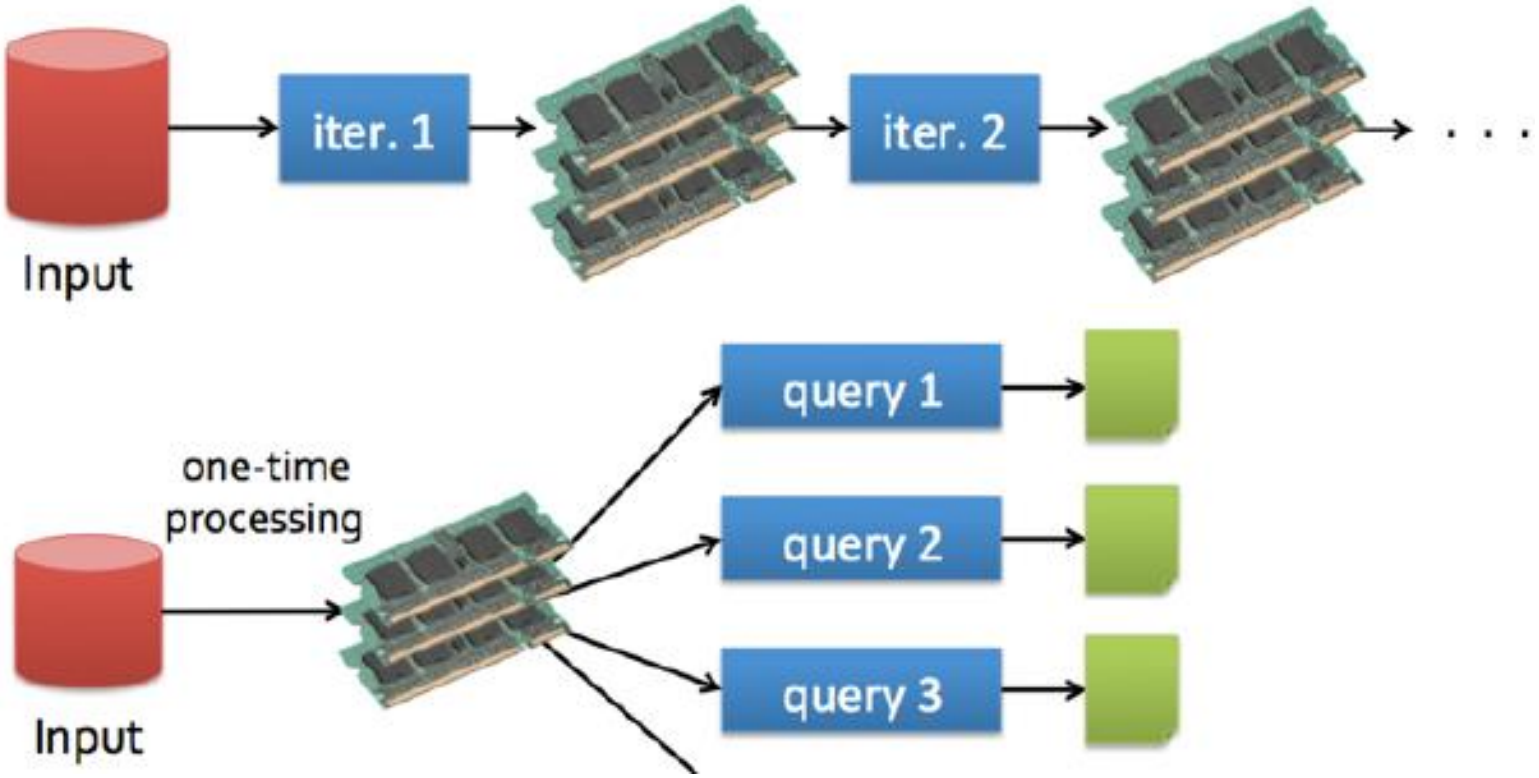
Hadoop的共享数据慢

- 为什么慢？？？额外的复制，序列化，磁盘IO开销



Spark的共享数据快

- 快只是因为内存计算？当然还有DAG



Spark API

- 支持3种语言的API
  - Scala（很好）
  - Python（不错）
  - Java（...）

通过哪些模式运行Spark呢

- 有4种模式可以运行
- Local　　多用于测试
- Standalone
- Mesos
- YARN　　最具前景