02

第2章 Spark的设计与运行原理

授课教师：[李阳源]
15138617705@163.com
主页：https://sites.google.com/view/dalguri



第2章 Spark的设计与运行原理

目录

CONTENT

Spark概述

Spark生态系统

Spark运行架构

Spark的部署方式



2.4 Spark的部署和应用方式

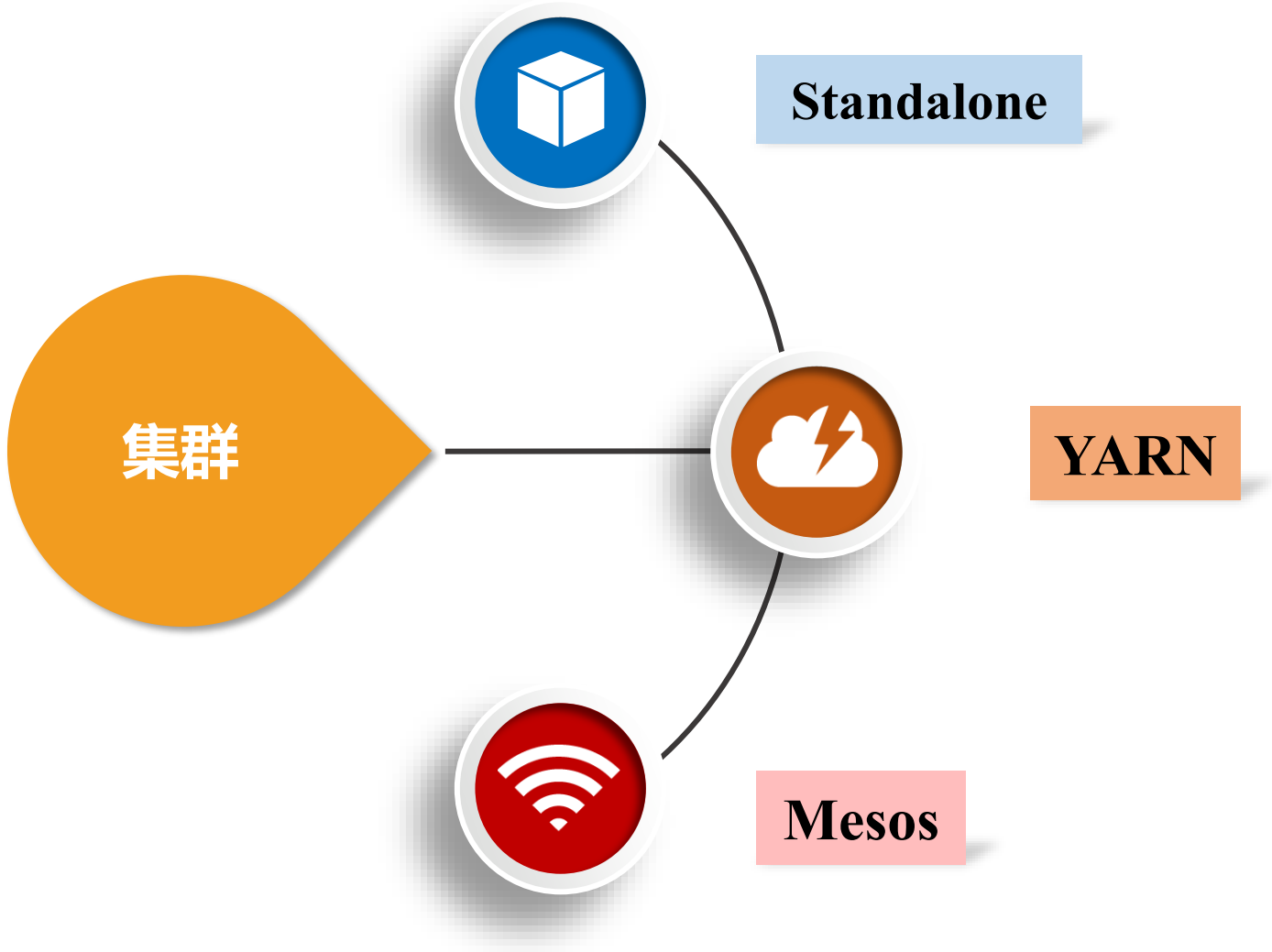
2.4 Spark的部署和应用方式



单机

集群

2.4 Spark的部署和应用方式



2.4 Spark的部署和应用方式





Standalone

Spark自带

效率不高

2.4 Spark的部署和应用方式





Mesos

血缘关系

性能匹配好

2.4 Spark的部署和应用方式



Spark

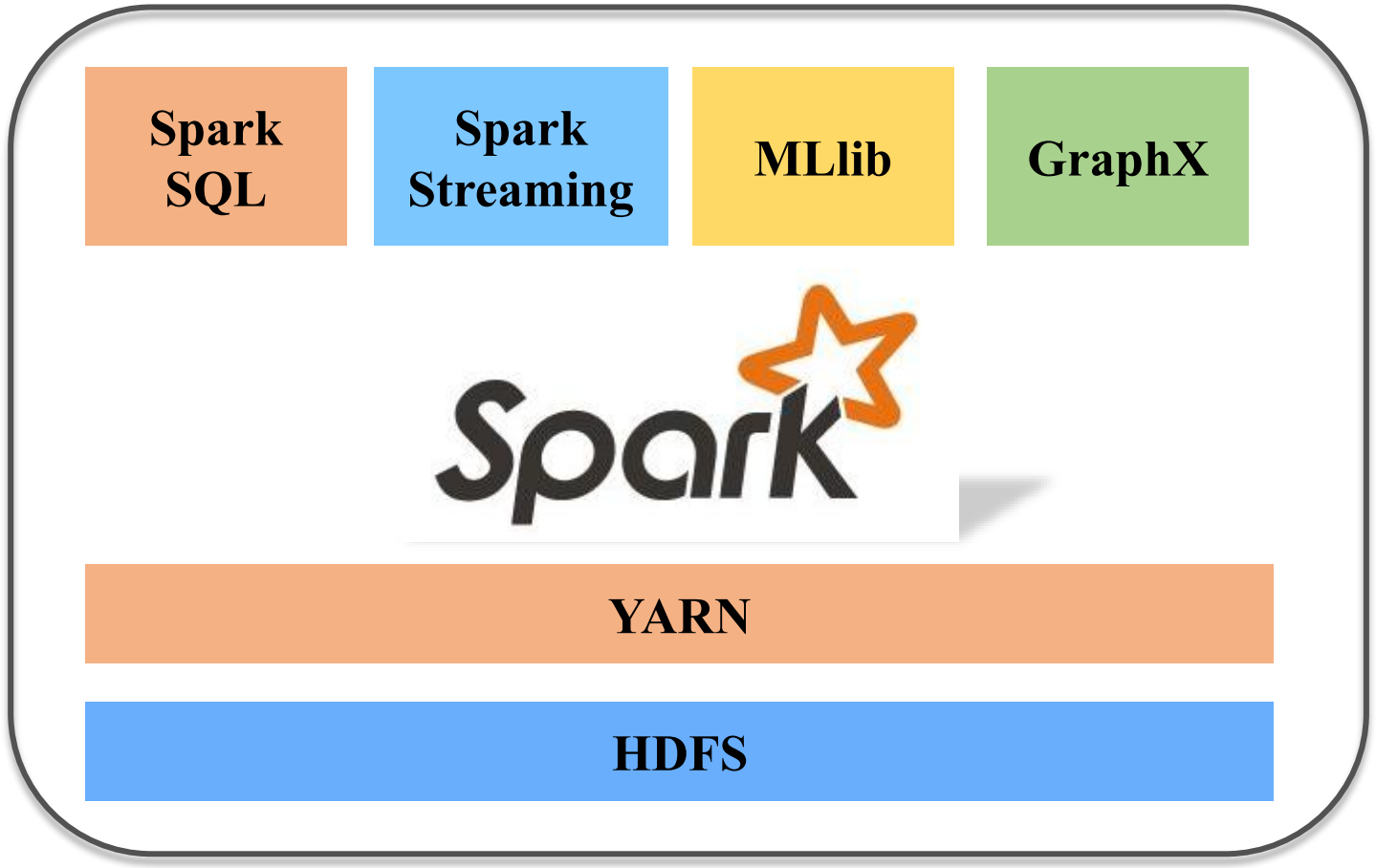
hadoop MapReduce

APACHE STORM™
Distributed • Resilient • Real-time

hadoop

YARN

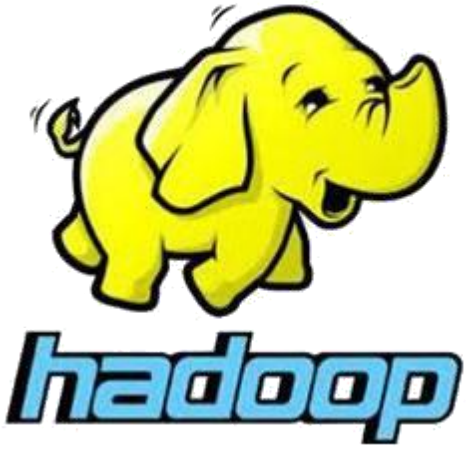
2.4 Spark的部署和应用方式



2.4 Spark的部署和应用方式







2.4 Spark的部署和应用方式



2.4 Spark的部署和应用方式





Spark做批处理比MapReduce获得更好的性能

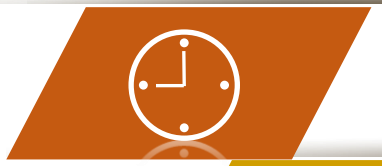
2.4 Spark的部署和应用方式



组件多



SparkContext



Spark SQL　查询计算



Spark Streaming　流计算



MLlib　机器学习

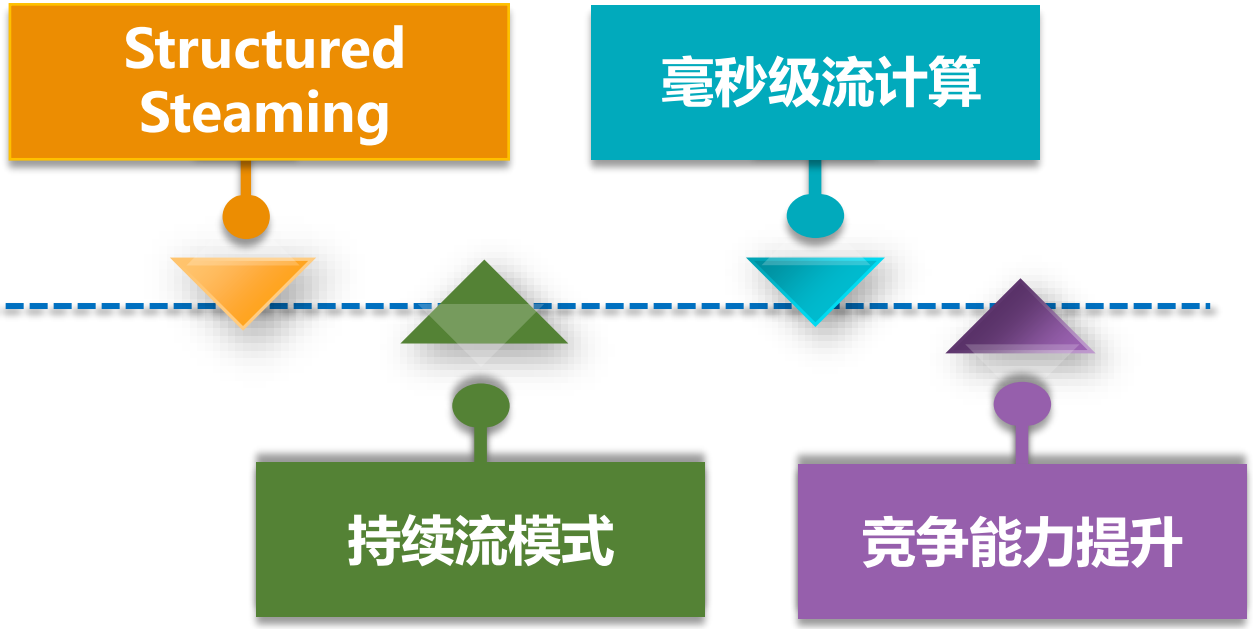


GraphX　图计算

2.4 Spark的部署和应用方式

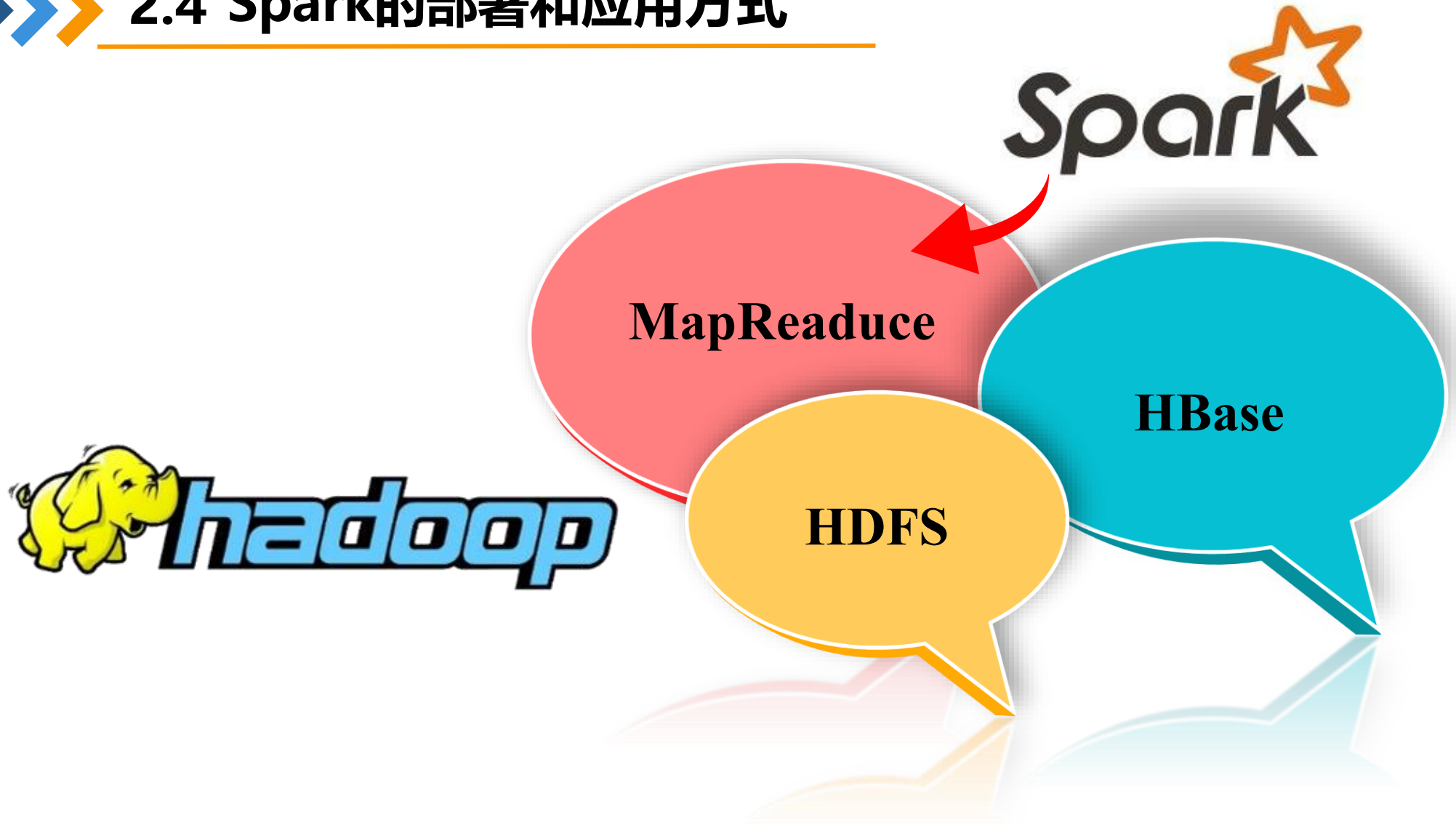








2.4 Spark的部署和应用方式



2.4 Spark的部署和应用方式





HDFS

HBase

它们共同来满足企业的相关应用场景需求

2.4 Spark的部署和应用方式理解测验（选择题）



问题1

Spark集群部署模式中，哪种方式是Spark自带的集群管理器？

A. YARN
B. Mesos
C. Standalone
D. Kubernetes

2.4 Spark的部署和应用方式理解测验（选择题）



问题2

以下哪个不是Spark的核心组件？

A. Spark SQL
B. Spark Streaming
C. MLlib
D. MapReduce

2.4 Spark的部署和应用方式理解测验（选择题）



问题3

Spark与Mesos之间性能匹配好的主要原因是什么？

A. 都使用Java语言开发
B. 都是Apache基金会的项目，有血缘关系
C. 都支持分布式计算
D. 都基于内存计算

2.4 Spark的部署和应用方式理解测验（选择题）



问题4

Spark 2.3版本引入的新功能是什么？

A.Spark SQL

B. MLlib

C. Structured Streaming

D. GraphX

2.4 Spark的部署和应用方式理解测验（选择题）



问题5

在YARN管理的Spark生态系统中，底层使用的分布式文件系统是什么？

A.HDFS

B. HBase

C. Cassandra

D. MongoDB



Thank you