1 Spark 概述

1.1 Spark 定义

构建与计算集群之上支持大数据集的快速的通用的处理引擎

a)快速： DAG、Memory

b)通用：集成Spark SQL、Streaming、Graphic、R、Batch Process

c)运行方式：

StandAlone

YARN

Mesos

AWS

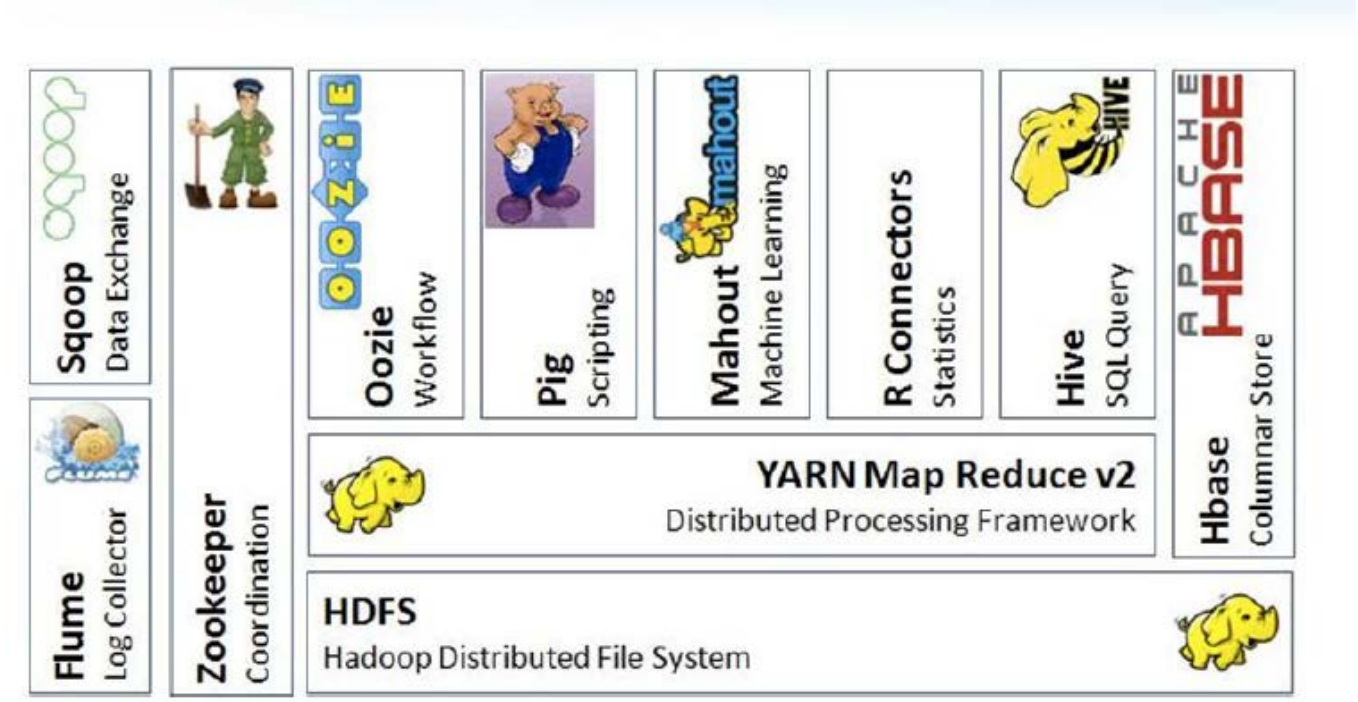
d)数据来源：

Hdfs Hbase Tachyon Cassandra Hive

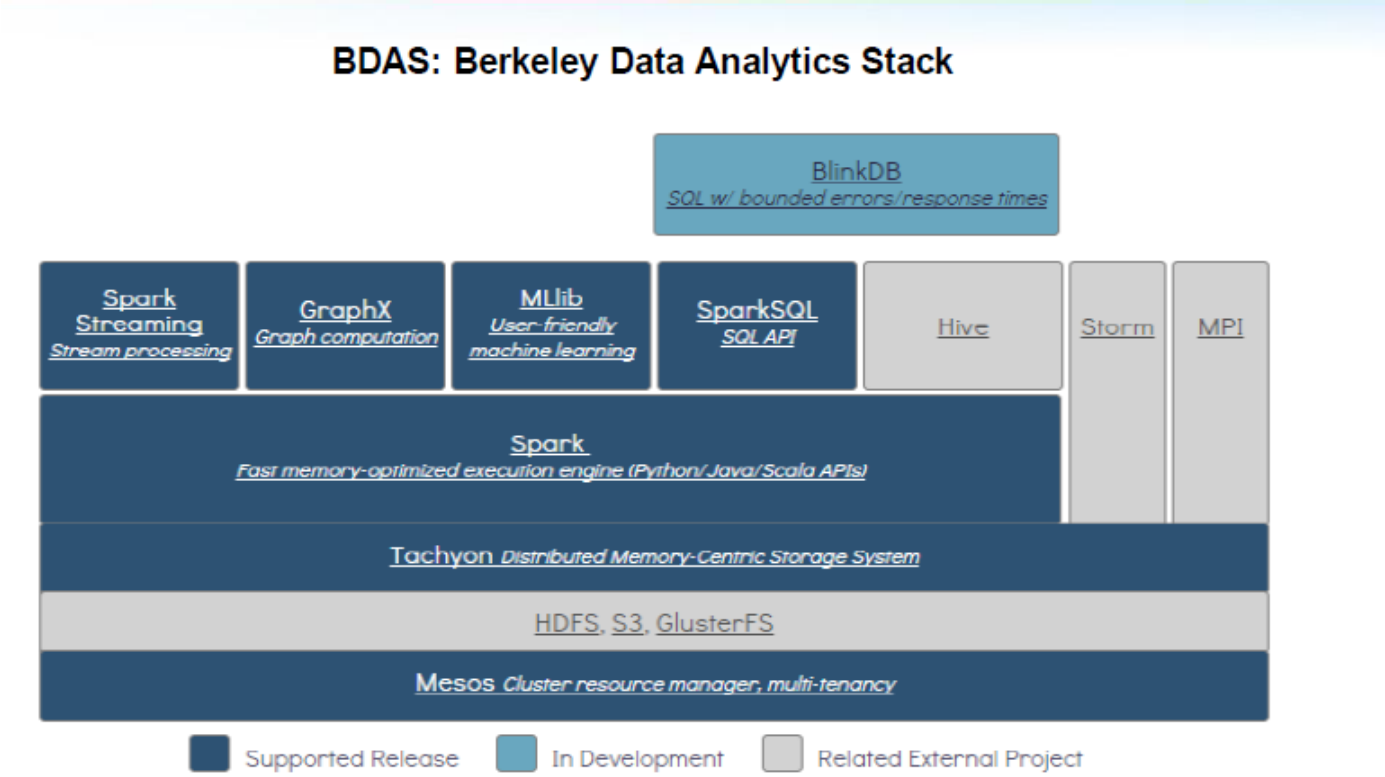
and Any Hadoop Data Source

1.2 Spark 协议栈

1.2.1 Hadoop 生态系统



1.2.2 Spark 协议栈



1.2.3 Spark VS Mapreduce

MapReduce	Spark
数据存储结构：磁盘hdfs文件系统的split	使用内存构建弹性分布式数据集RDD，对数据进行运算和cache
编程范式：Map + Reduce	DAG(有向无环图)：Transformation + action
计算中间数据落磁盘，io及序列化、反序列化代价大	计算中间数据在内存中维护，存取速度是磁盘的多个数量级
Task以进程的方式维护，任务启动就有数秒	Task以线程的方式维护，对小数据集的读取能达到亚秒级的延迟

MapReduce 与Spark比较
1.what? 处理对象
a)MapReduce:基于磁盘File的大数据处理系统
b）Spark：基于RDD(弹性分布式数据集)，可以显示的将RDD数据存储到磁盘和内存中
2.where(软硬件上下文)？
a)MapReduce: Disk
b)Spark: Mem
3.when? (应用场景)
a)MapReduce：可以处理超大规模数据，适合日志分析挖掘等迭代较少的长任务需求，结合了数据的分布式的计算
b)spark：适合数据的挖掘，机器学习等多伦迭代式计算任务

容错性：
a)数据容错性
MapReduce：容错性基于HDFS 冗余机制 ->安全模式->数据校验->元数据保护
spark：容错性基于RDD，spark容错性比mapreduce容错性低，但在处理效率上优势比较明显
b)节点容错性