大讲台
dajiangtai.com

Hadoop生态圈及企业应用

主讲人：杨老师

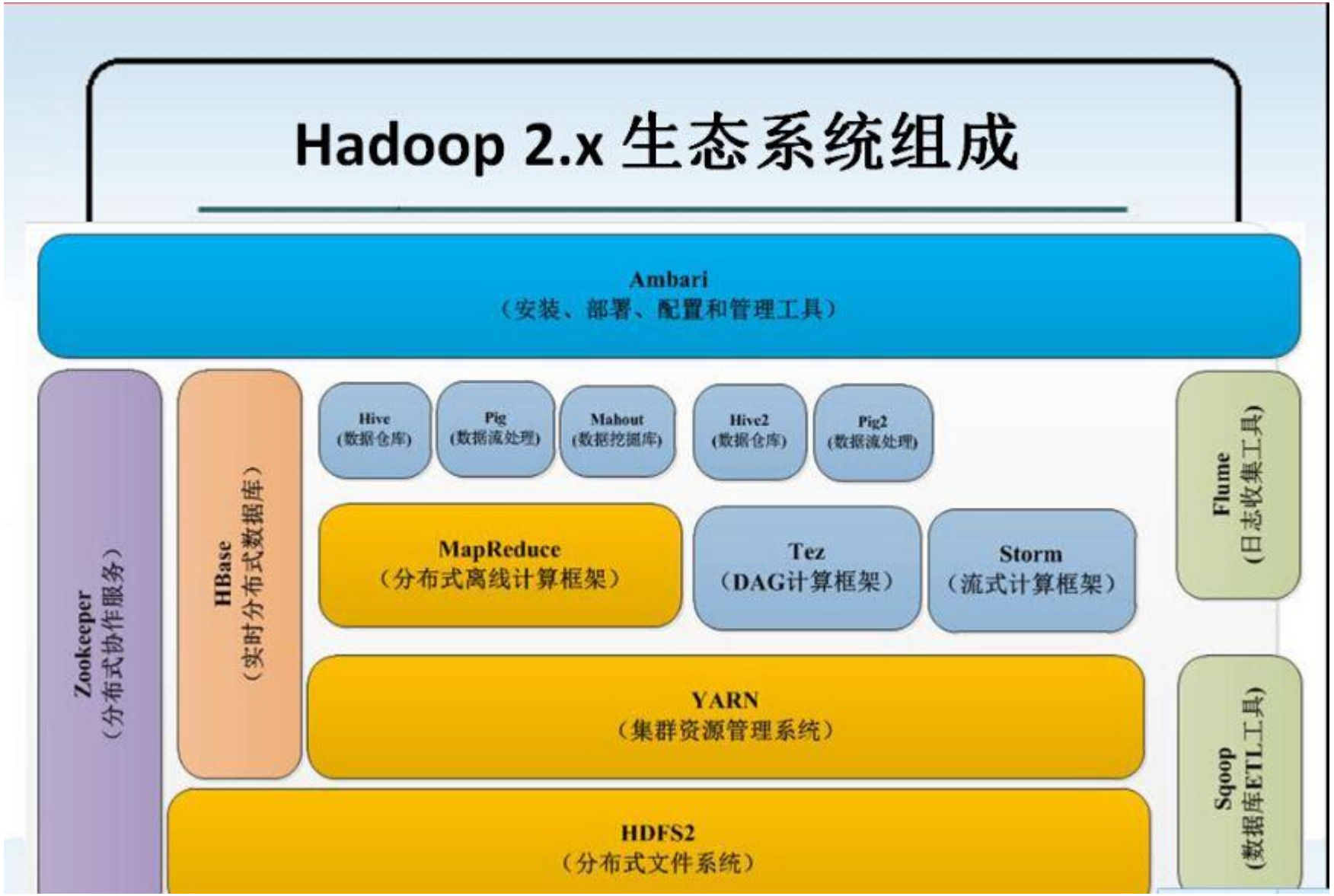
课程介绍

课程章节划分

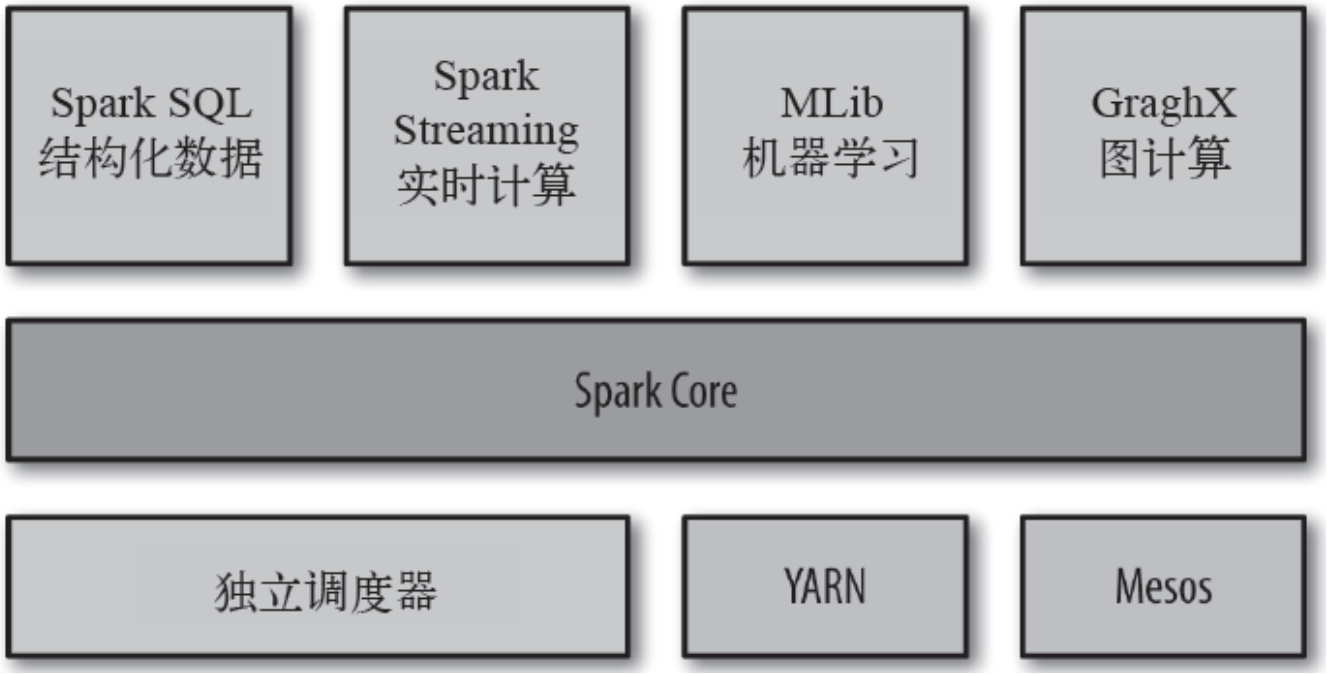
- Hadoop生态圈
- 企业应用

Hadoop生态圈

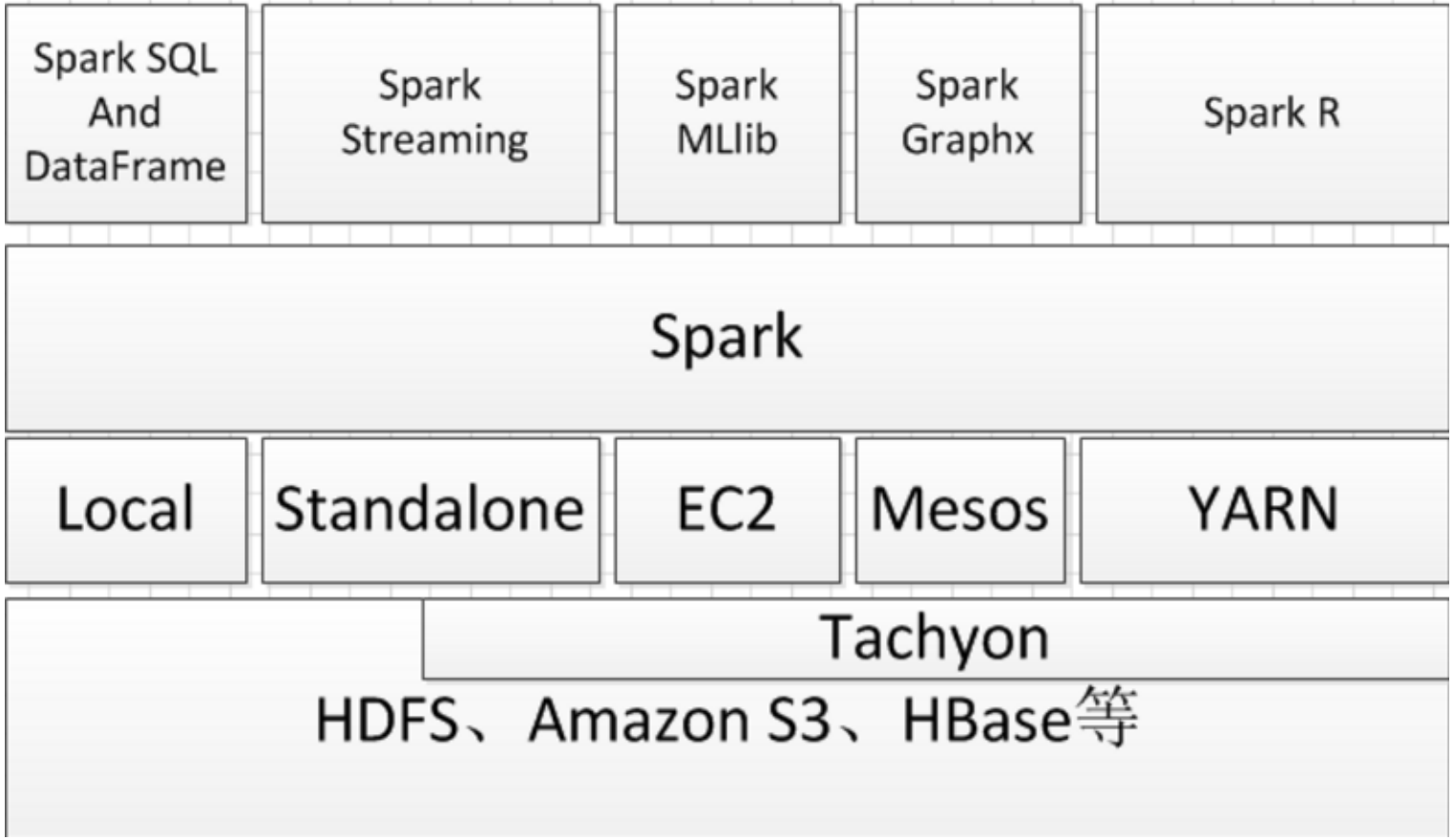
Hadoop生态系统



Spark组成---大一统的软件栈（计算）



Spark自有生态圈(以Spark为核心)

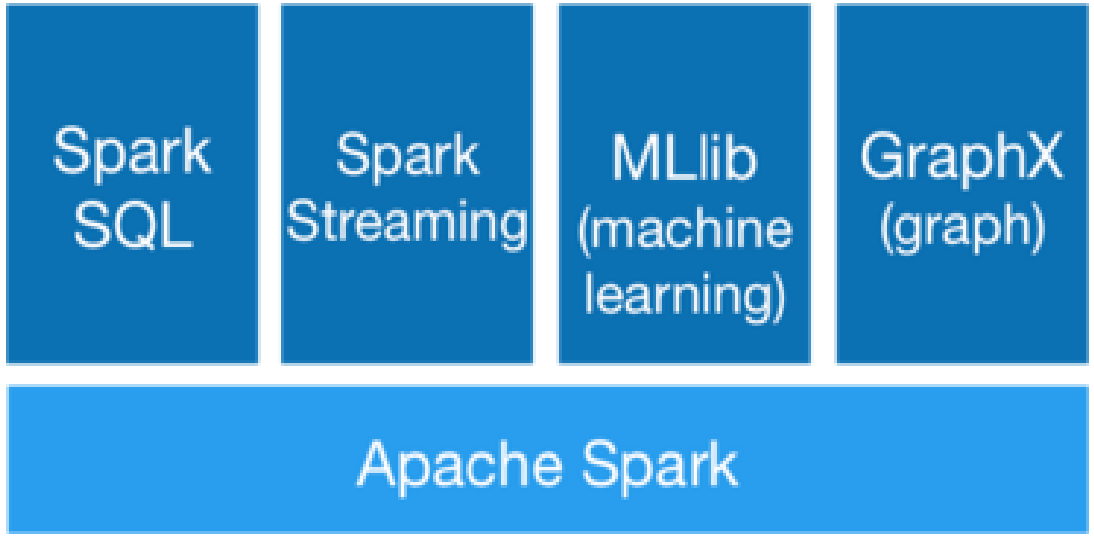


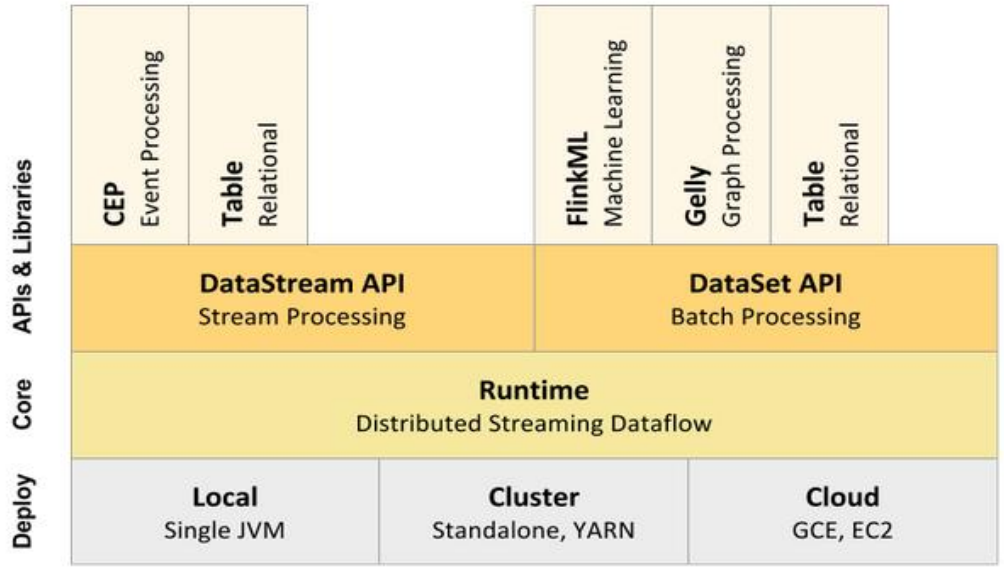
- 1.3.0 及后续版本中，SchemaRDD 已经改名为DataFrame
- 1.4才有Spark R，他是一个R语言包，它提供了轻量级的方式使得可以在R语言中使用Spark
- 1.6引入Dataset接口

Spark与Hadoop的关系---青于于蓝（仅仅是计算）

	spark	Storm	hadoop
流式计算	Streaming	Storm	无
批计算/离线运算	Core	无	Map/reduce
图计算	GraphX	无	无
机器学习	MLlib	无	Mahout
Sql	DataFrame	无	Drill、hive

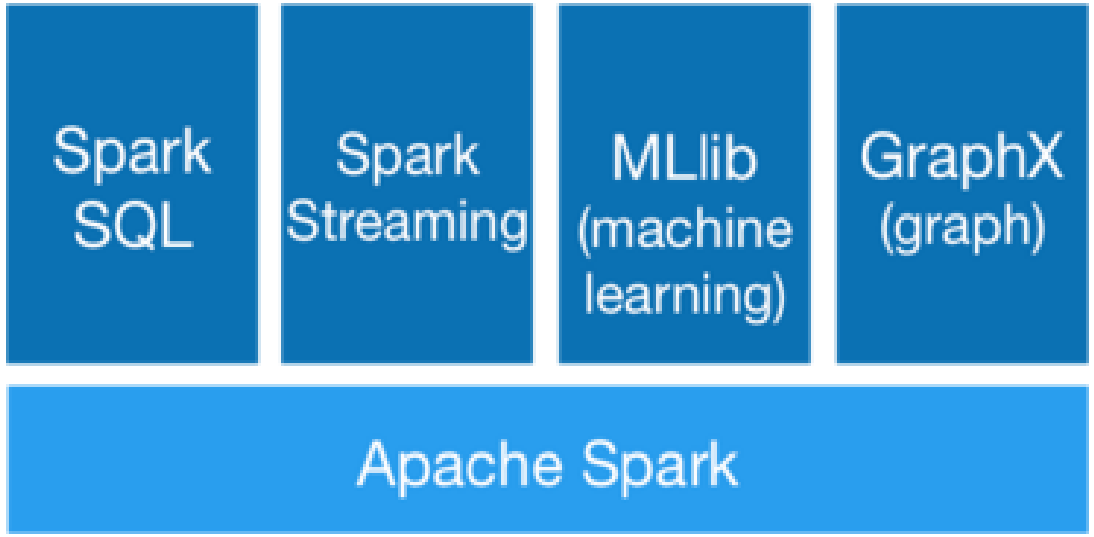
Spark的竞争对手---Flink

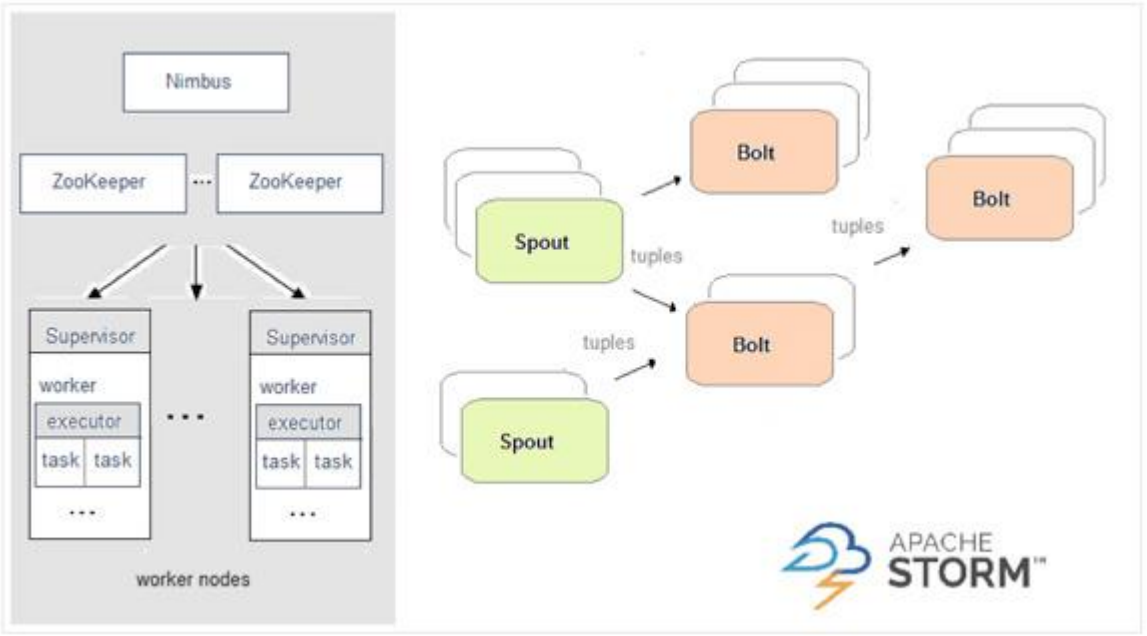




- Flink是先有流处理后有批处理

Spark的竞争对手---Storm/JStorm





- Storm仅限于流计算(topology)
- JStorm参照Flink改进了Storm

- Hadoop3.x

http://news.cnw.com.cn/news-international/htm2016/20160603_327510.shtml

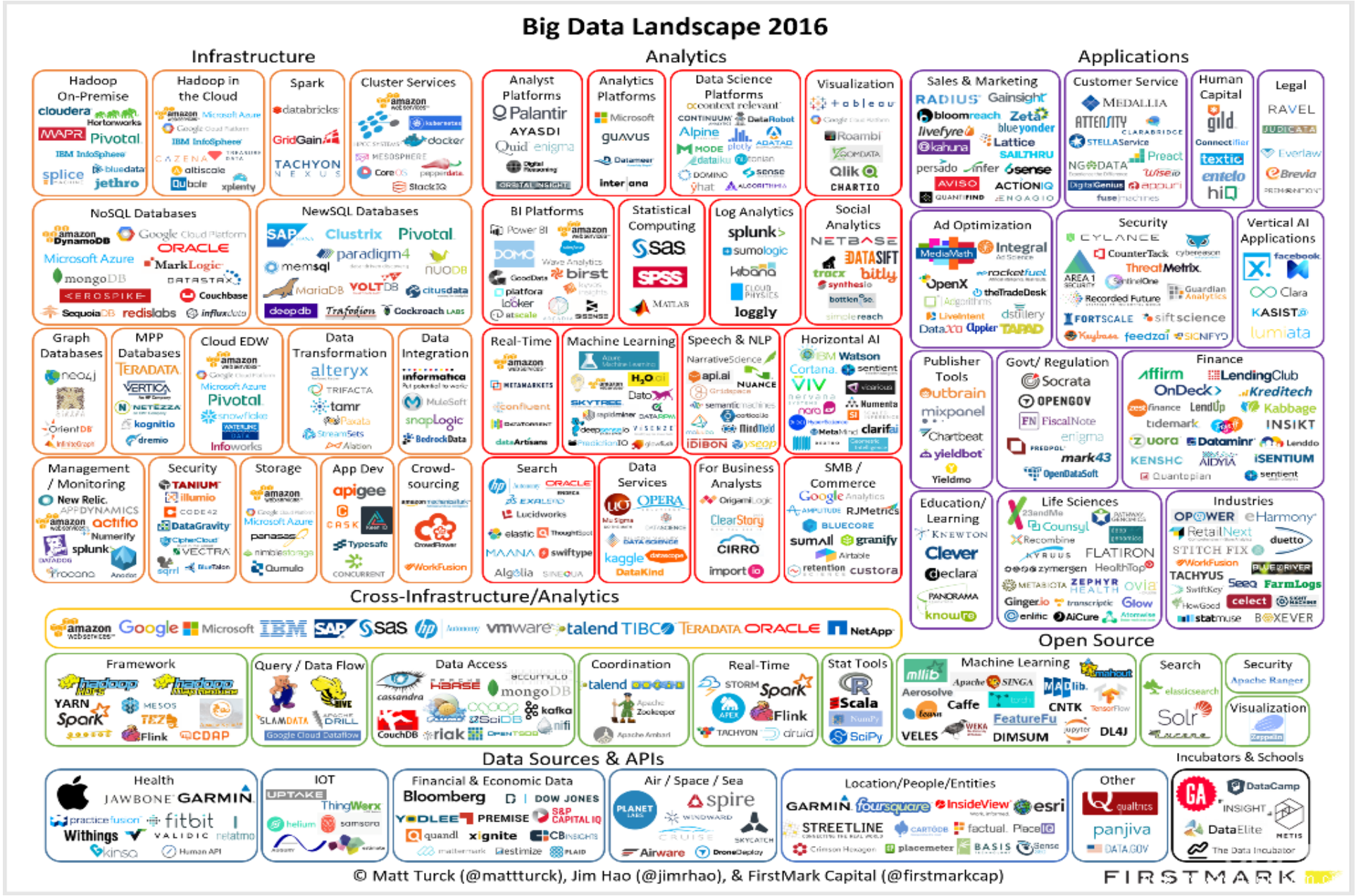
- Apache Beam

http://baike.baidu.com/link?url=8lvF2s3g_nUIo-yy52Xn3hOo9k8-ny0GEc0Fa0rL3-4iqxU49GoRZAq07UsYlRLYlNnQkuKT-xpa95Liw28gQ2L1mi1j1V1d1hjllLRXtHe

- Angel

http://baike.baidu.com/link?url=P1SCRfSLfA3CwpNSX5lRAs41HZLEQwQFdk13kn85B-s7WV7FmWQtgpPxW1twmcHW95kT4yvzJucTORZt8PoRB_

更大的生态圈--Spark与Hadoop的关系---相辅相成



我们该如何对待

- 无需纠结谁替代谁(计算方面相互补充)
- 以某一个组件为突破口逐渐深入
- 大局观和生态意识
- 一切以应用场景为出发点

企业应用

企业应用概况---云计算几种模式

- Iaas—比如阿里云（主机）
- Paas—比如进一步部署好环境（目前大数据部门干的事情），然后直接跑应用
- Saas—直接把应用部署好，封装成一个成熟的产品或者软件，直接购买账号使用

企业应用概况---计算场景

- 离线计算
- 实时计算
- 即席查询
    1）对即席探索—hbase
    2）搜索—solr、ElasticSearch
- 迭代计算—机器学习推荐（mahout，mlib）

企业应用概况---应用场景

- 日志存储—历史日志oracle、mysql存不下
- 日志分析—了解运行情况，为用户提供报表、决策等
- 风控—比如贷款，通过大数据手段搜集所有用户信息，觉得是否放贷
- 广告平台—精准投放
- 推荐系统—电商网站、视频网站、简历推荐（拉勾，58同城）
- 舆情分析—美国大选舆情监控

企业应用概况---产生价值

- 直接变现—京东、淘宝推荐购物
- 间接变现—广电收视率，可以为电视台提供决策，提高节目收视率，增加赞助商和广告收入，间接变现

大讲台
dajiangtai.com

THANKS