

Big Data2.0: 未来数据计算

王国仁

北京理工大学

2018年6月9日

- 一 大数据技术发展现状
- 二 未来数据计算发展趋势



一. 大数据技术发展现状

□ 数据思维已成为重要的科学研究方法

科学
一范式

实验思维-科学归纳

1000年前



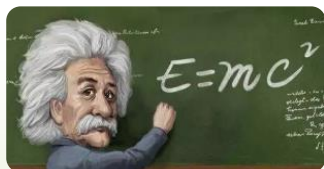
- 对自然现象的描述论证
- 对自然现象进行系统归类

牛顿三大定律提出

科学
二范式

逻辑思维-模型推演

数百年前



- 采用建模方式
- 由特殊到一般进行推演

爱因斯坦相对论提出

科学
三范式

计算思维-仿真模拟

几十年前



- 用计算方式模拟复杂现象
- 科学数据可以用模拟的方法获得

阿波罗登月计划成功

科学
四范式

数据思维-数据密集型科学

2007年以后



- 与大数据密切相关
- 采用IT技术获取、处理、存储、统计分析数据，从中获取知识

AI进入高速发展期



一. 大数据技术发展现状

□ 信息技术社会正在经历从IT时代到DT时代的转变

➤ 近40年以来，信息技术社会的演进经历了五大形态

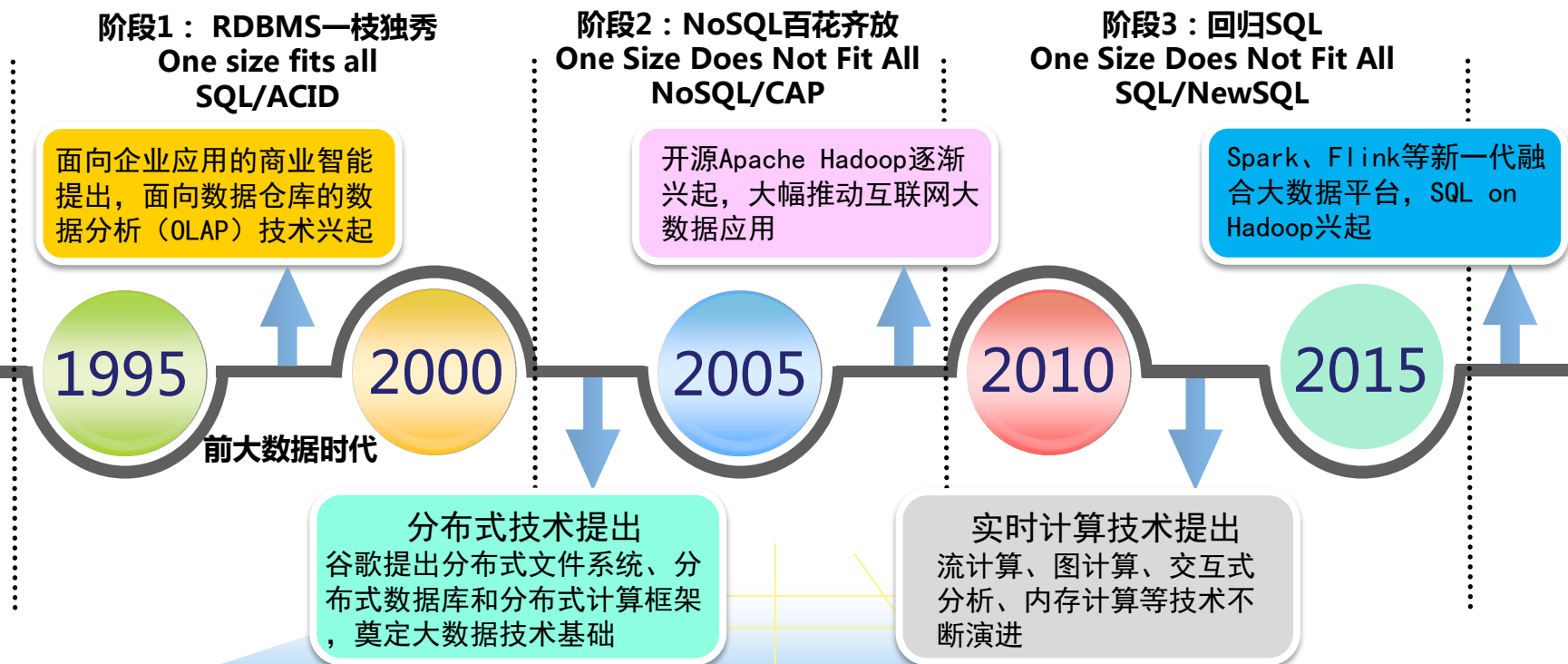




一. 大数据技术发展现状

□ 大数据技术创新日新月异

➤ 大数据处理技术



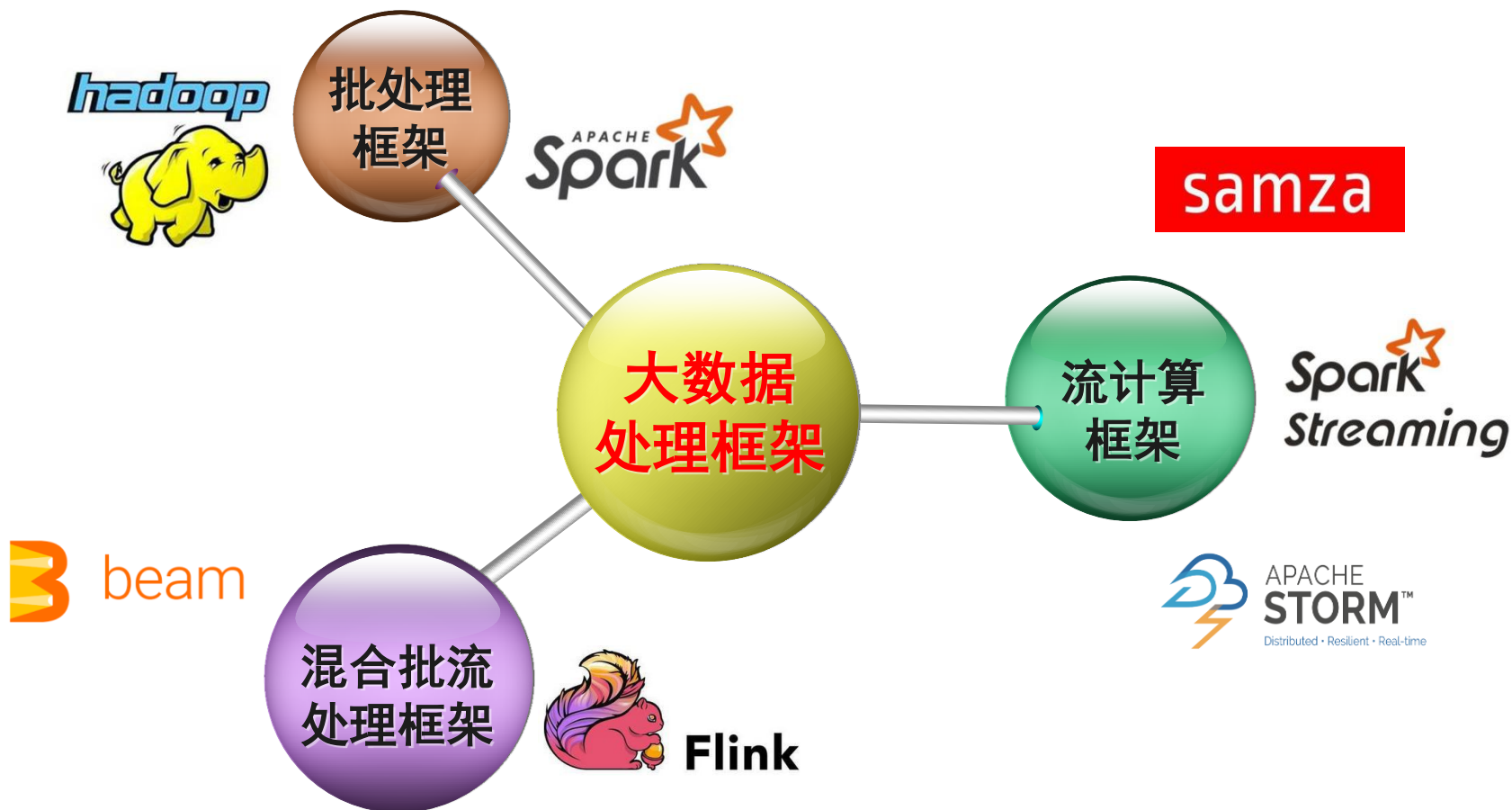
底技术日趋成熟，释放上层数据分析技术的创新活力：
实时流分析、交互分析、机器学习



一. 大数据技术发展现状

□ 大数据技术创新日新月异

➤ 大数据处理平台





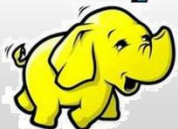
一. 大数据技术发展现状

□ 大数据技术创新日新月异

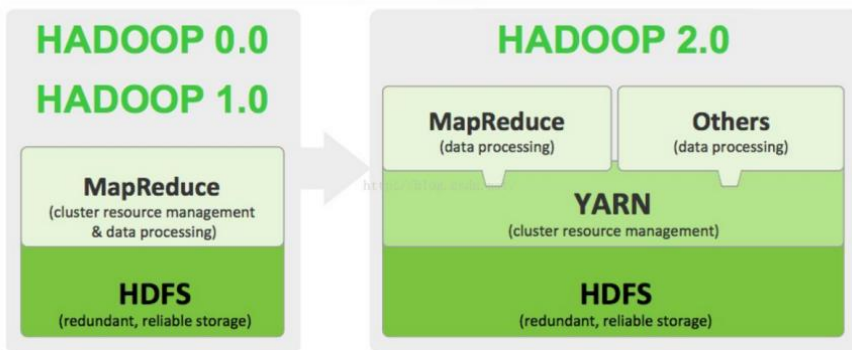
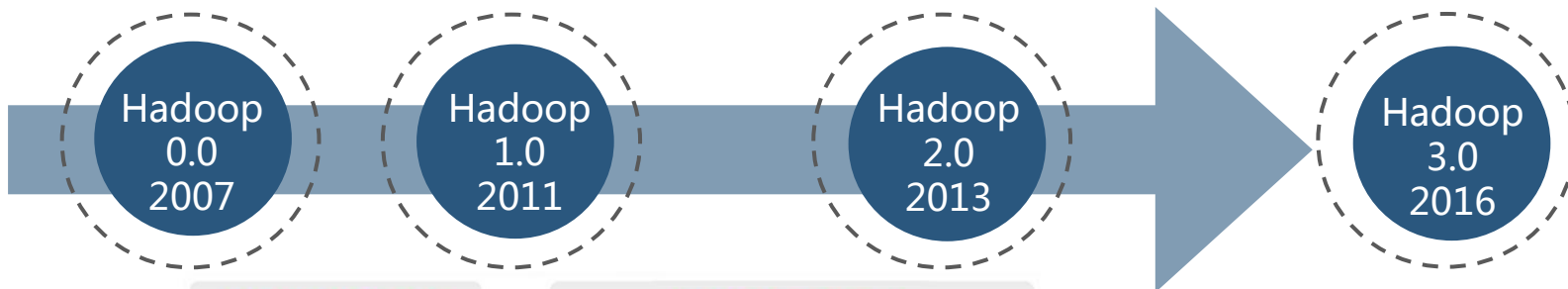
➤ 大数据处理平台

✓ 批处理：Apache Hadoop

hadoop



- 由Apache基金会所开发的分布式系统基础架构
- 是一种专用于批处理的处理框架
- 分布式文件系统HDFS具有很高的容错性，能提供高吞吐量的数据访问





一. 大数据技术发展现状

□ 大数据技术创新日新月异

➤ 大数据处理平台

✓ 批处理：Apache Hadoop-0.0/1.0

Hadoop0.0

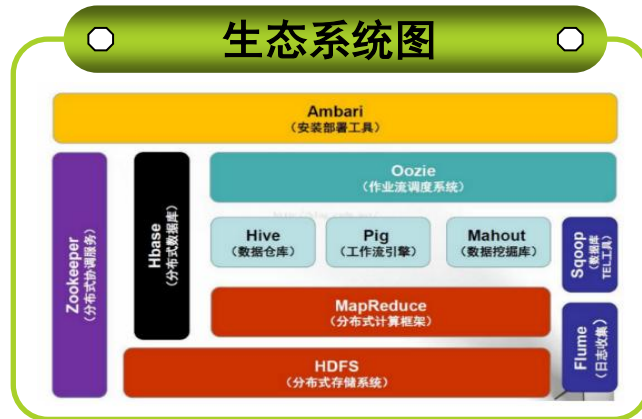
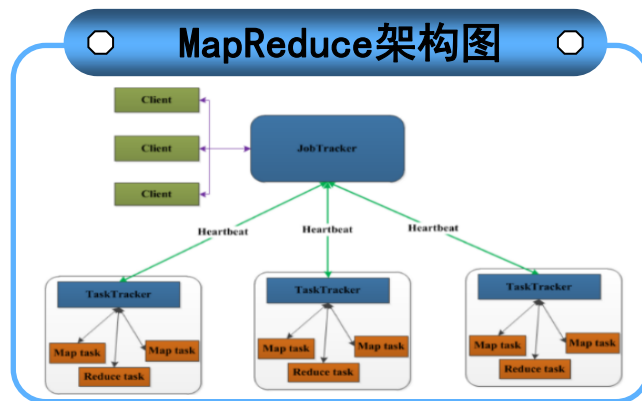


Hadoop1.0

- 加强了对安全方面的支持
- 提升HBase访问本地文件的性能，支持针对事务日志操作的sync和flush
- 改善兼容性和bug修复

存在不足

- 单点故障：JobTracker只有一个，JobTracker故障会导致整个集群瘫痪
- 扩展性受限：JobTracker接收其他JobTracker节点的RPC请求，压力大，限制集群扩展；随着节点规模增大，JobTracker成为瓶颈
- 计算框架支持受限：难以支持MR之外的计算框架



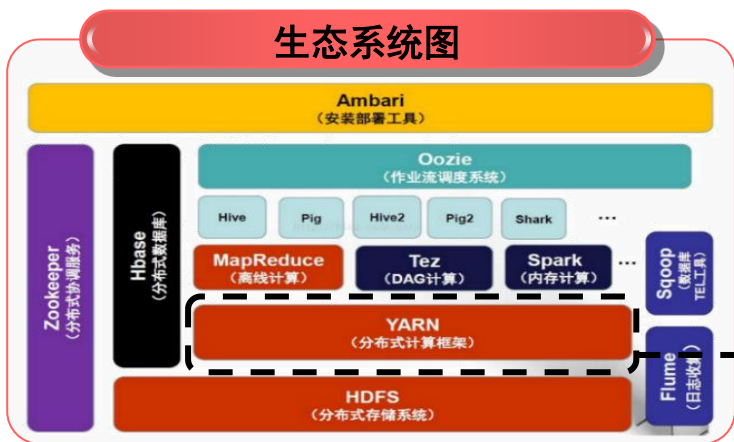


一. 大数据技术发展现状

□ 大数据技术创新日新月异

➤ 大数据处理平台

✓ 批处理：Apache Hadoop-2.0

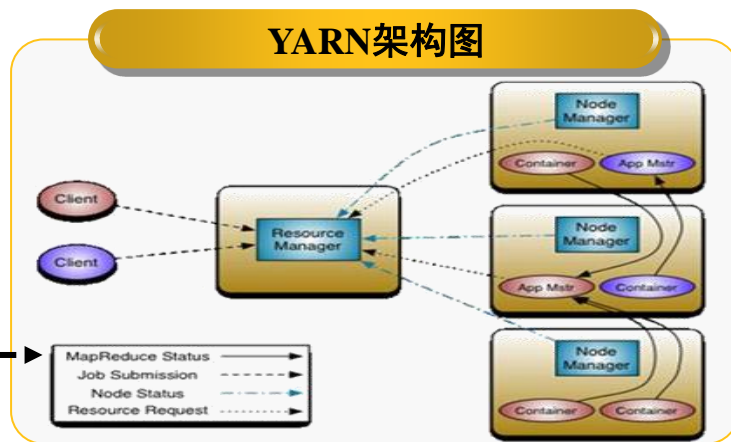


新特性

- 由HDFS、MapReduce和YARN三个分支构成
- 支持离线计算、DAG计算、内存计算等多种计算框架

优势

- 一个集群部署多个版本
- 计算资源按需伸缩
- 不同负载应用混搭，集群利用率高
- 共享底层存储，避免数据跨集群迁移



新特性

- ResourceManager
 - 处理客户端请求
 - 启动与监控ApplicationMaster，监控NodeManager
 - 资源分配与调度
- NodeManager
 - 单个节点上的资源管理
 - 处理ResourceManager与ApplicationMaster的命令
- ApplicationMaster
 - 数据切分，资源分配等



一. 大数据技术发展现状

□ 大数据技术创新日新月异

➤ 大数据处理平台

✓ 批处理：Apache Spark

文件/DSet → X → DSet/文件

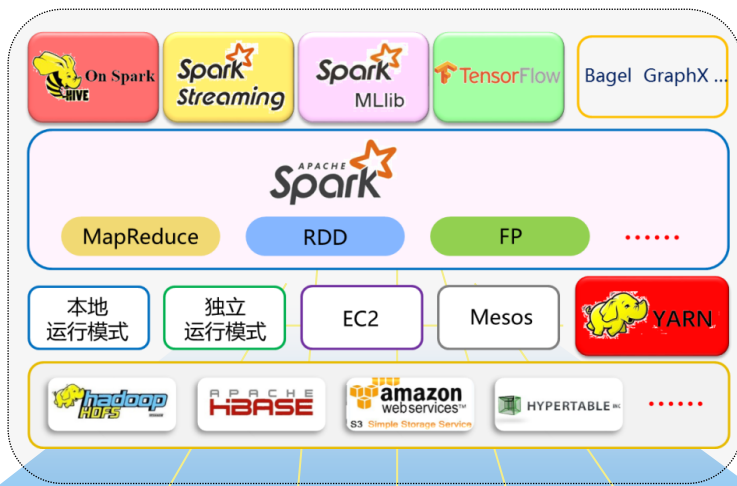
文件/RDD → Y → RDD/文件



- 围绕速度、易用性和复杂分析构建的大数据处理框架
- Job中间输出结果可以保存在内存中，速度可比Hadoop提升100倍
- 将数据的运算过程抽象为RDD的变换过程

Spark特性

- 目前提供了针对Scala、Java和Python的API，2.0版本提供针对其它语言（例如R）的支持
- 可以和Hadoop生态系统和数据源（HDFS、Amazon S3、Hive、HBase、Cassandra等）进行集成
- 可以运行在由Hadoop YARN或者Apache Mesos管理的集群上，也可以运行在单独的集群上
- Spark Streaming可模拟流计算，响应时间可达百毫秒



生态系统图



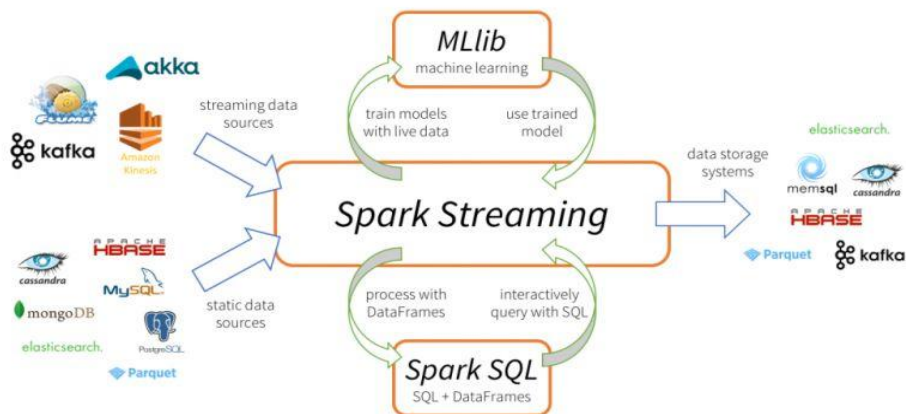
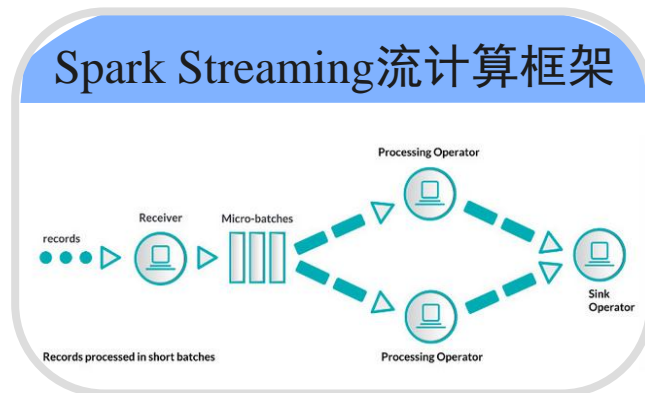
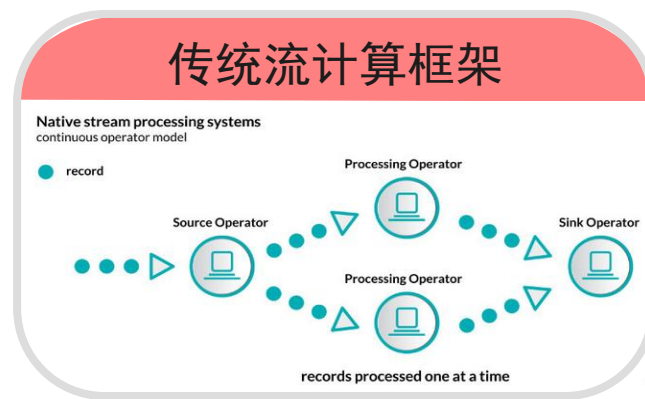
一. 大数据技术发展现状

□ 大数据技术创新日新月异

➤ 大数据处理平台

✓ 流计算：Spark Streaming

- 建立在Spark上的实时计算框架
- 将流式计算以时间片 Δt （秒级）分解成一系列短小的批处理（Micro-batch），本质上是“更细粒度的批处理”
- 每块数据作为一个RDD，使用RDD操作处理每一小块数据，每个块都会生成一个Spark Job处理，最终结果返回多块





一. 大数据技术发展现状

□ 大数据技术创新日新月异

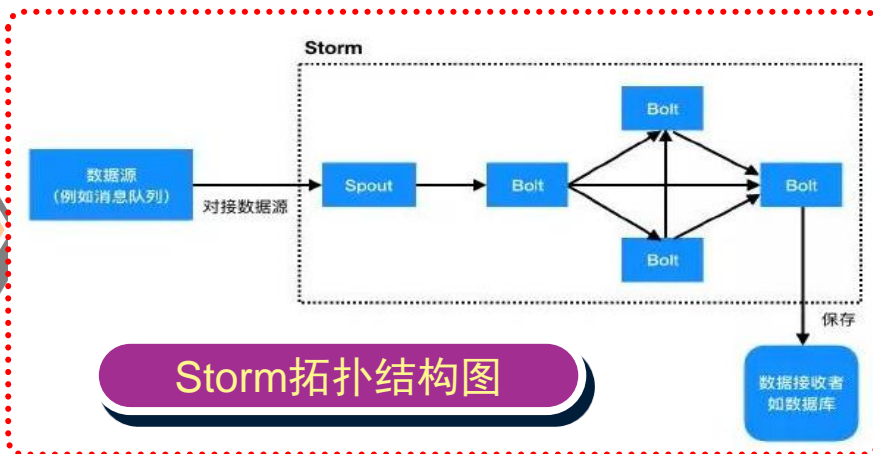
➤ 大数据处理平台

✓ 流计算：Apache Storm

- 由Twitter开源的分布式实时计算系统
- 本身无状态，通过Apache ZooKeeper管理
- 应用场景广泛，比如实时分析、在线学习、分布式RPC、ETL等
- 具有良好的扩展性、容错性，保证数据至少被处理一次



- Spouts（数据源头）将数据发射到一个或多个Bolts（数据处理单元）
- Bolt表示具有最小处理逻辑的节点，Bolts的输出可以发射到另一个Bolts作为输入
- Storm的主要工作是运行拓扑，并在给定时间运行任意数量的拓扑；拓扑始终运行，直到用户终止拓扑





一. 大数据技术发展现状

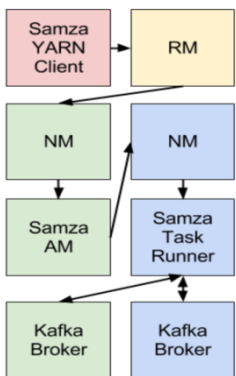
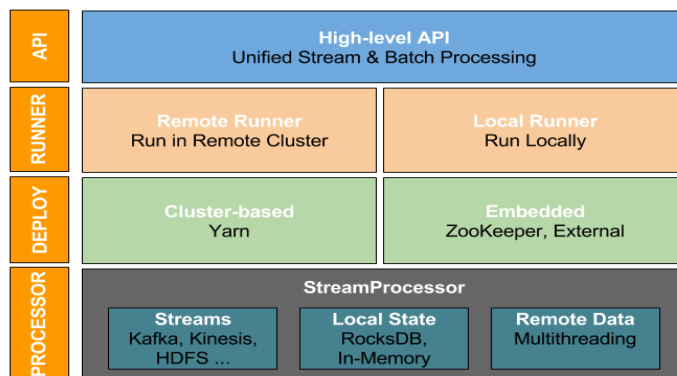
□ 大数据技术创新日新月异

➤ 大数据处理平台

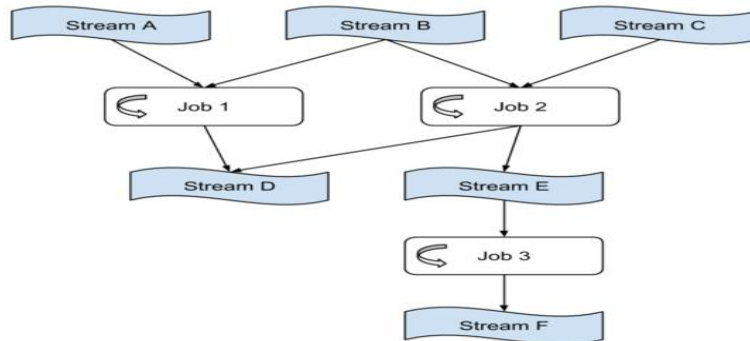
✓ 流计算：Apache Samza

Samza

- 一种与Apache Kafka消息系统紧密绑定的流处理框架
- 可通过Kafka提供容错、缓冲，以及状态存储
- 可使用YARN作为资源管理器，直接使用YARN丰富的内建功能；默认情况下需要具备Hadoop集群（至少具备HDFS和YARN）



- 数据流输入到Kafka的Brokers。
- YARN启动并且监控一个或者多个Samza Containers，业务处理逻辑代码运行在这些容器里，处理结果输出到Kafka的Brokers





一. 大数据技术发展现状

□ 大数据技术创新日新月异

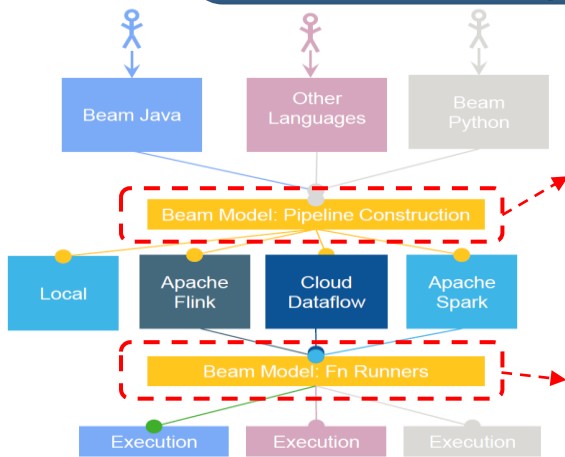
➤ 大数据处理平台

✓ 混合批流处理：Apache Beam



- 基于谷歌的Dataflow模型创建的项目
- 按照如何对数据进行计算、何时将计算结果输出、迟到数据如何处理四个维度调用具体的API
- 将流式数据当作一个无穷数据集，将离线数据当作一个有穷数据集

Beam架构图



针对不同的编程语言构建一组用于定义Pipeline相关抽象，提供编程API，被称为Beam SDKs。用户可以基于这些抽象的Beam SDK来构建数据处理Pipeline。

用来对接底层的计算引擎，执行上层用户开发好Pipeline程序。



- 统一：实现批处理、流处理统一
- 可移植：在底层计算平台升级的过程中无需重写上层已定义的数据Pipeline
- 可扩展：实现任意可以共享的Beam SDK、IO connector、Transform库



一. 大数据技术发展现状

□ 大数据技术创新日新月异

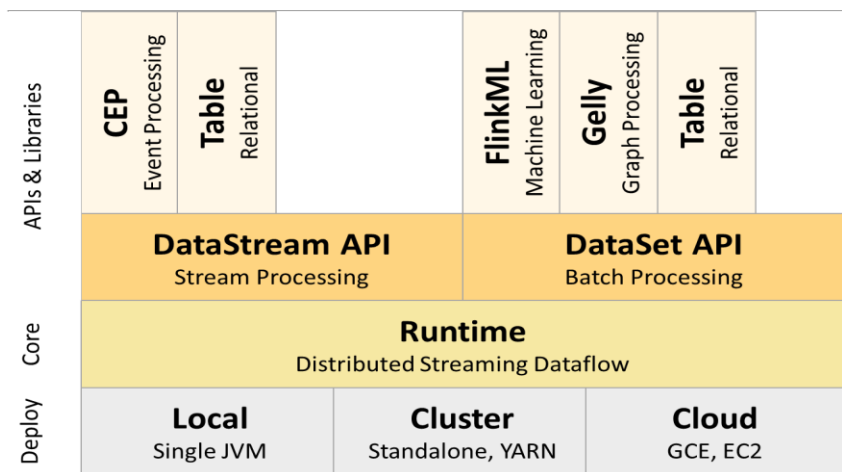
➤ 大数据处理平台

✓ 混合批流处理：Apache Flink



Flink

- 一个面向分布式数据流处理和批量数据处理的开源计算平台
- 本质是一个流式计算引擎，运行时分别搭建了批处理(DataSet)和流式计算(DataStream)的编程接口和相配套的生态系统
- 支持Redis, Kafka, Casandra, Elasticsearch, Flume, Rabbitmq, Twitter, HDFS等数据存储形式



Flink生态圈

- 支持本地、集群、云等多种部署模式
- 运行时以JobGraph形式接收程序，DataStream API和DataSet API都会使用单独编译的处理方式生成JobGraph
- DataStream API和DataSet API支持Java、Scala和Python
- 提供了类似SQL的查询，图形化处理，以及机器学习库

- 一 大数据技术发展现状
- 二 未来数据计算发展趋势



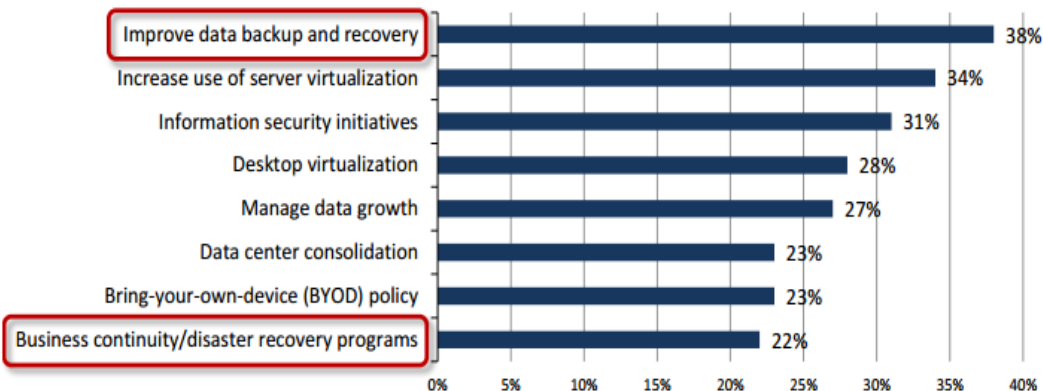
二. 未来数据计算发展趋势

□ 跨域数据计算(Geo-distributed Data Computing)

- 异地容灾是重大战略需求

业务连续性与灾难恢复建设是重要任务

Top eight most important IT priorities for midmarket organizations (100 to 999 employees) over the next 12 months. (Percent of respondents, N=213, ten responses accepted)



Source: Enterprise Strategy Group, 2015

发生重大灾难后，如果不能在**24小时**之内恢复数据访问，**40%**的公司将走向破产。

Gartner

调查问卷显示，只有**31%**的CIO认为他们的业务连续性计划是**行之有效的**。

IDG

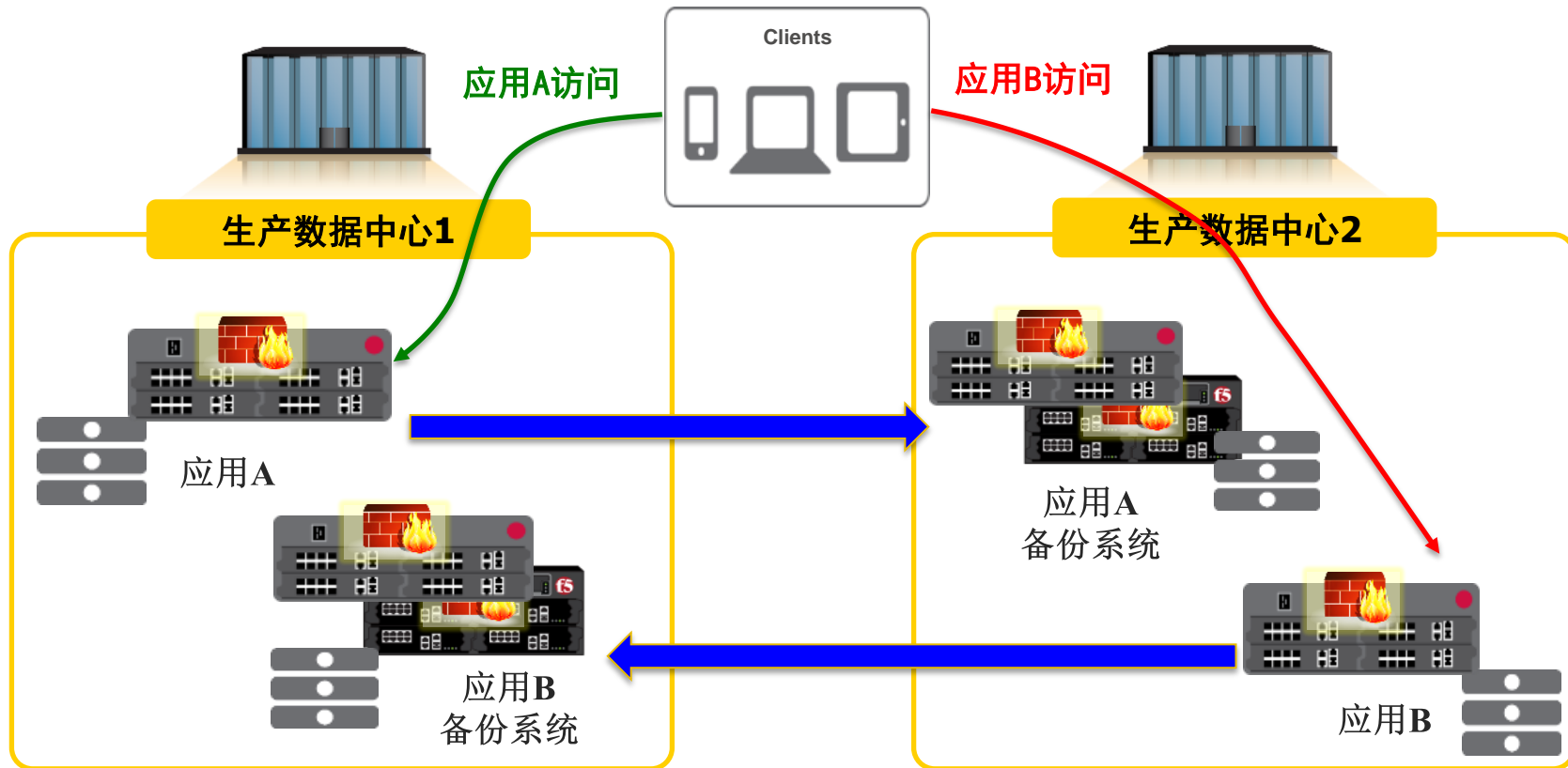


二. 未来数据计算发展趋势

跨域数据计算(Geo-distributed Data Computing)

➤ 异地容灾是重大战略需求

✓ 异地容灾双活模式：**主备模式，数据容灾**



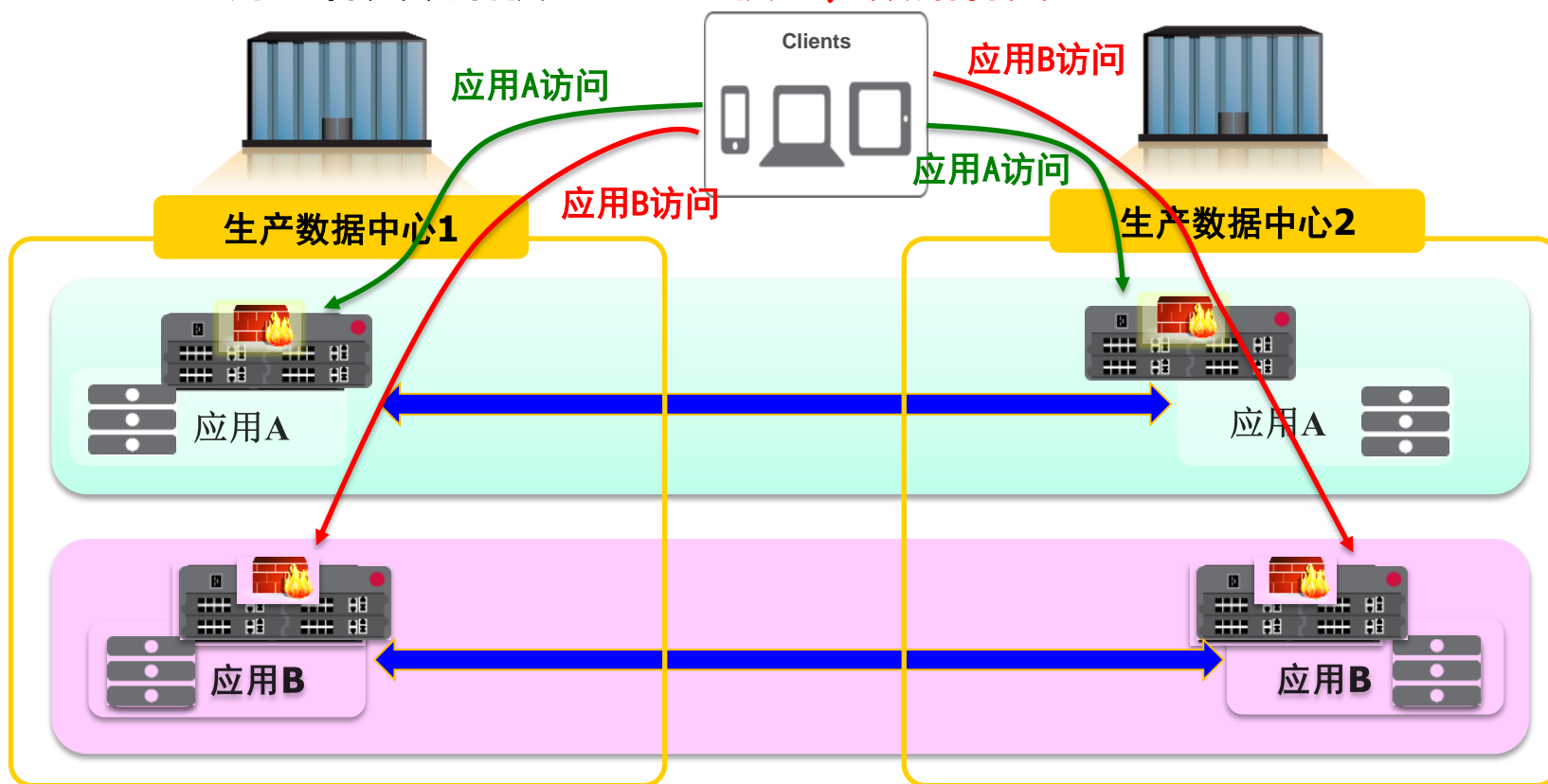


二. 未来数据计算发展趋势

跨域数据计算(Geo-distributed Data Computing)

异地容灾是重大战略需求

- ✓ 异地容灾双活模式：**主主模式，数据容灾**





二. 未来数据计算发展趋势

跨域数据计算(Geo-distributed Data Computing)

地理分布的数据中心是大数据企业的现实部署



- ✓ Microsoft和Google在全球有数十个数据中心
- ✓ Google在全球有1500+边缘数据中心
- ✓ 数据中心之间的通信和数据交换是跨域的

跨域数据计算是大数据企业的重大需求



用户日志分析：广告决策



网络日志分析：DoS攻击检测



系统日志分析：故障诊断、容错分析、性能优化



二. 未来数据计算发展趋势

□ 跨域数据计算(Geo-distributed Data Computing)

➤ 传统做法是集中处理

代价高昂



网络空间疆域的管理



异构巨大的网络延迟

```
bytes=32 time=247ns TTL=124
bytes=32 time=172ns TTL=124
bytes=32 time=579ns TTL=124
bytes=32 time=1204ns TTL=124
bytes=32 time=1827ns TTL=124
bytes=32 time=1497ns TTL=124
bytes=32 time=421ns TTL=124
bytes=32 time=136ns TTL=124
bytes=32 time=286ns TTL=124
```

➤ 跨域数据计算的前沿技术

- ✓ 跨区域数据副本复制策略：提高可用性、容错能力，降低网络延迟
- ✓ 跨域数据存储技术
- ✓ 数据一致性技术：因果一致性技术和并行快照隔离技术
- ✓ 地理分布的查询处理和优化技术
- ✓ 域间作业与任务调度技术.....



二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

➤ 物联网时代大数据处理需要边缘计算



- ✓ IDC的统计数据显示，到2020年全球将有超过500亿的终端和设备联网。我们预计，2016年我国物联网连接数约8.4亿个，预计2020年将增长317%，达到35亿



- ✓ 根据IDC的预计，在2018年将有40%的数据需要在网络边缘侧分析、处理与储存



- ✓ “移动边缘计算”作为一种将计算、网络、存储能力从云延伸到物联网网络边缘的架构，遵循“业务应用在边缘，管理在云端”的模式



二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

➤ 移动边缘计算是5G的核心技术之一



- ✓ 5G技术：大容量、大带宽、大连结、低延迟、低功耗
- ✓ 一段网红的短视频约为10MB，如果一个区域内有1000个人观看这段视频就会产生10GB的网络流量。在这过程中，实际上这段视频内容从互联网到移动网络内被重复发送了1000遍，99.9%的网络带宽被浪费了，如果将视频缓存在靠近边缘侧的节点，将大量节省带宽。
- ✓ 业务完全在云端终结并非完全有效





二. 未来数据计算发展趋势

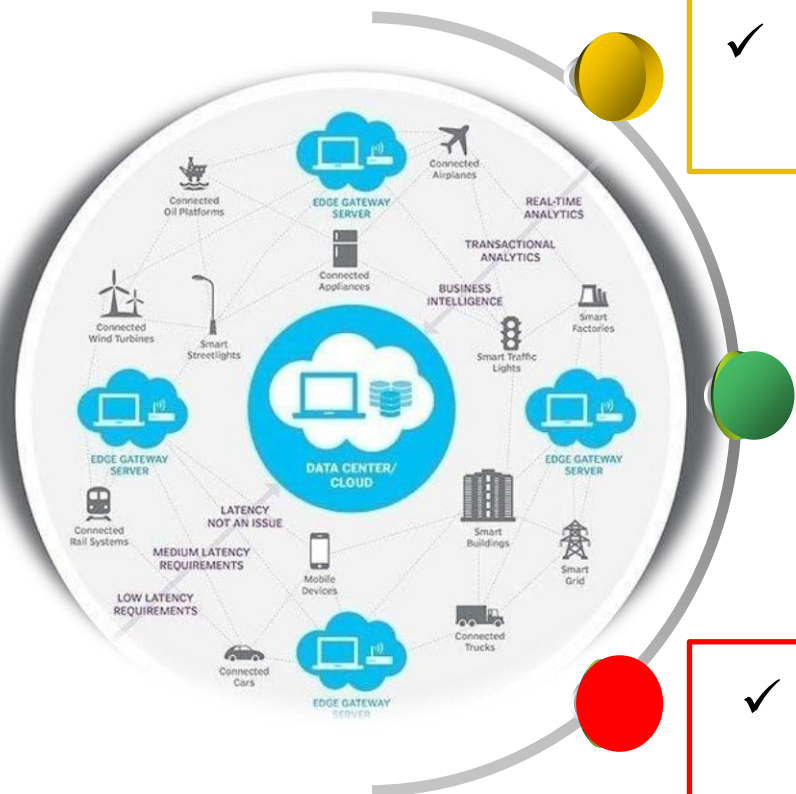
□ 边缘数据计算(Edge Data Computing)

➤ 边缘计算是云计算的有效补充

✓ 边缘计算可以配合云计算进行数据的预处理、特征提取等功能

✓ 能在边缘处理的业务尽量在边缘处理，边缘存储、数据和计算是离用户最近的地方

✓ 边缘计算和云计算的关系好像集团公司与地方办事处

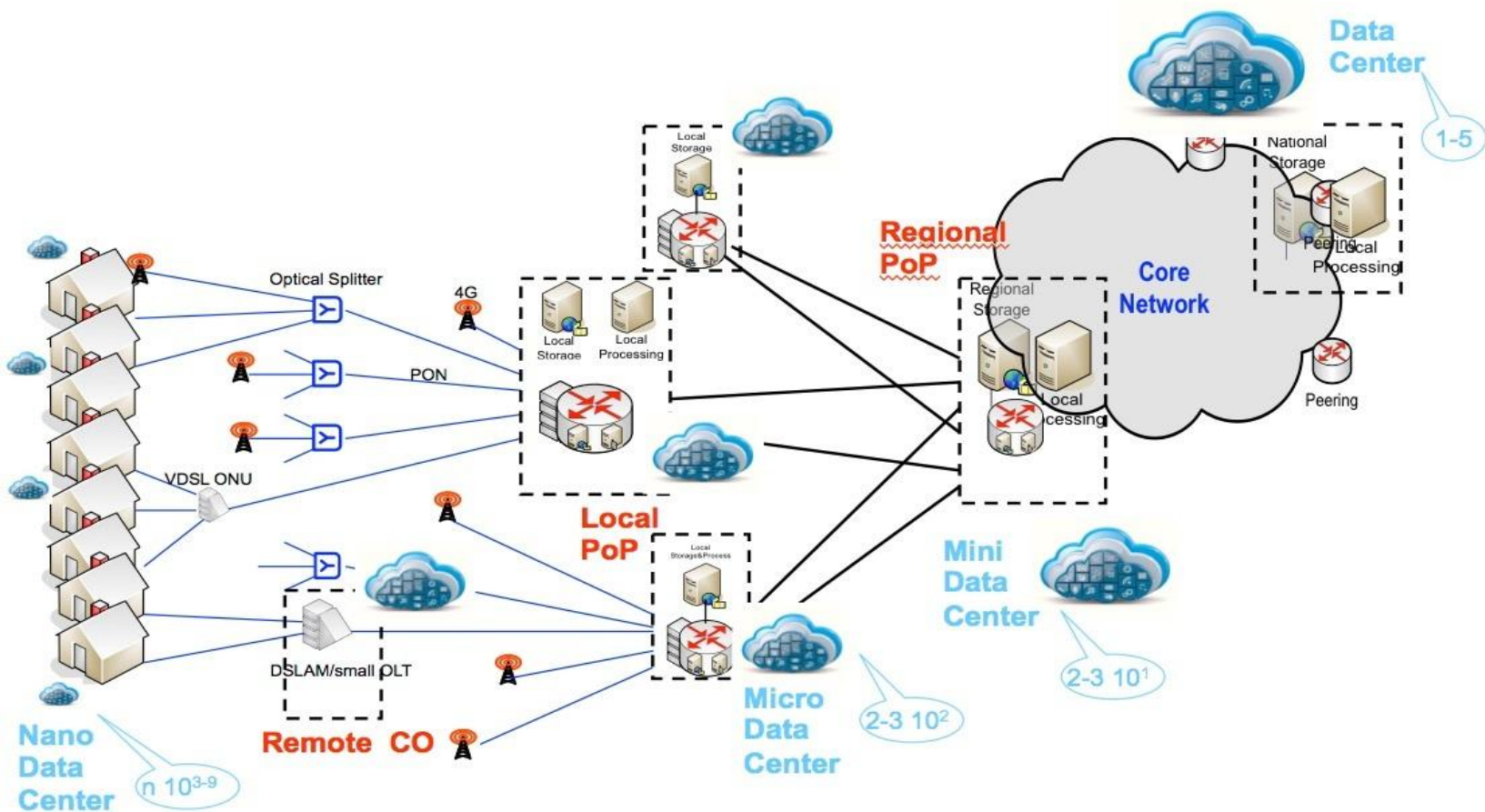




二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

➤ 边缘计算是云计算的有效补充





二. 未来数据计算发展趋势

边缘数据计算(Edge Data Computing)

边缘数据计算的研究内容

研究内容

数据缓存技术：静态内容

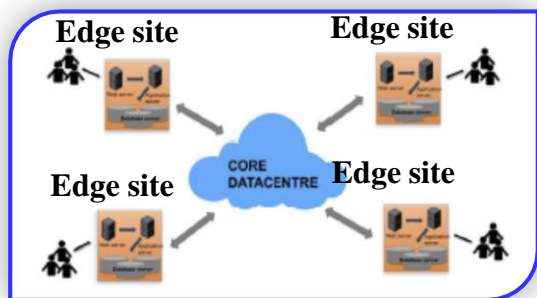
边缘计算体系架构研究

数据复制策略：Lazy、主副本策略

一致性保障机制

数据路由策略 (routing)

查询处理与优化技术……



ITEMS						
id	quantity	reserve_price	nb_of_bids	max_bid	seller	category
5	1	5382	12	4833	995858	5

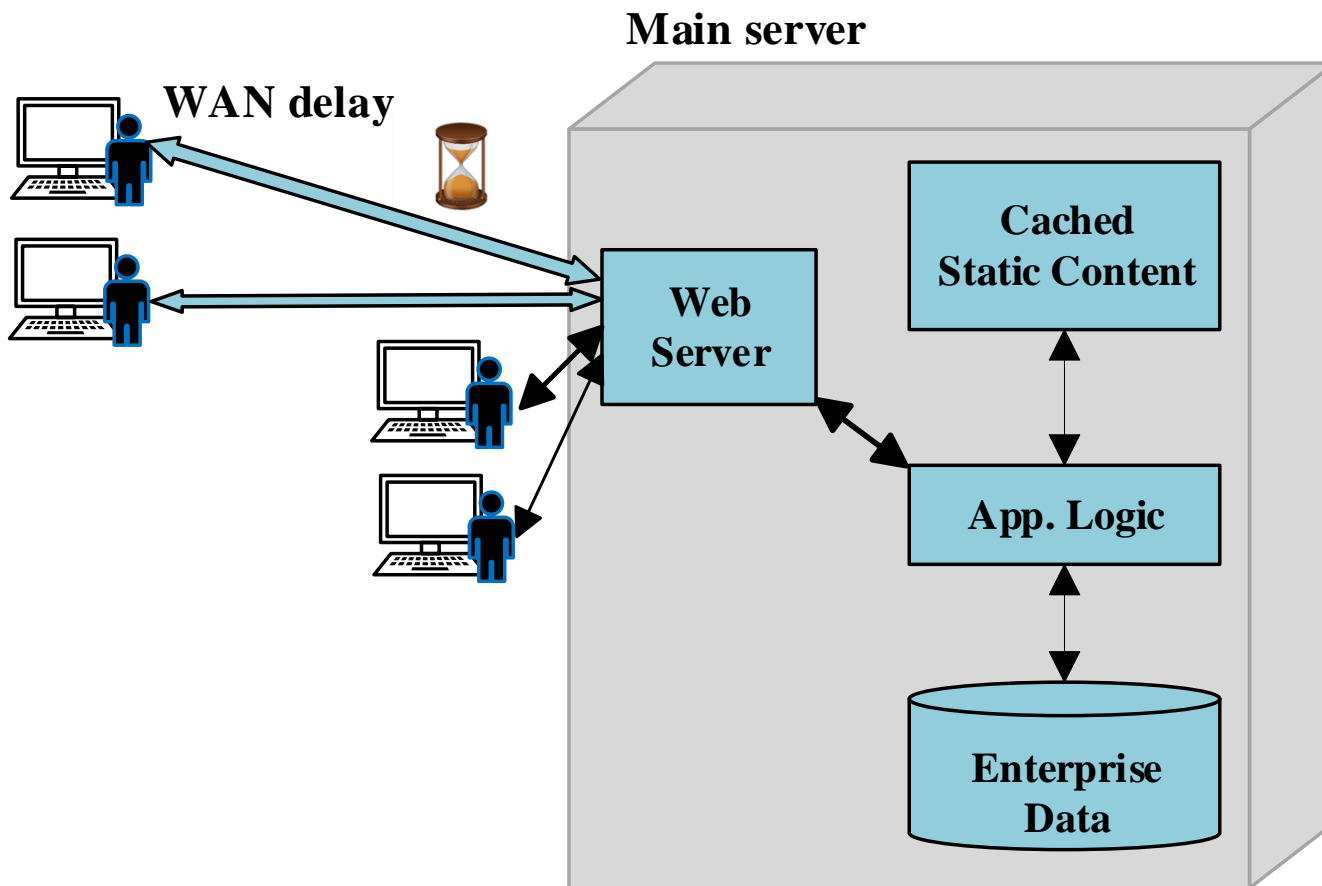
BIDS						
id	user_id	item_id	qty	bid	max_bid	date
23	607382	3	1	4771	4776	2014-10-10 09:48:09
24	842589	3	1	4776	4785	2014-10-10 09:48:09
25	358294	3	1	4787	4796	2014-10-10 09:48:09
26	466670	3	1	4793	4799	2014-10-10 09:48:09
27	674629	3	1	4882	4811	2014-10-10 09:48:09
28	969515	3	1	4889	4836	2014-10-10 09:48:10
29	756439	3	1	4812	4815	2014-10-10 09:48:10



二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

- 边缘数据计算体系结构：集中式

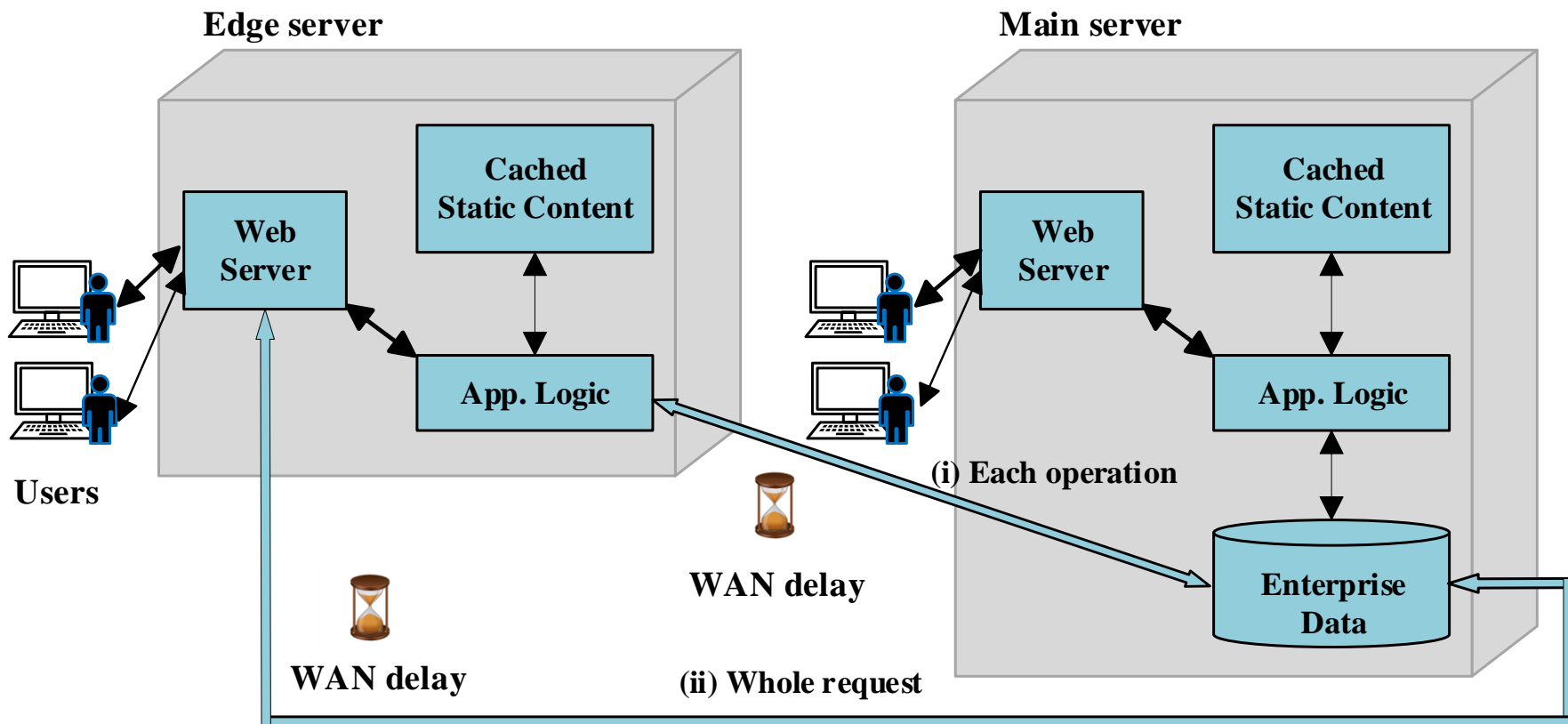




二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

➤ 边缘数据计算体系结构：静态数据缓存

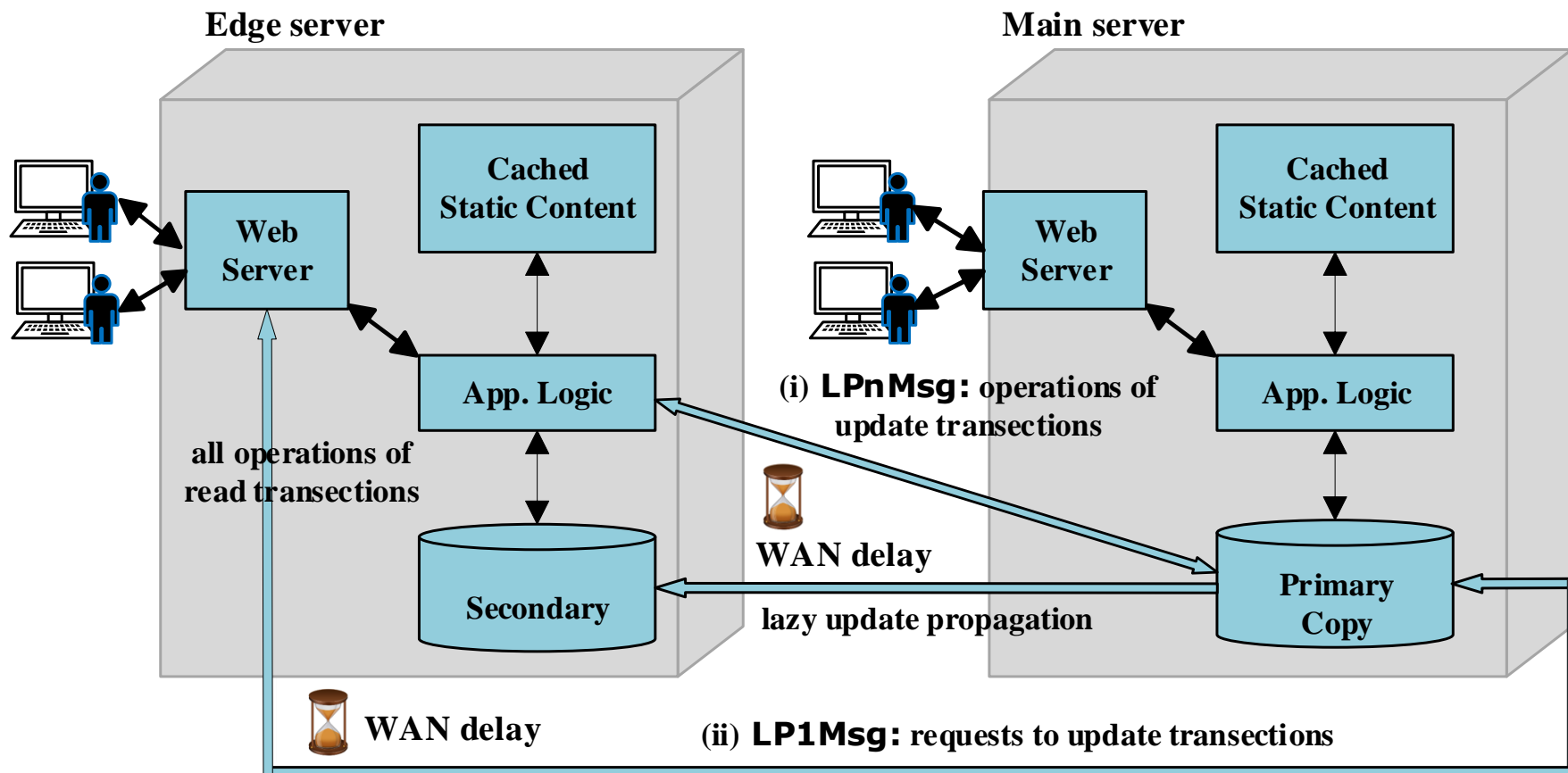




二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

- 边缘数据计算体系结构：动态数据复制（Lazy一致性）

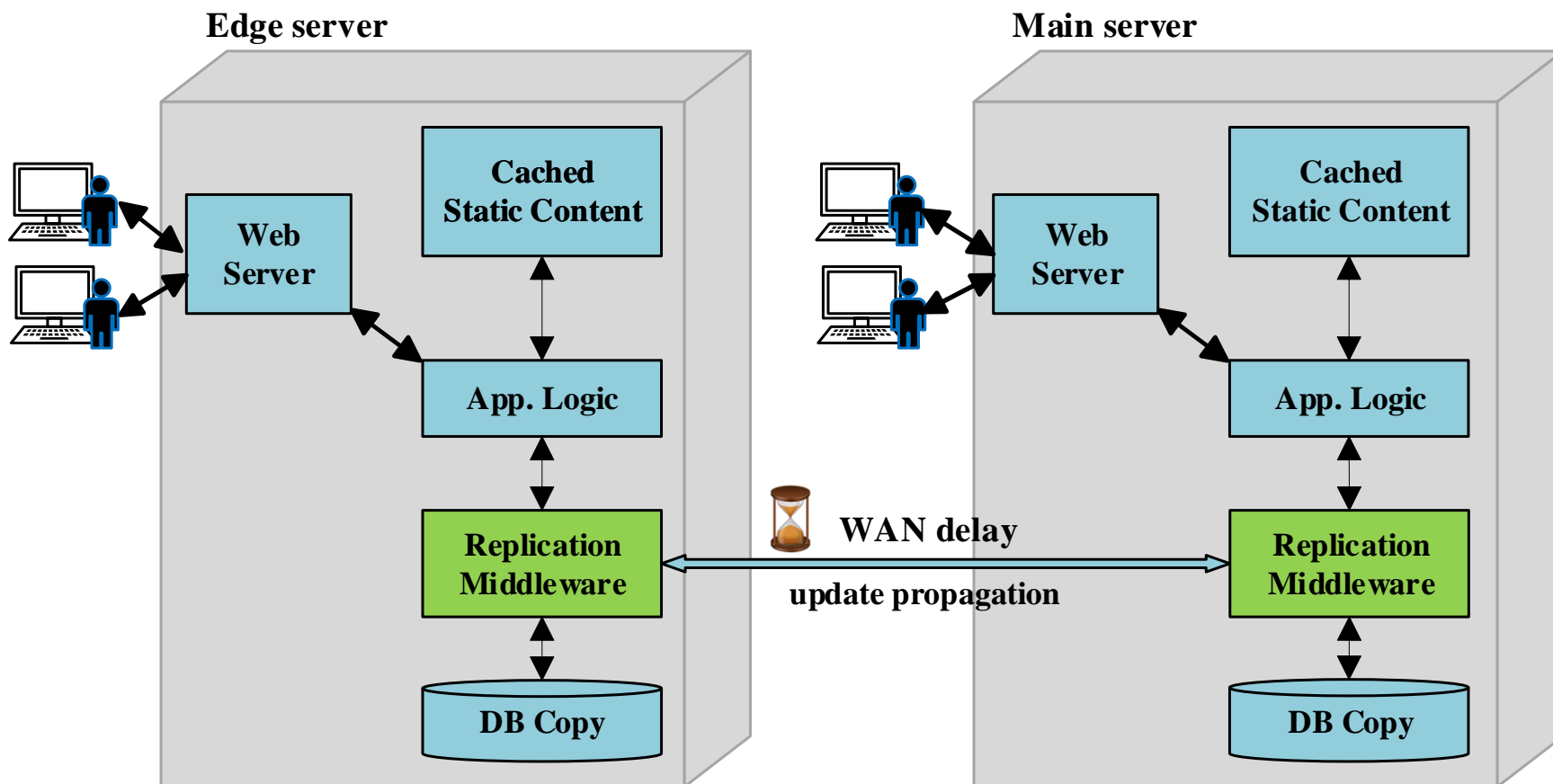




二. 未来数据计算发展趋势

□ 边缘数据计算(Edge Data Computing)

- 边缘数据计算体系结构：动态数据复制（强一致性）





二. 未来数据计算发展趋势

□ 智慧数据计算(Smart Data Computing)

➤ 感知数据特征

多元异构

- 到2020年将有260亿个传感设备连接上网 (Garter 报告, 2014)
- 智慧汽车、智慧城市、智慧农庄、环境与气候监测、健康穿戴设备、GPS、ZigBee等
- 类型包括数值、文字、图像、音频、视频等

巨量动态

- 实时、高速、连续的数据流

数据特征

弱语义性

- 原始的感知数据具有弱语义性：原始感知数据层次比较低，数据之间通常只有时空关联性，高层语义是从底层时空关联语义中通过事件进行抽象的

数据劣质

- 通常的感知系统仅能采集到1/3的正确数据



二. 未来数据计算发展趋势

智慧数据计算(Smart Data Computing)

物联网数据管理环境的特点与挑战



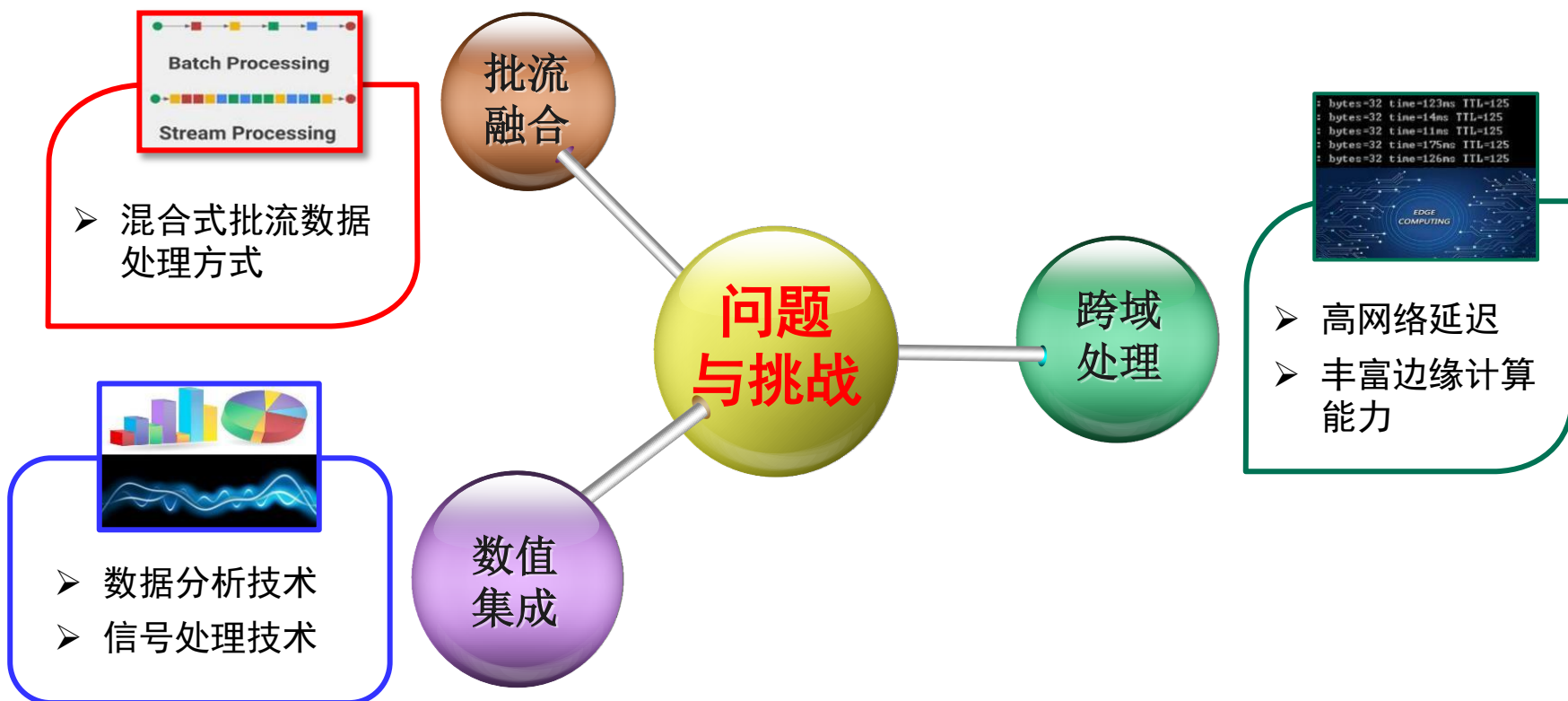
如何提供优质的用户体验带来非常大的挑战



二. 未来数据计算发展趋势

智慧数据计算(Smart Data Computing)

值得关注的物联网数据管理环境的特点与挑战

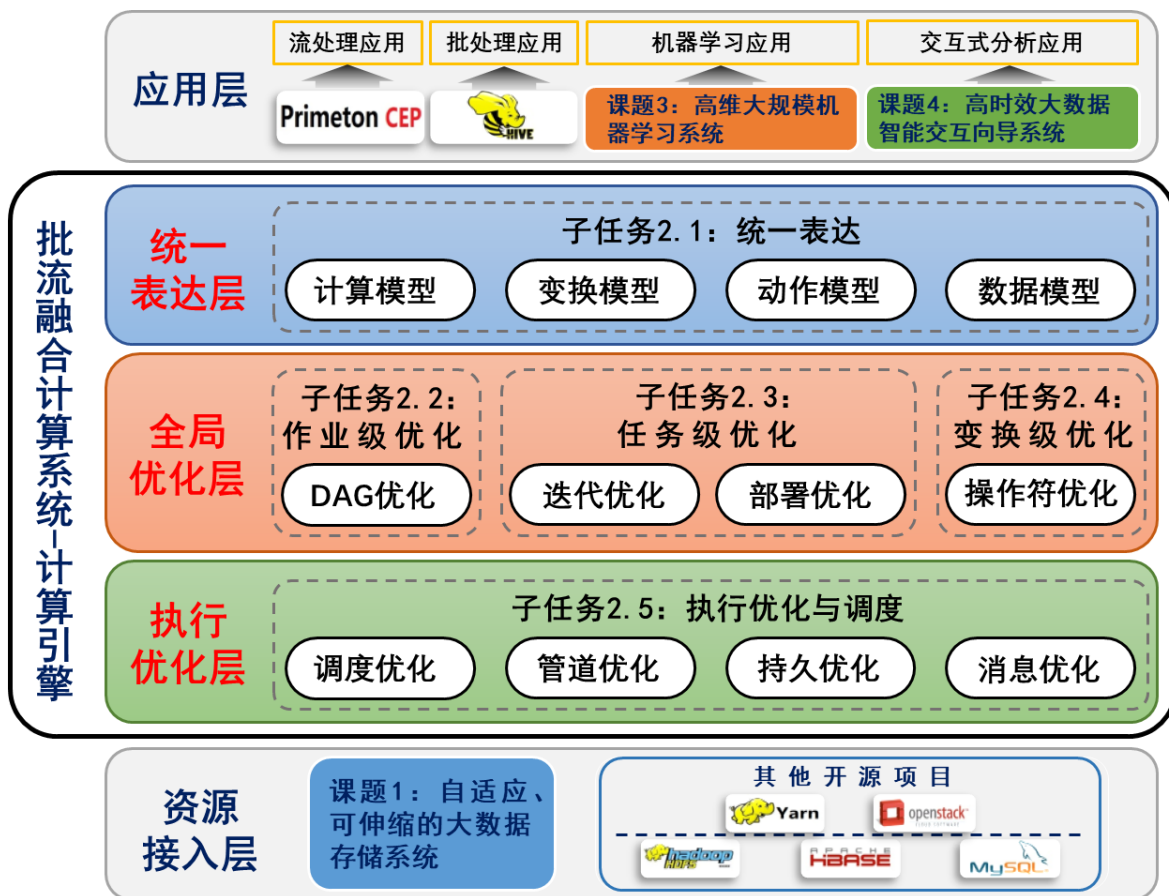




二. 未来数据计算发展趋势

智慧数据计算(Smart Data Computing)

批流融合



- ✓ 批流融合计算框架是新趋势：统一数据模型、统一操作模型、统一计算模型、统一用户接口
- ✓ 三大优点：方便用户使用、简化系统部署、带来优化机会



二. 未来数据计算发展趋势

□ 智慧数据计算(Smart Data Computing)

➤ 数值集成

- 物联网传感器的不稳定性等带来严重的数据质量问题
- DSP（数值计算方法）是当前物联网数据解决质量问题的主流技术
 - 有限差分方法（FDM）：丢值
 - 快速傅里叶变换（FFT）：丢值
 - 数字滤波技术，如小波变换等：去噪
 -
- DSP处理算法的数据模型是数组（Array）
- 物联网传感器的不稳定性等带来严重的数据质量问题
- DSP处理算法都是面向领域的
- 数值计算工具：MATLAB、R语言等

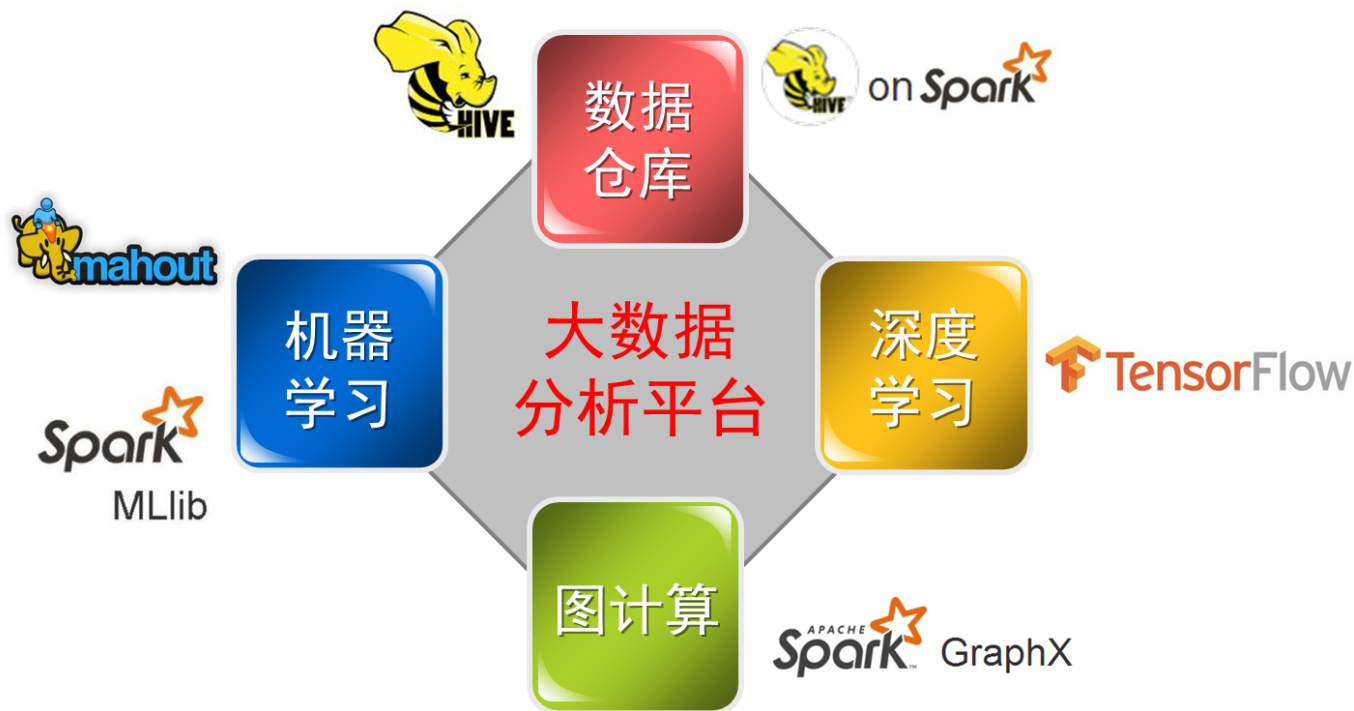


二. 未来数据计算发展趋势

□ 智慧数据计算(Smart Data Computing)

➤ 数值集成

- 数据分析是另一个物联网数据处理的重大需求
- 大数据分析技术：通用架构





二. 未来数据计算发展趋势

□ 智慧数据计算(Smart Data Computing)

➤ 数值集成

- 物联网数据处理量大需求：数据分析+信号处理
- 数据模型不同：关系模型、数组模型
- 处理架构不同：
 - 数据分析技术：通用架构，利用优化技术
 - 信号处理技术：专用架构，利用领域知识
- 为了数据存储需要统一两种处理方式
 - 松散耦合：数据转换带来巨大的性能问题
 - 深度集成：是数值集成的未来方向，统一数据模型并同时有效支持分析操作和信号处理算法

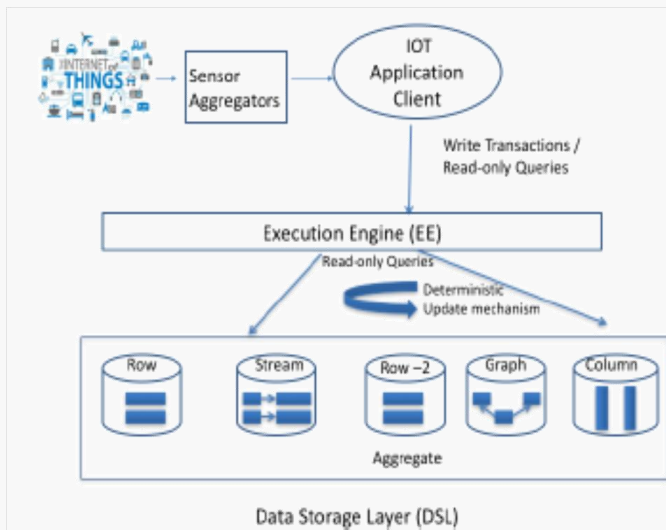


二. 未来数据计算发展趋势

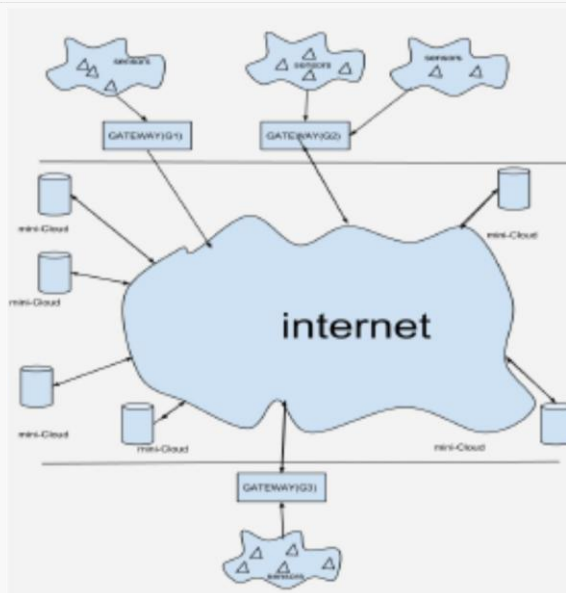
智慧数据计算(Smart Data Computing)

物联网数据计算案例

多表达的数据存储

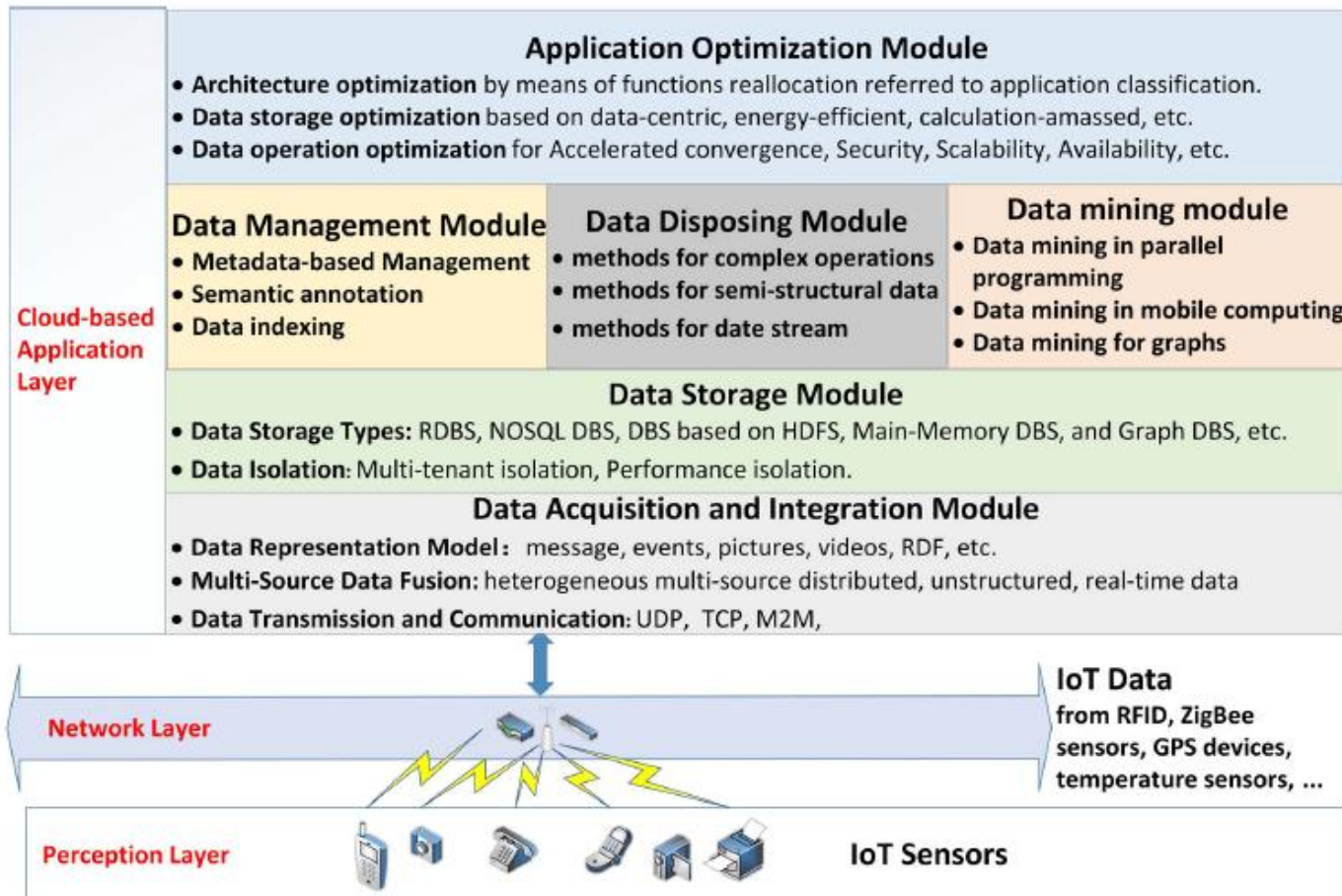


物联网数据复制技术





□ 边缘计算数据管理





工业革命的变革历程

□ 工业革命的变革历程

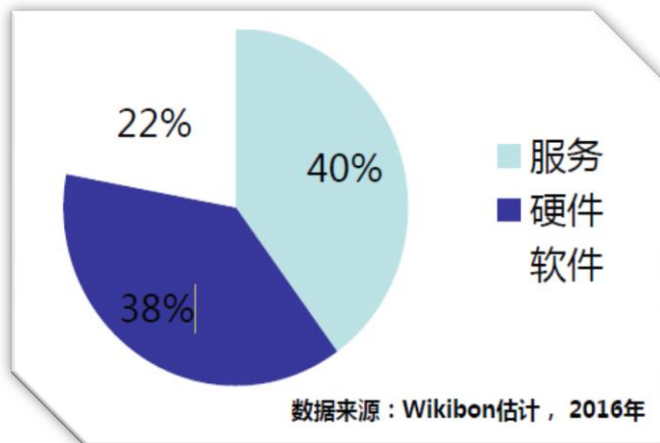
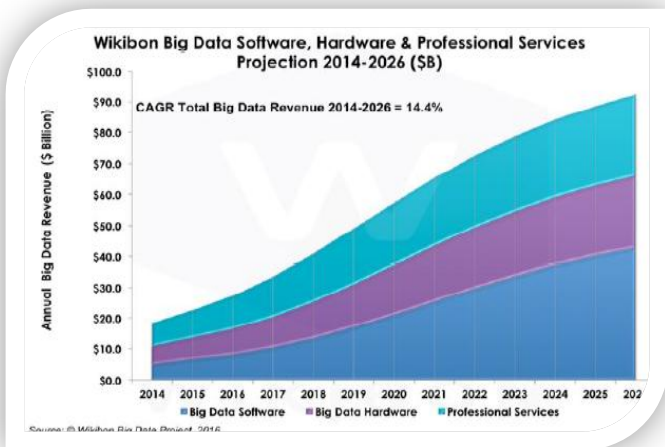




一. 大数据技术发展现状

□ 大数据发展概况—产业收益

➤ 狭义大数据产值口径：专门用于大数据的软件、硬件及服务，300亿美元



➤ 大数据应用于各个行业，产生的外部价值受到各国重视

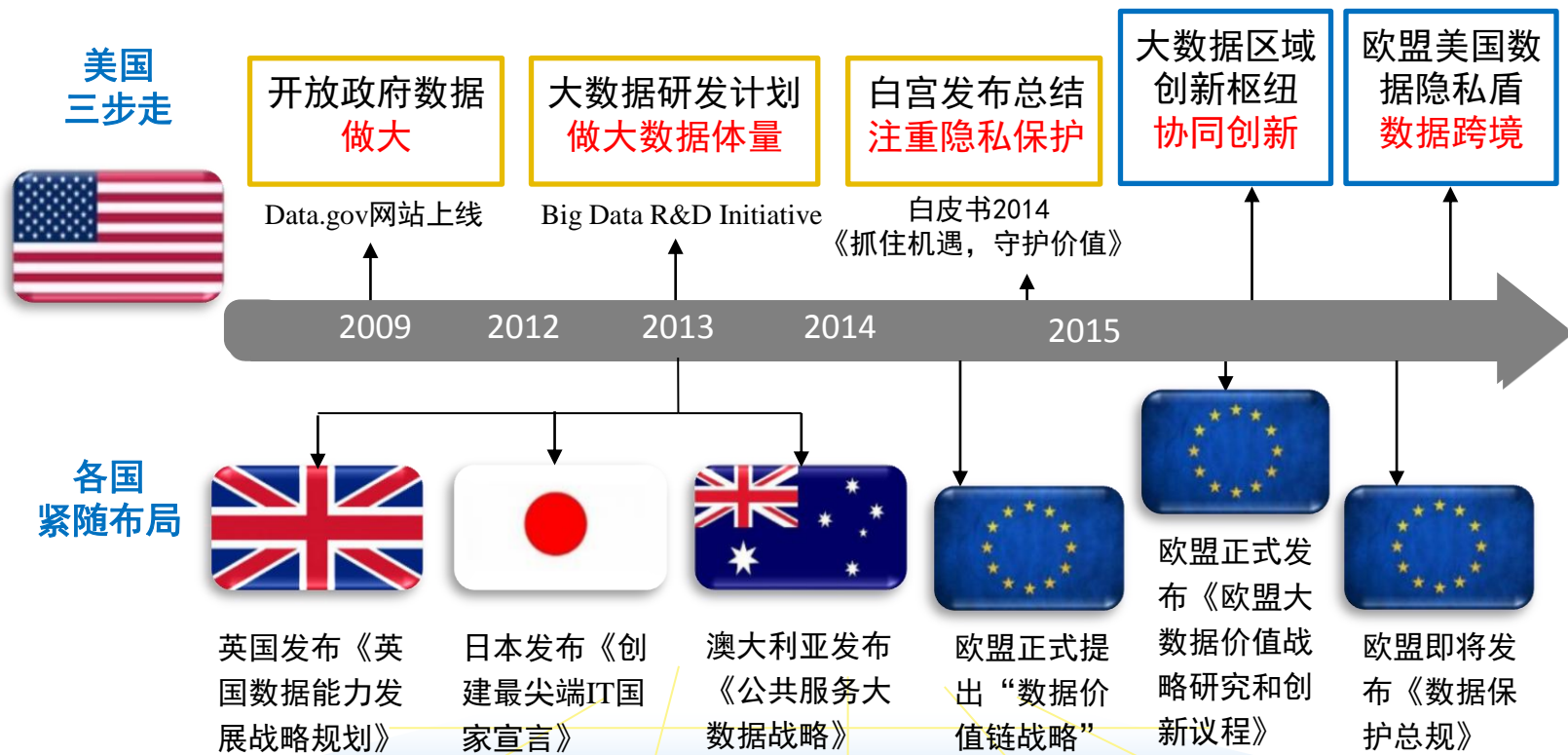
The EU economy will grow by an **additional 1.9% by 2020** thanks to big and open data

区域	测算机构	带动GDP增
欧盟	华沙经济研究所	1.9% (2020)
美国	麦肯锡	2%-4% (2020)
英国	CEBR	2.3% (2017)
中国	CAICT	0.6%-0.85% (2013)



一. 大数据技术发展现状

□ 大数据发展概况—各国政策



各国从推进数据开放、支持技术创新、打造协作环境等方面多措并举，促进大数据快速发展



一. 大数据技术发展现状

云计算发展概况—国外发展

- 2016年以IaaS、PaaS和SaaS为代表的典型公有云服务市场规模达到710.26亿美元，增速35.97%。市场份额进一步向巨头企业集中。

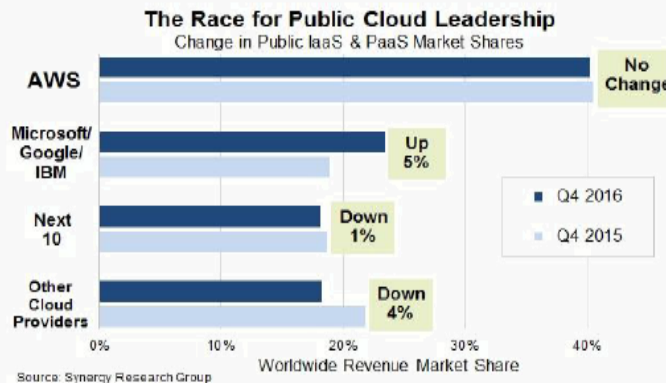
市场规模



数据来源: Gartner 2017年3月

预计全球公有云市场规模2020年将达到1620.84亿美元，年复合增长率达20%

市场份额



数据来源: Synergy 2017年2月

公有云 (IaaS、PaaS) 的集中度进一步增大。其中AWS稳占40%全球市场份额，第二梯队微软、谷歌和IBM进一步巩固地位



一. 大数据技术发展现状

云计算发展概况—国内发展

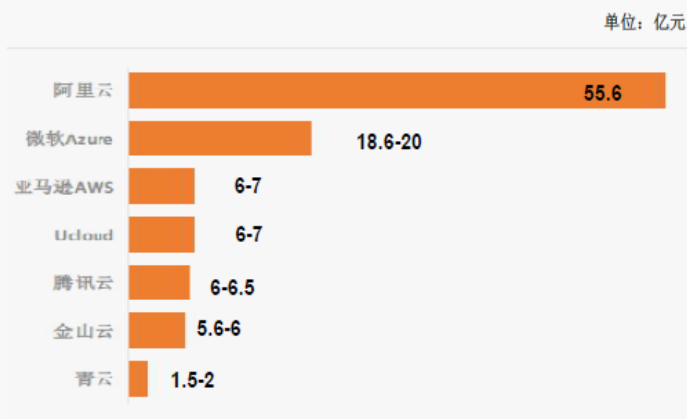
- 2016年我国公有云计算市场规模为147.6亿元，私有云为345.8亿元，同比增长分别为44.1%和25.5%。阿里云已占我国云计算市场的主导，远远领先跟随者。

市场规模



“十二五”以来我国云计算高速增长，公有云以40%以上复合增长率增长，私有云以25%以上增长

市场份额

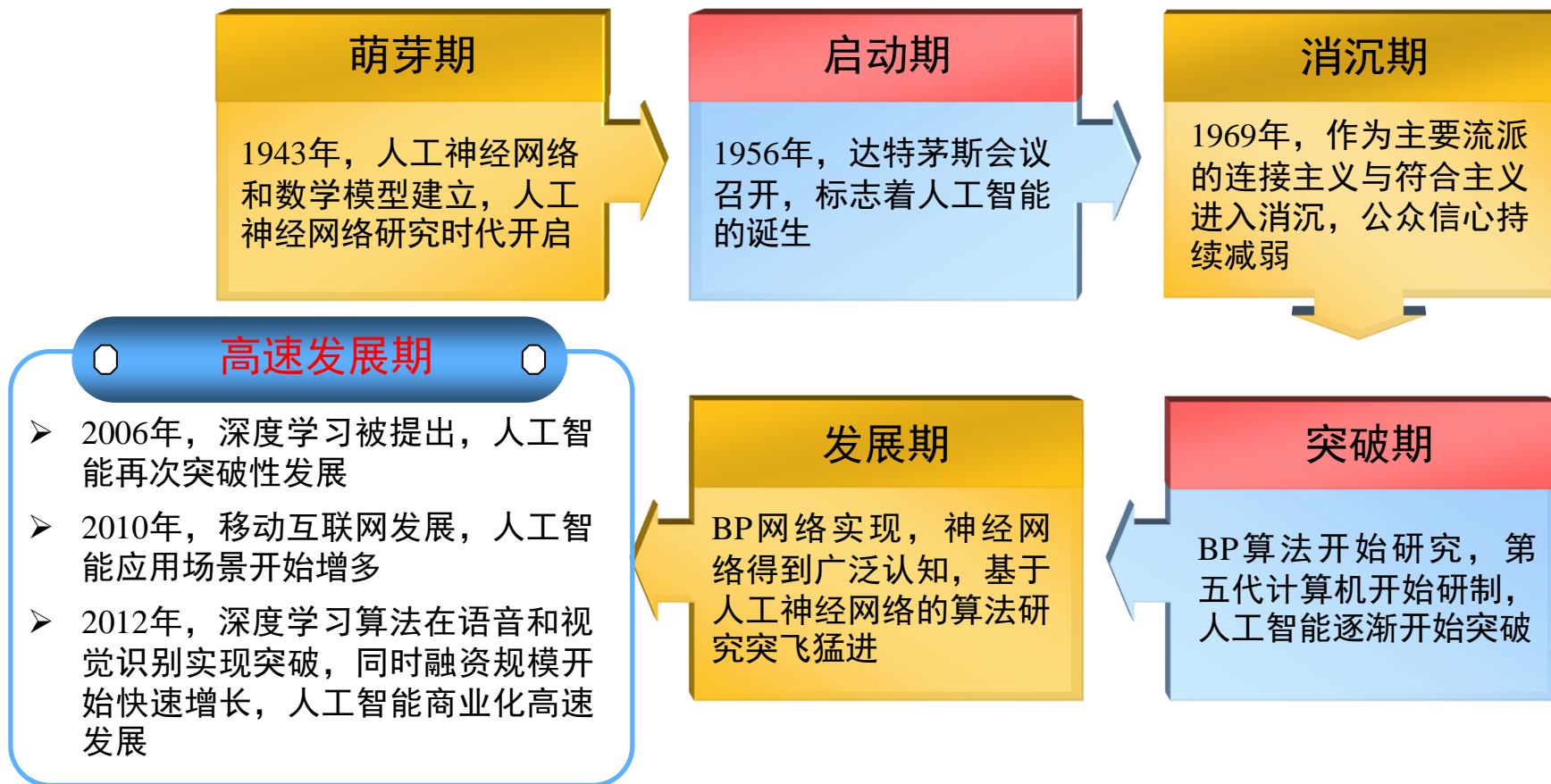


在中国的云计算市场中，阿里巴巴已经成为占主导地位。一家营收比之后6名的总和还多



一. 大数据技术发展现状

□ 人工智能发展概况—发展历程





一. 大数据技术发展现状

人工智能发展概况—高速发展期

全球人工智能融资额持续增长



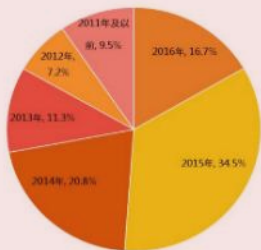
人工智能学术研究处于爆发期



高速发展

人工智能发展潜力巨大

中国人工智能创业公司成立年份分布



中国人工智能核心产业规模规划



人工智能领域生态繁荣

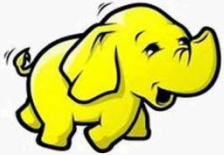




一. 大数据处理平台

□ 批处理框架——优势与局限

hadoop



- 可以处理非常海量的数据集，可以在廉价硬件上运行
- 生态系统辽阔，成为使用不同技术的多种工作负载处理平台的底层基础
- 严重依赖持久存储，需要多次执行读取和写入操作，速度相对较慢

APACHE
Spark

- 在内存计算策略和先进的DAG调度等机制的帮助下，更快速处理数据集
- 可为机器学习、交互式查询等任务提供更好的支持
- 资源消耗更大，可能会对需要在同一时间使用集群的其他任务产生影响

在应对大量持久数据方面的表现极为出色，经常被用于对历史数据进行分析

大数据批处理



Batch Processing

大量数据的处理需要付出大量时间，不适合对处理时间要求较高的场合

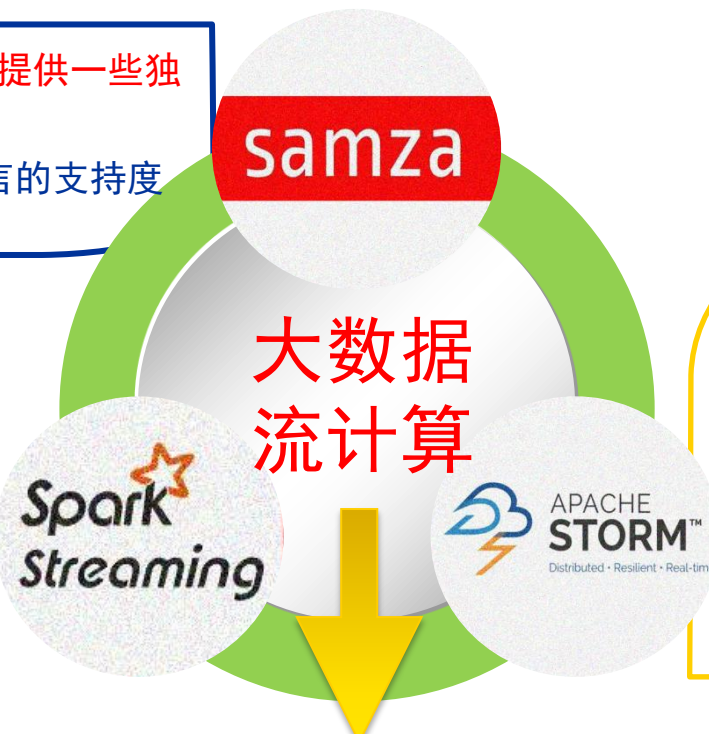


一. 大数据处理平台

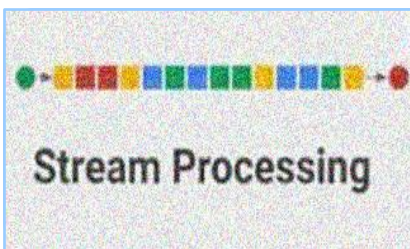
□ 流计算框架—优势与局限

- 与Kafka绑定，可以为系统提供一些独特的保证和功能
- 成熟度较低，对平台和语言的支持度不如其他平台灵活

- 使用Spark作为执行引擎，生态系统广阔，具有高效和容错的特性
- 不适合处理对延迟有较高要求的工作负载，相比真正的流处理框架性能弱



- 延迟极低，近实时处理领域的最佳解决方案
- 与Hadoop的YARN资源管理器进行集成此可以很方便地融入现有Hadoop部署
- Core Storm无法保证消息的处理顺序，可能发生重复

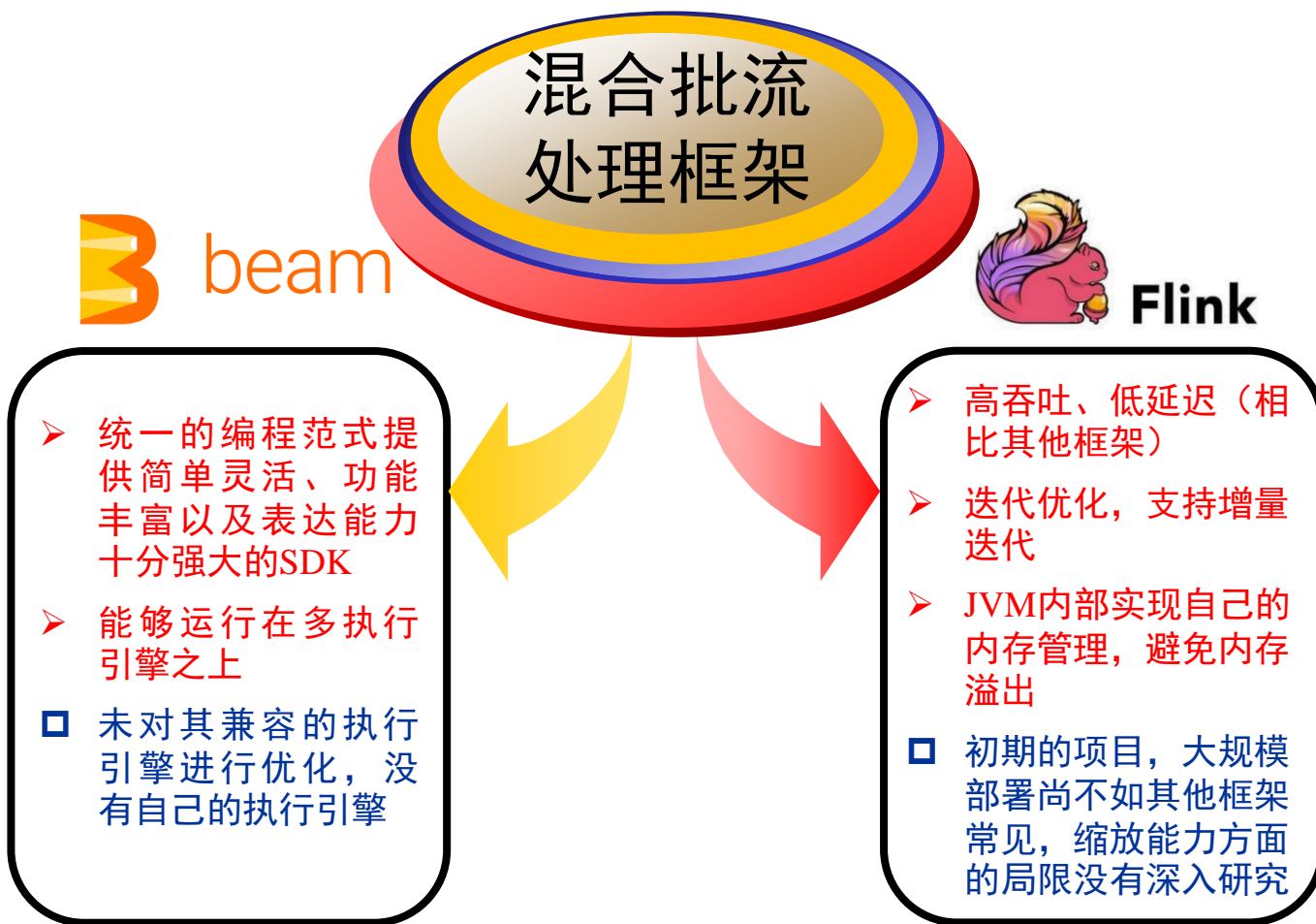


- 近实时处理需求的任务和应用场景很适合使用流处理模式
- 面对流数据出现的乱序性、突发性和易失性等问题，无法做出很好的响应



一. 大数据处理平台

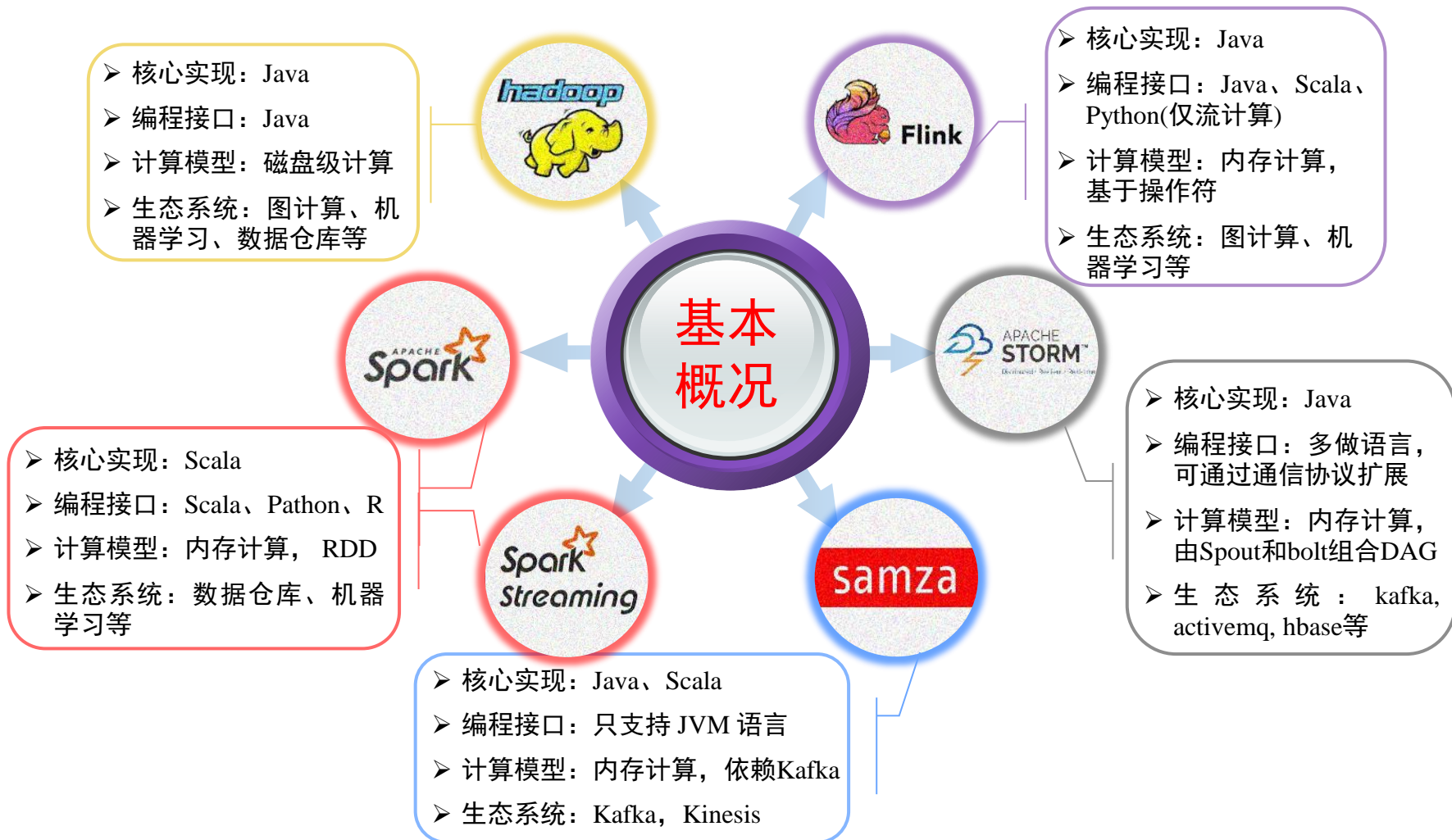
混合批流处理框架—优势与局限





一. 大数据处理平台

□ 大数据处理平台比较—基本概况





一. 大数据处理平台

□ 大数据处理平台比较—批处理



VS



VS



1



➢ 内存计算基础上进行迭代计算优化，支持增量迭代

2



➢ 数据处理工作全部在内存中进行，所有中间态的处理结果均存储在内存中

3



➢ 磁盘计算，I/O开销大

1



➢ 可稳定运行在数万节点的集群之上

2

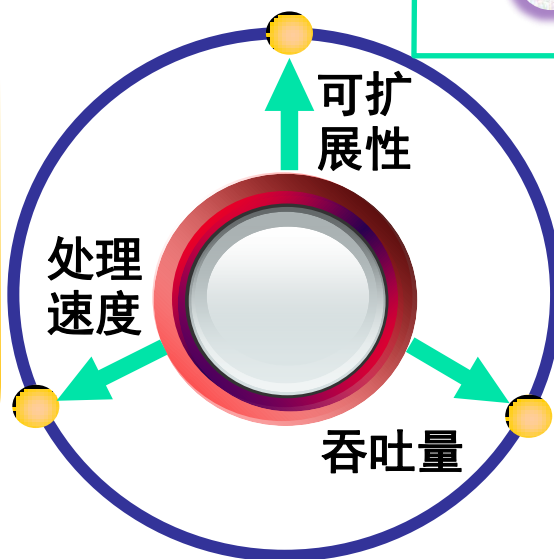


➢ 目前最大已知Spark集群是8000个节点，规模仍在扩大

3



➢ 平台处于发布初期阶段，大面积部署和优化仍需时间



1



➢ 吞吐量高，但不适合处理大批量小文件

2



➢ 相比Spark有优化，但吞吐量由其pipeline中最慢的部件决定

3



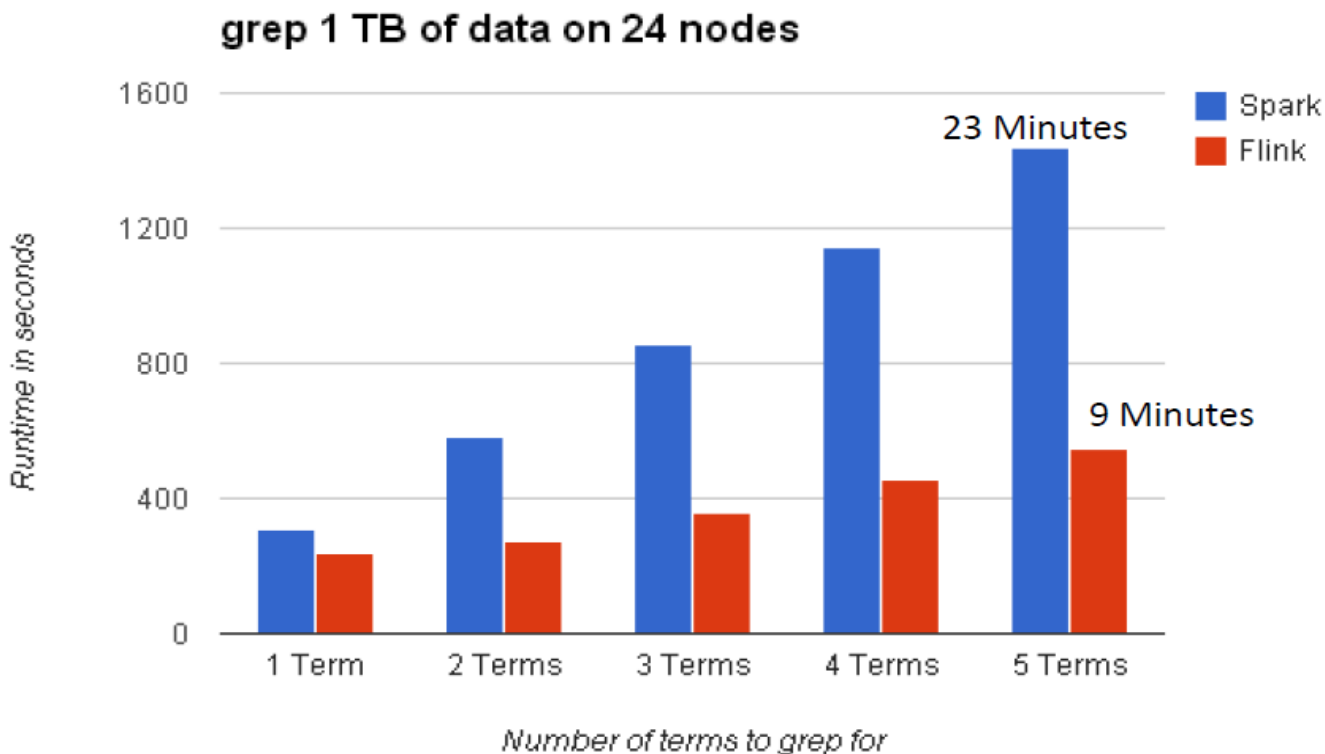
➢ 据分片不均衡可能导致吞吐量下降



一. 大数据处理平台

□ 大数据处理平台比较—批处理

Distributed Grep

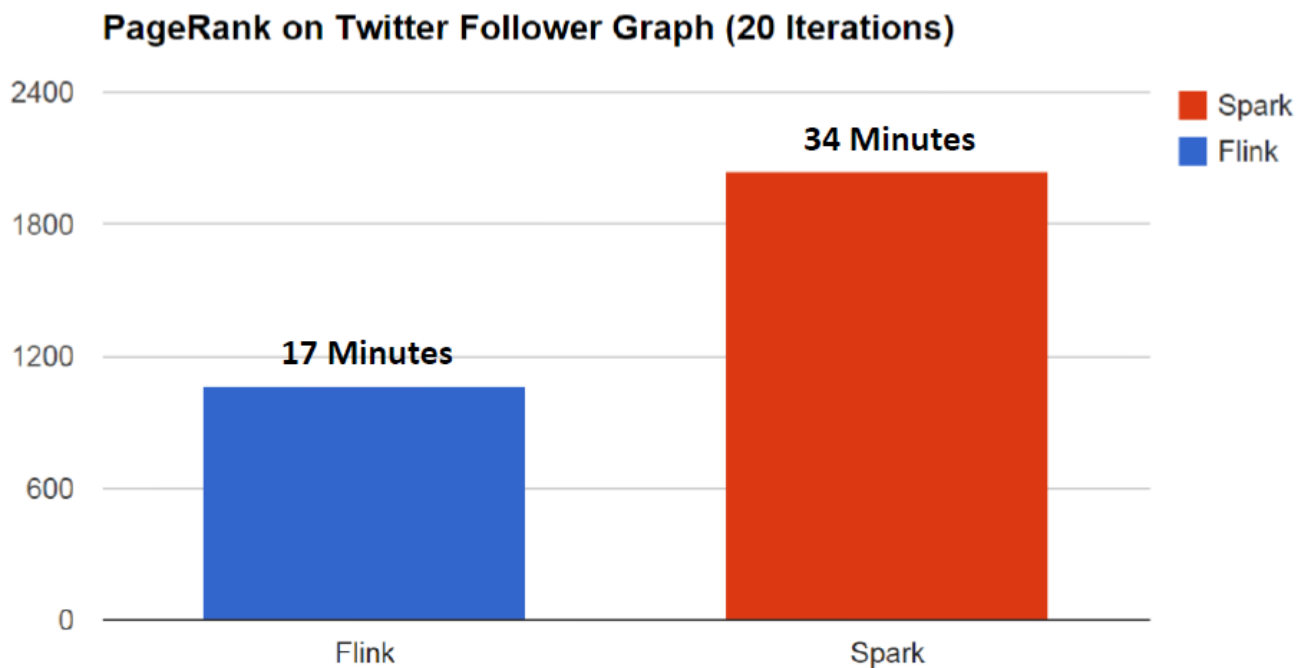




一. 大数据处理平台

□ 大数据处理平台比较—批处理

PageRank results





一. 大数据处理平台

□ 大数据处理平台比较—流计算



延迟时间

吞吐量

可扩展性

- 原生流处理平台，调优后最快可达毫秒级
- 原生流处理平台，延迟时间可达10毫秒级
- 微批处理，一般在秒级或亚秒级

- 本质是批处理，吞吐量高
- Kafka可帮助提升吞吐量
- 基于缓存区的数据交换提升吞吐量
- 更注重实时性，吞吐能力有所牺牲

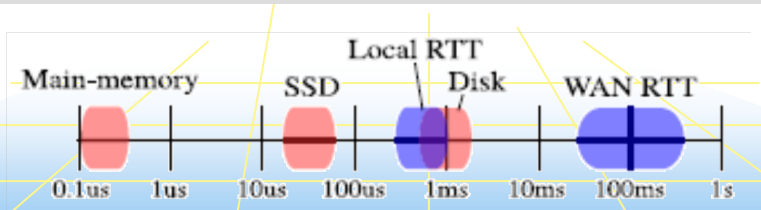
- 最大已知Spark集群是8000个节点
- 第一个主流流计算框架，可达数千节点
- 新兴平台，有待大面积部署
- 使用较少，扩展能力有待进一步调研

跨域数据处理的主要挑战

```
bytes=32 time=247ns TTL=124
bytes=32 time=172ns TTL=124
bytes=32 time=579ns TTL=124
bytes=32 time=1204ns TTL=124
bytes=32 time=1827ns TTL=124
bytes=32 time=1497ns TTL=124
bytes=32 time=421ns TTL=124
bytes=32 time=136ns TTL=124
```



- 巨大的网络延迟
 - 彻底颠覆传统的大数据处理技术
- 分散性大范围网络用户
 - 给提供优质的用户体验带来巨大挑战
- 丰富的边缘计算
 - 给系统的结构优化和过程优化带来不可估量的机会



主流的处理方法：副本复制技术



跨域数据处理的前沿技术

前沿技术

跨区域数据副本复制策略：提高可用性、容错能力，降低网络延迟

跨域数据存储技术

数据一致性技术：因果一致性技术和并行快照隔离技术

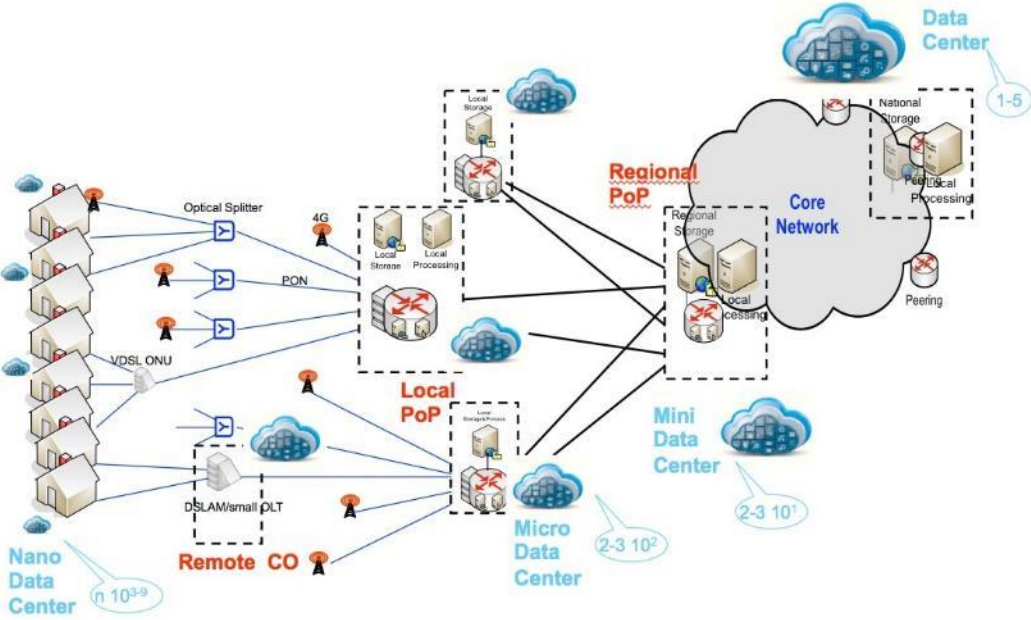
地理分布的查询处理和优化技术

域间作业与任务调度技术

边缘计算能力的有效利用

边缘计算能力的有效利用

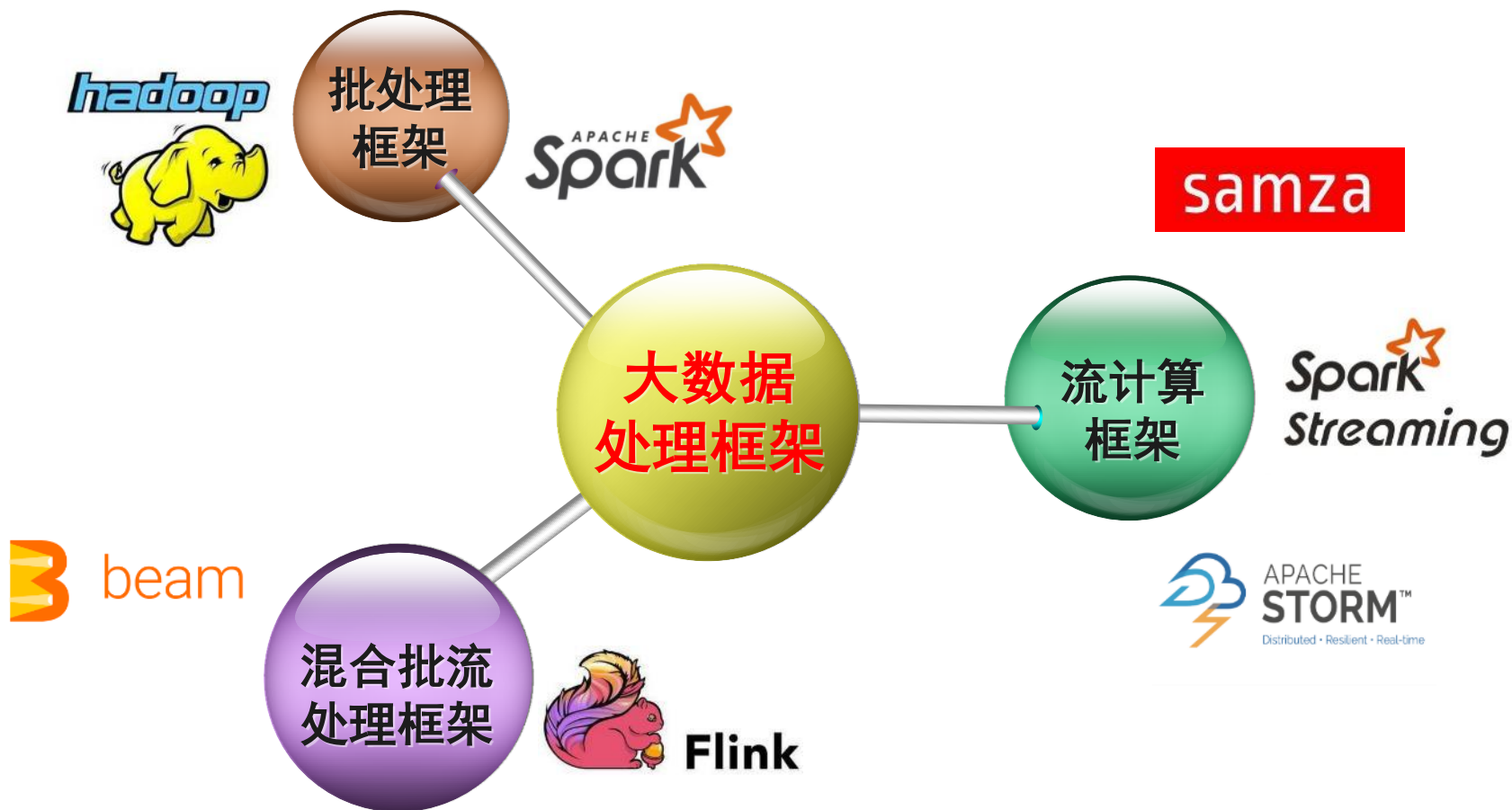
- **分布式低延迟计算**: 边缘计算聚焦实时、短周期数据的分析, 能够更好地支撑本地业务的实时智能化处理与执行
- **资源部署靠近用户**





批流融合

- 物联网数据计算的两种主流框架与模式：批计算、流计算





物联网数据管理前沿技术

前沿技术

跨区域数据副本复制策略：提高可用性、容错能力，降低网络延迟

数据一致性技术：因果一致性技术和并行快照隔离技术

地理分布的查询处理和存储技术

边缘计算能力的有效利用

在线增量计算技术

信号处理技术（DSP）

数据分析技术

