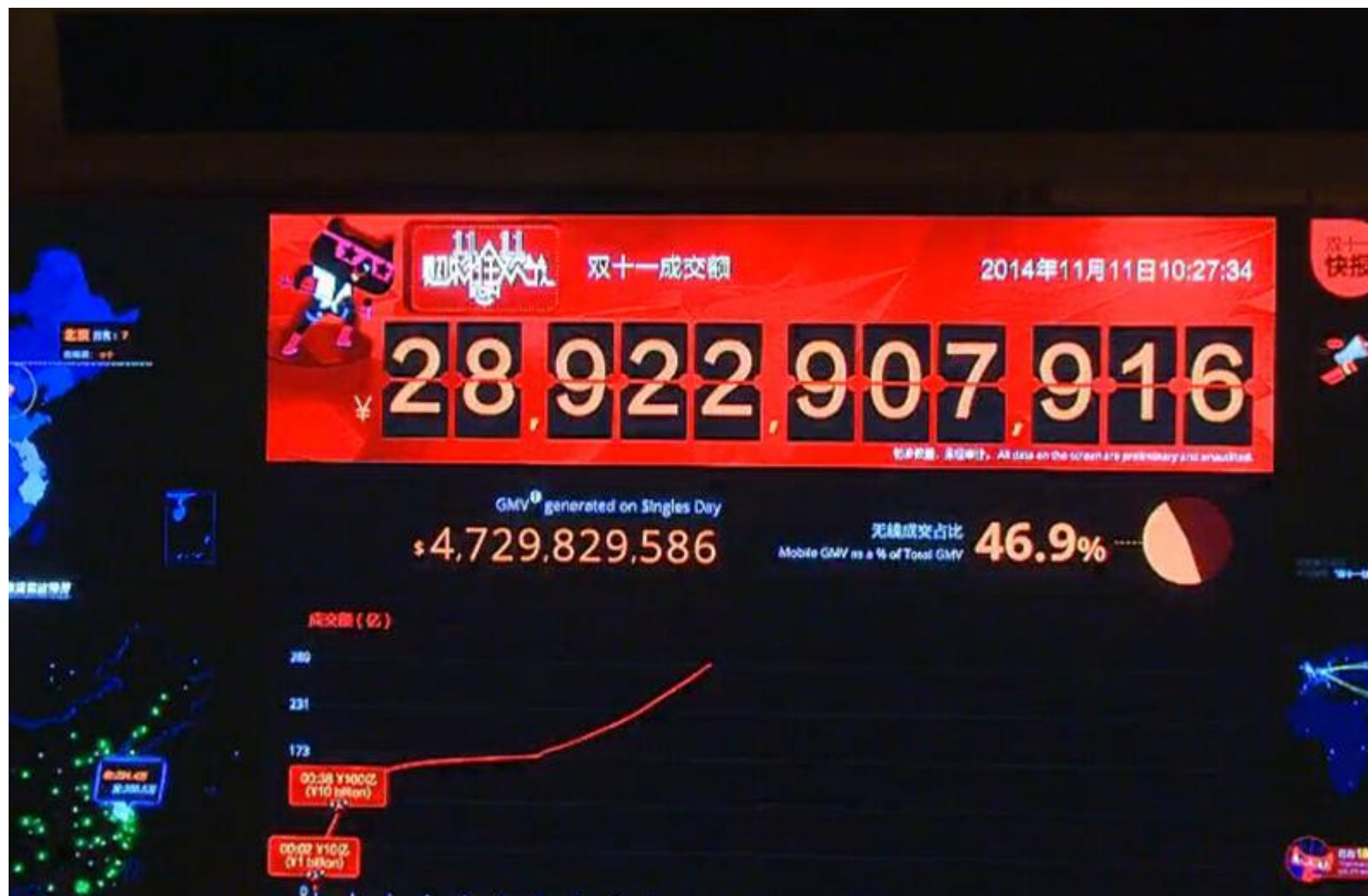


Storm实时计算详解

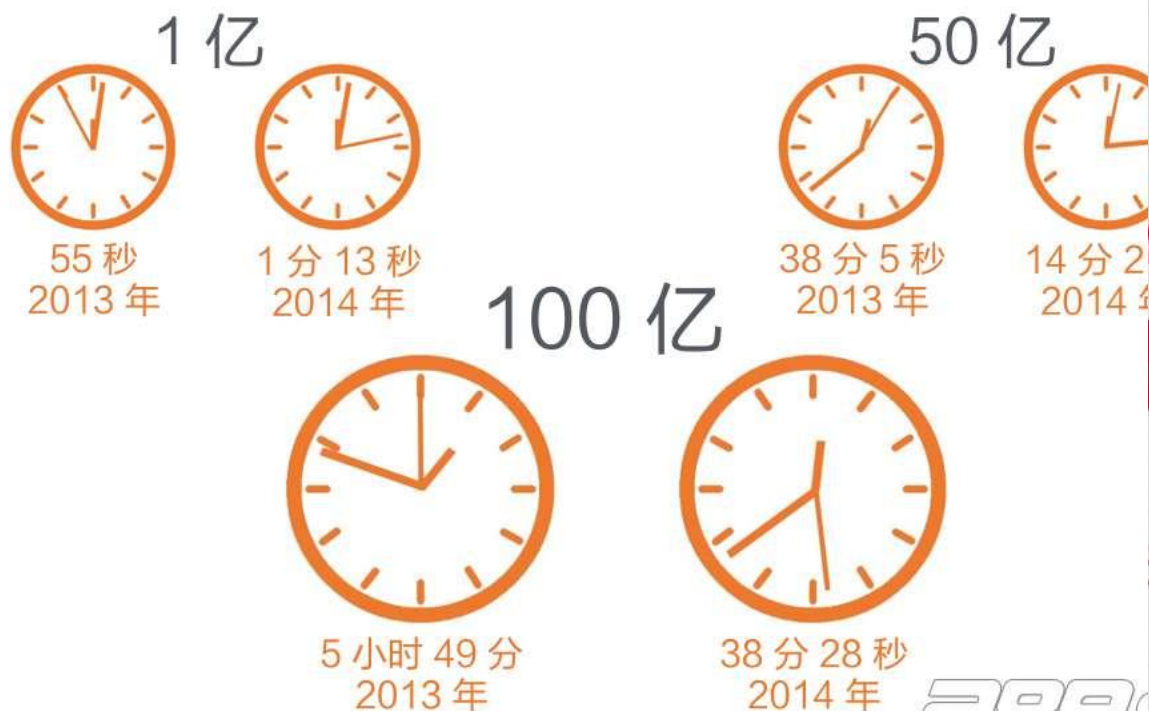
讲师: Yasaka





天猫双十一大屏幕

2014 阿里巴巴双十一之 交易额里程碑



数据来源：阿里巴巴 好奇心实验室



截止 03:00

天猫1111购物狂欢节
省份购买排名

广东省	912,088,366
浙江省	866,004,840
江苏省	785,598,039
上海	680,749,295
北京	547,641,712
四川省	504,118,389
山东省	421,682,351
湖北省	405,981,034
湖南省	369,851,972
福建省	366,121,688



- 什么是实时(流式)计算？
- 公司里为什么需要实时计算？
- 哪些技术支持实时计算？
- 什么是Storm？解决了哪些问题？
- Storm架构是什么？
- 如何搭建Storm？
- 运行样例程序
- 剖析API
- 中国移动项目展示



什么是实时(流式)计算?

Hadoop



Storm



什么是实时(流式)计算?

- 信息时效性的要求越来越高，随着时间的流逝，数据也在流逝
- 目标是随着数据流的实时到达，实时处理
- 采集、计算、查询
- Kafka/Flume/Scribe/TimeTunnel/Chukwa
- Storm/Spark/Samza/S4/Puma/JStorm
- Redis/Memcache/MongoDB/BerkeleyDB/HBase



- 在数据持久性建模不满足现状的情况下，急需数据流的瞬时建模或计算处理
- 应用实例有金融服务、网络监控、电信数据管理、Web应用、生产制造、传感检测等
- 对于大型互联网网站来说具有重大实际意义，实时的数据计算和分析可以动态实时地刷新用户访问数据，展示网站实时流量的变化情况，例如网站的访问PV/UV、用户访问的内容、搜索的内容、页面服务的质量、带宽使用情况等





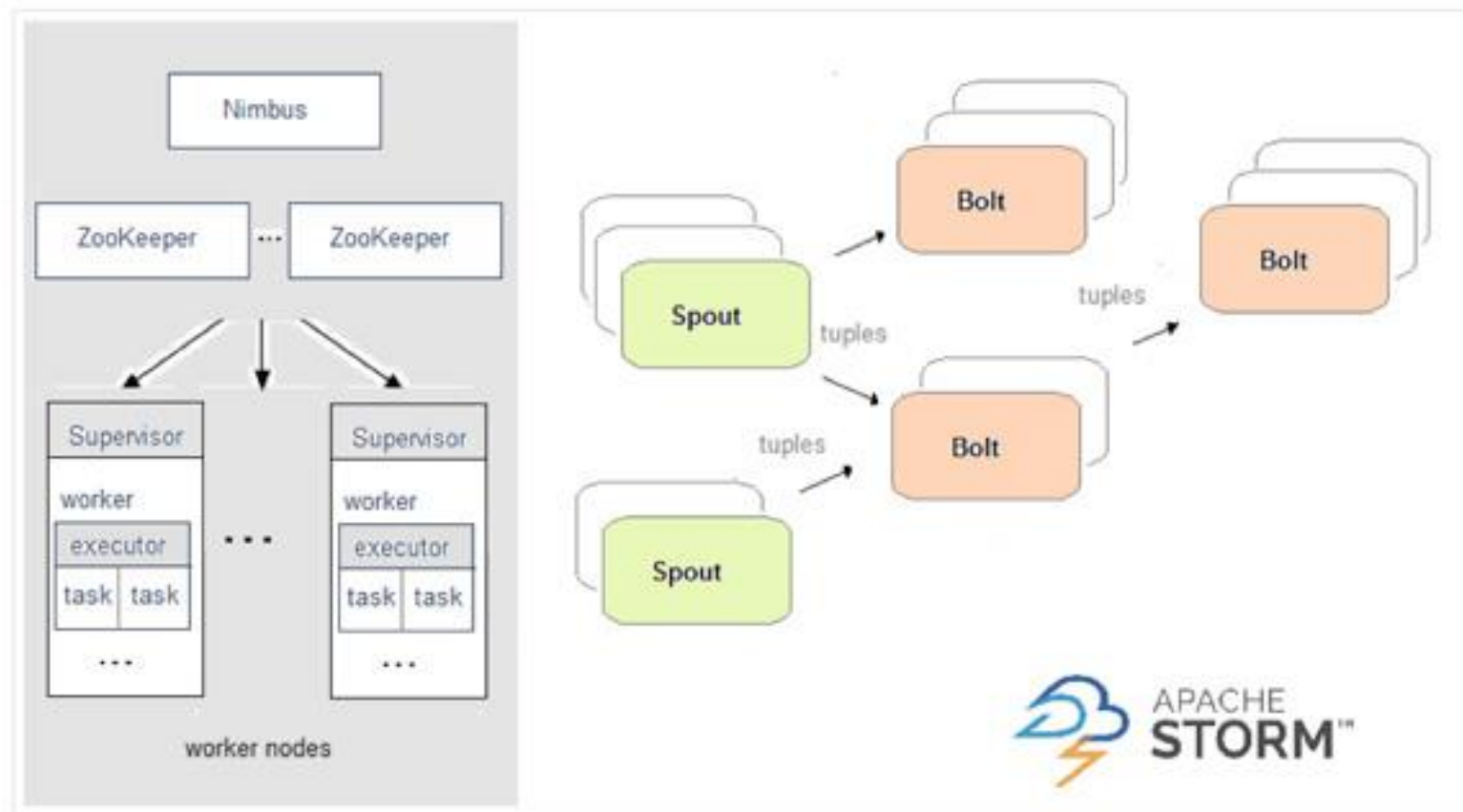
腾讯QQ在线人数统计



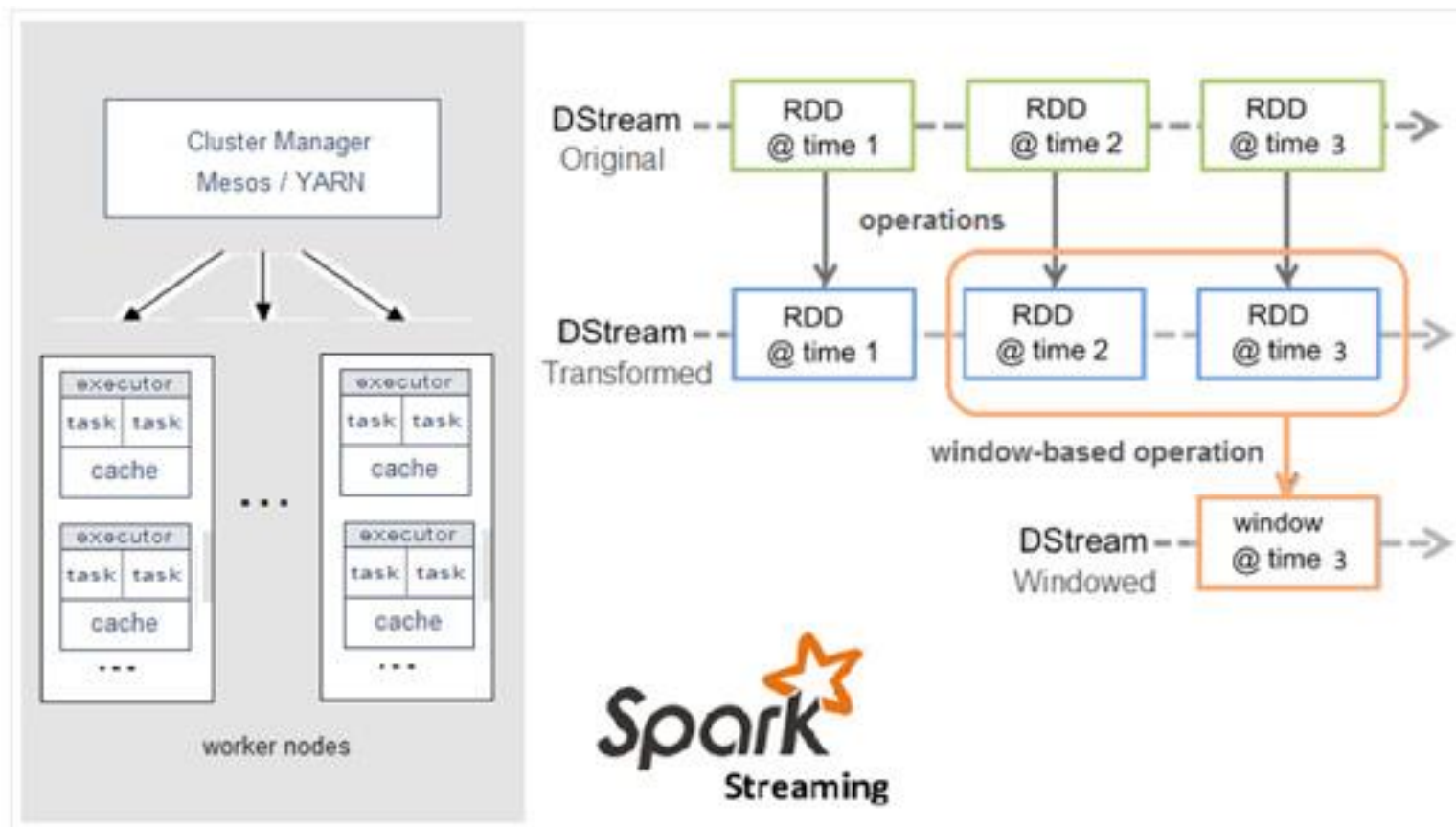


哪些技术支持实时计算？

- Storm

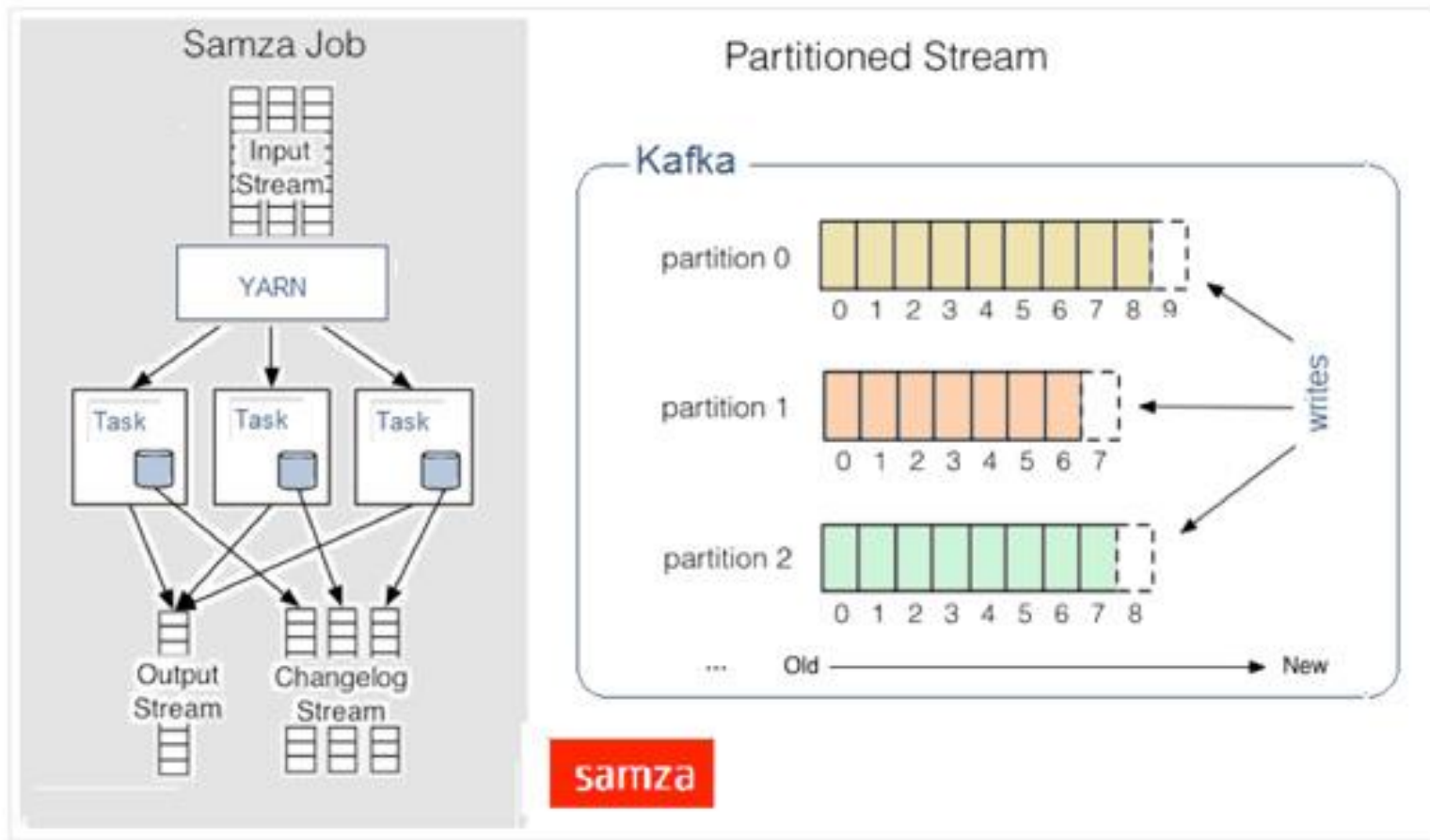


- Spark

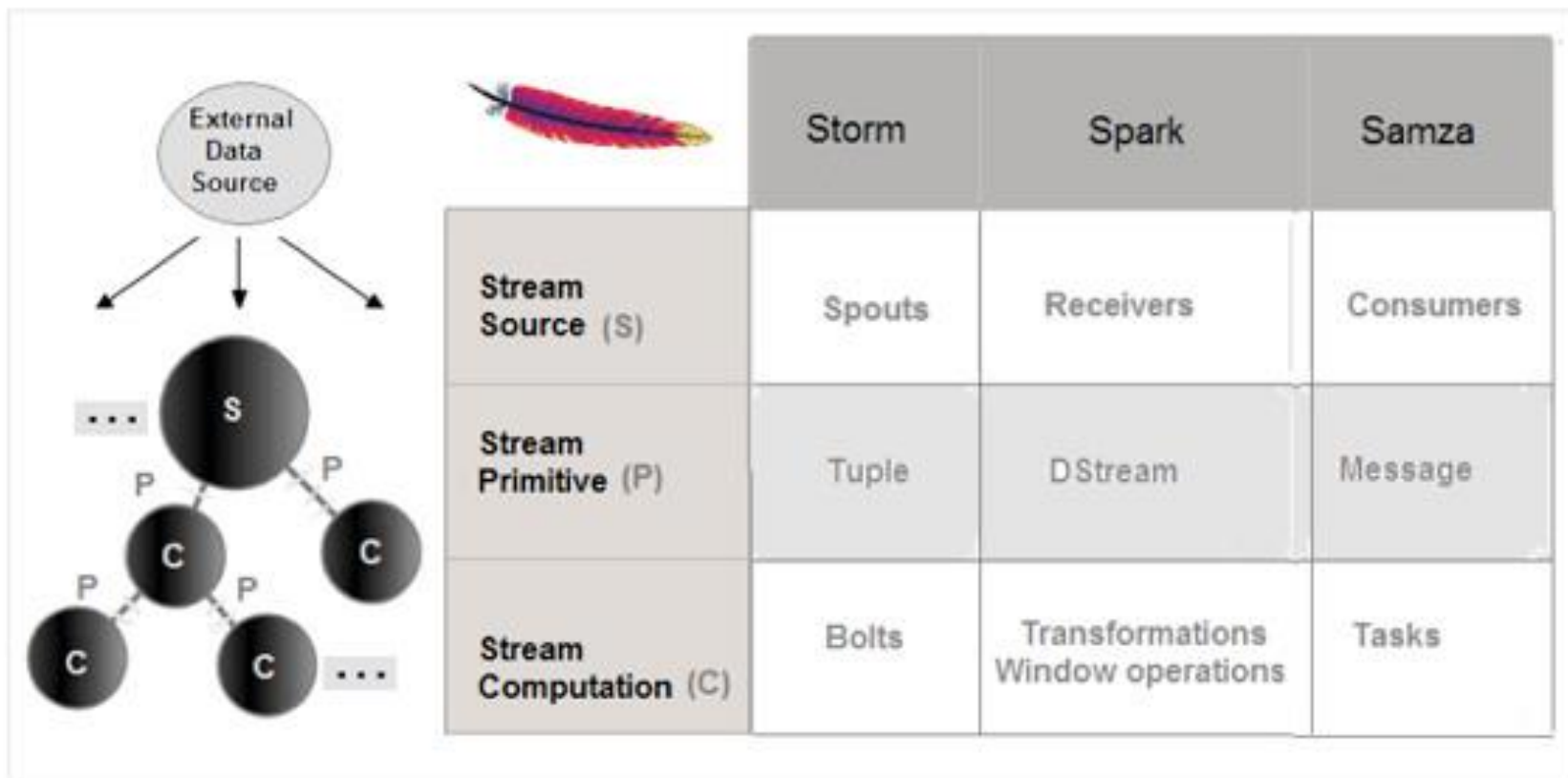


哪些技术支持实时计算？


- Samza



哪些技术支持实时计算？



哪些技术支持实时计算？



	Storm	Spark	Samza
Delivery Semantics	At Least Once Exactly-Once with Trident	Exactly Once Except in some failure scenarios	At Least Once
State Management	Stateless Roll your own or use Trident	Stateful Writes state to storage	Stateful Embedded key-value store
Latency	Sub-Second	Seconds Depending on batch size	Sub-Second
Language Support	Any JVM-languages, Ruby, Python, Javascript, Perl.	Scala, Java, Python	Scala, Java JVM-languages only



- Storm是实时流计算的一员，具备了与Hadoop MapReduce不同角色的能力——低延迟、高可靠性和容错
- 2011年由Twitter开源，使用Clojure语言实现的。Lisp，JVM
- 支持多语言
- Storm保证每个消息都会得到很快处理，一个小集群中，每秒可以处理数以百计的消息
- Storm的处理速度非常惊人：经测试，每个节点每秒可以处理100万个数据元组

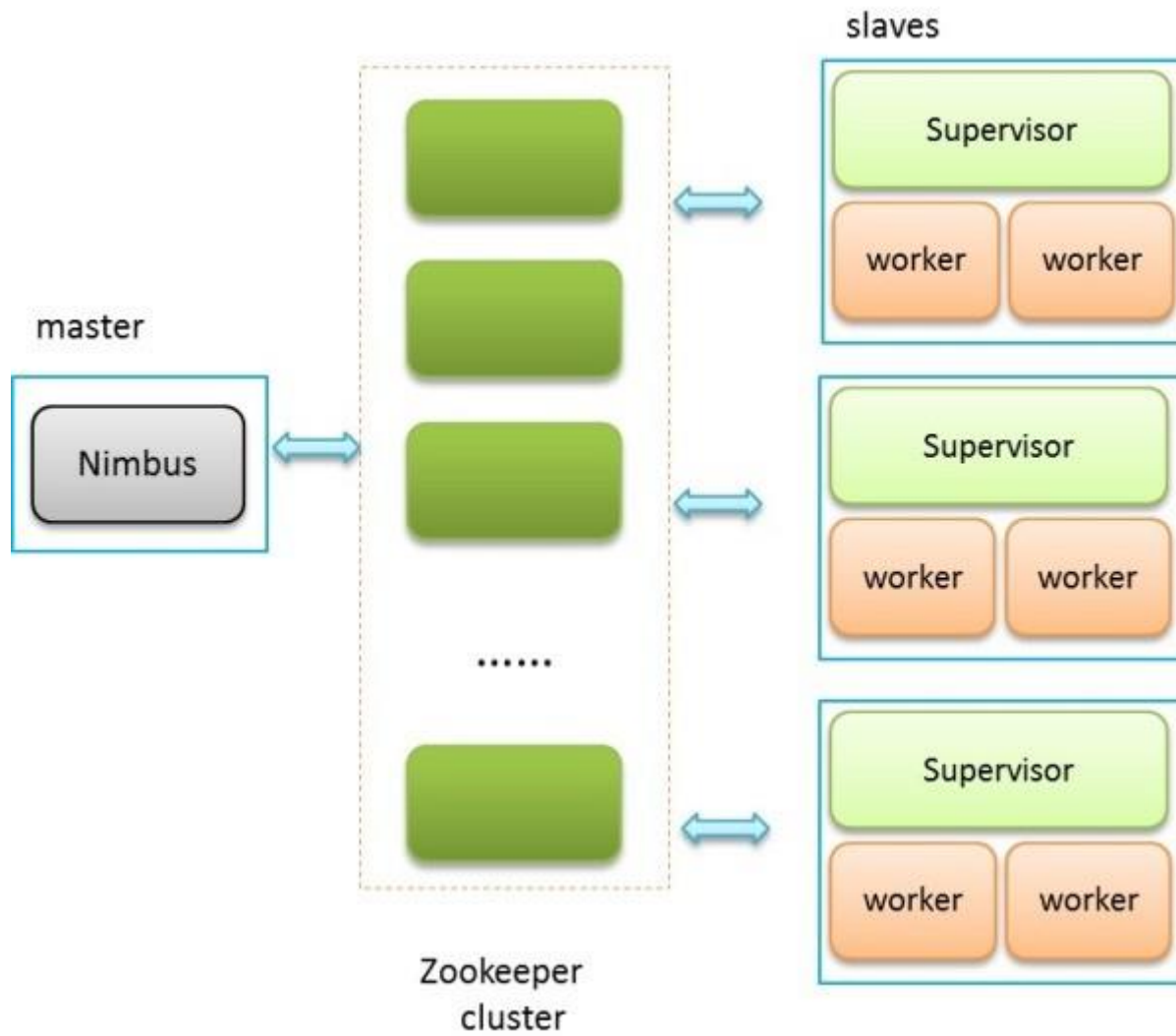


解决了哪些问题？

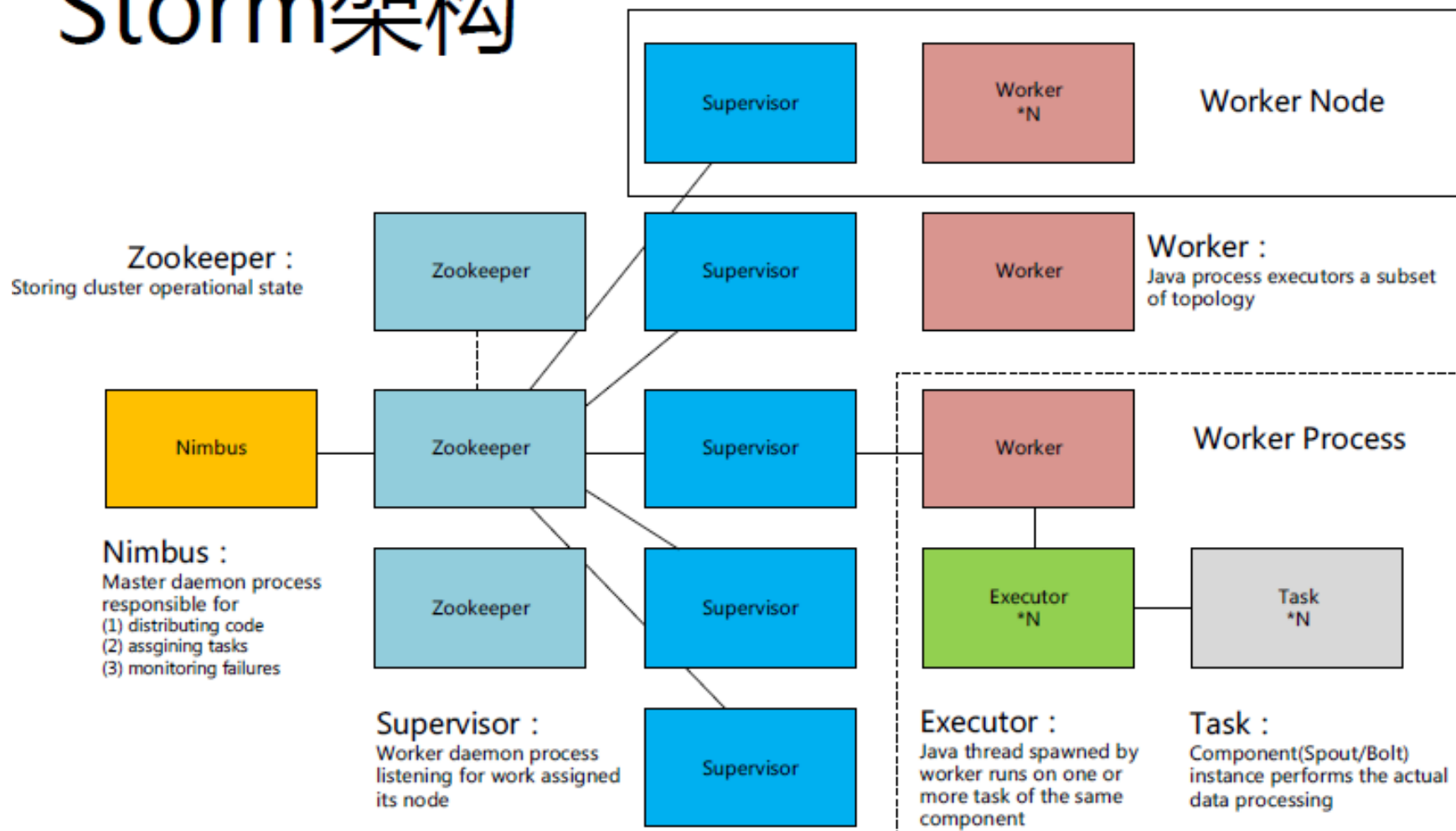
- 互联网广告实时流量统计
- 互联网数据质量实时监控
- 交通超速频发路段监控
- 交通基于GPS的实时路况分析
- 移动互联语音实时墙
- 运营商网络流量流向实时分析
- 中国移动小区基站预警



Storm架构是什么？



Storm架构



- 环境准备(JDK,SSH服务,Python,ZK)
- 单机
- `./bin/storm dev-zookeeper >> ./logs/zk.out 2>&1 &`
- `./bin/storm nimbus >> ./logs/nimbus.out 2>&1 &`
- `./bin/storm ui >> ./logs/ui.out 2>&1 &`
- `./bin/storm supervisor >> ./logs/supervisor.out 2>&1 &`
- `./bin/storm logviewer >> ./logs/logviewer.out 2>&1 &`
- 验证：访问<http://localhost:8080>，运行wordcount example



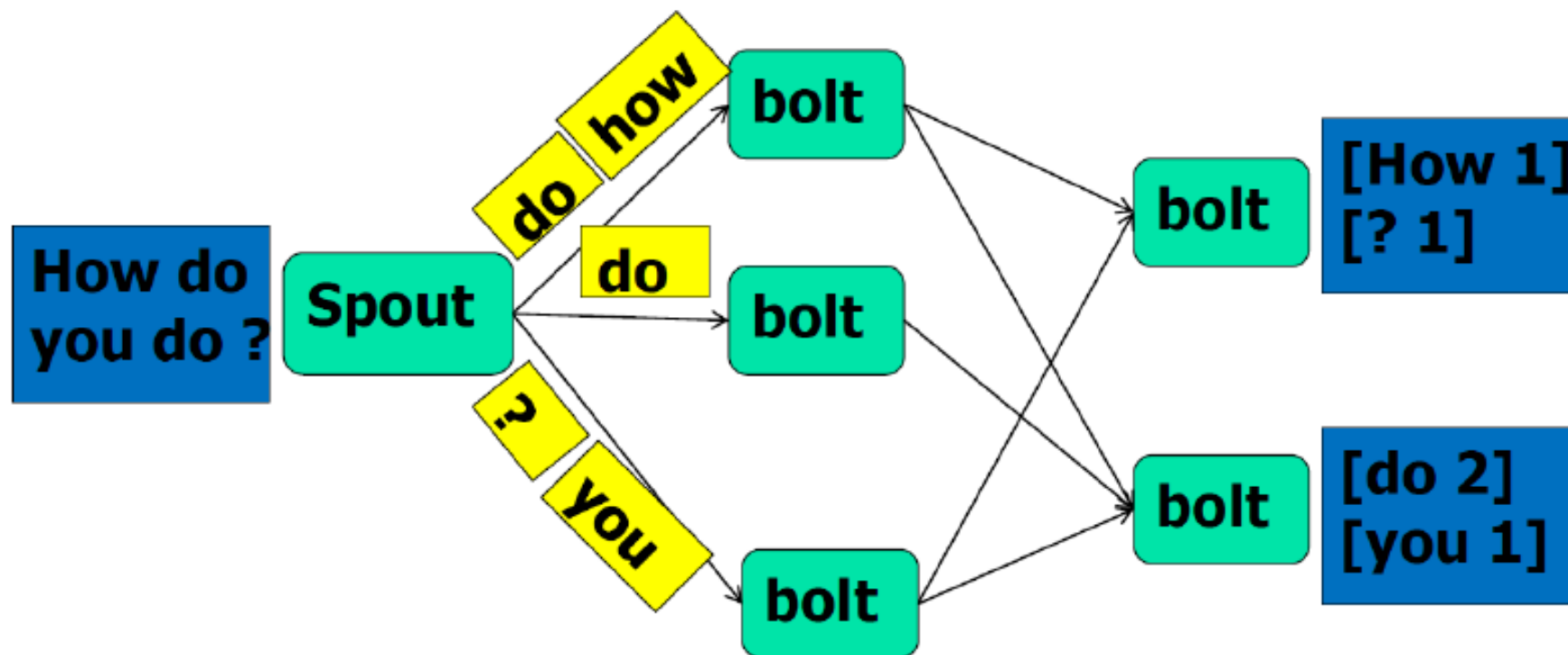
- `ls examples/storm-starter`
- `jar tvf examples/storm-starter/storm-starter-topologies-0.9.5.jar | grep WordCount`
- 下面我们来提交这个topology
- `./bin/storm jar examples/storm-starter/storm-starter-topologies-0.9.5.jar storm.starter.WordCountTopology wordcount`



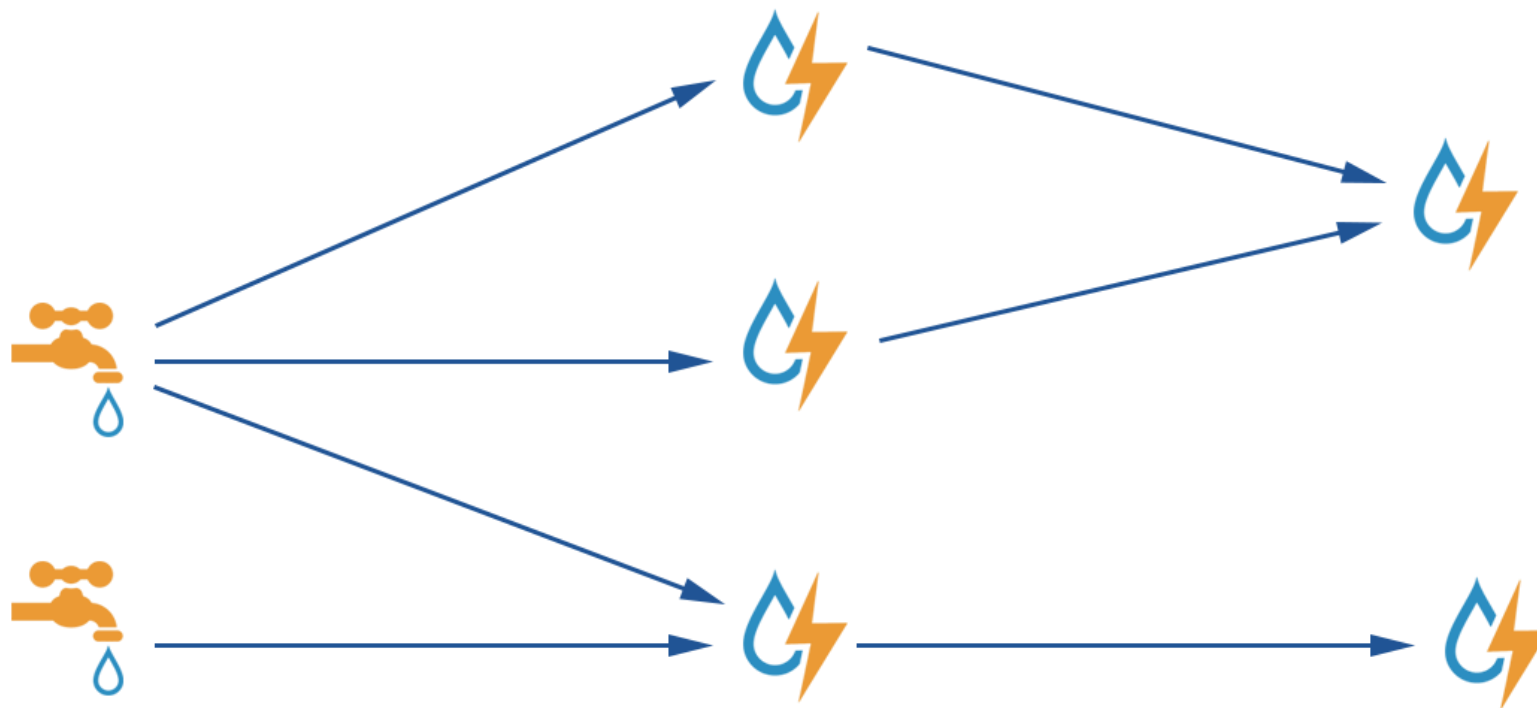
- 分布式
- 分发Storm到所有nimbus/supervisor服务器
- conf/storm.yaml
- storm.zookeeper.servers:
 - - "spark001"
 - - "spark002"
 - - "spark003"
- nimbus.host: "spark001"



- `./bin/storm jar examples/storm-starter/storm-starter-topologies-0.9.5.jar storm.starter.WordCountTopology wordcount`



- Spout / Bolt



- kafka(移动的程控交换机获取数据)
- storm (实时计算)
- hbase+zookeeper (数据存储)
- servlet+ehcart(数据展示)
- get 'cell_monitor_table','29448-000001_2015-07-24'



- 分久必合，合久必分
- Storm绝对是一个相当“有内涵”的系统，能把这么复杂的事情抽象得很完美，就像河水从曲曲折折的河道一直流向大海一样，所以也称这种数据处理方式为流式计算

