

2011/12/3

NoSQL/NewSQL在传统IT产业的机遇与挑战

钟颢 华为IT产品线 企业数据解决方案

walkerzhong@huawei.com



免责声明

本胶片仅代表个人对于业界技术的分析与观察，不代表作者所服务组织华为技术有限公司的观点。

本胶片提及的华为技术有限公司的相关产品，并不代表任何承诺，对于胶片中提及产品的深入交流，请查询华为技术有限公司网站，或者联系华为技术有限公司销售代表获得准确信息。

目录

- 传统IT产业的演变与挑战
- NoSQL/NewSQL的兴起、挑战与演进
- 可信赖的华为企业数据解决方案

历史数据管理

在很多企业中，历史数据管理，还没有作为一个重要的流程出现，这导致很多历史数据可能还没有到生命周期的时候，就丧失了，但是在一些对于国计民生非常关键的领域，比如涉及到财务的领域，国家有要求，需要至少保持十五年，由于中国“大”的特征，特别是由于电子商务等交易形式的出现，历史数据的管理，已经远不是传统的**RDBMS**加高端的**PC Server**或者小型机就能解决的，垂直扩展在技术架构和成本上已经走不下去，横向扩展的方案势成必然。

过去:1



昨天:10



中长期:100



电信系统容量的增加和多系统融合

电信系统是一个分布式系统，构成电信网络。

随着竞争的激烈，为了减少资本投入、运维投入，甚至仅仅是为了腾出机房，减少占地，单一电信系统的容量越来越大。历史上一个系统**20~30**万人，目前甚至有的系统需要做到可以支持数亿人。

在数据进行融合管理的趋势下，电信网络中同类项合并，比如都是属于用户鉴权认证的系统，合并到一起。

以上的动作，都对传统上依赖于**RDBMS**的系统形成挑战，尽管特别高端的**RDBMS**在这种情况下仍有可能，但是成本已经无法接受。

信令监测系统的演变—抽样到全量

信令监测系统可以算是一种传统的**IT**系统，规模的产品化始于十多年前，一直是传统的**RDBMS**的架构，后来采用了传统的列存储**RDBMS**。

随着运营商优化网络需求的提高以及监管要求，对于信令系统的全量监测成为必然，这种需求与系统容量本身的增长，数据量出现了百倍的变化。

数据中心安全与监控

信息系统的监控与安全防范，是一个公司**IT**最基本和最重要的功能之一。

传统的数据是用**RDB**，据有的大企业反馈，很多情况下，每天的此类型数据量已经多达数十个**GB**，而且随着系统的复杂化，数据量还有增加的趋势。传统**RDB**的扩展已经到了尽头，目前架构也有逐渐切到**NoSQL**的趋势。

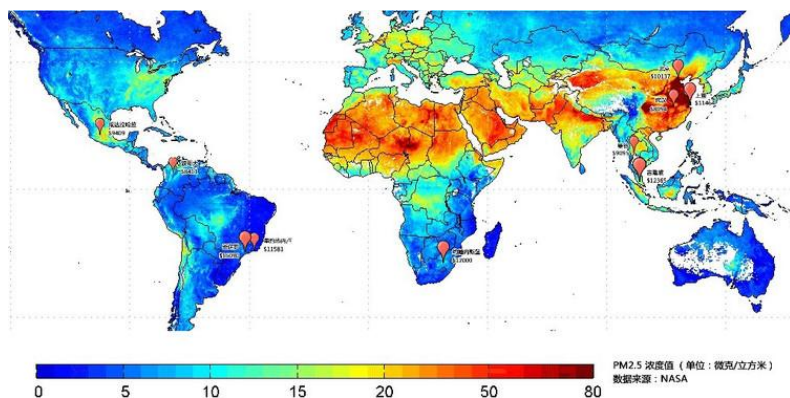


数据要求在更短时间内得到处理

传统IT企业涉及到数据处理的工具，主要是**RDBMS**，还有内存数据库（**IMDB**）。

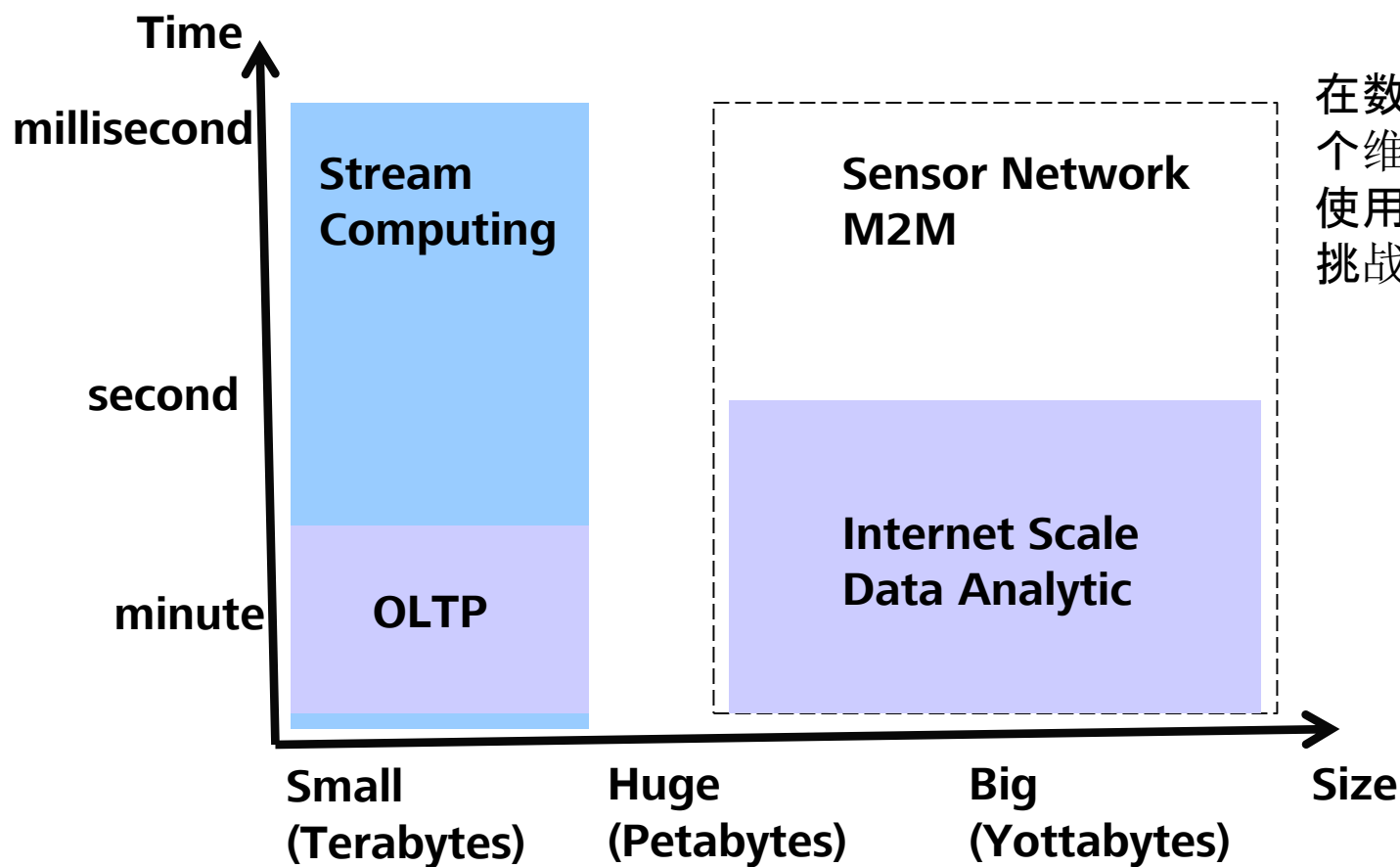
随着短信、邮件等工具的广泛使用，个人要求对涉及到自己的事务处理，有更多的知情和干预，比如账户变化短信通知等目前已经广泛获得应用。

更多的，由于个性化的发展，未来将会出现一种趋势，就是个人对于与自己相关的所有数据具有完全的掌控，设想一种可能性：未来整个城市的空气监测点数据实时发布，个人可以完全根据自己的需求来渲染自己的空气质量地图，从而制定自己的出行计划，提供此类型服务的应用必然吸引更多客户。



人均GDP在\$10,000美元附近城市PM2.5 浓度值分布图

传统IT产业演进对数据处理挑战总结



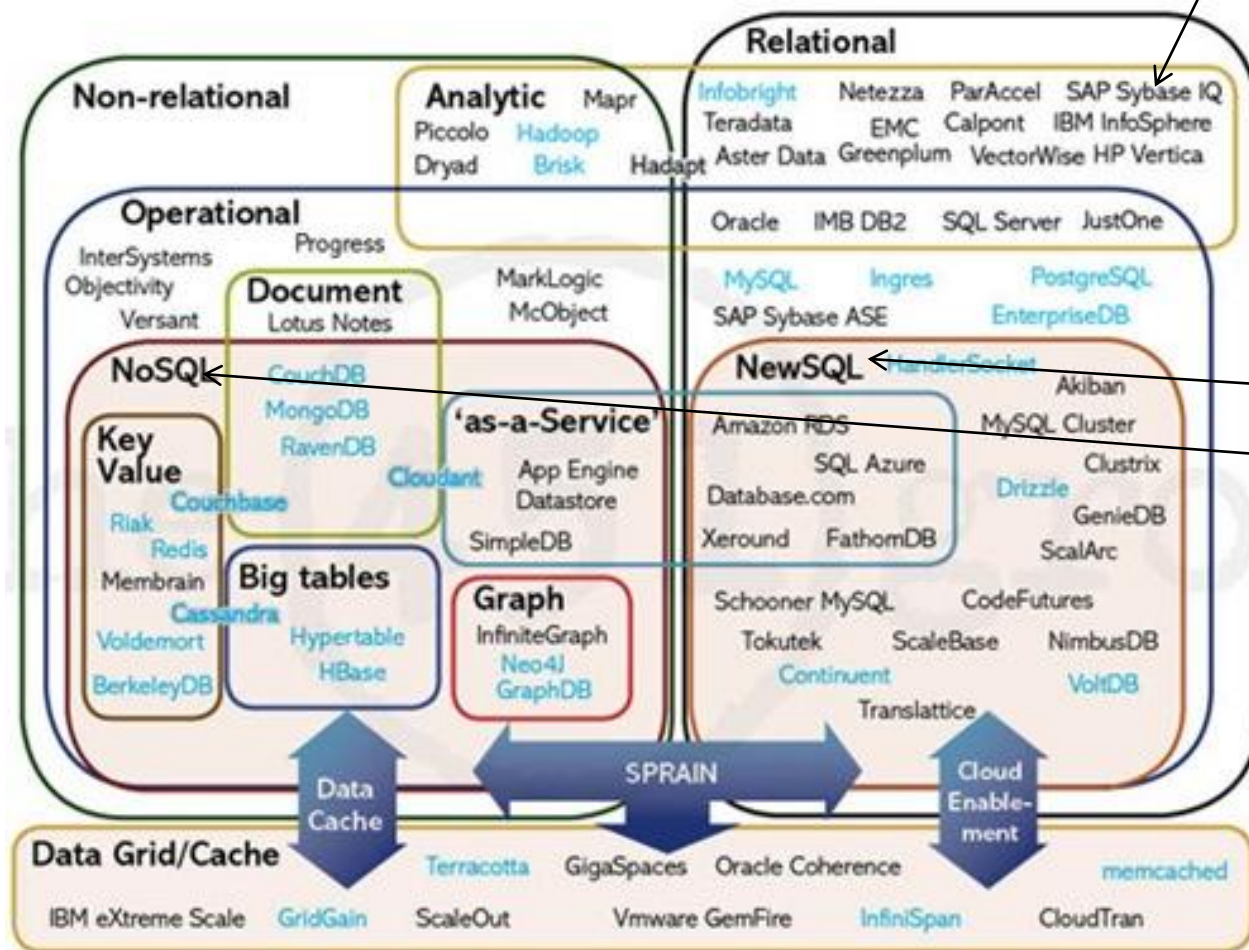
在数据量和响应时间两个维度, 新的需求对传统使用RDBMS的产品形成挑战。

目录

- 传统IT产业的演变与挑战
- NoSQL/NewSQL的兴起、挑战与演进
- 可信赖的华为企业数据解决方案

从一幅技术地图说起

传统上主要是内存数据库、RDBMS两条线，九十年代出现列存储数据库一条线。



直到2000年后，由于互联网、传感器网络预期的兴起，才大量出现了新兴的数据库或者说数据管理技术，两大类即nosql和newsq1，原则上来说，MPP RDB以及流处理技术也要归于newsq1一类

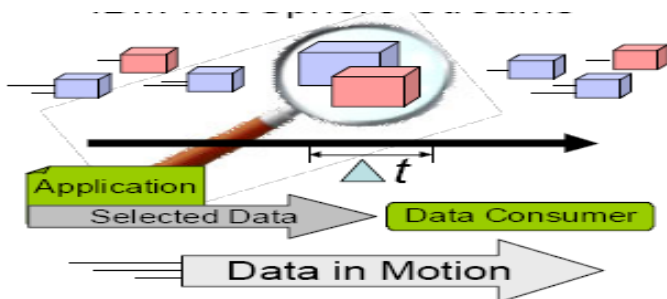
NoSQL/NewSQL技术的机遇（1）

从机遇的视角来看，基本上是不言自明的，那就是新兴的**NoSQL/NewSQL**能够解决前面提出来的一些需求场景导致的问题，对于几个基本的技术可以总结如下：

NoSQL：从构建分布式系统的视角出发，首先解决量的问题

NewSQL：从继承**SQL/ACID**处理能力的视角出发，构建分布式系统

流处理技术：将数据从先存储，后处理的模式，转换为先处理的模式



为应对海量数据实时处理的需求，业界引入了流处理的机制，特点在于：

- 在数据流动的过程中分析和计算
- 分析只对一定时间段内（ Δt ）的数据进行处理
- 事件/数据触发分析，分析过程始终在线

NoSQL技术的机遇（2）

- 卸载RDBMS的load
- 传统RDBMS中大量的数据，并不是都需要ACID保证的，这种情况下，可以将原来存储在RDBMS中，不需要ACID保证的数据都卸载到NoSQL
- 但是这种情况下，系统中往往有部分还是需要ACID保证的，所以是一个与传统的RDBMS技术联合的过程

1、卸载RDBMS load的例子

- 例一：信令监测系统
- 例二：IT安全监控与分析系统

2、Data Mart迁移的例子

与卸载的例子比较，一个差别是基本上系统中的RDBMS几乎可以完全的被NoSQL替换，如果不是考虑到与上层报表系统的对接

NewSQL技术的机遇（3）

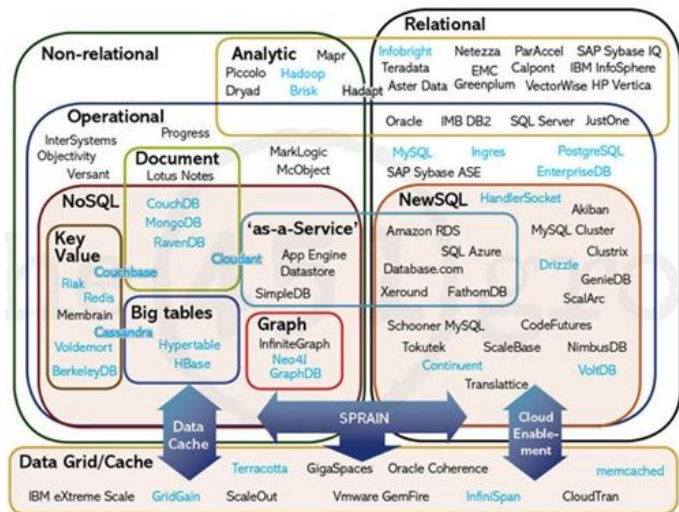
- **NewSQL**的技术优势在于几乎可以无缝的完整的替换原有系统的**RDBMS**

Enterprise Edition

- ✓ Breakthrough performance for high scaling database applications
- ✓ Ideal for production deployments
- ✓ Relational, SQL, ACID consistency
- ✓ Built-in HA, crash recovery and Hadoop integration

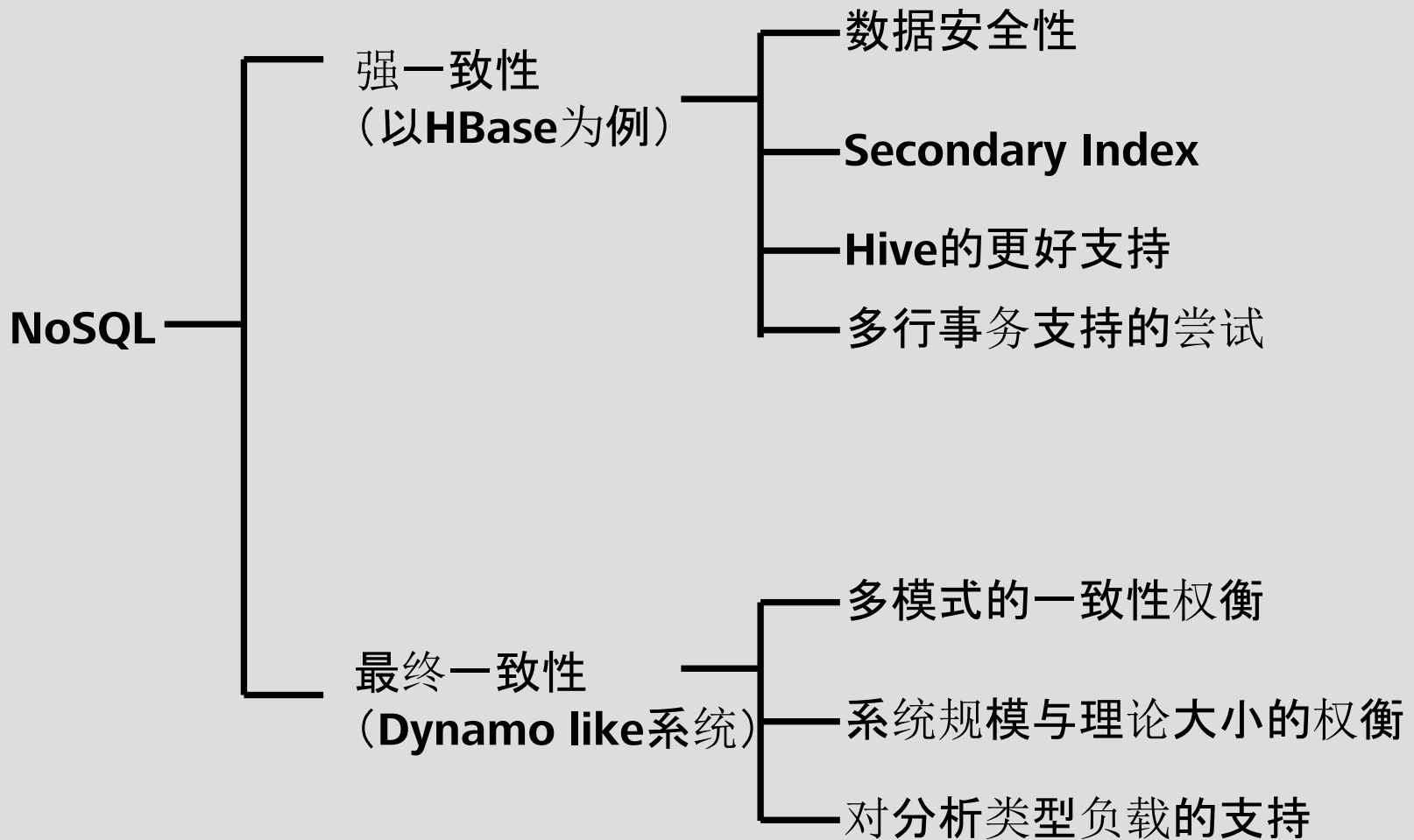
NoSQL面临的挑战（1）

分析挑战的视角：一个技术领域能否有成熟的产品供应者，教育者，消费者组成的完善生态系统，并且形成若干个主要商用（不一定是闭源）产品支撑的体系，对于数目庞大的企业客户，可以以合理的价格获得相关的产品。



NoSQL第一挑战是不够标准化，第二挑战是工程化方面。

NoSQL面临的挑战（2）



NoSQL面临的挑战（3）

1、值得关注的进展一

俄亥俄大学计算机团队做的YSmart的工作，实际上是在一个大规模分布式系统上进行查询优化的工作，其工作是优化不同分支间数据访问的性能。这可以视为是一种自底向上的努力，优化到最后，是否会出现事实上的分布式的分析性数据库系统？

2、值得关注的进展二

普林斯顿大学一个团队牵头做的COPS的工作，其试图在目前广为人知的最终一致性上和强一致性之间，寻求一个中间地带的实现形式，对应用实现的约束更小。（<http://sns.cs.princeton.edu/projects/cops/>）

Linearizability > **Sequential** > Causal+ > Causal > FIFO
> Per-Key Sequential > Eventual

NewSQL面临的挑战（1）



NewSQL面临的挑战（2）

NewSQL实际上是采取的是和**NoSQL**相反的方向来思考问题，同时由于其往往对**SQL**、**ACID**的支持，更容易切入传统IT产业的产品，同时在推广中也比较容易得到共鸣，因为大家都是在学习**SQL**的过程中成长起来的。

NewSQL的风险在于过度承诺(**Over Promise**)而产生的风险。

NewSQL如果能够直面自己不能解决所有大数据问题的现实，转而将自己放在一个与传统**RDBMS**、**NoSQL**合作分工的位置，将很好的找到自己的市场定位。

目录

- 传统IT产业的演变与挑战
- NoSQL/NewSQL的兴起、挑战与演进
- 可信赖的华为企业数据解决方案

Hadoop企业级应用存在的挑战

基础设施建设和规划	在Hadoop上构建业务系统	性能调优	运维	Hadoop社区跟进
● 选择适合Hadoop的硬件。 ● 根据Hadoop特征进行组网规划。	● 系统设计需充分考虑Hadoop的优劣势。 ● 培训开发人员基于MapReduce、HBase等API开发业务系统。	● Hadoop默认的参数配置通常难以满足业务的性能需求，需要进行配置参数调优。 ● 分析和优化业务层代码以达到性能最佳。	● 社区版本默认的OM功能难以满足企业级应用高效、可靠的运维需求，需要进行增强。	● 分析、同步、解决社区版本bug。 ● 分析和同步社区新特性，在业务系统中使用新特性以达成业务目标。

丰富运维经验与业界技术的最佳结合（1）



自动化OS安装
Web界面Hadoop集群安装
Web界面数据配置

新增服务器

* 服务器名称:

* 维护 IP:

* 类型:

描述信息: 1号数据中心 第1机房
第1机架 第6号服务器



应用名称	服务器名称	软件类型	模块状态	操作
hadoopstub_1	appserver1	datanode	正常	启动 停止 复位
hadoopstub_1	appserver2	datanode	正常	启动 停止 复位
hadoopstub_1	appserver3	datanode	正常	启动 停止 复位
...				

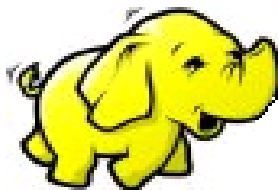
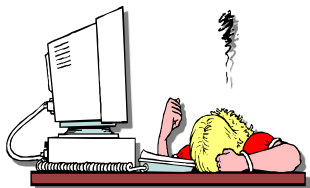
硬件状态监控
软件状态监控
软、硬件故障告警

■ 紧急: 0 ■ 重要: 0 ■ 次要: 0 ■ 提示: 0		
确认 恢复 刷新		
应用名称	告警级别	告警名称
hadoopstub_1	重要	进程故障
hadoopstub_1	重要	进程故障
hadoopstub_1	重要	进程故障

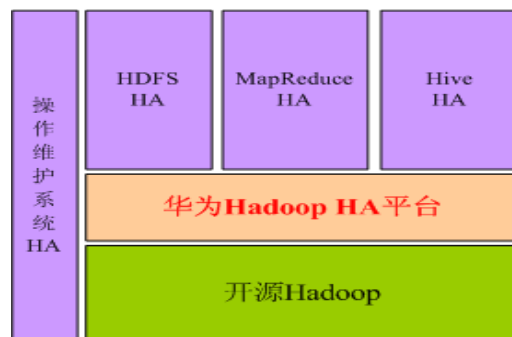
一键式集群规模调整

丰富运维经验与业界技术的最佳结合（2）

Apache Hadoop



Huawei Hadoop



- Apache Hadoop的NameNode、JobTracker、HiveServer进程存在单点故障问题，故障后需要人工干预才能恢复服务。
- 各进程故障后无法告警和自动拉起进程。

- 基于自研的Hadoop HA平台，构建NameNode、JobTracker、HiveServer的HA功能，进程故障后系统自动Failover，无需人工干预。
- 各进程故障后提供告警机制通报故障，同时OM系统自动拉起故障进程。

专业团队（1）

▣ 专业的国际化团队

- 在美国硅谷成立了研究中心，从事Hadoop领域研究的高端专家若干名，为总部研发团队提供源源不断的业界前沿技术。

▣ 一流的本地化研发

- 在深圳总部组建从事Hadoop研究、开发和产品交付的团队。
- 有专门的小组负责Hadoop开源社区技术动态、版本演进的跟进和分析同步。
- 具备提供从开源社区同步补丁和发布独立版本的能力。
- 研发团队构建了一套自动化版本构建和自动化测试的系统，能保证系统持续处于能发布状态。

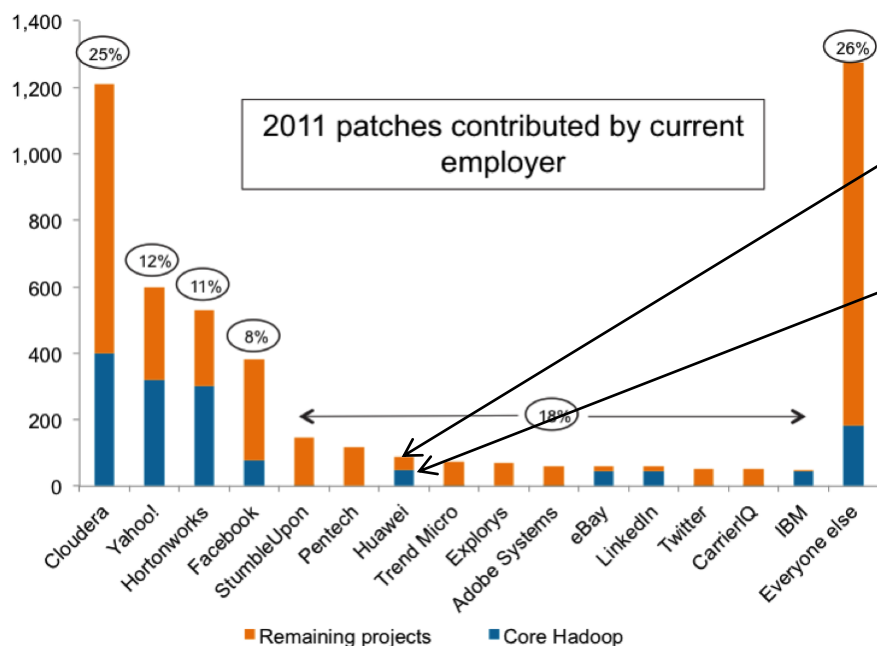
▣ 专业咨询与服务

- 提供实时咨询和运维服务，华为数据解决方案团队持续致力于集群故障检测和修复的研究，目前提供多种故障场景的检测、自动修复的手段。
- 提供客户系统运维培训，包括Hadoop系统入门、进阶和常见问题FAQ。
- 提供客户基于Hadoop系统的二次开发接口和开发指南，同时华为也有能力为企业提供服务，包括基于Hive、MR的开发。

专业团队（2）

● 团队对社区的回馈与贡献

- 在Hadoop领域，8名工程师被开源社区正式接纳为Contributor，以Hadoop社区中HBase项目为例，在最新发布的0.90.4版本中，华为在社区贡献的比例占**24%**。1名工程师成为HBase Committer（全球共有13名）
- 由于华为在Hadoop社区的贡献，华为已经成为Hadoop In China组织的关键成员，并且将在2011年年会上发表主题演讲。



按照业界的公开排名，华为Hadoop团队在全球综合排名第七

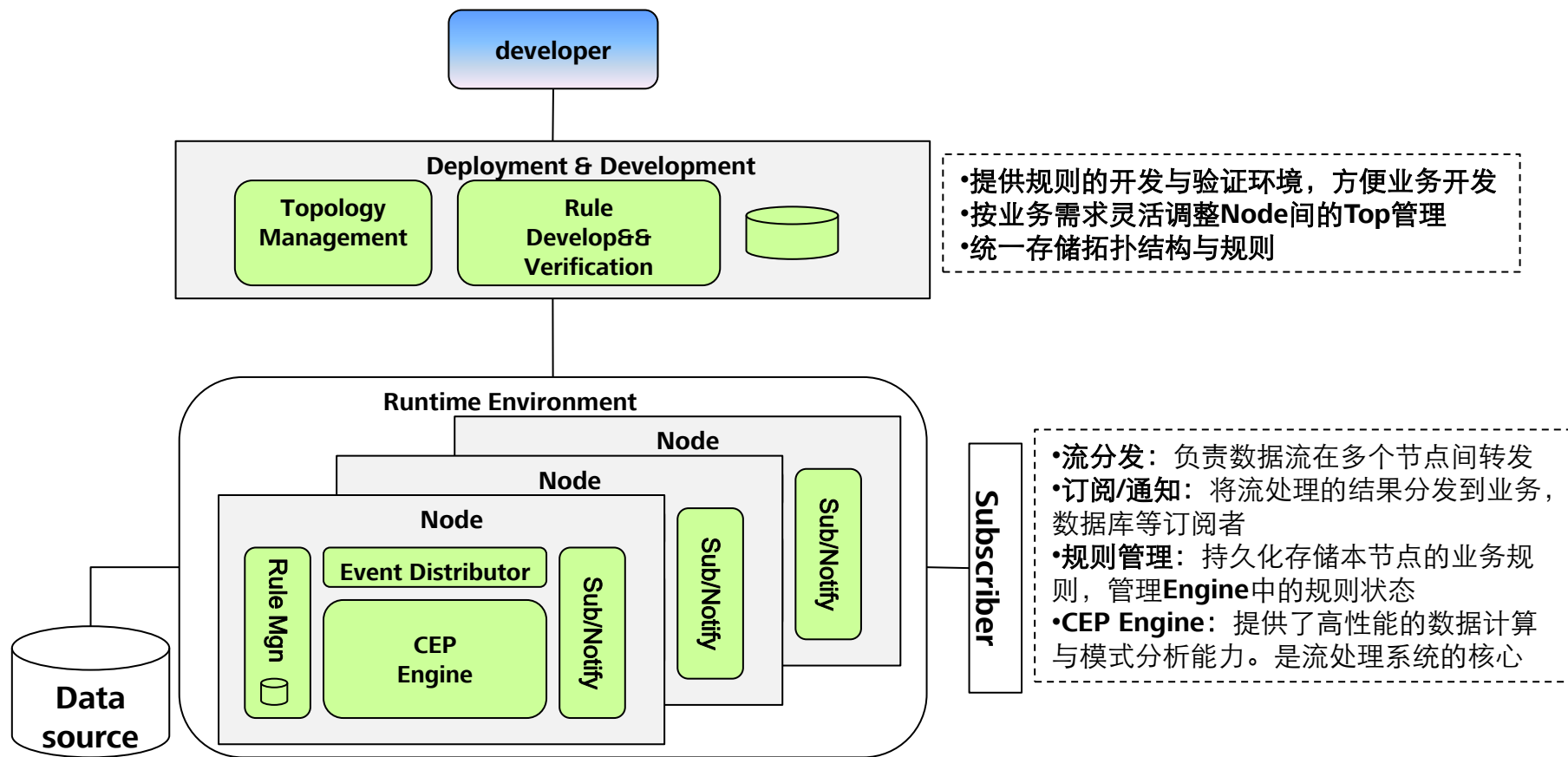
华为Hadoop团队的贡献是均衡的，同时在Hadoop内核和周边项目都有贡献

专业团队（3）

得益于专业团队的长期投入，华为**Hadoop**团队能够洞察**Hadoop**核心技术的方方面面，在通过工程化产品为客户提供价值的同时，持续的回馈社区，如下是对社区贡献的一个不完整小结。

	问题报告	补丁贡献	Committers
HDFS & Core	53	61	0
MapReduce	54	40	0
Hive	39	24	0
HBase	92	55	1
ZooKeeper	11	6	0
Total	196	160	1

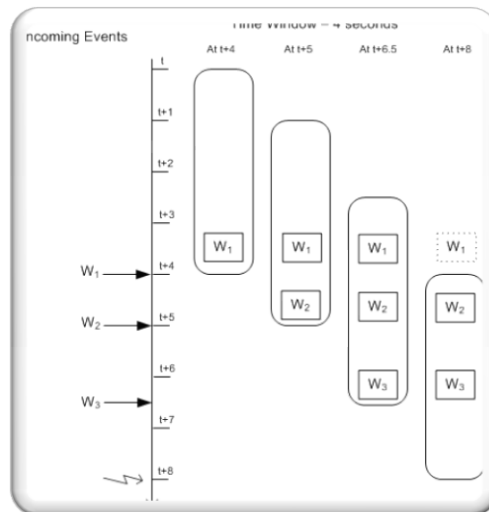
华为流处理解决方案



- 高性能的处理能力：基于简单规则单节点处理能可以大多100万条/秒
- 高可靠性：节点故障之后，业务可以自动切换到备用节点
- 开发简单：具备SQL基础的开发人员，可以很快上手开发业务

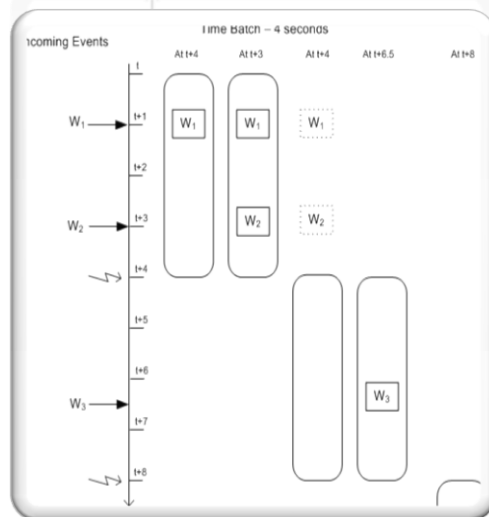
华为流处理的功能介绍

- 结果的获取
 - 使用监听器(Listener), **PUSH**
 - 使用Iterator, **PULL**
- 使用类SQL语言定义规则
 - **SELECT, FROM, WHERE, GROUP BY, HAVING, ORDER BY**
- 支持多种视图(view, window), 如
 - 时间窗口(滑动时间窗口, 跳动时间窗口)
 - 长度窗口(滑动, 跳动)
 - 排序视图(对时间排序, 对数值排序, 分组等)
 - 统计视图(规模, 平均值等)
- 支持模式语法—检测流中的复杂模式
 - 模式: 表达式重复模式、对事件顺序的操作(跟随)、时间段
 - 匹配识别(使用正则表达式): 针对单一类型事件
- 支持对多种流的多种处理
 - 支持同时对多个流中的数据进行选择
 - 创建新的流: **INSERT INTO**
 - 可以从多个流中提取数据插入到新的流中
- 支持窗口操作
 - 类似SQL中的表操作
 - 创建窗口, 插入、更新、删除事件



滑动时间窗口:
根据系统时间, 具有特定时间间隔的移动的窗口(在过去的4秒钟里...)

```
select * from  
Withdrawal.win:time(4 sec)
```



跳动时间窗口:
缓存事件, 并在特定的时间间隔后释放(在每个相邻的4秒钟里...)

```
select * from  
Withdrawal.win:time_batch(4 sec)
```

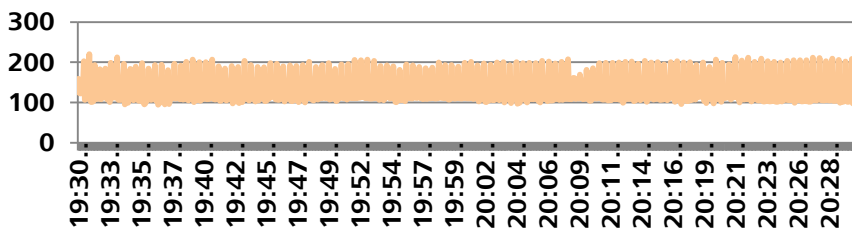
华为流处理的性能

测试环境

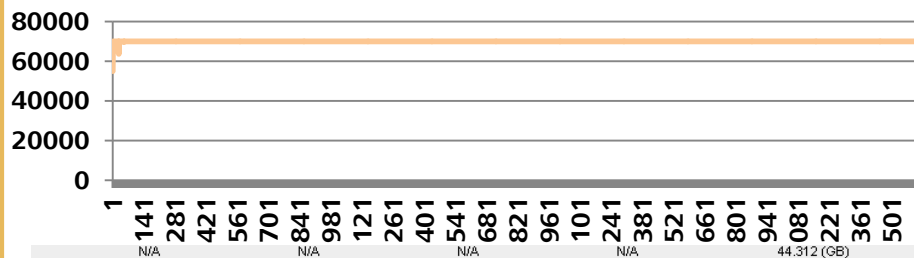
CPU: 16 Intel(R) Xeon(R) CPU E5620 @ 2.40GHz 内存: 16 G

30组KPI计算 单进程最大处理能力在 7万条/秒左右

%CPU

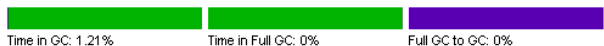


Read Input



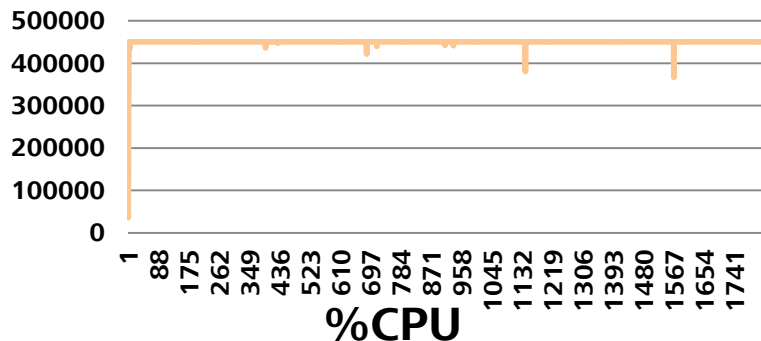
Last occurrence (s)	Count	Average interval (s)	Average duration (s)	Average rate of collection
3,597.519 (s)	121	29.739 (s)	0.367 (s)	36.262 (GB/s)

	Value	Name	Value
nt	3,668.894 (s)	Time spent in GC	44.348 (s)
	1.582 (TB)	Percentage of time in GC	1.209%
	121	Time spent in Full GC	0 (s)
er GC	13.385 (GB)	Percentage of time in Full GC	0%
	457.568 (MB/s)	Avg. allocation rate	452.037 (MB/s)
	11.483 (GB)		

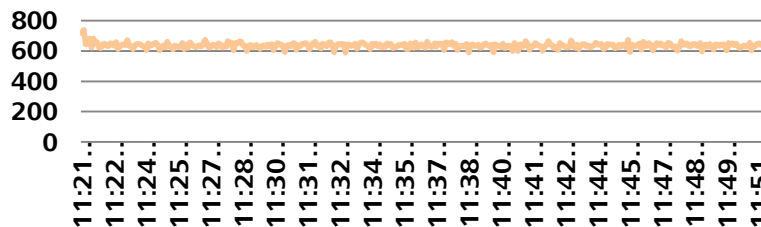


简单过滤规则 单进程最大处理能力在 45万条/秒

Read Input



%CPU



- 以上指标均为单进程测试结果
- 每个物理节点部署有三个同样的进程，实际单物理节点吞吐量为以上值的2.5倍左右
 - 30组KPI计算为 $7\text{万} \times 2.5 = 17.5\text{万}$
 - 简单过滤规则为 $45\text{万} \times 2.5 = 112.5\text{万}$

NoSQL+NewSQL技术的实际应用案例

运营商要求通过对从移动通信网络中采集的信令面和用户面数据进行分析、处理及存储，从而协助运营商全方位的分析网络运营情况，同时需要提供接口，供运营商根据自己需求定制第三方应用。

- 全方位的网络运营分析能力：

- ✓网络分析：主要从服务器（服务器的IP）、网站、协议、终端类型、接入网络类型、位置信息、用户组信息等维度呈现整个网络业务情况的概貌。

- ✓用户分析：包括Top N用户分析和用户详细分析。

- 客户化的操作维护系统

支持近端、远端多用户的同时访问，以及对服务器集群的集中管理。

- 强大的报表处理能力

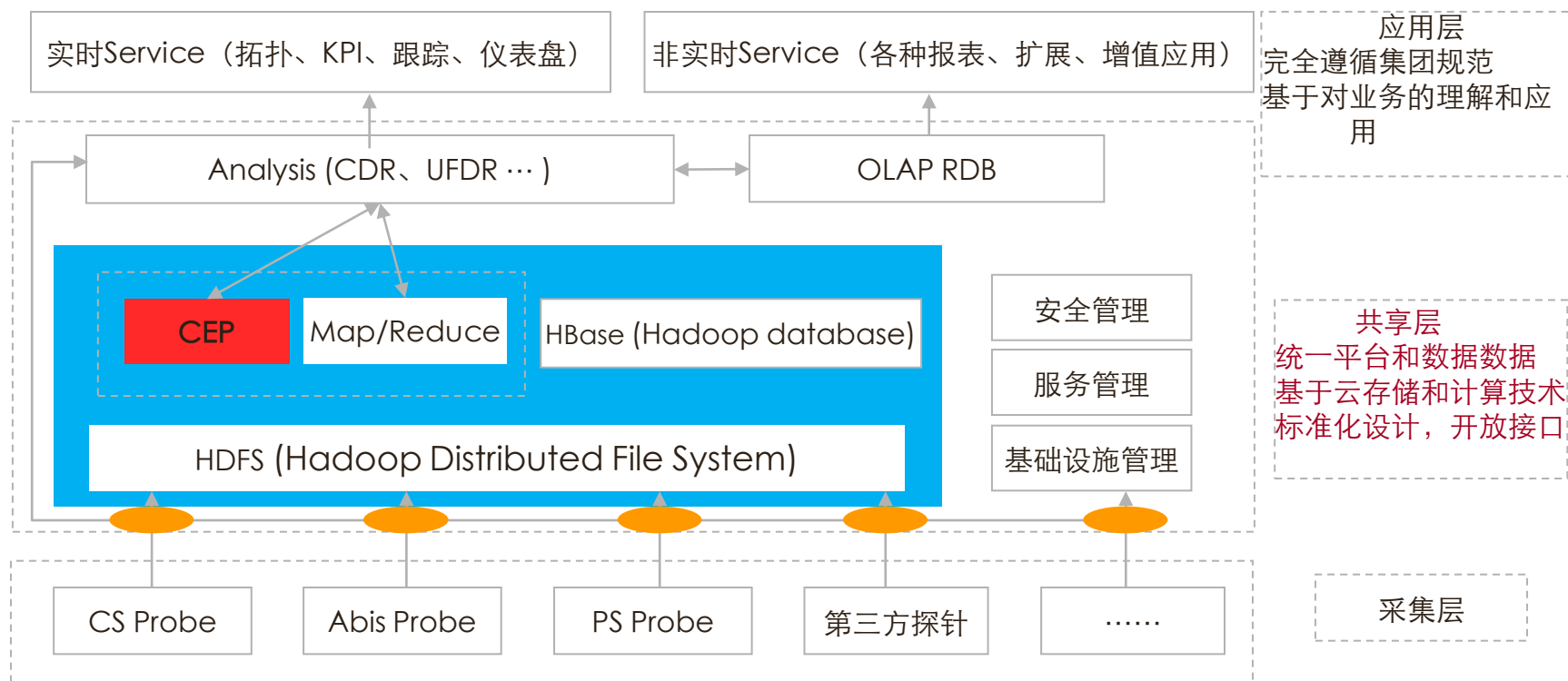
支持多种报表展现方式，如表格、折线图、饼图、面积图、堆积面积图、簇状柱形图、堆积柱形图等，方便进行分析。

提供周期报表生成和自定义桌面以及自动邮件发送报表功能。

- 大容量的数据存储能力

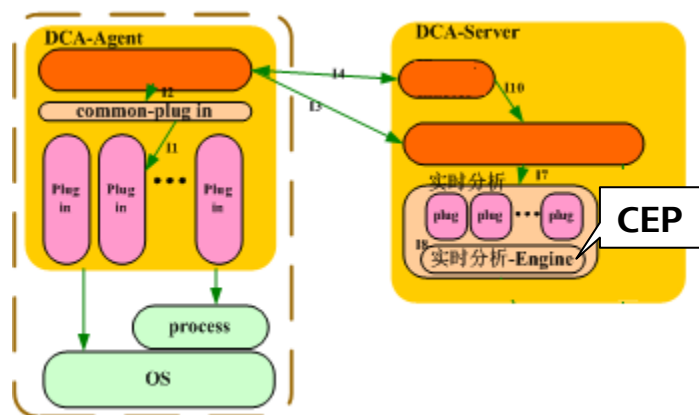
支持本地硬盘存储和磁盘阵列存储。

华为SmartCare信令监测平台战略



- 移动网络业务监测系统需要对大量信令数据进行分析、挖掘、存储和管理，对异常做出实时响应
- 其中恶意呼叫、恶意短信、疑似单通、乒乓切换等，需要复杂逻辑判断、有时序性与实时性要求的监测分析业务均由CEP实现

华为云计算数据中心管理实时监控



- 云OM实时监控，对云系统的运行信息进行收集，并进行实时统计分析，以便了解系统整体运行情况，为系统扩容、资源调度、新业务的开发提供数据支撑和快速相应
- 其中对虚拟机、主机、集群、业务等维度的，CPU、内存、IO、状态等指标的实时KPI计算由CEP实现
- 目前单进程实现了十数万个虚拟机下，几十个维度共数百个监控指标

华为大数据产品成为云计算战略的重要成员

云操作系统**Galax**中已经包含**NoSQL/NewSQL**部件：

1、Hadoop

2、ESP

Q&A

向各位专家学习 共同携手，应对大数据的挑战

会后交流，欢迎联系：walkerzhong@huawei.com



Thank you

www.huawei.com