

延云

千亿大数据实时解决方案

2015/11/07

延云云计算技术有限公司
大数据的未来在大索引,大索引的未来在延云

议程

第一节

- 传统常规解决方案所面临的问题。
- 大索引技术对企业大数据的影响。

第二节

- 关于我们：延云云计算。
- 公司主要产品：YDB的介绍。

第三节

- 应用场景-YDB能做什么？
- 联系我们。

传统常规解决方案所面临的问题

超**千亿规模**的数据，数据库根本就运行不了，怎么办？

数据从产生到能够查询，要延迟一天才能看到，如何能做到**分钟级延迟**？

50台规模的hadoop集群，几亿条数据，一个MR任务要运行几小时，每天也就能进行几百次查询。

如何能让任务的执行时间缩短到**秒级响应**，每天能执行**千万次查询**。

Hbase只接受KV形式的存储，数万个维度的大宽表，如何进行**多维索引**？

Storm流计算能预计算固定的维度、粒度，但业务千变万化，突发事件很多，如何对**任意维度**的组合进行筛选、钻取、统计？

硬盘坏了，机器宕机，怎样做到**数据可靠不丢失**？

小型机太贵，我们买不起，怎么办？

大索引技术对企业大数据的影响

为什么要使用大索引？使用后会有什么好处？

1. 索引大幅度的加快数据的检索速度。
2. 索引可以显著减少查询中分组、统计和排序的时间。
3. 索引大幅度的提高系统的性能和响应时间，从而节约资源。

正因为大索引技术可以显著的降低大数据的处理成本，显著提高大数据系统的执行效率，延云自主研发了自己的分布式大索引系统YDB。

大数据的未来在大索引，大索引的未来在延云。

公司简介

- 延云云计算技术有限公司：
国内第一家“**千亿级实时多维分析**”解决方案提供商。
旨在为企业提供千亿级大数据的高可靠、低成本、高性能的实时大数据解决方案。
- 我们的核心产品YDB：
是我们自主研发的一个大型分布式索引系统。旨在为总量在万亿级别、每天千亿级别数据增量的项目提供近似实时的数据导入，并提供近似实时响应的多维查询与统计服务。

我们的团队

- 延云的核心开发团队成员大多出自阿里腾讯的高级开发工程师，大多都有超10年的从业经验，也有一些留英归来的硕士。
- 团队成员中也不乏一些开源项目的核心committer，如 [JStorm](#) 与 [Mdrill](#)。
- 我们的成员曾在阿里腾讯内部设计出多个千亿规模的大数据系统，部分项目在业界也是大名鼎鼎。

11月19日阿里巴巴JStorm正式成为Apache Storm的子项目

10月7日腾讯Hermes集群单日数据规模突破3600亿，规模超过600台。

YDB的特点

千亿规模

每天千亿增量，总数据量可达几万亿

低延迟

数据从产生到能查询，根据配置的不同一般在十几秒到几分钟

查询快-高性能

常规查询毫秒级响应
常规统计秒级响应。

实时搜索

长文本字段可以进行根据关键词进行全文检索模糊匹配，并且有较高的性能

多维钻取

支持上万个维度，任意组合查询，任意维度组合过滤、分组，统计、排序。

容灾可靠

索引存储在分布式文件系统中，不因硬件的损坏或异常宕机而丢失数据。

YDB的功能

- 多维检索

支持 =,<>,>,>=,<,<=,in 以及全文检索, 模糊匹配

- 统计

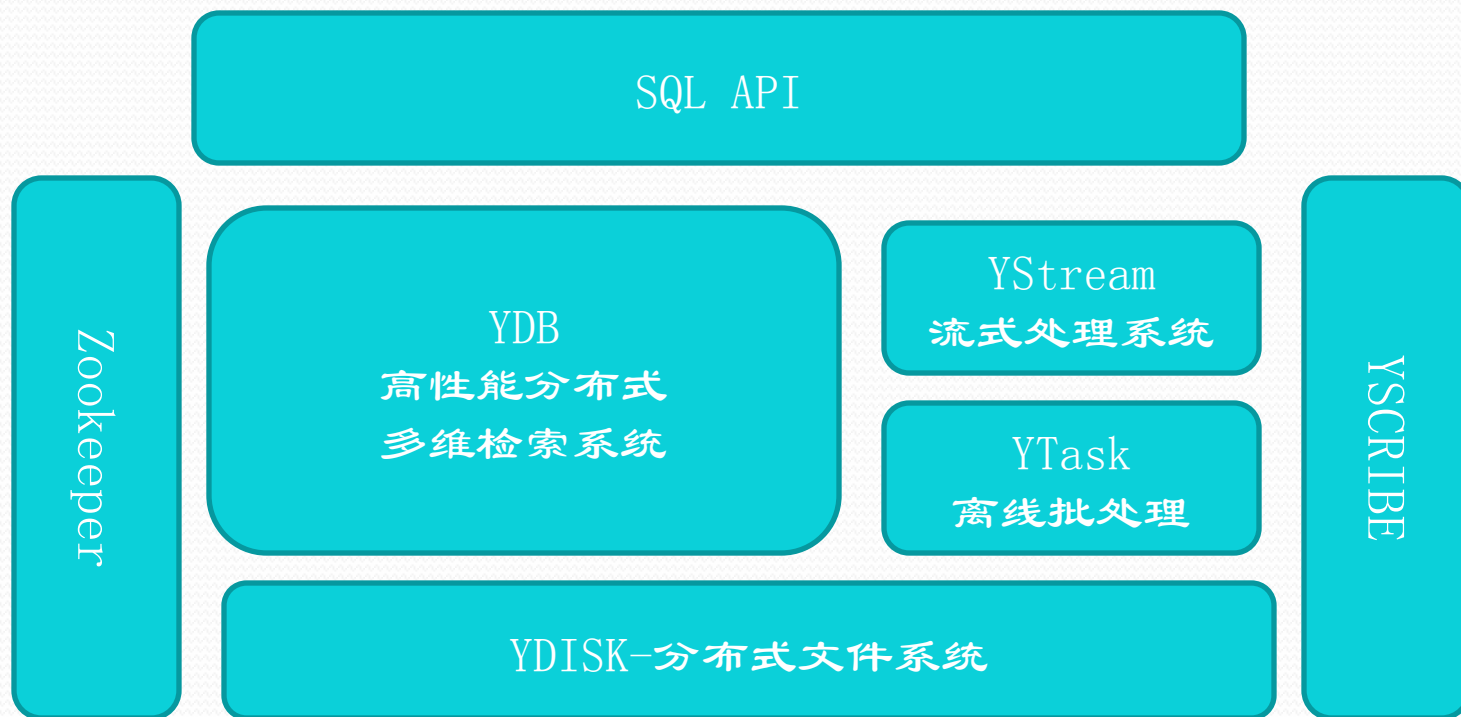
支持count,sum,max,min,avg

- 多列分组 (group by)

- 多列排序 (order by)

更多的用法请去官方下载 延云YDB安装与使用说明书

YDB系统成员概览



YDB检索过程示意图

请求分发与合并

独立计算

分布式文件系统

用户查询
SQL API

查询引擎

子查询一

查询引擎一

子查询二

查询引擎二

子查询...

查询引擎...

子查询N

查询引擎N

读取索引一

索引分片一

读取索引二

索引分片二

读取索引...

索引分片...

读取索引N

索引分片N

YDB对故障的处理

指标	描述
存储节点故障	数据存储在YDISK上，多份冗余，数据不受机器硬件故障影响
查询服务节点异常	服务异常可以感知的自动重启，硬件损坏，服务可以自动迁移到其他机器上继续服务。

应用场景-YDB能做什么？

应用场景-用户画像



熟话说：知己知彼，百战不殆。用户画像就是在基于大量真实数据的基础上，从多个纬度真实反应用户特征，挖掘用户需求。从而在产品的设计、运营策略等环节上，根据目标受总用户的特征进行更符合用户需求的设计和运营方式。具体到不同行业会有不同的应用方式和应用场景。例如：广告行业会进行精准广告推送，交易类平台进行用户个性化推荐，内容类网站进行内容优化和内容推荐。总之用户的特征和需求明确后，所有的活动将具有针对性，有助于提高服务的质量和产品的ROI

棱镜门-大数据监听计划



棱镜计划（PRISM）；是一项由美国国家安全局自2007年起开始实施的绝密电子监听计划。

根据斯诺登披露的文件，棱镜"监控的主要有10类信息：电邮、即时消息、视频、照片、存储数据、语音聊天、文件传输、视频会议、登录时间和社交网络资料的细节都被政府监控。通过棱镜项目，国安局甚至可以实时监控一个人正在进行的网络搜索内容

YDB可以为这其中的海量数据提供实时的存储以及即席的搜索服务。

因YDB的数据时效性较高，并且检索速度很快，该领域未来在工信部以及公安系统上会有较大的应用前景。

应用场景-精准广告营销

● 个性化推荐

是否想过有一天当你在地铁上、公交上、电视上、马路上的大型电子显示屏或墙壁广告，他们可以感应到你，播放的广告都是为你量身定制的，都是你真正需要的。

● 富余服务能力的消化与精准投放

你是一个咖啡店主，当你的店比较清闲的时候，是否想过使用YDB搜索下周围3公里内的小资人群，告诉他们你这里有一个比较雅致的咖啡店，而且给他打5折，而当你的咖啡店人员比较满的时候就不在推送这些折扣服务。

● 阿里和腾讯已经涉水，你还在等什么？

腾讯的ADS与阿里的达摩盘升级为公司的战略产品。

YDB为此而生，立即构建属于你的DMP（大数据市场）吧。

智慧运营商

- 运营商拥有多年的数据积累，其数据资源的广度和深度是移动互联网企业难以相提并论的。每天数千亿的数据如何及时并快速的分析，是摆在运营商面前的一个难题。
YDB本身就是一个千亿级的实时即席分析系统，可以在如下几个方面帮助运营商管理数据。
- **流量精细化**
利用YDB保存详细的客户终端信息、手机上网行为轨迹，上网时长等数据。可以即席的检索出用户的上网记录，可以针对众多的客户投诉，提供更精细化的客户服务。
- **套餐精准营销**
基于客户的位置、话单，上网行为等，实时的筛选用户，进行更加精准的电话营销。
- **舆情监测-预防犯罪**
对短信，用户行为，位置进行实时的即席分析，进行多维关联，打击与预防犯罪。
- **基站优化**
实时的用户流量监控，对链接负载严重的基站进行升级扩容。
- **对外数据服务**
运营商可以利用得天独厚的大数据资产优势，利用YDB千亿级的大数据处理能力，将数据封装成服务，提供给相关行业的企业用户，为合作伙伴提供数据分析开放能力。

应用场景-图书馆与专利检索系统

- 仅仅根据书名，类目的检索太弱了，有没有想过我输入一段话，或一个人名，告诉我在哪本书的第几页里出现过。
- 一本书通常有几百万字，一个图书馆通常有数百万的藏书，这个搜索量不是普通数据库所能解决的。
- 国家专利局达几百T的专利数据的全文检索。

换成YDB吧，轻松帮你解决。

智慧城市

- 交通流量热点展示

利用延云的流式系统可以实时的对每个监控点的数据进行采集、统计与计算，从而实时的获取车辆流动信息以及交通拥堵情况。

- 交通规划

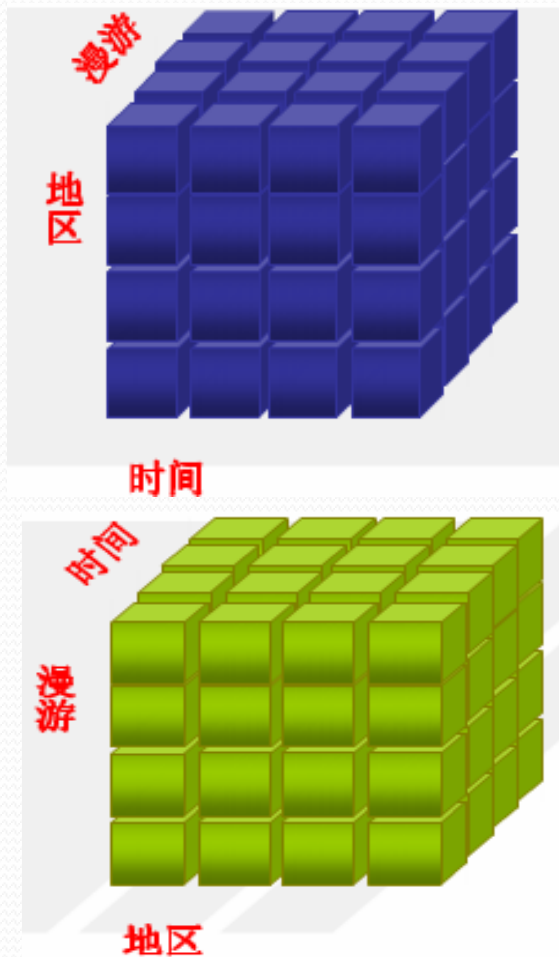
通过延云的 交通流量热点展示 大屏。

可以了解人群动向，对交通道路规划建设有参考作用。

千亿日志全文检索与多维分析

- 物流系统，网站，运营商，证券交易所，零售商每天有大量的销售，访问日志。经常需要对这些日志进行分析、过滤、筛选 从而发现系统潜在的问题。
- YDB在承接了大量千亿级的系统，对千亿级别的数据进行低延迟的导入，快速的多维分析，多维筛选。
- YDB可以做到每天接入 3000亿的系统日志，总数据量可达3万多亿。

OLAP-多维分析



- 切片与切块

在确定某些维数据的情况下对其他维进行观察，在多维结构中对二维数据切片，三维数据切块。如对城市，时间，漫游维度进行切块，可以得到城市的漫游费用情况

- 钻取

- 可以再一个维度从高到低或者从低到高钻取，了解不同深度的数据情况

- 旋转

对数据按照不同维度组织与考察

业务合作

- 电话：024-83653716
- 微信：ycloudnet
- 邮箱：ycloudnet@163.com
- QQ：1820150327
- 主页：<http://www.ycloud.net.cn>
- 地址：沈阳市浑南区新隆街万科明天广场



谢谢