CloudyBI

基于云计算技术的海量数据解决方案

通过日均百亿世界级大数据考验！

上海云数信息科技有限公司

海量数据的挑战

- 复杂查询耗时过长，甚至无法完成
- 系统面对大并发时，性能急剧下降
- 传统数据库可扩展性差，增加硬件无法有效提高性能
- 海量数据的存储和处理，需要昂贵的投资及维护费用

怎么办？

Teradata?——太贵了！

Hadoop？NOSQL?——好像这个解决关系数据库的复杂查询问题

也不是，再说会用这些技术的人也很贵啊！

???

CloudyBI——基于云计算技术解决海量数据挑战！

上海云数信息科技有限公司 http://www.cloudybi.com
上海市联航路 1588 号（上海市 863 软件园）1B408 电话：54325046-608 手机:18616507502

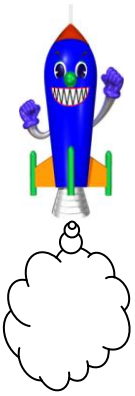
CloudyBI 产品特点：

■ **快：独特的并行 BI 架构，革命性的提升查询速度！**

云计算的核心在于数据的分布式存储与大规模并行计算，google,yahoo,百度，facebook 等互联网巨头正是利用这一技术来处理它们后台的海量数据，从而获得巨大商业利益。CloudyBI 将这一技术与传统的关系数据库技术相结合，创造出独特的基于云计算技术的 BI 系统，为基于关系数据库的海量数据处理带来巨大的突破。

海量数据处理速度对比

传统数据仓库与 BI 解决方案



CloudyBI 解决方案

■ **稳定可靠：智能节点替换技术，系统性能稳定可靠。**

在节点坏掉的时候，可以自动用备份节点替换掉故障节点，保证系统的稳定性。互联网等巨头使用的 hadoop 分布式系统，通常对一份数据保存 3 次，每份数据都放在不同的服务器上，当某个服务器宕机后，系统仍然能通过其他备份节点访问数据。CloudyBI 采用和 hadoop 类似的多重备份模式，一份数据，多重备份(CloudyBI 可以由系统管理员指定采用几重备份)。当有节点宕机后，系统自动用备份节点替换掉故障节点，保证系统稳定运行。

■ **并发能力：多项负载均衡技术，轻松应对海量数据大并发访问。**

通过 dispather 指定任意节点承担 Master 任务，有效消除并行计算中 master 节点工作负荷太重的问题；通过内容索引，二次查询等技术，有效减少按条件查询时对全节点进行扫描，运算的系统资源浪费等问题。在多节点配置下，可以轻松应对上千并发。

■ **可扩展性：核心架构可线性扩展，成倍提升系统性能**

由于系统是完全基于并行的模式设计，因此，当节点增加时，整个系统的 I/O,CPU,内存等呈线性增长，因此，系统性能也能成比例提高。

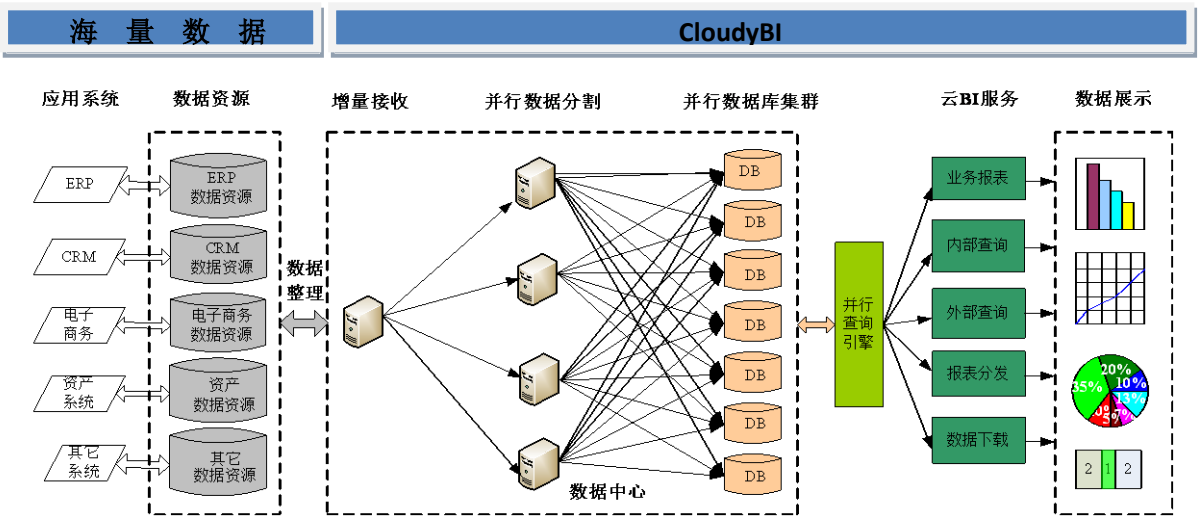
■ **零客户端：完全 BS 架构设计，部署使用方便快捷**

整个系统完全按照云计算的架构，从系统管理，报表和查询的开发，用户使用完全基于 Web 完成。所有操作界面都通过浏览器完成，用户可以使用私有云在企业内部部署，也可以使用公有云的模式，将数据上传到云 BI 中心，让后通过云的方式使用，分析数据。

CloudyBI 产品技术架构

■ **总体架构**

传统的数据仓库是集中在一台大型服务器上的，CloudyBI 则是将数据分割到相互联系的一个集群上。每个服务器上有一小部分数据，整个集群的数据组合成一个完整的数据集。当系统并行运算时，整个系统的 I/O,CPU,内存都远远高于单服务器的架构，从而为数据处理速度带来巨大的提升。



和云计算的 map/reduce/merge 架构相对应,可以将数据分割理解为 map,将每台服务器单独处理的模块理解为 reduce,最后在汇总节点进行再处理则理解为 Merge。

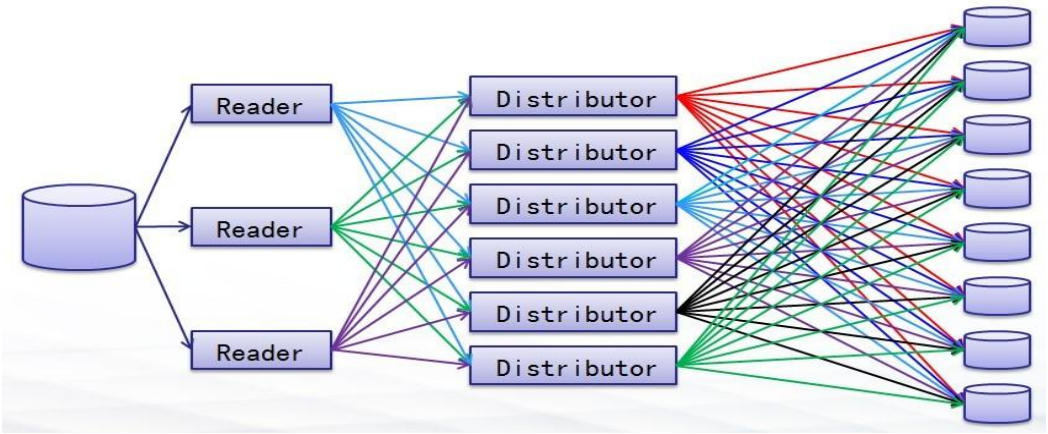
■ **并行数据分割技术**

在并行计算系统中，如何分割数据是整个并行计算的核心问题之一。

简单按字段做 hash 分割，可以快速分割数据，但是对系统带宽有很高的要求，而且系统的并发，可扩展性都有很大限制。(某些国外的系统，超过 20 个节点就要是用万兆网络，且不可以并发)

按业务规则进行复杂的数据分割可以极大的减少节点间数据的交换，降低并行计算系统对带宽的要求。但同时又会带来一个新的问题，就是分割数据的运算量非常巨大（当对一个 1.3 亿条的数据按业务规则进行分割时，单服务器进行的分割时间需要 20 多个小时)。

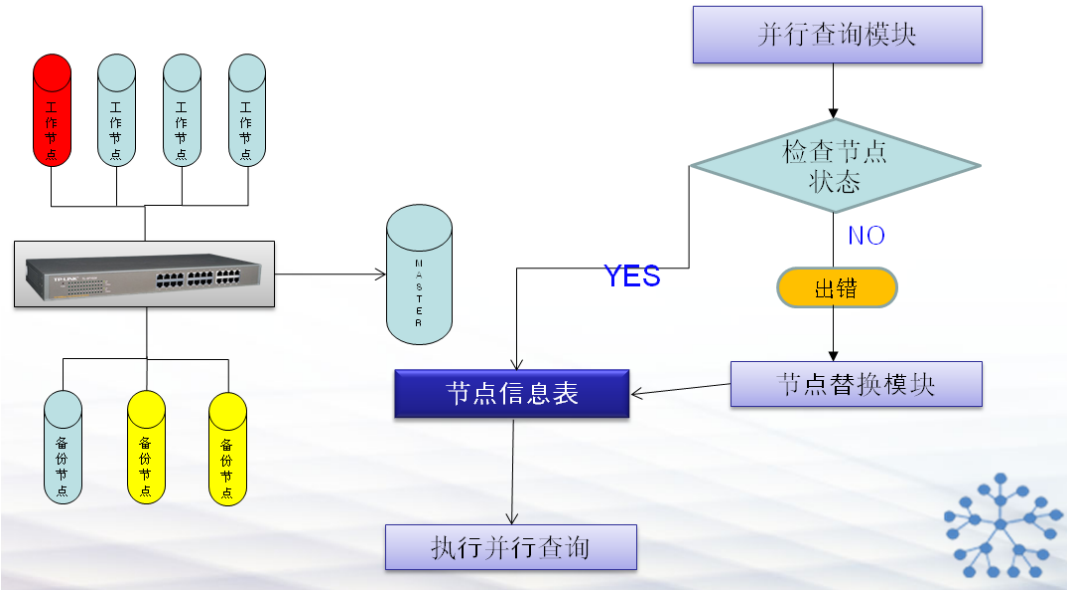
为了有效解决这个两难的问题，CloudyBI 开发了并行数据分割系统，用多机并行的模式按业务规则对数据进行有效分割(对 1.3 亿条数据，9 台机器进行分割，可在 24 分钟内按成)。



■ **容错技术**

系统采用多重备份模式，一份数据，多机存储。当某个节点出现故障时，系统的节点替换

模块会自动更新节点信息，用备份节点替换掉故障节点。用户在前台操作时，丝毫体会不到系统在后台的操作。对于云计算系统来说，容错技术是系统保持稳定的技术基础。

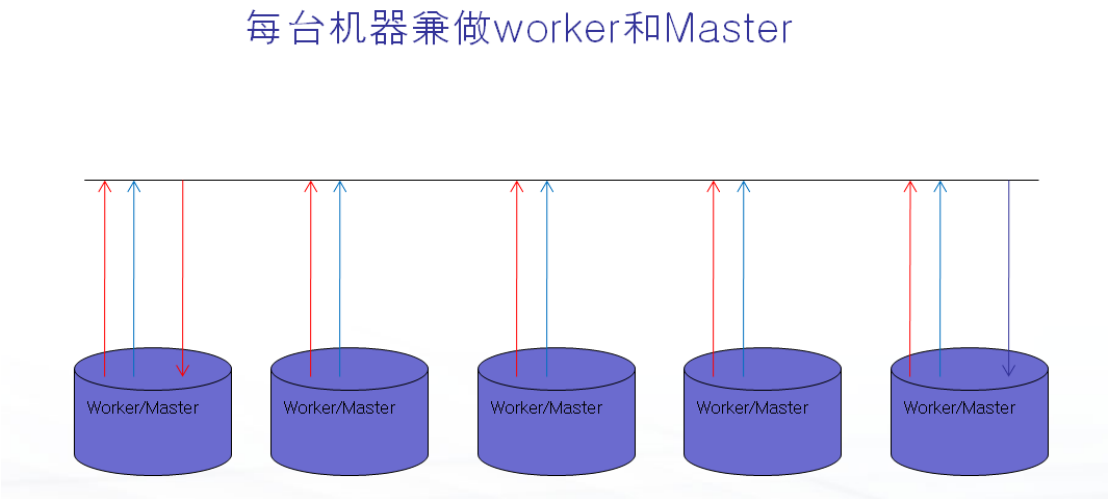


■ **负载均衡技术**

云计算技术面临的一个技术挑战就是要应对大并发的访问。

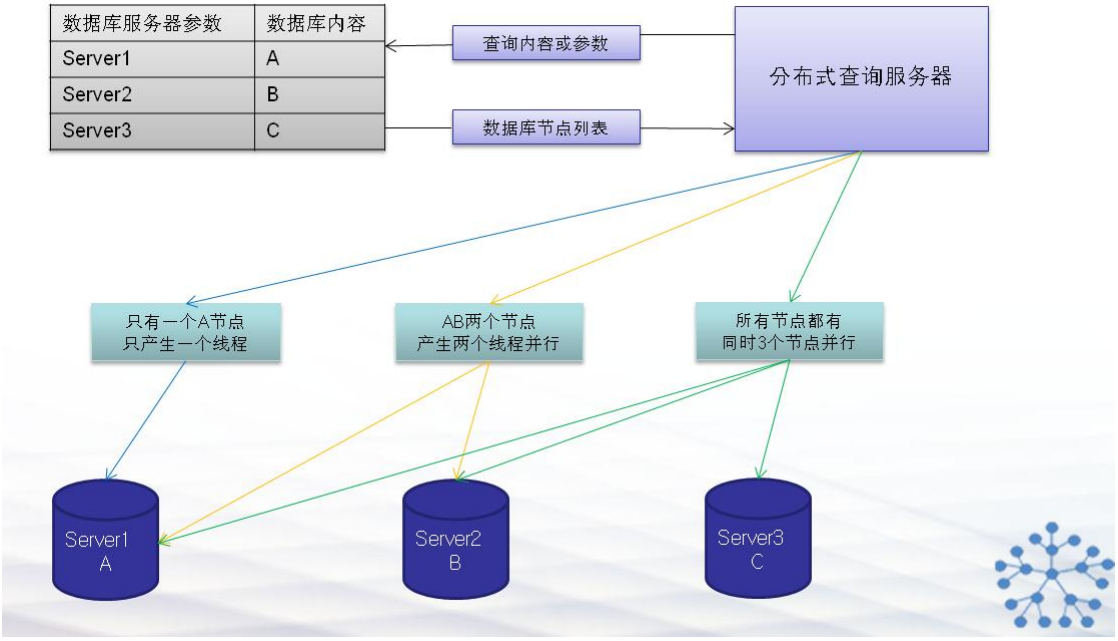
在并行计算时，所有子节点的运算结果，需要由某个汇总节点进行集中再处理。在大并发的条件下，如果这个汇总节点是固定的，那么它的任务负荷一定会非常重，将会造成整个系统的崩溃。 CloudyBI 采用 Master 节点按任务进行负载均衡的技术，可以让任意节点担任 master 的工作，从而极大的提高了并行计算系统应对大并发的能力。

如图所示：红色任务和蓝色任务 2 个任务同时并发，红色任务的汇总任务由 1 号节点完成，蓝色任务的汇总任务则是由 5 号节点完成。

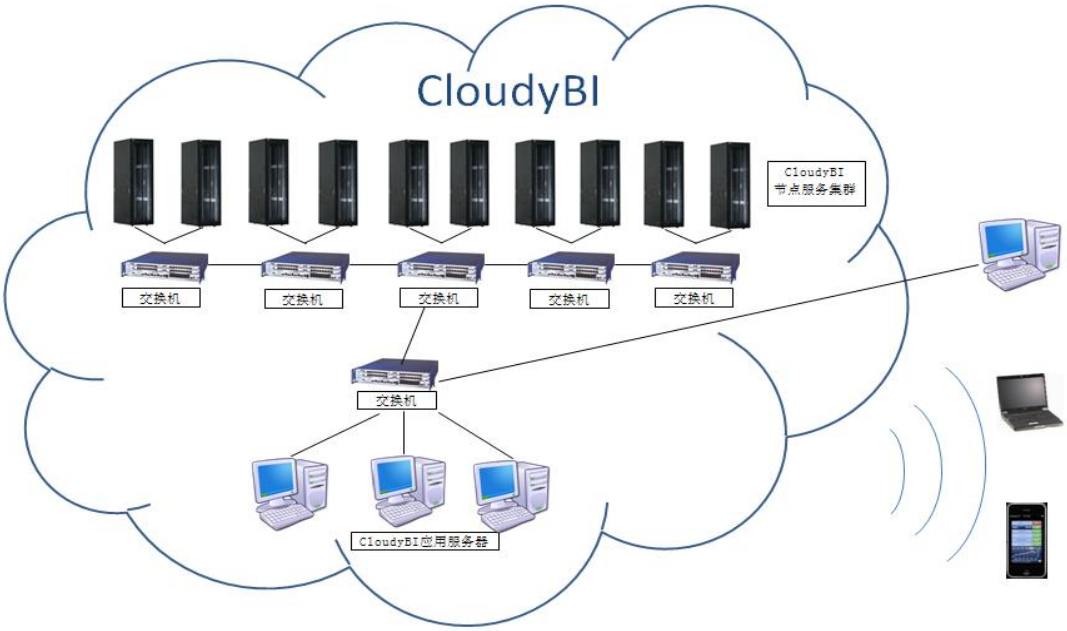


查询时按内容条件进行负载均衡的技术。查询往往包含很多过滤条件，如果能有效的利用

这些过滤条件，锁定他们所在的节点，则可以有效的减少对所有节点进行扫描，运算所带来的巨大浪费。通过指定过滤条件和数据内容索引间的关联，查出查询所需要访问的节点，再针对性的向对应节点发出请求，大大减少系统资源的使用，同时也为大并发提供了更多可用的计算资源。



CloudyBI 网络架构图



CloudyBI 系统配置表(示例)

硬件	硬件配置(可选任意普通 PC 服务器或者 PC)	软件
CBI 应用服务器	PowerEdge R515 2U 机架式服务器 AMD 皓龙™ 4100 x 2/8G 内存/ SAS (15K RPM)：146 GB X 4	CloudyBI Application Server Windows 操作系统 MS Sql server
CBI 工作节点服务器	PowerEdge R515 2U 机架式服务器 AMD 皓龙™ 4100 x 2/8G 内存/ SAS (15K RPM)：146 GB X 4	CloudyBI Node server Windows 操作系统 MS Sql server
CBI 预备服务器	PowerEdge R515 2U 机架式服务器 AMD 皓龙™ 4100 x 2/8G 内存/ SAS (15K RPM)：146 GB X 4	CloudyBI Node server CloudyBI Node server
交换机	GS748TS 48 口千兆智能交换机 可堆叠	

(各种服务器的数量和配置，可根据企业实际的数据量，并发访问数等参数进行具体配置)

公司简介

上海云数信息科技有限公司是全球领先的海量数据处理云计算技术提供商，位于上海 863 软件高科技园区内。公司拥有雄厚的技术积累，尤其在云计算、海量数据处理、数据挖掘、商业智能等方面，产品与技术在国际上具有领先优势，并已申请多项发明专利。我们的 CloudyBI 海量数据处理平台可以帮助企业有效的利用其数据资源，将数据转化为真正的商业价值。 对于大中型企业，我们提供一站式的 CloudyBI 私有云海量数据解决方案。对中小型企业，我们以公有云的方式提供快速，功能强大，价格便宜的云 BI 服务。

云数科技通过与 863 园区和上海市软件技术创新服务平台的合作，基于一流的数据中心、完善的 ITIL 和 ISO20000 运营保障体系、200 多台高性能服务器，构建了以云计算技术为核心的智能云 BI 应用平台，可以承担海量数据处理、测试、查询、统计和分析任务。

“服务客户，持续创新”是我们的使命，云数科技坚持以服务客户为导向，坚持自主核心技术研发，不断引领海量数据处理技术创新，从而为客户提供技术先进、平台开放、价格优惠的云数据仓库与 BI 解决方案。

上海云数信息科技有限公司
上海市联航路 1588 号(上海市 863 软件园)1B408
联系人：李晓华
电话：54325056-608
手机：18616507502
Email:lixiaohua@cloudybi.com