中国航天 神舟通用

神通数据库管理系统
（MPP集群版）
技术白皮书

天津神舟通用数据技术有限公司

目　录

第 1 章 产品简介

神通数据库管理系统（MPP 集群版）【简称“神通 MPP”】是天津神舟通用数据技术有限公司【简称“神舟通用公司”】所拥有的一款具有自主知识产权的高性能和高扩展性的并行数据库管理系统。神通 MPP 以多年大型通用数据库领域的研发实力为基础，集深厚的航天信息化建设经验，集成多项先进技术。为满足航天、政府、金融、电信、电力、能源、网安、审计等行业的海量数据分析统计应用需求而打造的大数据分布式并行计算数据库集群软件。是国家“核高基”重大专项、国家 863 计划在基础软件所取得的一项重要成果。基于神通 MPP，可形成支持交易处理、数据分析与处理等综合解决方案，满足多种应用场景需求。

神通 MPP 主要面向 OLAP 联机分析应用场景，同时兼具 OLTP 联机事务交易应用场景，充分满足混合事务和分析处理（HTAP）应用场景。系统采用 MPP（Massively Parallel Processing，大规模并行处理系统） Share-nothing 架构设计与实现。在数据存储层实现了行列混合压缩存储引擎，在每个服务器内使用多级行列混合压缩技术，存储成本降低 70%以上。提供了强大的吞吐和计算性能。系统提供平滑无感的在线扩容能力，通过数据的水平扩展满足大数据计算需求。系统采用高效的压缩存储引擎和多维度并行计算引擎，通过智能索引技术、MPP 多级并行技术，满足系统的快速检索和复杂统计类查询业务。

秉承实用和务实的精神，神通 MPP 自诞生伊始就结合了行业真实需求，经历了诸多行业典型用户多年的实践和考验。在国内并行数据库管理领域中居于领先水平。可靠支撑了并行计算、海量数据存储等数据业务平台。

第 2 章 体系结构

神通数据库管理系统（MPP 集群版）在整体架构上分为三个层次：接口层、管理服务层、数据节点层。如下图所示：

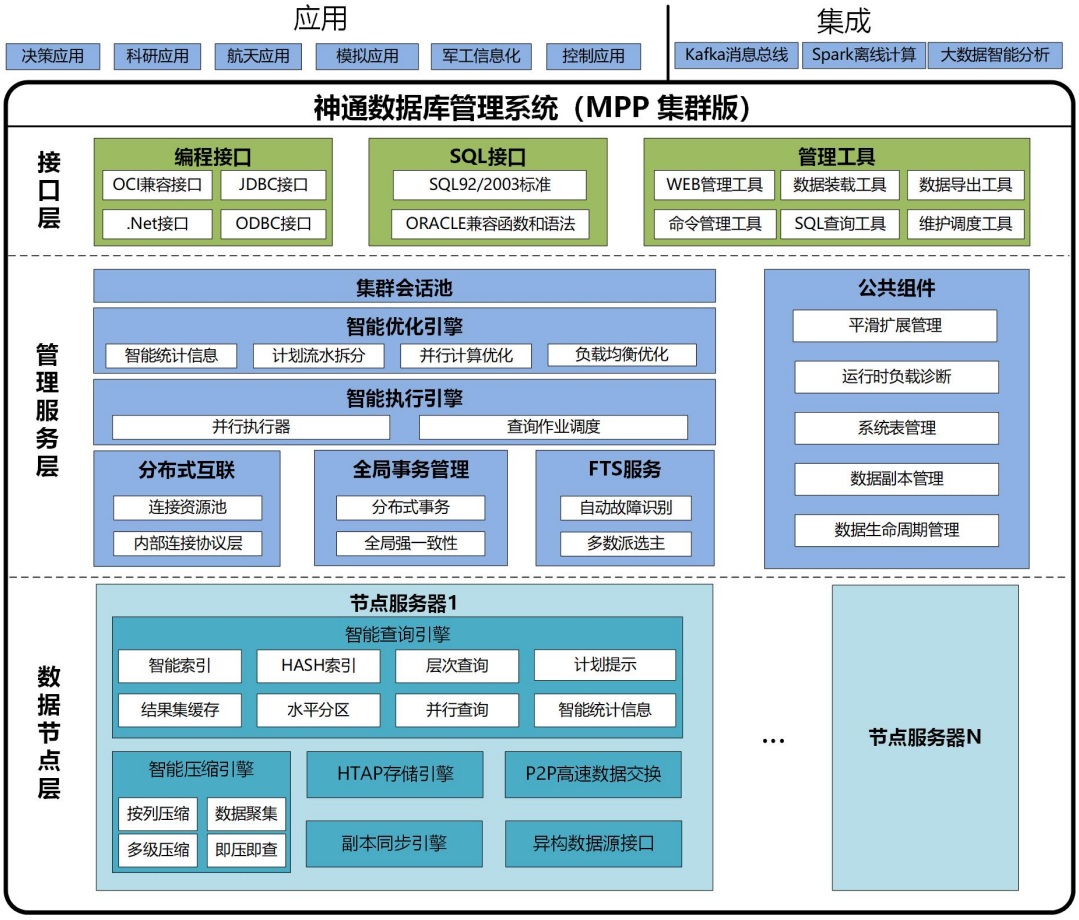

图 2-1 神通 MPP 体系架构图

说明：

- **接口层：**主要完成统一编程接口、统一 SQL 查询接口和统一管理维护界面的功能，达到易管理、标准化和开发性的目的。
- **管理服务层：**主要查询计划的生成和计划执行调度功能，在设计时充分降低了中心节点的计算负载压力，将所有计算、网络和磁盘 IO 负载充分下降，达到系统性能无限扩展的目的。
- **数据节点层：**由多个分布式服务组成，通过专用的内部高速网络进行互联通信，是整个查询真正的执行单元，包括对数据的压缩存储、快速检索和并行计算。

第 3 章 部署架构

神通 MPP 搭建采用在数据平台每个存储端部署计算子系统的方式，通过 MOVE CODE 优化，从而可在网络拓扑层面尽可能的降低网络分发，并提升独立主机的磁盘存储利用率。

在每个系统子网内利用高速万兆网络实现 P2P 进行数据交换。在跨地域模型下，通过定制实现节点网络代价评估模型解决跨地域条件下的查询计划优化。在每个查询地域，部署独立的中心服务节点，以满足本地查询访问需求。

神通 MPP 拓扑结构图如下所示：

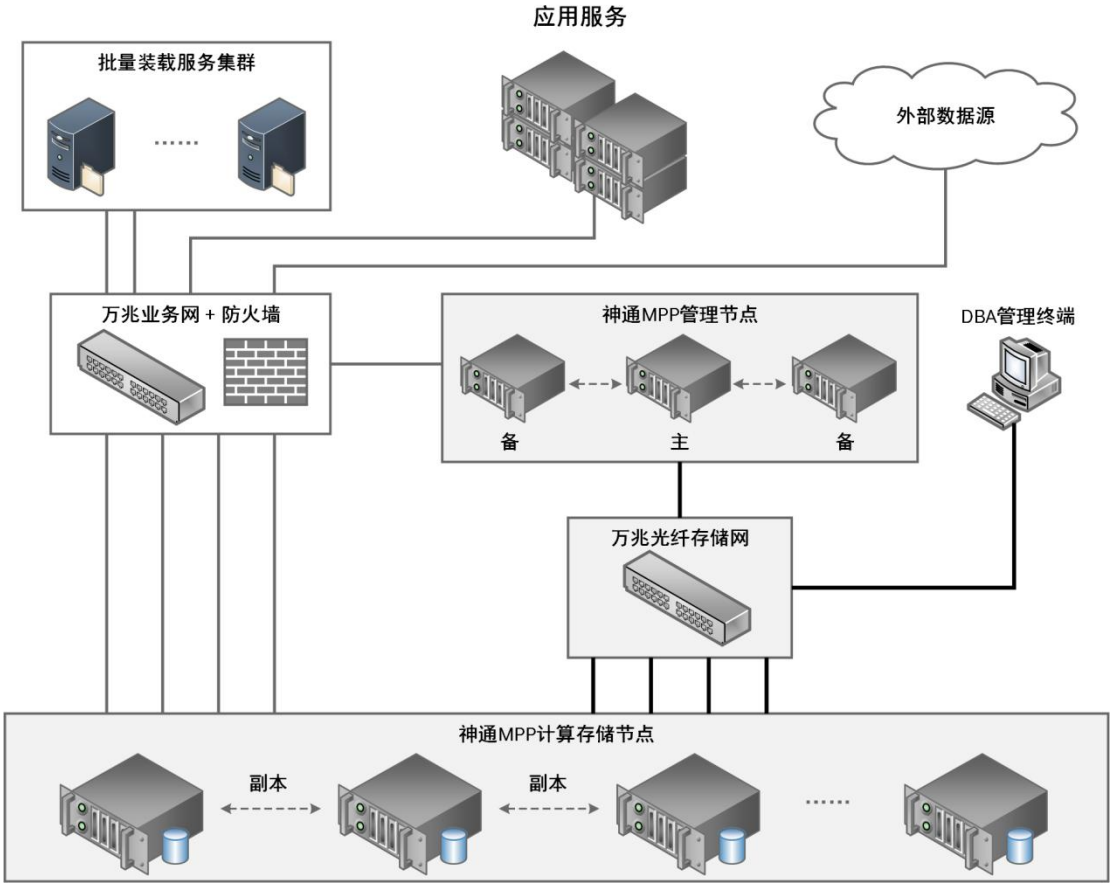

图 3-1 神通 MPP 拓扑结构图

神通 MPP 的部署拓扑如上图所示包括两部分：

- **神通 MPP 后端的部署**

包括管理节点、计算存储节点、存储网络、DBA 管理终端。管理节点采用 1 主 2 备方式部署在 3 台服务器。计算存储节点分别部署在多台服务器，其副本交

叉部署在这多台服务器。存储网采用大带宽低延迟的万兆光纤交换机。管理节点、计算存储节点和 DBA 管理终端全部接入存储网。

- **神通 MPP 前端的部署**

包括应用服务、批量装载服务、外部数据源、业务网络。应用服务需要安装神通 MPP 的驱动或工具。批量装载服务由多台服务器共同组成集群。业务网采用大带宽的万兆交换机并安装防火墙服务。应用服务、批量装载服务、外部数据源全部接入业务网。

第 4 章 产品特点

神通 MPP 经历了审计、电信运营商、农业大数据、公安大数据、互联网安全等诸多行业典型用户多年的实践和考验，日趋完善和成熟。在神通 MPP 设计过程中，充分考虑集中存储的大数据中心和海量结构化数据统计分析的实际场景，拥有如下显著的特点：

- **系统扩展能力：**能够实现系统运行态的资源在线平滑扩展，系统扩展不再需要暂停业务，实现硬件资源的即插即用；
- **数据查询能力：**兼具高效的精确查询和统计查询计算能力，可使用一套存储引擎满足两种类型应用的需求特征。已在多个实际系统中进行使用，综合评测指标优于国内外同类产品；
- **数据管理能力：**能够按照数据的近线和远线特征进行自动生命周期管理，降低系统运维成本；
- **存储成本降低：**采用行列混合压缩存储引擎，通过压缩机制降低系统存储采购成本。当前在实际项目中已实现对电信详单类数据降低存储空间 90%以上，互联网日志类数据降低 75%以上存储采购成本的实际效果；
- **高可用增强：**采用基于全路径优化的多副本策略-高可用设计方案，可实现系统的快速故障转移。在原有的主备机设计方案基础上对系统的高可用性进行了大幅度的提升；
- **方案整合能力：**已打通与 Spark、Kafka、HDFS、Lucene 等优秀开源产品的高速数据交换接口。可实现大数据分析计算、大数据消息队列、海量非结构化数据的整合。

第 5 章 基础功能

神通 MPP 遵循国际标准 SQL 语法，支持丰富的数据类型、内置函数、索引类型、开发接口，并提供了便捷的图形化管理工具和容灾备份策略等诸多基本功能。

5.1 SQL 标准

神通 MPP 为用户提供了方便、灵活的数据查询访问接口。用户可将符合 SQL92、TPC-DS 等标准的查询语句直接在神通 MPP 上执行。神通 MPP 支持 CREATE、ALTER、DROP 等 DDL 语法，支持 SELECT、INSERT、UPDATE、DELETE、MERGE 等 DML 语法，支持 BEGIN、COMMIT、ROLLBACK 等事务控制语法。

5.2 数据类型

神通 MPP 提供丰富的数据类型支持，包括：整型、浮点数值类型、定长变长字符串类型、位串类型、精确数值类型、布尔型、近似数值类型、日期类型、时间间隔类型、时间戳类型、变长二进制类型、二进制&字符型大对象数据类型等。

5.3 内置函数

神通 MPP 提供丰富的内置函数，可根据用户的实际需求定制特殊函数。包括：数学函数、字符函数、时间函数、类型转换函数、聚集函数、分析函数、HASH 函数、IP 操作函数、获取定义函数、系统管理函数、大对象函数等。

5.4 数据库对象

神通 MPP 提供了表空间、表、视图、索引、约束、存储过程、自定义函数、匿名块等常用数据库对象的创建、修改和删除操作，支持数据库用户的创建、删除操作，以及用户权限的分配与回收。

神通 MPP 支持单表最多 20000 个列属性的宽表功能，单列最大数据长度支持 16MB，以此来满足海量数据宽表宽列的存储、管理、分析的业务需求。

5.5 数据分布

神通 MPP 提供多种数据分布方案，以满足不同应用业务类型的数据分布需求。具体如下：

- **基于某一列或几列进行 HASH 分布**

适用于可能会进行复杂连接操作的大规模数据对象，这种分发方式主要用于并行查询的智能连接。

- **循环分布到所有节点（Round robin）**

数据均匀分布各个节点，方便并发。

- **复制到所有节点（Replicated）**

另外，神通 MPP 支持节点内的多级数据分区管理，数据分区模式包括范围、哈希、列表三种水平数据分区模型，可以为每个分区指定物理存储位置。

5.6 存储方式

神通 MPP 支持多种数据存储模式，可同时支持表级行存、列存和行列混合存储，并支持三种存储模式的表混合查询，可根据不同业务场景选择合适的存储方式。

神通 MPP 通过独有的智能压缩专利技术（按列压缩、按行级存储的行列混合压缩存储）为用户提供全生命周期的数据压缩处理功能。可实现近期热点变更数据不压缩、中期数据低级别压缩、长期历史数据高级别压缩的多级压缩配置模式。在保证原有业务需求的前提下，大幅降低企业存储设备采购成本。

神通 MPP 不依赖于存储类型，支持本地存储、SAN 存储、以及分布式存储。支持 NTFS、extX 等主流的文件系统。

5.7 查询优化

神通 MPP 支持了多种查询优化策略（如：逻辑代数优化、规则优化、基于代价的物理优化、基于遗传算法的物理优化、基于成本的全局优化等），可生成最高效的查询执行方案，实现基于成本的查询机制。不依赖用户输入的 SQL，按 SQL 语义信息进行分析和重写，自动选择合适的查询计划，从而提高查询的执行效率。

5.8 计划提示

计划提示功能为应用开发人员提供一种手段用以控制生成计划的方式。基于代价的优化器在绝大多数情况下都会生成正确的执行计划，但是有时也会选择出较差的执行计划。如果优化器生成的计划并不是最优的，应用开发人员可以通过计划提示（hint）功能干预优化器，从而生成更优计划，以满足应用对性能的要求。

5.9 开发接口

神通 MPP 为用户提供了便捷的标准驱动开发接口。支持多种数据库开发接口，包括：ODBC2.X/3.X、JDBC3.0、OLE DB2.7、Unix ODBC、ADO.NET、C API、ESQL(PRO*C)、QT、ACI(OCI)、STCL(C++)、PL/SQL 和.Net Provider 等数据库访问接口，并提供高性能的直接数据访问接口。满足 C、C++、JAVA、PHP、R、Python、Perl 等语言开发需求。

同时，神通 MPP 也提供各种框架开发的支撑接口，例如：与 Hibernate 集成的方言包、与应用服务器集成的数据源接口等。通过使用神通 MPP 提供的各种标准驱动开发接口，用户可将原来使用各种应用框架开发的系统平滑移植到神通 MPP 上。

5.10 在线扩展

神通 MPP 提供了集群在线扩展功能，当集群的负载增加可能超出其承受能力时，可以通过添加节点数据库来提升集群性能，而不需要停止应用业务。集群性能随着节点数的增加呈线性增长。

5.11 容灾备份

神通 MPP 为用户提供方便、灵活的数据备份功能，包括物理备份、逻辑备份，确保系统在出现系统崩溃、服务器掉电、存储介质错误等各种软硬件故障时实现数据的及时恢复。同时神通 MPP 还提供基于日志模式的异地准实时灾备机制。用户可自行根据需要设计并建立完善系统容灾备份体系。

5.12 数据迁移

数据迁移的主要功能是进行神通 MPP 与其它各种异构数据源（如各种关系数据库系统、文本文件等）之间数据的迁移、转换以及合并。

5.13 图形化工具

神通 MPP 提供图形化的管理工具、监控工具、客户端查询分析工具，满足用户对集群日常的监控、管理、以及相关的查询分析业务。

神通 MPP 提供可视化的 Web 管理工具，可设计和管理表、视图等各类数据对象，可执行 SQL 语句进行数据查询。

神通 MPP 提供基于命令行的数据管理工具、查询工具、以及远程数据导入导出工具等，数据导入导出工具可满足多节点并行导入导出需求。

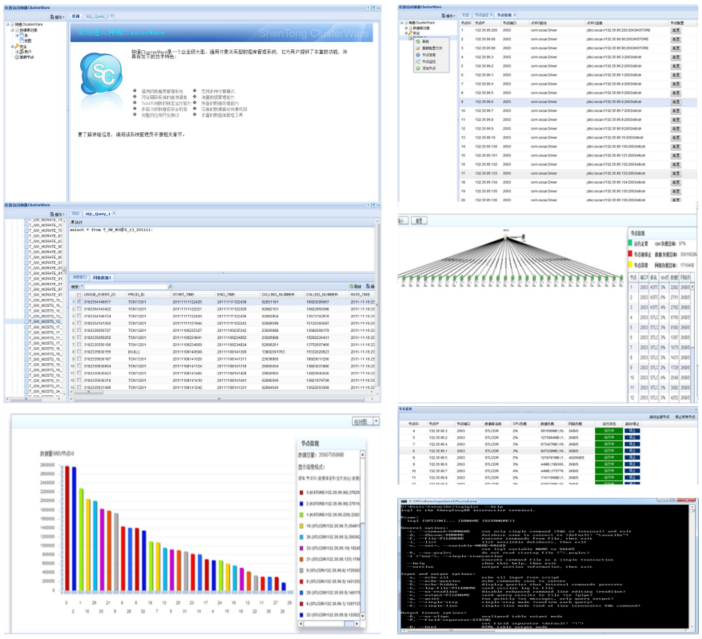

图 5-1 神通 MPP 管理工具

神通 MPP 提供丰富的数据库运行状态监控工具，监控项目包括集群管理服务的运行状态、各存储计算节点的运行状态、各类服务器的硬件资源的使用状态、SQL 语句执行状态、锁状态等。

图 5-2 神通 MPP 运维工具

神通 MPP 提供异构数据库统一访问平台工具，在访问神通数据库的同时，也支持访问 Oracle、SQLserver、MySQL、DB2、SYBASE 等异构数据库，为用户在跨库数据访问操作时，提供统一的访问工具，简化用户操作。

5.14 辅助工具

神通 MPP 提供命令行版的 SQL 查询工具、导入导出工具、备份恢复工具及数据库维护工具，满足了无图形界面服务器环境中的常用管理需求，同时也避免了图形化界面所造成的性能损耗。

- **isql 工具**

采用类似 Oracle SQLPlus 的交互式命令行界面风格，为用户提供良好的 SQL 语句执行环境，同时也提供大量和 SQL Plus 相似的命令模式，例如 desc table 等。

- **brcmd 工具**

以命令行方式完成数据备份和恢复工作。

- **datamigrate 工具**

datamigrate 工具是一个采用纯 C 语言开发的数据导入导出工具。Datamigrate 支持从文本文件与数据库互导；也支持数据库之间的数据迁移处理。由于 datamigrate 在源端和目的端都采用了多线程异步操作模式，能充分利用平台的硬件资源来提高数据处理效率，因此它具备极高的性能，被广泛应用在大数据的导入导出场景。

➢ **数据库维护工具**

通过此工具可实现命令行模式的数据库维护处理，例如：建库、删库、建立自启动服务、实现 mount 模式数据库维护管理。

第 6 章 特色功能

为满足用户实际项目需求，神通 MPP 集成出了诸多贴近项目需求的特色功能，如：软硬件兼容性、Spark、Hive、Kafka 等的集成、混合负载优先级调度、分布式事务强一致、内存总量控制、数据联邦查询等功能。

6.1 软硬件兼容性

神通 MPP 支持多种软硬件平台。支持的硬件平台包括主流的 x86 平台和全国产化 CPU 硬件平台（龙芯、飞腾、申威、兆芯等）。支持的软件平台包括主流 Windows、Linux、Unix、AIX 和国产操作系统平台（中标麒麟、普华、中科方德、凝思磐石、深度、思普等多种通用/专有国产操作系统），提供对应 32 位版本、64 位版本。各种平台上具有一致的数据存储结构和通信协议，使得产品各种组件或工具均可跨不同软硬件平台与数据库服务器进行交互。同时兼容主流的 Hadoop、kafka 等开源系列软件。

➢ **支持多种中间件**

支持东方通、金蝶、中创、IIS、WebLogic、JBoss、Tomcat 等主流的应用服务器。

➢ **支持多种开发语言**

支持 J2EE、PowerBuilder、Delphi、JBuilder、VB、VC、VS .NET、XML、ADA、powerdesigner 等多种具有广泛适应性的开发语言和工具。

➢ **数据源集成**

通过使用神通 MPP 集成的数据源接口，神通 MPP 可作为分析平台软件、数据挖掘软件（如神通 K-Miner 数据挖掘分析系统）、互联网信息爬取软件（如神通 T-Bees 信息采集系统）、全文检索软件（如神通 T-Search 文本检索系统）和文本挖掘软件（神通 T-Miner 文本挖掘分析系统）等系统的数据源。神通 MPP 进行

了深度集成优化，通过就近读取策略和高效数据加载接口进行数据加载。相比普通关系数据库系统，大大提高了数据加载速度，提升了系统效率。

6.2 Spark 集成

神通 MPP 针对自身多种数据存储分布特点，提供基于数据存储位置感知算法的 Spark RDD 集成访问接口，支持以 RDD、Dataframe 和 Spark SQL 三种方式进行数据访问。

6.3 Hive 集成

通过神通 MPP 提供的接口驱动，可支持 Hive 的元数据注册，使 Hive 平台可访问神通 MPP 中的数据。Hive 可通过驱动程序获取神通 MPP 的元数据信息，然后 Hive 的 RawStore 通过神通 MPP 实现的 SZMPPHandler 从神通 MPP 查取数据，通过 InputFormat 和 OutputFormat 接口对数据格式进行规范。

6.4 Kafka 集成

神通 MPP 为提高与 Kafka 间的数据交互性能，定制了可扩展、高可用的专用接口。可将 Kafka 的数据快速消费到神通 MPP 对应表中，专用接口内部直接将数据加载至各个存储节点而不经过 MPP 管理服务，提高数据加载性能，并实现线性扩展。

6.5 开放能力

神通 MPP 支持数据挖掘，可实现利用 MADlib、神通 K-mine 等数据挖掘工具对数据进行深度挖掘。同时支持多机并行数据挖掘技术，提升数据挖掘性能，支持多种数据挖掘算法，包括：时间序列分析，回归模型，朴素贝叶斯分析，决策树，随机森林等。

6.6 正则函数

正则表达式是一种用于文本搜索和处理的语言，它可以用来匹配、查找、替换和提取字符串中的特定模式。在正则表达式标准方面，POSIX 和 Perl 是两个常见的标准兼容性规范。Perl 正则表达式标准是一种功能强大、灵活且广泛使

用的正则表达式标准，因此在很多领域被广泛采用。

神通 MPP 采用了 PERL 正则表达式标准，并提供了一系列正则匹配函数，如 REGEXP_LIKE、REGEXP_REPLACE、REGEXP_SUBSTR、REGEXP_INSTR、REGEXP_COUNT 等，以便用户能够在查询和处理数据时灵活地使用正则表达式进行文本模式匹配和操作。

6.7 混合负载优先级调度

在长时作业和短时作业混合负载下，长时统计分析型查询对 CPU、IO 负载的需求较高，可能会影响短时查询的执行，降低用户的使用体验。为了解决这一问题，神通 MPP 基于 Cgroup 实现了自适应的优先级调度功能，使系统能更好地适应混合负载环境下不同查询需求，从而提高系统的性能和可靠性。

6.8 分布式事务

大多数 OLAP 产品不支持事务，或者支持的事务功能较弱。神通 MPP 提供了集群级别完备的事务特性，用户可以像使用单个数据库一样使用神通 MPP 集群，在应用层无需处理分布式集群各节点数据不一致的问题。

6.9 HTAP 增强

神通 MPP 主要面向 OLAP 联机分析应用，但随着用户需求的变化，越来越多的用户希望产品能够满足包含 OLTP 业务在内的混合负载需求。

为了满足这一需求，神通 MPP 新版本在原压缩存储的基础上，提供了一种新的存储方式，即 HTAP 存储。HTAP 存储在继承原有存储的压缩存储等功能特性的同时，进一步完善了数据的多版本控制机制和对更新/删除等操作的支持，以实现对混合业务负载的更好支持。

6.10 无感 GIC 压缩

为了更好地支持流式加载业务场景，神通 MPP 提供了后台异步线程来处理由流式加载产生的非压缩数据包 GIC。这些异步线程能够高效地处理大量的数据，并且不会影响到系统的正常运行。通过使用这种方法，神通 MPP 可以实现高效而稳定的流式加载数据处理，从而满足用户对于实时数据处理和分析的需求。

6.11 内存总量控制

神通 MPP 引入会话内存总量控制和实例内存总量控制两种机制来避免 OOM 问题出现，并对大 AP 类查询的内存资源使用进行上限控制，避免单条语句消耗过多的内存资源。同时引入内存安全阈值管理策略，尽可能在资源紧张时响应小 AP 类或 TP 类请求。

6.12 数据联邦

神通 MPP 具备访问外部数据源的能力，支持神通同类产品 KSTORE、MPP、OSCAR 和其他异构数据源，如：HIVE、HDFS、ORACLE、MYSQL 等。用户只需将相关数据源注册到神通 MPP 中，就可以通过一条 SQL 语句实现多个数据源之间的联接查询，无需进行数据汇集或复杂的数据迁移操作。另外，神通 MPP 针对同构外部数据源，内部做了极致优化，查询性能接近本地存储查询性能。

6.13 高可用

神通 MPP 的管理节点、数据节点均采用日志传输技术，通过在多台服务器间进行日志传输来实现元数据、用户数据的高可用。神通 MPP 高可用支持故障自动切换、故障恢复再平衡、副本自动上线等功能。

- **故障自动切换**

神通 MPP 架构采用分布式设计，当某个节点发生故障时，系统能够自动切换到其他可用副本节点，确保服务的持续性，从而提高系统的可用性和容错性。

- **故障恢复再平衡**

在节点故障修复后，为了达到故障前的负载性能，神通 MPP 提供特定的 SQL 实现故障恢复再平衡功能。

- **副本自动上线**

无需外部干预，主本节点会内部启动 watch_dog 线程来实时监控其副本的状态。

综上所述，神通 MPP 的高可用功能能够提高系统的可用性和容错性，为企业业务提供更好的应用体验。

6.14 维护性

神通 MPP 能够查看当前执行作业的加锁情况，能够自动分析作业间的锁等待链，并提供运行时诊断视图以便 DBA 进行锁等待分析。能够实现段（表、分区、索引、分区索引等）级细粒度缓存策略的设定和切换。

神通 MPP 提供丰富的系统表和系统视图，可使维护人员快速获取系统元数据信息，包括但不限于：集群节点信息、集群节点状态信息、表信息、列信息、表数据节点分布规则信息、用户信息。

神通 MPP 提供表级数据生命周期管理和规划功能，能够达到自动灵活配置物理层（文件分布）、逻辑层（分区分布）和业务层（存储周期、预分配策略）的生命周期管理能力。

神通 MPP 支持节点穿透查询，支持查看数据库集群中任意单节点数据库上的数据、空间使用情况和语句的执行计划等。

神通 MPP 支持在线进行动态诊断事件下发，对常用内存申请、磁盘 IO 访问等常用操作在线启动/停止诊断事件，启动诊断事件后自动生成该事件的诊断 Trace 信息。

神通 MPP 支持按表空间进行磁盘存储分配规划，能够设定初始化大小、是否自动扩展、扩展大小、最大大小等信息，避免系统长时间运行产生磁盘碎片化问题。能够按段级指定表空间分配方案。

神通 MPP 支持数据节点内哈希、范围、列表和时间间隔 4 种分区类型，支持分区管理，包括分区表导入，分区创建，分区删除，分区 split，分区 merge，和分区 truncate 等。

神通 MPP 支持自动存储管理、自适应内存管理和自调优管理，支持内存自调优、存储自调优和执行计划自调优。

神通 MPP 提供集群拓扑结构监控界面，对节点数据分布情况和空闲空间等进行可视化展示，支持对数据节点和管理节点的 CPU、内存、网络、磁盘 IO 等关键指标的可视化监控管理，自动进行问题告警，并能够对告警进行分级管理。

第 7 章 关键技术

7.1 内外两级缓存技术

神通 MPP 运行过程中，大数据的统计分析往往会占用大量的内存。当内存资源占用过多时，会导致操作系统将某些不常用进程占用的内存 SWAP OUT 到硬盘上，在需要的时候再 SWAP IN 到内存中，由于内存不足而频繁地 SWAP OUT/IN 会对系统性能有较大影响。Linux 操作系统的 OOM 的机制，会监控那些占用内存过大（尤其是瞬间消耗掉大量内存）进程，在必要的时候（为了保证系统进程的稳定和防止内存耗尽）会把该进程 KILL 掉。

因此，为保证神通 MPP 系统的稳定运行，同时避免被操作系统杀掉，神通 MPP 实现了完善的内存总量控制，确保使用的内存在用户限定的范围，同时支持内外存两种缓存机制，来保证当内存不够时，可使用外存来完成计算。神通 MPP 的外存实现是以临时表空间的方式来满足缓存的空间扩展，能够达到缓存空间的相对不限额的目的。

神通 MPP 的排序以及聚集都采用了内外两级缓存技术进行中间结果的缓存，来实现大规模数据处理，满足单节点百亿行级（甚至更大）数据排序、聚集需求。

7.2 并行数据装载技术

神通 MPP 除了提供和 DBMS 一样的 INSERT 数据插入方式之外，还可以通过高速并发导入工具，或者 JDBC 等驱动专门提供的大容量数据导入接口加载数据，用户可以根据实际情况选用。

INSERT 数据插入方式是通过神通 MPP 管理节点将数据分发给计算节点，管理节点很容易成为数据导入的瓶颈。高速并发导入工具或驱动经管理节点引导，可以直接同数据节点通信，并发的向各数据节点导入数据，避免经过管理节点的分发处理和网络开销。管理节点负责生成分布式计划并进行全局事务管理，数据节点负责数据在各节点间的重新分布和数据导入。通过数据节点直连模式，在网络承受范围内，神通 MPP 可以实现数据导入性能和节点数目成正比。

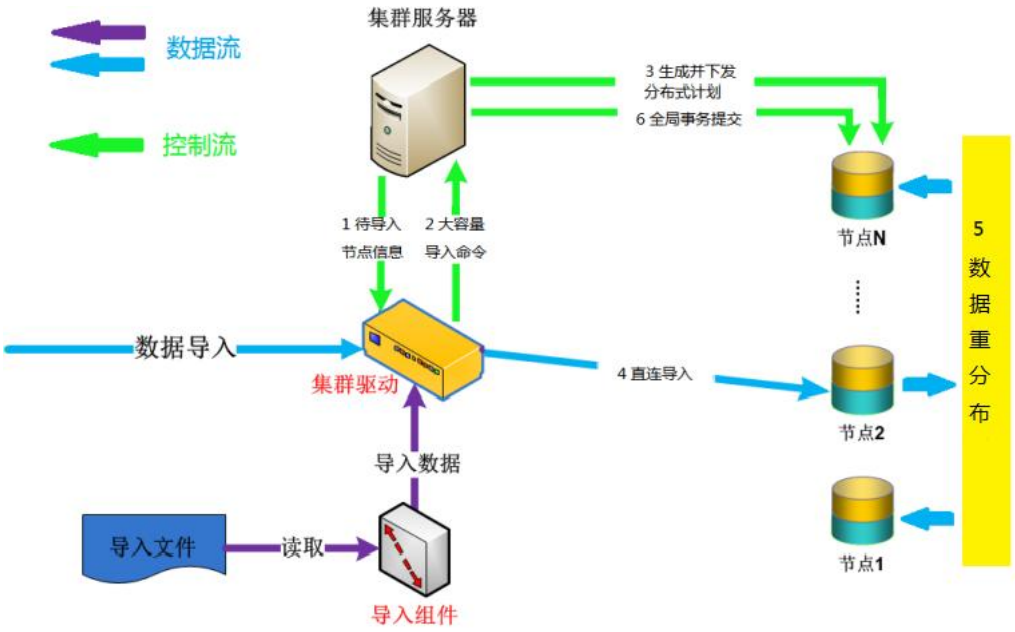

图 7-1 数据高速并行装载

上图所示是数据高速并行装载的流程图。该流程下的数据装载有如下特点：

- **基于分布键的数据重分布**

数据分布计算下降到集群内部，降低驱动负载，防止出现驱动端的性能瓶颈；数据重分布对驱动端透明；分离数据接收和数据导入，两者并发执行，并能平摊负载到各个节点，提高加载性能。

- **面向管理节点的分布式计划生成和全局事务管理**

驱动的数据导入行为、全局事务下的节点数据一致性均由管理节点统一控制，方便管理；数据导入并发量很高时可能会加重管理节点的负载，统一控制方便后续优化。

7.3 WAL 日志流复制技术

为实现节点的高可用，神通 MPP 采用 WAL 日志流复制技术，实现主体和副本之间的数据同步，并保证数据的正确性、完整性和最终一致性。为满足不同应用场景对可用性和性能的不同需求，神通 MPP 提供如下两种模式：

- **系统可用性优先模式**

全部副本完成同步后提交事务，在任何时刻保证副本数据的完全一致，牺牲部分性能，获得更高的系统可用性。

➢ **系统性能优先模式**

遵循“多数派”机制，半数以上副本完成同步后即可提交事务，提升系统性能，同时最大限度的保证系统可用性。

故障数 / 模式	故障数=0	故障数<1/2	故障数>=1/2	
			同步数据部分故障	同步数据全部故障
系统可用性优先	W	R	R	F
系统性能优先	W	W	R	F
R：只读，W：读写，F：禁止访问				

图 7-2 系统性能优先模式

7.4 硬件按需扩展技术

神通 MPP 基于细粒度的数据切片的策略，可以快速实现节点的动态扩展。原有数据进行有限移动，无需停止运行，即可实现计算节点扩展，平滑提升在线系统性能。

神通 MPP 在节点在线扩展的过程中，不需要停止应用业务，只对应用的性能有很低的影响。扩容包括如下两个过程：

➢ **新增节点的实例构建**

基于已有节点的实例库备份和恢复。在实例构建过程中需要在集群级别加 DDL 锁，根据业务类型可能会有或大或小的影响，但不影响数据装载、DML、查询等。

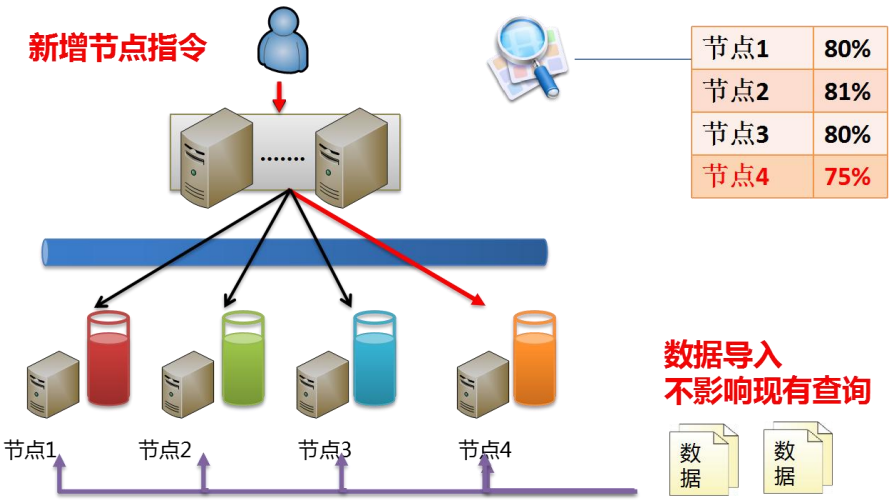

图 7-3 硬件按需扩展—实例构建

➢ **表级的用户数据迁移**

采用基于 DP 增量复制的技术，实现部分原有数据到新增节点的有限移动。在数据迁移过程中，在表级别加 DDL 锁。为了不影响数据装载和查询，分别提供了无锁和加锁模式的数据迁移模式，通过多次的无锁模式迁移，将上一次迁移过程中的增加数据迁移到新节点，最后采用加锁模式迁移（DML 锁），将尽可能少的增量数据迁移到新节点，并删除原有节点数据，实现新增节点和原有节点的数据完整性。上述过程均不会影响查询，对数据装载和 DML 的性能影响非常小。神通 MPP 提供了节点扩容的自动化脚本，以降低上述迁移过程引入的人工操作复杂性。

为了降低数据移动量，神通 MPP 基于表切片的逻辑 DP 为最小复制单元，保障移动切片的逻辑数据一致性，并基于逻辑 DP 更新的 SCN 编号提供数据异步增量识别能力，自动识别事务增量。逻辑 DP 指基于行列混合存储的压缩数据包，SCN 表示 DP 包最后一次更新的序列号，序列号全局唯一且递增。神通 MPP 通过记录源端和目的端 SCN 的同步位置，实现 DP 包的增量移动。

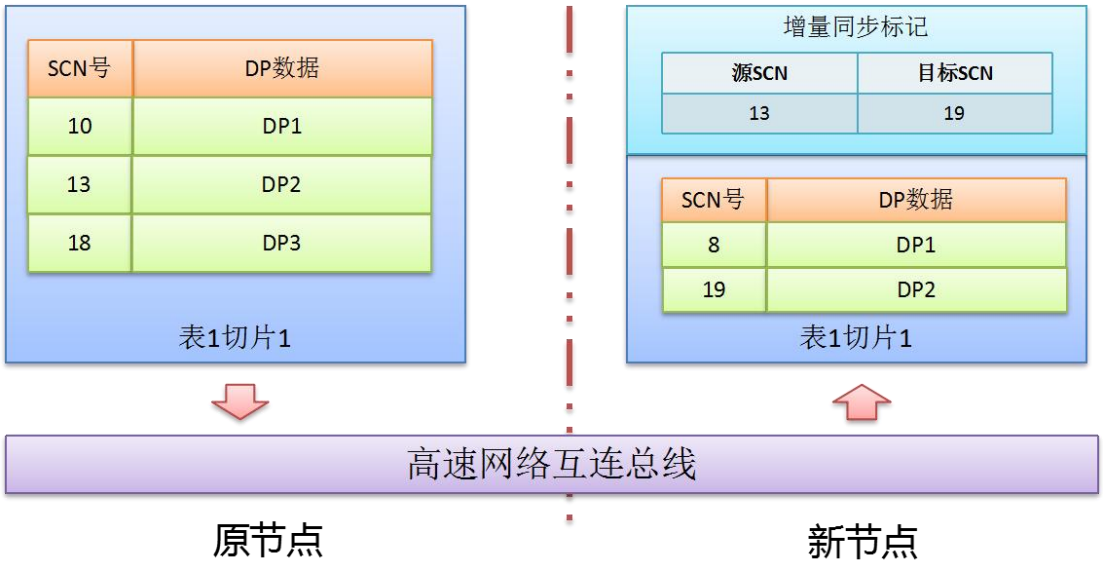

图 7-4 硬件按需扩展—数据迁移

7.5 并行计算技术

神通 MPP 采用全节点的 MPP 架构技术，实现统计查询多节点并发执行；神通 MPP 采用查询计划的多阶段处理策略，实现节点间的全流水化作业模式；神通 MPP 采用并行查询技术，实现统计查询单节点内多线程并发执行。上述并行技术能够充分利用节点计算资源，实现查询性能随集群规模的横向扩展。在计算优化方面基于 MOVE CODE TO DATA 的优化策略，实现低网络负载优化技术。

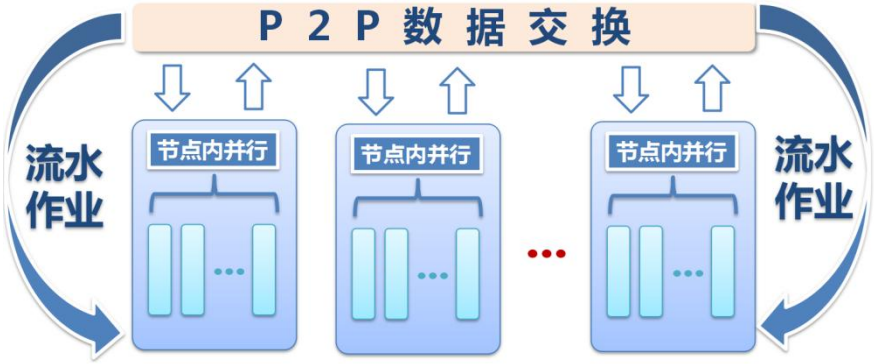

图 7-5 MPP 多机并行

7.6 矢量化运算技术

神通 MPP 主要针对 OLAP 型应用，属于数据密集型计算，查询语句本身比较复杂，采用矢量化迭代模型，可大幅提升神通 MPP 查询性能。矢量化迭代模型与传统数据库采用的按行迭代模型的不同之处在于每次迭代算子执行以矢量为单位，而不是以行为单位。矢量是矢量迭代模型中数据维护和传输的基本单元。

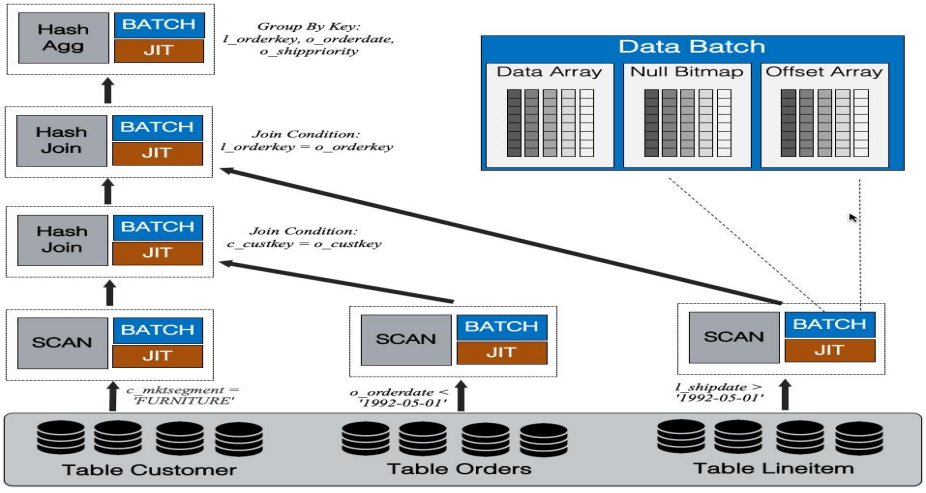

图 7-6 矢量迭代模型

神通 MPP 矢量化迭代模型较按行迭代模型有如下优势：

- **指令/数据 Cache Miss 减少；**
- **相关函数调用次数减少；**
- **处理单行元组的 CPU 消耗较少；**
- **充分地利用 CPU 的 SIMD 指令优势。**

7.7 多级压缩存储技术

神通 MPP 采用 HCC（行列混合压缩）技术实现数据在线实时压缩功能，支持在数据压缩装载的同时进行数据查询。神通 MPP 采用多级压缩（0-9 级）机制，极大的提升了系统运维人员对系统优化的调整弹性。同时在不同压缩级别，神通 MPP 内部会采用具有专利技术的智能压缩算法，按列选择对该列数据最优的压缩算法，整个过程无需任何人工干预，极大的降低了系统的运维难度。在实际的应用环境中最高可以达到 50:1 的压缩比，电信详单类型应用可以达到 15:1，电信账单类应用可以达到 10:1 以上的压缩比。支持压缩数据更新，且满足 MVCC 读不等待特性。

7.8 大数据索引优化技术

神通 MPP 在数据检索方面，基于不同的数据存储模型和计算需求，设计不同的索引优化策略：

- **索引优化技术**

提供基于行列混合压缩存储模型数据的近似有序列范围查询索引优化技术，能够满足基于时间等类型列进行范围查询的需求。

在大数据中心，其数据大多以日志类数据为主，此类数据的数据加载过程中有较强的天然近似有序性。因此如果能够较好的利用这种天然近似有序性，将对系统的查询性能有较大提升。智能索引的设计思想其实十分简单，其核心思想就是先将数据分片（一定数量的行集），然后在每个数据分片的每一列上建立最大值、最小值的统计信息，这样在查询的时候就可利用此统计信息进行数据裁减了，具体设计原理如下图所示：

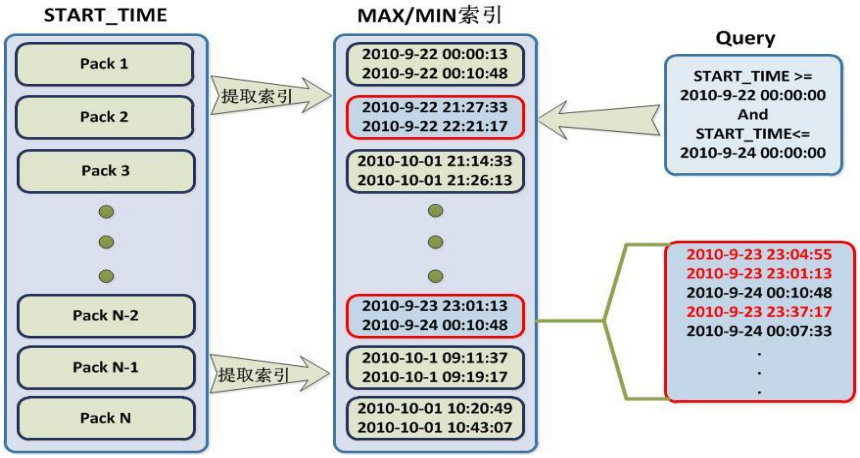

图 7-7 智能索引优化演示图

智能索引的设计目的主要是为了优化天然近似序列的查询性能。同时还可弥补原有关系数据库系统对这种列的检索方案的欠缺。在原有数据库系统中，对于涉及到类似于时间列时，需要建立 B-Tree 索引，而与此同时，在存储和内存开销方面都极大的提高了。在涉及的查询跨度较大时，关系数据库采用水平分区的优化方案，而水平分区方案具有只能基于其中某一个具有这种特征的列进行分区，因此当系统中具有多个这种类型的列时，将无法处理。

智能索引通过简单的优化手段，在仅引入了近似可忽略的存储和内存消耗的情况下，近乎完美的解决了近似有序列的检索问题。

➢ **精确查询索引**

提供基于行列混合压缩存储模型数据的精确查询索引，能够满足高效等值精确查询需求。

- 基于 HASH 稀疏 Bitmap 索引技术

智能索引解决了具有近似有序关系的列的检索问题，那么在大数据的实际应用场景中还存在大量的基于 Key 值的查询，此类查询具有 Key 范围大、引选择率高的特点。原有的关系数据库中的 B-Tree 索引恰恰就是应用于这种场景的，但在大数据情况下，其存在内存开销大，随机访问的磁盘 IOPS 压力大的缺点。

针对这种模式，Hash 索引吸收了原有 BTree 索引和 BitMap 索引的优点，并结合使用 Hash 技术，以一种稀疏模式建立了 Sparse Bitmap 索引。这种索引以数据的 Hash 值为 Bitmap 的入口。在 BitMap 大小方面，采用基于 HCC 的 Package ID 做为其内部键值，从而极大的降低了索引存储成本。Sparse Bitmap 索引的原理如图所示：

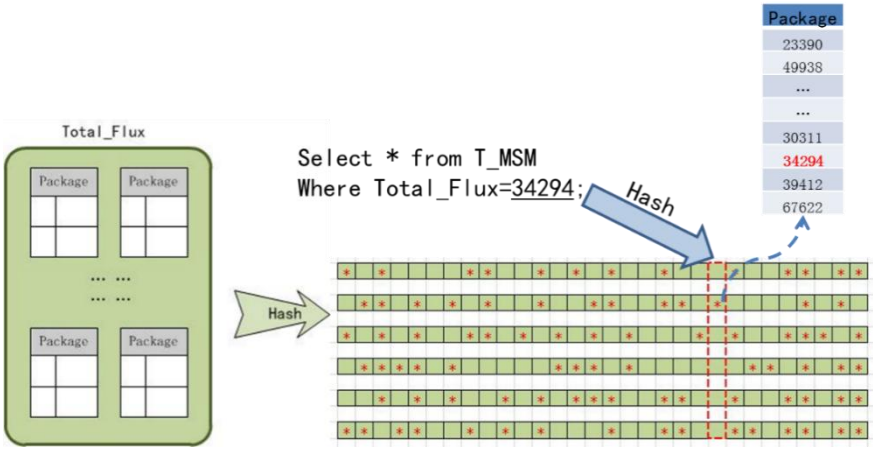

图 7-8 中低密度 Sparse Bitmap 索引原理图

通过这种优化手段，能够实现索引存储空间的大幅度降低，将索引空间由原有的 B-Tree 索引的 30 字节/行，缩小到 2.5 字节/行，这样即使是百亿行数据，单列索引的内存开销也仅为 25GB，完全可适应当前大数据应用的实际需求。

- 前缀树索引

前缀树索引是一种特殊的优化索引技术，这种索引适用于类似手机号码、身份证号等具有前缀编码特点的数据，这种数据的前缀具有非常高的重复率，因此通过前缀编码模式构件前缀树索引，将极大的提高系统的索引压缩率，降低索引的存储成本和内存消耗。在查询性能方面其查询路径等于查询子串的长度，因此其查询效率甚至高于 B-Tree 索引，前缀树索引的存储结构设计所示：

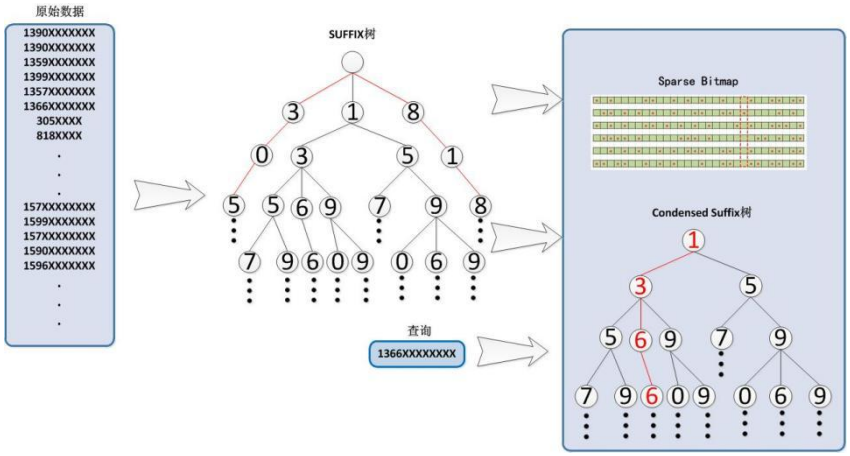

图 7-9 高密 Suffix Tree/Trie 索引原理图

第 8 章 产品指标

参数	限制类型	参数值
集群节点规模	最大值	1024
单表容量	最大值	12288PB
数据压缩比	最大值	与数据间相关性有关，一般在 15:1 左右
块大小	内置	8KB
区大小	内置	64KB
数据文件大小	最小值	4M
	最大值	操作系统支持的文件大小与 32T 之间的小者
单个存储节点的数据文件个数	最大值	30000
单个存储节点的日志文件个数	最小值	1
	最大值	256
表空间数量	最大值	30000
文件路径长度	最大值	255
SQL 语句长度	最大值	2G
存储过程长度	最大值	4G
文本数据的长度	最大值	16MB
VARCHAR 类型列长度	最大值	16MB
CHAR 类型列的长度	最大值	8000
每张表的列数	最大值	20000
查询表的数目	最大值	128
连接并发数	最大值	65535
数据容量	最大值	无限制，受限于操作系统和硬件环境。
单节点的数据库实例大小	最大值	无限制，受限于操作系统和硬件环境。
数字精度	最大值	65

表的个数	最大值	1000 万
表中一行的内部长度	最大值	128M
一个 INTEGER 类型列的长度	最大值	8 字节
数据库名长度	最大值	63 字符
用户名包含字符的个数	最大值	63 字符
表名长度	最大值	63 字符
列名长度	最大值	63 字符
索引名长度	最大值	63 字符
别名长度	最大值	63 字符
MTTR(平均修复时间)	最大值	<1 小时
软件安装时间	最大值	<1 小时
软件升级时间	最大值	<2 小时 100 节点内

第 9 章 运行环境

9.1 硬件要求

神通 MPP 数据库服务器端、客户端管理工具部署硬件要求：

硬件	服务器端	客户端
处理器	--处理器类型：Intel Xeon 兼容处理器或 速度更快的处理器，鲲鹏处理器，飞腾 ARM 架构处理器； --内核数：8 或以上； --处理器速度：2.0GHz 或更快。	--处理器类型：Intel Xeon 兼容处理器或速度更快的处理器，鲲鹏处理器，飞腾 ARM 架构处理器； --内核数：2 或以上 --处理器速度：2.0 GHz 或更快
内存	--最小：4GB --推荐：64G 或更大	--最小：4GB --推荐：8GB 或更大

硬盘	--最小：1GB 以上的空闲空间 --推荐：500GB 以上的空闲空间	--200MB 以上空闲空间
显示器	--VGA 或更高分辨率	--VGA 或更高分辨率
定位设备	--Microsoft 鼠标或兼容设备	--Microsoft 鼠标或兼容设备
驱动器	--从磁盘进行安装时需要相应的 CD 或 DVD 驱动器	--从磁盘进行安装时需要相应的 CD 或 DVD 驱动器
网络	--标准以太网 推荐至少 10Gbps	--标准以太网

9.2 软件要求

神通 MPP 数据库服务器端、客户端管理工具部署软件要求：

软件	操作系统	网络支持
服务器端	--Linux（含国产操作系统） --Microsoft®Windows 系列（Windows NT 4.0 以上版本）	TCP/IP 、UDP
客户端	--Linux（含国产操作系统） --Microsoft®Windows 系列	TCP/IP

根据操作系统的要求，可能需要容量至少为 512MB 的额外虚拟内存（根据实际应用增加）。

实际的要求因系统配置和选择安装的客户端部件的不同而异。在服务器端，随着神通 MPP 中数据量的增加，可能需要更多的磁盘空间。

第 10 章 联系方式

如果您对我们的产品和服务有任何疑问（或合作意向），欢迎随时联系我们：

➢ 神舟通用数据技术有限公司（天津公司）

- 电话：4006-198-288
- 传真：022-5822-1101
- 邮编：300384
- 地址：天津滨海高新区华苑产业区工华道 2 号 8 号楼-1-1

➢ 神舟通用数据技术有限公司（北京公司）

- 电话：4006-198-288
- 传真：010-5989-5589
- 邮编：100094
- 地址：北京市海淀区永丰路 28 号

➢ 神舟通用数据技术有限公司（河北分公司）

- 电话：4006-198-288
- 邮编：050091
- 地址：石家庄桥西区新石中路 377 号物联网大厦西配楼三楼

我们期待为您提供卓越的数据库产品和优质服务，与您共同实现业务的成功！

中国航天

神舟通用

神舟通用 通用神州

4006-198-288