华为数据仓库服务（DWS）
day04 高性能加载数据

目录

HUAWEI

DWS：实时、简单、安全可信的企业级数据仓库

数据实时洞见

- 支持流式数据**实时入库**、业务数据准**实时同步**
- 数据**入库即可查，零等待**
- 万亿数据查询分析**毫秒级**响应

极简易用

- 数据迁移工具，最大化降**低TTM**
- 兼容标准**SQL 2003**，内置丰富OLAP函数
- TPC-H、TPC-DS真正**100%**支持

企业级、安全可信赖

- 支持分布式事务**ACID**，数据强一致保证
- 满足史上最严安全合规要求**GDPR**
- **业界唯一**数据库防火墙服务

HUAWEI

华为云围绕DWS提供数据全生命周期解决方案

Data Sources

IoT data

RDS

On-Premise Database

Ingest

- DIS数据集成
- DLF数据调度
- CDM数据迁移
- DES数据快递
- DLI实时流加载

Data Store&Analyze

- DBSS数据库安全
- **DWS数据仓库服务**
- OBS对象存储

Visualize

tableau SOFTWARE

FineReport 帆软报表

YonghongTech Talk with Data

sas

Superset

ZOOMDATA

DATA → INSIGHT → ACTION

HUAWEI

目录

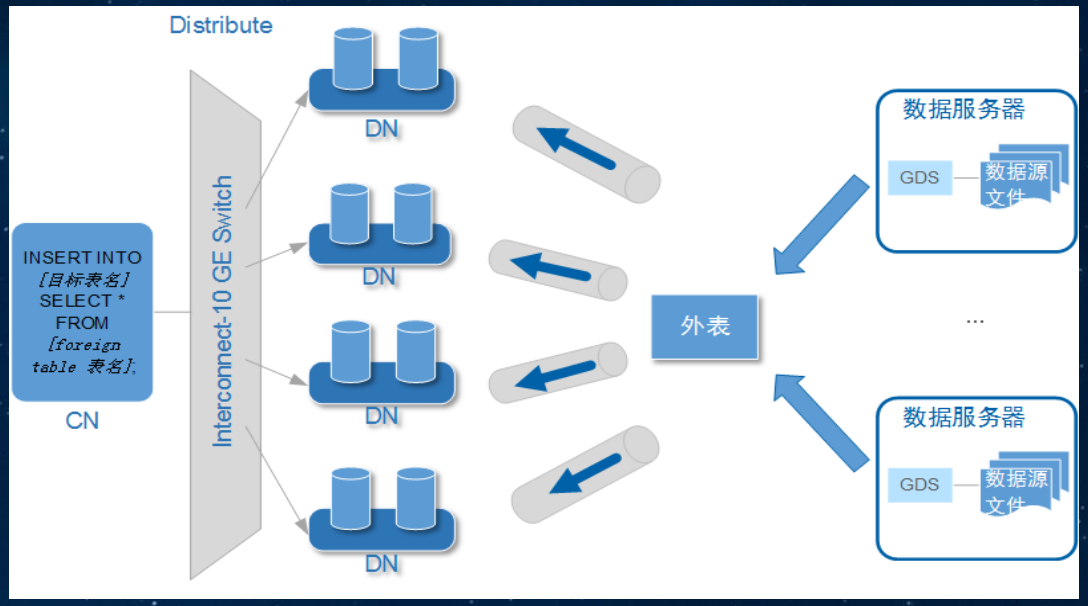
HUAWEI

导入数据方式

方式	特点
GDS	通过GDS工具，采用多DN并行导入，导入效率高。适用于大批量数据入库。
INSERT	通过INSERT语句插入一行或多行数据，及从指定表插入数据
COPY	通过COPY FROM STDIN语句直接向LibrA写入数据。 通过JDBC驱动的CopyManager接口从其他数据库向LibrA写入数据时，具有业务数据无需落地成文件的优势。
Gsql工具元命令\copy	与直接使用SQL语句COPY不同，该命令读取/写入的文件只能是gsql客户端所在机器上的本地文件。

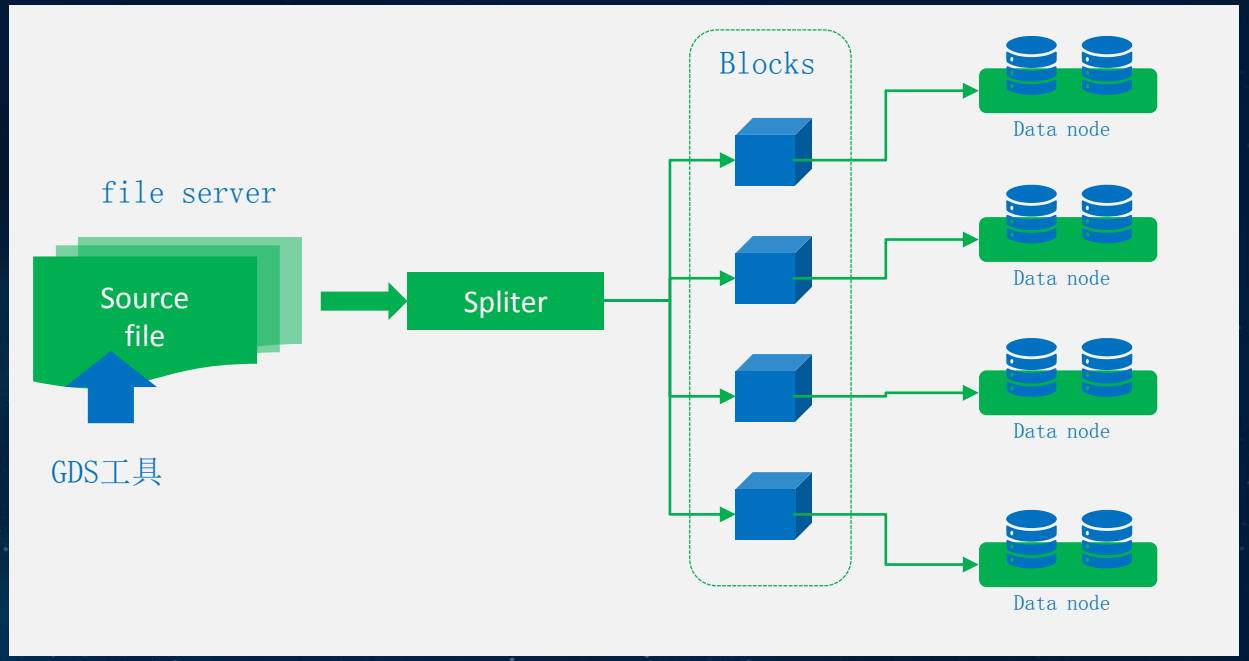
导入数据方式

- **CN（Coordinator Node）**：DWS协调节点。在导入场景下，接收到应用或客户端的导入SQL指令后，负责任务的规划及下发到DN。
- **DN（Datanode）**：DWS数据节点。接收DN下发的导入任务，将数据源文件中数据通过外表写入数据库目标表中。
- 数据源文件：存有数据的文件。文件中保存的是待导入数据库的数据。
- 数据服务器：数据源文件所在的服务器称为数据服务器。基于安全考虑，建议数据服务器和DWS集群处于同一内网。
- 外表**Foreign Table**：用于识别数据源文件的位置、文件格式、存放位置、编码格式、数据间的分隔符等信息。是关联数据文件与数据库实表（目标表）的对象。
- 目标表：数据库中的实表。数据源文件中的数据最终导入到这些表中存储，包括行存表和列存表。



极速并行Bulk Load工具-GDS

并行Bulk Load工具GDS， 实现 x100 TB/天 数据导入



- 利用集群并行数据导入能力，平衡了网络、CPU、IO的资源占用，实现了x100 TB/天的数据导入速度，且随着集群规模的扩展，导入性能线性提升。
- 针对列存、宽表（80+列以上）、数据压缩级别为Low/Middle级别的，对导入性能要求高的场景进行了增强。

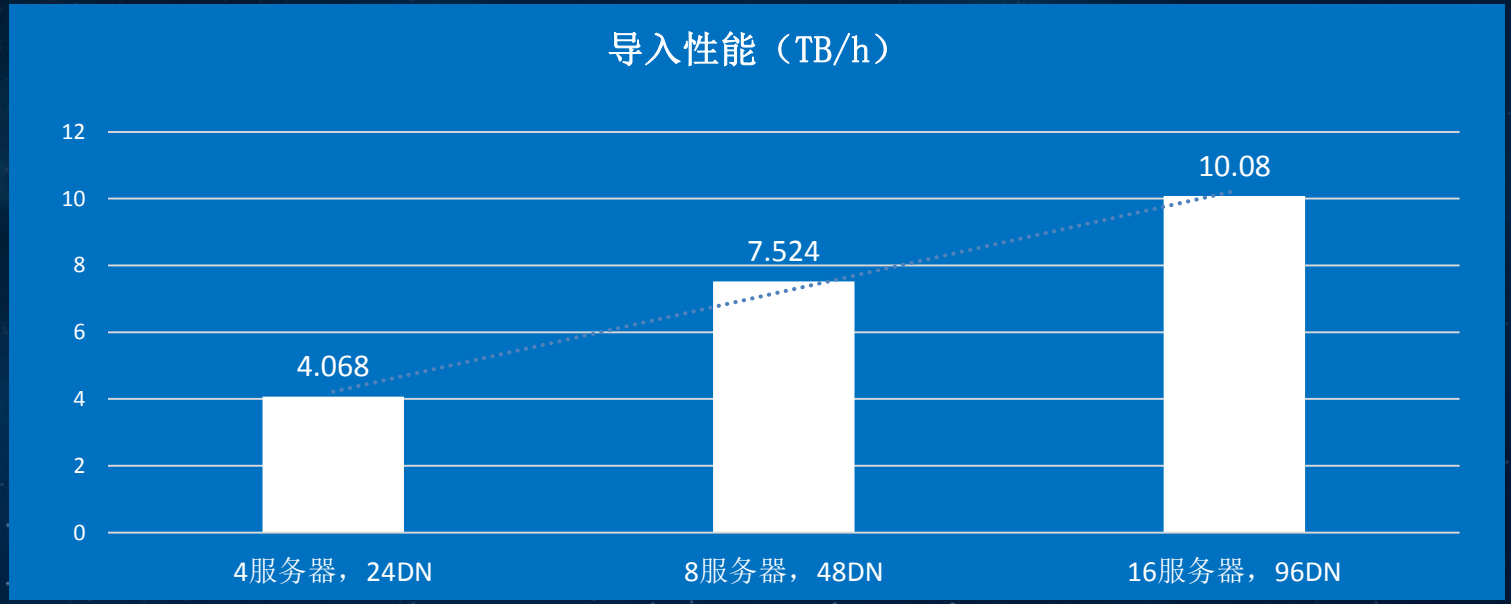
导入性能

业务场景：

将客户OLTP（Oracle等）系统中生成的数据，在**指定时间段**内，导入到GaussDB 200中。

导入性能实测：

随着数据节点数的增加，数据导入性能稳定增长，每日可完成**数百TB**数据导入。



测试环境：4/8/16台RH2288高性能服务器，搭建GaussDB集群。

测试数据：TPC-H 3000X，常见场景——分区表、低压缩级别的数据导入。

Analyze，VACUUM

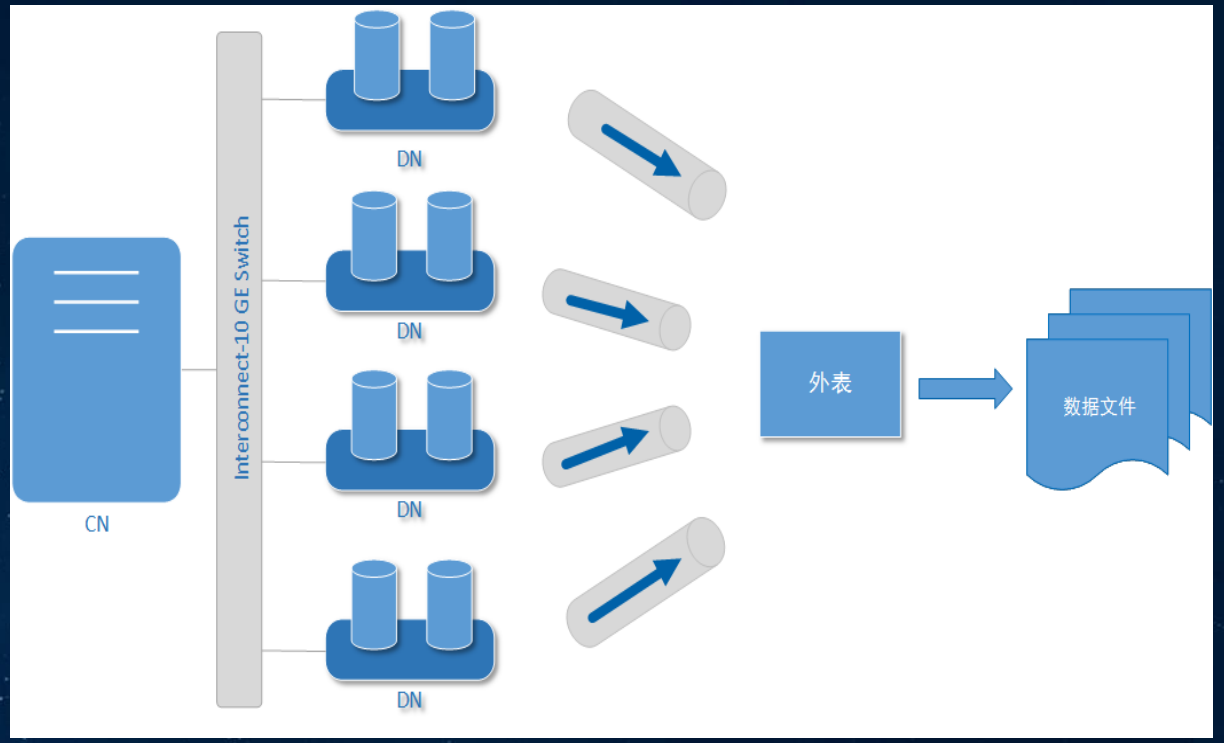
在数据导入完成后，执行ANALYZE语句生成表统计信息。执行计划生成器会使用这些统计数据，以生成最有效的查询执行计划。

如果导入过程中，进行了大量的更新或删除行时，应运行VACUUM FULL命令，然后运行 ANALYZE 命令。大量的更新和删除操作，会产生大量的磁盘页面碎片，从而逐渐降低查询的效率。VACUUM FULL可以将磁盘页面碎片恢复并交还操作系统。

外表并行导出数据

通过外表导出数据：通过外表设置的导出模式、导出数据格式等信息来指定待导出的数据文件，利用多DN并行的方式，将数据从数据库导出到数据文件中，从而提高整体导出性能。不支持直接导出文件到HDFS文件系统。

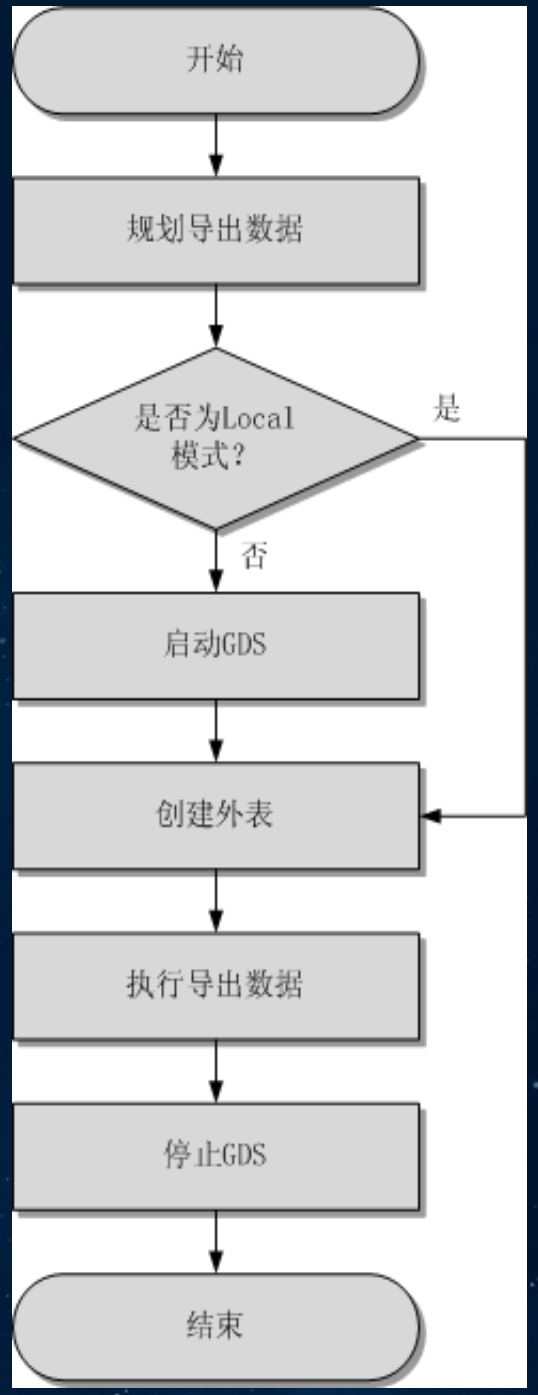
- CN只负责任务的规划及下发，把数据导出的工作交给了DN，释放了CN的资源，使其有能力处理外部请求。
- 通过让各个DN都参与数据导出，充分利用各个设备的计算能力及网络带宽。



外表并行导出数据

数据文件：存储有数据的TEXT、CSV或FIXED文件。文件中保存的是从DWS数据库导出的数据。

- 外表：用于规划导出数据文件的数据文件格式、存放位置、编码格式等信息。
- **GDS**：数据服务工具。在导出数据时，需要将此工具部署到数据文件所在的服务器上，使DN可以通过该工具导出数据。
- 表：数据库中的表，包括行存表和列存表。数据文件中的数据从这些表中导出。
- **Local**导出模式：将集群中的业务数据导出到集群节点所在主机上。
- **Remote**导出模式：将集群中的业务数据导出到集群之外的主机上。



JOIN US IN

BUILDING A BETTER CONNECTED WORLD

THANK YOU

HUAWEI