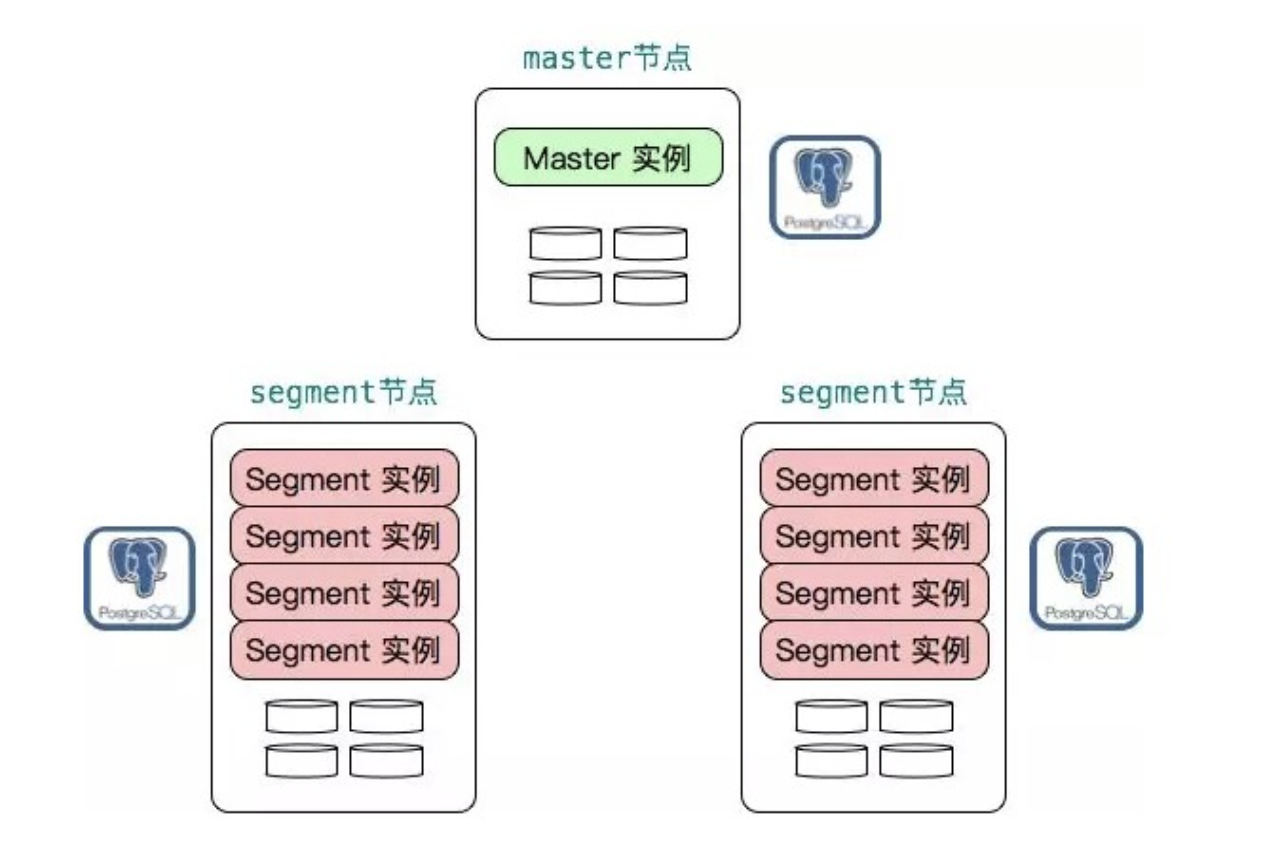
# Greenplum

## Greenplum架构

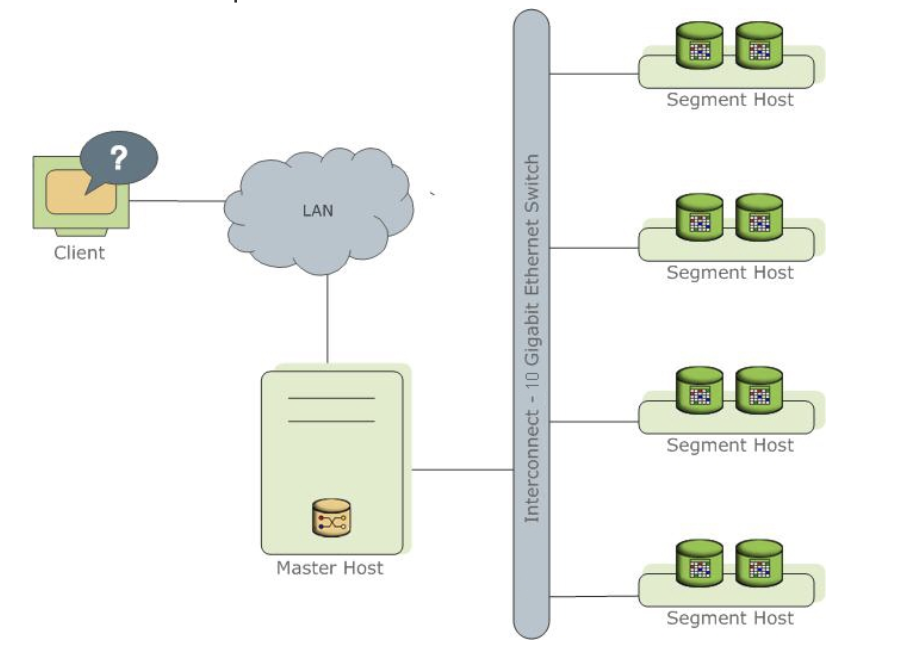
Greenplum数据库是一种大规模并行处理（MPP）数据库服务器，其体系结构专门设计用于管理大规模分析数据仓库和商业智能工作负载。

MPP（也称为无共享体系结构）是指具有两个或两个以上处理器的系统，这些处理器合作执行一项操作，每个处理器具有自己的内存，操作系统和磁盘。Greenplum使用这种高性能的系统体系结构来分配多TB数据仓库的负载，并且可以并行使用系统的所有资源来处理查询。

Greenplum是典型的Master/Slave架构，在Greenplum集群中，存在一个Master节点和多个Segment节点，其中每个节点上可以运行多个数据库。Greenplum采用shared nothing架构（MPP）。典型的Shared Nothing系统会集数据库、内存Cache等存储状态的信息；而不在节点上保存状态的信息。节点之间的信息交互都是通过节点互联网络实现。通过将数据分布到多个节点上来实现规模数据的存储，通过并行查询处理来提高查询性能。每个节点仅查询自己的数据。所得到的结果再经过主节点处理得到最终结果。通过增加节点数目达到系统线性扩展。



如下图Greenplum主要由Master节点、Segment节点、interconnect三大部分组成。



Greenplum master是Greenplum数据库系统的入口，接受客户端连接及提交的SQL语句，将工作负载分发给其它数据库实例（segment实例），由它们存储和处理数据。

Greenplum interconnect负责不同PostgreSQL实例之间的通信。

Greenplum segment是独立的PostgreSQL数据库，每个segment存储一部分数据。大部分查询处理都由segment完成。

Master节点不存放任何用户数据，只是对客户端进行访问控制和存储表分布逻辑的元数据。

Segment节点负责数据的存储，可以对分布键进行优化以充分利用Segment节点的io性能来扩展整集群的io性能。

存储方式可以根据数据热度或者访问模式的不同而使用不同的存储方式。一张表的不同数据可以使用不同的物理存储方式：行存储、列存储、外部表。

## 分布式数据存储

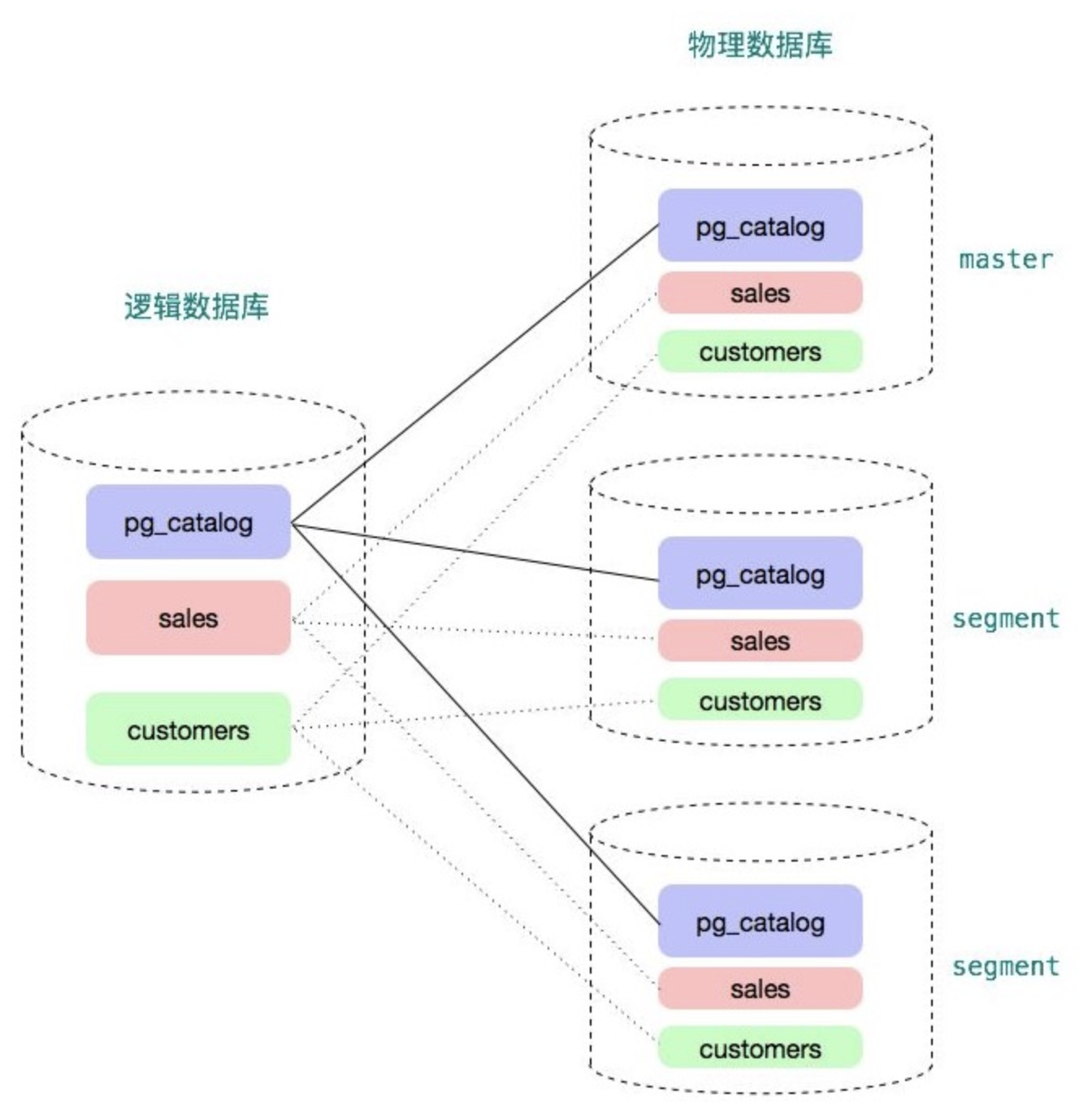
Greenplum 不单单做到了基本的分布式数据存储，还提供了很多更高级灵活的特性，譬如多级分区、多态存储。Greenplum 6 进一步增强了这一领域，实现了一致性哈希和复制表，并允许用户根据应用干预数据分布方法。

（1）Greenplum数据库通过将数据分布到多个节点上来实现规模数据的存储。数据库的瓶颈经常发生在I/O方面，数据库的诸多性能问题最终总能归罪到I/O身上，久而久之，IO瓶颈成为了数据库性能的永恒的话题。

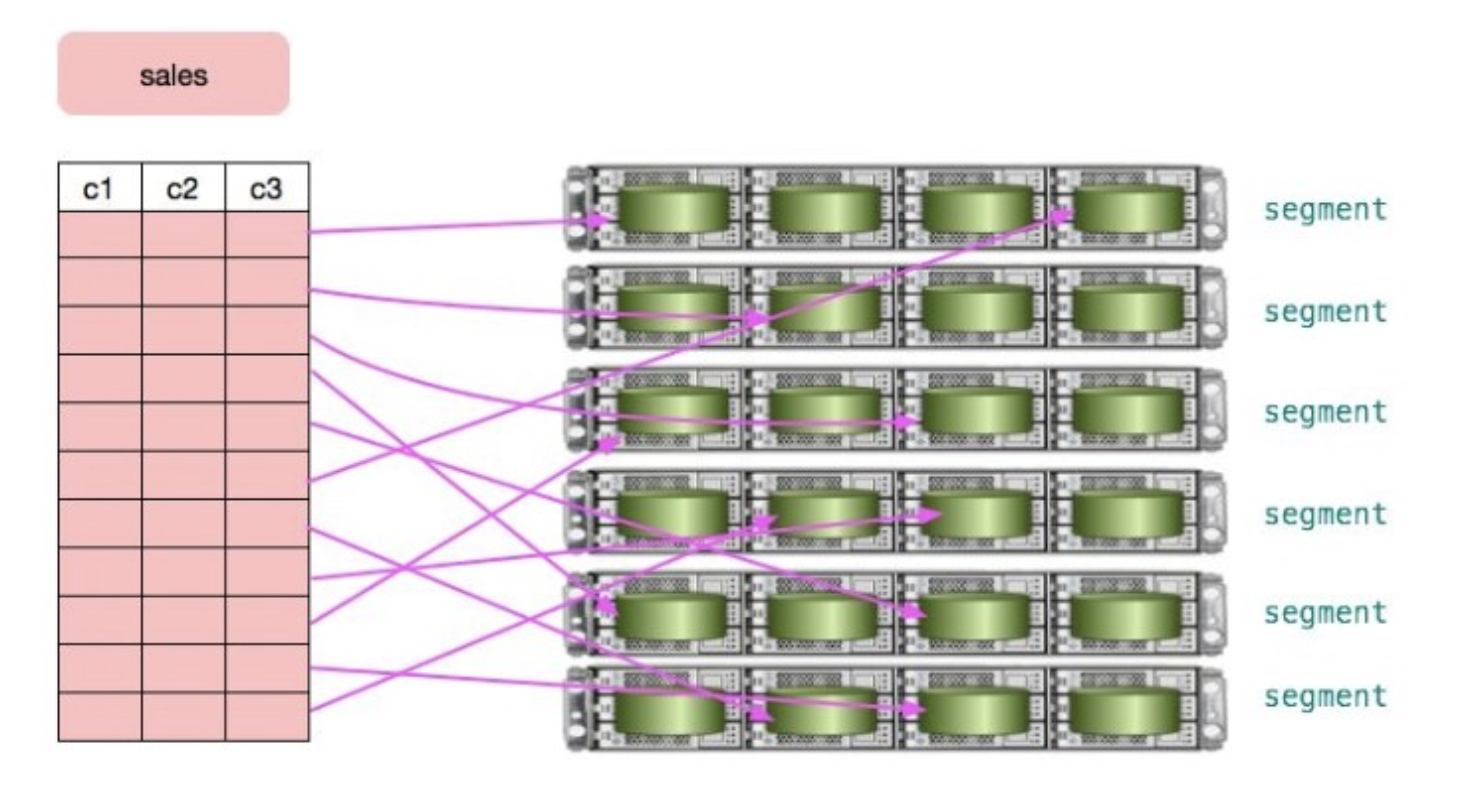
（2）Greenplum采用分而治之的办法，将数据规律的分布到节点上，充分利用Segment主机的IO能力，以此让系统达到最大的IO能力（主要是带宽）。

（3）在Greenplum中每个表都是分布在所有节点上的。Master节点首先通过对表的某个或多个列进行hash运算，然后根据hash结果将表的数据分布到Segment节点中。整个过程中Master节点不存放任何用户数据，只是对客户端进行访问控制和存储表分布逻辑的元数据。

如下图所示，用户看到的是一个逻辑数据库，每个数据库有系统表（例如 pg\_catalog 下面的 pg\_class, pg\_proc 等）和用户表（下例中为 sales 表和 customers 表）。在物理层面，它有很多个独立的数据库组成。每个数据库都有它自己的一份系统表和用户表。master 数据库仅仅包含元数据而不保存用户数据。master 上仍然有用户数据表，这些用户数据表都是空表，没有数据。优化器需要使用这些空表进行查询优化和计划生成。segment 数据库上绝大多数系统表（除了少数表，例如统计信息相关表）和 master 上的系统表内容一样，每个 segment 都保存用户数据表的一部分。



在 Greenplum 中，用户数据按照某种策略分散到不同节点的不同 segment 实例中。每个实例都有自己独立的数据目录，以磁盘文件的方式保存用户数据。使用标准的 INSERT SQL 语句可以将数据自动按照用户定义的策略分布到合适的节点，然而 INSERT 性能较低，仅适合插入少量数据。Greenplum 提供了专门的并行化数据加载工具以实现高效数据导入，详情可以参考 gpfdist 和 gpload 的官方文档。此外 Greenplum 还支持并行 COPY，如果数据已经保存在每个 segment 上，这是最快的数据加载方法。下图形象的展示了用户的 sales 表数据被分布到不同的 segment 实例上。



除了支持数据在不同的节点间水平分布，在单个节点上 Greenplum 还支持按照不同的标准分区，且支持多级分区。

Greenplum提供称为“多态存储”的灵活存储方式。多态存储可以根据数据热度或者访问模式的不同而使用不同的存储方式。一张表的不同数据可以使用不同的物理存储方式。支持的存储方式包含：

行存储：行存储是传统数据库常用的存储方式，特点是访问比较快，多列更新比较容易。

列存储：列存储按列保存，不同列的数据存储在不同的地方（通常是不同文件中）。适合一次只访问宽表中某几个字段的情况。列存储的另外一个优势是压缩比高。

外部表：数据保存在其他系统中例如HDFS，数据库只保留元数据信息。

## Greenplum优势

Greenplum之所以能成为处理海量大数据的有效工具，与其所具备的几大优势密不可分。

### 优势一：计算效率提升

Greenplum的数据管道可以高效地将数据从磁盘传输到CPU，而目前市面上常用的计算引擎SPARK在传输数据时，则需要为每个并发查询分配一个内存，这对大型数据集的查询十分不利，而Greenplum所具备的实时查询功能，能够有效对大数据集进行计算。

### 优势二：扩展性能增强

Greenplum基于的MPP架构，节点之间完全不共享，同时又可以达到并行查询，因此在进行线性扩展时，数据规模可以达到PB级别。目前，Greenplum已经实现了开源，并且社区生态活跃，对于使用者而言，也会觉得更为可靠。

### 优势三：功能性优化

Greenplum可以支持复杂的SQL查询，大幅简化了数据的操作和交互过程。而目前流行的HAWQ、Spark SQL、Impala等技术基本都基于MapReduce进行的优化，虽然部分也使用了SQL查询，但是对SQL的支持十分有限。

## Greenplum的容错机制

Greenplum数据库简称GPDB，它拥有丰富的特性，支持多级容错机制和高可用。

### 主节点高可用

为了避免主节点单点故障，特别设置一个主节点的副本（称为 Standby Master），通过流复制技术实现两者同步复制，当主节点发生故障时，从节点可以成为主节点，从而完成用户请求并协调查询执行。

### 数据节点高可用

每个数据节点都可以配备一个镜像，它们之间通过文件操作级别的同步来实现数据的同步复制（称为filerep技术）。故障检测进程（ftsprobe）会定期发送心跳给各个数据节点，当某个节点发生故障时，GPDB会自动进行故障切换。

### 网络高可用

为了避免网络的单点故障，每个主机会配置多个网口，并使用多个交换机，避免网络故障时造成整个服务器不可用。同时，GPDB具有图形化的性能监控功能，基于此功能，用户可以确定数据库当前的运行情况和历史查询信息，同时跟踪系统使用情况和资源信息。

## Greenplum在业务场景中的应用

总得来说，Greenplum帮助开发者有效解决了处理数据库时遇到的一些难点，比如跨天去重、用户自定义维度、复杂的SQL查询等问题，同时，也方便开发者直接在原始数据上进行实时查询，减少了数据聚合过程中的遗失，Greenplum仍存在着一些问题需要去完善，例如在节点扩展的过程中元数据的管理问题，分布式数据库在扩展节点时会带来数据一致性，扩展的过程中有时会出现元数据混乱的情况等等。在Greenplum有很多优秀的运维工具，能够帮我们在发生问题及时进行排查，更好的保障业务的稳定性。