**通过HAWQ与MADlib，深入学习大数据分析技术**

- HAWQ安装、连接、对象与资源管理、查询优化、备份恢复、高可用性、运维监控
- ETL处理、自动调度系统、维度表与事实表技术、OLAP与数据的图形化表示
- 降维、协同过滤、关联规则、回归、聚类、分类等常见数据挖掘与机器学习方法


Apache Hadoop Native SQL
Advanced Analytics MPP Database for Enterprises

# HAWQ

# 数据仓库与数据挖掘实战

王雪迎 著

清華大學出版社

# HAWQ
## 数据仓库与数据挖掘实战

王雪迎 著

清华大学出版社
北京

## 内 容 简 介

Apache HAWQ 是一个 SQL-on-Hadoop 产品，它非常适合用于 Hadoop 平台上快速构建数据仓库系统。HAWQ 具有大规模并行处理、完善的 SQL 兼容性、支持存储过程和事务、出色的性能表现等特性，还可与开源数据挖掘库 MADlib 轻松整合，从而使用 SQL 就能进行数据挖掘与机器学习。

本书内容分技术解析、实战演练与数据挖掘三个部分共 27 章。技术解析部分说明 HAWQ 的基础架构与功能特性，包括安装、连接、对象与资源管理、查询优化、备份恢复、高可用性等。实战演练部分用一个完整的示例，说明如何使用 HAWQ 取代传统数据仓库，包括 ETL 处理、自动调度系统、维度表与事实表技术、OLAP 与数据的图形化表示等。数据挖掘部分用实例说明 HAWQ 与 MADlib 整合，实现降维、协同过滤、关联规则、回归、聚类、分类等常见数据挖掘与机器学习方法。

本书适合数据库管理员、大数据技术人员、Hadoop 技术人员、数据仓库技术人员，也适合高等院校和培训机构相关专业的师生教学参考。



**图书在版编目（CIP）数据**

HAWQ 数据仓库与数据挖掘实战 / 王雪迎著. — 北京：清华大学出版社，2018
ISBN 978-7-302-49802-5

I. ①H… II. ①王… III. ①数据库系统②数据采集 IV. ①TP311.13②TP274

中国版本图书馆 CIP 数据核字（2018）第 037177 号

**责任编辑：** 夏毓彦
**封面设计：** 王　翔
**责任校对：** 闫秀华
**责任印制：** 沈　露

**出版发行：** 清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

**印 装 者：** 三河市铭诚印务有限公司
**经　　销：** 全国新华书店
**开　　本：** 190mm×260mm　　**印　　张：** 37　　**字　　数：** 947 千字
**版　　次：** 2018 年 4 月第 1 版　　**印　　次：** 2018 年 4 月第 1 次印刷
**印　　数：** 1~3000
**定　　价：** 98.00 元

产品编号：077883-01

# 推荐序

回想过去几年，从我在 EMC（Greenplum）启动 HAWQ 项目开始，到全球多个世界 500 强公司使用 HAWQ，后来把 HAWQ 开源到 Apache 社区，现在又基于 HAWQ 创立“偶数”，时光荏苒。今天非常高兴能够看到雪迎这本关于 HAWQ 的书出现。

数据仓库的架构发展经历了几个阶段，第一代数据仓库是基于传统交易型数据库的共享存储（Share Storage）架构，比如 Oracle，这种架构的缺点是基于专有高端存储，价格昂贵，可扩展性差，扩展到十几个节点往往就会撞到存储的瓶颈。

第二代数据仓库称为 MPP（Massively Parallel Processing），采用无共享架构（Share Nothing），最早商业化的 MPP 产品为 20 世纪 80 年代出现的 Teradata。Teradata 当时基于大型机和专有硬件。在 2000 年左右又出现了几个基于普通 x86 服务器的 MPP 数据仓库创业公司，比如 Greenplum、Vertica 和 Netezza，这几个创业公司后来分别被巨头 EMC、HP 和 IBM 收购。MPP 架构解决了专有硬件的问题，可扩展性也得到了一定的提高，一般可以扩展到 100 节点左右。这种架构的缺点是在执行查询时，无论查询多大，所有节点都同样执行查询中均匀划分的一小部分，在节点数特别多的时候，很难协调保证所有节点的状态和工作都是均匀一致的。就像几个人一起干活，大家分工协调起来容易，如果几千人一起干活，人与人之间的不同以及协调问题就会突显起来。这也是 MPP 架构很难扩展到大规模的一个重要原因。

MPP 之后的新一代数据仓库（New Data Warehouse）都采取了存储与计算分离架构。正是因为存储与计算分离，计算可以访问存储在任何节点的数据，并在任意节点进行调度，从而可以实现高可扩展性。存储与计算分离的另外一个好处是管理的简单性，比如扩容不再需要像 MPP 一样重新分布一遍数据。

新一代数据仓库根据存储实现方式的不同也可以分为三大类：SQL on Hadoop、SQL on Object Store 以及 SQL on Global Store。Hive、SparkSQL 和 HAWQ 2.x 版本属于典型的 SQL on Hadoop，存储为 HDFS；像 Amazon 的 Athena 和 Snowflake 则属于 SQL on Object Store，数据存储在 S3 对象存储中。一般 SQL on Hadoop 和 SQL on Object Store 都有着兼容性不好、性能一般或者对 Update/Delete 以及混合工作负载支持不好的缺点，但 HAWQ 因为从开始就定位为下一代的 Greenplum Database 和语法解析器等源于 Greenplum Database，所以在兼容性和性能等方面表现得

很优秀。HAWQ 社区现在正在开发的 HAWQ 新版本将会创新性地提出 SQL on Global Store 架构，HAWQ 将会具有一个可以全球规模部署、多数据中心、多活的存储。这样 HAWQ 就可以更加高效地支持各种传统数据仓库可以实现的功能，比如 Update/Delete 等，还可以更好地支持传统数据仓库做不到的功能，比如多数据中心、多活等，从而彻底取代传统数据仓库。

雪迎的这本书很好地介绍了 HAWQ 的基本技术，并从用户角度详细给出了如何使用 HAWQ 来构建数据仓库、进行机器学习和数据挖掘的方法，非常全面，是一本很好的 HAWQ 入门书籍。人工智能的流行以及数据驱动的方法是企业能够在新的数据和 AI 时代取得成功的关键，相信这本书的读者一定会从中受益，掌握最新的技术发展趋势与潮流。

Apache HAWQ 创始人
常雷
2018 年 1 月于北京

# 前 言

从 Bill Inmon 在 1991 年提出数据仓库的概念，至今已有 27 的时间。在这期间人们所面对的数据，以及处理数据的方法都发生了翻天覆地的变化。随着互联网和移动终端等应用的普及，运行在单机或小型集群上的传统数据仓库不再能满足数据处理要求，以 Hadoop 及其生态圈组件为代表的新一代分布式大数据处理平台逐渐流行。

尽管大多数人都在讨论某种技术或者架构可能会胜过另一种，而我更倾向于从“Hadoop 与数据仓库密切结合”这个角度来探讨问题。一方面企业级数据仓库中已经积累了大量的数据和应用程序，它们仍然在决策支持领域发挥着至关重要的作用；另一方面，传统数据仓库从业人员的技术水平和经验也在逐步提升。如何才能使积累的大量历史数据平滑过渡到 Hadoop 上，并让熟悉传统数据仓库的技术人员能够有效地利用已有的知识，可以在大数据处理平台上一展身手，才是一个亟待解决的问题。

虽然伴随着大数据的概念也出现了以 MongoDB、Cassandra 为代表的 NoSQL 产品，但不可否认，SQL 仍然是数据库、数据仓库中常使用的开发语言，也是传统数据库工程师或 DBA 的必会语言，从它出现至今一直被广泛使用。首先，SQL 有坚实的关系代数作为理论基础，经过几十年的积累，查询优化器也已经相当成熟。再者，对于开发者，SQL 作为典型的非过程语言，其语法相对简单，但语义却相当丰富。据统计 95%的数据分析问题都能用 SQL 解决，这是一个相当惊人的结论。那么 SQL 怎样才能与 Hadoop 等大数据技术结合起来，既能复用已有的技能，又能有效处理大规模数据呢？在这样的需求背景下，近年来涌现出越来越多的 SQL-on-Hadoop 软件，比如从早期的 Hive 到 Spark SQL、Impala、Kylin 等，本书所论述的就是众多 SQL-on-Hadoop 产品中的一员——HAWQ。

我最初了解到 HAWQ 是在 BDTC 2016 大会上，Apache HAWQ 的创始人常雷博士介绍了该项目。他的演讲题目是“以 HAWQ 轻松取代传统数据仓库”，这正是我的兴趣所在。HAWQ 支持事务、性能表现优良，关键是与 SQL 的兼容性非常好，甚至支持存储过程。对于传统数据仓库的开发人员，使用 HAWQ 转向大数据平台，学习成本应该是比较低的。我个人认为 HAWQ 更适合完成 Hadoop 上的数据仓库及其数据分析与挖掘工作。

## 本书内容

一年来，我一直在撰写 HAWQ 相关的文章和博客，并在利用 HAWQ 开发 Hadoop 数据仓库方面做了一些基础的技术实践，本书就是对这些工作的系统归纳与总结。全书分为技术解析、实战演练、数据挖掘三个部分，共 27 章。

技术解析部分说明 HAWQ 的基础架构与功能特性，包括安装部署、客户端与服务器连接、数据库对象与资源管理、查询优化、备份恢复、高可用性等。

实战演练部分通过一个简单而完整的示例，说明使用 HAWQ 设计和实现数据仓库的方法，包括初始和定期 ETL 处理、自动调度系统、维度表与事实表技术、联机分析处理与数据的图形化表示等。这部分旨在将传统数据仓库建模、SQL 开发的简单性与大数据技术相结合，快速、高效地建立可扩展的数据仓库及其应用系统。

数据挖掘部分结合应用实例，讨论将 HAWQ 与 MADlib 整合，MADlib 是一个开源机器学习库，提供了精确的数据并行实现、统计和机器学习方法，可以对结构化和非结构化数据进行分析。它的主要目的是可以非常方便地加载到数据库中，扩展数据库的分析功能。MADlib 仅用 SQL 查询就能做简单的数据挖掘与机器学习，实现矩阵分解、降维、关联规则、回归、聚类、分类、图算法等常见数据挖掘方法。这也是 HAWQ 的一大亮点。

## 本书读者

本书适合数据库管理员、数据仓库技术人员、Hadoop 或其他大数据技术人员，也适合高等院校和培训学校相关专业的师生教学参考。

## 代码、彩图下载

本书代码与彩图文件下载地址如下（注意数字与字母大小写）：

https://pan.baidu.com/s/1bpppAj1（密码：r7er）

如果下载有问题，请联系电子邮箱 booksaga@163.com，邮件主题为本书书名。

## 致谢

在本书编写过程中，得到了很多人的帮助与支持。感谢清华大学出版社图格事业部的老师和编辑们，他们的辛勤工作使得本书得以尽早与读者见面。感谢 CSDN 提供的技术分享平台，给我有一个将博客文章整理成书的机会。感谢我在优贝在线的所有同事，特别是技术部的同事们，他们在工作中的鼎力相助，使我有更多的时间投入到本书的写作中。感谢 Apache HAWQ 的创始人常雷先生在百忙之中为本书写推荐序。最后，感谢家人对我一如既往地支持。

因为水平有限，错漏之处在所难免，希望读者批评指正。

著　者

2018 年 1 月

# 目 录

## 第一部分 HAWQ 技术解析

# 第一部分

# HAWQ 技术解析