

# 《Greenplum企业应用实战》

## 图书基本信息

书名：《Greenplum企业应用实战》

13位ISBN编号：9787111481003

出版时间：2014-10-1

作者：何勇,陈晓峰

页数：348

版权说明：本站所提供下载的PDF图书仅提供预览和简介以及在线试读，请支持正版图书。

更多资源请访问：[www.tushu000.com](http://www.tushu000.com)

# 《Greenplum企业应用实战》

## 内容概要

这是国内首本Greenplum著作，国内最早开始使用Greenplum的企业是阿里巴巴，本书的两位作者是阿里巴巴最早负责使用和维护Greenplum的技术工程师，权威性毋庸置疑。本书完全立足于阿里巴巴的企业应用实践，不仅系统介绍Greenplum的功能特性、使用方法、高级应用，而且还详细讲解Greenplum的系统架构、运维管理、性能优化和各种技巧。最重要的是，包含大量企业级应用案例，每个案例都进行了详尽的讲解和实操指导。

全书一共15章，分为三个部分：基础篇（第1~3章）首先介绍了Greenplum的应用场景、功能特性以及与PostgreSQL的关系，然后讲解了Greenplum的安装配置、语法以及相关操作，最后通过两个具体的数据仓库ETL案例加强读者对Greenplum的功能特性的了解和操作能力；进阶篇（第4~7章）围绕数据字典、执行计划、系统架构、高级特性等主题对Greenplum进行了更深入地讲解，不仅能让读者更深入理解Greenplum的工作原理，也能让读者游刃有余地应对各种日常操作；管理篇（8~15章）从运维和管理的角度讲解了Greenplum的线上部署、数据库管理、脚本维护、监控、权限控制、容灾/扩容、备份恢复、性能调优、常用技巧和常见问题等。

## 作者简介

### 陈晓峰

陈晓峰 资深数据库专家和高级开发工程师，对Greenplum和PostgreSQL等数据库以及Hadoop和Storm等大数据技术有非常深入的研究和丰富的实践经验。曾就职于阿里巴巴数据平台部和数据平台事业部，负责数据仓库Greenplum计算集群、报表集群的维护及调优，担任RTDC项目和天罡项目的技术负责人，以及负责双十一的交易直播间项目；现就职于阿里巴巴小微金服集团保险事业部，负责保险事业部所有险种的核保核赔。熟悉Java、C、C++、Python，以及数据挖掘和数据分析相关技术。

### 何勇

何勇 资深数据库专家和软件架构师，对Greenplum、Oracle、Teradata、MySQL以及各种NoSQL都有非常深入的研究，实战经验丰富。曾就职于阿里巴巴和盛大，从事数据库系统架构、软件架构和数据中心相关的工作。熟悉Perl、Python、Java、C，以及移动开发。杭州遥指科技有限公司联合创始人兼CTO。

## 书籍目录

### 目 录

#### 前言

#### 上篇 基础篇

#### 第1章 Greenplum简介

2

##### 1.1 Greenplum的起源和发展历程

2

##### 1.2 OLTP与OLAP

3

##### 1.3 PostgreSQL与Greenplum的关系

3

###### 1.3.1 PostgreSQL

3

###### 1.3.2 Greenplum

5

##### 1.4 Greenplum特性及应用场景

6

###### 1.4.1 Greenplum特性

6

###### 1.4.2 Greenplum应用场景

7

##### 1.5 小结

8

#### 第2章 Greenplum快速入门

9

##### 2.1 软件安装及数据库初始化

9

###### 2.1.1 Greenplum架构

9

###### 2.1.2 环境搭建

11

###### 2.1.3 Greenplum安装

13

###### 2.1.4 创建数据库

20

###### 2.1.5 数据库启动与关闭

20

##### 2.2 安装Greenplum的常见问题

22

###### 2.2.1 /etc/hosts配置错误

22

###### 2.2.2 MASTER\_DATA\_DIRECTORY设置错误

24

##### 2.3 畅游Greenplum

25

###### 2.3.1 如何访问Greenplum

25

2.3.2	数据库整体概况	27
2.3.3	基本语法介绍	28
2.3.4	常用数据类型	35
2.3.5	常用函数	37
2.3.6	分析函数	43
2.3.7	分区表	46
2.3.8	外部表	49
2.3.9	COPY命令	51
2.4	小结	52
第3章 Greenplum实战		53
3.1	历史拉链表	53
3.1.1	应用场景描述	53
3.1.2	原理及步骤	54
3.1.3	表结构	55
3.1.4	Demo数据准备	57
3.1.5	数据加载	58
3.1.6	数据刷新	61
3.1.7	分区裁剪	64
3.1.8	数据导出	64
3.2	日志分析	65
3.2.1	应用场景描述	65
3.2.2	数据Demo	65
3.2.3	日志分析实战	66
3.3	数据分布	68
3.3.1	数据分散情况查看	

69	
3.3.2	数据加载速度影响
69	
3.3.3	数据查询速度影响
72	
3.4	数据压缩
73	
3.4.1	数据加载速度影响
73	
3.4.2	数据查询速度影响
74	
3.5	索引
75	
3.6	小结
75	
中篇	进阶篇
第4章	数据字典详解
78	
4.1	oid无处不在
78	
4.2	数据库集群信息
80	
4.2.1	Gp_configuration和gp_segment_configuration
80	
4.2.2	Gp_id
82	
4.2.3	Gp_configuration_history
84	
4.2.4	pg_filespace_entry
84	
4.2.5	集群配置信息表转化
84	
4.3	常用数据字典
85	
4.3.1	pg_class
85	
4.3.2	pg_attribute
88	
4.3.3	gp_distribution_policy
89	
4.3.4	pg_statistic和pg_stats
90	
4.4	分区表信息
90	
4.4.1	如何实现分区表
91	
4.4.2	pg_partition
91	
4.4.3	pg_partition_rule

92	
4.4.4	pg_partitions视图及其优化
93	
4.5	自定义类型以及类型转换
94	
4.6	主、备节点同步的相关数据字典
95	
4.7	数据字典应用示例
96	
4.7.1	获取表的字段信息
96	
4.7.2	获取表的分布键
96	
4.7.3	获取一个视图的定义
97	
4.7.4	查询comment (备注信息)
98	
4.7.5	获取数据库建表语句
99	
4.7.6	查询表上的视图
103	
4.7.7	查询表的数据文件创建时间
104	
4.7.8	分区表总大小
106	
4.7.9	如何分析数据字典变化
108	
4.7.10	获取数据库锁信息
111	
4.8	Gp_toolkit介绍
112	
4.9	小结
114	
第5章	执行计划详解
115	
5.1	执行计划入门
115	
5.1.1	什么是执行计划
115	
5.1.2	查看执行计划
116	
5.2	分布式执行计划概述
116	
5.2.1	架构
116	
5.2.2	重分布与广播
117	
5.2.3	Greenplum Master的工作
119	

5.3	Greenplum执行计划中的术语	120
5.3.1	数据扫描方式	120
5.3.2	分布式执行	121
5.3.3	两种聚合方式	122
5.3.4	关联	123
5.3.5	SQL消耗	126
5.3.6	其他术语	126
5.4	数据库统计信息收集	128
5.4.1	Analyze分析	128
5.4.2	固定执行计划	129
5.5	控制执行计划的参数介绍	130
5.6	规划器开销的计算方法	131
5.7	各种执行计划原理分析	133
5.7.1	详解关联的广播与重分布	133
5.7.2	HashAggregate与GroupAggregate	137
5.7.3	Nestloop Join、Hash Join 与Merge Join	141
5.7.4	分析函数：开窗函数和grouping sets	142
5.8	案例	144
5.8.1	关联键强制类型转换，导致重分布	144
5.8.2	统计信息过期	145
5.8.3	执行计划出错	145
5.8.4	分布键选择不恰当	147
5.8.5	计算distinct	148
5.8.6	union与union all	150
5.8.7	子查询not in	

152	
5.8.8	聚合函数太多导致内存不足
154	
5.9	小结
155	
第6章	Greenplum高级应用
156	
6.1	Appendonly表与压缩表
157	
6.1.1	应用场景及语法介绍
157	
6.1.2	压缩表的性能差异
157	
6.1.3	Appendonly表特性
158	
6.1.4	相关数据字典
164	
6.2	列存储
165	
6.2.1	应用场景
165	
6.2.2	数据文件存储特性
166	
6.2.3	如何使用列存储
166	
6.2.4	性能比较
166	
6.3	外部表高级应用
168	
6.3.1	外部表实现原理
168	
6.3.2	可写外部表
171	
6.3.3	HDFS外部表
173	
6.3.4	可执行外部表
177	
6.4	自定义函数—各个编程接口
179	
6.4.1	pl/pgsql
180	
6.4.2	C语言接口
182	
6.4.3	plpython
185	
6.5	Greenplum MapReduce
187	
6.6	小结
193	

## 第7章 Greenplum架构介绍

195

### 7.1 并行和分布式计算

195

### 7.2 并行数据库

197

### 7.3 Greenplum架构分析

198

### 7.4 冗余与故障切换

199

### 7.5 数据分布及负载均衡

200

### 7.6 跨库关联

202

### 7.7 分布式事务

203

### 7.8 其他大数据分析方案

205

### 7.9 小结

208

## 下篇 管理篇

## 第8章 Greenplum线上环境部署

210

### 8.1 服务器硬件选型

210

#### 8.1.1 CPU

211

#### 8.1.2 内存

211

#### 8.1.3 磁盘及硬盘接口

211

#### 8.1.4 网络

213

### 8.2 服务器系统参数调整

213

#### 8.2.1 Solaris参数修改

214

#### 8.2.2 Linux参数修改

216

#### 8.2.3 系统参数及性能验证

217

### 8.3 计算节点分配技巧

221

### 8.4 数据库参数介绍

221

### 8.5 数据库集群基准测试

225

### 8.6 小结

227

第9章 数据库管理	228
9.1 用户及权限管理	228
9.1.1 Greenplum数据库逻辑结构	228
9.1.2 Grant语法	229
9.2 登录权限控制	231
9.3 资源队列及并发控制	232
9.4 Greenplum锁机制	236
9.5 数据目录结构	238
9.6 数据文件存储分布	240
9.7 表空间管理	241
9.8 小结	244
第10章 数据库监控及调优	245
10.1 Linux监控工具介绍	245
10.1.1 监控磁盘	245
10.1.2 监控网络	246
10.1.3 监控CPU	247
10.1.4 监控内存	247
10.2 安装Performance Monitor	248
10.3 监控Segment是否正常	252
10.4 VACUUM系统表	253
10.5 数据倾斜排查	255
10.6 查看子节点的SQL运行状态	258
10.7 自动加分区	261
10.8 自动赋权	266
10.9 清理过期数据	

266	
10.10	小结
267	
第11章	解读Greenplum维护脚本
268	
11.1	添加Greenplum Contrib模块
268	
11.2	启动和关闭脚本gpstart和gpstop
270	
11.3	初始化系统脚本gpinitssystem
272	
11.4	集群操作脚本gpssh和gpscp
274	
11.5	数据库状态检查脚本gpstate
275	
11.6	数据库升级脚本gpmigrate
276	
11.7	参数修改脚本gpconfig
281	
11.8	数据库一致性检查脚本gpcheckcat
282	
11.9	小结
284	
第12章	备份及恢复策略
286	
12.1	Greenplum 3.x
286	
12.2	Greenplum 4.x
287	
12.3	gp_dump和pg_dump
290	
12.4	Greenplum Master备份策略
294	
12.4.1	增加Standby Master
295	
12.4.2	重新同步Standby Master
296	
12.4.3	启用Standby Master
296	
12.5	小结
297	
第13章	数据库扩容
299	
13.1	迁移计算节点
299	
13.1.1	两种备份方案
300	
13.1.2	数据迁移实战
301	

13.2 增加计算节点	306
13.3 小结	311
第14章 基于Greenplum的海量数据实时分析服务平台	312
14.1 需求概述	312
14.2 典型方案	313
14.2.1 NoSQL	313
14.2.2 分布式数据库/集群	314
14.2.3 分表分库	315
14.2.4 方案优劣分析	316
14.3 基于Greenplum的混合架构	316
14.3.1 架构分析	317
14.3.2 实施要点	317
14.4 小结	318
第15章 使用Greenplum的常见报错及小技巧	319
15.1 分析常见报错	319
15.1.1 找不到类型705对应的操作符	319
15.1.2 SQL占用的资源超过了资源队列限制	321
15.1.3 自定义函数不能在Segment上执行	321
15.1.4 子查询没有加别名	322
15.1.5 字段名有歧义	322
15.1.6 字段重名	323
15.1.7 gpfdist错误：无法读取文件	323
15.1.8 事务被中止	324
15.1.9 网络异常错误	324
15.1.10 无法删除表	

324	
15.1.11	内存不足
325	
15.1.12	文件名在pg_class中已存在
325	
15.1.13	不能对分布键执行Update
325	
15.1.14	网络错误
326	
15.1.15	无法找到数据文件
326	
15.2	常见问题及解决办法
326	
15.3	常用的一些小技巧
329	
15.3.1	显示SQL执行的时间
330	
15.3.2	获取某个schema下所有的表或视图
330	
15.3.3	查找分区最多的表
330	
15.3.4	连接Segment节点
331	
15.3.5	psql默认密码登录
331	
15.3.6	查看数据库启动时间
331	
15.3.7	查看在psql中\d到底查询了哪些数据字典
331	
15.4	小结
332	

### 精彩短评

- 1、很少买出版时间超过1年之久的书，但这次破例，事实证明是值得的！
- 2、已经挑了几章快速看了一下，对PostgreSQL维护很有帮助。尤其是数据字典那块，蛮不错的。硬件方面和常见报错及小技巧等章节，看了也蛮有收获的。总体上来说，值得入手的。
- 3、赞~~~
- 4、一本很有诚意的书，内容很实在，都是实战干货，缺点就是纸张可以更好点。
- 5、写的比较深入。但是现在大数据方案比较多，gp感觉是逐渐衰落
- 6、greenplum唯一的入门书籍了吧，还可以就是讲的太浅了。
- 7、浙江温州数仓厂，核心开发带着小姨子跑路了，Pivotal没有办法，原本都是几十万的数据仓库，开源甩卖啦。
- 8、工具类书不用切不可乱评价、闲着没事儿翻是很乏味的。这次也是赶鸭子上架，要用gp。去北京的高铁上看了前半本、到了就开始现学现用了，心中窃喜。回来的高铁上，再翻完后半本、真正的实践者组织过的内容是真知灼见，值得尊重。算是顿悟。
- 9、皆为实战，如果未接触过GP，以此作为入门，那价值远远超过书本媒介所能带给你的。
- 10、还是开源的hadoop更深入，商业的东西总是有块扯不掉的遮羞布，不爽。

## 版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:[www.tushu000.com](http://www.tushu000.com)