HZ BOOKS
华章IT

数据库
技术丛书

Oracle Exadata 技术详解

李亚 著

Understanding The Technical Details Of Oracle Exadata

- 国内第一本Oracle数据库一体机运营实践领域的原创著作，Oracle数据库一体机分析专家以真实客户环境为基础撰写。
- 根据大型客户的实践经验及案例详细剖析Oracle Exadata一体机的重要特性，围绕与Exadata相关的数据迁移、并行、安全加固、备份与恢复等展开分析，并针对维护和使用过程中最常见的问题进行了解答。

机械工业出版社
China Machine Press

数据库
技术丛书

Oracle Exadata 技术详解

Understanding The Technical Details Of Oracle Exadata

李亚 著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Oracle Exadata 技术详解 / 李亚著 . —北京：机械工业出版社，2015.9
（数据库技术丛书）

ISBN 978-7-111-51706-1

I. O… II. 李… III. 关系数据库系统 IV. TP311.138

中国版本图书馆 CIP 数据核字（2015）第 237051 号

本书作为国内第一本关于 Oracle Exadata 一体机的中文教程，偏重于实践方面，同时加入了更多 V2 版本以后的新内容。全书可分为三个部分。第一部分为基础篇（1 ~ 5 章），主要介绍了 Oracle Exadata 一体机的配置、架构、安装、升级，帮助读者对 Exadata 有一定程度的认识。第二部分为功能篇（6 ~ 14 章），详细介绍了 Oracle Exadata 一体机的特性以及与 Exadata 相关的数据迁移、并行、安全加固、备份与恢复等课题。第三部分为实战篇（15、16 章），主要针对 Exadata 一体机管理员在维护和使用过程中常见的场景以及问题进行解答。

本书以 Exadata 独有的特性介绍开篇，以最佳实践结尾，内容翔实，覆盖全面，适用的读者范围包括：数据库管理员、应用开发者、数据库开发者、存储管理员、系统架构师，以及广大的数据库兴趣爱好者。

Oracle Exadata 技术详解

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：杨绣国　　责任校对：殷　虹
印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2015 年 11 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：36.75
书　　号：ISBN 978-7-111-51706-1　　定　　价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

Preface 前　言

为什么要写这本书

2010 年我刚接触 Exadata 的时候，当时国内还只有屈指可数的几个客户，Exadata 的版本还是第二版 V2。短短四年多时间过去了，国内 Exadata 一体机的客户已经可以使用千位来计数了，而且每年还在以较大幅度增长。Oracle Exadata 一体机的市场占有率已经远远甩开竞争对手，稳坐国内数据库一体机市场的头把交椅，其产品发布的速度也比较惊人，2015 年初已经发布了第六代数据库一体机产品 X5。

在科技领域，近几年有几个趋势越来越明显。

第一个趋势是集成化，也就是我们所说的软硬件一体化。将软件与硬件结合起来，一并创造最佳的体验。苹果公司引领了智能手机软硬件一体化的趋势，而特斯拉在智能汽车方面创造了另外一个神话，同时其他各行各业的领军人物也正以相同的方式颠覆着传统的孤岛模式。抓住了软硬件一体化，就抓住了下一个商业模式的趋势。Oracle 公司也正是这样做的，除了在传统强项数据库领域的深耕，同时还顺势拓宽了其他领域的集成系统，推出了中间件一体机 Exalogic、数据分析一体机 Exalytics、备份一体机 ZDLRA，以及面向中小型企业的数据库一体机 ODA。

第二个趋势是云化。经过几年的迅猛发展，云计算已经从最初的漂浮不定到现在逐步开始落地。现在业界谁都在抢占云计算的制高点。除了传统的 SaaS、PaaS 和 IaaS，Oracle 又提出了 DBaaS，即数据库即服务，并且将其思想精髓逐步地融入 Oracle 数据库产品与 Exadata 一体机，省略了大量纷繁复杂的部署流程，向用户提供“开箱即用”的云。

第三个趋势是开放化。小型机、中型机在企业级信息系统基础架构中日渐式微已是不争的事实。大量用户已经完成了从小型机运行专有程序到 x86_64 架构运行通用程序的转变，当然还有更多的正在向其靠拢。这主要得益于 x86_64 的开放性，使得运行维护的成

本大大降低，同时 x86_64 平台的稳定性与性能的提升，也大大促进了这一趋势的蔓延。Oracle Exadata 一体机构建于 x86_64 平台，很多用户的应用程序不需要任何修改就可以无缝迁移到 Exadata。

在本书之前，国内市场上已经有译作《深入理解 Oracle Exadata》一书。而本书作为国内第一本关于 Oracle Exadata 一体机的中文教程，更多偏重于实践方面，同时加入了更多 V2 版本以后的新内容。

读者对象

这里根据需求划分出了一些能使用 Exadata 的用户团体：

- 数据库管理员；
- 应用开发者；
- 数据库开发者；
- 存储管理员；
- 系统架构师；
- 数据库兴趣爱好者。

如何阅读本书

本书假定读者对关系型数据库，尤其是 Oracle 数据库有一定程度的了解，否则有可能对书中的某些知识点的介绍感到困惑。如果你是一名初学者，建议先学习 Oracle 数据库的一些基础知识。

本书共包括 16 章，可以将其大致分为三个部分。

第一部分为基础篇，包括第 1 章到第 5 章，这些章节主要介绍了 Oracle Exadata 一体机的配置、架构、安装、升级，帮助读者了解一些基础知识，对 Exadata 有一定程度的认识。

第二部分为功能篇，包括第 6 章到第 14 章，这些章节详细介绍了 Oracle Exadata 一体机的特性以及与 Exadata 相关的数据迁移、并行、安全加固、备份与恢复等课题。在这部分中，每个章节都是独立的，没有严格意义上的依赖关系，所以读者可以任意选其中自己感兴趣的话题进行阅读。

第三部分为实战篇，包括第 15 章与第 16 章，这两个章节主要是针对 Exadata 一体机管理员在维护和使用过程中常见的场景以及问题进行解答，涉及 Exadata 日常运维的方方面面，并且其中的每一节都是互相独立的。

附录 A 为 Exadata 默认密码一览表。

附录 B 为缩略语中英文对照表。

勘误和支持

由于本人水平有限，编写时间也很仓促，所以书中难免会出现错误或者不全面的地方，在此恳请读者批评斧正。你可以将书中的错误发布在 Bug 勘误表页面中，同时，书中的源文件也将发布在华章公司的网站上，并及时更新相应的功能。如果你有任何意见或问题，也欢迎发送邮件至我的邮箱 steven.ya.li@gmail.com，我很期待听到你们的真挚反馈。

致谢

感谢 Oracle 公司内部 Exadata 邮件列表的许多专家对本人提出问题的耐心解答。感谢我在 Oracle 中国公司同事的无私帮助，尤其是来自高级服务团队同事的帮助。他们包括胡奇虎、陈伟、王劲松、顾水林、罗敏、孙建光、蒋健、祁琪、张毅宁、彭玉周、吕春雷、王辉、郭忠伟、王福龙、林宇泽、蔡磊、刘建军、张润平、杜平、刘相兵、金丹、张大鹏、程飞、沈杰、李纯香、郑伯欧等。

同时也感谢 Oracle 社区和 Oracle 上海用户组的大力支持，需要额外感谢的人包括罗炳森、徐浩然、李德鹏、刘斌、赵欣等。

感谢 Oracle 美国总部研发团队的 Michael Chen，谢谢你提供的 Exadata 测试环境，让我得以验证本书中的案例。

感谢机械工业出版社华章公司的编辑杨绣国老师，你的专业与细心深深地感染了我。同时感谢你对我因工作繁忙而将交稿日期一再推迟的理解。

最后要感谢我的父母与家人，为了编写本书，我牺牲了大量本该陪伴你们的时间，正是你们的理解与鼓励使我能够顺利完成此书。

谨以此书献给那些工作多年还依然热爱技术，奋战在技术一线的朋友们。

李亚

2015 年 7 月于上海

目　录 *Contents*

第 1 章 Chapter 1

Exadata 概述

1.1 Exadata 的诞生

Exadata 是什么？几乎每个新接触 Exadata 的人都会问到这个问题。权威百科全书维基百科（wikipedia）对 Exadata 的定义为：Exadata 是 Oracle 公司推出的针对联机事务处理系统（OLTP）和联机分析处理系统（OLAP）的软件和硬件结合的 Oracle 数据库一体机。事实上，我们很难用几句话简单地概括 Exadata 的所有特征。因为 Exadata 涉及的知识面非常广，不仅包括数据库，还包括主机、存储、操作系统、网络等各个方面。而本书的所有内容都将围绕着 Exadata 进行，相信读者在读完本书以后，会对 Exadata 有一个更加全面而深刻的理解。

在此之前，首先还是要介绍 Exadata 的背景知识。Exa 读音为“艾克萨”，表示一个单位，度量是 10 的 18 次方。一般认为，Exadata 这个名字源自于 Oracle 的竞争对手“ Teradata”（中文名为天睿）。Teradata 成立于 1979 年，是美国前十大上市软件公司之一，它于 2007 年从其母公司 NCR 独立出来，是世界上最早提供数据仓库一体机的厂商。事实上，从名字上就可以看出它们之间的一些渊源——tera 表示 10 的 12 次方。Exadata 最初的目标就是要超越 Teradata 在数据仓库一体机方面的垄断地位，在尽管小众但是利润率很高的一体机市场分得一杯羹。当然 Oracle 是一家“野心勃勃”的公司，超越 Teradata 并不是它的最终目标，但是 Exadata 这个名字却一直被沿用下来。

与 Exadata 的诞生密切相关的另外一家巨头公司是 Oracle 公司的启蒙老师 IBM。很早的时候，IBM 就将其数据库与硬件、操作系统及自家的服务作为一个整体打包推销给其客户。例如 DB2 for zOS 以及 DB2 for OS/400。Oracle 公司对这片市场自然是“垂涎已久”，

按照 Oracle 公司向来“咄咄逼人”的架势，免不了会朝这个领域发力，挑战 IBM 公司的霸主地位，抢占高端市场。

另外，不得不提到的一个关键性人物就是 Larry Ellison 的老友、也是他曾经的老邻居——Steve Jobs。iPhone 的成功不仅在电子消费品市场开创了一个崭新的时代，还给这个浑浑噩噩的市场注入了一剂强心剂，也彻底颠覆了行业中盛行的企业级硬件市场为不盈利的鸡肋这一想法。同时还给很多产品经理上了宝贵的一课：单凭卖硬件不足以获得很高的利润，只有把软硬件结合起来，才能获得更好的用户体验，从而获取更高的利润。

1.2 Exadata 设计哲学

Exadata 不是无缘无故产生的，更不是出自“要有光，于是便有了光”的无所不能的上帝之手。Exadata 的产生源于很多方面的专家多年宝贵经验的积累，同时它更是为了解决长期困扰很多 Oracle 用户的特定难题而来。

长期以来，受制于传统的 Oracle 数据库自身架构的局限，Oracle 数据库在处理某些类型的请求时效率并不高。例如在数据仓库架构中有这么一种很典型的场景：当需要从一个很大的结果集中过滤一少部分数据时，首先数据库会发出一个请求，把大量数据从存储端读到数据库服务器端；然后由数据库服务器应用过滤条件，对这大批量的数据库进行条件过滤；最后才能将过滤后的结果返回给最终用户。可以看到：整个数据库的瓶颈在于存储端向数据库服务器端内存传输数据的这一阶段。如果结果集非常大，同时过滤返回的数据量并不多，那么这种方式是非常低效的。

由于数据量的爆炸性增长，单纯依靠增加存储到数据库服务器端的带宽显然已经无法满足这种需求了。要解决这个问题，得从减少从存储到数据库的流量着手。那么怎样才能减少这一段的流量呢？首先想想传统的架构为什么无法做到，原因在于传统的存储是“死”的，不够智能，即存储并不能识别数据库段发送过来的过滤请求，最终的数据过滤操作还得交由数据库引擎来处理。如果有这么一种智能存储，能够识别数据库服务器发送过来的数据过滤请求，那么能大大地降低存储端到数据库端的数据流量，从而大大提高这种场景的效率。Exadata Smart Scan 就是用来解决此类问题的。这种架构可以看作是一种分布式架构，可以认为是 Oracle 公司为了解决 RAC shared disk 架构的局限所做的一种尝试。

还有就是传统的数据仓库应用的数据量可能非常大，并且随着业务的增长，数据量在进一步地膨胀。一方面，会导致用户需要购买更多的存储设备，带来了更高的成本。另一方面，随着数据量的增加，带来的是扫描效率的降低，因为更大的数据量意味着需要扫描更多数据块，所以数据压缩的比例和效率在数据仓库领域显得至关重要。传统的行式数据库例如 Oracle 数据库通常无法提供特别高的压缩比，而列式数据库虽然能提供较高的压缩比，但是对 DML 性能影响较大。如果有这么一种压缩算法既能保证较高的压缩率，同时也能将 DML 操作性能的影响降到最低，那不就两全其美了吗？没错，它就是 Exadata 的

Hybrid Columnar Compression（混合列式压缩）！

在过去的十多年中，机械硬盘的容量在逐渐增大，同时可靠性也变得非常高了。但是令人遗憾的是，硬盘的速率却没有随之增高，硬盘的读 / 写速率相比 CPU 和内存要慢好几个数量级，因此绝大多数性能瓶颈最终可能都出在 I/O 上，而当前硬盘的读 / 写速率几乎已经达到机械设备的物理极限，想要继续增加必然是难上加难，硬盘的读 / 写速率不可避免地成为阻碍性能提升的一块短板。幸好，固态硬盘的出现让人又看到了一丝曙光。相比机械硬盘，固态硬盘最大的优势在于能够提供几十甚至上百倍的读 / 写速率。随着闪存技术的不断进步，固态硬盘必然是未来存储发展的一大趋势。对于 Oracle 数据库而言，并非通过简单地将机械硬盘替换为固态硬盘就能获得非常大的性能收益。必须要从最底层的内核、架构方面针对闪存存储进行全面优化。根据大量真实客户的实践经验，Oracle 开发和性能优化团队把所有固态硬盘的最佳实践都融合到了 Exadata 的闪存技术上，从而使得固态硬盘的威力在 Exadata 上发挥得淋漓尽致。

以上就是 Exadata 的三大核心技术 Smart Scan、Hybrid Columnar Compression 和 Exadata Smart Flash Cache 产生的背景。可以看出，尊重客户需求和敏锐的市场眼光决定了“Exadata 将成为 Oracle 30 年发展史中最成功的新产品”。

1.3　Exadata 的演化与发展

Exadata 这个名词第一次被业界所知晓是在 2008 年的 Openworld 大会（Oracle 的年度技术大会，通常在 Oracle 的 Openworld 大会上，Oracle 公司的高级管理层都会公布一些新的产品以及宣布一些产品发展方向）。Oracle 公司的创始人 Larry Ellison 在该会上宣布与 HP 公司一起发布第一代 Oracle 数据库一体机：HP Oracle Database Machine。但是实际上，Exadata 的历史可以追溯到更早的时候，一个产品从创意到最终成为产品往往需要经历很长的时间。

据 Oracle 系统技术高级副总裁 Juan Loaiza 回忆：2000 年到 2005 年的时候，Oracle 公司就有一个名叫 SAGE 的内部项目已经在秘密研发。（SAGE 是 Storage Applicant Grid Environment（网格环境存储设备）的缩写，这就是 Exadata 的前身。至今还有部分 Oracle 文档中有 SAGE 的影子，例如介绍 diskmon 进程的时候。）当时产品经理内部有两种思路：第一种是提供智能存储设备，直接接入客户已有的 Oracle 数据库系统上；第二种是提供主机、存储、网络、操作系统、数据库等一整套设备的解决方案。刚开始，第一种思路在产品经理内部占据了主流的位置，因为用户无需对已有的数据库系统做任何修改就能将已有的 Oracle 数据库系统迁移到 SAGE 存储上，并且当时 Oracle 的优势也不在做硬件产品上，相对于一体机而言，智能存储的客户接受程度应该会更高。但是随着时间的推移，产品经理发现这种看似“无痛”的方案几乎无人问津，鲜有客户表示有意购买 SAGE。最主要的原因在于整个方案太过复杂，并且为了可以跨平台，还要解决各类令人讨厌的平台兼容性问题。相反，产品经理在和一些客户的接触过程中发现，他们竟然对 Oracle 公司提供一整套

设备的解决方案似乎更感兴趣。

第一种方案很快就被产品经理们否决掉了。于是 B 计划开始顺理成章地实施，即做数据库一体机。但是同时新的问题又来了：Oracle 是一家软件公司，在硬件设计方面毫无经验可言，即使重新招聘一批顶级的硬件工程师，也可能需要相当长的一段磨合期才能出成果，如果掌握的火候不够，甚至有可能导致整个计划流产。所以这时 Larry Ellison 找来了和自己私交不错的时任 HP 公司总裁的 Mark Hurd（现任 Oracle 公司总裁），希望 HP 公司能和 Oracle 公司合作，一道研发数据库一体机。Mark Hurd 爽快地答应了，并成立了一个专门的硬件专家团队与 Oracle 公司的软件团队一起研发数据库一体机 Exadata。经过数年的潜心研发，第一台 Exadata V1 终于问世。

1.3.1 Exadata V1

第一代 Exadata，产品经理们将其取名为 HP Oracle Database Machine，下文简称 Exadata V1。Oracle 公司的高层对该产品十分重视，且寄予了很大的希望。一方面，希望借助 Exadata 巩固自己在数据库行业老大的地位，至少不让核心客户一个个流失。另一方面，也觊觎 Teradata 金融、电信行业的一些优质客户。进可攻，退可守。

Exadata V1 选择的操作系统是 x86_64 的 Oracle Linux 5，数据库版本为 11.1。技术支持服务的接口统一由 Oracle 公司来进行，只有遇到与硬件相关的问题才会以内部协作的方式转给 HP 公司。

从现在的角度来看，与其竞争对手 Teradata、Netessa 相比，Exadata V1 并不算是成功的产品，最多只能算得上是 Oracle 公司的一次试水，当时真正拿得出手的客户成功案例寥寥可数。主要原因一是产品还处于初级研发的阶段，产品太新，很多功能测试不充分并且缺乏稳定性；二是 Exadata V1 针对的客户群体有限，因为产品经理给它的定位是针对数据仓库的客户，因此没有针对其他类型的应用，例如 OLTP，进行过优化，所以效果往往很糟糕；三是设计方面确实存在不成熟、不完善的地方。例如，因为 Exadata V1 的机柜较小，并且服务器的布局很密集，所以其机器的散热功能饱受诟病，甚至有的客户调侃 Exadata 的机柜可以用来煎鸡蛋。另外，还有整个 Exadata V1 的部署十分复杂，软件部署所有的步骤都需要使用一条条命令来完成，除了美国 Exadata 核心研发部门，几乎没有人能够顺利安装好一台 Exadata。即使由 Exadata 研发部门来部署安装一套 Exadata，至少都需要几周时间，并且还无法保证不出现人为方面的失误。

购买 Exadata V1 的客户大多数是 Oracle 的战略性客户，Oracle 没有理由让这些客户蒙受损失。据说所有的 Exadata V1 的客户后来都在 Oracle 的帮助下成功升级到了 Exadata V2，免费或者仅仅支付了很少的一部分费用。

1.3.2 Exadata V2

生意场上没有永远的朋友，也没有永远的敌人，只有永远的利益。对于 Oracle 而言，

2009 年注定是不平凡的一年。2009 年 4 月，Oracle 宣布以总计 74 亿美金收购 Sun，打入企业硬件服务器市场，此时命运无法避免地将 Oracle 与 HP 的关系由昔日亲密的队友变成了直接的竞争对手。这自然也给 Exadata 的发展带来了一些不确定性因素，Oracle 与 Sun 在同年联合推出的 Sun Oracle Database Machine 就是在这样的背景下诞生的。事实上，在收购以前，Exadata V2 的研发就在秘密进行了。可以想象这样的场景，突然某一天，Oracle 的老板告诉 Exadata 研发工程师：从今天开始，中止所有与 HP 的合作，因为来自 Sun 的另外一个团队会接替 HP 来完成所有与硬件有关的工作。因为 HP 的离开和 Sun 的加盟，很多以前做过的工作需要推倒重来。在我看来，这个版本注定是仓促而缺乏诚意的，因为产品不是由技术因素在主导，而是市场决定这一切，所以没过多久它就被下一任 X2 所取代。

与上一代产品相比，除了硬件全部替换为 Sun 的硬件外，最大的亮点是每台存储服务器上添加了 4 块 96G 的 PCI-e 闪存卡，并且在软件上开发了全新的与之配套的闪存技术——Exadata Smart Flash Cache。Oracle 的宣传册已经悄悄地由“为数据仓库设计”修改为“支持 OLTP”了。

同时，另外一个重要的更新是数据库版本由原来的 11.1 更新到了更主流的 11.2 版本，因为绝大多数客户习惯性地让生产系统运行在第二个版本上。因为第二个版本更稳定，产品支持的周期也更长。并且在这个版本中，引入了新的“一键安装”部署工具 onecommand。onecommand 原本是 Oracle 的一个内部项目，可通过预先设置一些配置文件来批量快速地部署 Oracle RAC 数据库。Exadata 有了 onecommand 以后，极大地减少了安装部署的工作量，降低了出现人为错误的概率，使得整个安装过程变得简单化、标准化。相比 Exadata V1，用在安装部署上的时间缩短了一大半。

1.3.3　Exadata X2

Exadata X2 是一个里程碑似的版本，因为这个版本标志着 Exadata 已经逐渐走向成熟和稳定。硬件方面，Exadata X2 进一步增强了 CPU 的处理能力和内存的容量，同时更好地处理了各项性能指标之间的平衡关系。另外，将原来的 SATA 接口磁盘全部替换为了 SAS 接口的磁盘，从而大幅提高了磁盘的可靠性和速率。软件方面，继续对其进行了完善，开发出了针对闪存的新功能 Exadata Smart Flash Logging。此功能在某些情况下能大幅提升 I/O 性能，极大地减少了 log file sync 之类的与 I/O 相关等待事件的出现。

Exadata X2 包括两个版本——X2-2 和 X2-8。前者依然是按照常规的扩展方式，主要针对数据仓库用户。后者减少了数据库节点，一来减少了可能的数据库节点之间的 global cache fusion，二来增强了单个节点的处理能力，适合于 OLTP 或者负载混合型系统。

另外，在以前的 Exadata 中，只能按照 Oracle 特定的型号进行扩容，例如在 Exadata V2 的时代，如果用户觉得 1/4 配的存储空间不够用，需要对空间进行扩展，那么唯一的选择就是升级到 1/2 配或者 1/1 配。如果升级到 1/2 配，用户不仅需要购买 4 台存储服务器，同时也需要购买 2 台数据库服务器。而很多时候用户并不需要增加数据库的计算能力，只