

实验室验证 报告

EMC Greenplum 数据计算应用装置

作者: Ginny Roth 和 Tony Palmer

2011年5月

目录

简介	3
EMC Greenplum 数据计算应用装置.....	4
ESG 实验室验证	5
物理测试环境	5
数据模型	5
加载并执行	6
线性扩展能力	8
分析就绪性	10
ESG 实验室验证要点	12
要考虑的问题	12
重要事实	13
附录	14
DCA GP1000 测试环境硬件配置详细信息（全机架）	14
RAID 配置详细信息	15
测试查询 SQL 代码	15

ESG 实验室报告

ESG 实验室报告的目的是，让 IT 专业人员了解存储、数据管理和信息安全领域内新兴的技术和产品。ESG 实验室报告不是为了替代在做出采购决定前应当进行的必要评估过程，只是为了让您了解这些新兴技术。我们的目标是介绍一些更有价值的产品特性/功能，展示如何使用它们解决真实的客户问题，并确定需要改进的地方。ESG 实验室专家的第三方观点是基于我们的亲手测试以及在与生产环境中使用这些产品的客户进行交流得出的。此 ESG 实验室报告由 EMC Greenplum 赞助。

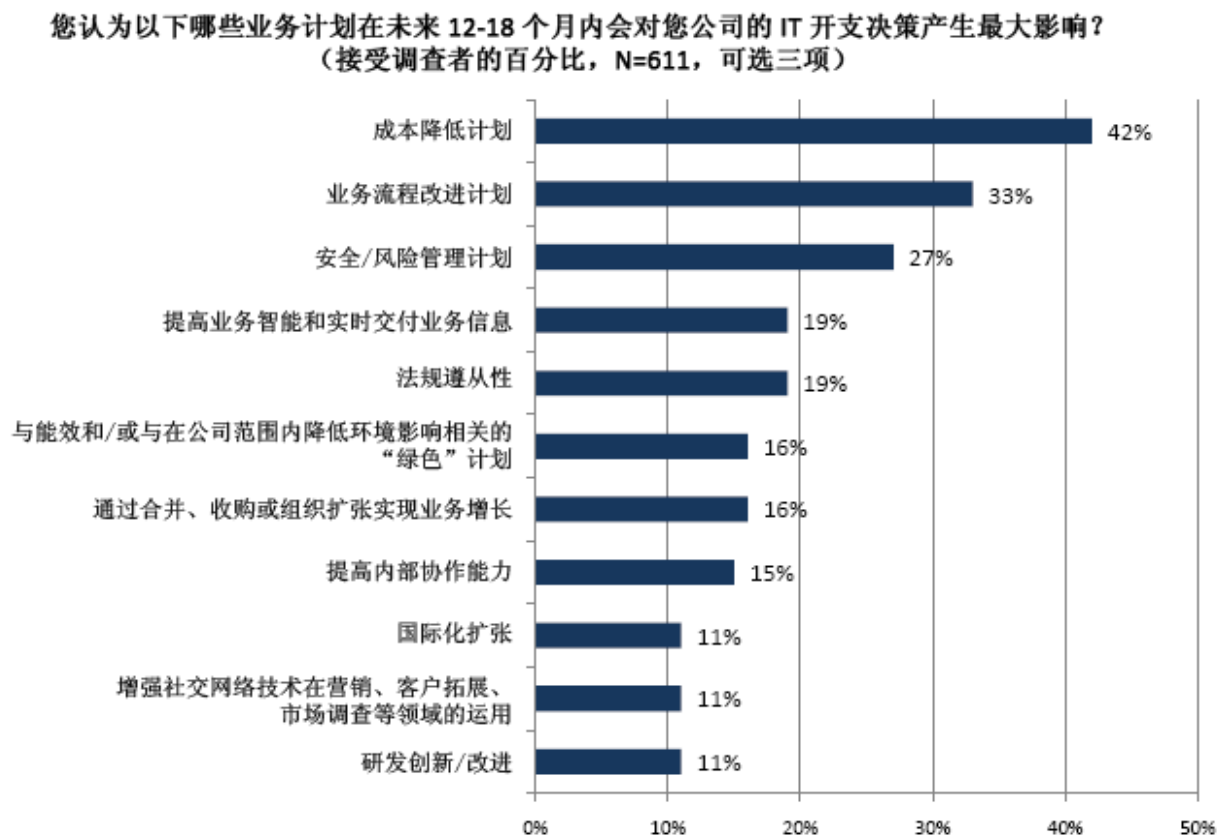
所有商标名称是其各自公司的资产。本出版物中包含的信息是由 Enterprise Strategy Group (ESG) 认为可靠的来源提供的，但 ESG 不保证其可靠性。本出版物可能包含 ESG 的观点，这些观点会随时发生改变。本出版物的版权归 Enterprise Strategy Group, Inc. 所有。未经 Enterprise Strategy Group, Inc. 明确许可，不得对本出版物的整体或部分以硬拷贝方式、电子方式或其他方式进行复制或将其分发给未经授权的任何个人，否则都将违反美国版权法并将引起民事损害诉讼，乃至刑事诉讼。如有疑问，请与 ESG 客户关系部门联系，电话：(508) 482.0188。

简介

近几年来，业务数据正以指数级速度增长，并将在未来可预见的一段时间内继续维持这种增长势头。企业希望整顿其基础架构和流程，从而通过这种调整在无需付出等价的巨额成本的情况下适应急剧上升的增长率，这时“大数据”也就带来了大挑战。考虑到这一点，ESG 实验室验证了 [EMC Greenplum](#) 数据计算应用装置 (DCA) 的实际性能和功能。测试方式是采用真实零售数据模型对运行大型数据集的 DCA 平台的易用性、性能、扩展能力和分析就绪性进行评估。

ESG 的调查表明，2011 年，对 IT 开支最有影响的前两项业务计划是降低成本和改进业务流程（见图 1）¹。紧随其后的第四项计划是提高业务智能和实时交付业务信息。

图1. 对 IT 开支最有影响的业务计划



来源：Enterprise Strategy Group，2011 年。

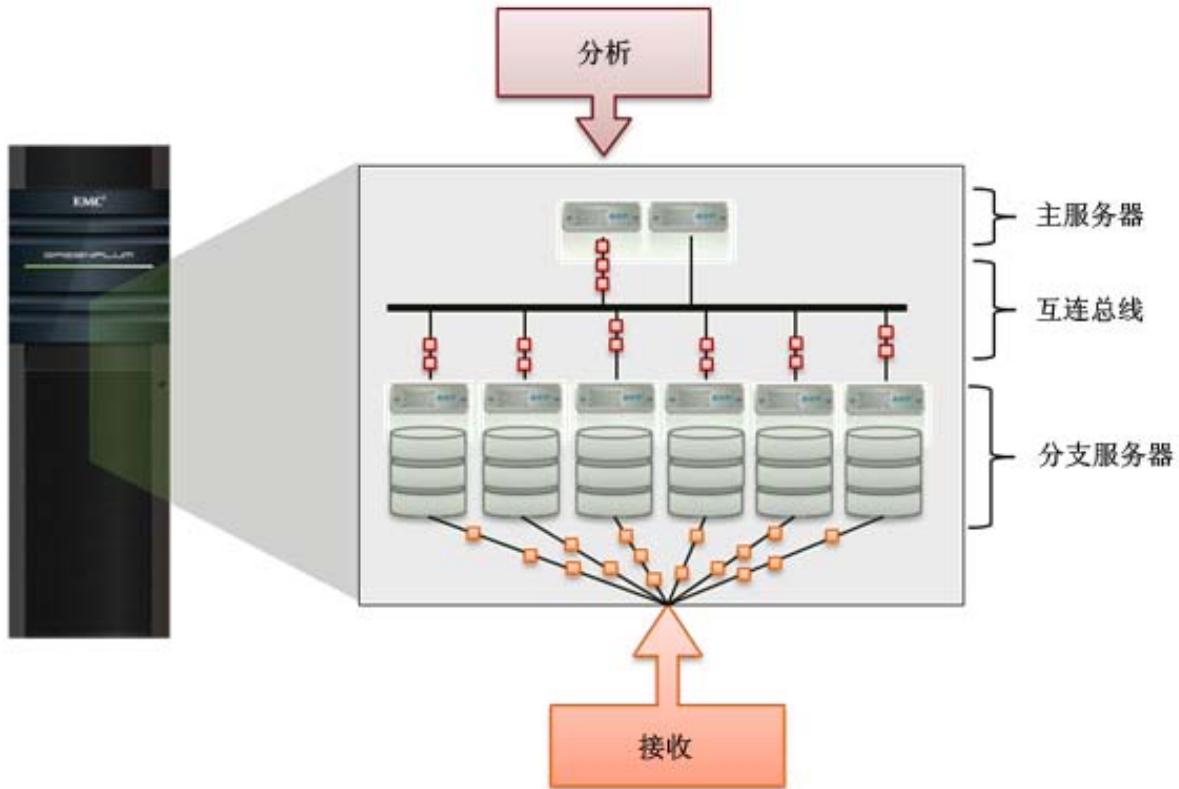
存储大量数据所产生的投资回报率 (ROI) 与公司的数据利用能力有直接关系。换言之，数据只有在您能对其加以适当利用的前提下才会算作一种资产。随着“数据科研人员”等新兴职位的出现，这种关系变得尤为明显。“数据科研人员”的部分工作为黑客性质，部分工作为定量分析，他们可以帮助公司利用所存储的丰富数据获得竞争优势。较为传统的报告行为由数据分析和数据挖掘这两方面进行了补充完善，这两者要求企业具备较强的培训和技术体系。该职位之所以涉及编程，部分原因在于越来越少的公司将传统数据仓库模型作为针对大数据的解决方案。数据的增长速度和变化速度太快，导致设计师无法将数据预先组织并整合到一个符合要求的模型中。这种情况虽然很适合报告行为（问题已知并已标准化），但不利于数据分析和挖掘，因为对于后两者，访问原始表格中的所有数据是非常关键的，而问题是其临时性的或者完全不可知性。从分析层面来讲，尝试根据活动调整数据只会是徒劳无功。在分析层面，数据使用模式并无任何界限。支持分析所需的处理能力、容量和吞吐量很容易就会超出传统堆栈式基础架构的成本/收益。

¹ 来源：ESG 调查报告，[2011 年 IT 开支意向调查 \(2011 IT Spending Intentions Survey\)](#)，2011 年 1 月。

EMC Greenplum 数据计算应用装置

EMC 和 Greenplum 携手合作，旨在利用数据计算应用装置 (DCA) 来解决上述业务问题。DCA 是一种一体化数据分析解决方案，能够以市场上最高的性价比提供处理最为繁杂的大数据分析所需的容量和功能。DCA 中有一种高级设计，那就是 MPP 体系结构。该体系结构采用主服务器（一台处于活动状态，一台备用）进行统筹，采用分支服务器处理繁重工作负载。所有的服务器都通过一条互连总线利用 24 个 10 GbE 端口和 10 个光纤通道端口实现内部连接。型号 GP1000 是一种配备 2 台主服务器（一台处于活动状态，一台备用）和 16 台分支服务器的全机架，其中每台分支服务器都能生成多个处理线程。

图2. EMC Greenplum 数据计算应用装置概况



本报告通过验证 DCA 的同类最佳接收速度、线性扩展能力以及对 TB 级海量数据进行复杂分析的能力测试，展现了 DCA 现有的性能和功能。所有这一切使用的都是可容纳超过 10 TB 或 500 亿行数据的真实零售电子商务数据模型。具体而言，本报告展示了单个 DCA 如何实现：

- 在无需宕机且无需重新加载数据的情况下从四分之三机架快速扩展到全机架。
- 成为一种带有线性扩展能力（性能的提高程度与增加的资源数量等效）的真正的 MPP 体系结构。
- 在不使用高速缓存的情况下利用短短几秒的时间完成数十亿行数据的扫描。
- 以本机方式集成 Alpine Miner 等主流分析工具。
- 凭借无需索引等传统性能调整对象的优势实现管理上的简易性。

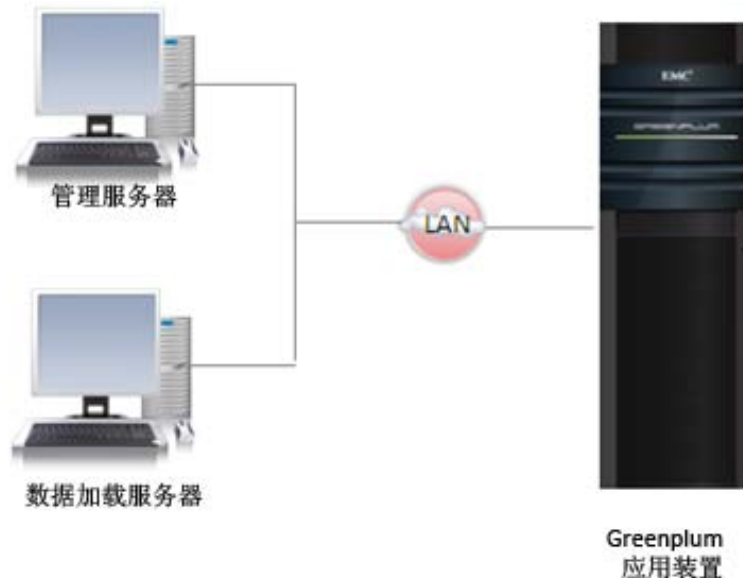
ESG 实验室验证

EMC Greenplum DCA 的实际性能和功能评估由 ESG 实验室在位于美国加利福尼亚州圣马特奥市的 EMC Greenplum 工作场所通过实际操作测试完成。本报告中提出的方法是采用实际零售数据模型对运行极大型数据集的单个数据计算应用装置 (DCA) 的性能和功能进行评估。此数据模型旨在通过从 Amazon、美国国家统计局、美国邮政总局及其他可用资源收集到的真实数据来展示电子商务系统的购买模式和周期性特征。该项目的实施得到了测试时在场的 EMC Greenplum 营销、工程、体系结构及管理帮助。

物理测试环境

图 3 中展示的起始测试环境是一套现成的 DCA GP1000，带有四分之三机架配置，能够产生 108 TB 的可用压缩后容量²。每套 GP1000 都拥有两台主服务器；四分之一机架、二分之一机架、四分之三机架和全机架之间的主要区别在于分支服务器的数量。

图 3. EMC Greenplum DCA 实验室配置



数据模型

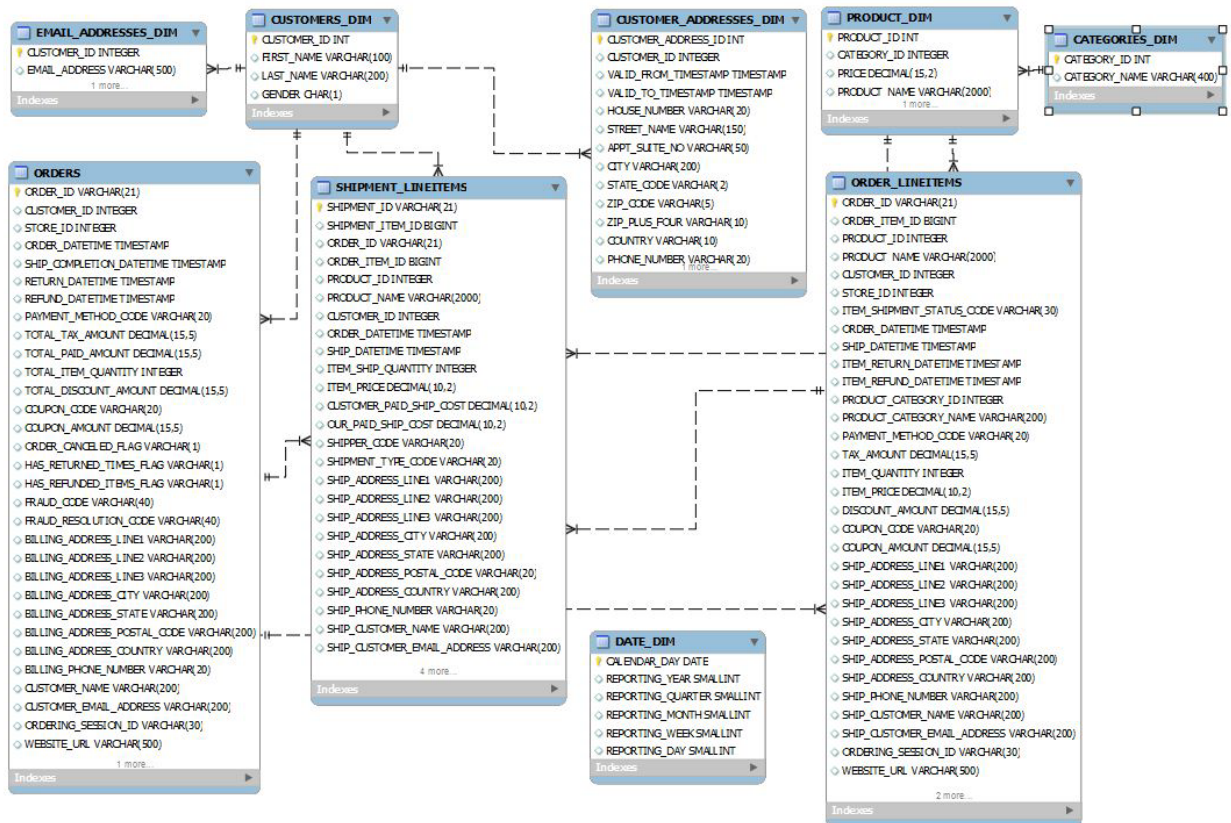
安装在测试计算机上的数据模型由 EMC 设计，代表一种基本的零售电子商务应用程序。其中一共包含七个维度表和三个非常大的事实数据表，前者涵盖基本的参考信息（客户、产品类别查找等），后者涵盖所有的订单、订单行项目和运输行项目（见图 4）。表是数据库中创建的唯一对象 — 数据库中没有索引、视图、具体化视图或其他常见的调整对象。实体关系图上的虚线表示表键之间的逻辑引用，而不是物理（主键/外键）引用。数据库中预先加载了约 10 TB（500 亿行）的数据，涵盖了五年的事务。前四年的三个事实数据表按月划分范围，当前的一年则按周划分范围。这些数据旨在反映零售业务周期伴随假期订单量增加、夏季订单量减少及其他特征（例如，平均订货量）而呈现的典型起伏趋势。其他特征包括：

- 销售量逐年增长，反映出业务不断发展
- 数据的月份和季度差异
- 真实的产品、名称和用户统计信息
- 有些产品的销量不错，有些产品则相反

² DCA 配置为在默认情况下使用“快速压缩”算法来压缩数据，该算法在写入时执行压缩。未压缩的可用容量在四分之三机架中为 27 TB 或压缩后的大小的四分之一。

- 绝大部分都是单项订单，只有个别是大额订单

图4. 验证测试期间使用的数据模型



加载并执行

“加载并执行”测试用于衡量数据库应用装置消耗（或接收）新建数据的速度有多快。ESG 实验室团队在 DCA 四分之三机架测试环境中额外加载 825 GB 的数据（约为 6 个月的事务量）并对结果计时，以此验证 DCA 的接收速度。这相当于在包含 24 个（6 个月 x 4 个按周划分的区域）物理对象的单个逻辑表中加载 19 亿行数据。

通过记录从 ETL 服务器上的源表格 (order_lineitems_load) 向 DCA 上的目标表格 (order_lineitems) 插入 825 GB 数据所花费的时间，可以测算出 DCA 的接收性能。通过直接连接至 DCA 互连总线的 10 GbE 链路来连接 ETL 环境。所用的插入语句是：

```
INSERT INTO retail_demo.order_lineitems
SELECT * FROM retail_demo.order_lineitems_load;
```

完成数据加载所用的时间是 351 秒，速率约等于每小时 8.26 TB。这比每小时 7.3 TB³ 的预期加载速率高出了将近 1 TB。

³ 计算依据是 2010 年 10 月发布的《EMC Greenplum 数据计算应用装置：数据仓库存储和业务智能方面的高性能解决方案 — 体系结构概述》(EMC Greenplum Data Computing Appliance: High Performance for Data Warehousing and Business Intelligence – An Architectural Overview) 第 42-43 页上公布的二分之一机架和全机架性能。

表1. “加载并执行”性能测试结果

DCA 接收性能	二分之一机架	四分之三机架	全机架
加载速率（基准值）	4.77 TB/小时	7.285 TB/小时（估计值）	9.8 TB/小时（估计值）
加载速率（测试结果）	不适用	8.263 TB/小时	不适用

意义

向应用装置上快速转移数据的能力对生产效率而言非常重要。在分析应用程序中，提取和转移大型数据集就是一种时间上的浪费，因为这样会导致生产效率基本为零。快速接收大型数据集使得分析人员可以腾出更多的时间来进行分析，同时减少处理数据收集工作所浪费的时间。对于要求实时报告的任务关键型业务流程，数据加载速率更是至关重要。缩短确定潜在欺诈事务、监控风险措施和确定客户订单系统问题过程中产生的延迟直接意味着需要改善业务流程和最终结果。

线性扩展能力

数据库应用装置的扩展能力是管理组织内指数级数据增长所必不可少的条件。通过额外添加四台分支服务器将 GP1000 从四分之三机架（起始配置）扩展为全机架，以此测试了容量和性能的横向扩展。如前所述，16 台分支服务器全都在测试环境中预先接线，其中只有 12 台正在由主服务器使用。这样就可以在测试过程中更多地侧重于在 DCA 仍然处于联机和使用状态的情况下对这 16 台分支服务器均匀地重新发现有数据（超过 10 TB）所需的步骤和时间。

ESG 实验室测试

ESG 实验室使用一组在添加分支服务器之前、期间和之后所运行的基准测试查询来验证 GP1000 的线性扩展能力；图 5 展示了使用 Greenplum Query Plan GUI 监测的某次查询的执行情况。此外，ESG 实验室还通过同时运行额外的查询验证了在 DCA 仍然处于联机和开放状态（可供用户查询）时将数据从 12 台分支服务器重新分发到 16 台分支服务器的情况。记录完成这一过程所需的时间采用的方法是：使用某个终端的标准输出中的时间戳，并测算从执行重新分发命令到完成该命令二者之间的时间差。只有在新的分支服务器联机后快速弹跳数据库时才需要宕机，这样是为了让主服务器识别出新配置。根据记录，重新分发超过 10 TB 的数据所需的时间为 1 小时 22 分钟。用户无需重新加载数据，因为这些操作已经全部由 DCA 动态处理了。

图 5. 监控“杀手锏”查询的执行情况



衡量之前和之后的 DCA 性能结果时采用了复杂性和数据行扫描数量都各有不同的四种基准查询。GP1000 DCA 不在内存中高速缓存查询结果，所以在执行第二次或第三次查询时没有性能偏差。下面的表 2 详细列出了完整的结果和性能差异（增量）。

表2. 从四分之三机架到全机架的性能横向扩展结果

测试查询 ⁴	扫描的数据行	四分之三机架	全机架	增量
单份订单的详细数据	5,253,880	6160 毫秒	4600 毫秒	-25.32%
订单行项目计数（1 个月）	369,789,719	13028 毫秒	9770 毫秒	-25.01%
订单明细项目计数（6 个月）	1,993,167,486	64394 毫秒	45548 毫秒	-25.08%
“杀手锏”查询	数十亿	178 秒	109 秒	-35.88%

意义

IT 和财务部门常常疲于应对堆栈式基础架构的容量规划，这种规划更多地侧重于技巧而非技术。这类基础架构拥有的移动部件过多，仅仅是为了作出一种有根据的推测就需要大量的开销和计算。此外，尝试确定是否需要更多的 CPU、更高的吞吐量、更快的磁盘转速或更大的容量也会是一种颇有难度的权衡之举，这通常会造成因过度分配资源而投入过多，更糟的情况是扩展不足以至于无法满足需求。采用了基于无共享体系结构的 MPP 设计的一体化解决方案的优势在于，在增加 25% 的容量和处理能力时，可额外获得 25% 的容量和处理能力。这是一种简单的运算，无需考虑配置决策 — 您只需决定是要增加 25%、50% 还是 100%。再加上此操作能在系统处于联机状态的情况下完成，而且无需通过物理方式迁移数据，因此您就拥有了一种易于管理且具备可预知性能的解决方案。

⁴ 计算依据是 2010 年 10 月发布的《EMC Greenplum 数据计算应用装置：数据仓库存储和业务智能方面的高性能解决方案 — 体系结构概述》(EMC Greenplum Data Computing Appliance: High Performance for Data Warehousing and Business Intelligence – An Architectural Overview) 第 42-43 页上公布的二分之一机架和全机架性能。

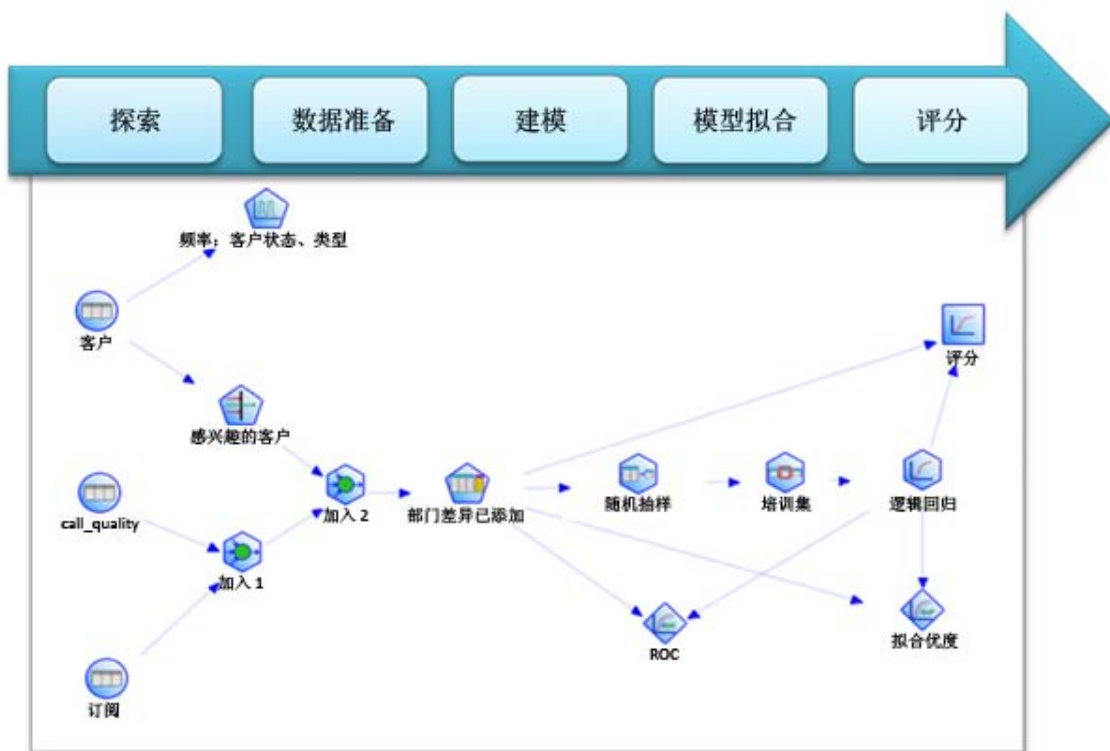
分析就绪性

要了解某种应用装置的价值，关键是要了解它的集成度是否足够高以及是否有助于促进分析。从功能角度来看，可以考虑在所有数据都位于 DCA 中的情况下，分析人员将可以多么轻松地插入所选工具并着手开展业务。从性能角度来看，可以考虑 DCA 是否能够处理支持完整分析生命周期所需的高要求计算、IO 和吞吐量。ESG 实验室测试中该部分的目标是利用主流分析工具从最终用户的角度检查 DCA 的功能。

ESG 实验室测试

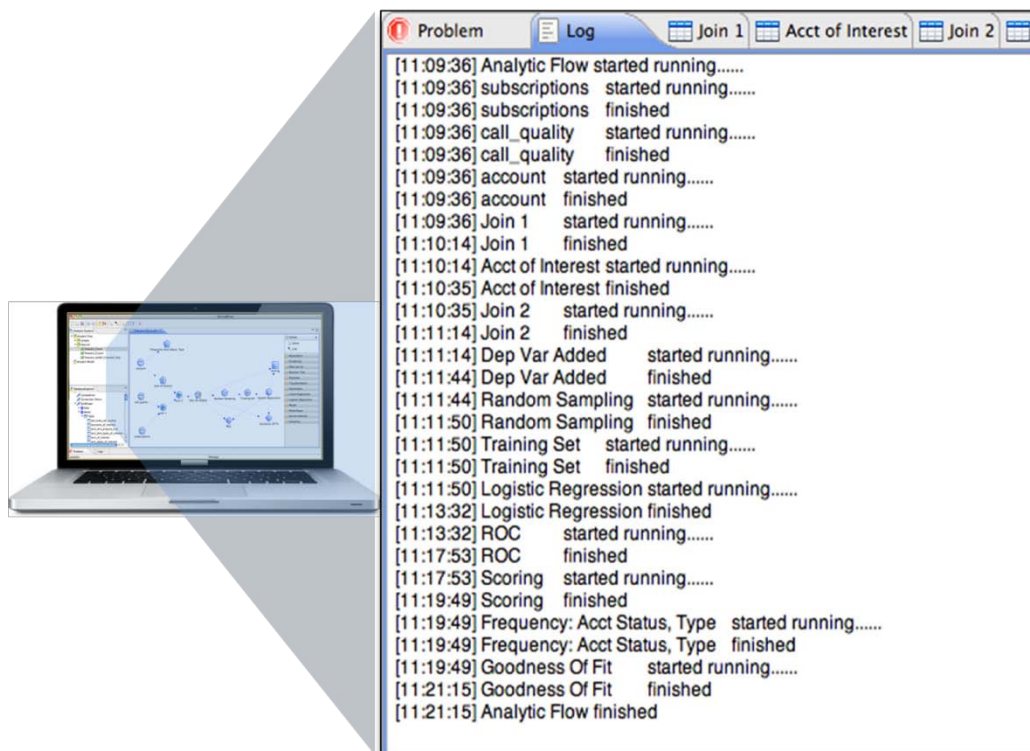
ESG 实验室利用 Alpine Miner 设计和构建端到端分析，并将 DCA 作为数据存储库和计算引擎。Alpine Miner 提供快速且易于使用的界面来构建多阶段分析，包括数据准备、数据转换、数据建模和数据评分。所用的“搅拌”模型旨在为高风险客户挖掘多个数据表。图 6 显示了从原始数据开始的各个分析阶段的详细信息。ESG 实验室使用 Alpine Miner 执行日志来记录模型的处理时间。

图 6. 实验室验证测试期间使用的分析 workflow



这种 14 阶段模型对大约 180 万客户运行计算，并在 11 分 29 秒内完成计算：

图 7. Alpine Miner 模型日志结果



意义

数据已从资产发展成为**战略性**资产。报告和业务智能活动继续在组织的管理和流程改进方面发挥关键作用。拥有成熟的数据报告技术的组织正将其战略扩展到分析领域，以便寻求新的机遇并在问题真正发生之前预先发现问题——从本质上来讲，这是一种未雨绸缪之举。分析和数据挖掘正快速成为一种核心竞争力和战略计划。对 IT 部门而言，在数据集相同的前提下，临时分析所占用的资源远远超出（大约是几个数量级之多）重复性报告活动。EMC Greenplum DCA 专为处理最为繁重的分析工作负载而全新构建，并通过集成 Alpine Miner 等行业标准工具为从探索到评分这整个分析周期提供了支持。

ESG 实验室验证要点

- ☑ 现成四分之三机架配置接收速度已确认超过 8 TB/小时。
- ☑ ESG 实验室已验证 EMC Greenplum MPP 体系结构在容量和性能方面都具备可预见的线性扩展能力，可简化容量规划。
- ☑ EMC Greenplum DCA 提供了与 Alpine Miner、SAS 和 R 等主流分析套件的即插即用型集成功能。
- ☑ ESG 实验室已确认无需移动数据、为数据建模或调整数据库即可进行复杂的分析。
- ☑ 性能测试显示 EMC Greenplum 的本机写入时压缩能够增加可用容量，并且不会对性能产生负面影响。
- ☑ ESG 实验室能够在只需短暂弹跳数据库而不必进一步宕机的前提下扩展基础架构。

要考虑的问题

- ☑ 目前，EMC Greenplum DCA 的任意扩展都必须由 EMC 工程师进行处理。根据客户的观点，由专家到现场负责处理扩展可能有助于降低风险。
- ☑ DCA 目前不含集成分析包。任何超出 SQL 能力范围的分析都需要借助外部分析套件完成。虽然数据库内分析包表现良好并提供了一定程度的便利性，但 EMC Greenplum “不限分析工具”的方式使分析员能够使用其擅长的工具和技巧。
- ☑ 截至本文撰写时，仅采用直接连接的 600 GB SAS 驱动器作为内部 DCA 存储介质，这样很好地组合了速度和容量。EMC 已指出，EMC Greenplum DCA 的战略路线图中包括扩展的后端存储选项。
- ☑ 公司从堆栈式基础架构转移到应用装置后，可能会产生一定的前期投资成本，但随着环境的扩展将来所需的成本会较低。

重要事实

各组织在扩展基础架构以处理激增的数据量和高要求的分析方面正面临着一些挑战。依靠数据建模、调整和预聚合数据的传统 BI/报告体系结构不适合大数据的分析。

方法	BI/报告	分析/挖掘
业务驱动因素	问题已知	问题可能未知
活动特征	重复性和可预知	临时性和高度可变性
数据粒度	聚合和多维数据集	原始 — 全部
有益于 SQL	高度地	最低限度地

报告通常是一种“一次构建、多次运行”的操作，分析则是一种“以多种方式构建、少量几次运行”的活动。IT 部门必须融合各种技能，以便运行和优化快速扩展的 IT 基础架构堆栈，该堆栈由应用程序、数据库、服务器、存储设备和网络等组成。因为对实时信息的渴望持续增长，所以即使投入许多运营成本并混合各种复杂的解决方案也不足以适应快速发展的业务需求。数据科研人员现在需要融合多种类型的源数据 — 包括结构化和非结构化的外形 — 以便生成综合全面的视图。针对当前数据类型开发的复杂查询可能无法执行，或者无法轻松适应多种数据类型。此外，生成复杂时间系列分析（回归、评分、移动平均数等）的需求与主流的关系数据库设计背道而驰。

EMC Greenplum 正以一种易于管理、性能强大且扩展能力极强的解决方案来解除这些限制。ESG 实验室团队不仅验证了大数据日益增长时 DCA 所展现的无缝扩展能力，而且验证了 DCA 在加载和分析大型数据集方面的惊人性能。最低的管理和调整量、快速的接收能力以及强大的分析性能相结合，意味着数据团队可将更多的时间用于分析问题和提供深入见解。

EMC Greenplum DCA 提供了一种低风险的纵向集成式数据库应用装置，简化了 IT 部门在支持大数据对性能和扩展的要求时执行的工作，而且不会像传统数据库管理解决方案那样增加复杂性。因为数据的加载和分析可在数据库中采用并行处理体系结构执行，所以企业体系结构在系统和服务器整合方面获得了新的机遇。高级分析算法和工作负载管理功能可在数据库内运行，使得数据设计师和开发人员能够继续采用最符合其需求与技能的语言和应用程序进行编程，同时尽可能再次利用 MPP。

EMC Greenplum DCA 从各个角度推动了生产效率。从降低维护开销到加快提供业务见解的速度，这些优势体现了该数据计算设备更注重结果。对于需要提高大数据分析性能和效率的 IT 组织，建议其了解 EMC Greenplum DCA。

附录

表3. 高级硬件配置详细信息

GP1000	起始（四分之三机架）	结束（全机架）
主服务器数量	2	2
分支服务器数量	12	16
CPU	144	192
内存大小	576 GB	768 GB
未压缩/压缩后容量	27 TB/108 TB	36 TB/144 TB

DCA GP1000 测试环境硬件配置详细信息（全机架）

主服务器配置（2 台）

- 处理器—2 个 Intel X5680 3.33 GHz（6 核）
- 内存—48 GB DDR3 1333 MHz
- 双端口聚合网络适配器—2 个 10 Gb/s
- RAID 控制器—双通道 6 Gb/s SAS
- 硬盘—6 个 600 GB 10k SAS
- 操作系统—RHEL 5.5

分支服务器配置（16 台）

- 处理器—2 个 Intel X5670 2.93 GHz（6 核）
- 内存—48 GB DDR3 1333 MHz
- 双端口聚合网络适配器—2 个 10 Gb/s
- RAID 控制器—双通道 6 Gb/s SAS
- 硬盘—12 个 600 GB 15k SAS
- 操作系统—RHEL 5.5

互连总线

- 24 个 10 GbE 端口
- 8 个光纤通道 (FC) 端口

RAID 配置详细信息

表 4. DCA GP1000 RAID 配置

服务器类型	RAID 组	物理磁盘数量	虚拟磁盘	功能	文件系统
主服务器	组 1 RAID 5 (4+1)	5	虚拟磁盘 1	ROOT	ext3
			虚拟磁盘 2	SWAP	SWAP
			虚拟磁盘 3	DATA	XFS
	热备盘	1	不适用	DATA2	不适用
分支服务器	组 1 RAID 5 (5+1)	6	虚拟磁盘 1	SWAP	SWAP
			虚拟磁盘 2	DATA2	XFS
	组 2 RAID 5 (5+1)	6	虚拟磁盘 1	SWAP	SWAP
			虚拟磁盘 2	DATA2	XFS

测试查询 SQL 代码

-- 单份订单的详细信息（测试查询 1）

```
SELECT * FROM retail_demo.order_lineitems
WHERE order_id = '145338570'
AND order_datetime BETWEEN date '2006-01-27' AND date '2006-01-28';
```

-- 订单行项目计数（测试查询 2）

```
SELECT TO_CHAR(COUNT(*), '999,999,999') AS cnt
FROM retail_demo.order_lineitems
WHERE order_datetime BETWEEN date '2010-11-01' AND date '2010-11-30';
```

-- 订单行项目计数（测试查询 3）

```
SELECT TO_CHAR(COUNT(*), '999,999,999,999') AS cnt
FROM retail_demo.order_lineitems
WHERE order_datetime BETWEEN date '2010-01-01' AND date '2010-06-30';
```

-- “杀手锏” 查询（测试查询 4）

-- 以下客户：

-- 在之前 2 年（2008 年和 2009 年）的节假日期间购买了 DVD

-- 在过去的 6 个月中购买了蓝光播放器

-- 自购买蓝光播放器后未曾购买蓝光光盘

```
SELECT customers.customer_id
, email.email_address
, customers.Last_BluRay_2010
, customers.First_Player_2010
, SUM(RFMT.num_purchases) as total_purchases
FROM (SELECT oli.customer_id
, SUM(CASE WHEN cat.category_name = 'DVD'
AND oli.order_datetime BETWEEN date '11-01-2008' AND date '12-24-2008'
THEN item_quantity ELSE 0 END) AS DVDs_2008
, SUM(CASE WHEN cat.category_name = 'DVD'
```

```
        AND oli.order_datetime BETWEEN date '11-01-2009' AND date '12-24-2009'
        THEN item_quantity ELSE 0 END) AS DVDs_2009
,   MAX(CASE WHEN cat.category_name = 'DVD' AND prod.product_name LIKE '%Blu-ray%'
        AND oli.order_datetime BETWEEN date '05-01-2010' AND date '10-31-2010'
        THEN order_datetime ELSE NULL END) AS Last_BluRay_2010
,   MIN(CASE WHEN cat.category_name = 'CE'
        AND prod.product_name LIKE '%Blu-ray%'
        AND oli.order_datetime BETWEEN date '05-01-2010' AND date '10-31-2010'
        THEN order_datetime ELSE NULL END) AS First_Player_2010
FROM order_lineitems oli
,   products_dim prod
,   categories_dim cat
WHERE oli.product_id = prod.product_id
AND   prod.category_id = cat.category_id
AND   cat.category_name IN ('DVD', 'CE')
AND   (oli.order_datetime BETWEEN date '11-01-2008' AND date '12-24-2008'
OR     oli.order_datetime BETWEEN date '11-01-2009' AND date '12-24-2009'
OR     oli.order_datetime BETWEEN date '05-01-2010' AND date '10-31-2010')
GROUP BY oli.customer_id
) AS customers
,   email_addresses_dim email
,   customer_RFMT_scores RFMT
WHERE customers.customer_id = email.customer_id
AND   customers.customer_id = RFMT.customer_id
AND   DVDs_2008 > 0
AND   DVDs_2009 > 0
AND   Last_BluRay_2010 < First_Player_2010
GROUP BY customers.customer_id
,   email.email_address
,   customers.Last_BluRay_2010
,   customers.First_Player_2010
limit 100
```



Enterprise Strategy Group | **Getting to the bigger truth.**