甲骨文白皮书

一月，2012

甲骨文: 企业版大数据

执行摘要

如今，大数据这个术语吸引了大量的注意，但是在所有的炒作的背后有一个简单的故事。数十年来，公司一直在依靠存储在关系型数据库中的交易数据来制定商业决策。除了对关键数据，一些非传统的不太结构化的数据：博客、社交媒体、电子邮件、传感器、还有一些可以获得有用信息的图片也是一个潜在的宝库。存储和计算能力成本的下降是我们能够收集这些数据，这在几年之前是要被扔掉的。因此，越来越多的公司正在寻求那些在他们传统企业商业智能分析上的数据，包括非传统的但可能是非常有价值的数据。

要从大数据获得真正的商业价值，你需要合适的工具来捕捉和组织来自不同来源的各种数据类型，能够很容易的分析范围内的所有的企业数据。甲骨文提供了最广泛最完整的产品组合，帮助您获得并组这这些不同的数据类型和他们一起现有的数据进行分析，从隐藏的关系中找到新的见解。

介绍

在甲骨文最近推出了甲骨文大数据装置和甲骨文大数据连接，甲骨文成为第一家提供一个完整和集成的解决方案来满足全方位的企业大数据的要求的供应商。甲骨文大数据战略的重点是，你可以发展你的大数据架构，将大数据转化为商业价值的想法。通过发展你的企业架构，你可以充分利用你的甲骨文系统久经考验的可靠性、灵活性和性能，来解决大数据的需求。

大数据定义

大数据通常指以下类型的的数据：

传统的企业数据 - 包括客户信息的 CRM 系统，ERP 的事务数据，网上商店交易，总账数据。

机器生成/传感器数据 - 包括呼叫详细记录（CDR），博客，智能电表，制造

传感器，设备日志（通常称为数字垃圾），交易系统的数据。

社会资料 - 包括客户的反馈流，像 Twitter 的微博网站，如 Facebook 的社交媒体平台。

麦肯锡全球研究院估计，数据量每年增长 40%，2009 年和 2020 年之间增长 44 倍。但是，尽管它常常是最明显的参数，数据量并不是唯一重要的特性。事实上，有四种关键的特性来定义大数据：

数量。机器生成的数据比传统的数据数量更大。例如，一个单一的喷气式发动机可以在 30 分钟内生成 10TB 数据。每天拥有超过 25000 航班，每天仅这一个数据源长生的数据量就达到 PB 级别。智能电表和炼油厂和钻机等重型工业设备产生类似的数据量，使问题复杂化。

速度。社会化媒体数据流 - 而不是大量的机器生成的数据 - 大量涌入的选项和有价值的客户关系。即使每个 Tweet 只有 140 个字符，高的速度（或频率）的 Twitter 数据，确保将会产生大量的数据（每天超过 8 TB）。

多样性。传统的数据格式往往是相对较好的描述，并慢慢地改变。相比之下，非传统的数据格式都表现出了令人眼花缭乱的变化率。随着新服务的添加，部署新的传感器或执行新的营销活动，新的数据类型都需要捕捉得到的信息。

值。不同的数据的经济价值变化显著。通常情况下，有好的信息隐藏在非传统数据的机体之后。非传统的数据面临的挑战是确定什么是有价值的，然后转换和提取的数据进行分析。

为了更充分的利用大数据，企业必须发展他们的 IT 基础设施来处理能产生极端大量数据的速度，这些数据有不同的数据类型，它们可以集成到一个组织的其他企业数据进行分析。

大数据的重要性

对大数据结合传统企业数据进行提炼和分析，企业可以更加彻底和深入的理解他们的业务，这可以导致生产力的提高，更强的竞争地位和更大的创新----所有的这些都可以在这个底线上有一个显著的影响。

例如，在提供医疗服务上，慢性管理或者长期条件都是昂贵的。使用家庭监控设备来测量生命体征，检测进展情况仅仅是用传感器数据来改善病人的健康和

减少诊所就诊和住院的方式。

制造企业在他们的产品中部署传感器来返回检测的数据流。有时，这是用来提供服务的，像 OnStar 系统， 它提供通讯安全和导航服务。也许更重要的是，遥测也揭示了使用模式，故障率和产品的改进，可以减少开发和组装的成本。

智能手机和其他 GPS 设备的扩散，为广告商提供了一个机会去锁定消费者，当他们靠近商店，咖啡厅或者餐厅。这为服务提供商提供了新的收入，也为很多企业提供锁定新客户的机会。

零售商通常知道谁买他们的产品。使用社交媒体和他们自己电子商务网站上的网络日志文件，可以帮助他们了解谁没有买，为什么他们不买，还有今天不能向他们提供的信息。这可以更有效的进行微客户细分和开展针对性的营销活动，以及提高供应链的效率。

最后，如果没有大数据的话，社交网站例如 Facebook 和 LinkedIn 根本不会存在的。他们的商业模式是需要在网络上收集所有可用用户和成员的数据来提供个性化的体验。

建设大数据平台

伴随着数据仓库，网上商店和任何的 IT 平台，大数据的基础设施都有独特的要求。在考虑大数据平台的所有组件时，最重要的是要记住，最终目标是要能够轻松的集成你的大数据和企业数据，让你能在合并后的数据集上进行深入的分析。

基础设施的要求

一个大数据的基础设施的要求包扩数据采集，数据组织和数据分析。

获得大数据

在大数据出现之前，数据采集阶段是基础设施的重大变化之一。因为大数据是指具有更好的速度和更快的变化性的数据流，所需要的用来支持大数据采集的基础设施必须延迟低，在采集数据时延迟可预见，并且执行短，能进行简单的查

询，能够支持很高的交易量，并且经常在分布式环境中支持灵活的动态的数据结构。

非关系型数据库经常被用作采集和存储大数据。他们非常适用于动态的数据结构，并且是高度可扩展的。一个非关系型数据库中存储的数据通常是具有很高的变化性的，因为这个系统的目的是简单的捕获所有的数据，无需分类和分析数据。

例如，非关系型数据库经常用来收集和存储社交媒体数据。尽管面向客户的应用程序经常改变，底层的存储结构确实一直保持尽量简单。这些简单的结构往往只是包含数据的一个关键点，然后一个容器持有有关数据，而不是设计一个对应于实体之间的关系模式。这个简单的动态的结构，可以事改动无需支付昂贵的存储层的重组就能生效。

组织大数据

在传统的数据仓库的术语中，数据组织被称为数据整合。 正是因为大数据有这么大的数据量，才有一种趋势是在数据的原始存储位置进行数据的组织，从而节省了金钱和时间，因为不需要移动那么大量的数据了。组织大数据所需的基础设施必须能够处理和操作数据的原始存储位置，并且支持非常高的吞吐量（通常是批处理）来处理大型数据处理步骤，处理从结构化到非结构化的各种大量的数据格式。

Apache Hadoop 是一个新技术，它可以允许处理大数据量的同时保持数据的原始存储集群。Hadoop 分布式文件系统（HDFS）是一个长期存储系统，例如存储网络日志。通过在群集上运行 MapReduce 程序，生成汇总结果，在同一群集上，这些网络日志都变成了浏览行为（会话）。这些汇总的结果，然后加载到一个关系数据库管理系统。

分析大数据

由于数据在组织阶段并不总是在移动的，所以分析也可以在分布式的环境

中，其中的一些数据将保留在他最初存储的位置，并且透明的访问数据仓库。大数据分析所需要的基础设施必须能够支持更深入的分析，如统计分析和数据挖掘，在更广泛的多种类型的存储在不同系统上的数据上，规模达到极端的数据量，用行为的改变来带动更快的响应时间，和在建立在自动化分析模具基础上的自动化分析决策。最重要的是，基础设施必须能够结合大数据和传统企业数据的分析。新的认识不仅来自于新的数据分析，更来自于在该范围内从老的问题中得到的新的视角进行的分析。

例如，在自动售票机所在的场地中，结合自动售货机的库存数据进行分析，我们就能为自动售货机决定最佳产品组合和补充计划。

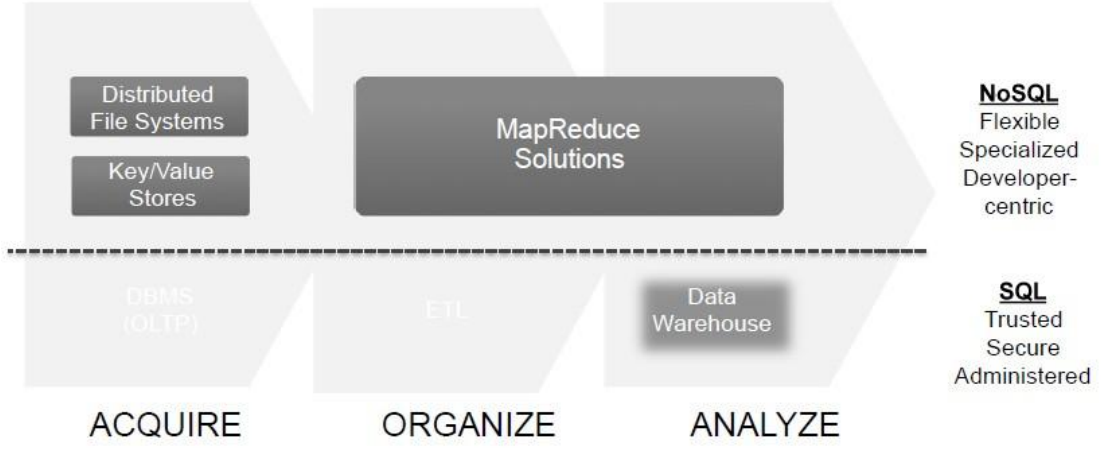
解决范围

许多新技术已经成为 IT 基础设施，以解决上述要求。在最近的一次统计中，有超过 120 个开源的键—值数据库，用来获取和存储大数据，正在兴起的 Hadoop 作为主系统来分析大数据，并且关系型数据库也将其范围扩大到不太结构化的数据集中来分析大数据。这些新的系统已经创建了一个分化的解决方案系列：

不只有 SQL（NoSQL 的）解决方案：开发人员为中心的专业系统。

SQL 的解决方案：通常等同于世界的可管理性，安全性和值得信赖的性质，关系数据库管理系统（RDBMS）

NoSQL 系统的设计是捕获所有的数据，不管它的分类和进入系统的入口，因此数据具有多样性。另一方面，SQL 系统通常将数据的结构定义零号，并且处理捕获数据的元数据，以确保数据的一致性和验证数据的类型。



图片一：分割的解决频谱

分布式文件系统和事务（键—值）存储主要用来捕捉数据，并且都是和文本前面讨论的要求保持一致。为了解释和提炼这些解决方案中的数据信息，使用名为 MapReduce 的一种编程范式。MapReduce 程序是运行在分布式数据节点的自定义编写的程序。

键—值存储或者 NoSQL 数据库是大数据世界的 OLTP 数据库，他们对快速的数据采集和简单的查询模式进行了优化。NoSQL 数据库可以提供非常快速的性能，因为被捕获的数据是用一个单独的标记键进行快速存储的，而不是一个被解释的架构。通过这样做，NoSQL 数据库能迅速的存储大量的事务。

然而，由于存储在 NoSQL 数据库中的数据性质的改变，任何数据的组织工作都需要编程来解释所用的存储逻辑。这一点，加上缺乏对复杂查询模式的支持，使得它很难成为最终用户提供一个 NoSQL 数据库中的数据提取值。

为了从 NoSQL 的解决方案中获得更多，并把它们从一专业开发人员为中心的解决方案转变到企业化的解决方案，它们必须与 SQL 的解决方案结合成一个单一的成熟的基础设施，才能满足当今企业对可管理性和安全性的要求。

甲骨文公司的大数据解决方案

甲骨文公司是第一个就各公司对大数据的要求提出详尽而又综合的解决方案的销售商。它的大数据战略的中心理念是使你在现有的公司数据架构的基础上，通过整合大数据，传递商业价值，利用已经验证的可靠性，灵活性和甲骨文系统的性能，来满足你对大数据的要求。

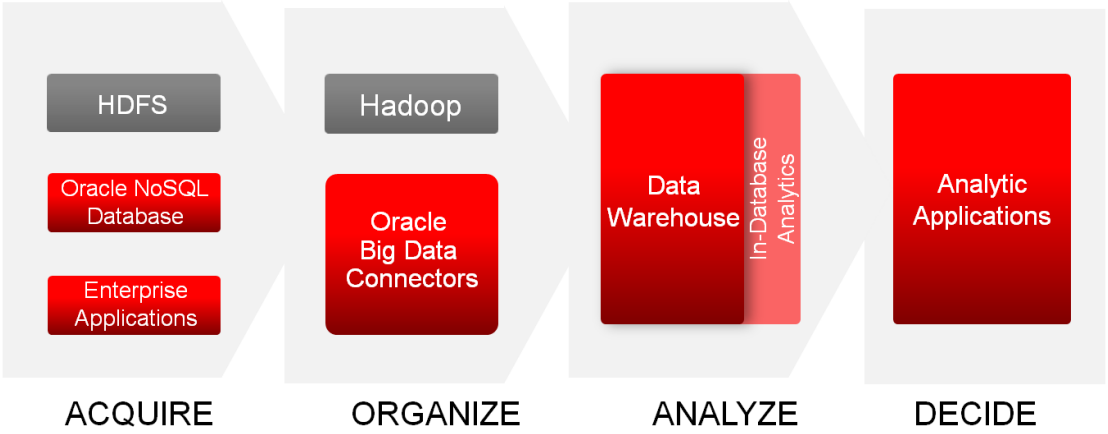


图 2---甲骨文大数据解决方案

甲骨文公司有这一独特的能力：它能结合所需来完成大数据这一挑战—把软件和硬件都整合到一个工程系统中。甲骨文公司的大数据设备就是这样一个由完善的硬件和最全面的软件栈相结合的系统，这一系统由甲骨文公司开发，为将大数据组织并加载到甲骨文系统提供了一个完整，易获取，易于部署的解决方案。这一设计通过企业级的性能，效益，耐力以及安全性对所有数据的种类做出了详尽地分析。借助大数据连接器的帮助，这一方案与 Oracle Exadata 和甲骨文数据库紧密结合，因此你能结合一些极端的性能来分析数据。

甲骨文大数据设备

甲骨文大数据设备有 18 台 Sun 服务器的全机架配置的总存储容量为 648TB。在机架中的每台服务器有 2 个 CPU，每个有 6 芯的为全机架，共 216 个核心。每个服务器都具有每满架的内存，864GB，总 48GB 内存。

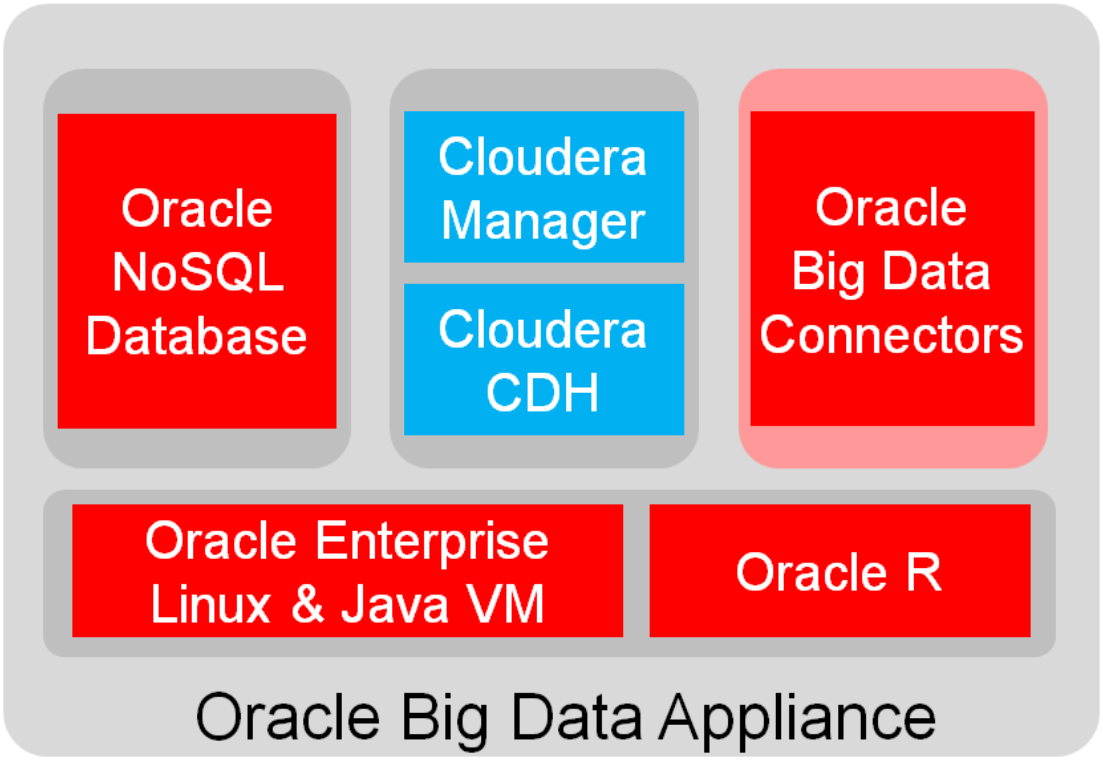


图 3---大数据设备上的软件的高级概述

这一由甲骨文公司开发的甲骨文大数据设备通过开放的软件资源与特殊软件相结合来满足企业的大数据要求。

甲骨文大数据设备集成软件包括：

Cloudera 分布式中的全分布，包括 Apache 的 Hadoop（CDH）；

Cloudera 的管理器来管理所有方面的 Cloudera 的 CDH；

开源分布式的统计分析 Oracle 大数据设备的未过滤的数据包 R；

甲骨文 NoSQL 数据库社区版；

Oracle Enterprise Linux 操作系统和 Oracle 的 Java 虚拟机；

CDH 和 Cloudera 管理器

甲骨文大数据设备包含 CDH 和 Cloudera 管理器。无论是基于 Apache Hadoop 的商业还是非商业的环境上，CDH 都是排在首位的。CDH 由 100％开源的 Apache Hadoop 加上 一整套需要使用 Hadoop 的开源软件组件。

Hadoop 的甲骨文加载器

使用者能够通过 OLH，经过 Hadoop MapReduce 的处理来创建优化的数据集，使得高效地下载和分析 Oracle Database 11g。与其它 Hadoop 加载器不同的是，甲骨文加载器能产生其内部格式能够更快速地加载数据，并且将占用更少的数据库系统资源。OLH 是作为单独的一部分最后加入 MapReduce 变换的。这最后一步通过使用 Hadoop 集群中的芯片将数据的格式转换为甲骨文系统可读的数据，来产生甲骨文集群中的更低的 CPU 加载和更高的数据采集速率，这是由于数据已经转换成甲骨文数据库所需的格式了。一旦数据被加载，将存在数据库中永久可用，并且为使用 SQL 或者商业智能工具的一般数据库使用者提供得到这个数据的快捷渠道。

Hadoop 分布式文件系统（HDFS）的甲骨文直接连接器

Hadoop 分布式文件系统（HDFS）的甲骨文直接连接器是一个从甲骨文数据库中直接得到数据的高速连接器。Hadoop 分布式文件系统（HDFS）的甲骨文直接连接器让使用者能够根据他们设备所需随时从 HDFS 中查询想要的数据，灵活而又方便。

这一部件使使用者能在甲骨文数据库中创建外部表，使得直接 SQL 能访问存于 HDFS 中的数据。然后存于 HDFS 中的数据便能通过 SQL 进行查询，与存储在甲骨文数据库中的数据放在一起，或是被加载进入甲骨文数据库。并在自动负载平衡的环境下，对 HDFS 中数据的访问方式将被优化为快速的数据移动并且将共同运作。HDFS 中的数据能在分隔的文件中或在由甲骨文 Hadoop 加载器创建的 Oracle 数据泵文件中。

甲骨文 Hadoop 应用程序配适器的数据集成器

甲骨 Hadoop 应用程序配适器的数据集成器通过甲骨文数据集成中易于使用的界面从 Hadoop 和甲骨文数据库上简化了数据集成。

一旦数据进入数据库，终端使用者将能够使用 SQL 和 Oracle BI 企业版来获取访问数据。已经在使用 Hadoop 解决方案并且不需要像甲骨文大数据设备这样的集成的产品的公司也能够通过 HDFS 与大数据连接器的结合作为独立的软件解决方案来合并数据。

甲骨文 Hadoop R 连接器

甲骨文 Hadoop R 连接器是一个提供进入 Hadoop 并获取存储于 HDFS 中数据的 R 包。甲骨文 Hadoop R 连接器给使用者提供了一个开放资源数据环境 R，它能够分析存储于 HDFS 中的数据，能在不要求 R 用户学习另一种 API 或是语言的情况下，利用 MapReduce 处理大量数据并运行规模上的 R 模型。终端使用者可以利用 3500 多个的开源 R 包对存储于 HDFS 中的数据进行分析，而管理员不需要在生产环境中学习 R 来设置 MapReduce 模型。

甲骨文非关系型数据库

甲骨文非关系型数据库是一个基于甲骨文伯克利数据库的，分布式的，高度可扩展的，键—值数据库。它达到了一个通用的目的，即企业级键值库在分布式的伯克利数据库之上添加了智能驱动程序。这种智能驱动器能跟踪底层的存储拓扑结构，粉碎数据，用最快的速度得出数据的准确位置。与同类竞争解决方案不

同的是，甲骨文 NoSQL 数据库是易于安装，配置和管理的，它支持大范围的工作量，并提供由甲骨文企业级所支撑的可靠性。

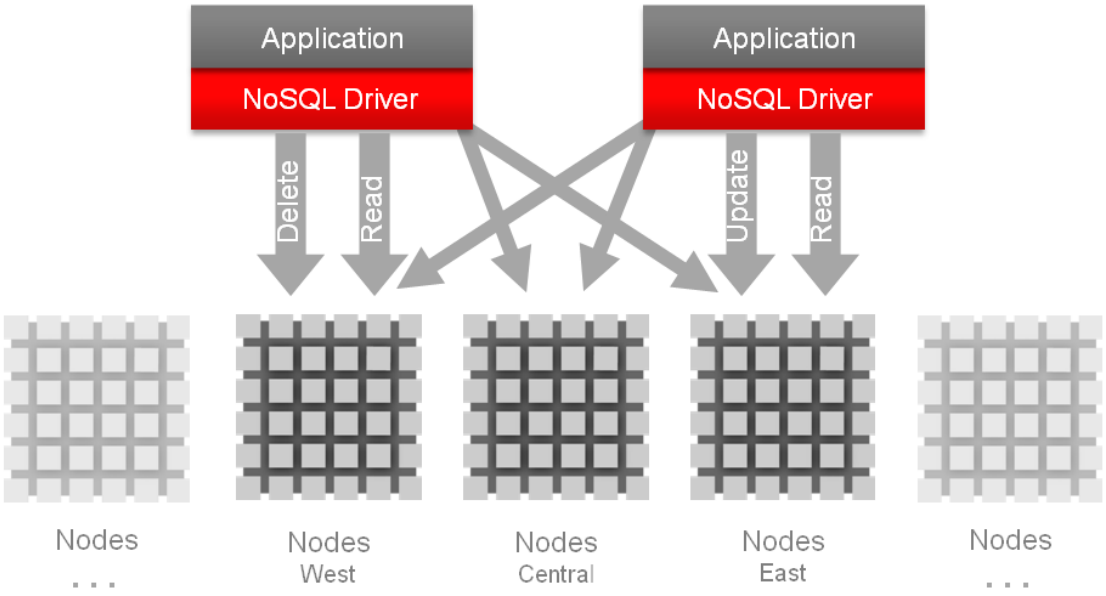


图 4　非关系型数据库结构

在甲骨文非关系型数据库中，主要使用的案例是通过键查找进行数据的低延迟采集和快速查询。在甲骨文 NoSQL 数据库中能很方便地使用 Java API 和管理框架。该产品在开源社区版和付费的企业版大型数据分布中心中均可获得。而前者是作为大数据设备集成软件的组成部分而安装的。

数据库内分析

一旦数据从甲骨文的大数据设备中被加载到甲骨文数据库或者甲骨文 Exadata 中，终端用户就可以使用以下任一简单的工具进行数据库内的深入分析：

一旦数据从 Oracle Big Data Appliance 中被加载到 Oracle Database 或 Oracle Exadata 中，终端用户可以使用以下任一操作简单的工具进行 in-database 而又深入的分析：

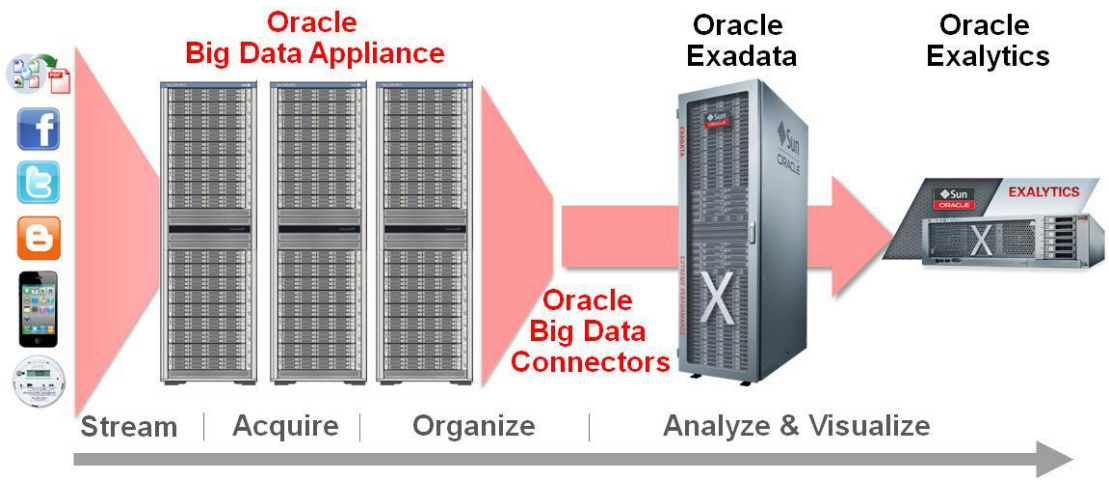


图 5　大数据设备和 Exadata 的使用模型

结语

通过对新型且多样的数字数据流的分析，能看出经济价值的新来源，给消费者和市场走向提供一个新视角。但是这一新数据的涌入也为 IT 部门带来了新的挑战。为了从大数据中获得真正的商业价值，你还需要正确的工具来获取和整合大量来源不同的数据类型，并且能够结合该公司数据情况进行简单的分析。通过使用甲骨文大数据设备和甲骨文大数据连接器，并在 Oracle Exadata 的帮助下，企业便能获取，整合并且分析他们所有的企业数据，不论是是结构化的还是非架构化的，由此做出最明智的决定。