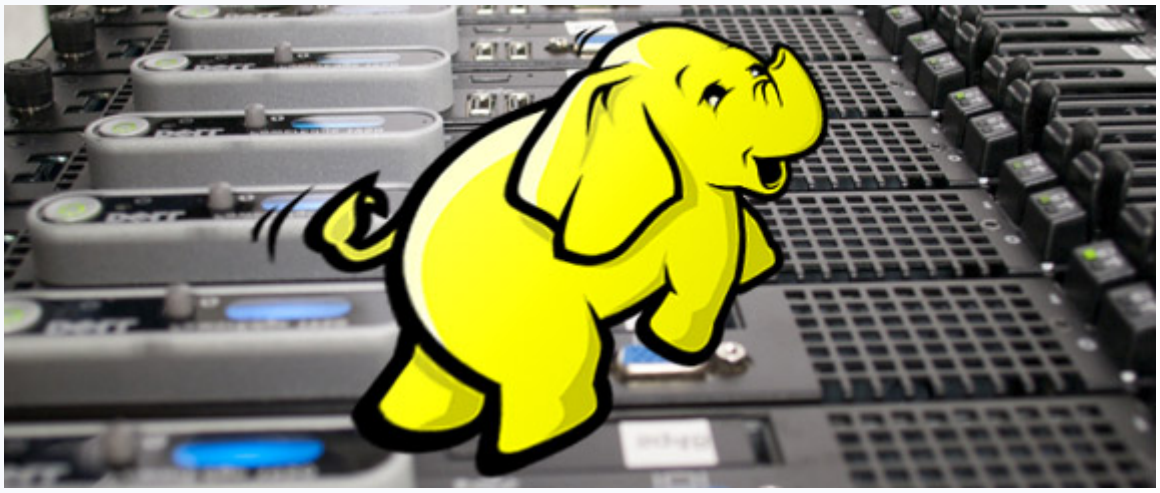
大数据处理三大瓶颈：大容量、多格式和速度

导读：Yahoo CTO Raymie Stata 是领导海量数据分析引擎的关键人物。IBM 和 Hadoop 将更多的精力专注在海量数据上，海量数据正在潜移默化的改变企业和 IT 部门。



越来越多的大企业的数据集以及创建需要的一切技术，包括存储、网络、分析、归档和检索等，这些被认为是海量数据。这些大量信息直接推动了存储、服务器以及安全的发展。同时也是给 IT 部门带来了一系列必须解决的问题。

信息技术研究和分析的公司 Gartner 认为海量数据处理应该是将大量的不同种类以及结构化和非结构化的数据通过网络汇集到处理器和存储设备之中，并伴随着将这些数据转换为企业的商业报告。

海量数据处理的三个主要因素：大容量数据、多格式数据和速度

大容量数据（TB 级、PB 级甚至 EB 级）：人们和机器制造的越来越多的业务数据对 IT 系统带来了更大的挑战，数据的存储和安全以及在未来访问和使用这些数据已成为难点。

多格式数据：海量数据包括了越来越多不同格式的数据，这些不同格式的数据也需要不同的处理方法。从简单的电子邮件、数据日志和信用卡记录，再到仪器收集到的科学研究数据、医疗数据、财务数据以及丰富的媒体数据（包括照片、音乐、视频等）。

速度：速度是指数据从端点移动到处理器和存储的速度。

Kusnetzky 集团的分析师 Dan Kusnetzky 在其博客表示“简单的说，大数据是指允许组织创建、操作和管理的庞大的数据集和存储设施工具”。这是否意味着将来将会出现比 TB 和 PB 更大的数据集吗？供应商给出的回应是“会出现”。

他们也许会说“你需要我们的产品来管理和组织利用大规模的数据，只是想想繁杂大量的维护动态数据集带来的麻烦就使人们头疼“。此外海量数据的另外一个价值是它可以帮助企业在适当的时机作出正确决策。

从历史上看，数据分析软件面对当今的海量数据已显得力不从心，这种局面正在悄然转变。新的海量数据分析引擎已经出现。如 Apache 的 Hadoop、LexisNexis 的 HPCC 系统和 1010data（托管、海量数据分析的平台供应商）的以云计算为基础的分析服务。

101data 的高级副总裁 Tim Negris 表示海量数据的收集以及存放和利用海量数据实际上完全是两回事。在做任何事前需要大量（准备数据）的工作是像 Oracle 和大多数数据库厂商所面临的难题之一。我们正是要消除这个难题，并把数据直接交到分析师的手中。Hadoop 和 HPCC 系统做到了这一点。这三个平台都着眼于海量数据并提供支持。

开源的 Hadoop 已经在过去 5 年之中证明了自己是市场中最成功的数据处理平台。目前 Cloudera 的首席执行官和 Apache 基金会的 Doug Cutting 是 Hadoop 的创始人，他曾在 Yahoo 工作过。

Hadoop 将海量数据分解成较小的更易访问的批量数据并分发到多台服务器来分析（敏捷是一个重要的属性，就像你更容易消化被切成小块的食物）Hadoop 再处理查询。

“Gartner 和 IDC 的分析师认为海量数据的处理速度和处理各种数据的能力都是 Hadoop 吸引人们的地方”。Cloudera 的产品副总裁 Charles Zedlewski 说到。

在 Cutting 和他的 Yahoo 团队提出 Hadoop 项目之后，在 Yahoo IT 系统测试并广泛使用了很多年。随后他们将 Hadoop 发布到开源社区，这使得 Hadoop 逐渐产品化。

在 Cutting 和 Yahoo 在开发、测试并内部运行代码时，他们了解到使用起来还是很复杂的。这导致他们马上意识到如果在未来提供周边服务（例如提供直观的用户界面、定制部署和附加功能软件）可赚取更多的资金。

在 2009 年 Cloudera 作为一家独立公司开始运营，公司产品采用开源并产品化 Hadoop 分析引擎和 Cloudera 企业版（Cloudera Enterprise 整合了更多的工具，包括 Hive、HBase、Sqoop、Oozie、Flume、Avro、Zookeeper、Pig 和 Cloudera Desktop）。

Cloudera 得到了大量投资者的青睐，这其中包括 VMware 的创始人和前首席执行官 Diane Greene、Flickr 的联合创始人 Caterina Fake、MySQL 前首席执行官 Marten Mickos、Linkedln 总裁 Jeff Weiner 和 Facebook CFO Gideon Yu。

自从 Cloudera 成立以来，只有少数的顶级公司和初创公司免费提供他们基于 Hadoop 开放源代码架构制作的自己的版本。

这是一场真正的企业科技的竞争。就像在一场接力赛中，所有选手都必须使用同一种类型的接力棒（Hadoop 的代码）。企业竞争主要集中在处理数据的速度、敏捷性和创造性上。这场竞争是迫使大多数企业在海量数据分析市场有所作为最有效的方法。

IBM 提供了基于 Hadoop 的 InfoSphere BigInsights（IBM InfoSphere BigInsights 是用于分析和虚拟化海量数据的软件和服务，这款新产品由 Apache Hadoop 提供技术支持。）基本版和企业版。但公司有更大的计划。

IBM CEO Sam Palmisano 表示 IBM 正在将新一代数据分析作为公司的研发重点，IBM 在此项目上投资了 1 亿美元。IBM 院士和计算机科学研究室主任 Laura Haas 表示 IBM 实验室的研究远远超出了海量数据的范围，并已经着手”Exadata“分析研究。Watson 就是 IBM 在数据海量数据研究的成果，Watson 将用于更多用途，包括卫生保健、科学研究等。

其他 Hadoop 版本

MapR 发布了一个分布式文件系统和 MapReduce 引擎，MapR 还与存储和安全的领导厂商 EMC 合作向客户提供了 Greenplum HD 企业版 Hadoop 存储组件 。EMC Hadoop 的另一个独特之处在于它没有采用官方版本的 Apache 代码，而是采用 Facebook 的 Hadoop 代码，后者在可扩展性和多站点部署上进行了优化。

另一家厂商 Platform Computing，Platform 提供了与 Apache Hadoop MapReduce 编程模型完全兼容的分布式分析平台，并支持多种分布式文件系统。

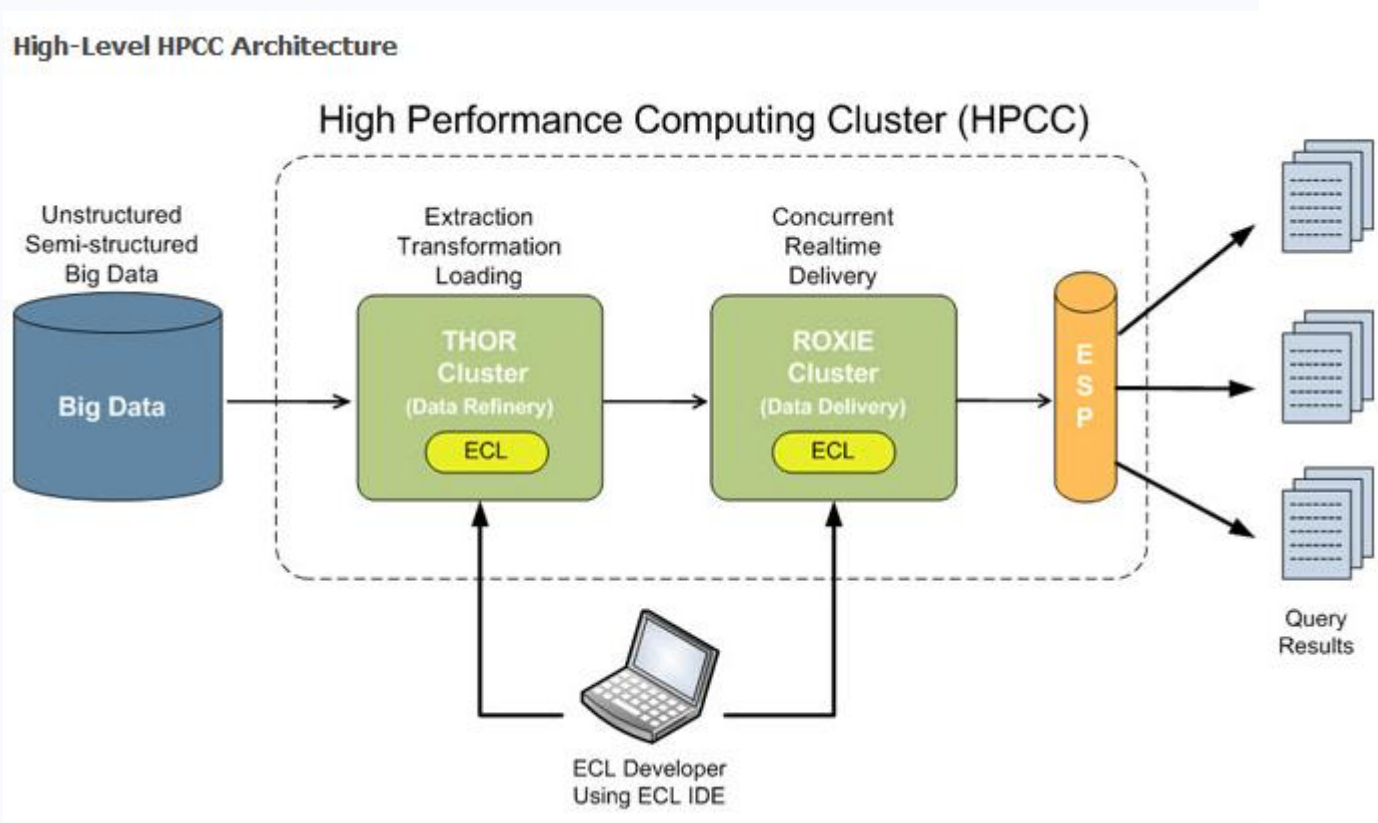
SGI（Silicon Graphics International ）提供基于 SGI Rackable 和 CloudRack 服务器产品实施服务的 Hadoop 优化解决方案。

戴尔也开始出售预装该开源数据处理平台的服务器。 该产品成本随支持选项不同而异，基础配置价格在 11.8 万美元至 12.4 万美元之间，包含为期一年的 Cloudera 支持和更新，6 个 PowerEdge C2100 服务器（2 个管理节点，1 个边缘节点和 3 个从站节点，以及 6 个戴尔 PowerConnect 6248 交换机）。

替代品浮出水面。包括 1010data 的云服务、LexusNexis 公司的 Risk，该系统在 10 年间帮助 LexusNexis 公司分析大量的客户数据，并在金融业和其他重要的行业中应用。LexusNexis 最近还宣布要在开源社区分享其核心技术以替代 Hadoop。LexisNexis 公司发布一款开源的数据处理方案，该技术被称为 HPCC 系统。

HPCC 可以管理、排序并可在几秒钟内分上亿条记录。HPCC 提供两种数据处理和服务的方式——Thor Data Refinery Cluster 和 Roxy Rapid Data Delivery Cluster。Escalante 表示如此命名是因为其能像 Thor(北欧神话中司雷、战争及农业的神)一样解决困难的问题，Thor 主要用来分析和索引大量的 Hadoop 数据。而 Roxy 则更像一个传统的关系型数据库或数据仓库，甚至还可以处理 Web 前端的服务。

LexisNexis CEO James Peck 表示我们认为在当下这样的举动是对的，同时我们相信 HPCC 系统会将海量数据处理提升到更高高度。



在 2011 年 6 月 Yahoo 和硅谷风险投资公司 Benchmark Capital 周二联合宣布，他们将联合成立一家名为 Hortonworks 的新公司，接管被广泛应用的数据分析软件 Hadoop 的开发工作。

据一些前 Yahoo 员工透露，从商业角度来看 Hortonworks 将保持独立运营，并发展其自身的商业版。

在转型时期，Yahoo CTO Raymie Stata 成为关键人物，他将负责公司所有 IT 项目的发展。Stata 表示相对于 Yahoo，在 Hortonworks 我们会投入更多的精力在 Hadoop 的工作和相关技术上，我们认为应加大对 Hadoop 的投资。我们会将一些关键人员指派到 Hortonworks

公司，但这既不是裁员也不是分拆。这是在加大对 Hadoop 的投入。Yahoo 将继续为 Hadoop 的发展做出更大的贡献。

Stata 解释说，Yahoo 一直有一个梦想，就是将 Hadoop 变为大数据分析软件的行业标准。但是这必须将 Hadoop 商业化。Stata 表示创建 Hortonworks 的主要原因是因为 Yahoo 已经看到了未来企业分析（感谢 Hadoop 6 年以来的发展）的未来，并知道该怎样去做。我们看到海量数据分析将很快成为企业非常普遍的需求。

我们将 Hadoop 部署在企业之中，我不认为所有人都否定这样的解决方案。我们要通过 Hadoop 为我们的股东创造价值。如果某一天 Hadoop 成为海量数据处理的行业标准，这将是对我们最好的奖赏。