数据仓库与大数据融合的探讨

于 鹃

(中国水利水电第三工程局有限公司 西安 710016)

摘 要:针对传统企业对非结构化数据缺乏有效利用的问题,探讨了基于数据仓库与大数据融合的企业大数据解决方案。根据数据仓库建设理论和下一代企业级数据仓库架构,利用 Hadoop 技术进行非结构化数据的收集、处理及存储,通过与传统数据仓库和 BI 工具共同协作,解决企业大数据应用困难的问题。数据仓库与大数据技术的融合解决了传统企业数据处理的障碍,从而推动大数据项目的实施。

关键词:大数据;数据仓库;Hadoop;架构

doi: 10.11959/j.issn.1000-0801.2015066

Discussion on Integration of Data Warehouse and Big Data

Yu Juan

(Sinohydro Bureau 3 Co., Ltd., Xi'an 710016, China)

Abstract: In view of the traditional enterprise couldn't use such unstructured data efficiently, a big data solutions based on integration of data warehouse and big data was discussed. Based on the theory of data warehouse construction and architecture of the next-generation EDW, Hadoop was used to collection, processing and storage for unstructured data. By big data technology work together with traditional data warehouse and BI, to help enterprises to solve difficulties that applied. Integration of data warehouse and big data can resolve the difficulty on data processing and promote the implementation of big data projects.

Key words: big data, data warehouse, Hadoop, architecture

1 引言

数据库技术从诞生到现在,已形成了成熟的理论基础、实践方法以及技术产品,并已在此基础上建立了覆盖各个行业、各个领域的各类业务系统,数据库技术是信息技术的重要组成部分,它让人们可以将纷繁复杂的信息按规律进行保存、使用和管理。而随着数据库系统的应用,如何使用和分析已有的数据库又成为一个新问题,在这种背景下,数据仓库应运而生[1]。

很多人以为数据仓库就是"数据库的集合"或者是更大规模的数据库,其实数据仓库是利用已有数据库,对其中的数据进行再一次抽取、加工和使用,并最终用于管理决策,并不是简单的数据复制或数据累加。另一方面,在数据仓库中会使用数据库技术对其中的数据进行管理,因此也有一种看法认为数据仓库是数据库技术的升级。数据仓库与数据库技术息息相关,但又不仅是数据库技术,它是以数据库技术为核心,涉及元数据、数据挖掘、BI 等多技术领域的综合应用。

收稿日期:2014-09-15;**修回日期**:2015-02-27

论文引用格式:于鹃.数据仓库与大数据融合的探讨. 电信科学,2015066

Yu J. Discussion on Integration of Data Warehouse and Big Data. Telecommunications Science, 2015066

在国内大多数企业还在集中精力进行系统整合、数据仓库建设的时候，“大数据”这个名词似乎一夜之间名传天下，其受追捧程度比前几年的云计算有过之而无不及，按照 Gartner 公布的新兴技术炒作周期分析报告显示，大数据在 2013 年已经处于期望膨胀期的顶端[2]，但在 2014 年 8 月公布的报告中，大数据就已进入了幻觉破灭期[3]，Gartner 预计大数据要在 5~10 年才能到达稳定期。大数据虽然在降温，但一个与大数据密切相关的“数据科学”又出现在今年的技术成熟度曲线中，这说明大数据的出现不但加速了信息技术的发展与融合，同时对自然科学与社会科学领域产生了正面的影响。

对大数据的需求主要集中在分析上，即对规模巨大、结构复杂的数据进行管理与处理，以达到预测和决策的目的。从背景和目的来说，大数据和数据仓库很相似，但其处理的数据量、数据类型、处理速度、结果的准确性等都不是现在的数据仓库技术所能比拟的，所以有人预测大数据时代的到来以及相关技术的发展会导致数据仓库的消亡。

2 大数据技术架构及应用困局

大数据为什么会这样火爆，其根本原因在于近几年包括移动应用在内的互联网的快速发展，这些应用产生了比任何时候都多的数据，这些海量的数据包括社交网络、移动设备和传感器等新渠道以及新技术使用所带来的半结构化或非结构化的数据，而想要挖掘利用这些数据并通过预测分析产生价值，传统的数据库运算和处理能力无法实现，在这种情况下大数据技术产生了。以 Hadoop 为代表的大数据技术在互联网企业的成功使用，极大地刺激了业界对大数据的热情，似乎只要是有关大量数据的分析预测都是大数据，在这种情绪下唱衰数据仓库也就可以理解了。

按照科尔尼咨询公司的预测，全球用于大数据的软件、硬件以及服务费用将以近 30%的复合年增长率增长，到 2018 年将达到 1 140 亿美元[4]，而数据指数级的增长也必将改变传统数据存储与分析方法。关于大数据的架构，科尔尼也做了总结，如图 1 所示。

大数据技术架构可分为存储、处理、应用、展示以及整合 5 个部分，并可根据数据的结构化程度对相关技术进行选择和组合。每个部分包含一些技术要素，而某些要素又可根据结构化程度共同作用形成特定的功能，如图 1 中的行业应用、决策支持、并行和分布式处理与存储、报告及可视化以及分析服务。另外，考虑到安全问题，还应加入一个专门的数据安全与隐私部分。

大数据的架构反映出它的复杂性，大数据不是一个单独的产品或技术，而是传统 DBMS（database management system，数据库管理系统）数据库技术与非结构化数据库、

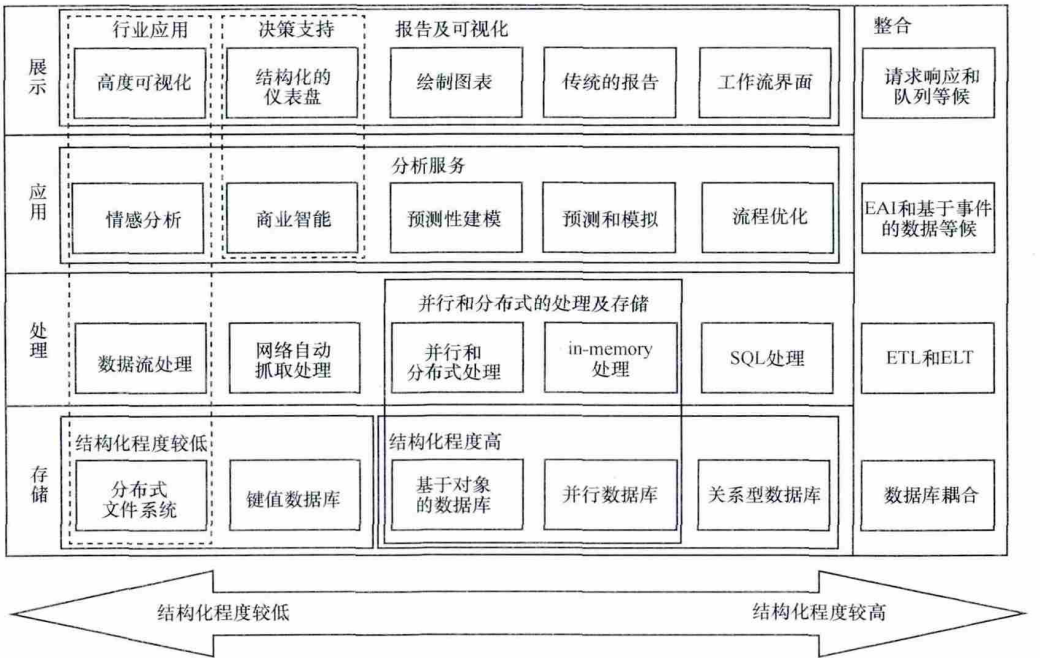

图 1 大数据架构

BI 和数据科学等新技术的集合，这些新技术必将给数据的存储、处理和分析带来根本性的改变，传统企业几乎不可能独立进行大数据项目的建设，这不仅仅是资金投入的问题。在技术领域持续发烧的背景下，对企业来说大数据应用面临的困难如下。

首先，目前关于大数据的话题还主要集中在技术层面，而工程和方法问题并没有解决。也就是说，虽然和大数据有关的技术问题已经基本解决，但如何将技术具体应用到实际企业或组织以及相应的工程学并未解决。特别在国内，虽然个别大型互联网企业有大数据成功的案例，但这些案例和经验无法或者很难复制到传统行业。而且目前为止没有体系化的应用模型，虽然开源的 Hadoop 是免费的，且对硬件要求也并不高，但想要使用这些技术以及维护这类项目，开发和维护成本并不见得更便宜，甚至可能更高，所以大数据目前还只是一个看起来很美的“水中之月”。

其次，大数据技术的优点可能会成为缺点。目前 Hadoop 框架几乎统一了大数据技术的天下，虽然 Hadoop 非常优秀，具有创造性，但同样也具有缺点。例如，它天生就是为了处理海量数据的，对一些相对“少”的结构化数据，反倒不如关系型数据库灵活、性能高，因此不适合处理需要及时响应的任务，且不便于设计，对于一些基础数据相对并不算“大”的企业和组织，如果需要对数据进行分析，目前大数据的解决方案可能就显得大材小用了。

第三，目前大数据技术的安全性缺乏有效的保证。与任何新技术一样，大数据相关的新技术及其伴随而来的安全问题并没有得到有效的重视与解决，人们的关注点主要集中在大数据解决方案，而 Hadoop、MPP 数据库、NoSQL、流处理以及相应基础设施等方面的安全性目前都还没有得以印证，NoSQL 没有经过长期的完善，Hadoop 这种开源框架安全性更是难以保证。除技术安全问题之外，大数据对于个人隐私保护问题也没有明确的监管[5]。大数据技术的安全会逐渐得到完善，但这个过程不会很短。

第四，市场对大数据的应用态度不明朗。与前两年对“云”概念的追捧一样，IT 业界因为通过“云”解决了企业 IT 基础建设难、维护难、浪费大、能耗高等几乎所有难题，一厢情愿地认为云计算的优势必定会马上被企业接受，并很快得以产业化、利润化。但市场反应并非如业界猜测，绝大多数企业出于安全和稳定性等顾虑，根本不接受将业务放到商业性的云服务器上去，虽然后来针对企业应用，也出现了一些诸如企业云的建设方案，但出于成本和技术的原因，并没有呈现爆炸式的发展。

最后，大数据对决策的影响是否能有想象的那么大。大数据产生的一个基础是挖掘海量数据所包含的信息价值，在这个理论基础上数据都有其隐含的价值，所以每一个数据都需要被“加工处理及分析”，因而才有了怎么样处理这些海量数据的技术问题。但这样就产生了一系列疑问：是否真的有必要对每一个数据都进行加工和分析、其准确性怎么验证、领导者是否愿意采信其预测结果、大数据又是否可以解决业务问题。分析及预测是一种技术手段，但未必会影响决策。另外，大数据的目的本是对各类源数据进行统计及分析，但在这个过程中本身就已经又产生了一系列数据，而结果也是一系列数据，这些数据的存储和处理又将产生不菲的成本，因此基于投入和产出的考虑，目前大数据技术的应用环境并不乐观，当然这些怀疑本身并非技术层面的。

3 数据仓库发展趋势及与大数据技术的融合

数据仓库经过多年的发展，已经具备了完整的架构理论、方法及商业化产品，有了诸如 Ralph Kimball 所提倡的项目全生命周期的方法论，技术基础和人才储备也相对完善，并有着大量的行业和企业成功案例。

因此，在大数据还未形成完整应用理论和体系时，DBMS 厂商在传统数据仓库产品功能上，针对大数据分析需求和 Hadoop 进一步融合，加强对列式数据库、数据库内分析、in-memory、数据压缩等技术的研究，以应对更大规模数据的实时分析和处理。根据这种趋势，Forrester 提出了下一代企业级数据仓库的平台架构[6]，如图 2 所示。

在下一代架构中，除了传统的业务数据源之外，加入了来自社交网络、传感器、地理信息等方面的非关系型数据，利用 Hadoop 进行处理。通过可提供云服务的企业级数据仓库平台，结合数据虚拟化整合不同数据源，使用数据压缩技术更有效地管理更大的数据集，以便提供实时或近实时的分析预测。并可利用 in-memory 数据库内分析技术处理更复杂的应用，包括同时进行分析和事务处理。而其中的 DWaaS 代表可以提供多个厂商的数据仓库产品，根据用户需要自动配置，从而提供给企业更经济的部署方式。

在 Forrester 的报告中，特别强调该架构并非单纯的软件架构，而未来的企业级数据仓库供应商应具有更强大的软硬件集成能力，可提供基于硬件的企业级数据仓库的解决方案。从 Forrester 提出的下一代数据仓库平台架构也可看出数据仓库与大数据理念及技术深度融合的发展方向，

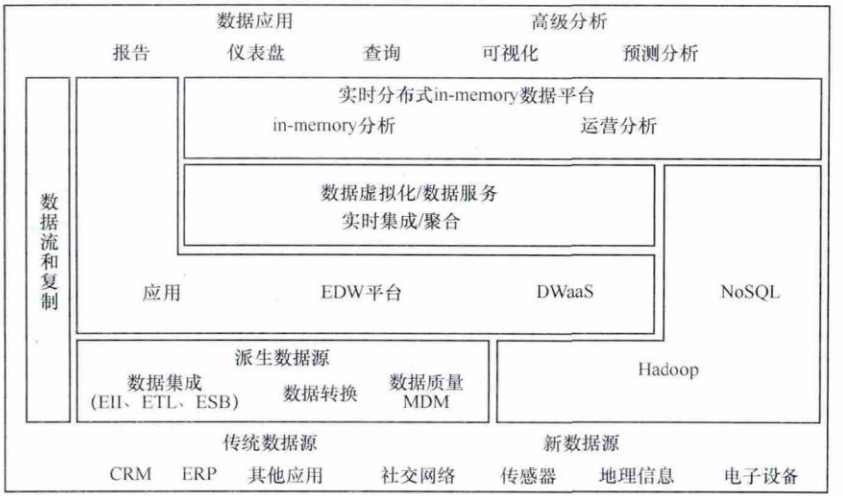

图 2　下一代企业级数据仓库平台架构

结合图 1 的大数据架构来看，这个下一代数据仓库几乎可以说已经是一个大数据方案了。

大数据需求的产生背景与数据仓库类似，人们希望利用新技术处理越来越多的数据、挖掘更大的数据价值。因此，从需求角度来说，无论是数据库、数据仓库还是大数据都是解决不同需求、处理不同级别数据量的技术，它们之间并无冲突，所以短期内并不会出现由谁取代谁的结果，而应该是针对不同需求和现状进行技术选择，各种技术相互补充、相互协作。

目前阶段对于大部分企业来说，想要开展一个全新的大数据项目似乎无从下手。从现有数据仓库建设理论和经验入手，引入部分大数据技术，特别是实现非结构化数据的收集、存储和处理是一种比较可行的方法。例如，将 Hadoop 技术应用于对数据的采集、ETL、存储、处理，开发提供给传统的数据仓库 BI 工具，其架构如图 3 所示。

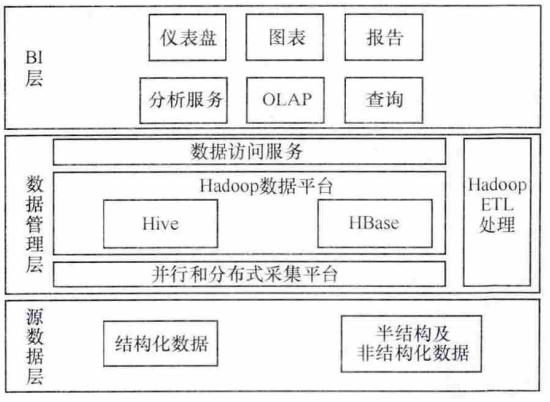

图 3　数据仓库与 Hadoop 技术融合的平台架构

在这个架构中，主要改变了传统数据仓库单节点数据处理和存储的方式，利用了 Hadoop 强大的数据处理能力，将各类数据处理成结构化数据，向上提供给传统 BI 工具，对数据进行分析和结果展示。在这个基础架构之上，可以根据数据处理速度及分析响应能力，逐层进行细化及分解，优化组合 MPP 数据库、内存数据库等各类技术，从而满足 BI 层分析展示的需要[7]。

另外，还可以在数据管理层利用传统数据仓库和 Hadoop 共同合作[8]，由传统数据仓库工具对结构化数据进行处理，由 Hadoop 对更大规模的非结构化数据进行预处理，并将两者处理后的数据存储至结构化数据库中，以便于 BI 层进行分析和展示。

除了技术层面上数据仓库与大数据的融合之外，非常重要的一点是传统数据仓库在具体应用理论、方法和实施上的成功经验，如基于数据驱动的螺旋式开发方法、调研及需求设计、ETL、数据建模、元数据管理[9]等各个方面同样具有很多的可借鉴性与融合性。

有人可能对利用传统数据仓库技术实现企业大数据愿景表示不屑，虽然传统数据仓库厂商并不是当前大数据技术的引领者，但对于大多数企业来说，把有关大数据的赌注放在自行开发和管理上是不可能的，与数据仓库技术融合、与传统厂商进行合作，是目前较为可行的选择。

4　结束语

人们对大数据的追捧反映了市场的需要，除大数据技术的主要领导者谷歌公司之外，传统技术厂商也纷纷通过

收购或者技术融合，推出一些技术方案或数据仓库一体机，以解决企业面临的数据分析问题，但没有哪家能够独立解决大数据问题。本文首先介绍了大数据的技术架构，分析了其应用的困境，然后结合数据仓库和大数据技术的优点，探讨了数据仓库与大数据技术融合的方案，介绍了数据仓库与 Hadoop 技术融合的平台架构，解决企业大数据应用困难的问题，从而推动大数据项目的快速实施。

参考文献

1 陈继东. 数据库发展史. 程序员, 2004(6): 46~50
Chen J D. History of the development of database. Programmer, 2004(6): 46~50

2 Gartner. Hype cycle for emerging technologies. http://www.gartner.com/newsroom/id/2575515, 2013

3 Gartner. Hype cycle for emerging technologies. http://www.gartner.com/newsroom/id/2819918, 2014

4 Kearney A T. Beyond big: the analytically powered organization. http://www.atkearney.com/analytics/featured-article/-/asset_publisher/FNSUwH9BGQyt / content / beyond-big-the-analytically-powered-organization/10192, 2014

5 王倩, 朱宏峰, 刘天华. 大数据安全的现状与发展. 计算机与网络, 2013(16): 66~69
Wang Q, Zhu H F, Liu T H. Current status and development of big data security. Comput & Network, 2013(16): 66~69

6 Yuhanna N, Gualtieri M. The forrester wave: enterprise data warehouse. http://www.forrester.com/pimages/rws/reprints/d ocument/86621/oid/1-M6RP7C, 2013

7 辛晃, 易兴辉, 陈震宇. 基于 Hadoop+MPP 架构的电信运营商网络数据共享平台研究. 电信科学, 2014, 30(4): 135~145
Xin H, Yi X H, Chen Z Y. Design of telecom operators' network data sharing platform Based on Hadoop +MPP architecture. Telecommunications Science, 2014(4): 135~145

8 John Kreisa. Hadoop and the Data Warehouse: When to Use Which.http://hortonworks.com/blog/hadoop-and-the-data-warehouse-when-to-use-which/, 2013

9 沈雷明, 别志铭. 基于电信大数据的数据建模平台研究. 电信科学, 2014, 30(6): 138~141
Shen L M, Bie Z M. Research on data modeling platform based on big data of telecom. Telecommunications Science, 2014, 30(6): 138~141

[作者简介]

于鹃，女，中国水利水电第三工程局有限公司高级工程师，主要从事企业信息化发展与规划、IT 项目设计与实施工作。

(上接 2015083-5)

Meng L L, Chen N N, Wang J, *et al*. A performance evaluation method of power distribution and utilization communication network based on improved TOPSIS. Telecommunications Science, 2014, 30(4): 167~172

5 张高记，吕建东，陈文学. 通信基站节能减排评估指标及评估方法研究. 电信科学，2011, 27(3): 48~51
Zhang G J, Lv J D, Chen W X. Study on energy conservation and emission reduction's evaluation index and methods about communication base station. Telecommunications Science, 2011, 27(3): 48~51

6 李家姿，文涛. 端到端业务质量指标体系建立及评测方法研究.电信科学，2012, 28(11): 21~25
Li J Z, Wen T. Research of end to end service quality indicator system establishing and evaluating method. Telecommunications Science, 2012, 28(11): 21~25

[作者简介]

周静，女，博士，国家电网公司智能电网研究院信息通信研究所主任工程师，现从事国网公司基础性前瞻性新技术预研、电力通信网络仿真与规划、通信技术管理与科技规划等工作。

刘国军，男，硕士，国家电网公司智能电网研究院信息通信研究所项目经理，现从事电力通信安全、网络评估、指标体系建设等工作。