



大数据架构在企业中的应用

王永峰,程新洲,高洁

(中国联合网络通信有限公司网络技术研究院 北京 100048)

摘要:数据是资源已经成为共识,如何获取数据、挖掘数据、把数组资源盘活,成为每个企业都需要深度思考的问题。按照不同层级剖析了大数据架构的关键技术,从企业应用的角度出发,探讨了不同技术的特点与应用范畴。

关键词:大数据;分布式系统;数据挖掘;互联网+;架构

doi: 10.11959/j.issn.1000-0801.2015229

A Brief Probe into Big Data Architecture of Enterprise

Wang Yongfeng, Cheng Xinzhou, Gao Jie

(Network Technology Research Institute, China United Network Communications Co., Ltd., Beijing 100048, China)

Abstract: While data value is deeply appreciated nowadays, enterprises are considering about extracting and digging data value. The architecture of big data was analyzed from the aspect of enterprises. And the characteristic and the application of data were also deeply considered.

Key words: big data, distributed system, data mining, internet+, architecture

1 引言

随着各行业信息化速度的加快,不同类型的数据皆呈现出爆发性增长的趋势。对于这些数据,企业往往用来进行运营、策划、销售等方面的应用,得到不同层面的技术指标,产生系列的报表并反馈到生产和运营中。但是,当这些数据的量增长到一定程度后,量变引起质变,原有的信息化系统和工具渐渐无法承载如此庞大的数据存储和运算,分析效率逐步降低,以至于无法胜任数据分析的需求。在这种处境下,决策者们往往向大数据系统架构靠拢,对原有的信息系统进行改造,去拥抱真正意义的大数据资产。在大数据系统的引入过程中,纷繁复杂的系统实现方法是较为迷惑人的,不同层面、不同技术的选择也会因为需求和目的的不同而影响到实际的决策。因此,深入地了

解大数据系统凸显出了它的意义。

2 企业大数据应用的关键环节

大数据本身是一个较为概括的称谓,并没有对数据的量级做明确的界限。在实际应用中,通常把采用大数据架构进行存储、分析和应用的场景统称为大数据。

大数据在实际的企业应用中,包含多个环节的处理,最终形成监控运行状态、支撑方案决策的大数据应用。在大数据分析的全链条中,比较关键的两个环节是大数据存储和大数据计算。如图1所示,描述了大数据转化的周期,包括数据的采集、解析、入库、存储、分析和应用的各个环节。

大数据采集即数据的初步收集过程。一个公司的数据价值是否明晰,主要从两个方面判断,其一是数据源的获取是否稳定,其二是其数据价值的变现过程是否可持续。

收稿日期:2015-06-19; 修回日期:2015-09-07

论文引用格式:王永峰,程新洲,高洁. 大数据架构在企业中的应用.电信科学, 2015229

Wang Y F, Cheng X Z, Gao J. A brief probe into big data architecture of enterprise. Telecommunications Science, 2015229

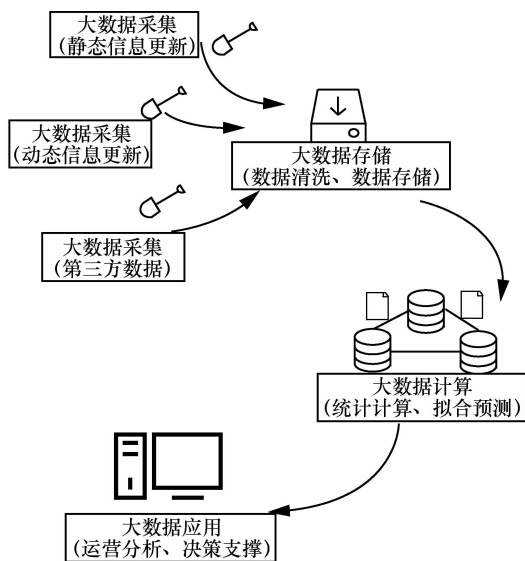


图1 基于自组织的分布式网络管理模型

因此,大数据采集是至关重要的,数据采集决定了整个系统的输入、数据采集的深度和广度,决定了整个大数据分析链条的价值导向。大数据采集包含很多类别,如自身用户数据、系统运营数据产生的自有数据,或者从其他用户庞大用户群体的公司处获取的第三方数据,抑或一些静态的数据信息,如用户身份信息、乡镇街道信息等。

大数据存储是大数据有别于传统数据分析的标志性特征,服务于大数据的存储方式以及相应的工具也蜂拥而至。如何低风险、低成本地建立一套大数据存储体系是企业构建大数据架构考虑的首要问题。从 HDFS (Hadoop distributed file system, Hadoop 分布式文件系统),到接踵而至的 HBase、Hive、Impala 等,纵向细分的差异性的系统和工具给人们更多的选择空间,企业可根据自身大数据的应用场景,从成本、时效性、数据规模等多个方面选择,进而因地制宜地搭建适合本企业数据类型、分析方式的大数据存储架构。

数据存储和数据计算是相辅相成的,存储是为了计算,计算中还需要存储,因此大数据计算不是独立设计的。在大数据计算中,需要考虑多个方面的因素。数理统计是大数据分析中较为常用的分析方法,例如频次统计、分布统计等,是支撑企业运营分析的一项重要的重要途径。在运营分析中,数据运营是重要的组成部分,通过日报表、周报表、月报表等,企业可以提炼出数据中的价值,明确地看出企业运作中的不同趋势,进行用户对比、内容策划、成本控制、风险预警等。除了数理统计以外,更高一层面的、复合型的分析也是

大数据胜任的内容,包括数据的聚类分析、关联分析、回归分析、拟合分析等,通过神经网络、机器学习、迭代等步骤得到更智能、更具有应用价值的计算结果。

诸多的计算步骤,只有一个目的,就是为大数据应用服务。大数据应用就像企业运作中的顾问,用数据说话,给决策者一个参考的依据。大数据应用可以分为前台和后台,前台通过用户友好的界面,分层次成体系的向用户展示不同维度的大数据分析结果,提供定制化分析模板,辅助以不同的应用工具,将后台的运算的数据通过图表、应用、地理化呈现等方式展示大数据计算的结果。

纵观整个大数据链条可以发现,与大数据核心技术密切相关的环节主要在大数据存储和大数据计算两个环节。企业在“互联网+”战略计划中,势必要加大在大数据分析方面的投入,改造原有的大数据仓库,或者建造新的大数据仓库,部署分布式系统架构,搭建大数据平台。而这些投入不是盲目跟风的,需要针对使用的场景以及预算的成本来综合抉择,最终形成企业的大数据方案。

3 典型的架构与差异化策略

大数据典型架构按层级划分,可分为物理层、存储层、计算层、业务层和应用层,如图2所示。层与层之间并没有天然的隔阂和明显的划分标准,只是将逻辑的功能范围作为划分的依据。多层的架构承担着不同的功能,每层都对上层提供可靠的服务,因此形成了一个有机的功能体。

企业在构建大数据架构时,是需要谨慎考虑的,尤其

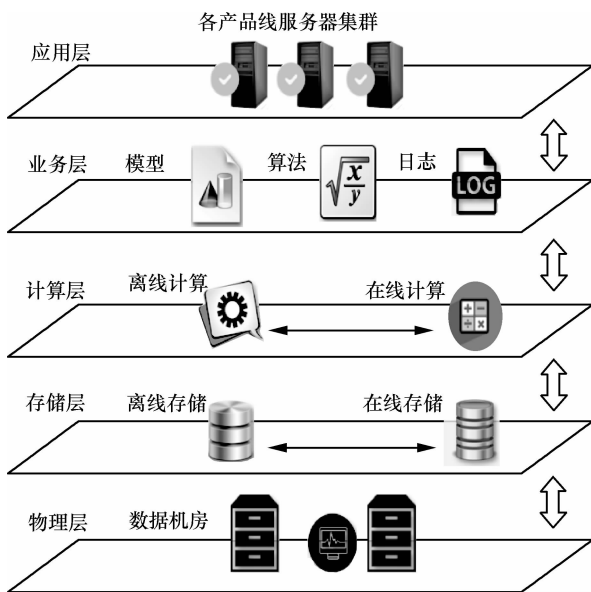


图2 大数据架构层级

是底层的构建很大程度影响了上层的业务、应用性能。这里着重从差异化的策略来阐述几种典型技术对大数据架构的影响。

3.1 硬件策略

3.1.1 数据备份与存储方案选择

存储是首要考虑的问题,作为企业级应用,存储更需要慎重抉择。历史经验证明,数据资产的安全对一个企业是至关重要的,尤其是在如今,数据存储危机给企业带来的影响是致命的,数据可靠性对企业的重要性不言而喻。

硬盘作为存储的重要介质,其类型也是多种多样的。在做选择的时候,可依据不同应用场景进行选择。

(1) SATA 硬盘

对于自身容错备份机制较好的大存储系统,SATA 硬盘是一个很好的选择。SATA 即串口硬盘,现已基本取代了传统的 PATA 硬盘。Intel、APT、Dell、IBM、希捷和迈拓几大厂商组成的“Serial ATA 委员会”正式确立了规范,SATA 硬盘采用串行连接方式,串行 ATA 总线使用嵌入式时钟信号,具备了更强的纠错能力,与以往相比其最大的区别在于能对传输指令进行检查,如果发现错误会自动矫正,这在很大程度上提高了数据传输的可靠性。串行接口还具有结构简单、支持热插拔的优点。

(2) 列阵磁盘

通常,不是所有大数据应用场合都适合建立自身容错备份机制较好的大存储系统,在不具备这样的容错机制的时候,列阵磁盘是一个很好的解决方案。磁盘阵列是由很多价格较便宜的磁盘,组合成一个容量巨大的磁盘组,利用个别磁盘提供数据所产生加成效果提升整个磁盘系统效能。利用这项技术,将数据切割成许多区段,分别存放在各个硬盘上。数组中任意一个硬盘出现故障时,仍可读出数据,在数据重构时,将数据经计算后重新置入新硬盘中。

(3) SAS 硬盘

SAS 硬盘即串行连接 SCSI (small computer system interface,小型计算机系统接口),是新一代的 SCSI 技术,和现在流行的 serial ATA(SATA)硬盘相同,都是采用串行技术以获得更高的传输速度,并通过缩短连结线来改善内部空间等。SAS 硬盘是并行 SCSI 之后开发出的全新接口,此接口的设计是为了改善存储系统的效能、可用性和扩充性,并提供与 SATA 硬盘的兼容性。

因此,在大数据架构搭建的底层,根据自身场景的需求,做出相应的判断,进而建立或升级存储结构是十分必要的。

3.1.2 计算性能的需求与硬件的匹配

虽然计算机硬件的发展速度十分迅猛,但是仍然不能小觑日益增长的计算需求。每当硬件水平上一个新的台阶,计算的需求会迅速地填补上来,因为算法的进步是永无止境的。

随着语音识别、图像识别等模式识别的发展以及迭代算法需求的增加,大数据平台的计算数据量是异常庞大的。因此,当计算性能成为瓶颈的时候,相应的提升工作就成为必要考虑的内容。

(1) 处理器加速

随着非结构化数据的爆发性增长,计算量异常巨大,需要依靠 GPU 加速或者重核卡的加速才能在可容忍的时间内完成计算,不少企业的大数据集都采用了 GPU 加速或重核卡。

(2) SSD

某些应用场景下,如快速的迭代算法会有频繁的读取与存储操作,这使得传统的 I/O 暴露出速度慢的弊病,严重影响了运算的速率。SSD(solid state drives,固态硬盘)有传统机械硬盘不具备的快速读写、质量轻、能耗低以及体积小等特点。随着目前 SSD 价格的不断下滑,SSD 逐渐成为一种十分有吸引力的选择。

(3) 虚拟机

服务器的资源常常由于得不到有效的复用,导致资源无法发挥最大的性能或者资源部分闲置。虚拟机的引入解决了这个问题,它能够精细化地划分和管理服务器的资源,实现合理地配置、充分地复用,从而整体提升计算的性能。虚拟机技术正逐步成为一种趋势,目前很多企业级的大数据平台都选择搭建在虚拟机集群上。

以上只做了一些常用的技术的描述,实际需根据企业构建大数据平台的真实情况,选择合适的技术方案来提升系统性能。

3.2 云存储策略

在选定了存储的硬件介质之后,接下来就是从软件层面上定义存储的方式。按照时效性划分,云存储可分为离线存储和在线存储两种类型,如图 3 所示。选择何种存储

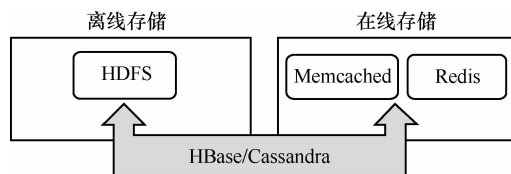


图 3 云存储的两种方式



类型,对于企业整体大数据架构而言十分重要。选择存储类型一方面要满足应用时效性的需求,另一方面要和存储介质以及算法匹配。

离线存储一般适用于超大规模的数据量,从时间上来说一般是长时间的存储,往往应用于对时效性要求不高的场景。例如对于企业应用中一些海量的过程类数据,不需要频繁访问的,即可采用离线存储方式。这种方式最典型的案例就是 HDFS,它的部署成本和技术门槛相对较低,因此得以普及。

在线存储顾名思义需要频繁的互动,及时性要求较高,同时又具备大数据的海量存储特征。最常用的 Memcached 是一个高性能的分布式内存对象缓存系统,常用于动态 Web 应用以减轻数据库负载。它通过在内存中缓存数据和对象来减少读取数据库的次数,从而提高动态数据库驱动网站的速度。其本质上是一套分布式的快取系统,但是不提供持久化。与之类似的 Redis 是一个开源的使用 ANSI C 语言编写、支持网络、可基于内存亦可持久化的日志型 key-value 数据库,能够提供持久化的能力。

3.3 计算层策略

计算层也有离线和在线之分,其实与存储层区分是类似的,只是表现的技术不同而已,如图 4 所示。在分布式系统架构中,不同的任务之间是需要进行数据传递的,一部分是采用存储来传递,即存储和读取方式实现,另一部分采用数据管道系统来完成。

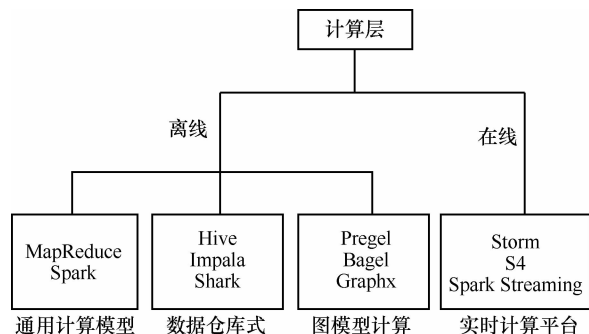


图 4 计算层的不同计算方式

对于数据量庞大、运算耗时长任务,通常会采用离线计算的方式进行,特别是对于那些需要深度发掘、多次迭代的算法而言。离线计算以 MapReduce 为代表,MapReduce 的设计,采用了很简化的计算模型,只有 map 和 reduce 两个计算过程,中间用 shuffle 串联。用这个模型,可以处理大数据领域很大一部分问题。第二代的 Tez 和 Spark,除了内存缓存之类的新特性以外,让 MapReduce 模

型更通用,让 map 和 reduce 之间的界限更模糊,数据交换更灵活,更少的磁盘读写,以便更方便地描述复杂算法,取得更高的吞吐量。

某些应用场景下需要较短的时延,需要实时性较高,比如智慧城市中的路况监控,这是离线计算无法胜任的。因此,流计算应运而生了,Storm 是最流行的流计算平台。流计算的思路是,如果要达到更实时的更新,则在数据流读取的时候就直接进行处理。流计算虽然快速,但它不灵活、统计的内容必须预先知道,因此虽然功能强大,但是无法替代数据仓库和批处理系统。

在企业构建大数据架构的过程中,多种计算方式都是要复合使用的,通过扬长避短达到发挥大数据平台最高性能的目的。

3.4 业务层策略

构建大数据平台之前,必须要做的工作是清楚大数据平台应用的目的、对象以及算法。否则软硬件设备购置部署完成后,终究是骨架,而没有血肉。

业务层承载着大数据的核心思路,包括核心的算法、基础的分析模块等。一般来说,这些业务应该包括日志处理、离线分析、深度挖掘、分类聚类 and 预测建模等离线业务,也有实时挖掘、实时监控等实时处理业务。大数据的核心就是预测,它通常被视为人工智能的一部分,或者更确切地说,被视为一种机器学习。大数据大大解放了人们的分析能力,一是可以分析更多的数据,甚至相关的所有数据,而不再依赖于随机抽样;二是研究数据如此之多,以至于人们不再热衷于追求精确度;三是不必拘泥于对因果关系的探究,而可以在相关关系中发现大数据的潜在价值。因此,当人们可以放弃寻找因果关系的传统偏好,开始挖掘相关关系的好处时,一个用数据预测的时代才会到来。

业务层可以实现丰富的数学方法、挖掘算法,常用的介绍如下。

(1) 聚类算法

企业级应用中的大数据分析经常用到聚类算法,比如针对某些特征对用户群体进行划分,如按照用户标签对预测其偏好类别,淘宝商铺将用户在一段时间内的购买情况划分成不同的类。聚类方法有许多种,例如基于球邻域的空间划分、仿生模式识别、视觉分类方法等。

(2) 回归分析

预测对于企业的意义比较深远,回归分析是进行大数据预测的有效方法。回归分析用函数表达式形式,反映

了值与值、属性与属性之间的相互关系。当人们最大似然地获得了其中的关系,就可以用其中一部分数据来预测另一部分数据。在市场营销的很多方面都可以运用到回归分析,例如可以通过回归分析对当月的销售规律进行挖掘,从而预测出下个月的发展趋势,进而做相应的策略上的变更或保障。

(3)神经网络

神经网络是一种人工智能技术,其优势在于处理非线性以及那些以模糊、不完整、不严密的知识或数据为特征的问题。随着数据的增长与发展,企业中非结构化数据呈爆发增长趋势,而神经网络在分析这类数据上有着天然的优势。神经网络具有分布存储和高度容错等特性,与大数据架构不谋而合,因此部署神经网络算法来解决数据挖掘的问题成为企业很好的选择。典型的神经网络为三大类:第一类是前馈式神经网络模型,其特征是分类预测和模式识别;第二类是反馈式神经网络模型,擅长联想记忆和优化算法;第三类是自组织映射方法,在聚类方面有较好的应用。

业务层的方法是十分丰富的,企业在构建自己的大数据平台时,不应局限于个别几类方法,而是需要充分挖掘和汇聚不同方法的潜能,使得统计和预测的结论以更大的概率接近实施情况,从而真正体现企业的数据价值。

4 结束语

数据是资源已经成为共识,如何获取数据、挖掘数据、盘活数组资源,成为每个企业都需要深度思考的问题。

在反馈经济中,数据是一面镜子,企业通过数据了解客户,了解自身,洞悉整个行业乃至整个产业生态环境,对未来发展有更充分的把握。工欲善其事,必先利其器。做好企业的运营,大数据平台是利器。随着大数据各领域技术的日臻成熟,选择适合自身发展的技术是要解决的当务之急。存储、计算、业务的策略是使用者需要考虑的重中之重,是大数据系统得以协调运作的基础。好马配好鞍,满足需求的硬件配置,也是不容忽视的刚性条件。相信只要把这些关键环节的工作做到位,大数据给企业带来的可预见的增益就近在咫尺了。

参考文献

- 张宁,贾自艳,史忠植. 数据仓库中的 ETL 技术的研究. 计算机工程与应用, 2002, 38(24): 200~211
Zhang N, Jia Z Y, Shi Z Z. Research on technology of ETL in data warehouse. Computer Engineering and Applications, 2002, 38(24): 200~211
- 董翔英. SQL Server 基础教程. 北京: 科学出版社, 2005
Dong X Y. Beginning SQL Server for Developers. Beijing: Science Press, 2005
- 陶冶, 范玉顺, 罗海滨. 分布式 workflow 系统的可靠性研究. 计算机科学, 2001, 28(5): 6~10
Tao Y, Fan Y S, Luo H B. Research on the reliability of distributed workflow system. Computer Science, 2001, 28(5): 6~10
- Baru C, Bhandarkar M, Nambiar R, et al. Big data benchmarking. Proceedings of the 2012 Workshop on Management of Big Data System of ACM, San Jose, CAL, USA, 2012: 39~40
- Rogers Y, Sharp H, Preece J. Interaction Design: Beyond Human-Computer Interaction. West Sussex: Wiley, 2011
- 罗军舟, 金嘉晖, 宋爱波等. 云计算: 体系架构与关键技术. 通信学报, 2011, 32(7)
Luo J Z, Jin J H, Song A B, et al. Cloud computing: architecture and key technologies. Journal on Communications, 2011, 32(7)
- 王劲. 大数据时代的管理变革. 中国商贸, 2013(2): 189~190
Wang J. Management revolution in big data era. China Business & Trade, 2013(2): 189~190
- 崔杰, 李陶深, 兰红星. 基于 Hadoop 的海量数据存储平台设计与开发. 计算机研究与发展, 2012(49)
Cui J, Li T S, Lan H X. Design and development of the mass data storage platform based on Hadoop. Journal of Computer Research and Development, 2012(49)
- 张辉, 赵郁亮, 徐江等. 基于 Oracle 数据库海量数据的查询优化研究. 计算机技术与发展, 2012, 22(2)
Zhang H, Zhao Y L, Xu J, et al. Query optimization research on mass of data based on Oracle database. Journal of Computer Research and Development, 2012, 22(2)
- 李成华, 张新访, 金海. MapReduce: 新型的分布式并行计算变成模型. 计算机工程与科学, 2011, 33(3)
Li C H, Zhang X F, Jin H. MapReduce: a new programming model for distributed parallel computing. Computer Engineering and Science, 2011, 33(3)

[作者简介]



王永峰, 男, 中国联合网络通信有限公司网络技术研究院工程师, 主要研究方向为通信大数据挖掘及应用。

程新洲, 男, 中国联合网络通信有限公司网络技术研究院高级工程师, 主要研究方向为通信大数据挖掘及应用。

高洁, 男, 中国联合网络通信有限公司网络技术研究院工程师, 主要研究方向为通信大数据挖掘及应用。