大数据时代勘探云建设模式探索①

林 茂, 塔依尔·伊布拉音, 许 涛, 宋雪峰, 姜元刚

(新疆油田勘探开发研究院 地球物理研究所, 乌鲁木齐 830013)

摘 要: 按照大数据定义, 油气勘探数据处理工作显然是一种大数据应用模式, 而作为油气勘探核心工作平台的勘探云计算中心, 其建设目的首先是为了满足企业内部业务数据处理工作需求, 在确保计算力分享和协同工作的基础上, 勘探云更注重数据存储、数据管理以及数据应用业务的支持, 对照一般公用云建设模式, 勘探私有云建设具备其独特的建设模式. 笔者企业结合自身特点, 对照当前大数据应用需求开展勘探私有云建设工作, 取得了一定的效果, 对企业主营业务的支持效果明显.

关键词: 计算机系统结构; 云计算; 大数据; 数据管理; 作业调度

Construction Model of Cloud Center for Oil-Gas Exploration in Big-Data Era

LIN Mao, Tayir IBRAHIM, XU Tao, SONG Xue-Feng, JIANG Yuan-Gang

(Research Institute of Exploration and Development of Xinjiang Oilfield, Urumqi 830013, China)

Abstract: According to big data definition, the data processing of exploration of oil and gas is obviously a big data application mode. Computing center is designed to meet data processing needs of the company as the main exploration platform of oil and gas exploration cloud. Under the primise of ensuring the sharing computing power and Collaborative work,exploration cloud focus on support of data storage, data management and data applications. Compared with the general public cloud, the Construction of private exploration cloud has its unique mode. The author company combined with own company characteristics and current big data application needs to construct large data private cloud, and it achieved certain results. The support for the main business is obviously effect.

Key words: computer architecture; cloud computing; big data; data management; job scheduling

云计算概念在业界已经流行很长一段时间了, 各地各企业纷纷根据需要建设各自的云计算产业, 一批云计算中心相继投入运行应在实际工作中发挥作用. 然而随着“大数据”概念的提出, 过去几年中以“计算力”为核心的云计算中心建设模式发生了一些技术上的变化. “所谓地震勘探, 就是通过人工方法激发地震波, 研究地震波在地层中传播的情况, 以查明地下的地质构造, 为寻找油气田或其他勘探目的服务的一种物探方法. 地震勘探资料处理过程是利用先进的计算机数据处理能力对野外收集的原始资料进行各种去粗取精、去伪存真的加工[1].”从本质上说油气勘探工作就是一个典型的数据处理、数据应用的过程. 正是由于这个工作性质的确认, 作为油气勘探核心工作平台的勘探云计算中心,在建设之初其实现目标和建设模式就与一般公用云有所差别. 勘探云的定位是一朵企业自建私有云, 因此它的建设除了是为实现“计算能力”分项和异地多学科协同工作外, 更侧重于企业主业数据(地震勘探数据)的管理、存储和应用. 这一点恰巧符合了目前“大数据”概念对云计算中心的要求.

1 勘探云建设需求分析

业界对“大数据”(Big data)的定义可以描述为: “狭义的大数据概念, 主要指大数据技术及其应用, 是指从各种各样类型的数据中, 快速获得有价值信息的能力. 大数据一方面反映的是规模大到无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集合; 另一方面, 主要是指海量数据的获取、存储、管理、分析、挖掘与运用的全新技术体系. 事实

① 收稿时间:2014-09-01;收到修改稿时间:2014-10-16

上，大数据的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理. 换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”[2].

油气勘探工作整个过程恰巧是产生数据、管理数据、处理数据、数据模拟显示、数据分析，通过这一系列工作后形成对某一个地质区块的地层内是否存在油气资源的认识，指导企业找到油气，创造企业价值. 而伴随着油气勘探技术的持续进步，整个工作环节中涉及到的数据种类越来越多、数据数量越来越大，处理时间周期越来越紧而处理方法越来越先进，这个处理过程正在不断地向大数据应用模式转换. 而随着这几年高密度宽方位采集技术的不断应用，笔者单位数据处理量从 2010 年每年处理的原始数据量大约在 25TB(中间过程所需要涉及的数据量大约为 250TB)跃升到 2013 的 150TB(中间过程数据量大约为 1.5PB)，6 倍的数据增量说明了大数据应用模式已经来临. 而另一个指标即计算机规模从过去 4TFLOPS 提升 150TFLOPS 的同时集群利用率从过去 30%左右上升到 80%也说明了这些数据处理过程更加复杂、繁琐. 与此同时寻找油气的任务又要求每一个处理任务花费时间越来越少，勘探数据处理方法亟待适应“大数据应用”模式. 作为油气勘探核心工作平台的勘探云，显然要适应这种工作的变化. 与公用云关注“计算力”建设以及用户管理不同，勘探云建设在保证计算资源共享和透明化的前提下更加注重以下几点:

1)确保油气勘探大块数据集中管理、整理、检索以及加载;

2)能够提供 PB 级别的存储能力，同时对多节点大块数据读写能力要求很高.

3)确保存储体系的延续性，确保企业核心资产--数据持续稳定的存储保障.

4)提供手段使得专业用户透明使用计算资源和存储资源，不将精力花费在信息系统架构上.

5)提供手段将数据成果有形化显示，使数据具有专业意义.

6)确保勘探应用系统在勘探云中的部署以及日常运行维护、资源配置.

可以看出，勘探云建设目标从其本身转变为作为勘探“大数据”应用的支持平台提供服务. 为适应油气勘探工作发展，信息人员有必要提供一种手段，使得勘探科研人员不再需要关心勘探数据的正确和完整，不再关心这些数据究竟存放在那个位置；不关心使用那种技术确保其安全；更无须关心使用那个计算平台进行数据处理，将工作精力重点投入处理方法的研究和实现，以及数据成果的分析研究工作当中. 从现金技术发展趋势看，信息人员提供的这种手段就是勘探云计算中心. 它是一个企业自建私有云，是综合应用各种信息技术集成而成的一个勘探应用平台. 勘探云存在的目的就是为应用人员提供一种近乎透明的便捷手段去把保存、处理、分析勘探数据，将他们的工作精力更多投放到到如何最快、最准确地处理越来越多、越来越大的勘探数据，寻找数据内在关系和存在规律，从中找到油气资源存在的蛛丝马迹，指引企业工作业绩和效益的获得.

2 勘探云建设中的几点认识

为实现企业盈利目标，笔者企业近几年致力于建设内部勘探云. 根据油气勘探实际工作需要以及发展状况，在云计算中心 SaaS 服务模式建设的同时，重点做好海量地震勘探数据管理、存储、处理和应用等工作的云端实现，在建设过程中取得了一些经验教训和方法认识. 具体包括以下几个方面:

2.1 数据管理中心与计算中心同步发展

以往云计算中心建设通常强调计算能力建设以及计算能力共享的建设. 云计算很容易被理解为用户从某个计算中心租用一部分计算资源，将自己所需的软件安装在这些资源上，或直接租借计算中心已经安装好的应用软件. 数据的组织和加载则需要用户自己进行，云计算中心最多只提供这些数据的存储介质. 油气勘探工作中需要使用到的数据种类多、数据大、处理过程复杂、耗时、对数据读写要求高、不适合关系型数据库管理. 尤其是伴随着勘探数据采集方法的不断进步，野外采集的原始数据越来越大，笔者单位 2013 年处理的一个数据体原始数据达到 65TB,这样一个数据体从高速并行存储当中加载到应用系统里就需要花费 2 天的拷贝时间，这种加载如果是通过局域网络进行显然所需要的时间是无法忍受的，与此同时产生的数据传输将对整个网络系统造成灾难性的影响. 这样大规模的数据加载如果需要科研用户自行完成显然不太合适. 勘探数据体的正确性、多种数据的组合、

分类、筛选、预处理都需要一些专业数据管理人员使用专业工具完成.

根据这一需求特点, 勘探云建设过程中必须:

1)在计算中心建设的同时同步发展数据中心, 在现有网络互联条件下, 数据中心和计算中心的物理位置最好建设在一个地方或一个局域网络内.

2)数据管理工作必须是信息人员和专业人员共同完成, 离开专业人员的指导, 数据管理工作没有意义.

3)数据管理工作需建立一部分数据筛选、预处理、质控、组合等手段和工作平台, 在确保数据“原样”保存的基础上为科研人员提供尽可能规范、简单、符合需要的数据.

4)数据组织形式是否使用数据库管理数据必须由数据本身和工作性质确认.

笔者企业正是这样做的, 数据管理中心建设工作始终由企业内部专业人员与数据管理人员共同开展工作, 通过勘探数据质控、预处理、筛选、捕捉关键中间成果数据、整理组合等多环节工作利用和管理油气勘探各种数据, 形成不同管理形式的八大基础库(如图 1 所示).

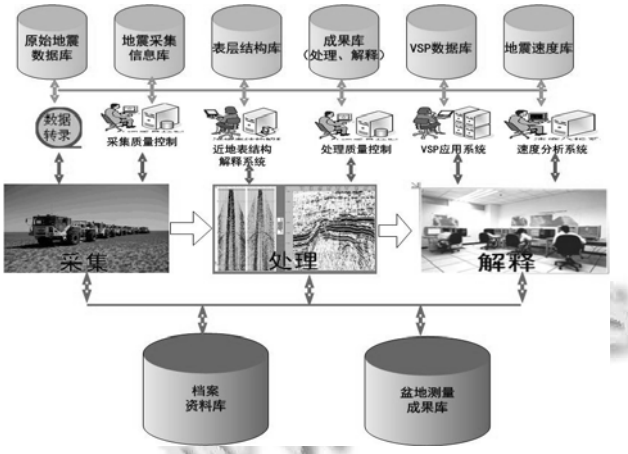

图 1　勘探基础库及其应用模式

数据管理中心的建设使得数据人员与计算机系统紧密结合, 对勘探企业日常使用的各类数据进行分门别类的筛选, 规范, 建立检索查询机制并按科研人员需求提供和加载数据. 这种机制是在暂时无法实现全面自动化的情况下先通过数据管理人员的手工劳动建立一套完整的数据管理体系, 对勘探大数据进行规范化管理和一定程度的预处理. 通过这种模式科研人员可以直接“透明”使用预处理、质控、筛选后的数据开展研究工作, 有效提升工作效率.

2.2 按需定制高可用存储体系

勘探数据是大块数据, 数据量通常是以 TB 作为计数单位的. 这种大数据应用要求勘探云计算中心的数据存储设备容量是 PB 级的. 而以 TB 作为单位的项目数据进行计算是一个复杂、耗时的过程, 对存储设备读写速度的要求非常苛刻. 通过对勘探数据处理作业计算过程进行统计分析后, 信息人员认为符合油气勘探数据处理工作要求的存储系统必须达到表 1 所指定的一些关键指标.

表 1　勘探处理用存储系统关键指标

参数	目标值	评估指标的计算依据
总的 IOPS	300 万次/秒	3 万次/秒(单个计算节点 I/O 请求数)*100(平均并发节点数)
存储空间	600TB	笔者单位处理中心年平均数据处理新增工作量
总读/写速度	6GB/5GB	10Gb/8(I/O 节点网络带宽)*80%(带宽平均利用率)*6(I/O 节点个数)
单节点最高速度	650MB	1000Mb(计算节点网络带宽)/8*50%(带宽平均利用率)
存储架构	并行存储	并行存储架构解决 PC 集群存储的特殊需要

从上表中可以看出, 勘探云所需要配备的存储设备必须是并行存储系统, 单节点读写速度为 650MB/S, 多节点同时访问存储时, 存储系统应该能够达到 5-6GB/S 的总读写带宽. 通过测试、比较目前市场上最常用的存储设备, 能够满足以上所有指标的品牌产品寥寥无几. 而就是这种能够满足指标的存储设备在引进时还存在引进费用高、扩充困难、技术不开放、维保困难、管理分散等这样或那样的问题, 很难充分满足勘探云建设的需求.

为提升勘探云建设步伐, 提升勘探大数据应用的支持, 笔者企业在勘探云建设过程中提出“存储现场定制”工作模式. 这种模式就是首先根据用户现场工作模式制定存储设备需求关键指标, 选择一个合适的存储集成厂商, 利用一些通用计算机设备和专业存储设备, 选择合适的存储管理软件, 按照并行存储系统架构定制统一的并行存储体系, 通过反复验证和实际工作考验, 配比存储系统内部各组成要件的不同比例关系, 从而集成出符合实际工作所需关键参数的存储系统. 图 2 所示的存储系统就是笔者企业利用不同厂商生产的通用存储设备, 选用合适的存储管理系统, 确定每个部件的配比关系后形成的一套并行存储系统.

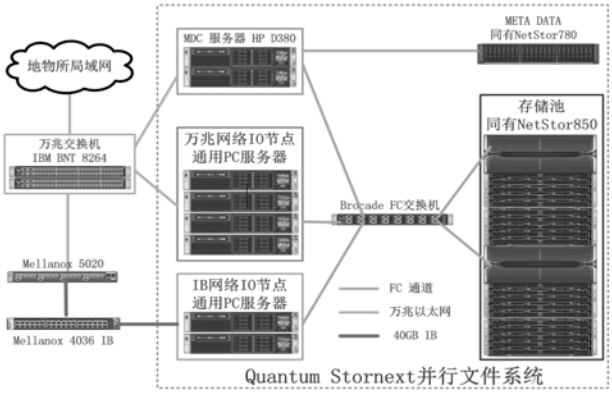

图 2 定制并行存储系统架构图

2.3 作业调度系统的研发

云计算最关键的就是使计算资源透明化，公用云通常采用虚拟化的手段将一个服务器分解成多台服务器供给不同的需求进行使用. 而作为油气勘探工作对计算能力的要求显然不是通常意义的网站应用完全不同，与之对应的经常是一个多节点 PC 集群或一组服务器群. 勘探云要实现的就是如何将数以千计的集群节点或几十台服务器群组合成为一个整体，使勘探科研人员面对的是一个海量计算单元为一体的应用平台. 实现这种“多对一”的服务器云化最有效的手段就是队列管理和作业调度. 这种工作模式首先是将服务器节点按照队列进行管理，将一个大的服务器群或集群形成一个或多个不同节点数量配置的队列. 科研人员只需要将自己需要计算和应用的过程编织成一个作业发送到某一个队列当中，随后寻找合适的计算资源进行计算的工作就全部交给作业调度系统完成. 这种模式中科研人员无需知道完成其计算请求的计算机设备是哪一个，其工作状态如何，何时以什么方式完成提交的作业. 再他面前显示的就是一个个待命的服务器队列，从而使得集群或服务器群云化成一个整体.

通常集群管理中会使用到作业调度体系，而笔者企业通过研究发现勘探中解释服务器群负载均衡同样也可以使用作业调度体系完成. 为降低集群和服务器群云化管理费用，笔者单位采用开源的 PBS 昨夜调度系统作为管理平台，通过定制开发符合科研人员工作习惯的作业脚本实现了处理集群以及解释服务器群的队列管理和作业调度. 下图 3 中显示的就是该系统实现的用户交互窗口的多节点负载均衡.

在下图 3 所显示的工作中，用户在登陆节点(hc03n15b)上根据它的使用习惯键入“focus &”命令启

图 3 Epos 交互作业的封装效果图

动处理交互窗口. 通过作业命令封装，该请求被改造为 PBS 作业脚本(作业号 5897)发送到 PBS_Server 上，通过 PBS 调度发现交互队列中 b50n04 节点比较空闲，因此将该请求发送到 b50n04 节点上打开交互窗口交付给用户使用. PBS 作业调度体系在这一过程中起到了很好的负载均衡和调度管理功能，而整个过程全部由计算机系统自动完成，用户无需进行任何干预[3].

2.4 远程 3D 显示提供应用平台

数据应用离不开数据可视化展示，只有将大数据转换成一种方便科研人员理解应用的显示形式才能够有助科研人员理解和研究分析这些数据，油气勘探工作当然也离不开这种大数据应用模式. 油气勘探应用系统的一个关键功能就是将各种勘探数据模拟转换成不同的地质模型和虚拟图像显示给勘探科研人员进行研究.

勘探工作中越来越强调的多学科、多人员协同研究模式也要求研究项目需要集中部署在一定范围的应用服务器上共享给所有参与人员共同使用. 种种要求都要求勘探云应用模式能够使科研人员工作平台更简单化，将所有计算能力都后移到后台高端服务器上进行，3D 图形显示功能当然也需要集中采用远程虚拟服务器形式实现.

为实现远程共享 3D 图像处理能力，勘探云需要建设 GPU 专用服务模式. 通过比较目前市场上现有的一些远程 3D 显示解决方案，笔者企业在勘探云建设过程中选择虚拟显卡+虚拟机+DCV/EOD 显示软件的建设模式，在勘探云中配置了专用 GPU 服务器，利用 GPU 服务器中配置的 Nvidia K2 显卡虚拟化技术配合虚拟机技术将 GPU 服务器虚拟成不同配置的多套 GPU 虚拟机. 在这些虚拟机上安装 DCV 远程虚拟显示软件或 EOD 显示软件. 其中 DCV 工作原理如图 4

所示:

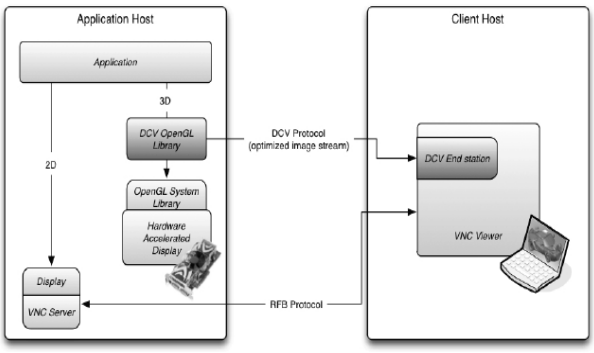

图 4　DCV 远程 3D 图形处理原理

正如上面原理图所示的，科研人员使用办公桌面系统开展勘探处理解释工作，当需要进行 3D 图像显示时，通过调度机制，将勘探云中某个配备有 GPU 虚拟图卡的 GPU 虚拟服务器配属给该任务，利用 GPU 显卡处理图像数据，生成 3D 图像后以图形压缩包的形式传送到科研人员桌面系统中. 桌面系统内嵌 DCV 客户端对图形包进行解压显示，将其中的 3D 图像展示给科研人员.

远程 3D 显示整个过程对科研人员桌面系统没有特殊要求，所有图形处理工作都交由后台 GPU 服务器完成，图形处理过程对科研人员而言全程透明. 这一过程反映了勘探云对科研人员大数据应用有效、透明的应用支撑.

3　勘探云建设成果

通过一段时间的摸索和研究，笔者所在企业建设了一套具备一定规模的勘探云计算平台雏形，初步实现了油气勘探工作异地资源共享和协同应用功能. 该勘探云计算平台由一套主要用于勘探数据处理的 7152CPU 核+64512GPU 核异构集群、主要进行地层显示以及地质解释的 32 套服务器组成的应用服务器群、负责高端三维图形处理和渲染的 7 套 GPU 图形服务器以及集中保存、管理勘探数据的总量 1.6PB 高性能并行存储系统构成. 所有集群节点和应用服务器安装部署了许可统一管理的主流油气勘探软件平台，由 PBS 作业调度系统统一调度、管理. 勘探科研人员在办公室内使用未安装专业应用软件的普通的桌面计算机系统，通过系统中内嵌浏览器或远程终端软件“近乎透明”地使用部署在核心机房当中的勘探云数据资源和计算资源，完成油气研究工作. 油气勘探数据加载、筛选、预处理以及备份、更新、转储等工作有专业数据管理人员利用自建勘探数据管理系统进行同意管理. 整套勘探云的使用过程如下图 5 所示.

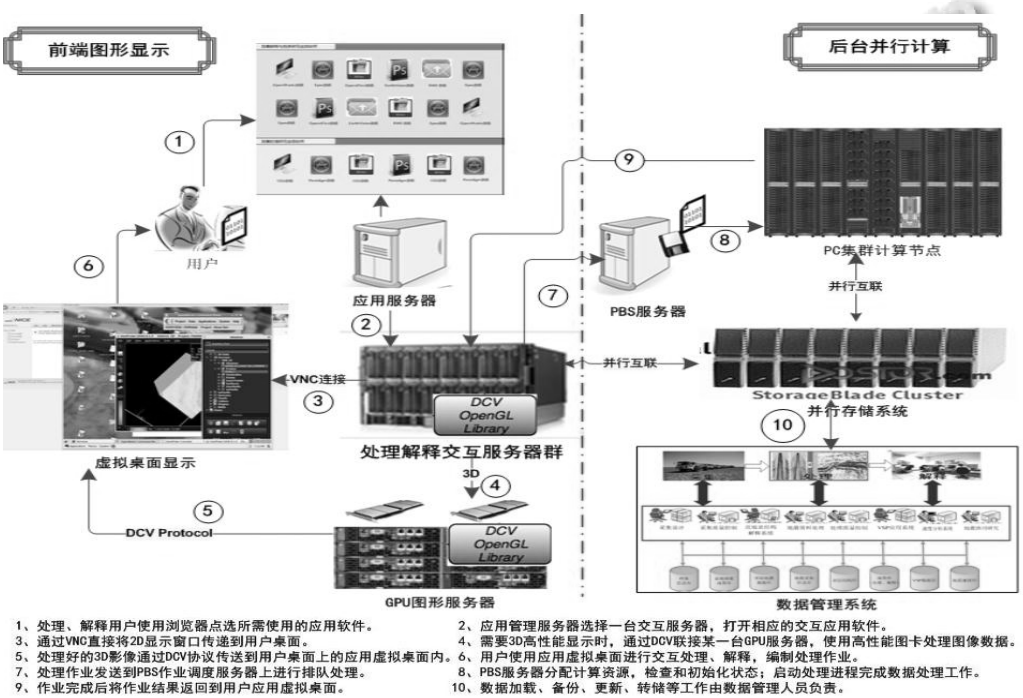

图 5　勘探云使用流程示意图

通过上图所示勘探云使用过程, 勘探科研人员可以在石油广域网络内部从异地登录使用勘探云计算资源而无需考虑所使用的数据是否准确、存放在什么地方、使用的计算资源是那一台计算机系统提供的. 科研人员主要工作精力逐步从勘探数据本身转移到这些数据的处理方法研究以及数据内在含义的解释上. 2013 年勘探科研人员利用勘探云完成了 113465 标准公里, 160TB(中间过程合计数据量为 1.6PB)野外原始数据的处理解释工作, 完成工作可折算成 1.4 亿人民币产值. 通过勘探处理解释和分析工作企业新发现了一系列优质油气资源, 圆满完成了油气储量勘探任务.

4 结论

通过以上勘探云的建设过程可以发现, 作为一种企业自建私有云, 勘探云主要目的就是保障勘探大数据处理、解释工作的顺利开展. 它是以企业应用为导向的信息支持系统. 业界很多人将“云计算”模式比喻为供电系统, 希望“计算力”能够象“电”一样统一产生, 按需分配. 然而正如离开了电视台、电视机以及电视节目, 电对人们娱乐方面的支持就变得毫无意义一样, 单纯强调“计算力”建设同样也没有多少价值, 信息系统最关键、最重要的是解决实际生产任务的应用系统. 以信息建设为主导的单纯软硬件堆叠和分配服务机制对企业效益的实现不会带来更多的便利. 这一点在面对大数据应用时更为重要, 云计算中心建设目的正在从其自身转换为企业践行大数据应用的支持平台.

参考文献

1 陆基孟.地震勘探原理.东营:石油大学出版社,1993.
2 钟瑛等.大数据的缘起、冲击及其应对.现代传播.中国传媒大学学报,2013,7:104–109.
3 邹杰等.基于 PBS 的勘探数据处理作业管理.计算机与现代化,2014,2:119–123.
4 邬贺铨.大数据时代的机遇与挑战.求是,2013,(4).
5 黄晓斌,钟辉新.基于大数据的企业竞争情报系统模型构建.情报杂志,2013(3).
6 汪圣利.大数据时代指挥信息系统发展分析.现代雷达,2013,(5).
7 曹建鹏等.云计算在石油勘探领域的应用.信息系统工程,2012,(6).
8 彭英等.一种用于石油勘探的云计算与虚拟存储平台设计.测绘与空间地理信息,2013,(11).
9 于会松.勘探协同研究云平台的设计及应用.计算机仿真,2014,(6).