DOI：10. 16516/j. gedi. issn2095-8676. 2015. S1. 049

电网非结构化数据管理平台研究与实现

冯国平，古明生，吉小恒
（中国能源建设集团广东省电力设计研究院有限公司，广州 510663）

摘要：随着大数据时代来临，数据被认为是企业重要的生产要素。提高非结构化数据的管理水平得到电网企业的重视。本文分析了电网企业非结构化数据的管理现状，提出采用大数据技术构建电网企业的非结构化数据管理平台。该平台采用 Hadoop 并行分布式计算架构和混合式存储采架构进行构建。文章阐述了非结构化数据管理平台的技术架构和应用架构，并结合电网企业实际简要论述了这种技术架构的先进性。

关键词：Hadoop；HDFS；NoSQL 数据库；非结构化数据管理平台
中图分类号：TP311　　**文献标志码**：A　　**文章编号**：2095-8676(2015)S1-0222-04

Research and Implementation of Unstructured Data Management Platform for Power Grid Enterprises

FENG Guoping, GU Mingsheng, JI Xiaoheng
(China Energy Engineering Group Guangdong Electric Power Design Institute Co., Ltd., Guangzhou 510663, China)

Abstract: With the advent of the era of big data, data is considered to be an important factor of production. Great importance has been attached to improving the management level of unstructured data by the power grid enterprises. This paper analyzes the current situation of the management of the unstructured data in the power grid enterprises, and proposes to construct the unstructured data management platform based on the big data technology. The platform uses the Hadoop parallel distributed computing architecture and the hybrid storage architecture. This paper describes the technical architecture and application architecture, and discusses the advanced nature of the technical architecture in the power grid enterprise.

Key words: hadoop; HDFS; NoSQL database; unstructured data management platform

随着大数据时代的来临，信息和数据以爆炸的方式增长。IDC 和 EMC 联合发布的《2020 年的数字宇宙》预测，到 2020 年全球数据存储量将会达到 40 000 EB，且以每两年翻一番的速度增长。企业数据按类型可以分为结构化数据和非结构化数据，在企业日常运营中产生的数据，能够采用关系型数据库处理的结构化数据约占企业数据总量的 20%，而其他 80% 的非结构化数据无法完全采用关系型数据库来处理。

数据成为重要的生产要素已经形成共识，在大型企业如何利用好企业最具价值的无形资产，发掘数据价值，提升企业核心竞争力已经成为最为紧迫的任务。整理、组织并分析非结构化数据，能够为企业带来更多的竞争优势。科学管理和合理应用这些非结构化数据已经成为企业正确决策、增强核心竞争力的关键。

1　基本概念

非结构化数据（Unstructured Data）是指无法用二维表结构化表示的一种数据类型，主要包括文本、音频、视频、图像、网页等。相比于交易型数据，非结构化数据的增长速度要快很多。非结构化数据的处理技术包括非结构化数据采集、存储、查询和管理，以及非结构化数据分析，包括文本挖掘、图像分析和视频分析。非结构化数据中至少 70% 上数据来源于人与人的互动与协作，是以人为中心产生的[1]。这些非结构化数据蕴涵着公司对提升业务质量的经验与思考，是宝贵的数据资产。

根据非结构化数据的四面体模型[1]，非结构化

收稿日期：2015-07-30
作者简介：冯国平(1980)，男，湖北云梦人，工程师，硕士，主要从事电电网企业信息化规划、大数据相关研究(e-mail) fengguoping@gedi. com. cn。

数据由原始数据、基本属性、语义特征和底层特征构成[2]。原始数据是指非结构化数据文件的本身，基本属性是指原始文件的一般属性，如文件名称、类型、创建时间等，语义特征是指原始文件的语义属性，如内容描述、主题说明等，而底层特征是指通过专用处理技术获得的一些数据特性，如颜色、形状、视频摘要等。一般的非结构化数据从上述四个方面进行描述。

2　电网企业非结构化数据管理现状

2.1　非结构化数据基本情况

电网企业非结构化数据的基本情况是数据类型多、数据量大、增长速度快，以及缺乏统一管理和充分利用。非结构化数据包括数字化的文书、图纸、照片、公司文件、报告等。从数据类型可将电网企业的非结构化数据分为企业经营管理数据、电网运行数据和外部环境数据。企业经营管理数据包括规划计划、工程建设、物资采购、设备运维、市场营销、人力资源和综合管理数据，电网运行数据包括电网运行、设备状态和运行环境数据，外部环境数据包括经济环境、宏观政策、行业对标和社交网络数据，如表 1 所示。

表 1　电网企业非结构化数据

Table 1　Unstructured Data of Power Grid Enterprises

数据分类	数据示例
规划计划	规划报告/计划报表
工程建设	设计图纸/施工现场视频
物资采购	合同文本/规格手册/资料图片
设备运维	监测视频/缺陷图片/试验报告
市场营销	用电合同/客服图片/营业厅视频
人力资源	劳动合同/照片/绩效报告
综合管理	设计报告/法律文件/案件影像
电网运行	运行报告/视频文件
设备状态	现场照片
运行环境	环境视频/山火监测
经济环境	各种经济报告/政策文件
宏观政策	政策文件
行业对标	各种文档报告/报表文档
社交网络	网页文件/网络图片/网络视频

2.2　非结构化数据存储

目前电网企业非结构化数据一般有四种存储方式。

方式一：将非结构化数据直接存储在关系数据库的二进制大数据对象类型 BLOB 字段中。这种方式的优点是调用文件的速度很快，维护和管理简单，缺点是使数据库迅速膨胀，导致数据库性能下降。

方式二：以文件传输协议(FTP)将非结构化数据文件保存到文件服务器中，文件服务器独立于应用系统服务器。采用 FTP 协议的方式进行文件传输，文件方便共享，维护相对简单。

方式三：以人工的方式通过文件系统直接存储在文件服务器中。这些数据没有应用系统管理，而是由人工管理，如 IT 工具软件、源代码、开发文档、技术资料、新闻素材等，通常都是将文件直接存储到文件服务器中。

方式四：非结构化数据零散存在于企业员工的各种工作终端上，由员工自行管理，如员工的工作报告、汇报材料等与员工工作内容相关的非结构化数据。

2.3　非结构化数据操作与利用

电网企业的非结构化数据操作一般有两种，一是基于非结构化数据的索引，另一种是基于非结构化文件本身。对于非结构化数据索引方式，采用各种工具，基于领域的知识抽取非结构化数据的描述信息并建立索引，基于索引进行文本的检索，实现对非结构化数据的利用。这种方式一般基于关系数据库，因此数据模型是核心[4]。对于非结构化文件本身方式，以图像、视频和音频等数据文件本身所包含的信息为基础进行检索，这种方式涉及到各种不同的多媒体文件处理、模式识别技术[5]。

目前开展非结构化数据统一管理的电网企业较少，数据利用方式基本是在本地生成、本地存储、系统内部使用。应用的方式仅是检索、查看等简单利用，没有与结构化数据融合，不仅没有实现企业级的资源共享，数据的价值也远远没有发掘出来，还存在一定程度的重复建设和资源浪费。

基于 EMC Documentum 等成熟商业产品构建非结构化数据管理平台 UDMP(Unstructured Data management Platform)是一般企业的首选技术路线，然而这种方式需要花费昂贵的采购成本，可扩展性差。在当前大数据技术日趋成熟的技术条件下，本研究了非结构化相关的存储、计算技术，提出基于大数据技术实现电网企业非结构化数据管理平台。

3 基于大数据技术的 UDMP

大数据技术是指以 MapReduce 并行分布式计算、NoSQL 数据存储为代表的从大数据中获取有价值信息的技术，包括数据采集、数据预处理、数据存储、数据分析等相关技术。

3.1 基于 Hadoop 的并行分布式架构

Hadoop 是 Apache 基金会的发起的一个分布式系统基础架构，基于该架构用户可以无需关注底细节，充分利用集群的资源进行高速的计算和存储。Hadoop 框架的核心是 MapReduce 和 HDFS。MapReduce 是 Google 实验室提出的分布式并行计算编程模型，用于大规模数据的并行处理，其技术原理是提供一个包含 Map 和 Reduce 两阶段的并行处理模型和过程，提供一个并行化编程模型和接口，以键值对数据输入方式处理数据，自动完成数据的划分和调度管理。HDFS 是一个分布式文件系统，具有高容错性、高吞吐量的特点，可部署在低廉的硬件设备上。对于电力企业而言，基于 MapReduce 的分布式计算框架适合于大规模数据量分析计算，如日志分析、电力系运行数据分析、设备状态监测数据分析、信息提取和数据挖掘等，但不适合实时性要求高的数据分析计算场景。HDFS 则适合于电网企业当前体量巨大、价值密度相对较低的非结构化数据存储。

本文基于 Hadoop 分布式框架，提出电网企业的非结构化数据管理平台的技术架构，如图 1 所示。

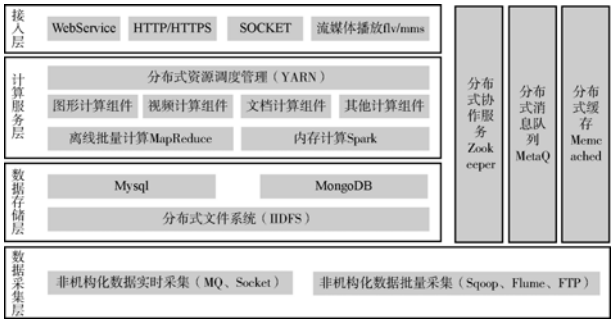

图 1 UDMP 技术架构

Fig. 1 Technical Architecture of UDMP

UDMP 采用 MapReduce 并行分布式计算框架，基于 Hadoop 架构实现电网企业非结构化数据的采集、存储和计算。UDMP 技术架构从下至上依次分为数据采集层、数据管理层、计算服务层和接入层。数据采集层通过 Sqoop、Flume 等数据收集和交换工具从关系数据库或文件系统抽取非结构化数据的元数据，通过 FTP 等方式实现非结构化文件的传输。数据管理层通过 Hadoop 框架实现非结构化数据的存储和计算。接入层以服务提供的形式对外提供非结构化数据的访问服务。

3.2 基于 NoSQL + MySQL 混合式存储架构

MongoDB 是应用广泛的文档型数据库。文档型数据库是 NoSQL 中非常重要的一个分支，主要用来存储、索引并管理面向文档的数据或者类似的半结构化数据。MongoDB 的特点是弱一致性、用户访问速度快、支持大容量的存储，性能优越，但是不支持事务操作，且占用空间过大。MySQL 是最为流行的关系型数据，由于其体积小、速度快、总体拥有成本低且开放源码收到广泛欢迎。关系型数据库的最大特点是支持事务型操作。

UDMP 内部的数据可分为三种类型：管理类数据、元数据和非结构化数据。管理类数据如用户、权限、系统配置、存储级别、数据生命周期节点等结构化数据。这类数据关系型操作较多，涉及到事务型操作，适合采用 MySQL 这类关系型数据存储（也可采用其它商业数据库产品，如 Oracle、SqlServer 等）。非结构化数据的元数据结构简单但是结构不固定、数据量大，事务操作较弱，因此适合采用 MongoDB 存储。同时，MongoDB 面向文档存储，模式自由，非常适合非结构化数据文件本身的存储。NoSQL + MySQL 混合式存储架构如图 2 所示。

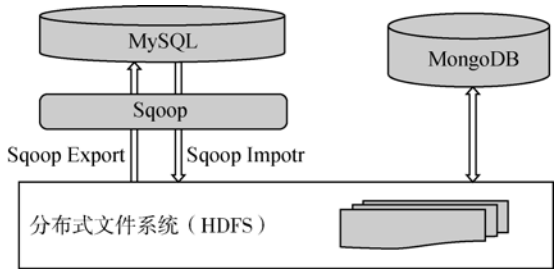

图 2 NoSQL + MySQL 混合式存储架构

Fig. 2 Hybrid Storage Architecture of UDMP

3.3 非结构化平台的应用架构

UDMP 包含 5 个模块，分别是数据存储、数据管理、数据应用、元数据标准管理和平台管理等模块，其应用架构如 3 图。依托与基础软硬件设施，UDMP 为业务管理系统、企业搜索引擎等提供非结构化数据存储、管理与利用服务。

数据存储模块提供非结构化数据存储的相关功能，包括数据存储管理、不同类型的存储管理以及数据路由与交换等功能。数据管理模块实现提供非结构化数据统一管理的相关功能，实现元数据、目录、路径及权限和数据全生命周期的管理等。数据应用模块提供非结构化数据利用的相关功能，如全文搜索、文档格式转换等。元数据标准管理模块实现非结构化数据标准的管理和编码管理功能。平台管理模块实现平台的配置和监控管理。

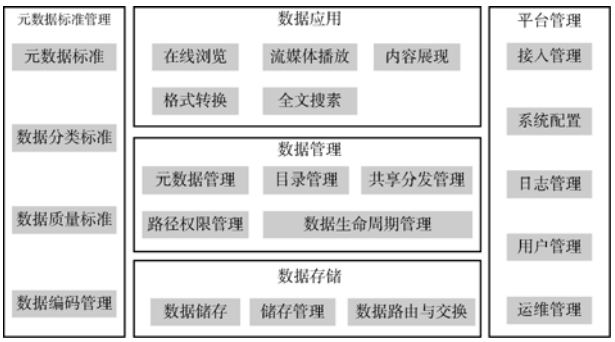

图 3　非结构化数据管理平台应用架构

Fig. 3　Application Architecture of UDMP

4　结论

以 Hadoop 分布式框架和 NoSQL 数据库为代表的大数据技术具备高效、可靠、可伸缩和低成本等特点，基于大数据技术的电网非结构化数据管理平台在架构上灵活、可扩展，具有一定的先进性。在技术实现上摒弃了基于商业产品构建，而是采用基于 Hadoop 框架和 MongoDB 自主开发的技术路线，减少了对 IOE 产品的依赖。基于大数据技术的非结构化数据管理平台必将在电力企业产生良好的应用效果。

参考文献：

[1] OASIS. Unstructured Information Management Architecture(UIMA)[M]. Working Draft 05, 2008.

[2] 李未，朗波. 一种非结构化数据库的四面体数据模型[J]. 中国科学：信息科学，2010，40(8)：1039-1053.
LI Wei, LANG Bo. A Tetrahedral Data Model of Unstructured Database[J]. Science China, 2010, 40(8): 1039-1053.

[3] 韦琳，袁泉，霍剑青，等. E-learning 非结构化数据管理系统的构建与实现[J]. 中国科学技术大学学报，2010，40(6)：623-628.
WEI Lin, YUAN Quan, HUO Jianqing, et al. The Design and Implementation of Unstructured Data Management System in E-learning Teaching System[J]. Journal of University of Science and Technology of China, 2010, 40(6): 623-628.

[4] 李威. 半结构化数据挖掘若干问题研究[D]. 长春：吉林大学计算机软件与理论，2013.

[5] 邹波. 海量非结构化数据的组织研究与实现[D]. 武汉：华中科技大学计算机系统结构，2008.

(责任编辑　郑文棠)

(上接第 217 页 Continued from Page 217)

[6] LIN Luo, GJALT H. Life Cycle Assessment and Life Cycle Costing of Bioethanol from Sugarcane in Brazil[J]. Renewable and Sustainable Energy Reviews, 2008, 3(8): 1-7.

[7] ARCHED M. A Novel Fuzzy Logic Technique for Power Transformer Asset Management[C]. In: Islam S. M, eds. Industry Applications Conference, 2006. The 41th IAS Annual Meeting, 2006, 276-280.

[8] RAY Mohapatra S K, SUBRATA Mukhopadhyay. Risk and Asset Management of Transmission System in A Reformed Power Sector[C]. In: power india conference, 2006IEEE, 2006, 725-730.

[9] 韩天祥，李莉华，余颖辉. 用 LCC 方法对 500 kv 变电站改造的经济性评价[J]. 华东电力，2007，35(8)：7-11.
HAN Tianxiang, LI Lihua, YU Yinghui. Economic Evaluation for 500 kv Substation Transformation in LCC Method[J]. East China Electric Power, 2007, 35(8): 7-11.

(责任编辑　黄肇和)