

文章编号:1671-251X(2019)08-0060-05

DOI:10.13272/j.issn.1671-251x.17436

基于 Hadoop 的煤矿数据中心架构设计

韩安^{1,2}

(1. 中煤科工集团常州研究院有限公司, 江苏 常州 213015;
2. 天地(常州)自动化股份有限公司, 江苏 常州 213015)



扫码移动阅读

摘要:根据煤矿安全生产业务需求及智慧矿山发展要求,新型的煤矿数据中心需满足对同一时空坐标体系下煤矿海量、多元数据的高效处理、缓存、计算、存储与发布。针对传统煤矿数据中心各类数据离散存储,数据集成、业务应用及数据分析难度大等问题,设计了一种基于 Hadoop 的煤矿数据中心架构。采用 Storm 实时数据流引擎进行数据实时计算,并应用 MapReduce, Spark 实现批处理计算和内存计算,解决高频时序数据存储与海量数据计算问题;采用 Hadoop 分布式文件系统(HDFS)实现文件的可靠存储,并采用 HBase 分布式存储数据库实现历史数据的持久化存储,利用其无模式稀疏设计满足同一时空坐标体系下的数据分析需求;采用 Redis 作为实时数据库存储数据快照进行最新数据检索。基于 Hadoop 的煤矿数据中心充分利用大数据技术的高速数据存取和压缩性能,通过统一的云计算数据中心,有效降低了煤炭企业数据应用的复杂度及数据中心日常运维难度,可为智慧矿山建设奠定数据基础。

关键词:智慧矿山; 煤矿数据中心; Hadoop; 大数据; 云计算
中图分类号:TD67 文献标志码:A

Architecture design of coal mine data center based on Hadoop

HAN An^{1,2}

(1. CCTEG Changzhou Research Institute, Changzhou 213015, China;
2. Tiandi (Changzhou) Automation Co., Ltd., Changzhou 213015, China)

Abstract: According to business requirements of coal mine safety production and development requirements of smart mine, the new coal mine data center needs to meet efficient processing, caching, calculation, storage and release of massive and multivariate data of coal mine under the same space-time coordinate system. In view of problems of discrete data storage and difficulties in data integration, business application and data analysis of traditional coal mine data center, an architecture of coal mine data center based on Hadoop was designed. Real-time data stream engine storm is adopted for real-time data calculation, and MapReduce and Spark are applied to implement batch processing and memory calculation to solve high-frequency time-series data storage and massive data calculation problems. Hadoop distributed file system (HDFS) is used for files reliable storage, and distributed storage database HBase is used for persistent storage of historical data, which can meet data analysis needs of the same space-time coordinate system by its modeless sparse design. Redis is used as a real-time database storage data snapshot for the latest data retrieval. Coal mine data center based on Hadoop makes full use of high-speed data access and compression performance of big data technology, complexity of data application in coal enterprises and daily operation and maintenance difficulty of data center are effectively reduced through unified cloud computing data center, which can lay a data foundation for the construction of smart mine.

Key words: smart mine; coal mine data center; Hadoop; big data; cloud computing

收稿日期:2019-05-13;修回日期:2019-07-16;责任编辑:胡娴。
基金项目:天地科技股份有限公司科技创新创业资金专项项目(2018-TD-QN019);天地(常州)自动化股份有限公司科研项目(2018GY117)。
作者简介:韩安(1982—),男,陕西泾阳人,工程师,现主要从事信息化产品规划与管理工作,E-mail:hanan@cari.com.cn。
引用格式:韩安. 基于 Hadoop 的煤矿数据中心架构设计[J]. 工矿自动化,2019,45(8):60-64.
HAN An. Architecture design of coal mine data center based on Hadoop[J]. Industry and Mine Automation,2019,45(8):60-64.

0 引言

随着“大物移云”(大数据、物联网、移动互联网、云计算)技术的蓬勃发展,智能化开采及智慧矿山相关技术也迅速发展,带来了煤炭行业数据量及业务应用模式上的逐步变化。传统的煤矿数据中心中,煤矿主数据、实时监测监控类数据、管理业务类数据及决策分析结果数据各自离散存储,数据集成、业务应用及数据分析难度大,数据中心日常运维困难,已经无法有效满足企业在自动化、信息化业务层面的应用需求,更无法满足智慧化决策分析的需求。这些问题给煤矿安全生产带来了极大隐患,也严重阻碍了智慧矿山的发展。本文在剖析煤矿数据中心应用需求及传统数据中心缺陷的基础上,提出一种基于 Hadoop 的煤矿数据中心架构设计方案。

1 煤矿数据中心需求

煤矿安全生产属于专业性、系统性工程,煤矿企业在安全生产过程中涉及以下数据:地理地测、井巷工程等固有地质基础数据,安全监控、人员定位、调度通信、供水网络、主要通风机监控、压风监控等安全保障数据,掘进机监控、工作面“三机”监控、带式输送机监控、电力监控、排水监控、提升机监控、称重计量、快速装车等生产过程控制数据,安全风险分级管控、安全隐患排查治理、瓦斯巡检、生产调度、设备巡检、生产技术、一通三防、地测防治水、安全培训等业务管理数据,人力资源管理、物资管理、财务管理等企业资源管理数据,企业安全、生产、运营各项关键指标数据。

根据数据类型、周期性及相互关系等特征进行归纳,可将煤炭企业数据分为企业主数据、空间基础数据、文件图纸类数据、实时监控类数据、业务类数据及决策分析结果数据。

目前,多源灾害综合性安全预防预警、安全生产运营管控一体化、企业内部市场化、矿山物联网、智慧矿山等成为煤炭行业的主要研究课题。煤矿安全生产在空间上立体分布,时间上动态发展,安全生产过程管理存在较大的动态性、随机性和模糊性,安全生产的影响因子具有复杂性、相关性和多元性,而课题研究均需要建立在同一时空坐标体系下的数据集。因此,新型的煤矿数据中心需满足对同一时空坐标体系下煤矿海量、多元数据的高效处理、缓存、计算、存储与发布,以有效支撑煤矿安全生产业务需求及智慧矿山发展要求。

2 传统煤矿数据中心

传统煤矿数据中心主要结合关系型数据库、实

时数据库进行构建。关系型数据库按照一定数据结构和模型来组织、存储和管理数据,将各种业务实体及实体之间的联系用关系模型进行表达,实现对企业主数据、空间基础数据及管理业务数据的管理。实时数据库主要采用实时数据模型处理不断更新、快速变化的数据及具有时间限制的事务,解决实时系统中的数据管理问题,满足海量数据的实时读写操作、大容量数据存储、集成工业接口数据采集及现场实时控制的要求。

传统数据中心解决了煤矿各类安全生产及管理业务类数据的存储与发布,支撑了煤矿安全生产业务的应用需求^[1]。然而,随着煤炭行业设备智能化及信息化的快速推进、安全生产标准化要求的提升、智能开采及智慧矿山的发展,传统数据中心已经很难支撑煤矿海量、多元数据的处理、缓存、计算、存储与发布的应用需求,存在以下缺陷:

(1) 由于煤矿安全生产是一个动态过程,煤矿井巷工程、传感器布置等均随工作面的推进而发生变化,实时数据库无法动态加载变化测点,需要人为进行维护,以实现煤矿井下动态生产过程监控^[2]。

(2) 人员/车辆定位数据属于连续记录型数据,无法通过实时数据库测点标签、点值、状态、时间戳的数据结构进行存储,主要采用关系型数据库进行存储,通过关系型数据分区技术来提升数据存储性能。但对于人员/车辆定位,特别是精确定位的海量轨迹类数据,存在数据处理、存储等性能瓶颈。

(3) 实时数据库基于时标进行处理,只能针对时间段进行数据查询,虽然各厂家定制了历史库工具以补充实时数据库在查询方面的不足,但无法与关系型数据库相比。实时数据库主要面向大型工业企业,价格昂贵,成本较高。实时数据传输主要依托工业网络,不适合云计算环境下的分布式部署,造成数据中心难以动态水平扩展^[3]。

3 基于 Hadoop 的煤矿数据中心架构设计

采用 Yarn(Yet Another Resource Negotiator,另一种资源协调者)实现资源管理与任务调度,以 Hadoop 生态为基础构建数据中心^[4],基于 Hadoop 的煤矿数据中心架构如图 1 所示。数据计算:采用 Storm 实时数据流引擎进行数据实时计算;应用 MapReduce,Spark 实现批处理计算和内存计算,解决高频时序数据存储与海量数据计算问题^[5]。数据存储:采用 HDFS(Hadoop Distributed File System,Hadoop 分布式文件系统)实现文件的可靠存储;采用 NoSQL 的 HBase 分布式存储数据库实现历史数据的持久化存储,并利用其无模式稀疏设

计满足同一时空坐标体系下的数据分析需求。数据检索:采用 Redis 作为实时数据库存储数据快照进行最新数据检索。通过采用统一的云计算数据中心,有效降低了煤炭企业数据应用的复杂度及数据中心日常运维的难度^[6]。

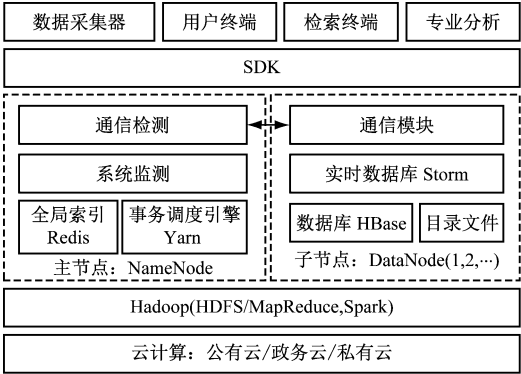


图 1 基于 Hadoop 的煤矿数据中心架构
Fig. 1 Architecture of coal mine data center
based on Hadoop

主节点运行在 Hadoop 的主节点计算机上,作为 HDFS 的 NameNode,同时也是 Master 节点,实现全局索引管理、数据块映射管理等全局事务管理,处理客户端的读写请求,配置副本策略,同时管理 HDFS 的名称空间。在具体应用中,为保证

NameNode 的可靠性和稳定性,同步建立一套 SecondaryNameNode,定时与 NameNode 进行同步,合并文件系统镜像和编辑日志,以备主节点故障后接管 Master 权限。

子节点运行在 Hadoop 的子节点计算机上,作为 Slave 的 DataNode,存在多个副本,负责存储主节点接收的客户端数据块,执行数据块读写操作;完成实时监测类数据的缓存与计算,图纸文档类、管理业务类、监测类历史数据及决策分析结果的存储;与其他 DataNode 进行通信,复制数据块,以保证数据的冗余性;响应主节点的监测,及时报告数据块修改信息^[7]。

数据中心对外以标准 SDK (Software Development Kit,软件开发工具包)进行接口发布,为数据采集器提供监测点、实时数据、历史数据的操作接口,为用户终端提供对企业主数据、图纸文档类数据、管理业务类数据、分析结果数据的操作接口,提供对外数据检索及专业应用分析的操作接口等。

4 数据流程设计

4.1 数据存储

数据存储流程如图 2 所示。

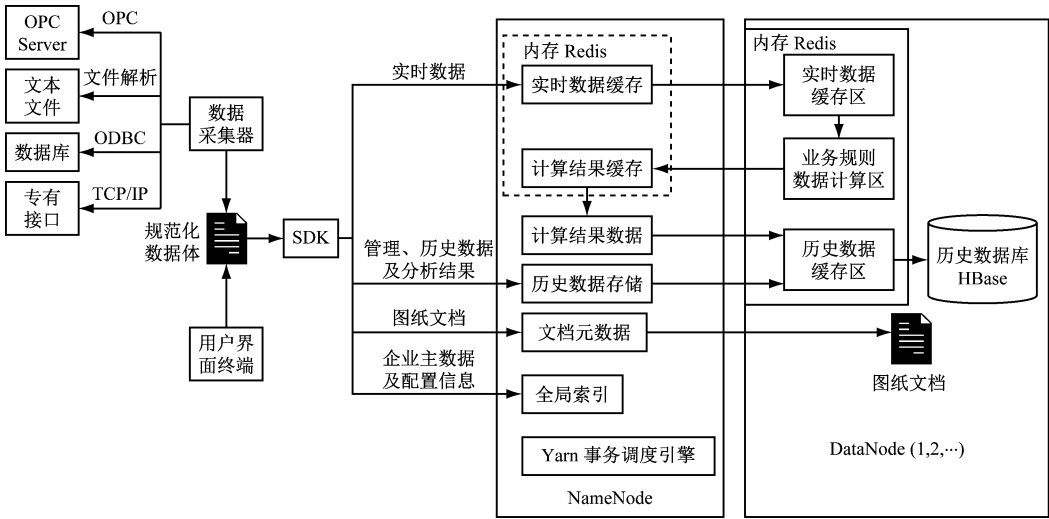


图 2 数据存储流程
Fig. 2 Data storage process

数据采集器通过 OPC (OLE for Process Control,用于过程控制的 OLE)、TCP/IP、文本、ODBC(Open Database Connectivity,开放数据库连接)等多种方式采集数据,根据不同的数据解析驱动完成实时数据解析,形成规范化数据体,并按照业务规则进行数据的一级压缩。

用户可通过界面终端对企业主数据、图纸文档类数据、管理业务类数据、分析结果数据进行操作。根据不同的业务规则请求不同的 SDK,将不同的数

据体提交给数据中心^[8]。

数据中心主节点 NameNode 接收到客户端请求后,通过 Yarn 事务调度引擎进行任务拆解分派:① 存储企业主数据及配置信息时,更新 NameNode 中的全局索引。② 存储实时数据时,由于监测监控类数据对存储时间及顺序要求较高,所以,需根据全局索引将数据存储任务按子系统或监测点映射调度多个计算节点 DataNode。DataNode 接收到请求写入的数据体后,根据业务应用需求完成数据计算,并

将结果返回 NameNode;Yarn 事务调度引擎根据需求完成计算结果的缓存,并根据业务规则将计算结果转为历史数据,调度多个 DataNode 完成历史数据存储。③ 存储管理业务类数据、历史数据及分析结果数据时,Yarn 事务调度引擎调度多个 DataNode 完成数据存储。④ 存储数据图纸、文档时,NameNode 完成文件的元数据存储,并根据全局索引调度多个 DataNode 完成文件存储。

DataNode 中的实时数据缓存接收到 Yarn 分解的实时数据时,按照数据流完成以下任务:① 实时数据缓存区按照数据体中的时间戳进行排序,将最新实时数据置于缓存区以便进行检索,并将排序后的数据体提交给业务规则数据计算区^[9]。② 业务规则数据计算区利用 Storm 实时流计算服务对时序数据体进行计算,如根据实时监测数据完成报警类数据计算、根据当前人员区域分布情况完成区域超员的计算等。③ 利用 Redis 实现对原始实时监测数据、实时报警信息、业务分析中间数据进行缓存^[10]。④ DataNode 通过规约将计算结果及完成计算指令提交给 NameNode。⑤ NameNode 根据业务规则将部分计算结果缓存转为计算结果数据,由 Yarn 根据资源调度多个 DataNode 进行存储。⑥ DataNode 接收 Yarn 调度指令,从 NameNode 接收各类历史数据块,并暂存至历史数据缓存区。当 HBase 资源充足时,直接存入 HBase 数据库中;如果 HBase 资源不足,则通过 Redis 暂时进行持久化存储,等资源充足时再存入 HBase 数据库,以保证历史数据的完整性^[11]。⑦ DataNode 从 NameNode 接收文档数据块,完成文件的存储。

4.2 数据查询

用户终端、检索终端及专业分析终端通过 SDK 实现数据查询:① 获取配置信息。通过接口获取 NameNode 中的对应全局索引。② 检索监测点实时数据。通过 NameNode 中的全局索引表及服务器映射表实现数据监测点的快速定位,避免轮询各服务器信息。全局索引表主要定义了该监测点所在服务器列表及各服务器所在地址。服务器映射表实现对服务器 IP 地址、名称、当前状态的记录,根据服务器状态及监测点所在地址,通过检索 Redis 中的监测点 Key 值获取实时数据。③ 检索监测点一段时间内的历史数据或人员在井下一定时间内的活动轨迹。通过全局索引表及服务器映射表检索到监测点或人员射频卡的 Key,通过 HBase 数据表中对应 Key 值及时间标签实现对应数据存储区内的检索。④ 查询同一时空坐标体系下监测点及人员数据。利用 HBase 无模式稀疏设计的特性^[12-13],将不同的

监测点数据放在同一行,以时间为主键,迭代查询监测点历史记录或某一区域内人员分布情况。⑤ 复杂性数据查询和统计。通过事务调度引擎实现复杂事务的作业和任务调度,分配和调度 DataNode,利用 MapReduce 实现任务的分级映射与归约、对各计算节点并行处理任务的管理及计算结果的收集。

5 特性分析

- (1) 易用性。采用 SDK 封装方式对外提供煤矿数据中心的系列操作,包括数据中心配置和实时数据更新、历史数据存储和检索、统计分析等,降低了业务系统与数据中心之间的耦合难度。
- (2) 较高的可靠性与扩展性。基于 Hadoop 的架构有效地避免了数据中心的单点故障,提高了可靠性,方便进行节点动态扩展^[14]。
- (3) 支持海量多元数据存储。通过 NoSQL 的 HBase 数据存储方式,满足煤矿主数据、动态变化的监测数据等海量、多元数据的统一集中存取^[15]。
- (4) 支持分布式大数据计算。利用 MapReduce、Spark 实现批处理计算和内存计算,具备分布式大数据计算能力。
- (5) 支持多维度数据统计分析^[16]。利用 Hbase 无模式稀疏设计特性提供同一时空坐标体系下的数据查询统计,并通过接口封装对外提供数据聚合、选择、转换等多种函数,支持多维度数据统计分析。

6 结语

结合煤炭行业设备智能化及信息化的推进,安全生产标准化要求的提升,智能开采及智慧矿山的发展,海量、多元数据处理、缓存、计算、存储与发布的应用需求,设计了基于 Hadoop 的煤矿数据中心架构,有效弥补了传统数据中心在应对统一数据存储、动态生产监控需求、连续记录型数据存取、动态扩展及海量数据计算分析方面的不足,满足了煤炭行业大数据在可靠性、扩展性、数据检索和计算能力方面的需求,提升了煤炭企业的智能化和信息化水平,并通过大数据手段挖掘潜在的数据价值,为智慧矿山建设奠定了数据基础。

参考文献(References):

[1] 张杰. 神华宁煤集团安全智能分析平台建设构想[J]. 工矿自动化,2012,38(9):83-86.
ZHANG Jie. Construction conception of intelligent safety analysis platform of Shenhua Ningxia Coal Industry Group[J]. Industry and Mine Automation, 2012,38(9):83-86.

- [2] 王渊,韩安,晁晓菲.煤炭集团公司生产安全调度指挥管理平台的建设构思[J].工矿自动化,2012,38(3):26-28.
WANG Yuan, HAN An, CHAO Xiaofei. Construction ideas of management platform of production safety dispatching of coal enterprise group [J]. Industry and Mine Automation, 2012, 38(3): 26-28.
- [3] 戚雪冰.风电监控系统中时序数据管理系统的设计与实现[D].南京:东南大学,2015.
- [4] 徐晓亮,温泉,姜彩玉.基于时序数据库的电网潮流全息全过程重演[C]//中国电机工程学会年会,海口,2010.
- [5] 林荫峰,林俊仁,杨俞平.基于时序数据库的空间数据处理[J].测绘与空间地理信息,2017,40(增刊1):178-180.
LIN Yinfeng, LIN Junren, YANG Yuping. Dealing with spatio-temporal data based on a new time-series database [J]. Geomatics & Spatial Information Technology, 2017, 40(S1):178-180.
- [6] 张华伟,陈勇,李海斌,等.基于HBase的工业大数据时序数据存储实现[J].电信科学,2017,33(增刊1):21-27.
ZHANG Huawei, CHEN Yong, LI Haibin, et al. Implementation of industrial big data storage in real-time database based on HBase [J]. Telecommunications Science, 2017, 33(S1):21-27.
- [7] 贺耀宜,王渊.煤炭企业集团公司生产综合调度指挥系统建设初探[J].工矿自动化,2010,36(5):23-27.
HE Yaoyi, WANG Yuan. Discussion of construction of integrated production dispatching and command system for coal groups [J]. Industry and Mine Automation, 2010, 36(5):23-27.
- [8] 郑晓露,粟勇,朱月梅.基于时标分层技术的时序数据存储引擎研究与实现[J].计算机应用,2017,37(增刊1):246-249.
ZHENG Xiaolu, SU Yong, ZHU Yuemei. Research and implementation of time series data storage engine layered by time scale [J]. Journal of Computer Applications, 2017, 37(S1):246-249.
- [9] 梅寒蕾.基于分段动态时间弯曲的时间序列查询[D].杭州:浙江大学,2017.
- [10] 黄红远,李高明,欧阳为年,等.一种新型时序数据库在大型配电网自动化系统数据中心中的应用[J].中国电机(技术版),2014(7):35-37.
HUANG Hongyuan, LI Gaoming, OUYANG Weinian, et al. Application on the new type database in large scale distribution automation system data center [J]. China Electric Power (Technology Edition), 2014(7):35-37.
- [11] 沈宇,王祺.基于大数据的煤矿安全监管联网平台设计与实现[J].矿业安全与环保,2016,43(6):21-24.
SHEN Yu, WANG Qi. Design and implementation of coal mine safety supervision networking platform based on big data[J]. Mining Safety & Environmental Protection, 2016, 43(6):21-24.
- [12] 荣雪琴,刘勇,刘昊,等.基于时序数据库的电力运维系统关键技术研究[J].电测与仪表,2018,55(9):48-51.
RONG Xueqin, LIU Yong, LIU Hao, et al. Research on key technology of power operation and maintenance system based on timing database [J]. Electrical Measurement & Instrumentation, 2018, 55(9):48-51.
- [13] 焦冬冬,张晨,房志奇,等.基于云技术的分布式实时数据库系统的设计与实现[J].电子技术应用,2017,43(11):112-115.
JIAO Dongdong, ZHANG Chen, FANG Zhiqi, et al. Design and implementation of the distributed real-time database system based on the cloud technology [J]. Application of Electronic Technique, 2017, 43(11):112-115.
- [14] 郭蒙雨,康宏,袁晓洁.基于流式计算框架的实时数据库分区系统[J].计算机工程,2017,43(11):8-15.
GUO Mengyu, KANG Hong, YUAN Xiaojie. Real-time database partitioning system based on streaming computing framework [J]. Computer Engineering, 2017, 43(11):8-15.
- [15] 庄杰.基于OPC的实时数据库关键技术研究[D].西安:西安石油大学,2015.
- [16] 吕鹏飞,何敏,陈晓晶,等.智慧矿山发展与展望[J].工矿自动化,2018,44(9):84-88.
LYU Pengfei, HE Min, CHEN Xiaojing, et al. Development and prospect of wisdom mine [J]. Industry and Mine Automation, 2018, 44(9):84-88.