

基于 Hadoop+MPP 架构的城市轨道交通大数据中心建设方案

朱嘉斌

(苏州市轨道交通集团有限公司, 215004, 苏州//高级工程师)

摘 要 分析了城市轨道交通大数据中心建设中面临的挑战,介绍了 Hadoop+MPP 技术架构的特点,介绍了基于 Hadoop+MPP 架构的大数据中心的系统逻辑架构、数据处理架构和数据处理流程。苏州轨道交通的项目实践表明,基于 Hadoop+Mpp 架构的大数据中心建设方案能够达到预期效果。

关键词 城市轨道交通;大数据中心;Hadoop;MPP;建设方案

中图分类号 U29-39

DOI:10.16037/j.1007-869x.2022.05.012

Construction Scheme of Urban Rail Transit Big Data Center Based on Hadoop+MPP Architecture

ZHU Jiabin

Abstract Challenges faced by urban rail transit big data center construction are analyzed. The characteristics of Hadoop + MPP technological architecture are introduced, as well as the system logic architecture, data processing architecture and data processing procedure of the big data center based on Hadoop + MPP architecture. The practice of Suzhou Rail Transit project has proven that the construction scheme of big data center based on Hadoop + MPP architecture can perform as expected.

Key words urban rail transit; big data center; Hadoop; MPP (massively parallel processing); construction scheme

Author's address Suzhou Rail Transit Group Co., Ltd., 215004, Suzhou, China

0 引言

城市轨道交通线网大数据中心统一收集、处理和储存各类数据,实现对线网内各个生产系统的监督、协调、监控、统计、分析和管理等。城市轨道交通数据来源广泛、数量庞大、类型多样、更新快,具有异构、量多、类杂和自组织等特点。

在城市轨道交通大数据中心建设过程中,面临的主要挑战为:①建设成本:数据中心处理的数据量大、处理要求高,且后续新建线路也需要接入数据中心,所以一次性建成数据中心的投资巨大,初期成本难以控制。②业务扩展:随着新线不断建设,以及技术不断创新和发展,数据中心应用软件系统也要不断升级,导致业务形态有很大的不确定性。③数据分析:大数据中心的数据分析旨在提取、挖掘海量数据背后的各种规律。核心问题在于如何有效地对海量数据进行组织、学习、计算、表达。设计同时适用于结构化数据和非结构化数据的组织管理系统是巨大挑战。④程序性能:如何构建高效自动化索引,如何优化组织、管理数据的工作流程,以便尽可能自动化处理各类事务,减少额外的资源占用,提高效率,是面临的重要挑战。

大数据中心处理的数据类型多样,业务广泛,彼此有千丝万缕联系,数据中心需要结合各专业信息对多种维度数据进行综合分析才能产生有价值的成果。本文针对以上挑战,结合大数据中心的现实需求,提出了基于 Hadoop+MPP 技术架构的大数据中心建设方案。

1 Hadoop+MPP 技术架构特点分析

Hadoop(一种分布式系统基础架构)是一个分布式系统基础架构。Hadoop 的整体优势是数据处理能力强、成本低、高可靠性和灵活的可扩充性。Hadoop 核心内容为:①HDFS(分布式文件系统)——是一种新型分布式文件系统,可提供高可靠、高扩展、高吞吐能力的海量文件存储业务。②Map/Reduce(映射/化简)模型——并行计算方式遵循 Map/Reduce 模型就可以实现分布式并行计算。③HBase 数据库——是非关系型数据库,主要依靠横向扩展,通过不断增加 PC 服务器就可增加计算

和存储能力。

MPP(一种海量数据实时分析架构)是通过一定的互联网节点连接多个 SMP(对称多处理)服务器协同完成工作任务。MPP 数据库将任务并行地分散到多个服务器和节点上,在每个节点计算完成后,将各自的结果汇总在一起从而得到最终结果。与传统的关系型数据库相比,MPP 在数据处理方面的优势为:①分布式架构。②处理数据量大,能处理 PB(千万亿)级数据。③更大的 I/O(输入/输出)能力。因为采用完全无共享的并行处理架构,所以能充分利用资源。④扩展能力好。⑤采用列存储,能节省更多的存储空间。

考虑到 Hadoop 和 MPP 的特性,将两者结合搭配使用是最佳方案。利用 x86 服务器搭建分布式数据库,利用 Hadoop+MPP 架构管理处理汇总的各类数据。Hadoop+MPP 架构的优点是:初期投资低、硬件方便扩展、容错性高、处理能力强;用户可以在不了解分布式底层细节的情况下,开发分布式程序,

充分利用集群功能进行高速运算和存储;可以同时结构化数据和非结构化数据进行在线交互处理。其缺点是对系统及软件开发人员的要求高。

选择批量的 x86 服务器搭建分布式的数据中心硬件平台,采用 Hadoop+MPP 架构交互处理各类实时和离线的结构化和非结构化数据,能大大降低初始建设成本,控制预算,而且能够较为贴切地解决城市轨道交通数据中心数据量大、关联性强、非结构化数据多等痛点,能很好实现数据挖掘分析,并在后期的发展过程中能根据业务需要灵活扩展硬件和系统软件以增加处理能力和升级业务。苏州轨道交通大数据中心项目就采用了该方案。

2 基于 Hadoop+MPP 技术架构的大数据中心系统技术实现

2.1 系统逻辑架构

根据城市轨道交通大数据中心的业务特点设计的大数据中心系统逻辑框架如图 1 所示。

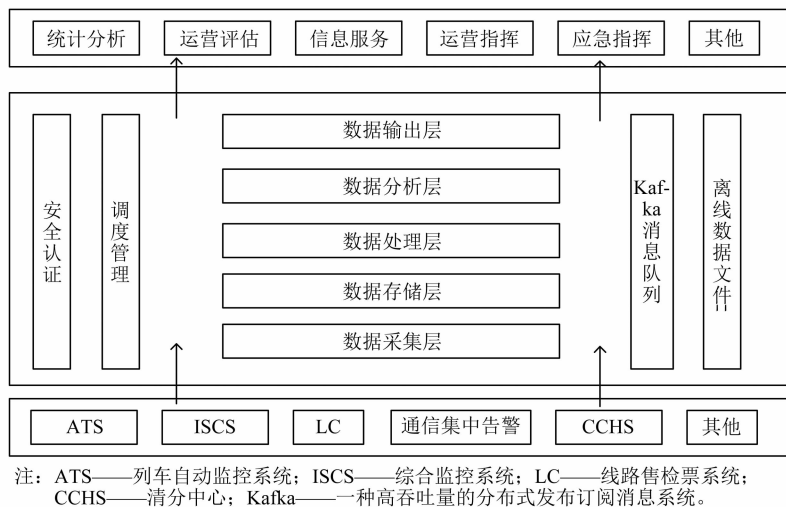


图 1 城市轨道交通大数据中心系统逻辑架构图

Fig. 1 Logic architecture of urban rail transit big data center system

城市轨道交通大数据中心系统包含采集层、存储层、处理层、分析层和输出层,各层功能主要如下:

1) 采集层:系统通过接口服务器与 ATS、ISCS、LC、通信集中告警、CCHS 线网清分中心等业务系统进行通信,对数据进行抽取、转换和校验。

2) 数据处理层:是大数据中心系统的核心,可将业务系统各类数据进行有效集成,满足海量数据管理需求。

3) 数据分析层:按照行车、设备、能耗、客流等专业划分进行数据挖掘、智能分析,从数据中发现

有价值的信息,以此作为预测、决策的数据支撑。

4) 数据输出层:主要由各开源大数据查询引擎构成,对外提供数据库查询服务。

2.2 数据处理架构

在苏州和青岛的轨道交通项目中,根据数据的流向,将大数据中心的处理架构设计为帖源层、基础层、汇总层和集市层。数据处理平台主要采用模块化、高可扩展的技术,如并行计算、并行装载、MPP 数据库、分布式存储等。应用平台获取大数据中心集市层的数据,展现方式采用基于 J2EE 的多层客

户/服务器模型。通过灵活的集成框架,不同的第三方插件、产品能够有效地集成到数据处理平台。

数据处理的核心区域为基础层、汇总层和集市层,其在整个数据架构中处于数据服务层,如图 2 所示。

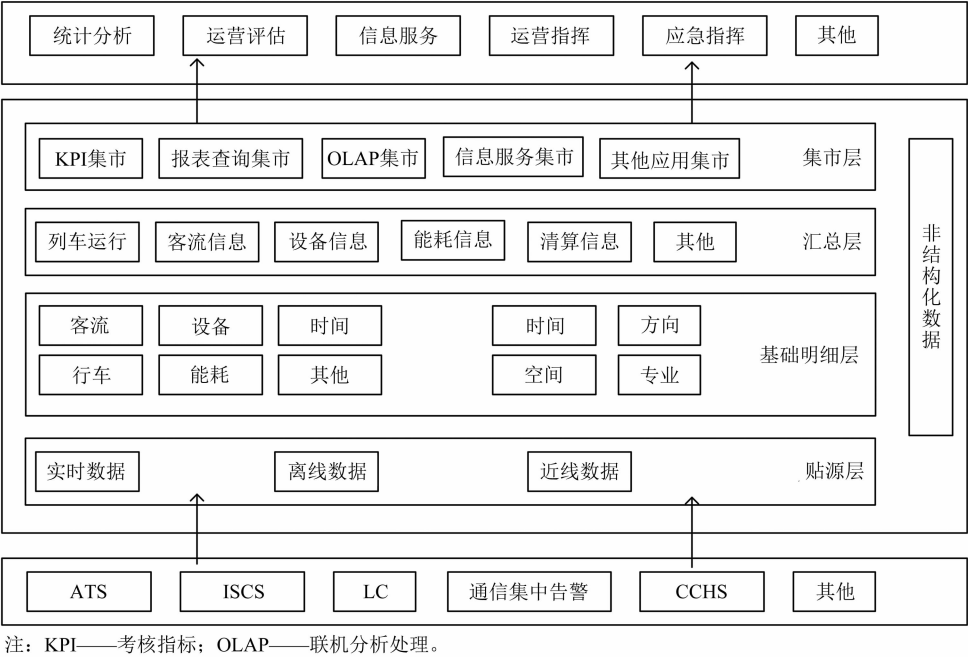


图 2 城市轨道交通大数据中心系统数据架构图

Fig. 2 Data architecture diagram of urban rail transit big data center system

1) 贴源层:实现采集到的文件数据到数据仓库的映射,为基础层数据的加工做好准备。

2) 基础层:是数据服务层中最重要的一个区域,按照数据标准的要求对贴源层数据进行统一加工和整合,存储明细粒度的历史数据区域,可为各个业务部门的不同业务需求提供一致规范的数据。同时,基础层数据可作为汇总层、集市层的数据源,并可直接向高级数据分析人员开放,进行深度灵活查询、数据挖掘和数据分析。

3) 汇总层和集市层:其数据是提供面向需求应用的、提供共享数据访问服务的公共数据。其数据流向是从基础层抽取数据,经过有针对性汇总加工后,满足上游应用的数据展示需求。

2.3 数据处理流程

为优化数据处理效率,将数据处理细分为实时数据流处理和离线数据流处理。对不同的数据流,根据其特点进行优化设计,利用数据库中不同的组件进行数据处理。如:对于实时数据,是采用 kafka 的方式将数据发送到处理层,再存储至 Redis(一种数据库)内存库;对于离线数据,是先将数据存入 HDFS,经大数据处理程序处理后再存入 MPP 数据库。

为提升任务管理的效率,考虑充分利用分布式系统的相关功能,如 Hadoop 中的 Map/Reduce 可以把一个任务分解为很多可以并行化处理的子任务,这些子任务被分配到不同服务器上并行计算,最后再把结果聚合到一起形成一个最终结果。

2.4 大数据分析

大数据分析的重点是对行车数据、客流数据、设备数据和能耗数据进行分析。

1) 行车大数据分析:主要功能是行车指标体系优化分析、运输计划调整分析、分时开行对数表分析、首末班车开行衔接分析、行车交路方案分析和停车方案分析。

2) 客流大数据分析:该项数据分析主要用于降低人均运输成本、引流提高运营收入、降低设备故障影响、大修计划安全评估、事故抢修及综合调度、应急故障方案、高峰集散方案、一日组织方案、特殊保障组织方案、降低建设成本、乘客行为分析与公共安全分析等方面。

3) 设备大数据分析:主要功能是可靠度分析、故障统计与回溯、智能维保、设备知识图谱分析和故障原因挖掘。

4) 能耗大数据分析:主要功能是空调通风能耗

分析、制冷系统能耗分析、牵引能耗分析、照明能耗分析、电梯能耗分析和能耗预测。

未来,大数据挖掘的可能发展方向为多专业相关性分析、客流预测、设备维修周期、状态监测和趋势预测等。

3 大数据中心建设工程实施要点

1) 制定数据源接口标准。大数据中心的数据来源于各条线路的各专业系统,连接、开发各专业间的接口和通信中间件十分重要。对于数据源的接口标准,建议在数据中心项目建设开始阶段就制定完成。

2) 保证数据的安全性。大数据中心是城市轨道交通的上层系统,有些城市甚至肩负着连接外部政府及互联网的重任,大数据中心系统的安全性至关重要。建议符合信息网络安全等保三级要求。在系统设计初期,建议请专业的信息安全咨询单位评估系统安全性,并严格按信息安全标准进行建设。

4 结语

本文分析了 Hadoop+Mpp 技术架构的优缺点。基于苏州轨道交通大数据中心项目的实践经验,分析了大数据中心的技术方案。苏州轨道交通的项目实践表明,基于 Hadoop+Mpp 架构的大数据中心建设方案能够达到预期效果。

参考文献

- [1] 王峰,雷葆华. Hadoop 分布式文件系统的模型分析[J]. 电信科学,2010(12):95.
WANG Feng, LEI Baohua. Modeling and analysis of Hadoop distributed file system[J]. Telecommunication Science, 2010(12):95.
- [2] 程莹,张云勇,徐雷,等. 基于 Hadoop 及关系型数据库的海量

数据分析研究[J]. 电信科学,2010(11):47.

CHENG Ying, ZHANG Yunyong, XU Lei, et al. Research on large-scale data processing based on Hadoop and relational database[J]. Telecommunication Science, 2010(11):47.

- [3] 田秀霞,周耀君,毕忠勤,等. 基于 Hadoop 架构的分布式计算和存储技术及其应用[J]. 上海电力大学学报,2011(1):70.
TIAN Xiuxia, ZHOU Yaojun, BI Zhongqin, et al. The technology and application of distributed computing and storage based on Hadoop architecture[J]. Journal of Shanghai University of Electric Power, 2011(1):70.
- [4] 陈梦杰,陈勇旭,贾益斌,等. 基于 Hadoop 的大数据查询系统简述[J]. 计算机与数字工程,2013(12):1939.
CHEN Mengjie, CHEN Yongxu, JIA Yibin, et al. A brief introduction Hadoop-based big data query system[J]. Computer & Digital Engineering, 2013(12):1939.
- [5] 李聪颖,王瑞刚,梁小江. 基于 HADOOP 的交互式大数据分析查询处理方法[J]. 计算机技术与发展,2016(8):134.
LI Congying, WANG Ruigang, LIANG Xiaojiang. An interactive processing method of analysis and query for big data based on Hadoop[J]. Computer Technology and Development, 2016(8):134.
- [6] 张雨,蔡鑫,李爱民,等. 分布式文件系统与 MPP 数据库的混搭架构在电信大数据平台中的应用[J]. 电信科学,2013(11):12.
ZHANG Yu, CAI Xin, LI Aimin, et al. Application of distribute file system & MPP database mashup architecture in telecom big data platform[J]. Telecommunications Science, 2013(11):12.
- [7] 吉增瑞. 基于 MPP 结构的计算机平台数据库管理系统设计技术探讨[J]. 计算机工程与科学,1998(71):117.
JI Zengrui. Discussion of the design technology of DBMS based on the computer platform of MPP architecture[J]. Computer Engineering & Science, 1998(71):117.
- [8] 王震华,田立军. 铁路一体化物联大数据运维平台的研究[J]. 铁路计算机应用,2019(2):26.
WANG Zhenhua, TIAN Lijun. Data operation and maintenance platform of railway integrated logistics of things[J]. Railway Computer Application, 2019(2):26.

(收稿日期:2021-12-10)

苏州轨道交通 S1 线——长三角一体化基础设施互联互通示范工程

S1 线为苏州市轨道交通第三轮建设规划的新建线路,全长 41.25 km,设站 28 座,起于苏州工业园区唯亭站,终于昆山花桥站,与上海轨道交通 11 号线衔接,为长三角一体化基础设施互联互通的示范工程,也是我国县域经济首条全城穿越的轨道交通线路。S1 线对增强苏州、昆山中心城区的辐射能力,引导城市空间布局优化,引领新型城镇化建设,促进交通、产业、空间一体化布局,实现长三角城市群的协同发展,有着极其重要的示范意义。

(来源:苏州市轨道交通集团有限公司)