构建“1+N大数据智能化运维平台”的具体实践

杨蕾蕾
身份证号码：410482198910254413

摘　要：“1+N”边缘计算大数据平台有效的提升了大数据平的数据处理能力，但同时也带来了边缘计算节点数据同步及软件高效运行之间的矛盾。其中大数据平台的千百节点及十余种软件统一高效管理，提升对外服务的效率，成为当前大数据平台运营亟待攻克的难题，也是本成果重点解决的问题。本成果提出构建1+N大数据平台智能化运维系统，从跨地域智能指标预测分析、跨地域智能日志分析、跨地域智能运维管理等三个层面解决了1+N大数据平台跨地域计算与运营能力亟待提升问题。

关键词：大数据；智能化

The specific practice of building “1 + N big data intelligent operation and maintenance platform”

Leilei Yang
Id No.: 410482198910254413

Abstract: The “1 + N” edge computing big data platform effectively improves the data processing capacity of big data flat, but also brings the contradiction between the data synchronization of edge computing nodes and the efficient operation of the software.Among them, the unified and efficient management of thousands of nodes and more than ten kinds of software and improving the efficiency of external services have become an urgent problem to be solved in the current operation of the big data platform, and also the key problem to be solved by this result.This achievement proposes to build the intelligent operation and maintenance system of 1 + N big data platform, which solves the urgent problem of improving the cross-regional computing and operation capability of 1 + N big data platform from three aspects: cross-regional intelligent index prediction and analysis, cross-regional intelligent log analysis, and cross-regional intelligent operation and maintenance management.

Keywords: big data; intelligent

一、项目背景

随着“全面实施国家大数据战略”的提出，以及大数据与人工智能技术的逐步成熟，中国移动G省公司大数据平台部承载了营业、计费账务、经营分析、财务审计等公司多项核心IT应用。为满足大数据业务日益增长的诉求，G公司大数据平台集群规模不断扩大，服务组件不断扩充，造成运维管理难度持续增加等困难，进而导致平台运营效率持续走低。为解决资源矛盾，只能单纯的依赖资源横向扩张来暂时解决问题，而增加资源又持续增大了运维难度，这要求必须从根本上进行技术改造才能从根本上改变现状。

基于上述背景，为提升现有大数据平台效能，技术部门规划构建了“1+N边缘计算大数据平台”。1+N边缘计算平台的快速落地，对重新定义运营商及大型央企IT系统架构具备非常重大的示范及指导意义，具备与国际最先进IT分布式架构对标的能力。

二、现有大数据平台面临的问题及挑战

为满足大数据业务日益增长的诉求，G省移动大数据集群规模快速扩充，特别是大数据边缘计算节点的加入，带来的是大数据平台集群规模庞大、跨地域部署、生态服务组件多、运维管理复杂等实际。随之而来的是营运瓶颈逐步显现，经过对大数据平台运营各方面的深入分析，总结出大数据平台运营主要面临的三方面问题与挑战：

1.性能及故障定位问题难

1+N边缘计算大数据平台软件涵盖包含HBase、

Hive、HDFS等十余种类软件，彼此之间关联交互情况较为复杂，需结合所有组件的运行情况综合判断当前大数据平台的性能或故障问题的根因，然而通过人工方式逐个分析千百个运行指标的工作量大、耗时长，同时结合多种组件复杂交互，难以找到最有效的解决方案。

2.提升大数据平台运行质量难

随着平台逐年来高负载运行，缺少对平台对年运行的隐藏问题及时发现，平台得不到及时有效的优化，往往累积到出现故障再事后弥补，会直接影响到客户感知和经营分析效率，而由于指标和日志数据的体量极为庞大，通过海量的运维数据难以精准定位改善性能的有效措施，往往单纯依赖服务器横向扩张提升性能。

3.日常运维管理难度大

生产系统稳定运行，节点或服务状态异常改变成为日常关注的重点，及时自动发现问题的手段显得尤为重要，然而大数据平台规模已达到成千百个数据节点，特别是异地部署的边缘计算节点多。平台组件包含多种软件类型，运营管理成本较高。同时对已经发生过的各类复杂技术问题，精准快速检索历史解决方案没有有效方法。

由于人工智能技术正加速探求可落地的应用场景，这为解决企业大数据平台当前运营问题提供了解决方案基础。针对大数据平台运营期间的集群规模庞大、生态服务组件多、运维情况复杂等三类问题，基于上述问题与挑战我们提出构建的大数据平台智能化运维系统，结合大数据平台日常运营管理方法进行深入分析，全方位提升大数据平台运营能力。

三、构建“1+N大数据智能化运维平台”

基于现有大数据平台面临的问题及挑战，技术团队提出构建的“1+N大数据平台智能化运维平台”。新平台从智能指标预测分析、智能日志分析、智能运维管理等三个层面解决了大数据平台运营能力亟待提升问题。通过事件驱动发现异常事件，自动分析事件根因，对未来可能发生的平台问题及时预警，并结合解决方案智能推荐形成系统内部智能运维体系闭环。采用AI算法挖掘海量日志中所蕴含的模式，通过聚类分析，将大量日志原文转化为少量日志模式，大幅缩减人工过滤时间。通过知识图谱方式自动呈现各组件服务关系。将日常技术文档形成知识库，借助AI语义识别技术的对知识相似匹配，自动推荐问题解决方案。

新平台中智能化运维系统采用了轻便灵活的服务端技术Flask作为主体应用技术栈，流处理技术Kafka进行日志收集，采用MongoDB完成日志数据存储，用Neo4J作为知识图谱的数据存储，MySQL作为事件数据存储，结合自主研发核心算法完成了智能化运维系统的研发。

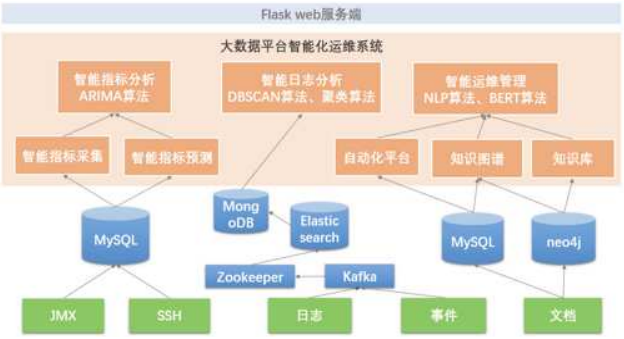

图1 “1+N”大数据智能化运维平台架构图

1.智能化指标分析

通过智能化指标分析对大数据平台的1200台节点，包括BOSS、CRM、财务集群、经营分析集群和收入保障集群等全部Hadoop各组件性能指标及主机指标进行机器学习，并对全量指标进行时间序列分析和预测，提供对未来将要发生的性能瓶颈和故障风险的预测能力。同时对全部指标形成多维对比分析，挖掘对比各指标的关联关系。将指标采集值、预测值以定制化报表形式对比、分析、呈现，并可以做原始数据导出查看。其中采集部分主要采用JMX接口数据调用，控制台数据获取和日志分析的方法，分析预测部分针对全量指标采用ARIMA算法。

2.智能日志分析

通过智能化运维系统监控Hadoop集群各组件生成的日志，作为数据源进行故障分析定位，并结合指标数据进行系统画像。接入BOSS集群、CRM集群、财务集群、收入保障集群和经分集群的各组件日志，包括：Namenode、Datanode、Journalnode、QuorunPeerMain、Resourcemanager、Jobhistoryserver、HMaster、Regionserver等，以及系统日志如用户登录日志等。智能化日志模块主要采用了基于注意力的相似度聚类算法及DBSCAN算法，将大量的日志原文转化为少量的日志模式，并提供对应模式在日志原文中的占比，大大减少了人工过滤时间，快速定位故障根因。本成果采用了文本切片和正则表达式的方式进行信息提纯，去除大量噪音信息，只保留描述性的文本信息，保证了信息的可靠性。

3.智能化运维管理

（1）自动化运维。通过自主开发的自动化平台执行日常的Hadoop组件、服务、集群、节点等的定期巡检工作，以及系统异常时组件、服务、集群、节点等的启动停止等工作。自动化巡检时发现状态异常，及时提醒运营人员及时处理异常。并可记录全部巡检、启停的事件、

日志和相关信息。提供历史数据为运维做数据依据。可提供Hadoop核心组件的配置文件、关键文件的定期备份、远端备份、历史版本情况和恢复。

（2）知识图谱与知识库。知识图谱模块可展示全部大数据平台Hadoop集群、服务、组件、节点之间的关联关系，配置信息与指标采集分析结果全量对应，可实现从CMDB图谱到生产监控的跳转。知识库部分可上传技术文档或知识记录，也可进行问答对的填写，实现针对知识库内包含的内容进行语义解析和查询能力。

四、应用情况及总结

系统改造上线至今近1年时间，已陆续在1+N跨域边缘计算大数据平台进行部署实施。通过智能指标预测、智能日志分析以及智能运维管理等能力，提升边缘生产系统大数据平台健康度35%，提升平台运营效率为45%，提高了其日常业务分析和对外服务能力。节约扩容服务器约200台，累计节约千万扩容成本，实现了大数据平台精细化运营管理，为AI技术应用落地树立典范。系统采用了业界最先进技术AICDE的AI人工智能与大数据技术，所采用的技术方法论对于全国其他应用都有良好的可移植性，可面向全国推广。

图2　系统运维关键资源运维监控图

AICDE人工智能、大数据、云计算等已成为国家战略。同时，互联网企业引领IT技术变革。面对机遇与挑战，通过对人工智能与大数据等技术发展战略合理规划与落地，本成果通过AI人工智能技术大幅提升了大数据平台运营能力，保障生产系统的稳定支撑，高效运营，满足企业的客户运营要求。从需求驱动角度，把握当前互联网形势下客户对产品、服务的需求变化，适应新的市场和客户需求，做到系统基于客户需求的灵活快速响应。

本项目通过实现大数据平台的智能化运维系统，同过智能指标预测与分析，智能化日志分析、智能化运维管理等能力落地应用，推动了技术研发与生产技术的全方面结合，通过投身高科技技术研发与应用，为整个技术行业应用树立良好典范，积极响应国家号召推动高新技术得到良好应用发展。

参考文献：

[1]冯永，钟将，王茜，李学明.共智融合的大数据智能化人才培养研究与实践[J].中国电化教育，2021（04）：16-25.

[2]王国胤，刘群，夏英，胡军，马彬，纪良浩.大数据与智能化领域新工科创新人才培养模式探索[J].中国大学教学，2019（04）：28-33.

[3]邓松，岳东，朱力鹏，胡斌，周爱华.电力大数据智能化高效分析挖掘技术框架[J].电子测量与仪器学报，2016，30（11）：1679-1686.