第 3 卷◆第 3 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

大数据驱动的云平台智能运维系统研究

黄立群　陆鋆
中移(苏州)软件技术有限公司
DOI:10.12238/acair.v3i3.15581

[摘　要] 云计算是信息时代的一个重要产业支柱,云计算平台为各行各业提供云计算服务。本文基于大数据技术和运维领域知识,设计针对云计算平台的运维系统。运维系统监控云计算平台的机器和服务并提供集群监控、异常发现、异常分类、实时告警、权限资源管理能力。运维系统为云计算平台的稳定提供保障。
[关键词] 云计算；大数据；运维领域
中图分类号：G633.67　文献标识码：A

Design and Research of an AI Operation and Maintenance System of Cloud Compute Platform Based on Big Data

Liqun Huang　Yun Lu
China Mobile (Suzhou) Software Technology Co., Ltd

[Abstract] Cloud computing is an import industry Pillars in present information age,cloud computing platform is well-serving in various of industries.the paper design an operation & maintenance system for cloud compute system by big data technology and professional operation & maintenance knowledge. Operation & maintenance system can well work for cloud computing, such as service and machine monitoring, anomaly detecting, anomaly clustering, alarm on time and manger permission resource.Operation System provide guarantees for the stability of cloud compute platform.
[Key words] cloud computing; big data; operation & maintenance domain

引言

云计算[1]为各行各业的用户提供可靠的各种资源服务,它是一种一切皆为服务的模式。为应对日益增加的数字化需求和全球竞争,云计算技术的创新和应用已经成为社会发展的重要支柱。大数据技术是一种处理海量数据,从庞大的数据中挖掘出具有价值的信息的技术[2]。云计算可以为大数据提供算力、存储支持,大数据技术能够处理云计算服务产生的庞大的数据,挖掘数据中的隐藏价值,通过这些数据提升云计算平台的运维质量和效率。

1 运维系统需求分析

云计算平台是一个分布式系统,云服务厂商在进行基础设施建设时会分可用区建设,一个可用区的物理节点数会达到数千台,每台物理节点能虚拟化出多台虚机,如此多的服务器其生产运维依托大量的人力资源,且存在设备状态评估错误,业务故障发现不及时,故障处理周期长问题。

云计算平台的智能运维系统以提升云计算服务的运维效率为目标。通过大数据技术和运维领域知识构建一个包括运维分析、资源权限管理、业务告警等模块的系统。云计算运维系统的核心需求如下：

(1)运维分析：运维系统通过云平台服务产生的业务日志以及部署云平台服务设备的性能指标实现设备、服务监控；根据服务日志链路和设备状态进行问题定位,异常分析。(2)资源、权限管理：围绕云平台产生的日志和部署服务的设备指标以及运维系统的功能进行资源划分。运维系统需要对资源进行分类,分级管理,针对不同角色的用户配置不同的资源访问权限。(3)业务告警：根据云计算服务的日志数据和部署云计算服务设备的性能指标数据做实时告警,及时发现故障。(4)数据管理：云计算服务产生的海量日志和部署服务的设备产生的指标数据需要持久化,且运维平台能对这些数据进行查询、聚合。

2 运维系统设计

2.1架构设计。运维系统架构如图1所示,系统采用了分层架构,自底向上包含了云服务层,数据源层,基础设施层,数据治理层,算法层和业务层。云服务层包含了云主机、云数据库、云存储、云网络等由云计算提供的服务,该层是运维系统的运维对象,其服务产生的数据是运维系统的数据源。数据源层包含了云计算服务的日志和部署云计算服务的设备的性能指标(cpu利用率、内存使用率、磁盘使用率、磁盘io、网络io等)。基础设施层包含了kafka、clickhouse、kubernetes、redis、mysql、haproxy、keepalived等核心中间件,为上层的数据治理、算法和业务层提供保障。数据

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

治理层通过Flink实现数据ETL能力,通过clickhouse存储海量数据。算法层通过聚类算法处理海量数据实现机器状态和服务状态评估。业务层主要实现了云服务性能监控查看、设备性能监控查看,服务和设备故障告警、故障分析等能力。

图1 基于大数据的云平台智能运维系统架构

2.2基础设施。云计算平台的运维系统依托于大数据能力,因此系统的基础建设和中间件选型尤为重要。以云主机指标为例,单可用区指标上报频率为百万条/周期。为了应对如此巨大的数据量和满足业务的实时性。在系统设计时选择Kafka消息队列作为指标传输通道,Clickhouse分布式数据库做指标数据持久化。业务等对事务要求较高的数据通过mysql存储。使用Kubernetes容器管理平台来管理计算集群。运维系统服务部署在kubernetes中,通过kubernetes的特性保证业务高可用[3]。

2.3关键技术。

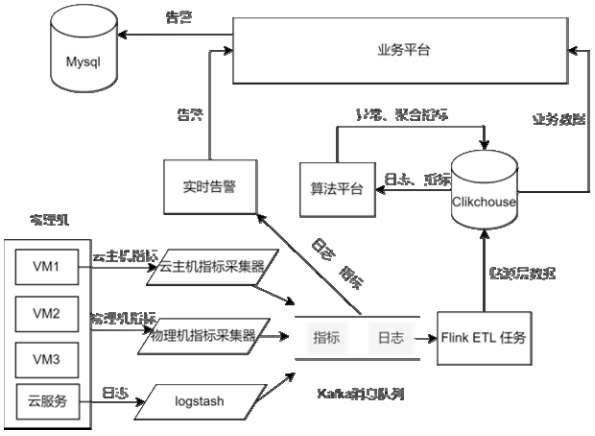

图2 大数据处理数据流图

2.3.1基于大数据技术的云计算平台运维系统实时数据处理。对于运维系统而言其服务对象的指标实时性和准确性极其重要,所以对指标、日志等数据的实时处理是一项基础且重要的工程。数据处理流图如图2所示,首先云计算平台的指标数据采集是通过指标采集插件采集的,日志数据是通过logstash采集,采集完的数据发送到kafka消息队列中。考虑到云计算平台的海量数据,消费端是通过Flink技术实现的指标和日志的ETL任务。Flink处理的指标、日志数据最后持久化到Clickhouse中,作为运维系统数仓的ods层。

算法平台通过数仓总ods层的数据通过聚类算法识别云平台的异常种类以及通过指标聚合规则对指标进行不同维度的汇总。实时告警模块对日志和指标进行实时告警,告警数据发送到业务平台,业务平台根据策略决定告警的后续处理。

运维系统需要处理的数据种类较多,这里以云主机的指标实时入库和5分钟汇总为例子,实现指标数据的读取、转换,过滤,聚合,入库等。数据处理流程如图3所示。

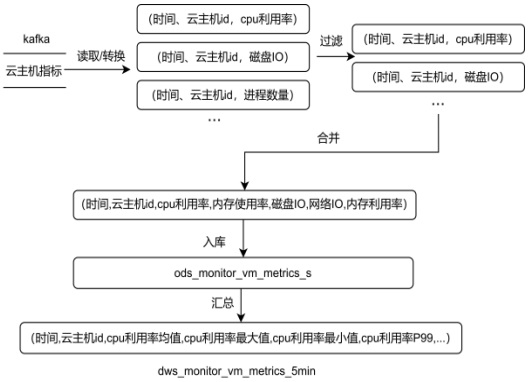

图3 数据处理流程图

(1)读取/转换。读取kafka消息队列中的云主机指标。通过实现Flink的Deserializer接口可以将kafka中Json字符串转换成定义好的Java对象。(2)过滤。步骤(1)中的每个Java对象是一台云主机的某一个时刻的某个指标。运维系统将不需要处理的云主机指标通过filter()算子过滤。(3)合并。通过Flink的TumblingEventTimeWindows窗口函数将同一周期内同一云主机的不同指标合并成一个Java对象。(4)入库,将处理好的指标数据写入到Clickhouse中的云主机指标表ods_monitor_vm_metrics_s中。(5)按时间维度汇总,利用Clickhouse的物化视图的特性,计算云主机指标表中的指标数据的最大值、最小值、均值、P99分位点,并将结果存入dws_monitor_vm_metrics_5min表中。

2.3.2基于大数据技术的云计算平台运维系统数据分析技术。数据分析技术是观察和总结数据的特性,目的是提取数据中有用信息,做出推断,并得出结论。数据分析技术有基于大数据的分类模型和回归模型,有根据数据的相似模式进行归纳到一簇的聚类模型,还有其他的关联规则分析模型等。

本文以评估部署云服务的设备的状态为例,阐述如何通过高斯分布模型判断设备的状态是否异常。云计算平台的运维系统某个周期采集的设备指标如公式1所示。表示集群中第i个设备的性能数据。

$$x=\{x_1,x_2\cdots,x_n\} \tag{1}$$

每台设备的性能指标有多个维度,如公式2所示,公式中$\{x_i^1,x_i^2,x_i^3\cdots,x_i^7\}$表示第i台设备的cpu利用率、内存利用率,内存可用空间,磁盘空间利用率、磁盘可用空间,磁盘io,网络io共7个指标。

$x_i = \{x_i^1, x_i^2, x_i^3 \cdots, x_i^7\}$ (2)

根据公式1和公式2，可以算出公式分布公式中的均值 μ 和标准方差 σ。两者的计算公式分布是公式三和公式4。

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j \tag{3}$$

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i^j - \mu^j)^2 \tag{4}$$

计算某个周期内某个节点的状态是否异常，通过公式5高斯分布概率公式得出p(x)

$$p(x) = \prod_{j=1}^{7} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

当 $p(x) < \varepsilon$ 时，说明该节点状态异常，运维人员应该重点关注该节点状态，通过观察公式5在多个周期数据中的计算结果的节点状态分布情况得出。

3 功能模块

3.1告警模块。告警模块是任何运维系统的基础功能。该模块包括告警策略管理、告警管理、告警触发功能。具体功能架构如图4所示。其中告警触发是基于大数据和规则引擎实现。以部署云服务设备的性能指标告警为例。研发或运维在运维系统中配置了磁盘利用率大于80%或者内存利用率大于70%的告警策略。基于Flink开发的实时告警模块会周期性的更新系统中的告警策略。并通过规则引擎将策略编译成disk_usage>0.8||cpu_usage>0.7表达式。运维系统每个周期采集的设备指标都会发送到kafka消息队列，Flink告警任务会实时消费设备的性能指标。判断该周期内哪些设备的指标满足上述表达式，满足表达式就触发告警到告警后置处理模块，该模块会将多条告警按规则进行合并成一条告警，防止产生告警风暴，同时将合并的告警根据告警策略通过消息发送给对应的运维人员。如此运维便可实时知晓集群中哪些节点的内存利用率或者磁盘利用率超过阈值，对该节点进行运维处理。

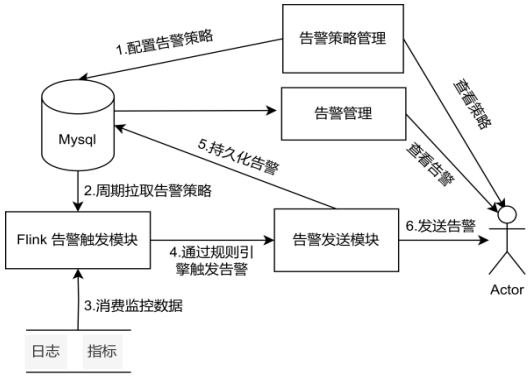

图4 告警功能架构图

3.2业务异常识别。云计算平台的核心是云服务业务，云服务业务系统是否正常直接影响了用户体验。运维系统通过采集云服务的日志、服务接口耗时、接口qps等指标并处理成数据 $X = \{x_1, x_2, x_3 \cdots, x_n\}$。运维系统通过聚类算法[4]模块将不同类型异常分到不同类簇中并记录类簇中心 $C = \{c_1, c_2, c_3 \cdots, c_k\}$，聚类算法如图5所示。公式中k代表有k类异常。$c_k$ 表示第k类异常的类簇中心。当有新的异常出现时运维系统会计算异常数据 X_{new} 和所有类簇中心的距离，判断其属于哪种异常，并将相应告警发送到研发和运维人员，研发和运维根据运维平台的日志模块和指标模块再次确认系统异常原因，缩短了异常发现和定位的时间。

```python
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score


def kmeans_clustering(data, n_clusters=3, random_state=42):
    scaler = StandardScaler()
    data_scaled = scaler.fit_transform(data)
    # 通过KMeans进行聚类
    kmeans = KMeans(n_clusters=n_clusters, random_state=random_state)
    kmeans.fit(data_scaled)

    return kmeans.labels_, kmeans.cluster_centers_

def load_data(file_path):
    """
    从数据文件中载入数据
    """
    return np.loadtxt(file_path, delimiter=',', skiprows=1)

import sys

if __name__ == "__main__":
    if len(sys.argv) < 3:
        print("Usage: python kMeans.py <data_file> <n_clusters>")
        sys.exit(1)
    data_file = sys.argv[1]
    n_clusters = int(sys.argv[2])
    print(f"Loading data from {data_file} with {n_clusters} clusters...")

    data = load_data(data_file)

    labels, centers = kmeans_clustering(data, n_clusters=n_clusters)
    print("类簇:", labels)
    print("列簇中心:", centers)
    print("轮廓系数:", silhouette_score(data, labels))
```

图5 业务异常分类算法

3.3权限、资源管理。云计算平台运维系统的使用对象是研发、运维以及管理员。因此需针对不同角色的用户设置不同权限。本系统使用的资源、权限管理模型是RBAC模型[5]。运维系统会对用户分配不同角色，每个角色对相同资源的访问权限是不一样的。比如研发是无法对物理机进行启停操作的，但是运维人员可以进行启停。

4 总结和展望

本文针对云计算平台的运维需求，设计了基于大数据能力的智能运维系统。该系统实现了云计算平台设备监控；设备、服务异常发现和识别；实时告警；权限资源管理以及海量数据的管理。未来，将融入大模型能力，构建一个通用性更强，运维效率更高的智能运维平台，助力云服务厂商的高速、健康发展。

[参考文献]

[1]彭好佑.云计算综述[J].福建电脑,2018,34(1):1-2,13.

[2]于骞翔,李擎,李琳琳.面向大规模工业生产过程的数据驱动故障诊断方法综述[J].工程科学学报,2025,47(4):780-793.

[3]盛乐标,游伟倩,张予倩,等.Kubernetes集群的高可用与负载均衡设计[J].电子技术与软件工程,2019(7):1-3.

[4]马燕,孙鲁鹏,刘家洛,等.基于Kmeans聚类对14个新疆陆地棉萌发期耐碱性评价[J].种子,2024,43(10):54-63,73.

[5]魏巍.基于RBAC模型云资源管理系统访问控制权限的设计与实施[J].科学技术创新,2025(2):122-125.

作者简介：

黄立群(1992--),男,汉族,江苏盐城人,硕士,软件研发工程师。