PROCEEDINGS OF SCIENCE

Operation and maintenance analysis platform at IHEP

Qingbao Hu[a,*] **and Lu Wang**[a] **and Xiangwei Jiang**[a] **and Wei Zheng**[a]

aInstitute of High Energy Physics, Chinese Academy of Sciences, Beijing, China

E-mail: huqb@ihep.ac.cn, wanglu@ihep.ac.cn, jiangxw@ihep.ac.cn, zhengw@ihep.ac.cn

Abstract. As the scale of computing facilities continues to grow and the computing environment becomes more and more complex, the difficulty of operation and maintenance of large-scale computing clusters is also increasing. Operation and maintenance methods based solely on configuration management automation technology cannot quickly and effectively solve various service failures in computing clusters. It is urgent to adopt emerging technologies to obtain comprehensive cluster operation and maintenance information, integrate monitoring data from multiple heterogeneous sources, and comprehensively analyze anomalous patterns in monitoring data. Based on the results of data analysis, it becomes possible to locate the root cause of service failures and help computing clusters quickly restore services. To provide a more stable cluster operating environment, IHEPCC combined big data technology and data analysis index tools to design and implement the operation and maintenance analysis toolkit (OMAT) as an open framework, which includes data collection, correlation analysis, anomaly detection, and alerting and other functions. This report introduces the architecture, processing capabilities, and some key functions of OMAT. Combined with the processing flow of monitoring data, it introduces the specific implementation of the system in data collection, data processing, data storage, and data visualization. The current OMAT platform has been applied to multiple cross-regional computing clusters including IHEP, covering about 5000 nodes. The collected information includes node status, storage performance, network traffic, user operations, account security, power environment, and other operation and maintenance indicators to ensure the computing cluster's performance and stable operation.

International Symposium on Grids & Clouds 2022 (ISGC 2022)
21 - 25 March, 2022
Online, Academia Sinica Computing Centre (ASGC), Taipei, Taiwan

[*] Speaker

PoS(ISGC2022)011

1. Introduction

The Institute of High Energy Physics of the Chinese Academy of Sciences (IHEP) is the largest and most comprehensive fundamental research center of high-energy physics in China. The major research fields of IHEP are particle physics, astrophysics and astroparticle physics, accelerator physics and technologies, radiation technologies, and their application[1]. At present, IHEP hosts and participates in more than 15 experiments in the field of high energy physics. such as BESIII, JUNO, HXMT, LHAASO and so on. IHEP also operates the BEIJING-LCG2 Tier-2 site and is actively involved in the computing of Atlas, CMS and LHCb. The Computing Center of the Institute of High Energy (IHEPCC) is at the center of IHEP's entire scientific, administrative, and computing infrastructure, and provides scientific computing services for the above experiments. In recent years, IHEPCC has launched a processing platform based on a multi-center design scheme, named "One Platform", which is intended to integrate computing resources distributed in different regions and provide more powerful computing capabilities for experiments. There are two large data center sites on this platform, one is IHEP Site located in Beijing with about 47k CPU cores and 200 GPU cards, and the other is Dongguan Site with more than 30k CPU cores and 80 GPU cards. The platform also has some sites located in Daocheng and Jiangmen. There are also some university site resources distributed in different places, such as Shandong University, Lanzhou University, etc. All sites in the platform have a very good network connection to IHEP, and are fully managed by IHEPCC.

Monitoring of One Platform for Multiple Data Centers

Sites	CPU Resources (CPU Cores)	CPU Resource Utilization	Disk Storage Capacity	Data Storage	Completed Jobs HTC&HPC	Job Run Time (CPU Hour)
IHEPCC	47,676	68.81%	70.94 PB	51.34 PB	3,473,456	5,394,639
DongGuan	32,080	47.93%	6.68 PB	4.02 PB	9,533	2,375,305
DaoCheng	3,392	21.05%	4.27 PB	3.77 PB	3,009	141,434
CSNS	5,572	6.644%	802.7 TB	400.8 TB	1,001	56,866
SDU	1,168	32.75%	352.9 TB	251.9 TB	1,071	64,769
USTC	3,714	31.76%	1.17 PB	876.5 TB	9,513	198,233
LZU	1,768	13.96%	341.8 TB	279.4 TB	3,269	36,869

Figure 1: Monitoring of One Platform for Multiple Data Centers (2022-08-15 ~ 2022-08-21)

As the scale of computing facilities continues to grow and the computing environment becomes more and more complex, the difficulty of operation and maintenance of large-scale computing clusters is also increasing. Operation and maintenance methods based solely on configuration management automation technology cannot quickly and effectively solve various service failures in computing clusters. To provide a more stable cluster operating environment, IHEPCC combined big data technology and data analysis index tools to design and implement the operation and maintenance analysis toolkit (OMAT) as an open framework, which includes data collection, correlation analysis, the strategy of monitoring data, anomaly detection, alerting and other functions.

Using the above functions, OMAT improves the quality of IHEPCC cluster operation and maintenance, and ensures the stability of computing services. As shown in figure 2, in the job scheduling scenario, OMAT detects anomaly services of worker nodes in time and corrects the job distribution strategy in time to reduce the number of job failures.

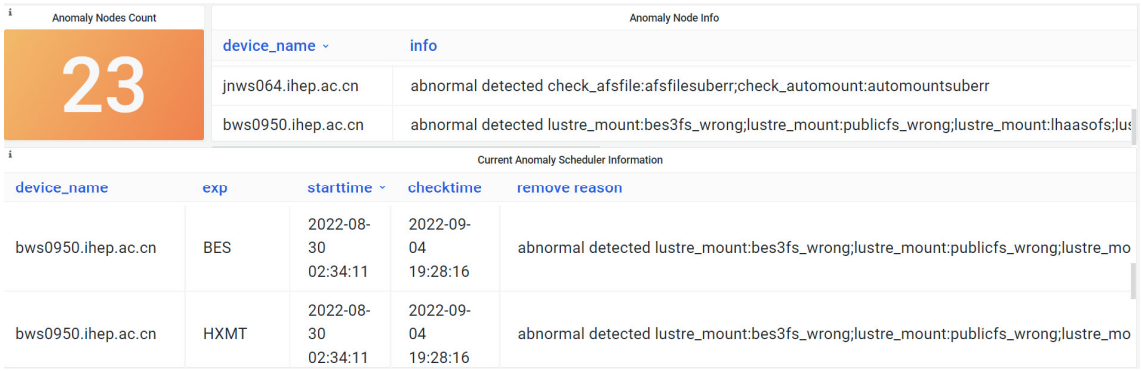

Figure 2: The bes3fs mount point of the bws0950 node is anomaly, and the scheduling policy is adjusted so that the bws0950 node does not receive the jobs of the BES experiment

As shown in figure 3, in the data access scenario, OMAT records the operation frequency of each user accessing the file storage server from the worker node, such as open, read, seek, etc., and calculates an anomaly score based on the anomaly detection model, and supports the query of the ranking results. When the performance of the file server is degraded, alarm information is sent to the storage administrator in a timely manner, and the storage administrator can view the anomaly value ranking of user access behaviors and quickly locate the cause of the fault.

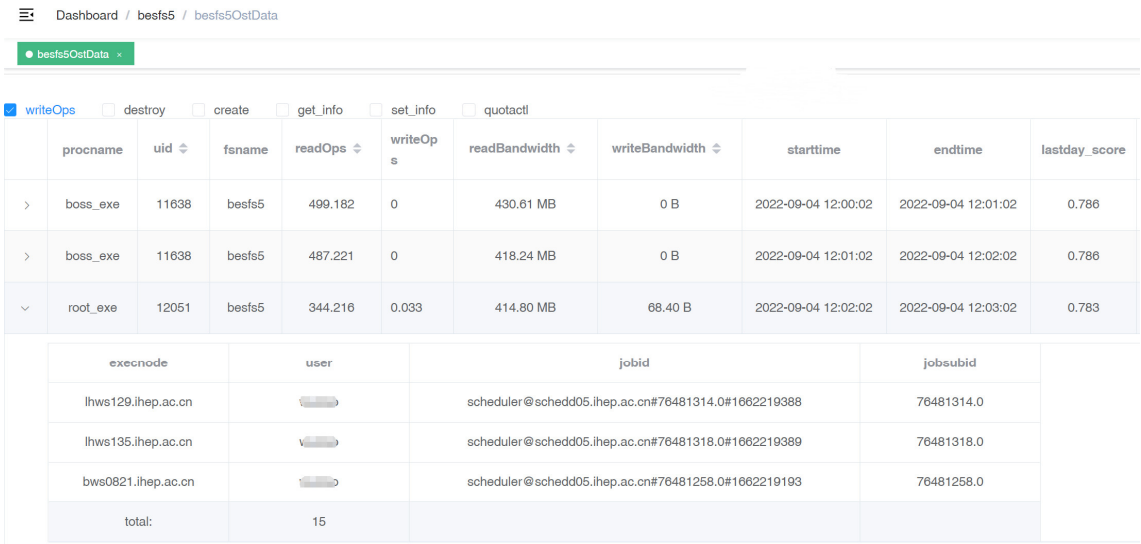

Figure 3: The real-time indicators of readops and writeops are recorded when a user with uid 12051 accesses the besfs5 file server. The job ids that produced the above access behavior, and which nodes are running these jobs. The anomaly scoring value of the current user indicator based on the anomaly detection model.

As shown in figure 4, in the multi-site operation and maintenance scenario, the online resource information and job running information of each site are collected and analyzed, and information such as the running status and resource utilization of the site is displayed.

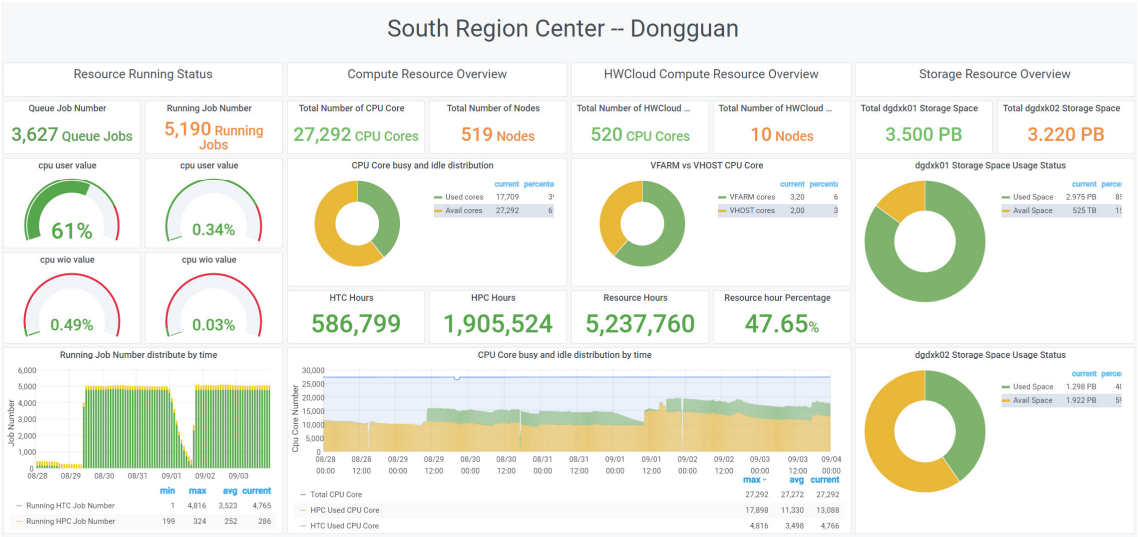

Figure 4: Monitoring information such as the number of job, resource utilization, and storage utilization at the Dongguan site (2022-08-28 ~ 2022-09-04)

This paper introduces the architecture, processing capabilities, and some key functions of OMAT.

1. Monitoring challenges

The sites that constitute the platform have, in an incremental and gradual manner, deployed several systems to address their monitoring needs. This has occurred both for the distributed sites and in particular at the IHEPCC: 1) Nagios[2] is used to check almost 10k computer center services for system and service level anomalies; 2) Syslog-ng[3] is used to collect user login logs of login nodes; 3) Ganglia[4] is used to show performance metrics covering 5k nodes; 4) Cacti[5] is used for network traffic monitoring; 5) Self-guard local monitoring on a server for system hardware (IPMI) events, temperature limits, Uninterruptible Power Supply (UPS) events; 6) Elasticsearch[6] and Kibana[7] are used for engineering infrastructure monitoring. Site administrators may also develop special monitoring tools for hardware and operating system or service level software according to some specific needs. The quantity and diversity of the monitoring tools have a profound impact on their effectiveness, at least in part because the system operators lose oversight, and their results can be combined only through visual inspection. It is difficult to use them as soon as the classic monitoring assumes the creation of a computer center control room with many monitors and at least two people to check all screens and react if needed. It is also difficult to use these programs for analyzing data and making decisions as there is no single API and single database for monitored data. All systems have their own bespoke data format, like rrd, sql tables, nosql indexes, text, JSON, and others. The monitoring systems that are used do not provide alarms that take event correlations into account and they have only simple threshold limits mechanisms to trigger alarm events. And finally, it is difficult to add new monitored parameters if they are not implemented by these tools.

1. Goals of the OMAT platform framework

Traditional operation and maintenance technologies cannot guarantee service quality. A more powerful maintenance analysis platform is needed. The primary goal of the monitoring system is not only to see nice graphs or alarms but rather to be a part of the self-management system. The new operation and maintenance monitoring system will be designed to meet the following functions.

3.1 Basic data management platform

As the first step in the operation and maintenance analysis platform, robust databases must be created to accumulate all information about resource site infrastructure, events, logs, and status, with the ability to continuously acquire the latest data in these areas. This part is used to collect as comprehensive monitoring data as possible, as the eyes and ears of the operation and maintenance robot, which the platform likened to. This component needs to meet the following characteristics: 1) Economical: Reuse existing operation and maintenance tools to save costs and support better maintainability. 2) Openness: Fast and flexible access to newly added monitoring data formats or data sources. 3) Unified interface: unified data aggregation, unified data sharing. 4) Reliability: Filter raw data to ensure data accuracy, and trigger alarms for missing data collection to ensure data integrity. 5) Low Latency: Process data quickly and provide near real-time data output. 6) Shareability: Multiple applications and multiple businesses can use data through a common "Data API". 7) Practicality: The data contains rich dimensional attribute information, which includes machine location, machine owner, machine power consumption, current visitors, etc. Based on these additional dimensional information, the data can support more monitoring scenarios. 8) Security: Control access to sensitive data fields, and support resilient data storage and archive storage. Based on the features listed above, the data management platform stably provides preprocessed structured data for the operation and maintenance platform

3.2 Intelligent analysis and decision-making platform

The management of resource sites by IHPECC integrates many technologies, including virtualization technology, automatic system deployment, distributed computing, access rights management, high throughput networking, highly reliable uninterruptable power systems, cooling systems, etc. The management system is highly redundant, and thus a failure of one subsystem will not be a problem. But the degraded system needs to be repaired as soon as possible to go back to its fully redundant state. The analysis and decision-making platform needs to quickly assess the stability of the site based on the collected data, and make corresponding decisions after identifying anomalies based on the accumulated expert operation and maintenance knowledge base or machine learning model. This part is used to implement the analysis, as the brain of the operation and maintenance robot. This process needs to meet the following characteristics: 1) It supports two data processing modes: calculation of background analysis and calculation of immediate response, which can analyze historical events and real-time analysis of newly collected data. 2) Supports data association analysis between the same data source or different data sources, including the association between different metrics, the association between metrics and events, the association between different events, the association between multiple metrics and an event, correlation between multiple events and fault propagation paths, etc. 3) Supports the identification of anomaly indicators based on machine learning models, including classic regression algorithms,

clustering algorithms, and classification algorithms. 4) Quickly identify and diagnose faults based on the known fault library, and perform fault feature deduction based on unknown faults combined with machine learning. Based on the above features, the timeliness of data analysis and the accuracy of decision-making results represent the core capabilities of the operation and maintenance analysis platform.

3.3 Agile automation control platform

A robust maintenance and operation management platform needs to have the ability to automate the management and use the issuing assertions that grant management rights and implementation methods to adjust the stability of the computing service of the site in a timely manner. This part is the embodiment of the platform's execution capabilities, similar to the hands and feet of a monitoring robot. There are various processing situations that may occur based on the analysis of the results of analyzing the monitoring data. These results may be the description of a common event or the status change information, which needs to be notified to the administrator; it may be a collection of alarms from many information sources, which needs to be classified and merged, it just needs to push urgent anomaly information to the administrator; it may also be a cluster fault identified based on anomaly characteristics or an expert experience database, which needs to be processed automatically according to a defined process; or a cluster fault without a processing plan; send an alarm and handle it manually. In this platform, the following functions should be satisfied: 1) Real-time information push function, ability to receive new push requests from multiple channels, support multiple message push methods, and be able to merge related alarm information; 2) The cross-host remote control function can quickly adjust the business strategy of the target server based on the analysis results of the monitoring platform, in order to reduce the impact of failures on-site service quality, and improve site resource utilization. For example, a machine has a job error due to an anomaly service, and the scheduling policy needs to be modified in time to stop sending jobs to the node.

3.4 Standardized intelligent maintenance and operation management process

The optimal effect of operation and maintenance analysis is to follow a set of standardized site operation and maintenance procedures to form a complete closed-loop of fault handling. As shown in figure 5, the site monitoring field is mainly divided into the following four parts, the bottom-level infrastructure and hardware resources, the middle-level computing platform, the upper-level application services, and the top-level user behavior. The complete process of fault handling in the operation and maintenance process includes the following 8 event state transitions and 7 processing processes between them. According to the processing sequence, the status of these events is the occurrence of potential anomaly indicators, the occurrence of faults, the discovery of faults, the reduction of the scope of faults, the repair of services, the analysis of fault characteristics, the optimization of processing procedures, and the improvement of the operation and maintenance standards. The processing process between different event states includes the following parts: analyzing historical monitoring indicators, predicting their changing trends, and eliminating hidden faults in advance; matching the fault signature database, identifying anomaly monitoring indicators, and pushing alarms in real time; generating corrected cluster management policies, trigger the corresponding cluster management means; correlate and analyze anomaly monitoring indicators to locate the root cause of the fault; summarize the fault handling process

and enrich the knowledge base of expert experience; generate a fault event processing report, archive and save it; simulate the fault scenario and test the complete fault processing process.

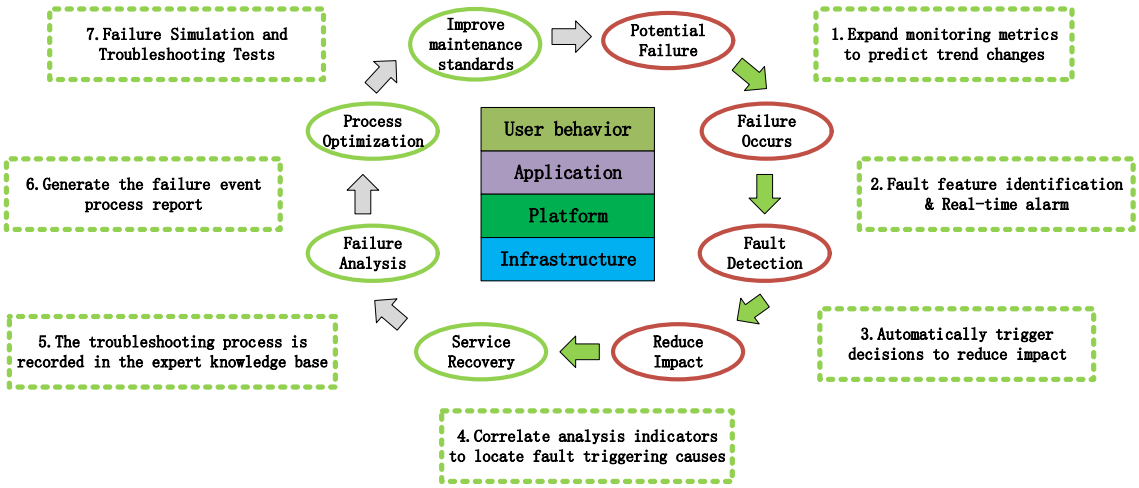

Figure 5: The closed loop of intelligent maintenance management

1. Platform framework and Implementation

Since 2017, in order to solve the problems of many types of site monitoring tools, isolated islands of monitoring data, and low analysis efficiency, IHEPCC designed and implemented the Operation and Monitoring Analysis Toolkit (OMAT), which is integrated with multiple open source tools in the field of big data. As shown in figure 6, OMAT realizes the functions of unified collection of site data, real-time analysis, and third-party application support. Gradually it has become the core of the daily operation and maintenance of IHEPCC.

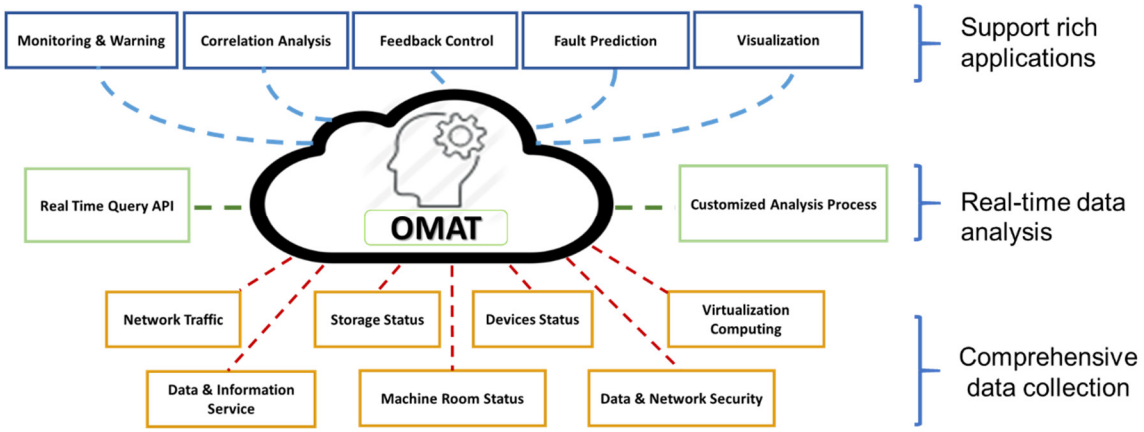

Figure 6: The framework of OMAT

In 2020, IHEPCC proposed a platform with a multi-center design. As the number of sites continues to increase, the amount of platform resources continues to expand, and the computing environment becomes more complex, increasing the difficulty of site maintenance. In order to improve the quality of operation and maintenance, combined with the operation and maintenance goals of the previous chapter, we have expanded the OMAT function and designed and implemented the operation and maintenance analysis platform of IHEP. Figure 7 shows the overall framework of the operation and

maintenance analysis platform. There are six main functional modules shown in this framework.

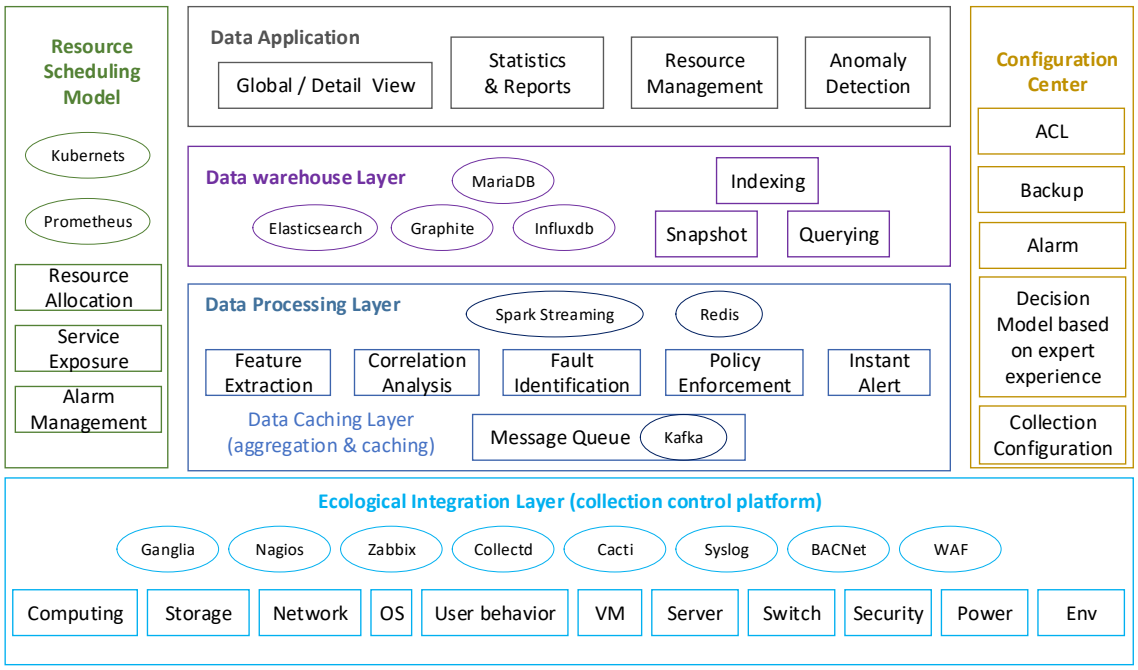

Figure 7: The overall framework of the operation and maintenance analysis platform

The bottom level is the collection control platform, which is the foundation of the entire framework. This level reuses the traditional monitoring tools, and realizes the monitoring data collection in multiple fields of the computing platform. The covered monitoring areas are as follows: 1) infrastructure part, which includes these functions: Computer room environment monitoring, such as power, UPS, temperature and humidity, air conditioning, etc. Hardware equipment monitoring, such as the monitoring of rack servers, blade servers, and network switching devices; virtual device monitoring, such as virtual machines, containers, etc. Operating system monitoring, such as Windows, Unix, and Linux 64-bit. 2) Cluster service part, which includes these functions: Cluster computing services monitoring, such as account authentication, job scheduling, resource management, operating environment configuration, etc. Cluster storage service monitoring, such as scientific software distribution, scientific data storage, database resource access, etc. Cluster network service monitoring, such as network connectivity, network bandwidth traffic, data transmission latency, etc. 3) User behavior part, such as user login behavior, job submission behavior, data file access behavior, website access behavior, etc. 4) Network security part, including password brute force cracking, DNS attack, mining attack, etc. Remote site monitoring areas, including communication status between sites, WAN transmission quality, etc.

The data processing layer is implemented by the data stream processing framework, which supports flexible and efficient data analysis, realizes fault determination, and executes policies, and real-time alarms. As shown in figure 7, this part uses Kafka[8] as a data cache module to receive different types of monitoring data from multiple topics; Spark Streaming[9] is used as the main data stream analysis tool to process the data; A redis[10] cluster is used as a high-performance database to temporarily store important data during analysis. The main functions of this layer

include feature extraction, correlation analysis, fault perception, policy application, and real-time alarms. The primary purpose of the data feature extraction state is to provide a structured analysis of the raw data coming from different monitoring sources, and to store the potentially reusable attribute information in the Redis cluster.

In the correlation analysis stage, it is used to realize the attribute enrichment of structured monitoring data. The enriched attributes may come from external static databases, such as the asset database that records the location of the machine, the environmental database that records the location of the sensor, etc. The enriched attributes may also come from the Redis cluster, which records the reusable attributes of other monitoring data sources. For example, the Nvidia-sim command collects the load information of the GPU card and uses the Redis cluster to query which HPC job is using the GPU card.

In the fault detection stage, offline data processing is used to train anomaly models for different operation and maintenance scenarios. Machine learning algorithms used at this stage include regression algorithms, such as logistic regression, and linear regression. classification algorithms, such as isolated forest algorithm, and clustering algorithms: such as K-means, etc.[11] Feature matching is performed on the preprocessed operation and maintenance data, and anomalies are automatically classified. Or according to the judgment rules provided by the expert experience database, key monitoring metrics are used for threshold judgment, historical data year-on-year, chain-month correlation calculation, and other processing methods to detect an anomaly. The anomaly processing stage includes processes such as automatic correction and sends the alarm. For anomaly accidents that have a clear processing method, the automatic correction operation is performed concurrently by coroutines according to the recorded correction process. For example, if the storage directory of the worker node is anomaly, it is automatically removed from the computing resource pool to stop receiving jobs. Anomaly incidents that need to be resolved manually shall be pushed to the relevant administrators in due course. For example, a storage server hardware failure requires an administrator to recover manually.

The data warehouse layer is the core of the platform, providing monitoring data storage, querying, and archiving services. This module includes Elasticsearch, Graphite[12], Influxdb[13], and MariaDB[14]. The Elasticsearch cluster is used to store monitoring data of log type or service status type, covering most of the data of the monitoring platform. Combined with Kubernetes[15], the Elasticsearch container node is divided into three zones, and data is stored in multiple copies. The copies of the same index are stored in different zones to avoid the risk of data loss and to ensure the high availability of data query functions. The Kubernetes combines the computing resources and memory resources with the node role attribute of Elasticsearch, and divide the nodes of each zone of Elasticsearch into master node, ingest node, transform node, data node, etc., which are used for cluster management, data writing, data querying, and storage.

Elasticsearch Cluster combines the persistent volume and storageclass attributes of Kubernetes, and divides the data node nodes of Elasticsearch into nodes with different attributes. The hot attributes nodes, which are deployed on container nodes with SSD storage, are used to store recently written and frequently accessed monitoring data. The warm attributes nodes, deployed on container nodes with SAS storage, are used to store monitoring data that may be queried to quantify. The cold attributes nodes, deployed on container nodes with SATA storage, are used to store the monitoring data

that is available and is only seldom queried, fewer than 100 times per year. By configuring the data migration rules for each index in the Elasticsearch cluster, the monitoring data can be migrated to data nodes with different roles, and the read and write efficiency of the monitoring data can thereby be improved. In addition, for the monitoring data from quantifying, combined with the snapshot function of Elasticsearch and the use of tape storage back-ends for archiving, the long-term preservation of historical monitoring data of Elasticsearch is realized. The automated scripts periodically perform snapshot operations on indexes that meet the backup rules on the cold node[16], move the data in the snapshot warehouse to tape storage, delete the indexes on Elasticsearch and data in the snapshot warehouse, complete the archive operation of monitoring data, and release disk storage space. For the index data that needs to be restored, remove the tape data to the snapshot warehouse and restore these snapshots. InfluxDB and Graphite have excellent features for data ingestion rates, disk data compression, and query performance for time series data. These databases are used to store numerical monitoring metrics, and use integral, derivative, median, and other statistical functions to aggregate and query indicators to meet the query requirements of the application layer. The application layer develops API using the Django web framework, to provide data query and data update services. The API receives query parameters of the application layer, and combines them into query statements to query data from different databases, and returns the results to the query requester.

The MariaDB database in the data warehouse layer is used to store the configuration information related to the operation and maintenance analysis platform and is configured and managed by the configuration center module of the operation and maintenance system. The configuration center provides operation and maintenance administrators with configuration interfaces for services at all levels of the data operation and maintenance platform, including functions such as collection configuration, expert experience record, decision configuration, alarm configuration, backup strategy, and authority control.

The collection configuration model provides the two core functions of creating a new monitoring strategy and selecting a deployment node. The administrator configures the data monitoring strategy and the corresponding deployment node according to the data collection requirements.

The decision model based on expert experience is used to record common problems and processing methods in daily operation and maintenance, and solidify the processing process to automatically repair the same type of anomaly detected. This model is based on the accumulated data obtained from administrator experience, and from the observed improvements in the processes used in the operation and maintenance work during fault repair by domain experts. According to various early warning rules and anomaly diagnoses, a repair strategy is generated, which is automatically implemented by the system. For example, the metadata query frequency of individual user jobs increases the storage service load, and the user job scheduling policy needs to be adjusted in time to improve the responsiveness of storage services.

Alarm configuration is a relatively independent configuration function used to create information such as alarm contacts, alarm methods, alarm time intervals, and alarm event hierarchy. It includes the alarm of missing collection of monitoring data in the collection control platform and the anomaly alarm of fault detection in the data analysis layer.

The backup strategy is a configuration function for the data warehouse layer. Combined with the data and query characteristics, it configures the migration strategy and snapshot archive strategy of index data. Permission control is used to configure the access rights management of the DATA API and limit the query scope of upper-layer applications.

The resource scheduling model is built based on Kubernetes, which is responsible for resource allocation and maintenance of the service liveness. Kubernetes implements functions such as automatic deployment, automatic expansion, and service liveness of container clusters at all levels of the platform. Through the resource configuration file, the CPU, memory, storage, and other resources of the newly added device are pooled into logical resources and allocated on-demand to achieve smooth expansion of operation and maintenance system resources. This model provides independent namespaces for resource isolation for different application services to ensure that application service resources are independent of each other. Storage resources are divided into different storage classes according to IO performance and replica status, and the data of stateful services is persisted to ensure the final persistence of application data and system disaster recovery capabilities. Using HAProxy[18] technology and ingress and headless[19] service discovery methods, this model configures the external exposure method of internal application services to achieve a balanced load and high service availability of various applications. At the same time, the Prometheus[20] monitoring component is deployed to monitor the status and operation indicators of the containerized application services, and its Alertmanager component is configured to notify and warn the operation and maintenance system of existing or potential application service failures in a timely manner to ensure the stability of the resource management layer.

The top layer is some application areas supported by the platform. In the field of information query and display, it includes two types of panels: The cluster panel for user is used to display the user's job information, accounting information, storage space usage, account login history, etc. The cluster panel for administrator is used to display the cluster's room environment, network quality, resource utilization, cluster service status, the status of remote sites, etc. By configuring role rights management on the visualization tool, the data display window displays different operation and maintenance data information for administrators and ordinary users respectively.

In the process of providing computing services by a computing cluster, there are usually two types of factors that affect the service quality. One factor is that cluster failure causes the operating environment to lack key services, which affects the job success rate. Another factor is the irregular data access behavior during the running of user jobs, which affects the performance of the storage server, which in turn affects the data access and data analysis efficiency of other user jobs. For the first factor, the HTCondor-based HTC computing resource management system based on operation and maintenance analysis platform has been developed[21]. Based on data stream processing technology, the service monitoring information, service and experiment mapping relationship, and node and experiment mapping relationship are combined to realize the dynamic adjustment of the node and experiment sharing policy. The system avoids the situation that anomaly service nodes cause error jobs, and ensures the stability of computing services. For the second influencing factor, anomaly detection of I/O behaviors in computing cluster based on unsupervised machine learning has been developed[22]. The collection control platform obtains access behavior indicators of user

jobs through the Lustre Collectd plugin, train Isolation Forest unsupervised models with data samples per week and per day, the data processing layer makes almost real-time anomaly detection by the anomalous score generated given by pre-trained models for a new data sample within a day and a week. Associated jobs and user information, the cluster panel for administrator, can help storage administrators quickly locate user and jobid that affect storage performance.

1. Conclusion and Future Research

IHEP has designed and implemented an operation and maintenance analysis platform providing a comprehensive monitoring process with data collection, data analysis, anomaly alarm, and fault self-healing, supports long-term storage of monitoring data, and supports data query of third-party applications. This platform provides flexible configuration modules, which can rapidly deploy and dynamically expand resources based on container clusters. It realizes the evolution of the value of operation and maintenance monitoring data, and solves the increasingly complex operation and maintenance problems of large-scale data centers. In the future, additional machine learning algorithms will be incorporated, and more anomaly detection models defined. The operation and maintenance analysis platform will expand the application scope, realize monitoring data analysis to drive business operation decisions, and ensure the stable operation of "One Platform".

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No.11805226, No.12105303, No.12075268, No.11775250).

References

[1] Computing A. An architectural blueprint for autonomic computing[J]. IBM White Paper, 2006, 31(2006): 1-6.

[2] Nagios Introduction, https://library.nagios.com/library/products/nagios-core/documentation, online, accessed 24-Apr-2022

[3] Syslog-ng Introduction, https://github.com/syslog-ng/syslog-ng, online, accessed 24-Apr-2022

[4] Ganglia Introduction, https://github.com/ganglia, online, accessed 24-Apr-2022

[5] Cacti Introduction, https://www.cacti.net, online, accessed 24-Apr-2022

[6] Elasticsearch Introduction, https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html, online, accessed 24-Apr-2022

[7] Kibana Introduction, https://www.elastic.co/guide/en/kibana/current/index.html, online, accessed 24-Apr-2022

[8] Kafka Introduction, https://kafka.apache.org/documentation/, online, accessed 24-Apr-2022

[9] Spark Streaming Introduction, https://spark.apache.org/docs/latest/streaming-programming-guide.html, online, accessed 24-Apr-2022

[10] Redis Introduction, https://redis.io/docs/, online, accessed 24-Apr-2022

[11] Chen J, Wang L, Hu Q. Machine learning-based anomaly detection of ganglia monitoring data in HEP Data Center [C]//EPJ Web of Conferences. EDP Sciences, 2020, 245: 07061.

[12] Graphite Introduction, https://graphite.readthedocs.io/en/latest/, online, accessed 24-Apr-2022

[13] Influxdb Introduction, https://www.influxdata.com/, online, accessed 24-Apr-2022

[14] MariaDB Introduction, https://mariadb.org/documentation/, online, accessed 24-Apr-2022

[15] Kubernetes Introduction, https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/, online, accessed 24-Apr-2022

[16] Elasticsearch node role Introduction, https://www.elastic.co/guide/en/elasticsearch/reference/current/modules-node.html, online, accessed 24-Apr-2022

[17] Elasticsearch snapshot and restore Introduction, https://www.elastic.co/guide/en/elasticsearch/reference/current/snapshot-restore.html, online, accessed 24-Apr-2022

[18] HAproxy Introduction, http://docs.haproxy.org/2.5/configuration.html, online, accessed 24-Apr-2022

[19] Kubernetes Ingress Introduction, https://kubernetes.io/docs/concepts/services-networking/ingress/, online, accessed 24-Apr-2022

[20] Prometheus Introduction, https://prometheus.io/docs/introduction/overview/, online, accessed 24-Apr-2022

[21] Q. Hu, W. Zheng, X. Jiang and J. Shi. Application of OMAT in HTCONDOR resource management. // 2021 International Symposium on Grids & Clouds 2021.

[22] L.Wang, Q. Hu, Juan Chen. Anomaly Detection of I/O behaviors in HEP computing cluster based on unsupervised machine learning. // 2021 Advanced Computing and Analysis Techniques in Physics Research 2021