JUNO raw data management system

Xiaomei Zhang[1]* Yifan Li[1] Xuantong Zhang[1] Yizhou Zhang[1]

[1] Institute of High Energy Physics, Beijing, China

*E-mail: zhangxm@ihep.ac.cn

Abstract. The Jiangmen Underground Neutrino Observatory (JUNO)[1] is a multipurpose neutrino experiment. JUNO will start to take data in 2024 with 2PB data each year. It is important that raw data is copied to permanent storage and distributed to multiple data center storage system in time for backup. To make available for re-reconstruction among these data centers, raw data also need to be registered into metadata and replicas catalogues of the JUNO distributed computing system. The raw data management system will take care of distributing raw data and running data processing activities in JUNO data centers. An automatic system based on JUNO distributed data management has been designed and developed to do registering, replicating, archiving and data reconstruction in a data-driven chain. The monitoring dashboard has been designed and developed to ensure the quality of data transfer and processing. The prototype of the system has been tested with commissioning data since 2023 and the system will continue to join JUNO data challenge in early 2024.

1. Introduction

JUNO is a versatile neutrino experiment with the primary physics goal of measuring neutrino mass hierarchy and mixing parameters. This includes studying various types of neutrinos such as solar neutrinos, supernova neutrinos, and atmospheric neutrinos. Located in Jiangmen, China, JUNO is set to commence data collection in 2024. TAO, a satellite detector situated near the JUNO site, is designed to precisely measure reactor energy spectra and enhance JUNO's sensitivity in studying neutrino mass hierarchy. JUNO is anticipated to collect data at a rate of 1kHz, resulting in a data volume of approximately 60 MB per second. As a result, the total data volume generated in a year is projected to be 2PB. The planned file size for each data file is 5 GB.

For data transfer from the JUNO and TAO sites to IHEP (Institute of High Energy Physics), a dedicated network with a speed of around 150Mb/s connects to CSNS (China Spallation Neutron Source), which in turn has a dedicated 20Gb/s line to IHEP. From IHEP to Europe, JUNO can utilize the LHCOPN network provided by WLCG (Worldwide LHC Computing Grid), which offers a speed of 80GB/s via GEANT.

JUNO employs distributed computing[2] technology to create a unique platform for organizing computing resources, distributing data to data centers, and scheduling and executing data processing activities. To take care of raw data distribution and processing, the JUNO raw data management system is being planned and developed using this platform, which is described in this paper. The JUNO distributed computing system is based on DIRAC[3], utilizing both the DIRAC workload management system[4] and data management system[5].

2. Raw data management

JUNO site is located in Kaiping, Guangzhou. Upon detection, the raw data from the JUNO detector is promptly transferred to IHEP using a dedicated non-grid replication system called Spade using a dedicated network. Once the JUNO raw data arrives at the IHEP data center, several activities related to data processing and replication are simultaneously triggered and initiated.

To coordinate and trigger these workflows, a message queue system based on Kafka[6] is employed, as shown in Figure 1. Messages are used to schedule and initiate the required processes upon receiving information about the raw data files in the IHEP EOS disk system. Upon receiving the messages, a format-transforming process is initiated to convert the raw data into RTRaw format, which is the ROOT format specific to the raw file. The RTRaw format data volume is approximately two-thirds of the raw data volume. The RTRaw data is directly utilized for subsequent reconstruction processes.

Simultaneously, upon receiving the RTRaw data, the first stage of reconstruction and sampling reconstruction for data quality monitoring is initiated. Additionally, both raw data and RTRaw data are replicated to European data centers. These data centers utilize distributed computing platforms and resources to carry out physics production.

The raw data is archived in the tape system at IHEP and other data centers as well, at least one copy in IHEP and another copy in Europe. All raw and RTRaw data are registered within the JUNO distributed computing system, ensuring efficient management, accessibility and authentication control.

To handle the replication of raw and RTRaw data, archival to tape systems, and processing within the JUNO distributed computing platform, an automated system for raw data transfer and data processing has been designed and implemented. This system streamlines and automates these critical tasks. The following sections of the paper will describe more details about this system.

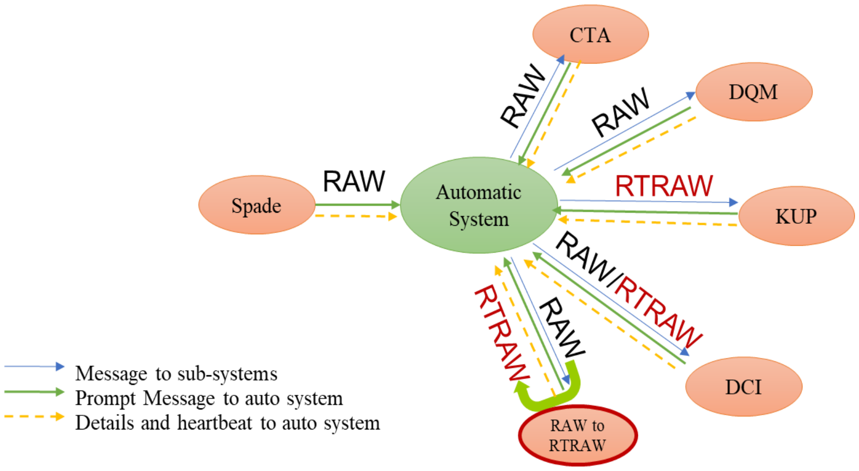

Figure 1. Raw data workflow management with MQ

3. Grid raw data transfer system

A grid raw data transfer system has been developed to facilitate the seamless flow of raw and RTRaw data from IHEP to data centers, enabling fully automated replication procedures as illustrated in Figure 2. This system incorporates automated checking and validation at each step of the replication process, ensuring robustness and accuracy. It also provides easy troubleshooting, error warning, and monitoring capabilities for both shifters and administrators.

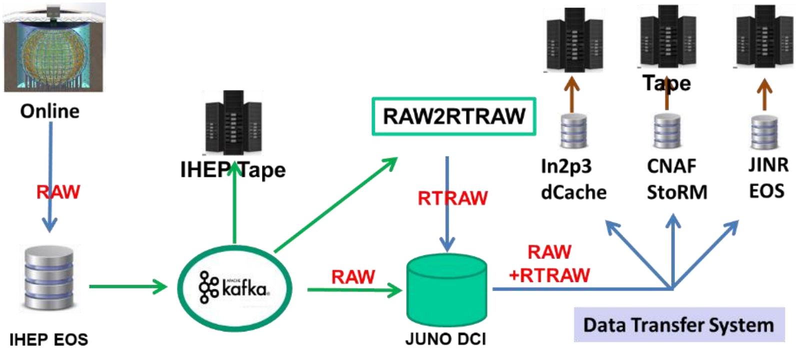

Figure 2. Procedure of raw data replication

To simplify control over the transfer process and enable efficient status checking, the system utilizes blocks as transfer units. A block represents a group of files with similar properties, and the size of each group is determined by the transfer efficiency. Depending on physics requirements, a dataset can be composed of one or more blocks.

The system comprises four main components: Data Receiver, File Register, Block Creator, and Block Transfer, as depicted in Figure 3. The Data Receiver module is responsible for receiving messages containing raw file information, performing checks on the messages, and decoding the file information. The Data Receiver also reports the problem to the responsible by email when met critical errors in files. Once all the file information is obtained and validated, the files in the EOS system are registered in the Dirac File Catalogue.

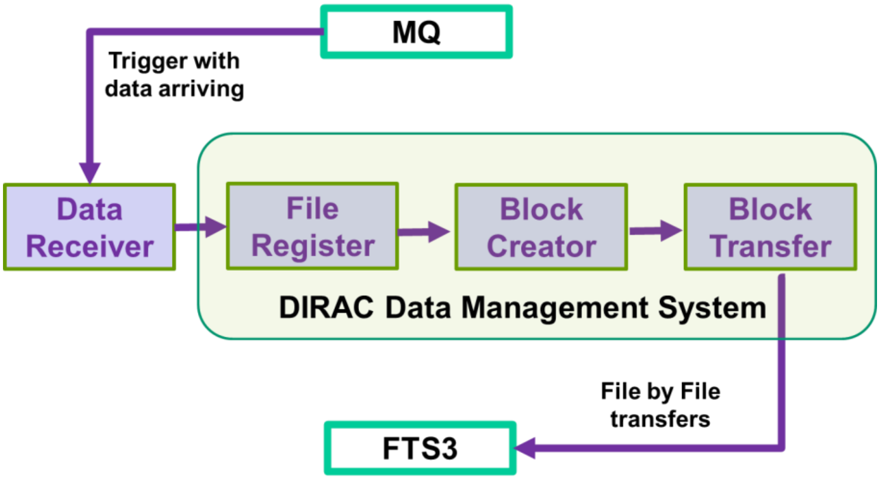

Figure 3. System architecture of raw and RTRaw data transfer system

The Block Creator module handles the creation of blocks, including the definition and assignment of metadata based on specified policies. This metadata enables querying of file lists within blocks from the Dirac File Catalogue. Meanwhile, the Block Transfer module takes charge of replicating files from IHEP to the data centers. This module is developed using the DIRAC data management system. Transfer tasks are automatically created, block by block. To ensure fault tolerance, each transfer channel has its own transfer queue, minimizing the impact of failures in one channel on others. Multi-channel transfers are supported in case of issues with specific channels. The Block Transfer module interacts with the transformation system, obtaining the file list for each block, initiating file transfer tasks and splitting block transfer tasks into individual file transfers accordingly. These file transfer tasks are sent to FTS3 (File Transfer Service 3) [7] for execution. Files are checked and validated on a block-by-block basis, enabling the system to identify any failed transfers. In the event of failed transfers, the system supports automatic rescheduling of these files into the transfer queue. This approach allows for retransferring of only the failed files within a block, reducing the impact of unstable factors such as network issues. The system leverages the data-driven nature of the transformation system, enabling transfer tasks to be initiated as soon as the data becomes available, without waiting for files to be fully registered within blocks.

To facilitate these processes, the File Register, Block Creator, and Block Transfer components have all been implemented using the DIRAC data management system. This choice ensures seamless integration and efficient management of the file replication and transfer operations.

4. Raw data processing

To facilitate the second reconstruction, also known as physics production (PP), which takes place in multiple data centers, an automatic reconstruction system has been developed to handle the entire workflow and dataflow in PP as shown in Figure 4. This system comprises a workflow component and a data flow component. The workflow component focuses on submitting and scheduling PP jobs in the data centers where RTRaw data is available. PP jobs process the RTRaw data and generate ESD (Event Summary Data) as output. The ESD data is then copied from the work node where the PP job is executed to the closest Storage Element (SE) within the respective data center. Simultaneously, the data flow component ensures the replication of ESD data from the temporary SE to the destination SE. This replication process is automated and takes place once the ESD data is available in the temporary SE. The workflow and data flow components are closely interconnected. They are triggered by the availability of RTRaw data in the Dirac File Catalogue (DFC) and connected through the ESD data present in the DFC. This integration enables a seamless and automated chain of operations, ensuring efficient processing and replication of the data throughout the physics production stage.

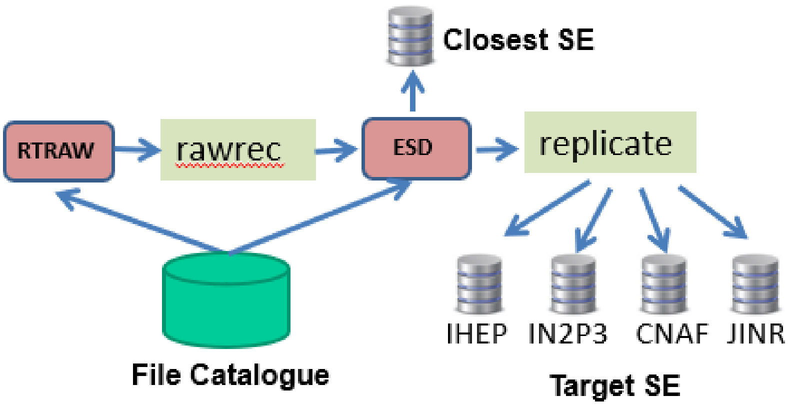

Figure 4. Automatic Physics production chain

5. Monitoring and testing

5.1 Monitoring

Monitoring plays a crucial role in ensuring the stability of the raw data management system. To enable effective monitoring, both file-level and block-level monitoring are provided for data transfers. File-level monitoring allows for detailed tracking of individual file transfers, providing easy access to status updates and error logs. Real-time monitoring of file transfers is achieved through the FTS3 web monitoring interface, which provides immediate visibility into ongoing transfers. For historical monitoring, the fts-msg-bulk agent is employed to send FTS transfer information in real-time to ActiveMQ. This information is then forwarded to logstash for filtering, and the resulting data is stored in ElasticSearch (ES) [8]. Grafana is utilized to generate informative plots based on the stored data. These plots include transfer status, transfer efficiency, transfer errors, and transfer throughput, as depicted in Figure 5. This visualization offers insights into the performance and progress of the data transfers. In addition to data transfer monitoring, the monitoring of PP jobs is also essential. Job information from the DIRAC monitoring and accounting database is sent to ElasticSearch (ES), enabling the generation of plots to monitor the status and progress of PP jobs. By utilizing these monitoring mechanisms, the raw data management system provides valuable information for analysis and troubleshooting.

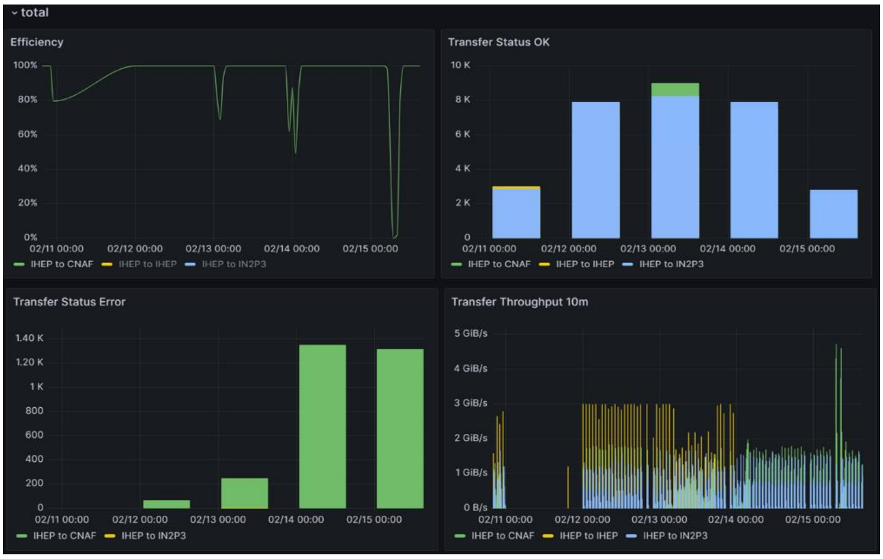

Figure 5. File transfer monitoring plots based on ES and Grafana

5.2 Testing

The raw data transfer system underwent testing starting from the end of 2023, focusing on commissioning and OSIRIS data. Throughout 2024, a continuous transfer of approximately 90,000 files and 36TB of data was conducted, as depicted in the left plot of Figure 6. Importantly, these transfers were accomplished without requiring human intervention. During the testing phase, although no major issues were encountered, a few irregular files from the commissioning and OSIRIS data did cause certain problems. Examples of such issues included files with zero size, problematic file information, and files larger than 100GB. To address these challenges and prevent them from affecting the registration and transfer queue, checks and protections were implemented. These measures ensured that the system could handle and resolve these irregularities effectively. Furthermore, it's worth noting that the network connection between IHEP and JINR had an impact on the transfer process between the two locations. The network conditions influenced the transfer efficiency, as shown in the right plot of Figure 6. Despite these factors contributing to less-than-perfect transfer efficiency, the system demonstrated overall robustness and the ability to handle the continuous transfer of data.

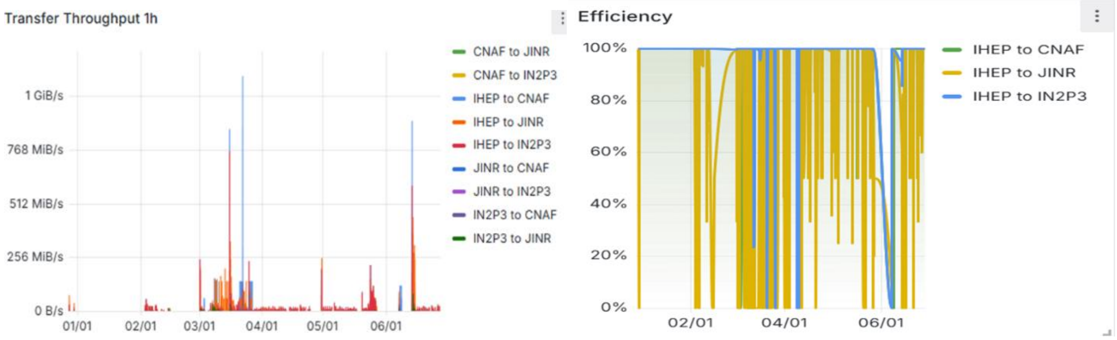

Figure 6. Raw data transfer test status

The first offline data challenge commenced at the end of 2024 with the primary objective of testing and validating the entire chain of raw data management. During this challenge, both the RTRaw data transfer and physics production processes were thoroughly tested, as depicted in Figure 7.

As part of the challenge, a total of 100,000 files were successfully registered in the Dirac File Catalogue (DFC). The RTRaw data, comprising 7 blocks and totaling 103TB of data, was transferred to the respective data centers within the designated timeframe. Remarkably, the transfer process achieved a 100% success rate, demonstrating the system's efficiency and reliability. Each block consisted of 14,400 files, and the entire dataset represented one week's worth of data.

In parallel to the data transfer, the offline data challenge also involved the submission and execution of PP (physics production) jobs. The PP jobs were initiated once all RTRaw data became available in DFC. Approximately 100,000 PP jobs were assigned and executed across four data centers. These jobs ran in an 8-core mode and exhibited an impressive success rate of 99.9%. No errors were encountered on the grid side during the execution of these jobs. Additionally, during the first offline data challenge, approximately 25TB of data was transferred from IHEP to three other data centers. This transfer was successfully completed without any reported errors. The results of the challenge demonstrate the robustness and effectiveness of the raw data management system, highlighting its ability to handle large volumes of data and execute complex physics production workflows with high success rates.

Overall, the testing phase provided valuable insights into the system's performance and identified areas for further optimization and improvement. The implemented checks and protections, along with ongoing monitoring and optimization efforts, contribute to the system's stability and reliability in handling large-scale data transfers.

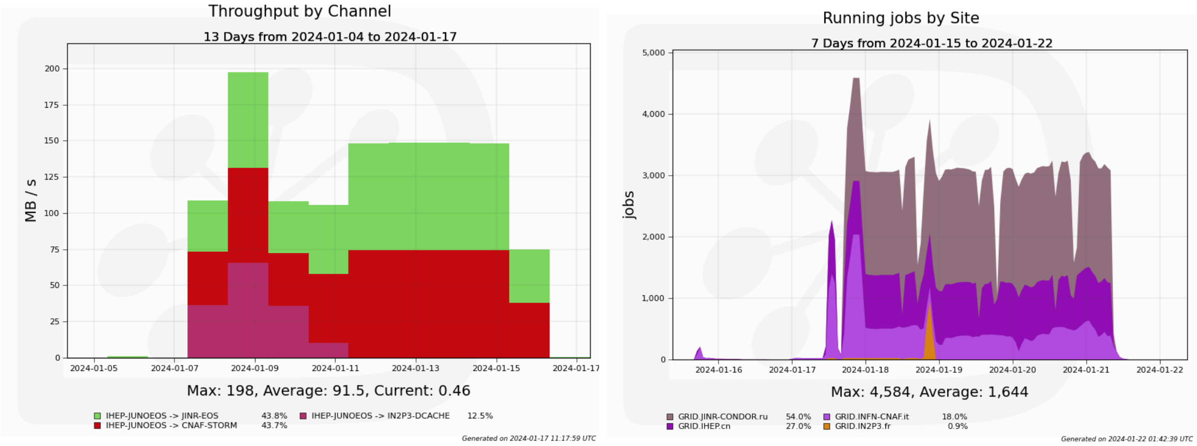

Figure 7. RTRAW transfer and PP job tests

6. Conclusions and future work

The readiness of the prototype for the JUNO raw data transfer and processing system is an important milestone. The seamless connection between the RAW-related offline data processing systems through MQ facilitates efficient communication and coordination. The utilization of DIRAC DMS provides a user-friendly and automated solution for managing the raw data in the JUNO experiment. In preparation for the upcoming data-taking phase, additional pressure tests will be conducted in 2024. These tests aim to simulate realistic data-taking scenarios, allowing for further evaluation and validation of the system's performance under increased load and scale. To ensure optimal scheduling efficiency, it is crucial to undertake larger-scale tests and investigations that closely resemble the real JUNO use cases. These comprehensive tests will provide valuable insights into the system's behavior, identify potential bottlenecks, and offer opportunities for further enhancements and optimizations.

Acknowledgments

The authors would like to thank the JUNO offline software team for their help with the JUNO software, the DIRAC group for the help with the DIRAC system, and the colleagues at the IHEP computing centre for their support. This paper is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences， Grant No. XDA10010900.

References

[1] F.P. An, Neutrino Physics with JUNO, J. Phys. G 43 030401 (2016)
[2] X. Zhang, JUNO distributed computing system, EPJ Web of Conferences 295, 04030 (2024) https://doi.org/10.1051/epjconf/202429504030
[3] Federico Stagni, Andrei Tsaregorodtsev, ubeda, Philippe Charpentier, Krzysztof Daniel Ciba, Zoltan Mathe, ... Luisa Arrabito. (2018, October 8). DIRACGrid/DIRAC: v6r20p15 (Version v6r20p15). Zenodo. http://doi.org/10.5281/zenodo.1451647
[4] A. Casajus1, R. Graciani, A. Tsaregorodtsev et al. DIRAC pilot framework and the DIRAC Workload Management System, J. Phys: Conference Series 219 (6) (2010)
[5] A. Tsaregorodtsev et al. DIRAC Data Management Framework. DOI: https://doi.org/10.22323/1.270.0035
[6] https://kafka.apache.org/documentation/

[7] A. Kiryanov, A. Alvarez Ayllon, O Keeble. FTS3/WebFTS – a powerful file transfer service for scientific communities. Procedia Computer Science Volume 66, 2015, Pages 670–678

[8] X. Zhang, Multicore workload scheduling in JUNO， EPJ Web of Conferences 214, 03048 (2019)