# TuPiX-Flow: Workflow-based Large-Scale Scientific Data Analysis System

Sunggeun Han, Jung-Ho Um, Hyunwoo Kim, Kyongseok Park
Scientific Data Technology Lab
Korea Institute of Science and Technology Information
245, Daehakro, Yuseong-gu, Daejeon, 34141
Rep. of Korea
{sghan, jhum, pardess, gspark}@kisti.re.kr

## ABSTRACT

With more research being actively conducted on big data, there has been a growing interest towards data-intensive data science in various fields. Against this backdrop, extensive efforts have been made to apply High Performance Computing (HPC) technology such as distributed computing or parallel computing. However, many researchers are unable to fully utilize such technology due to their lack of knowledge, programming skills, and analytical skills for large-scale data. TuPiX-Flow, which provides a workflow-based user interface, enables researchers to easily analyze large-scale data using workflow diagrams for end-to-end analytical processes even without programming knowledge or implementation technology. In addition, large-scale data can be efficiently analyzed in software compatible with HPC, including distributed computing and parallel computing. As a case study of large-scale data analysis using TuPiX-Flow, a model for the detection of red tides surrounding the Korean peninsula was analyzed based on ocean color satellite data.

## CCS CONCEPTS

• **Software and its engineering** → Model-driven software engineering

## KEYWORDS

Workflow-based analysis, Scientific data analysis, Data Science

## 1 INTRODUCTION

Today, there is a greater demand for big data analysis in various research fields, and researchers across disciplines such as science, engineering, and the humanities are showing increasing interest in data-intensive data science. While there are many platforms available for big data analysis like the Hadoop ecosystem, researchers are unable to fully utilize HPC technology due to their lack of knowledge or programming skills. In the case of researchers studying the marine ecosystem surrounding the Korean peninsula, the software used to analyze ocean color satellite data include the NASA-provided SeaDAS[1] and MATLAB. However, the limited features of such software make them incompatible for large-scale data processing. This highlights the need for a new system that offers high efficiency in large-scale data processing and analysis.

TuPiX-Flow was developed to resolve the aforementioned issues. Even without HPC technology or programming skills, researchers can use workflow diagrams to perform large-scale data analysis via the workflow-based user interface. To ensure high efficiency in large-scale data processing and analysis, the system features large-scale storage technology, distributed computing, and parallel computing. The advantages of TuPiX-Flow are as follows:

- *Workflow diagrams that constitute end-to-end analytical processes enable users to perform large-scale data analysis more easily.*
- *The analytical results can be visualized and exported in the form of images or documents.*
- *Users can collaborate with other researchers by sharing their workflow diagrams.*
- *Large-scale data analysis can be performed rapidly using the internal distributed and parallel processing system.*

## 2 SYSTEM CONFIGURATION & PROCEDURES

### 2.1 TuPiX-Flow Architecture

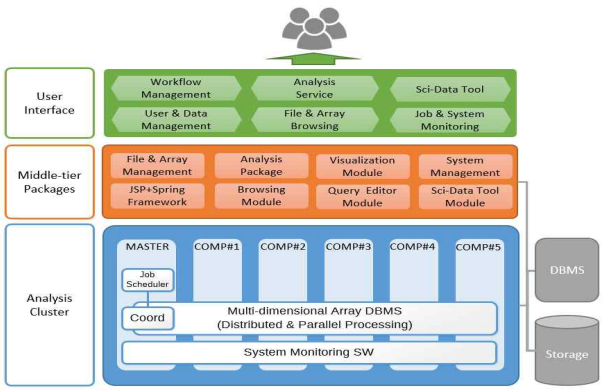


**Figure 1: TuPiX-Flow System Architecture**

TuPiX-Flow runs on a cluster-based system and supports system scalability. Fig. 1 shows the overall system architecture of TuPiX-Flow. The function of each layer is as follows.

• *Analysis Cluster.* The analysis cluster supports distributed and parallel processing of large-scale data[2]. Job scheduler distributes jobs to coordinator node, and the jobs are processed in the compute nodes.

• *Middle-tier Packages.* The package consists of modules that forwards requests sent via the user interface to the Analysis Cluster.

• *User Interface.* User can create workflow diagram, analyze data, and visualize the results on user interface. It also support monitoring for job

## 2.2 Analysis Procedure

For workflow-based analysis under TuPiX-Flow, the following procedures are required.

• *Making Blocks.* Individual analysis modules are created for the end-to-end analytical process. The input data, data processing function, and the output data are defined for each block. They are mapped to the program code for data processing.

• *Connecting Blocks.* Individual blocks are placed as necessary, and a workflow diagram is developed by connecting the blocks with lines in consideration of input and output.

• *Performing an Analysis.* Analysis commands are executed for the generated workflow diagram. Distributed and parallel processing is performed in response to analysis requests. Fig. 2 shows the analysis processing on user requests.

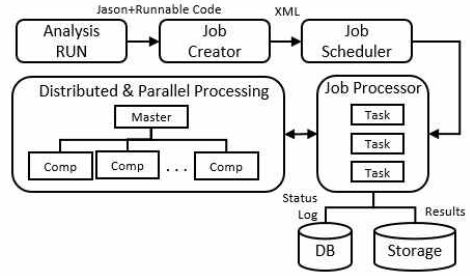


**Figure 2: Analysis processing**

• *Visualizing the Results.* The results are visualized. Users can re-generate the workflow diagram and repeat the analysis until satisfactory results are obtained.

• *Saving & Sharing the Results.* The results and completed workflow diagram can be stored and shared with other researchers.

## 3 Case Study : Ocean Color Analysis

Ocean color analysis was selected as a case study to verify the effectiveness of TuPiX-Flow in large-scale data analysis. Modis Aqua data[3], which is ocean color satellite data provided by NASA, was used to analyze a model for red tide detection in the southern coast of South Korea. The model of red tide detection was FRTD, SS, MOREL[4–6]. Fig. 3 is the workflow diagram for the red tide detection model.

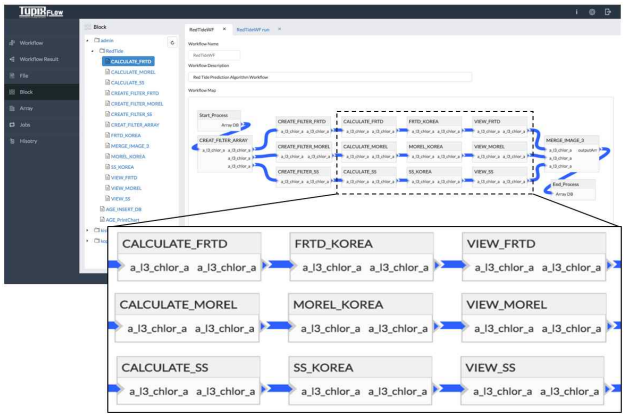


**Figure 3: Workflow diagram**

The results can be visualized as images. Fig. 4 shows the results of analyzing the red tide detection model.

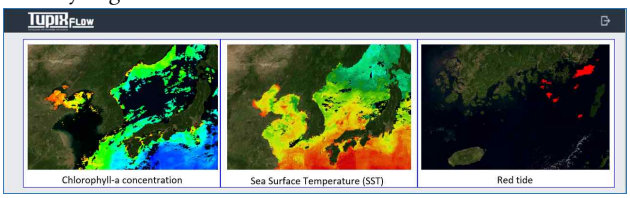


**Figure 4: Results of red tide detection analysis**

## 4 CONCLUSIONS

This study developed the TuPiX-Flow system for researchers to easily perform large-scale data analysis using workflow diagrams even without programming knowledge or implementation technology. The cluster-based TuPiX-Flow enables efficient analysis of large-scale scientific data by supporting distributed and parallel computations. The effectiveness of TuPiX-Flow in large-scale scientific data analysis was verified through a case study on ocean color analysis.

## ACKNOWLEDGMENTS

This work was supported by Korea Institute of Science and Technology Information (KISTI) grant funded by the Korea government (MSIT) (No.K-17-L03-C01-S03).

## REFERENCES

[1] Baith, K., Lindsay, R., Fu, G., & McClain, C. R. (2001). Data analysis system developed for ocean color satellite sensors. *Eos, Transactions American Geophysical Union*, 82(18), 202-202.

[2] Brown, P. G. (2010, June). Overview of SciDB: large scale array storage, processing and analysis. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 963-968). ACM.

[3] Modis-Aqua. https://oceancolor.gsfc.nasa.gov/data/aqua.

[4] Morel, A. (1988). Optical modeling of the upper ocean in relation to its biogenous matter content (case I waters). *Journal of Geophysical Research: Oceans*, *93*(C9), 10749-10768.

[5] Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., Warner, R. A., Tester, P. A., Dyble, J., & Fahnenstiel, G. L. (2008). Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes. *International Journal of Remote Sensing*, *29*(12), 3665-3672.

[6] Suh, Y. S., Jang, L. H., Lee, N. K., & Ishizaka, J. (2004). Feasibility of red tide detection around Korean waters using satellite remote sensing. *Fisheries and aquatic sciences*, *7*(3), 148-162.