

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

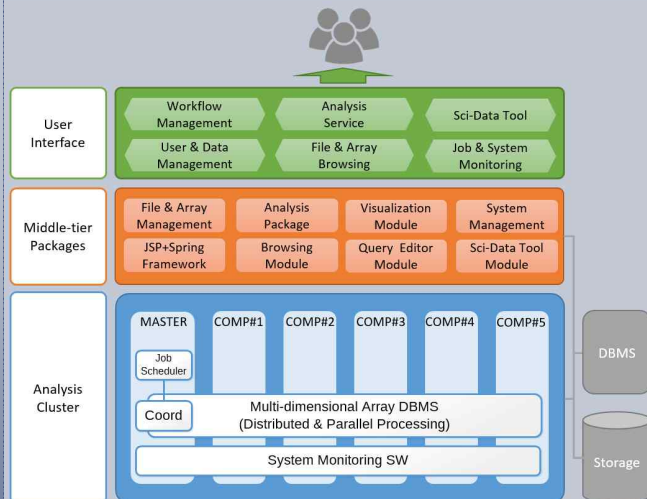
Sunggeun Han, Jung-Ho Um, Hyunwoo Kim, Kyongseok Park

Korea Institute of Science and Technology Information

Motivation

- Greater demand for big data analysis in various research fields (Science, Engineering, Humanities etc.)
 - Increasing interest in data-intensive Data Science
- Needs much knowledge of big data analysis technology
 - Hadoop Echo System
 - HPC Technology (Distributed Computing, Parallel Computing, etc.)
- Lack of knowledge of HPC technology or programming skills
 - Using the limited features of the software incompatible for large-scale data processing
- TuPiX-Flow supports Workflow-based analysis
 - Workflow diagrams : end-to-end analytical processes performing large-scale data analysis more easily.
 - Analytical results can be visualized & exported (images & documents format)
 - Sharing & Collaboration on their workflow diagrams.
 - Performance profit of the internal distributed and parallel processing system.

TuPiX-Flow Architecture



Analysis Cluster

- Supports distributed and parallel processing of large-scale data analysis

Middle-tier Packages

- Provides analysis tools and user & data management

User Interface

- Creates workflow diagram & Monitors job & system status

Analysis Procedure

Making Blocks

- Design blocks : Input, Output, and Functions matching runnable code blocks

Connecting Blocks

- Connecting the blocks with lines in consideration of input and output

Performing an Analysis

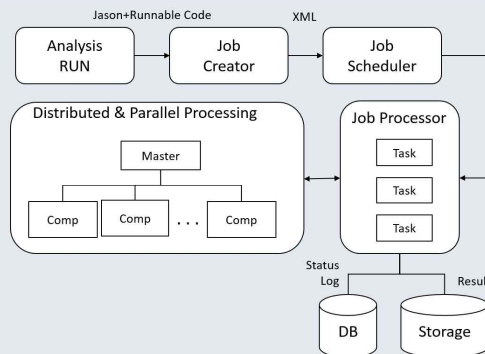
- Executing analysis commands for generated workflow diagram

Visualizing the Results

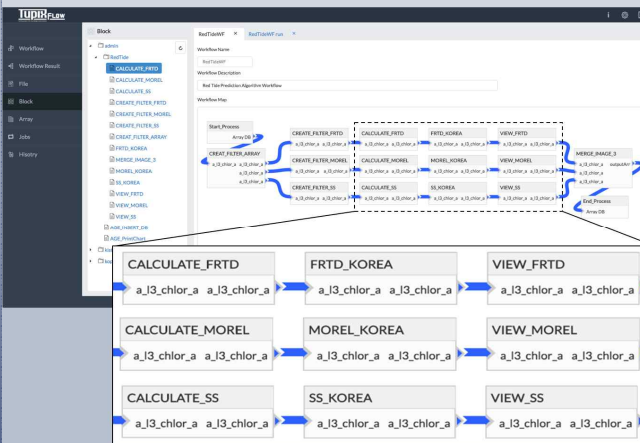
- Verification & Reconstruction workflow diagram

Saving & Sharing the Results

- Store the results and workflow diagram and share them with other researchers



Case Study : Ocean Color Analysis

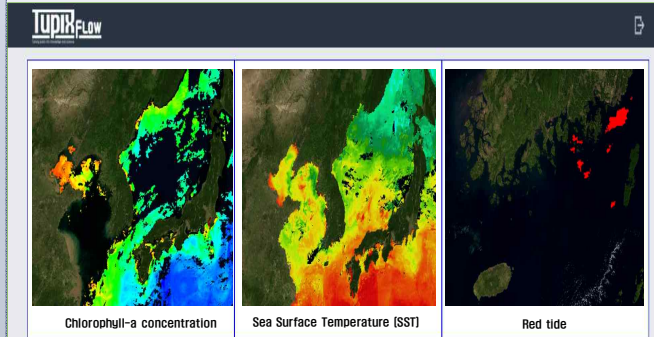


Test for large-scale data analysis

- Dataset : Ocean color satellite data (NASA)
- Analysis Model : Red tide detection

Analysis Results

- Red tide detection in the southern coast of South Korea (July, 2016)
- Algorithms : FRTD, SS, MOREL



CONCLUSIONS

Developed TuPiX-Flow: Large-scale Scientific Data Analysis System

- Performs large-scale data analysis using workflow diagrams
- Doesn't need programming knowledge or implementation skill for big data analysis
- Enables efficient analysis by supporting distributed and parallel computations

Verified our system through a case study on ocean color analysis

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.

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).

MULTIMEDIA MATERIALS

Demonstration of TuPiX-Flow

<https://youtu.be/ofQuBFdTCJg>

