In-place query driven big data platform: Applications to post processing of environmental monitoring

Jen-Gaw Lee¹ | Whey-Fone Tsai¹ | Lung-Cheng Lee¹ | Ching-Yao Lin¹ | Hsi-Ching Lin¹ | Ben-Jei Tsuang²

¹ National Center for High-Performance Computing, NARL, Hsinchu, Taiwan
² National Chung-Hsing University, Taichung, Taiwan

Summary

This paper describes the use of an experimental big data platform for applications of environmental monitoring associated with visualization of global climate forecast data and air quality model simulation and response. Environmental monitoring in general requires both capabilities of model simulation for forecast, and data processing for visualization and analyses. The in-place query driven big data platform, based on concepts of Query Driven Visualization and shared-nothing distributed database, thus is developed for the need. The system architecture of this experimental big data platform entails one master data node and 17 slave data nodes, while the system links to the National Center for High-performance Computing supercomputer, Advanced Large-scale Parallel Supercluster, and storage pool. For software implementation, the openSUSE operating system and MariaDB database are installed on all nodes. The master data node is responsible for metadata management and information integration and the 17 slave data nodes for distributed database and parallel model simulation, data visualization, and analyses. The application of global climate data visualization (Outgoing Longwave Radiation or OLR, temperature, rainfall, etc.) in the platform serves first to partition Network Common Data Form file data into shared-nothing distributed databases for partial visualization in slave data nodes, then integrated into whole visualization in the master node through Message Passing Interface communication.

For the application of air quality management, we first accessed Taiwan Environmental Protection Administration (EPA) observed data in the master node. EPA observed data are replicated to distributed databases in slave nodes; and the air pollution model, Gaussian plume trajectory model, is replicated in all slave nodes for model simulation, which produces output data and associated image files in the local file system. The master node is able to collect whole image files through the remote shared file system for display of the results. We can see the approach of data I/O access in 2 applications, due to individual problem features, each application is unique. Examples of benchmark cases reveal strong performance in accelerating computing speed and reducing the I/O operational time. It is found that the platform is able to accelerate climate data visualization processes, help research scientists gain the deep insights into data, and explore the potential phenomena and features, such as formation of Typhoon eddies. In air quality management applications, the platform is used to perform the air pollution model Gaussian plume trajectory model. Backward trajectory simulation of PM2.5 concentrations is used to identify the 30+ point source's contribution on 73 EPA monitor stations (receptors) in Taiwan. A user-friendly, web-service based big data presentation uses the heterogeneous observed and forecast pollutant data in space and time. The results support for air quality decision-making and emergency response. The limitation of data size for applications in the platform, the current users and future development of the platform, and the linkage of PRAGMA collaboration are also described in the paper.

Correction added on 13 May 2017, after first online publication: the order of authors has been corrected from “Jen-Gaw Lee, Ching-Yao-Lin, Whey-Fone Tsai, Hsi-Ching Lin, Lung-Cheng Lee, and Ben-Jei Tsuang” to “Jen-Gaw Lee, Whey-Fone Tsai, Lung-Cheng Lee, Ching-Yao-Lin, Hsi-Ching Lin, and Ben-Jei Tsuang”
INTRODUCTION

The advances in high-performance computing allow scientists using numerical modeling to enhance their understanding of natural phenomena and pinpoint engineering design problems. Due to the rapid progress of computing power and storage capacity in computers, scientists are able to predict physical phenomena and gain deep insights into the unknowns. However, due to the growth of computing scale and physical complexity, the data sizes produced by numerical simulation has increased explosively. As a consequence, high-performance computing accompanied by large datasets produced by model simulation has created new challenges in data processing and analysis. Under these circumstances, building a system to meet requirements in both high-performance computing (HPC) and data management and analyses becomes very crucial. An HPC-based big data platform is 1 of solutions to solve this type of problem.1

In general, numerical simulation of HPC uses domain decomposition methods to partition the computational domain, with the data in each domain being sent to separate computing nodes for computation. The results of computation are then sent back and stored in the data storage. This is the traditional approach in which data are bundled by the computing node. One example is the data server and processing procedure used in the parallel visualization software ParaView2 and VisIt.3 However, the approach differs in big data management and analysis in that the data are distributed to local data nodes for necessary processing. This is because it is more effective to do data processing in the data nodes rather than migrate large dataset to computing nodes. Parallel and distributed visualization is an important tool for exploring the unknowns from large, diverse data produced by climate and environmental forecasts. Accordingly, an HPC-based big data platform will be helpful for data processing in climate and environmental applications.

Regarding the climate and environmental-related applications, the output data produced by model simulations have a grid-field structure4 meaning that the data corresponding to spatial coordinates are coincident with those of computational grids. The grid-field structure is thus a structured data type. If we consider the programming adaptability of a numerical model, Hadoop streaming5 can be used to convert an existing numerical program, such as the popular Fortran language heavily used in engineering, into JAVA code. However, it is time-consuming to formulate Map-Reduce programming if the size of the existing numerical code is large. Consequently, Hadoop may not be ready for immediate application to climate data visualization and air quality model simulation problems. Basically, these problems are structured data problems, so a structured database (ie, relational Structured Query Language database) is sufficient for dealing with data processing in model simulation and visualization. Parallel processing of environmental big data for visualization is a challenging task. Data sizes can easily exceed hard memory limits. Some useful methods, such as In situ visualization6 or Query Driven Visualization (QDV)7-9 can be used in such case. In situ visualization implements a visualization function in the simulation model. The simulation produces a visualization, which can be saved in the output data of the model. The disadvantage of this approach is that it requires rerunning the model simulation if there is even a slight change in the visualization plan. The QDV method, in contrast, can be used to analyze existing data sets to read relevant data through the query of database.

Global climate forecasting produces large amounts of temporal and spatial data. The long-duration forecast, such as 45-day climate forecasts, can deliver climate phenomena longer than a 7-10 day forecast that weather agencies usually do, such that people can have more time to respond to changes in the weather. Visualization of climate data and overlaying onto earth geographical information system will make information more transparent and intuitive and provide greater context. Also, for air quality management and response in an area, it is crucial to assess the impact of air pollution caused by the geographically varied pollutant sources on locations of concern, ie, Environmental Protection Administration (EPA) monitoring stations. The uniqueness of climate data visualization and air quality model simulation in monitor stations is that it can be conducted by parallel processing without data communication among computing nodes. Therefore, the current proposed big data platform has adopted the concepts of QDV and shared-nothing distributed database,10-11 deploying both local distributed database, distributing the overall data to local data nodes for parallel post-processing, and for model simulations and analyses.

The system architecture of the experimental big data platform then is used in accordance with the above operation. Performance benchmarks associated with computation and data processing, applications to climate data visualization, and air quality simulation and the response can illustrate the capabilities of the platform. The platform users, future development, and associated PRAGMA collaborations will be highlighted as well.

SYSTEM ARCHITECTURE

2.1 Hardware

This big data platform is designed for experiments on 2 case studies: the climate data visualization and air quality simulation and response applications. The system architecture of this experimental big data platform is comprised of one master data node and 17 slave data nodes. The system links to the NCHC Advanced Large-scale Parallel Supercluster (ALPS),12 and storage pool, as shown in Figure 1. The ALPS supercomputer is responsible for global climate model simulation and the output is stored in the storage pool, which can hold 160 TB of data. The master node is the coordinator among ALPS, the storage
pool and 17 slave nodes. The network between ALPS and the master node is 10 GbE (Gigabit Ethernet; 10 Gbs bandwidth), the others are 1 GbE. For the air quality simulation and response application, simulation model and data analyses are both performed on slave nodes. The system’s total memory capacity is around 500 GB (16 GBx2x17). That means the data size limitation allowed in the platform is 500 GB. The platform is currently designed for accommodating this experiment. System scale-up performance is nearly linear due to the shared-nothing feature among data nodes. Therefore, upgrading the capacity of the hardware is straightforward when a production run is required in the future.

2.2 Software and methodology

The openSUSE operating system\textsuperscript{13} and the MariaDB database\textsuperscript{14} are installed on the master node and 17 slave nodes. The core concept of this big data analysis platform is in-place computing. The data are the key for in-place computing. Application programs are executed at the location where data are stored. That means data are distributed and stored on hard drives of slave nodes through the distributed file system. Then the application program is distributed to the data node to read local data and perform analyses locally. Each node has its own database installed. When the data analysis program carries out its application on the data node, it simply refers to the local database where data are situated. However, before the previously mentioned processes are performed, the whole data set should be partitioned into blocks among the database of each data node. Currently, the data are partitioned to blocks according to the timestamp attributes of the data. We use round robin as the policy for distributing blocks to slave nodes.

2.3 Application planning

Two applications of parallel post-processing: visualization of global climate simulation data, and air quality simulation and response; are selected for use as case studies for performance benchmark tests. Figure 2 shows the framework of these 2 applications on the big data platform. The global climate model simulation is executed on the ALPS supercomputer and post-processing is conducted simultaneously on the slave data nodes. The air quality model simulation and analyses are both performed in data nodes. A web service link to the database will be provided for viewing and exploring the insights of data.

The first step in the application of global climate data (OLR, temperature, rainfall, etc.) visualization is the partition of Network Common Data Form (netCDF) file data into shared-nothing distributed databases for partial visualization in slave data nodes. The results are then integrated into the overall visualization in the master node through Message Passing Interface communication. The air quality management application of the platform first accesses Taiwan EPA

---

**FIGURE 1** System hardware architecture of the big data platform. ALPS indicates Advanced Large-scale Parallel Supercluster

**FIGURE 2** Application framework in big data platform. ALPS indicates Advanced Large-scale Parallel Supercluster
observed data at master node. EPA observed data are then replicated to distributed databases in the slave nodes, and the air pollution model, Gaussian plume trajectory model (GTx), is replicated in all slave nodes for model simulation, which produces output data and associated image files on local file systems. The master node is able to collect whole image files through the remote shared file system for display of the results. The details are set forth below.

3 | APPLICATION 1: PARALLEL POST-PROCESSING OF GLOBAL CLIMATE DATA VISUALIZATION

3.1 | Workflow, benchmark, and application

The first case study of the platform application uses parallel post-processing for the visualization of global climate forecast data. Here, we have only selected limited cases for demonstration. The output data of climate model was in netCDF format, which is popularly used in atmospheric science. The grid size of the output data is 1536 × 768 in space, and 248 steps in time. The attributes of the simulation output include OLR, precipitation, temperature, wind field, and many others. The whole data in the netCDF dataset can be partitioned into local nodes, based on either spatial or temporal partition. In this case study, temporal partition has been adopted. For example, the output of climate model simulation produces 248 time series of data sets in netCDF format, which can be treated as 248 blocks. These blocks are in turn allocated to local databases on each data node, as shown in Figure 3, while the database on the master server node stores metadata, which contains the information on the data partition status (such as on which slave node each data block is stored, attribute names, dimensions, etc.) and integrates the results of the analysis accomplished in each slave data node.

Figure 4 shows the detailed processing workflow of the post-processing for visualization task. A client-server framework is implemented, allowing the user to communicate with the master server node. The user client then proposes an analysis demand to the server. Based on the demand, the master server node retrieves metadata from its database and distributes the information to the slave nodes. Consequently, slave data node is able to understand what data are required by using the metadata from the master node, and to retrieve data by querying local databases, and by then performing necessary analysis/computation. Because the data are partitioned in accordance with time steps, each dataset is independent of the others, and the only communications needed are between master and slave nodes. These communications are performed by Message Passing Interface. The analysis or computation results from the local nodes are then sent back to the master server node. After the master node collects the computation results from the slave nodes, the visualization assembling can be carried out on the master node and the resultant images sent back to the user. Alternatively, the user can retrieve all the computation results and can visualize the result on a personal machine.

The visualization of global climate forecast data starts with OLR data. For practical visualization of OLR data produced by Hiram model simulation, the data are stored in the slave nodes. Each temporal OLR data set is analyzed on a slave node to compute contour lines, which are then collected and integrated on the master node. The master server node, therefore, can display each temporal image, as seen in Figure 5, as well as display the sequential animation through the web service. In total there are 248 time steps of data, which are evenly distributed to 17 slave nodes. Each slave node is therefore responsible for processing around 14 ~ 15 temporal data sets.

In the left figure, the white color represents strong vertical convection, and the near transparent color indicates high-level clouds with a lower chance of precipitation. Precipitation in the right figure OLR data is represented by a color table. The result can be projected either on a 3-D sphere/earth (left) or shown on a 2-D map (right).

This approach can be applied not only for post-processing of OLR data, but also for other climate variables, such as temperature, precipitation, etc. Thus, the distributed system of big data platform can be further used to carry out the computation of contour lines for visualization of other variables simulated by European Hamburg TaiWan (EHTW) global climate model.

Table 1 shows the performance benchmark of 2 different cases. It takes ~670 seconds to compute the OLR simulation data set sequentially, but only takes ~58 seconds to do so on the present distributed parallel system. For a precipitation simulation data set, it took 13.2 seconds for sequential computation and only 2.3 seconds while using the distributed approach. The reason for the reduced time for a precipitation data set is because of the distributed nature, each data set is much smaller than the OLR data set.

For evaluating the I/O operation in the big data platform, we have also performed a case study, which let the distributed system query a single shared database (instead of distributed databases) to simulate a shared file system. It takes about 256 seconds for OLR visualization in this case; whereas, using the distributed system with distributed local database only takes 58 seconds. The reduction in total processing time from 256 seconds (shared database) to 58 seconds (distributed database) is largely due to reduced I/O operations. This is the primary advantage of “in-place computation” in the big data platform.

It is worth noting that, for purposes of this experiment, the current network interconnection is just 1 Gigabit Ethernet between the master node and the slave nodes. The performance could be even better if the interconnection were upgraded. The selected visualization results of

FIGURE 3 NetCDF file partition for parallel visualization. NetCDF indicates Network Common Data Form
post processing of temperature and rainfall are shown in Figures 6 and 7, respectively. This work demonstrates that the computing and data processing time can be accelerated by the proposed big data platform when dealing with global climate forecast data visualization. However, the performance benchmark only measures the overall computing (computation and data processing) time, it does not quantify the time spent on partitioning and distributing blocks to local slave nodes.

The visualization application was created using C/C++ and OpenGL. Two different display setups are available here. One displays the simulation results overlaid on a 2-D map (like Figure 6 and left in Figure 7); another displays the results overlaid on a 3-D earth (right in Figure 7).

### Table 1

| Data                     | OLR simulation | Precipitation simulation |
|--------------------------|----------------|----------------------------|
| Grid size dimension     | 1536 × 768     | 320 × 160                  |
| Time steps               | 248            | 181                        |
| Sequential computation   | ~670 sec       | ~13.2 sec                  |
| Parallel and distributed system |           |                             |
| Nodes                    | 17             | 17                         |
| Computation              | ~58 sec        | ~2.3 sec                   |
| Speedup                  | 11.55          | 5.74                       |

Abbreviations: OLR indicates Outgoing Longwave Radiation.
FIGURE 6  Schematic visualization of global climate rainfall data

FIGURE 7  Schematic visualization of global climate temperature data in 2-D map and 3-D earth

FIGURE 8  The tracking of typhoon eddy formation through the OLR visualization/animation. OLR indicates Outgoing Longwave Radiation
visualization/animation. The top portion of Figure 8 shows the formation of a typhoon with clear eddy formation located in the Pacific Ocean approaching Taiwan. The lower portion of Figure 8 shows that the outer ridge of the typhoon eddy has touched Taiwan.

4 | APPLICATION 2: AIR QUALITY SIMULATION AND RESPONSE

In this study, the air quality open data observed by the Taiwan EPA is used as inputs to run the air quality simulation model to explore the contribution of primary pollutants (point sources) on EPA monitoring stations (receptors). The model simulation is carried out on the slave nodes through the distributed and parallel processing in the big data platform. The contribution influence of pollutant point sources on the receptors can be assessed through a series of runs over a period of time, and the assessed results can be used to warn the public and to respond to severe air pollution situations in a city by controlling emissions of the pollutant source in accordance with government policy.

Currently, Taiwan EPA maintains 76 air quality monitoring stations. Each station provides pollutant standards index (PSI), SO₂ (Sulphur dioxide), CO (Carbon Monoxide), O₃ (Ozone), PM10 (Particulate Matter of less than 10 micrometers in diameter), PM2.5 (Floating Particulate Matter of less than 2.5 micrometers in diameter), NO (Nitrogen Oxide), NO₂ (Nitrogen Dioxide), NOₓ (Nitrogen Oxides), wind speed, and wind direction data. These are updated every hour and published on an EPA open data website to comply with the open governmental information policy in Taiwan. In addition to the EPA open data, a ground surface weather map made by the Taiwan Central Weather Bureau is updated daily; this can be used to support information on weather conditions, especially the wind field related data required for air quality model simulation.

4.1 | Air quality simulation model

The air quality simulation model used in the big data platform is developed based on the GTx, which has been improved by Tsuang et al. The advantage of GTx is that the source-receptor relationship can be determined in a single model run. The contribution of a pollutant point source to a receptor can be simulated in each model run. This is defined as forward trajectory simulation. Similarly, the specific receptor at which the air pollution is affected by various pollutant point sources can also be evaluated through model simulation; and it is defined as backward trajectory simulation. It is possible to run various forward and backward trajectory simulations to figure out the influence of major pollutant point sources. Nonetheless, the current application is focused on backward trajectory simulation trying to find which pollutant point source has had the greatest influence on given monitoring stations. When very poor air quality is found in the simulation, the response is either to warn the public or to enforce government regulation to reduce emission of pollutants.

4.2 | Workflow of air quality model simulation

The workflow of air quality model simulation is shown in Figure 9. The first step is to develop a program to read and parse the EPA open data through the web service, and then to write the EPA data to the master database. Once the master database node is updated, all slave database nodes will be updated automatically by the database automatic replication mechanism. The database provided incremental replication to all data nodes. This means that only newly added records are replicated to slave nodes.

The GTx simulation model, as introduced above, is installed on each slave node, and EPA monitoring data is also replicated to each slave node. Then, the GTx simulation model can read local data and execute simulation without needing to read input data from other nodes. The advantage of this in-place computation is that it saves time in moving huge data. All model simulation and analyses are conducted in the slave nodes, while web service of data presentation is executed in the master node, which collects the resultant data from the slave nodes.

4.3 | Performance benchmark

Figure 10 shows the detailed processing of GTx simulation model and the performance benchmark. The simulation model uses observed data from the past 2 days to predict air quality for the following day. The air quality data observed at EPA monitoring station is stored in a MariaDB. The “traj” subroutine reads EPA monitoring stations data as input to set up the initial condition. The output is stored on a local hard disk; then the GTx simulation is launched. This reads the traj’s output as input for computation. The GTx’s output is also stored on a local hard disk for “grads” subroutine to draw an output graph.

Here, only 73 EPA monitoring stations are taken into consideration for assessment; data from the other 3 stations are not sufficient. The 2007 wind field data observed by EPA are complete for the whole year. Therefore, the EPA 2007 observation wind field data are adopted.
FIGURE 10  Performance benchmark of sequential process, distributed process with shared file system, and distributed process with local file system (in-place I/O). Local HDD indicates Local hard disk drive; Remote HDD, Remote hard disk drive.

FIGURE 11  EPA monitoring stations: geolocation, IP camera and hourly air quality observed data. EPA indicates Environmental Protection Agency; IP, Internet Protocol.
for assessment in this study. In the benchmark, the first step is to repeat the sequential process by using the master node 73 times for the 73 EPA monitoring stations, this takes 540 minutes. Second, the 17 available slave nodes (each node has 4 cores and 8 threads) solve the same problem concurrently with a mounted shared file system at master node; this takes 50 minutes. Finally, we use the 17 slave nodes to solve the same problem concurrently with local database and file system, this takes only 10 minutes. The total processing time is therefore reduced from 50 minutes (shared file system: Remote hard disk drive) to 10 minutes (In Place I/O: Local hard disk drive) due to reduced I/O operations. That is the advantage of using in-place computation in the current developed big data platform.

Each slave node’s computation result is a text file, which is stored on its local hard disk. A simple script has been created to parse the output text file, extract each point source’s contribution value and then write back to the master node database. The master node database is used for web-based data visualization, which will be introduced in next section.

### 4.4 A user-friendly web-based data presentation

As mentioned above, the Taiwan EPA has maintained 76 air quality monitoring stations. Each station provides pollutant standards index (PSI), SO₂, CO, O₃, PM10, PM2.5, NO, NO₂, NOₓ, wind speed, and wind direction monitor data. The data are updated every hour and published on EPA open data website. The data are used as input to run the pollution model GTx backward trajectory simulation of PM2.5 concentrations to identify the contribution of each of 30+ point sources on 73 monitoring stations (receivers). To meet the need of practical operation, a user-friendly web service–based big data presentation uses the observed and forecast pollutant data in space and time.

The results support for decision-making and emergency response to the challenging air quality situations. This website is developed using ASP.Net 4.5 programming language and hosted on Microsoft Windows server 2012 R2; the database used is MariaDB. AMCHARTS JavaScript charting library is selected to draw different charts, while Google Map API is used to draw geolocation of monitoring stations. Significant effort was spent on interviewing domain experts and users to identify real world operational needs, to develop and revise a prototype, and to design a database scheme.

#### 4.4.1 EPA monitoring stations’ data

As shown in Figure 11, the Google map on the left hand side is used to show the geolocation of the EPA monitoring stations. Each monitoring station on the Google map is a hyperlink. When a user selects a monitoring station, the selected monitoring station’s IP camera image and every hour’s air quality data are displayed on the right hand side.

#### 4.4.2 Assessment of PM2.5 contribution by point sources at the selected monitoring stations

Figure 12 shows the ranking of PM2.5 concentrations contributed by 30+ major point sources to a selected monitoring station, Fengyuan. The total PM 2.5 concentration simulated on April 20, 1997 is 100.9 μg/m³ and 24.176 μg/m³ is contributed from the Taichung coal power plant. The user can use the scroll bar at the top to zoom in to a time-period of interest. The major pollutant point source listed at bottom is a hyperlink, which can be clicked to show or hide its concentration contribution on the chart. Figure 13 shows only the total concentration and concentrations contributed by the top 2 point sources. This chart indicates that the first one is the...
4.4.3 Air pollution response for decision maker

It is effective to respond to severe air pollution by reducing the emissions of air pollutant point sources. Consequently, a webpage is designed, as shown in Figure 14, to allow decision makers to adjust and reduce a point source’s PM 2.5 emissions, to alleviate the PM 2.5 concentration on the area of selected monitoring stations. The reduction of point source emissions at power plant or factories needs government regulation. This provides evidence of their impacts to regional air quality. The Google map on the left hand side shows the
geolocation of EPA monitoring stations. A user can select a station of interest; then the right hand side shows for the current day or a selected day in the past, the PM2.5 concentrations contributed by each pollutant point source to the selected monitoring station.

The user can change each point source’s contribution rate by selecting the reduction rate from 1.0 to 0.0, at intervals of 0.1. Then the system will calculate and show the new contribution values for each affected point source at the selected monitoring station. The case in Figure 14 shows that the original total PM 2.5 concentration was 100.9 μg/m3. The user tries to adjust by reducing the Taichung coal power plant by 70% and the Taiwan Chemistry Company at Changhua by 60%. The recalculated PM 2.5 concentration is 69.9 μg/m3 which is lower than the Taiwan EPA’s highest air quality standard of 72 μg/m3.

By default, the new contribution value is calculated using simple computation of the original contribution value and the reduction rate. If necessary, the system can invoke the simulation model to recalculate. This function allows local and central government to take quick action in response to urgent air-pollution situations.

5.1  Summary

Global climate forecast data visualization and air quality model simulation require a system that is capable of handling large data and high-speed computation. The proposed experimental big data platform uses the HPC distributed computing and in-place computing. Master and slave distributed databases are used in the data nodes. The database MariaDB is set up at slave cluster nodes, which are responsible for data storage and distributed parallel computing, and at the master data node, which is used for managing metadata and further processing the information obtained from the cluster nodes. The hardware and software system of the platform is easy to install. This platform currently is designed for large data and high-speed computation experiments. The system’s memory capacity in total is around 500 GB. System scale-up performance is nearly linear due to the shared-nothing feature among data nodes. Therefore, upgrading the capacity of hardware is straightforward when a production run is required in the future.

The benchmark performance of the big data platform has been carried out by using case studies in global climate data visualization and air quality model simulation. Both studies were conducted on 3 performance cases of benchmark: (1) sequential process, (2) distributed process with shared file system, and (3) distributed process with local file system (current big data platform solution). The computational time of case (1) compared with that of case (3) shows that the speedup ratio is 6 – 11 for global climate data visualization, and speedup ratio is 10 for air quality model simulation; while a comparison of the processing time between case (2) and case (3) infers that the reduction rate of I/O operation is about 5 for both cases. The benchmark results demonstrate the advantage of in-place computation in the current developed big data platform.

The total output variables of the global climate forecast model are over 20. In this paper, we only use the most important variables (OLR, precipitation and temperature variables) as a demonstration. The visualization of the application was created by using C/C++ and OpenGL. The visualization results can be overlaid on a 2-D map or on a 3-D earth. Based on the global climate data visualization, it is also shown that the eddy formation of potential typhoons can be tracked using OLR data animation with the current big data platform.

The air quality model GTx is used in the big data platform for backward trajectory simulations to isolate the contribution of specific major pollutant point sources to monitoring stations. The workflow from model simulation, data visualization, and web service has been integrated for automate execution and display. It is worth mentioning that the assessment of air quality data is relevant to 76 EPA sensor stations, 30+ pollutant point sources, and a few air quality indexes. A big data presentation web service is designed to illustrate the variety of types and multiple dimensions of data associated with air pollution. Managers and decision makers are able to use these results to assess the significance of air pollution from the web service and make appropriate decisions.

It is effective to improve severe air quality by reducing the emissions of pollutant point sources. Consequently, a webpage is designed to allow decision makers to adjust and reduce the PM 2.5 emissions of point sources to alleviate the PM 2.5 concentration in the area of selected monitoring stations. This has been tentatively applied in response to severe air pollution in Taichung city.

5.2  Lessons learned

The in-place query driven big data platform illustrated in this paper is an experimental one. The benchmark of examples shows strong performance in accelerating computing time and reducing the I/O operational time. However, the case studies are limited to a single run of data each. Daily forecasting and information services in environmental monitoring require the establishment of standardized procedures and upgrade platform capacity for series forecasts and automated processing of data. This can be done in conjunction with the research development of the newly awarded MOST (Ministry of Science and Technologies, Taiwan) project, “Development of global climate big data platform, visualization technologies, and associated value-added applications and service (July, 2016).” Automated processing system of climate data visualization will be established. Currently, selective climate visualization data have been provided to PRAGMA and CENTRA members, Osaka University, AIST and the University of Hawaii for collaboration on synchronous rendering of multiple climate variables in SAGE2 tile display walls. It is worth mentioning that the present users of global climate data visualization in the big data platform are research scientists of Chung-Hsing University and the Academia Sinica in Taiwan; when the automated processing system of climate data visualization is deployed, it will allow more users in Taiwan, as well as the PRAGMA community, to be served.

The current big data platform has been used for air quality forecast and response analyses in Taichung City and the Yun-Lin County, Taiwan. Another application of the platform is to identify countermeasures, through a series of simulations and analyses, for minimizing air pollution induced by straw burning in areas of southern Taiwan.
Air quality management is an important ingredient in smart city development. In addition to the air quality simulation and response topic described here, future applications are planned, such as air pollution sensor network. Currently, PRAGMA members, Academia Sinica, Osaka University, NCHC, and Nanyang Technological University in Singapore have agreed to collaborate on testing an air pollution sensor network under Internet of Things network environment, such as the Lo-Ra Wide Area Network.

ACKNOWLEDGMENTS

We appreciate Dr. Bill Chang for his suggestions on this paper. Dr. Chang recently retired from the National Science Foundation in the U.S. as the Head of the East Asia Pacific Regional Office. The authors also acknowledge the EPA and the Central Weather Bureau, Taiwan, for providing weather and air quality observation data.

ADDITIONAL AUTHORS

I-En Liao (National Chung-Hsing University, Taiwan, Email: ieliao@nchu.edu.tw); Chia-Ying Tu, Academia Sinica, Taiwan, Email: cytu@gate.sinica.edu.tw); Kai Chen Ku (National Chung-Hsing University, Taiwan, Email: dodo1654.ku@gmail.com); Shinyii Liu (National Chung-Hsing University, Taiwan, Email: shinyii.liu@gmail.com); T.-M. Chuang (National Center for High-performance Computing, NARL, Taiwan, Email: tmchuang@nchc.narl.org.tw); Wen-How Cheng (National Chung-Hsing University, Taiwan, Email: tom177832@gmail.com).

REFERENCES

1. Lee LC, Tsai WF, Lin SC, Hsu WC. Big data post processing for active magnetic regenerator thermal flow simulation. Paper No. 27-17, The 21st Taiwan CFD conference, Nantou, Taiwan. 2014.
2. ParaView software. http://www.paraview.org/
3. VisIt software. https://wci.llnl.gov/codes/visit/
4. Howe B, Maier D. Algebraic manipulation of scientific datasets. Proceedings of the 30th VLDB Conference, Toronto, Canada. 2004.
5. Hadoop: http://hadoop.apache.org/
6. Ma KL. In situ visualization at extreme scale: challenges and opportunities computer graphics and applications. IEEE. 2009;29:14–19.
7. Stockinger K, Shalf J, Wu K, Bethel EW. Query-driven visualization of large data sets. Visualization Conference; IEEE. 16th IEEE Visualization, p 22. 2005.
8. Gosink LJ, Anderson JC, Bethel EW, Joy KL. Query-driven visualization of time-varying adaptive mesh refinement data. IEEE Trans Vis Comput Graph. 2008;14:1715–22.
9. Zou HB, Slawinska M, Schwan K, et al. FlexQuery: an online query system for interactive remote visual data exploration at large scale. IEEE Cluster 2013, Indianapolis, IN. 2013.
10. Stonebraker M. The case for shared nothing. Database Engineering Bulletin. 1986. 1986;9(1):14–9.
11. Müseler T. A survey of shared-nothing parallel database management systems. IRCSE ’12 IDT Mini-conference on Interesting Results in Computer Science and Engineering, Nov. 8, Mälardalen University, Sweden. 2012.
12. Advanced Large-scale Parallel Supercluster (ALPS), http://www.nchc.org.tw/en/
13. OpenSUSE: https://www.opensuse.org
14. MariaDB: https://mariadb.com
15. Tu C-Y, Tsuang BJ. Cool-skin simulation by a one-column ocean model. Geophys Res Lett. 2005;32(22):L22602.
16. Tsuang BJ, Tsai JL, Lin MD, Chen CL. Determining aerodynamic roughness using tethersonde and heat flux measurements in an urban area over a complex terrain. Atmos Environ. 2003a;37(14):1993–2003.
17. NetCDF: http://www.unidata.ucar.edu/software/netcdf/
18. Tsuang BJ. Quantification on the source/receptor relationship of primary pollutants and secondary aerosols by a Gaussian plume trajectory model: Part I—Theory. Atmos Environ. 2003;37(28):3981–3991.
19. Tsuang BJ, Chen CL, Lin CH, et al. Quantification on the source/receptor relationship of primary pollutants and secondary aerosols by a Gaussian plume trajectory model: Part II. Case study. Atmos Environ. 2003b;37(28):3993–4006.
20. AMCHARTS JavaScript Charts V.3 charting library

How to cite this article: Lee J-G, Tsai W-F, Lee L-C, Lin C-Y, Lin H-C, Tsuang B-J. In-place query driven big data platform: Applications to post processing of environmental monitoring. Concurrency Computat: Pract Exper. 2017;29:e4135. https://doi.org/10.1002/cpe.4135