Scalable Computing: Practice and Experience
Volume 11, Number 2, pp. 131–136. http://www.scpe.org
ISSN 1895-1767

DISTRIBUTED DATA INTEGRATION AND MINING USING ADMIRE TECHNOLOGY*

ONDREJ HABALA†, MARTIN SELENG†, VIET TRAN†, LADISLAV HLUCHY†,
MARTIN KREMLER‡, AND MARTIN GERA‡

Abstract. In this paper we present our work on the engine for integration of environmental data. We present a suite of selected environmental scenarios, which are integrated into a novel data mining and integration environment, being developed in the project ADMIRE . The scenarios have been chosen for their suitability for data mining by environmental experts. They deal with meteorological and hydrological problems, and apply the chosen solutions to pilot areas within Slovakia. The main challenge is that the environmental data required by scenarios are maintained and provided by different organizations and are often in different formats. We present our approach to the specification and execution of data integration tasks, which deals with the distributed nature and heterogeneity of required data resources.

Key words: environmental applications, distributed data management, data integration, OGSA DAI

1. Introduction. We present our work in the project ADMIRE, where we use advanced data mining and data integration technologies to run an environmental application, which uses data mining instead of standard physical modeling to perform experiments and obtain environmental predictions. The paper starts with description of the project ADMIRE, its vision and goals. Then we describe the history and current status of the environmental application. The core of the paper then presents our approach to the integration of data from distributed resources. We have developed a prototype of data integration engine that allows users to specify data integration process in form of a workflow of reusable processing elements. This paper has been originally presented in [10].

1.1. The EU ICT Project ADMIRE. The project ADMIRE (Advanced Data Mining and Integration Research for Europe [1]) is a 7th FP EU ICT project aims to deliver a consistent and easy-to-use technology for extracting information and knowledge from distributed data sources. The project is motivated by the difficulty of extracting meaningful information by mining combinations of data from multiple heterogeneous and distributed resources. It will also provide an abstract view of data mining and integration, which will give users and developers the power to cope with complexity and heterogeneity of services, data and processes. One of main goals of the project is to develop a language that serves as a canonical representation of the data integration and mining processes.

1.2. Flood Forecasting Simulation Cascade. The Flood Forecasting Simulation Cascade is a SOA-based environmental application, developed within several past FP5 and FP6 projects [2], [3], [4]. The application's development started in 1999 in the 5th FP project ANFAS [5]. In ANFAS, it was mainly one hydraulic model (the FESWMS [6]). It then continued with a more complex scenario in 5th FP project CrossGrid, turned SOA in 6th FP projects K-Wf Grid and MEDIgRID, and finally extended the domain to environmental risk management in ADMIRE. The application is now comprised of a set of environmental scenarios, with the necessary data and code to deploy and execute them. The scenarios have been chosen and prepared in cooperation with leading hydro-meteorological experts in Slovakia, working mainly for the Slovak Hydrometeorological Institute (SHMI), Slovak Water Enterprise (SWE), and the Institute of Hydrology of the Slovak Academy of Sciences (IH SAS). We have gathered also other scenarios from other sources, but in the end decided to use the ones presented below, because they promise to be the source of new information for both the environmental domain community, as well as for the data mining community in ADMIRE. Together with the scenarios, we have gathered a substantial amount of historical data. SWE has provided 10 years of historical data containing the discharge, water temperature, and other parameters of the Vah Cascade of waterworks (15 waterworks installations and reservoirs in the west of Slovakia). SHMI has provided 9 years of basic meteorological data (precipitation, temperature, wind) computed by a meteorological model and stored in a set of GRIB (Gridded Binary) files, hydrological data for one of the scenarios, and also partial historical record from their nation-wide

*This work is supported by projects ADMIRE FP7-215024, APVV DO7RP-0006-08, DMM VMSP-P-0048-09, SMART ITMS: 26240120005, SMART II ITMS: 26240120029, VEGA No. 2/0211/09.
†Institute of Informatics of the Slovak Academy of Sciences, Dubravska cesta 9, 84507 Bratislava, Slovakia ({Ondrej.Habala, Martin.Selen, Viet.Tran, Ladislav.Hluchý}@savba.sk)
‡Comenius University, Faculty of Mathematics Physics and Informatics, Mlynska dolina, 84248 Bratislava, Slovakia

TABLE 2.1
Depiction of the predictors and variables of the ORAVA scenario

Time	Rainfall	$Temp_{Air}$	Discharge	$Temp_{Reservoir}$	$Height_{St}$	$Temp_{St}$
T-2	R_{T-2}	F_{T-2}	D_{T-2}	E_{T-2}	X_{T-2}	Y_{T-2}
T-1	R_{T-1}	F_{T-1}	D_{T-1}	E_{T-1}	X_{T-1}	Y_{T-1}
T	R_T	F_T	D_T	E_T	X_T	Y_T
T+1	R_{T+1}	F_{T+1}	D_{T+1}	E_{T+1}	X_{T+1}	Y_{T+1}
T+2	R_{T+2}	F_{T+2}	D_{T+2}	E_{T+2}	X_{T+2}	Y_{T+2}

network of meteorological data. They have also provided several years of stored weather radar data, necessary for one of the scenarios. The programs used by the application are in the context of ADMIRE described in Data Mining and Integration Lanuage (DMIL) [7]. The processes described in DMIL perform data extraction, transformation, integration, cleaning and checking. Additionally, in some scenarios we try to predict future values of some hydro-meteorological variables; if necessary, we use a standard meteorological model to predict weather data for these cases.

2. Environmental Scenarios of ADMIRE. In this chapter we present the suite of environmental scenarios, which we use to test the data mining and integration capabilities of the ADMIRE system. The scenarios are part of the Flood Forecasting and Simulation Cascade application, which has been in the meantime expanded beyond the borders of flood prediction into a broader environmental domain. There are four scenarios, which are in the process of being implemented and deployed in the ADMIRE testbed. These scenarios have been selected from more than a dozen of candidates provided by hydro-meteorological, water management, and pedological experts in Slovakia. The main criterion for their selection was their suitability for data mining application. The scenarios are named ORAVA, RADAR, SVP, and O3, and they are in different stages of completion, with ORAVA being the most mature one, and O3 only in the beginning stages of its design.

2.1. ORAVA. The scenario named ORAVA has been defined by the Hydrological Service division of the Slovak Hydrometeorological Institute, Bratislava, Slovakia. Its goal is to predict the water discharge wave and temperature propagation below the Orava reservoir, one of the largest water reservoirs in Slovakia.

The pilot area covered by the scenario (see Figure 2.1) lies in the north of Slovakia, and covers a relatively small area, well suitable for the properties of testing ADMIRE technology in a scientifically interesting, but not too difficult setting.

The data, which has been selected for data mining, and which we expect to influence the scenario's target variables—the discharge wave propagation, and temperature propagation in the outflow from the reservoir to river Orava—is depicted in Table 2.1.

For predictors in this scenario, we have selected rainfall and air temperature, the discharge volume of the Orava reservoir and the temperature of water in the Orava reservoir. Our target variables are the water height and water temperature measured at a hydrological station below the reservoir. As can be seen in Figure 2.1, the station directly below the reservoir is no.5830, followed by 5848 and 5880. If we run the data mining process in time T, we can expect to have at hand all data from sensors up to this time (first three data lines in Table 1). Future rainfall and temperature can be obtained by running a standard meteorological model. Future discharge of the reservoir is given in the manipulation schedule of the reservoir. The actual data mining targets are the X and Y variables for times after time T (T being current time).

2.2. RADAR. This experimental scenario tries to predict the movement of moisture in the air from a series of radar images (see for example). Weather radar measures the reflective properties of air, which are transformed to potential precipitation before being used for data mining. An example of already processed radar sample (with the reflection already re-computed to millimeters of rainfall accumulated in an hour) can be seen in Figure 2.2.

The scenario once again uses both historical precipitation data (measured by sensors maintained by SHMI) and weather predictions computed by a meteorological model. Additionally to these, SHMI has provided several years' worth of weather radar data (already transformed to potential precipitation).

2.3. SVP. This scenario, which is still in the design phase, is the most complex of all scenarios expected to be deployed in the context of ADMIRE. It uses the statistical approach to do what the FFSC application

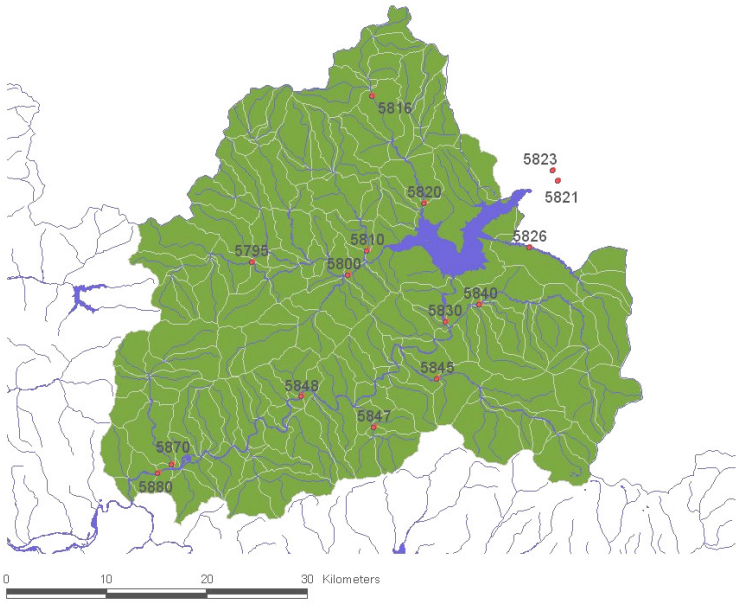

FIG. 2.1. *The geographical area of the pilot scenario ORAVA*

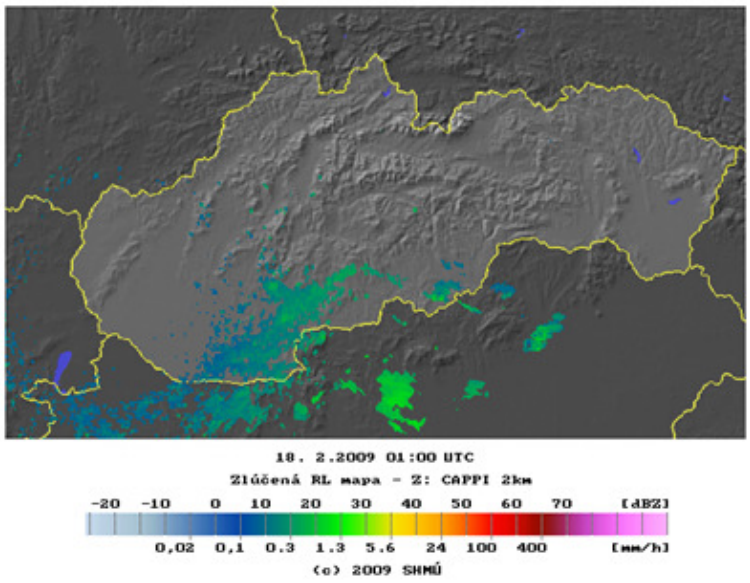

FIG. 2.2. *An example of weather radar image with potential precipitation*

did before ADMIRE—predict floods. The reasons why we decided to perform this experiments are mainly the complexity of simulation of floods by physical models when taking into account more of the relevant variables, and the graceful degradation of results of the data mining approach when facing incomplete data—in contrast to the physical modeling approach, which usually cannot be even tried without having all the necessary data.

For predicting floods, we have been equipped with 10 years of historical data from the Vah cascade of waterworks by the Slovak Water Enterprise, 9 years of meteorological data (precipitation, temperature, wind) computed by the ALADIN model at SHMI, hydrological data from the river Vah, again by SHMI, and additionally with measured soil capacity for water retention, courtesy of our partner Institute of Hydrology of the Slovak Academy of Sciences. We base our efforts on the theory, that the amount of precipitation, which actually

reaches the river basin and contributes to the water level of the river is influenced by actual precipitation and its short-term history, water retention capacity of the soil, and to lesser extent by the evapotranspiration effect.

3. Data integration engine for environmental data. In this section, we discuss the data integration engine designed for the environmental data integration and mining. It is motivated by the scenarios described in previous section. We first describe requirements that we took into account and then we present our approach to environmental data integration. In the discussion, we give examples mainly from Orava scenario; the first scenario implemented using our data integration engine.

In Orava river management scenario, the data from three different sources are used. The data are owned and maintained by different organizations. To allow the data mining operations proposed for this scenario, the data from those different sources must be integrated first. Furthermore, the data are kept in different formats. In the case of Orava scenario, two data sets are stored in relational database (waterworks data, water stations measurements) and one is kept in binary files (precipitation data are stored in GRIB files—binary file format for meteorological data). From technical point of view, we must be able to work with the heterogeneous data stored in distributed, autonomous resources. In our work, we have considered so far the data in the form of lists of tuples.

In the following, we use the term data resource to denote a service providing access to data, with a single point of interaction. We use the term processing resource to denote a service capable of performing operations on the input lists of tuples. Data resource can have capabilities of a processing resource.

Atomic units used for data access and transformations are called processing elements (PE). Following types of processing elements are needed:

- Data retrieval PEs—operations able to retrieve the data from different, heterogeneous data sources. Data retrieval PEs are executed at data resources. This class of PEs is also responsible for transforming raw data sets to the form of tuples.
- Data transfer PE—able to transfer list of tuples between distinct processing resources.
- Data transformation PE—operations that transform input list of tuples. These PEs can perform data transformation on per tuple basis, or can be used to aggregate tuples in the input lists.
- Data integration PEs—given input lists of tuples, data integration operations combine the tuples from input lists into a coherent form.

An operation has one or more inputs and one or more outputs. Inputs can be either literals or list of tuples and a outputs are list of tuples. Operations can be chained to form a data integration workflow—an oriented graph, where nodes are operations and edges are connection of inputs and outputs of the operations.

The term Application Processing Element (APE) will denote a data integration workflow that can be executed at a single resource. APE is a composition of atomic operations that provides functionality required by a data integration task. For example, in Orava scenario we use the precipitation data from GRIB files. The GRIB reader processing element extracts the data from GRIB files; it has two inputs—the first is a list of GRIB files and the second is a list of indexes in GRIB value arrays. The GRIB reader activity outputs all the values at input indexes from all the input files. We use an operation that queries the GRIB metadata database to determinate GRIB files of interest and another operation that transform given geo-coordinates in WGS84 to the indexes consumed by GRIB reader activity. This small workflow of three operations forms a single APE that provides precipitation data for given time period and geo-coordinates. The idea behind APE is to provide data integration blocks that can be executed at a single processing or data resource and can be reused for in multiple data integration tasks. Similarly to atomic PE, the inputs of APE can be literals or list of tuples and outputs are list of tuples.

The goal of our proposed data integration engine is to provide means of executing data integration tasks that are composed of multiple APEs and can integrate the data from distributed, autonomous and possibly heterogeneous data resources. Our data integration engine is designed to run the data integration tasks, given the input parameters and the APE workflow specification.

APE workflow specification is composed of four components: definition of APEs instances, mapping between inputs and outputs of connected APEs, mapping between the definition of integration task parameters and the parameter inputs of APEs in workflow and the definition of the result output.

In alignment with ADMIRE project vision, the APEs are specified in Data Mining and Integration Language (DMIL) [7] that is being developed within the project. The goal of DMIL is to be a canonical representation of data integration process, described in an implementation independent manner. The APE instance is specified by

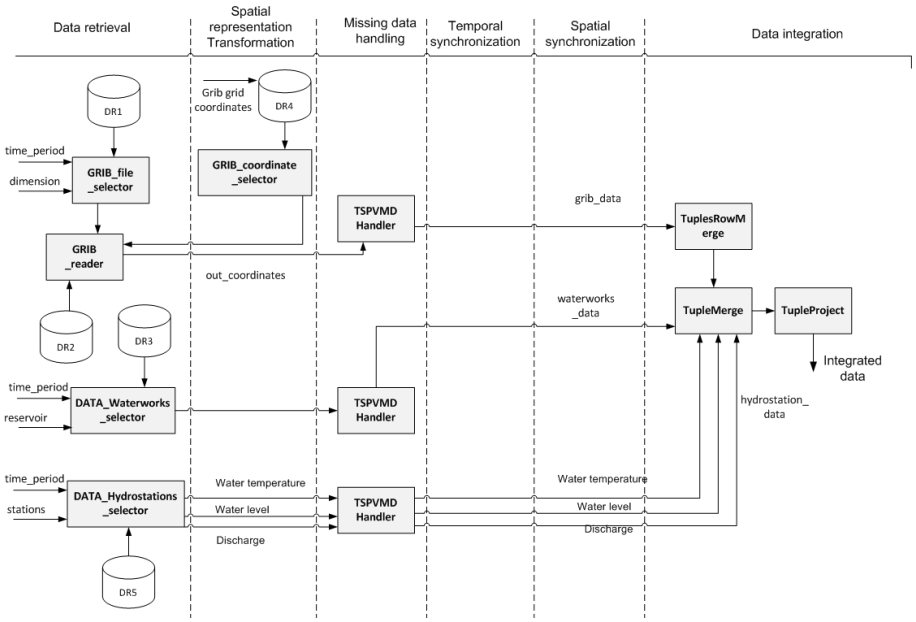

FIG. 3.1. *Orava river management scenario - APEs workflow*

the DMIL description of the process that should be executed, the specification of the data/processing resource it should be executed at and APE instance identifier that is unique within the APE workflow specification. Figure 3.1 depicts the APEs workflow of the Orava river management scenario.

In our view, the main advantage of proposed data integration engine is that user can specify sub-workflows that are executed on a separate data resources and the engine automatically connects the results of APEs executed on distributed resources. This helps to deal with the complexity of the distributed data integration.

3.1. Implementation. The prototype of proposed data integration engine for environmental data (DIEED) is implemented in JAVA programming language. It uses OGSA-DAI ([8], [9]) framework as the platform for exposing data resources in the distributed testbed and for executing the partial workflows of processing elements; it also provide us with the data transfer capabilities and streaming of the list of tuples between remote nodes. The data integration engine takes as inputs the integration task parameters and APE workflow specification. From the APE workflow specification, the engine constructs an oriented graph of APEs (defined by the mapping between inputs and outputs of APEs). For each node of the graph (containing an APE specified in DMIL) the DIEED performs following actions:

1. Compiles DMIL code—the DMIL specification of the node process is compiled to JAVA class that constructs an OGSA-DAI workflow.
2. JAVA class containing OGSA-DAI workflow is compiled by JAVA compiler, it is instantiated and OGSA-DAI workflow object is created
3. workflow object is submitted to OGSA-DAI service for execution
4. workflow execution on remote server is monitored

The whole APEs workflow is monitored during execution (providing information on the state of each of APEs); after execution is finished, the results can be retrieved in form of WebRowSet object.

DIE was integrated with the toolkit being developed in the project; this allows the user to submit APEs workflows, visualize the specified workflow and monitor its execution via graphical user interface based on Eclipse platform. Figure 3.2 depicts the graphical user interface for DIEED.

4. Conclusion. In this paper, we have presented preliminary results of our ongoing work on the data integration engine for environmental data that is being developed in the scope of ADMIRE project. We have first described four scenarios dealing with the integration and mining of environmental data. The main challenge is that the environmental data required by scenarios are maintained and provided by different organizations and are often in different formats. Our work concentrated on providing a platform that would allow integration

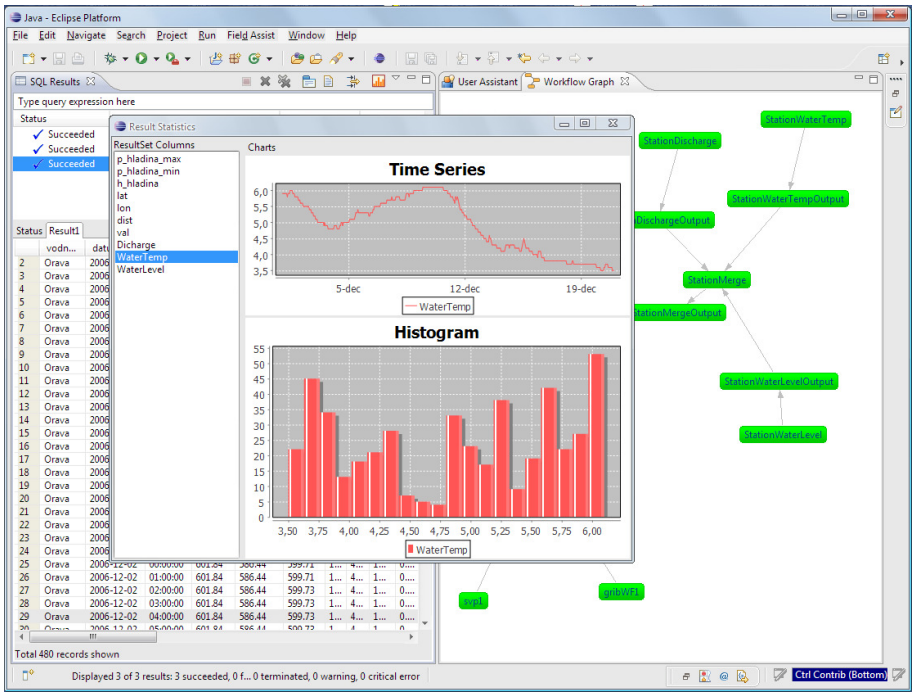

FIG. 3.2. *GUI showing results of DIEED—APEs workflow and its results*

of data from distributed, heterogeneous resources. Our results allow users to construct reusable application processing elements specified in DMIL [7] (language for data mining and integration, which is being designed within the project) and the engine executes them transparently on distributed data resources.

REFERENCES

[1] ADMIRE. EU FP7 ICT project: Advanced Data Mining and Integration Research for Europe (ADMIRE), 2008-2011. Grant agreement no. 215024.
[2] CROSSGRID. EU FP5 IST RTD project: Development of Grid Environment for Interactive Applications (2002-05) IST-2001-32243.
[3] K-Wf Grid. EU FP6 RTD IST project: Knowledge-based Workflow System for Grid Applications (2004-2007) FP6-511385, call IST-2002-2.3.2.8.
[4] MEDIGRID. EU FP6 RTD Sust.Dev. project: Mediterranean Grid of Multi-Risk Data and Models (2004-2006) GOCE-CT-2003-004044, call FP6-2003-Global-2.
[5] ANFAS. EU FP5 IST RTD project: datA fusioN for Flood Analysis and decision Support (2000-03) IST-1999-11676.
[6] Finite Element Surface Water Modeling System (FESWMS).
[7] Atkinson M. et al.: ADMIRE White Paper: Motivation, Strategy, Overview and Impact, v0.9. 2009.
[8] Antonioletti M., Atkinson M. P., Baxter R., Borley A., Chue Hong N. P., Collins B., Hardman N., Hume A., Knox A., Jackson M., Krause A., Laws S., Magowan J., Paton N. W., Pearson D., Sugden T., Watson P., Westhead M.: The Design and Implementation of Grid Database Services in OGSA-DAI. Concurrency and Computation: Practice and Experience, Volume 17, Issue 2-4, Pages 357-376, Feb. 2005.
[9] Karasavvas K., Antonioletti M., Atkinson M. P., Chue Hong N. P., Sugden T., Hume A. C., Jackson M., Krause A., Palansuriya C.: Introduction to OGSA-DAI Services. Lecture Notes in Computer Science, Volume 3458, Pages 1-12, May 2005, ISBN 978-3-540-25810-0.
[10] Ciglan, M., Habala, O., Tran, V., Hluchy, L., Kremler, M., Gera, M.: Application of ADMIRE Data Mining and Integration Technologies in Environmental Scenarios. To appear in Proc. of the 4th Grid Applications and Middleware Workshop (GAMW2009), Wroclaw, Poland, September 13-16, 2009, LNCS 6067-6068, Springer.

Edited by: Ewa Deelman
Received: March 30, 2010
Accepted: May 25, 2010