A Model to Capture Interaction between Data Provenance and Workflow Provenance

Zeeshan Ahmed[1], Syed Saif ur Rahman[2]

[1,2]*Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST) Karachi Pakistan*

[1]zeeshan.memonn@gmail.com

[2]saif.rahman@szabist.edu.pk

Abstract: Provenance means origin. There are two types of Provenance namely Data Provenance and Workflow Provenance. Data Provenance refers to the process of recording and tracking the source or origin of data while workflow provenance means history of workflows and their data that were used while performing an operation to get required result. The results which are obtained while performing operations are called datasets. On these datasets, the queries are performed to get required result. While exploring provenance topic in data and workflow, the author have tried to explore how a model can be presented and discussed which can capture interaction between data provenance and workflow provenance. In this paper, the author have discussed further about provenance, its types and its importance and also discussed how to execute query on provenance data. The main focus of this study was on the model which captures interaction between workflow and data provenance including the above mentioned topics.

***Keywords*-----Data Provenance, Workflow Provenance, Interaction between Data and Workflow Provenance, Scientific Workflow Provenance.**

I. INTRODUCTION

Provenance means source. Provenance is important when manipulated object is available and it is required to find the source and original shape of object. For example, a painting is converted into digital format through digitizing software; the provenance of that digital picture would be the originally painted picture. There are two types of provenance namely Data Provenance and Workflow Provenance. In data ware house, where summarized data is available and the source of actual data from which the data have been obtained is required, that original data is the provenance of summarized data and this term is called "Data Provenance".. The provenances in workflow are the steps or logs of all the operations that are performed in a workflow during its execution, its input data and output data. Today, workflow provenance is used widely and for that many tools have been invented like Kepler, MyGrid, and etc. In this paper, a model which captures interaction between workflow provenance and data provenance is discussed. This paper is divided into ten sections. Importance of provenance in section II will be discussed, applications of provenance in section III, management of scientific provenance data in section IV, provenance techniques in section V, model of Interaction in section VI, experimental results in section VII, related work in section VIII, future work in section IX and conclusion in section X.

II. IMPORTANCE OF PROVENANCE

The provenance is becoming important as it is growing and acknowledged extensively.. Today provenance plays a vital role in observation, analysis and in achieving required result. Provenance is used in every field of study. For example it is used in Science, Economics, Health, Defence, Weather, Astronomy and etc. Each aspect will be discussed in detail.

A. Provenance in Science

In science, provenance is used in observing and analysing different experiments like Big-Bang. In this regard, scientists' wanted to know how the earth has created. They observe every single data while performing the experiment. In case, if they want to enhance or add any new thing in their experiment, they must take all safety measure into consideration so that the experiment will not be affected by change and the provenance will accurate data to ensure calculations will be correct. If data which is under observation or analysis is not correct, then its outcome will not be trustable definitely.

B. Provenance in Economics

Provenance also plays an important role in economics as well. Economist share their views and analysis on the basis of the provenance. If data is not correct then decision made by authorities will not get the required result. A routine life example of an investor can be taken who wants to make investment in stock market, it is essential to know which company is profitable to avoid ending up in losses

To get the desired profitability, it is important that previous data available is correct.

C. Provenance in Health

Provenance plays a vital role in health related researches also. With the help of provenance, scientists are able to find a cure of numerous diseases. For example, cure of disease like Tuberculosis (T.B) was discovered after having studies of different cases in which patient is affected by T.B. After inventing the cure, it was applied on effected people for testing purpose to find out that it help people in recovering or not and then it was introduced to the world.

D. Provenance in Weather

With the help of provenance in environment, weather specialists announce what type of weather will be in coming days, week, etc. The weather specialists are able to predict and announce the weather conditions for the desired cities and regions.

E. Provenance in Astronomy

Provenance also plays a significant part in astronomy. With the help of provenance, scientists identify new galaxies, their movement pattern and effects.

III. APPLICATIONS OF PROVENANCE

Provenance system can be used for different purposes. According to Globe [1] some of them are Data Quality, information, Audit Trail, Attribution and Replication Recipes.

A. Data Quality

Data quality can be estimated with the help of provenance. Data reliability can also be measured with the help of provenance and it is based on the source of data and its transformation. On data derivations, provenance can also support in providing proof statements about the reliability and quality of the data [2]. Data provenance has a lot of importance. the popular example of UAL (United Airlines) will be discussed in this connection. In 2008, a back dated news article become very trendy on Google News about the bankruptcy of UAL organization (which actually happened in 2002) and due to this news, panic pick up its toll among investors and before it was realized and confirmed that the article is out dated, UAL shares dropped by 75 percent in market [3].

B. Informational

Provenance is generally used for the purpose of querying which is based on provenance metadata. Query is used to discover new data. A context for the interpretation of data can also be provided.

C. Audit Trial

Data audit trail can be traced with the help of provenance. Provenance information can be used by Audit Trail to determine about the resources and their uses and it can also detect errors in generation of data [2, 4, 5].

D. Attribution

Data ownership and copyright can be done with the help of provenance which helps in recording citation. The citation describes accountability if data is erroneous.

E. Replication Recipes

Derivation of data can be repeated if one has detailed information of provenance, which can help in maintaining the currency and it can be recipe for the replication.

IV. MANAGEMENT OF SCIENTIFIC PROVENANCE DATA

Process of scientific data orientation depends on fast and accurate analysis of data generated during experiments. The data increases simultaneously with experiments and it becomes troublesome to save and manage such huge data on available databases. For this, special types of databases are required which can fulfil the needs of scientists related to experiments. Scientists' demands more resources to store data as they present new theories to world and made discoveries by using and reusing that data. However, to fulfil the need, resources like good databases are required to fulfil the necessities of scientists.

According to Jim Gray (Microsoft researcher and ACM Turing award laureate)

"A fourth data intensive science is emerging. The goal is to have a world in which all the science literature is online and they interoperate with each other." [6]

In spite of the fact that scientific data has a variety, it still shares some aspects.

- Scalability of transactional datasets
- Data generated with the help of workflows
- Are multidimensional
- Physical models are also embedded
- Hold important metadata related to experiments and their provenance

It also includes common scientific data requirements like data automation and processing of metadata, parallel or online processing of data and efficient manipulation of data stored in files. As commercial software's' are unable to fulfil scientists' requirements, scientists' started to move towards application specific solutions for the fulfilment of their needs though some of them are built on the top of commercial databases. Further, the software which is built on top of commercial software is tightly coupled to that specific application and is not able to adapt changes. Scientific database demands high performance and data quality in every domain of science, scientific communities and big laboratories. Origination of scientific data is performed through simulation or observation. In observational data, detectors are used to collect data where input is digitized and output is raw observational. Simulators are used to produce data with the help of simulation parameters. In scientific research, both types of data are important. According to Anastasia *et al.* **"At Ohio state university and Johns Hopkins University, an astronomy project is running from a long time which calculates the stars and galaxies on sky. Currently, with the help of telescope they collect data of 25 percent of sky visible from earth. The data is saved on SQL Server databases which is periodically**

saved on delivered to tapes for archiving purpose. These tapes are then mailed to processing centre located at Fermalib in Batavia, IL, to be processed through automated software pipelines to identify astronomic objects. Many astronomical software's process the data in parallel and the output of that processing and the image generated is then stored in local archives at Fermalib. Metadata that is generated from it is then stored in SDSS database for analysis and observational purpose of SDSS experts. The schema of SDSS is on 70 tables but few are used mostly out of them while rest are dependent upon request." [7]

Similarly the same team has mentioned that data regarding earth quake is analysed by scientist. With the help of this data, scientists' try to find out the movement of the ground during earth quake. An octree based hexahedral mesh is modelled for ground motions which produce a mesh generator by taking soil density as input. A "solver" tool is used to propagate the waves through the earth. At each interval, the solver tool computes velocity of each node at spatial direction and writes result of it on disk. The result of this activity is a 4D image which describes the ground velocity. Different types of analysis can be carried out on this dataset for example time specific or space specific.

V. PROVENANCE TECHNIQUES

There are different tools for provenance. With the help of these tools, researchers observe the data, query that data, analyse results retrieved against these queries and come to the conclusion or try to reach at the conclusion. Today, there are many tools available in market but few important of them are Kepler, Chimera, myGrid, ESSW, CMCS, Trio, etc.

A. Kepler

Kepler is based on actor oriented approach and provides a user friendly IDE (Integrated Development Environment) to create scientific workflows. Kepler is designed to help scientists in creation and execution of workflows, and in the result the output that they receive is used for observation, analysis, querying and for other perspectives. Kepler also helps scientist in re-generating the workflow easily due to its user friendly environment. User just have to drag and drop the actors, it has to set the inputs and outputs and on the execution of that workflow result is generated which can be saved on different storage devices in different formates. For example, user can save its data in database, file, etc.

B. Chimera

Chimera is used for managing derivations of data objects and their analysis in accumulated environment. In chimera, provenance is collected in the form of derived data. Model of chimera is process oriented which help in recording of provenance data. In this tool, workflows are developed in VDL (Virtual Data Language) and are in the form of graphs known as "Derivation Graphs". Schema followed by VDL represents data as abstract datasets. Datasets can be in any format, for example it can be files, tables or objects in spite of the fact that prototype of chimera only support files. In chimera, computing process is written in files currently, but in future web services will also be used along with files. It is important to know that computing process of chimera is also called "Transformation". Parameterized instance of this transformation process is known as "derivations". On the execution of these workflows, objects are automatically created through derivation. Invocation objects links input and output of dataset and as a result provenance is received and analysed. VDL is responsible for the mapping results available through the help of query in relational database which is accessed with the help of Virtual Data Catalogue Service or VDC service. Storage of data (metadata) depends on user choice that whether user wants to store it in single repository or multiple repositories. In case if metadata is planned to be stored in multiple repositories, then an inter-catalogue will be maintained for the purpose of mapping. Storing data in multiple repositories provides scalability to user. In this case, user has the edge that whenever he needs more storage space, he can add a new repository. To access the VDL queries, a virtual query browser is already proposed [8]. A new idea for provenance is planning and cost estimation of regeneration of dataset(s). A dataset that is previously created and is required to be re-created, the workflow planner can take the help of provided provenance in selection of the best plan for allocation of resources.

C. MyGrid

MyGrid is based on service-oriented methodology. Workflows that are executed by MyGrid are written in Xscufl (XML based Simple Conceptual Unified Flow Language) language which uses Taverna engine. A log is maintained for the actions of workflows. The log contain details of how many services were invoked during execution of these workflows, what were their parameters, their starting and ending time for the purpose of TAT, what was the input data and what data was received as its output. This log is referred as provenance log. Provenance log is used to collect the provenance for intermediate and final data products. A data service called MyGrid Information Repository (MIR) is used to store metadata information in relational database related to experiments. The MIR is central repository service. There are numbers of ways to use provenance for finding new information or knowledge. For example, with the help of Haystack Semantic web browser, lineage information is available as RDF (Resource Description Framework) and that can be viewed as graphs (labelled graphs). Another tool is COHSE (Conceptual open Hypermedia Services Environment) which is used to build provenance on the basis of semantic web.

D. CMCS

CMCS (Collaborator for the multi-scale Chemical Sciences) is used to manage heterogeneous flow of data and metadata across multiple disciplines of sciences. Disciplines like combustion research which is supplemented to establish pedigree of data with the help of provenance metadata. CMCS store URL references files. The repository used by CMCS is SAM. SAM is the abbreviation of Scientific Annotation Middleware. CMCS project is a tool based on informatics and is used for the data management of metadata available for multi-science. With files that are available in SAM and XML (Extensible Mark-up Language), metadata properties are used which manage them through an interface called Distributed Authority and Versioning (WebDAV). Files available in SAM (Scientific Annotation Middleware) can be of any format. Files vary from granularity level to resources like web services, process, data objects or bibliographic records. DC (Dublin Core) verbs are used as properties of XML for processing of data files. In SAM (Scientific Annotation Middleware), it further relates them semantically via XLink (XML Linking Language) references. Additional contextual information is available from either metadata defined by user or DC (Dublin Core) elements. Example of Dublin Core elements are Title and Creator. For the purpose of mapping with standard DC (Dublin Core) metadata terms, there are number of heterogeneous metadata schemas available. Mapping is done with the help of XSLT (Extensible Style sheet Language Transformations) translators. In CMCS, for the purpose of provenance there is no automated facility available during execution of a workflow. DAV (Distributed Authority and Versioning) aware applications are used to populate metadata and data files. Another way is manual feeding of data which can be done with the help of portal. SAM (Scientific Annotation Middleware) is used for the purpose of query from provenance metadata with the help of WebDAV (Distributed Authority and Versioning) client. For a particular resource on portals, provenance metadata can be accessed by users from a webpage. Provenance information to RDF (Resource Description Framework) can also be exported so that semantic agents use the information for understanding purpose of the relationship between these resources. Metadata of provenance which points out that data is modified can generate notification which in result start the execution of workflow so that data should be updated related to data products which are dependent on it.

E. ESSW

ESSW (Earth Science System Workbench) is used for managing and storing (Meta) data related to earth for the researchers. Provenance is an important feature of metadata (data about data) that is created in workbench which is used to find error in data products which are derived and is also responsible for the datasets quality. A scripting model is used by ESSW for data processing. Entire data handling or manipulation is done with the help of scripts that wrap existing scientific applications [9]. A master workflow script is written to call all the scripts in a sequential manner and that forms a DAG. Scripts produce and consume data products. Every script and data product is uniquely labelled with a metadata object. As the master workflow script starts running, it calls other scripts which compose metadata (data about data) for themselves and also generate products of data or data products. The master script links the flow of data from one script to another in defined pattern. The linkage is done with the help of metadata ids so the provenance tracking for all data shall be noted. Balancing the ESSW between process and data oriented lineage is done with the binding of data and script in parent-child concept. Script in ESSW is responsible for recording metadata and provenance with the help of libraries and templates that were provided to it. Metadata objects are stored as files on a location that is easily accessible and the provenance is stored separately in a database having ability to keep relations.

This database is also known as relational database libraries are responsible for these linkages [2].

It is important to know that scalability is still not addressed in spite of the fact that it is proposed to combine the provenance information in different organizations.

The provenance information and its metadata can be navigated as a workflow DAG with the help of PHP scripts in a web browser; doing so will make the user able to access provenance information from database [10].

F. Trio

Trio project is implemented for data warehouses. The main purpose was to track the provenance information regarding data in data warehouses. It was started by Cui and Widom [11, 12]. The primary purpose was to track provenance information in data warehouse but any e-science systems which uses database objects like functions and queries for the modelling of workflows and transformation of data can also use these techniques. A query tree can be displayed from database view with bottom-up evaluation. It starts with leaf operators. Input of these leaf operators are tables and parent operators, following, the result of child operators taken as input [13]. It is possible for ASPG (“A” for Aggregate, “S” for select, “P” for project, “G” for join operator) type views to create a query which behaves like an inverse query for the query which operates materialized view. To identify the source recording of queries, inverse queries are written at the lower level of a row in a view. The information that is obtained is stored in a table know as lineage or provenance table. As lineage with rows are directly associated; therefore, it makes the provenance a data-oriented scheme of provenance. The mechanisms of handling rows that are not part of view are created by either insert or update queries.

In Trio, Provenance is quite simple as the rows from the source and query of the view create a row in the view which has no meaningful metadata (data about data) recorded. Scalability (expenditure of structure) is still not addressed in detail. Trio Query Language (TriQL) is used for retrieving lineage information by some special purpose constructs along with querying the lineage table.

Table 1. Comparison of Different Techniques of Provenance

	Chimera	myGrid	CMCS	ESSW	Trio
Applied Domain	Physics, Astronomy	Biology	Chemical Sciences	Earth Sciences	None
Workflow Type	Script Based	Service Oriented	Service Oriented	Script Based	Database Query
Use of Provenance	Informational; Audit; Data Replication	Context Information; Re-enactment	Informational; update data	Informational	Informational; update propagation
Subject	Process	Process	Data	Both	Data
Granularity	Abstract datasets (Presently files)	Abstract resources having LSID	Files	Files	Tuples in Database
Representation Scheme	Virtual Data Language Annotation	XML / RDF Annotation	DublinCore XML Annotation	XML / RDF Annotation	Query Invension
Semantic Info	No	Yes	Yes	Proposed	No
Storage Repository / Backend	Virtual Data Catalog / Relational DB	mIR repository / Relational DB	SAM Over DAV / Relational DB	Lineage Server / Relational DB	Relational DB
User Overhead	User defines derivations: Automated WF trace	User defines Service semantics: Automated WF Trace	Manual: Apps use DAV APIs; Users use portal	Use Libraries to generate provenance	Inverse queries automatically generated
Scalability Addressed	Yes	No	No	Proposed	No
Dissemination	Queries	Semantic browser: Lineage graph	Browser: Queries: GXL / RDF	Browser	SQL/TriQL Queries

G. Orchestra

Orchestra is a system which is used for data sharing between peers. The peers can be heterogeneous (mixed), for example, it can be used to share data between Windows, Linux, MAC and etc.

All these peers are connected with a network and they are mapped with a schema. Orchestra system is also known as Collaborative Data Sharing System or CDSS. Every peer has a local database which is controlled and edited by the local computer itself. For updates or new packages related to the required data, it contact other peers available in that network and if it founds new information, it updates the information in its local database. Torrent Downloader is a simple example of this system. Updates of each peer are translated and circulated within the mappings to the other peers in order to appraise the new information on a peer. Trust conditions are used as a filter for the exchange of the updates. The trust conditions translate which peer has the authority of particular data. A peer can accept or reject the update on this basis. To provide support to these filters, each update carries provenance information which is used to decide whether the update is required to be added in the information on local peer or not. The main purpose of Orchestra is sharing scientific data, but it is not limited to scientific data only. Other applications can also use the system for their required purposes if their prerequisite and characteristics are similar to scientific data. Figure 1 defines an example related to bio-informatics CDSS (Collaborative Data Sharing System) which is based on real world application. Both the application and databases are associated with the Penn Center for Bio-informatics.

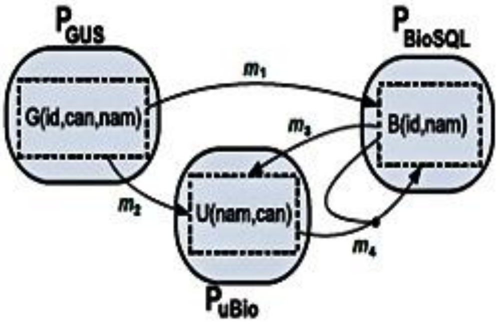
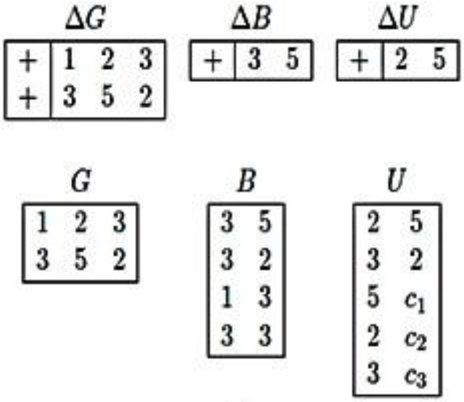


Fig. (1). Example of collaborative data sharing system for three bioinformatics sources [10]

Gene expression, organism and protein information are enclosed in Genomics Unified Schema (GUS). Another project named BioSQL, which is affiliated with BioPerl project contains very similar type of concepts. Another schema called uBio is used to create names and synonyms for taxa. Information related to taxon is stored in these cases of databases which are maintained independently but with mutual interest for the others. Figure 1 elaborates the above example.

According to above example, the uBio (PuBio) peer wants to import data from a peer GUS (PGUS). The transaction or movement of data is shown by an arc named m2 in Fig 1. Similarly, BioSQL peer PBioSQL also wants to import data from another peer PGUS. This Transaction is displayed by the arc named m1. Here, it can be seen that there are two transactions between UBio peer and BioSQL peer with the arcs m3 and m4. Both are providing input and output from each other. Arc m3 is illustrating that PuBio wants to import data from PBioSQL while arc m4 illustrats that PBioSQL wants to import data from PuBio. It is important that each peer has a certain trust policy. Depending on that policy, it decides whether to incorporate the data changes or not. For example, according to above Figure 1, one of the policies is that PBioSQL will import data from PGUS while PGUS will not import data from any peer. This Collaborative Data Sharing System (CDSS) allow flow of data between these systems. This system is managed with the help of policies defined for each peer and their mappings. These policies are defined by the administrators of these peers. Arcs that are defined between peers that are sets of TGDS (Tuple-generating dependencies), TGDS is one of the popular way of defining mappings and constraints [14, 15] while sharing the data. TGDS can also be termed as GLAV (Global Local as View). These types of changes or updates need new placeholder values which are known as “Labelled Nulls”. Updates exchanges that are defined in Figure 1 are further elaborated in Figures 2 and 3. For example, if it is assumed that local updates on a peer are shown on the top as plus (+) sign then the translation of these updates construct instances as are shown on the bottom (where c1, c2, c3 are labelled nulls).



ΔG

+	1	2	3
+	3	5	2

ΔB

+	3	5

ΔU

+	2	5

G

1	2	3
3	5	2

B

3	5
3	2
1	3
3	3

U

2	5
3	2
5	c_1
2	c_2
3	c_3

Fig. (2). Update Exchange and Resulting Provenance Graph Example Part 1

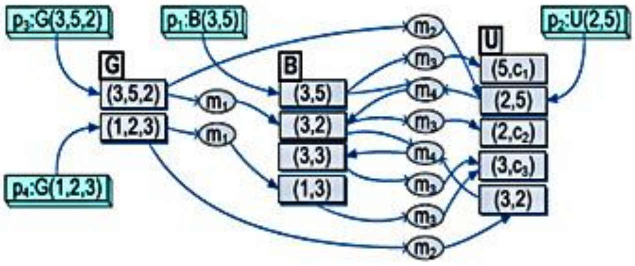


Fig. (3). Update Exchange and Resulting Provenance Graph Example Part 2

VI. MODEL OF INTERACTION

The above literature presents that both types of provenance: Data Provenance and Workflow Provenance work independently.

In this research, the author has certainly strained to figure out a possible way for the communication of two provenances. The problem is to work out the possibility of above condition. Is it possible? If yes, how will it behave? What will be the structure? What will be advantages and disadvantages?

To get the answers of these questions, a conceptual model has been created in which both the data provenance and workflow provenance attempts to communicate with each other. The idea has been illustrated in Figure 5. In the left box of figure 4, there are database tables while on right side there are a set or sequence of workflows which execute one after another.

While executing a workflow from a group of workflows, the workflow save its data in database tables and also stores an indicator or key to identify which workflow has been executed for the perusal of next workflow. It also keeps track of what were the inputs of that workflow and what are the outputs of the workflow. Storing the inputs and outputs of the workflow in database will help other researchers in re-iterating the same operation later to confirm whether the experiment was successful or not.

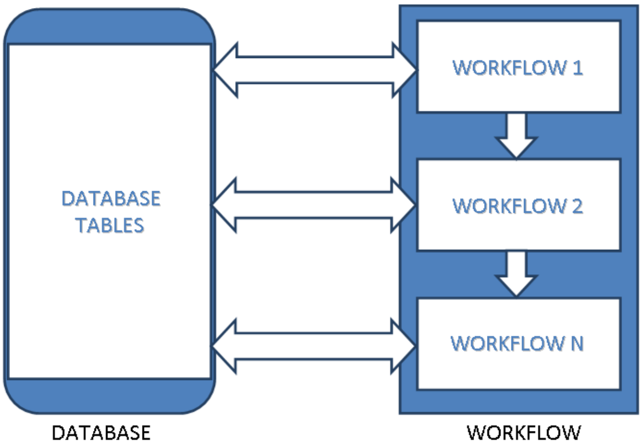


Fig. (4). Conceptual Model of Interaction

This will help researchers in various means. This system will save the time of researchers in performing the experiment. Previously, the researcher has to write down the entire process of workflows from workflow 1 to their desired workflow, and they can achieve only by executing all the previous workflows.

On the contrary, this system will save time, efforts; money of the researchers as the outputs of the last executed workflow is already saved in database

VII. EXPERIMENTAL RESULTS

All experiments required in this research were carried out in Kepler workflow. A simple workflow has been presented in Figure 5 to get reader familiar with the execution of workflow. The above workflow has few elements: SDF Director, a constant and a display element to show the result as output. When the workflow was executed, it displayed the output “Hello World”. With the above example, the author was introduced with the workflow execution in Kepler. In the next workflow, the data was read from file. It is demonstrated to illustrate that input can be provided to workflow in any format (like text files, from database tables, etc.). As it can be seen in figure 6, there is a SDF Director again, a File Reader and a Display control. File reader control is used to read data from file and perform the required operation in case if it is mentioned in workflow. In the this case, the data file has been forwarded to display control which displayed what data is in file. Figure 7 shows the popup window which opens to get input of file destination.

This popup window opens when file reader is double clicked in Kepler workflow window.

XML Data Transformation has been discussed in the next example (figure 8). In this example, data has been taken or read from an XML file and is displayed in three different formats. In first row of figure 8, same result is displayed as is in file. Second is Sequence display of it and third is HTML display of the XML file. In figure 8, it has been experienced how data can be transformed into three different formats and can be displayed.

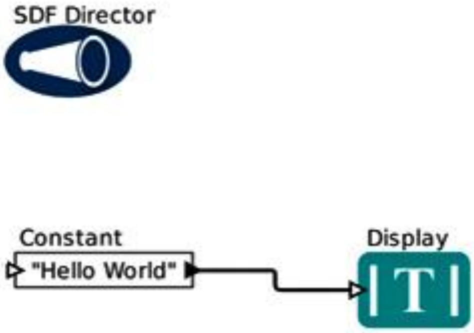


Fig. (5). Hello World Sample Workflow

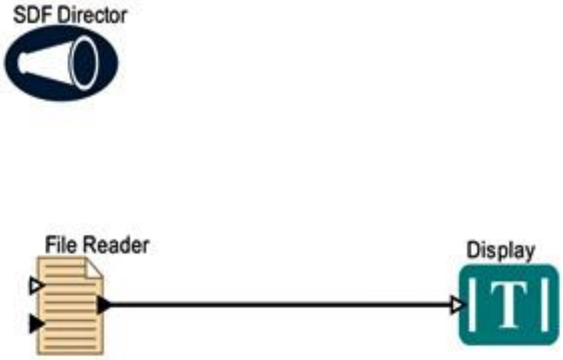
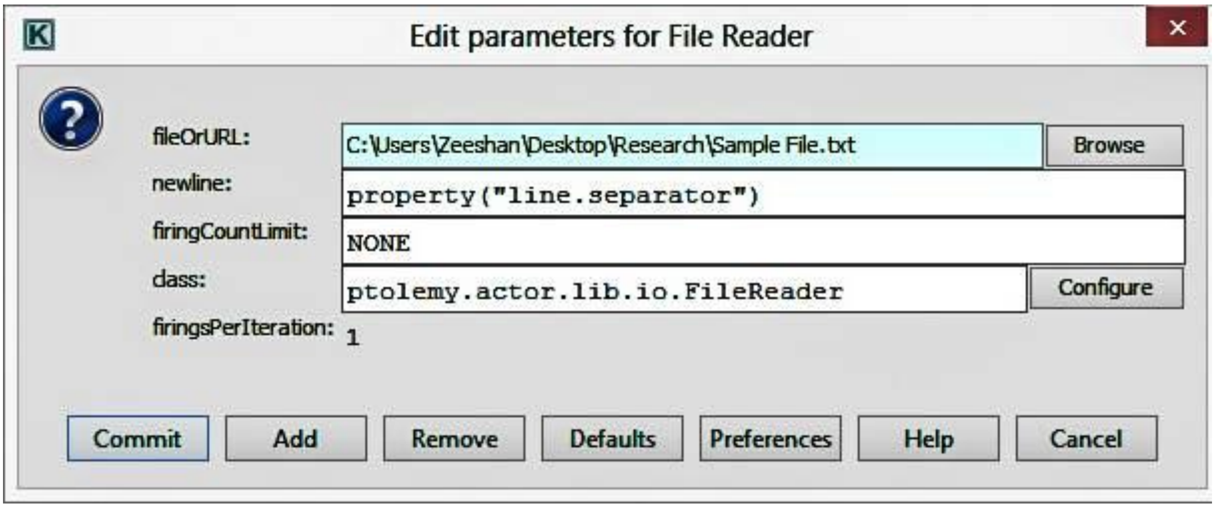


Fig. (6). File Reading Example

Next Example is related to research carried out in this paper and figure 9 shows how data can be fetched from database tables and displays it or manipulate it. In this workflow, information has been fetched from a table stored in MySQL database. The workflow drew the record, convert these records to XML, display the information and close the connection. From this example, the author have tried to show that how can the information be drawn from database and it can be displayed in records form and in XML format so that next workflow in a set of workflows can use the information and get its desired output.



Edit parameters for File Reader

fileOrURL: C:\Users\Zeeshan\Desktop\Research\Sample File.txt

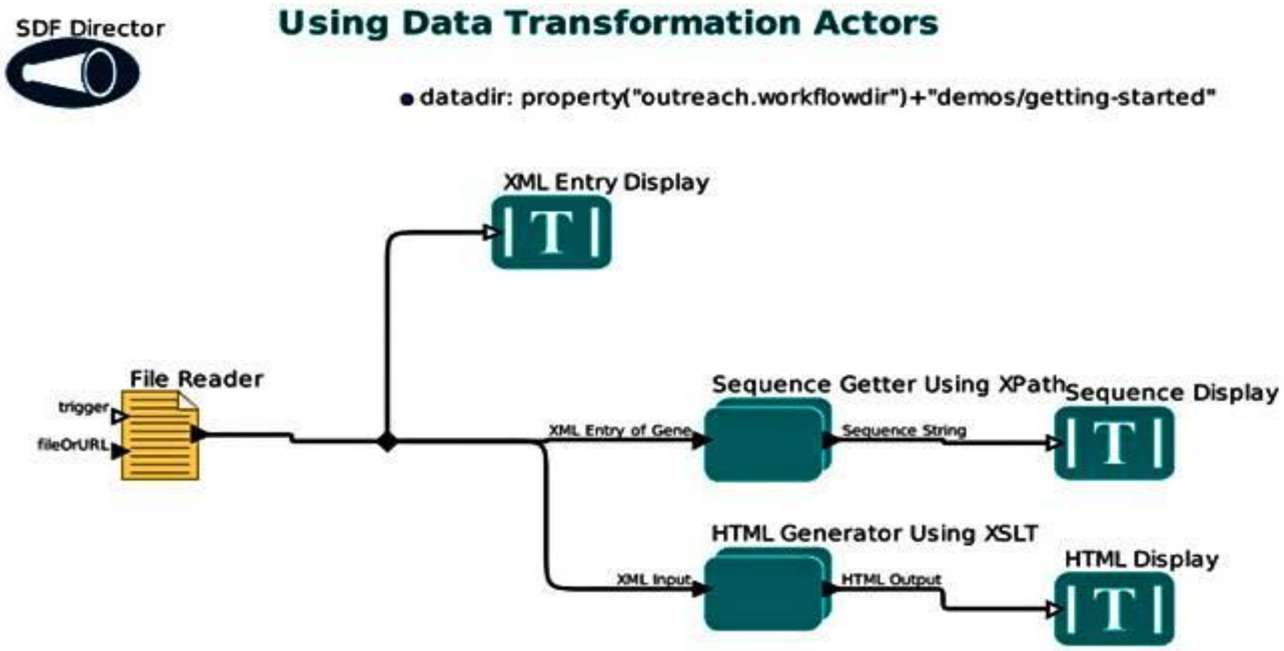
newline: property("line.separator")

firingCountLimit: NONE

class: ptolemy.actor.lib.io.FileReader

firingsPerIteration: 1

Fig. (7). File Loading Popup Window



SDF Director
Using Data Transformation Actors
datadir: property("outreach.workflowdir")+"demos/getting-started"
XML Entry Display
File Reader
trigger
fileOrURL
Sequence Getter Using XPath
Sequence Display
XML Entry of Gene
Sequence String
HTML Generator Using XSLT
HTML Display
XML Input
HTML Output

Fig. (8). XML Data Transformation

Figure 10 is displaying the output of the above workflow on its execution. Figure 11 shows yet another simple workflow through which a provenance can be recorded.

In this workflow, user defines a parameter. The parameter passes through a condition which checks whether the condition is true or false and on its behalf, concerned message is displayed. Information can also be tracked by recording the data. Tracking the interaction between data and workflow process can be carried out by storing the workflow image in database and also by writing the instructions in a manual or in database table with the inputs and outputs of that data. Output of the workflow illustrated in Figure 11 is displayed in Figure 12. Sub-workflows are also important to describe as this research is dependent on them. Sub workflows are workflows inside workflows. As presented in figure 4, the right box is the workflow provenance which is a series of workflows interlinked in sequential manner. Sub workflows helped in implementing and elaborating this research work. An example of sub-workflow is illustrated in Figure 13. There is a “User Customize Display Control” actor, this control act as a sub workflow. The control has simple actions for illustration purpose which received XML as input and displays it and has an output control for return purpose. The actors inside this workflow are shown in Figure 14.

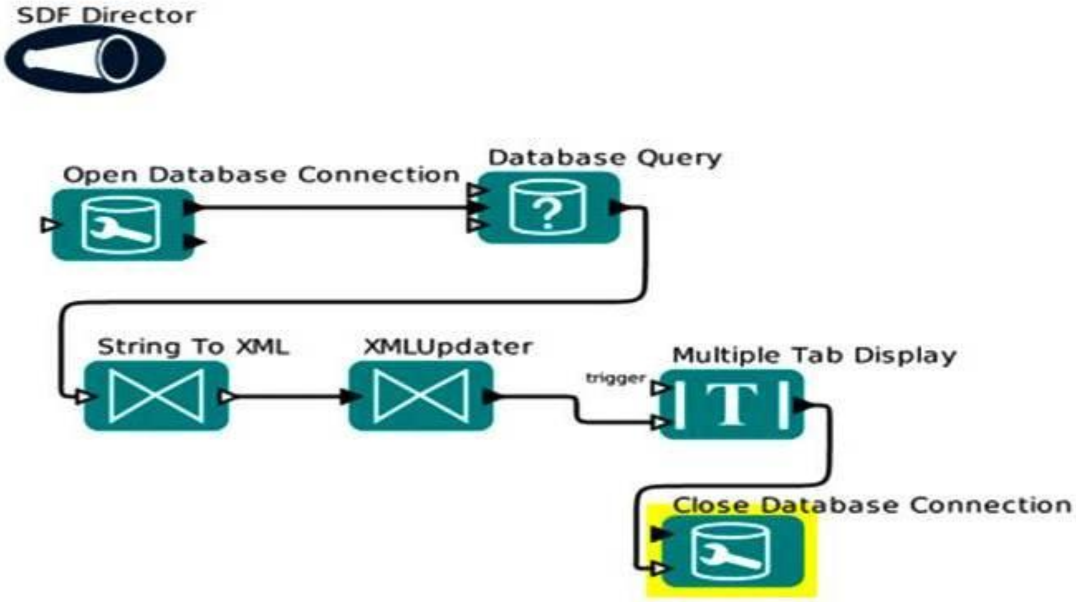
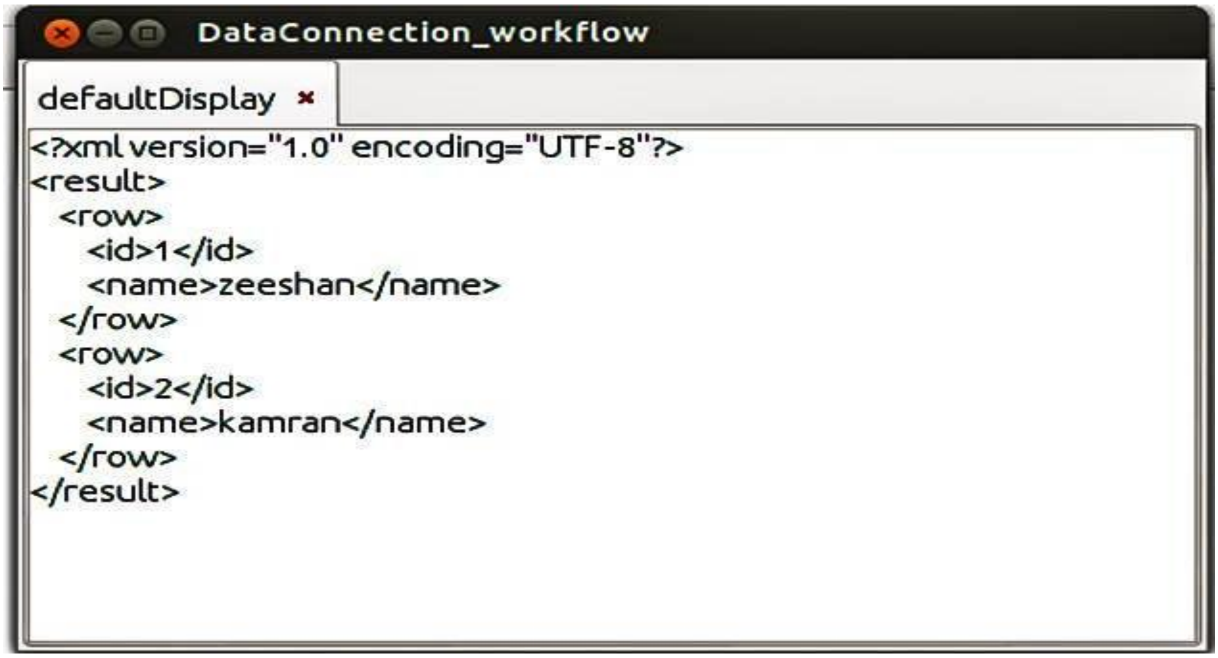


Fig. (9). Query Database Table



```
<?xml version="1.0" encoding="UTF-8"?>
<result>
  <row>
    <id>1</id>
    <name>zeeshan</name>
  </row>
  <row>
    <id>2</id>
    <name>kamran</name>
  </row>
</result>
```

Fig. (10). Output of Database Records to XML

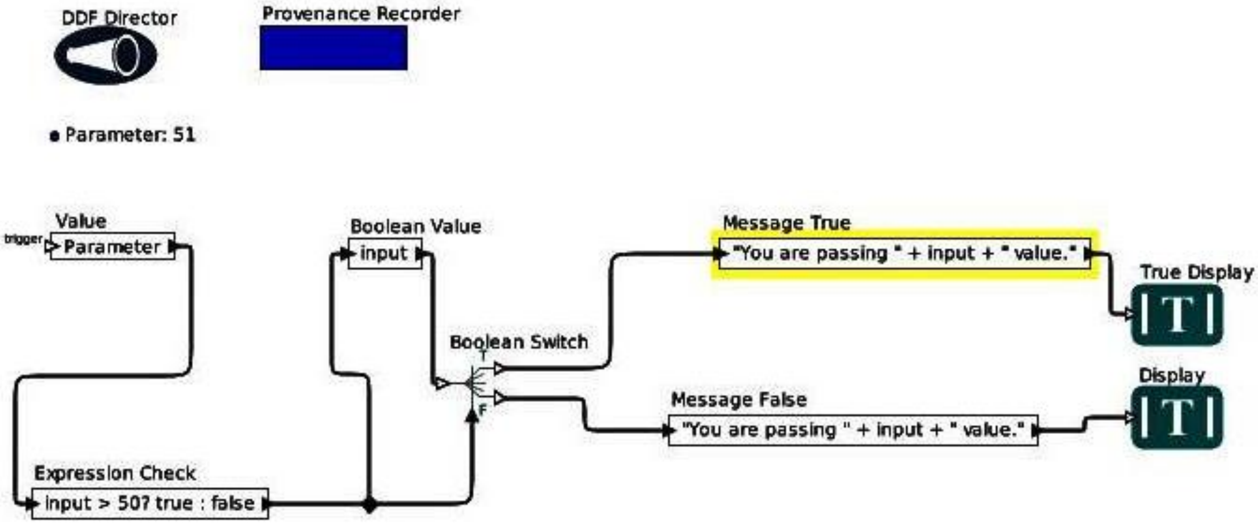
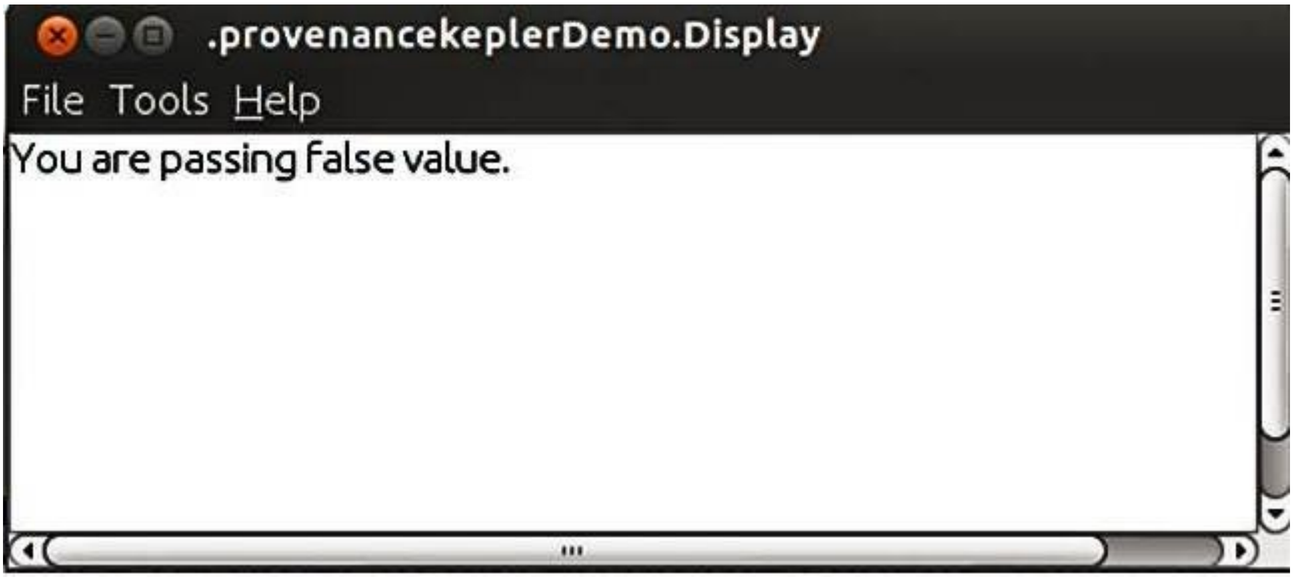


Fig. (11). Ternary Operation Example



```
You are passing false value.
```

Fig. (12). Output of Ternary Example

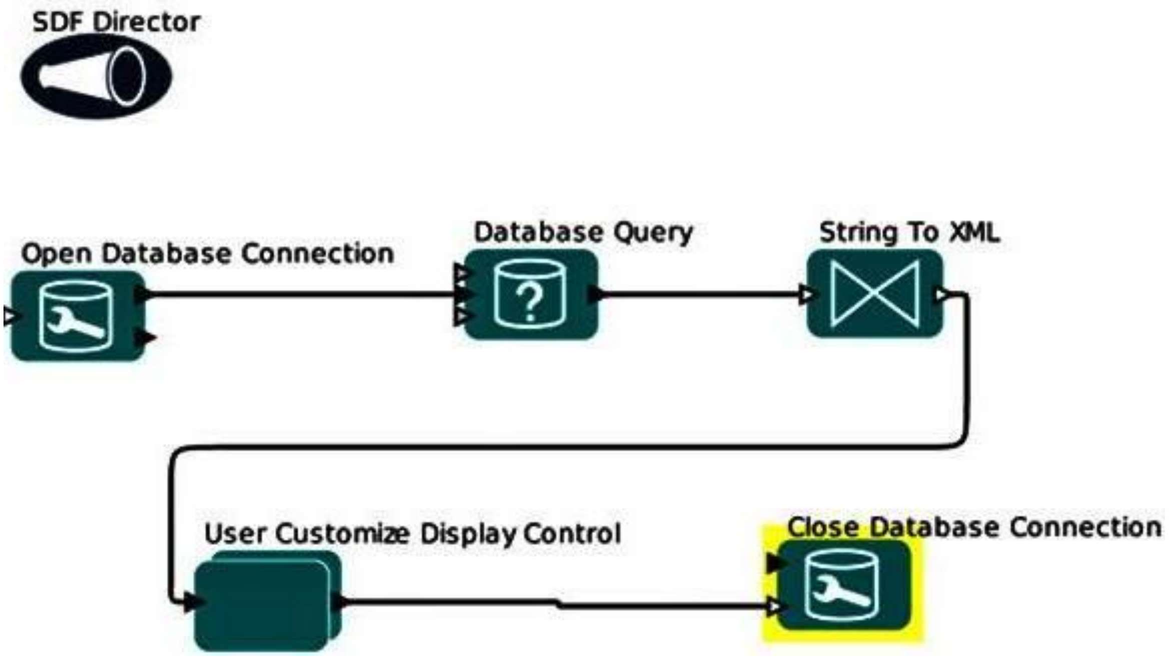


Fig. (13). Sub Workflow Example

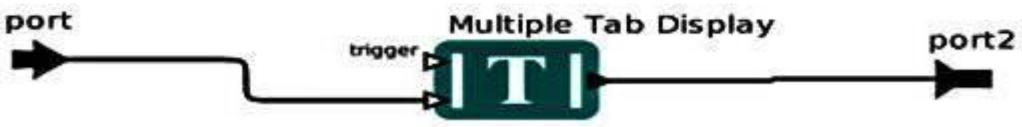


Fig. (14). Inside controls of sub workflow (User Customize Display Control)

VIII. RELATED WORK

Researchers have written a number of papers on database provenance and scientific workflow provenance; for example' "Provenance in Databases: Why How and Where by James Cheney and his team" [16], "Data provenance the foundation of data quality by peter Buneman and his team" [6, 9]. Some have described data provenance helpful in scientific researches like discovery of new stars in galaxy, environment behavioural changes, and etc. Some have described it helpful in other technologies like P2P (peer to peer) network.

Scientific workflow provenance has gained much attention [17]. Different types of organizations have worked on projects related to data provenance with their particular requirements. Researchers have queried the output to get their desired result. One common problem that was faced while querying was that as workflow becomes larger and complex, their querying cost also become higher [17]. In order to reduce the cost, researchers start focusing on other ideas and proposals were made which could help in reducing the cost and complexity of provenance graph. One of the simple solutions is creating views and displaying the provenance information according to the concerned view. Today, many workflows support the facility of view in the form of creating composite tasks. Kepler is one of the tools that support it. Others are myExperiment, myGrid and Taverna [17]. Files are also used as an alternate of views for data provenance purposes.

IX. FUTURE WORK

As this paper shares a conceptual idea of the interaction model, it's important to implement it in real world. Application should be designed which facilitate this model and help researchers more to focus their concentration on their actual work rather than implementing the workflow again to get at a particular point. Along with it, it is important to focus on data storage issues. Sizes of databases are required to be increased. Query related to provenance information should be optimized so that they can give quick output in provenance graphs.

X. CONCLUSION

Provenance is one of the hot topics in today's industry. Among Data and Workflow Provenance, workflow provenance has received more attention. Workflow provenance helps in keeping record of the executed workflows and their data. This paper includes all the necessary information related to provenance and. For example, it discussed tools relate to provenance, provenance applications, provenance importance, how it is helpful in science. More importantly, model suggested in this paper is also a proposal so that a physical application could be created based on this model which could help in querying data from workflow and also from databases.

REFERENCES

[1] C. Goble. "Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics". *In: Workshop on Data Derivation and Provenance*, Chicago. 2002.

[2] Y. L. Simmhan, B. Plale, and D. Gannon. "A survey of data provenance in e-science". SIGMOD Records, vol. 34, no. 3, pp: 31-36, September 2005.

[3] P. Buneman and S. B. Davidson. "Data provenance – the foundation of data quality". September 2010.

[4] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C-A. Saita. "Improving data cleaning quality using a data lineage facility". In *Proceedings of the 3rd International Workshop on Design and Management of Data Warehouses, DMDW'2001*, Switzerland, 2011, vol39 CEUR pp: 3.

[5] M. Greenwood, C. Goble, R. Stevens, J. Zhao, M. Addis, D. Marvin, L. Moreau, and T. Oinn. "Provenance of e-Science experiments experience from bioinformatics". In *Proceedings of the UK OST e-Science second All Hands Meeting*, 2003, pp: 223-226.

[6] Ailamaki, V. Kantere, and D. Dash. "Managing scientific data". *Communications of ACM*, vol. 53, no. 6, pp: 68–78, June 2010.

[7] P. Buneman, A. Chapman and J. Cheney. "Provenance Management in curated databases". In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, (SIGMOD '06), 2006, pp: 539–550.

[8] T. Foster, J.S. Vockler, M. Wilde and Y. Zhao. "Chimera: A virtual data system for representing, querying, and automating data derivation". In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management (SSDBM)*, 2002, pp: 37–46.

[9] R. Bose and J. Frew. "Composing lineage metadata with XML for custom satellite derived data product". *In Proceedings of 16th International Conference on Scientific and Statistical Database Management (SSDBM)*, 2004, pp: 275-284.

[10] Frew and R. Bose. "Earth system science workbench: A data management infrastructure for earth science products". In Proceedings of 13th International Conference on Scientific and Statistical Database Management (SSDBM), 2001, pp: 180–189.

[11] Y. Cui and J. Widom. "Lineage tracing for general data warehouse Transformations". *The VLDB Journal*, vol. 12, no. 1, pp: 41–58, May 2003.

[12] Y. Cui and J. Widom. "Practical lineage tracing in data Warehouses". In *Proceedings of 16th International Conference on Data Engineering (ICDE)*, 2000.., pp: 367–378.

[13] N. Foster and G. Karvounarakis. "Provenance and data synchronization". *IEEE Data Engineering Bulletin*, 2007.

[14] Deutsch, L. Popa, and V. Tannen. "Query reformulation with constraints". *ACM SIGMOD Record*, vol. 35, no. 1, pp: 65-73, 2006.

[15] R. Fagin, P. G. Kolaitis, R. J. Miller, and L. Popa. "Data exchange: Semantics and query answering". In *Proceedings of the 9th International Conference on Database Theory (ICDT '03)*, 2003, pp: 207 -224.

[16] J. Cheney, L. Chiticariu, and W-C. Tan. "Provenance in databases: Why, how, and where". *Foundations and Trends® databases*, vol. 1, no. 4, pp: 379–474, April 2009.

[17] Z. Liu, S. B. Davidson, and Y. Chen. "Generating sound workflow views for correct provenance analysis". *ACM Transactions on Database Systems*, vol. 36, no. 1, March 2011.