Lineage Tracing in a Data Warehousing System

(Yingwei Cui and Jennifer Widom)

Computer Science Department, Stanford University
{cyw, widom}@db.stanford.edu

1. Introduction

A data warehousing system collects data from multiple distributed sources and stores the integrated information as materialized views in a local data warehouse. Users then perform data analysis and mining on the warehouse views. In many cases, the warehouse view contents alone are not sufficient for in-depth analysis. It is often useful to be able to "drill through" from interesting (or potentially erroneous) view data to the original source data that derived the view data. For a given view data item, identifying the exact set of base data items that produced the view data item is termed the view data lineage problem. Motivation for and applications of lineage tracing in a warehousing environment are provided in [2, 3]. In the context of the WHIPS data warehousing project at Stanford [4], we have developed a system that performs efficient and consistent lineage tracing.

Some commercial data warehousing systems support schema-level lineage tracing, or provide specialized drill-down and/or drill-through facilities for multi-dimensional warehouse views. Our lineage tracing system supports more fine-grained instance-level lineage tracing for arbitrarily complex relational views, including aggregation. At view definition time, our system automatically generates lineage tracing procedures and supporting auxiliary views. At lineage tracing time, the system applies the tracing procedures to the source tables and/or auxiliary views to obtain the lineage results and to illustrate the specific view data derivation process.

2. Lineage tracing system

Given a view data item $I$, the exact set of source data that produced $I$ is called $I$'s lineage. We use an example to illustrate the concepts; a full formalization of the problem and its solutions are given in [3]. Consider a financial data warehouse with the three source tables and sample data as shown in Figure 1. A view Promising is defined to contain all "promising" industries, where an industry is regarded as promising if some stock in that industry is gaining money over all purchases, and the stock has a price-earnings ratio below 40:

```
CREATE VIEW Promising AS
SELECT e.industry
FROM Purchases p, Daily d, Earnings e
WHERE p.ticker = d.ticker
    AND d.ticker = e.ticker
    AND d.closing/e.earnings < 40
GROUP BY p.ticker, e.industry
HAVING SUM(p.price * p.shares)/SUM(p.shares) < d.closing
```

Over our sample source data the view contains two tuples, (computer) and (medicine). To learn more about why tuple (computer) is in the view, the user may choose to trace its lineage. The result is shown in Figure 2.

2.1. Tracing procedure

In general, to compute the lineage of a view data item, we need the view definition and the original source data, and perhaps some auxiliary information. (In Section 2.2, we describe how we can perform lineage tracing without access to the source data.) In our system, we first transform the view definition into a normal form composed of aggregate-project-select-join sequences, called ASPJ segments. The lineage of tuples in a view defined by a single ASPJ segment can be computed using relational queries over the sources, called tracing queries, which are parameterized by the tuple(s) being traced. To trace the lineage of a view defined by multiple levels of ASPJ segments, we logically define an intermediate view for each segment, and recursively trace through the hierarchy of intermediate views top-down. At each level, we use tracing queries for a one-level ASPJ view to compute the lineage for the current traced tuples with respect to the views or base tables at the next level below.

For example, consider the view Promising. The normalized view definition with two ASPJ segments and one intermediate view Stocks is:

```
CREATE VIEW Stocks AS
SELECT t.industry, t.closing / t.earnings < 40
    AND t.cost / t.shares < t.closing
FROM Daily d, Purchases p, Earnings e, Stocks s
WHERE d.ticker = p.ticker
    AND p.ticker = s.ticker
    AND s.ticker = e.ticker
    AND d.closing < e.earnings
    AND s.earnings < 40
GROUP BY t.industry, t.closing / t.earnings, t.cost / t.shares
HAVING SUM(t.price * t.shares) < t.closing
```

Let Split be an operator that breaks a table into multiple projections, i.e., $\text{Split}_{A_1, \ldots, A_m}(T) = (\pi_{A_1}(T), \ldots, \pi_{A_m}(T))$. We generate the tracing query $TQ_1$ for segment 1 as $\text{Split}_{\text{Daily}, \text{Stocks}, \text{Earnings}}(\sigma_{C_1}(\text{Daily} \bowtie \text{Stocks} \bowtie \text{Earnings}))$ where the selection condition $C_1$ is $(\text{price}/\text{earnings} < 40 \land \text{cost}/\text{shares} < \text{closing} \land \text{industry} = t.industry)$, and $t$ is the tuple to be traced. For segment 2, we generate the tracing query $TQ_2$ as $\text{Purchases} \bowtie T$, where we trace a tuple set $T$.

When tracing the lineage of tuple $t = (\text{computer})$ in Promising, we first use the tracing query $TQ_1$ to...
compute t’s lineage in Daily, Stocks, and Earnings. The result is \(\langle \text{Daily}^*, \text{Stocks}^*, \text{Earnings}^* \rangle\), where Daily* and Stocks* are as shown in Figure 2, and Stocks* = (BBB, 3600, 150). We then further trace the lineage of Stocks* in source table Purchases, obtaining Purchases* as in Figure 2.

2.2. Auxiliary views

In the example of Section 2.1 we introduced an intermediate view Stocks for the purpose of lineage tracing. In general, such intermediate views can either be materialized, or we can recompute the relevant intermediate results at tracing time. Because efficient incremental maintenance of multi-level aggregate views generally requires materializing the same intermediate views we need for lineage tracing, currently our prototype always materializes these views as auxiliary views in the warehouse. A second type of auxiliary view is motivated by the fact that in a distributed multi-source data warehousing environment, querying the sources for lineage information can be difficult or impossible; sources may be inaccessible, expensive to access, and/or inconsistent with the views at the warehouse. By storing additional auxiliary views in the warehouse based on the source tables, we can reduce or entirely avoid source accesses for lineage tracing. There are numerous options for which auxiliary views to store, with performance tradeoffs; see [1, 2] for details. All of the auxiliary views we create to support lineage tracing are maintained consistently with the user views in the warehouse by the WHIPS prototype [4].

2.3. System architecture

Figure 3 illustrates the architecture of our lineage tracing system in the context of a data warehousing system. When a view is defined through the View Specifier, if the view definition specifies that the view should be traceable, then the Auxiliary View Generator (AVGen) automatically generates the auxiliary views discussed in Section 2.2. The Maintenance Procedure Generator (MPGen) and the Tracing Procedure Generator (TPGen) then generate the maintenance procedures (see [4]) and lineage tracing procedures for the user view as well as its auxiliary views, and store them as part of the Metadata. When a user issues a request through the Tracing Interface, the Lineage Tracer is activated and calls the appropriate sequence of tracing procedures. The lineage results are then returned to the user as tables. If the user further requests to see the derivation process, the lineage tracer combines the lineage results and the view definition to generate a derivation tree for the user, showing the complete lineage information.

References


