

Streamlining Data Warehousing through Compilation

ABSTRACT

Performing analytics on corporate relational data is difficult; Transactional relational databases optimized for small, frequent updates, are ill suited for the aggregate-heavy demands of online analytics workloads. Datacubes and similar datastructures are better suited to dealing with such workloads, but create an update bottleneck, as even incremental cube update techniques are too slow to avoid batching. In this paper, we sidestep the issue of constructing and maintaining datacubes cube updates with the SpreadDB¹ system. SpreadDB aggressively compiles datacube construction queries into a distributed, main-memory data warehouse runtime. Incremental update rules generated by SpreadDB’s compiler make it possible to dynamically make the data warehouse mirror a corporate relational database in near-real-time. Combining low-delay synchronization with purely in-memory data storage makes it possible to evaluate OLAP queries on large, *live* datasets in a matter of seconds, rather than minutes or hours.

1. INTRODUCTION

Consider a typical corporate datacube infrastructure. At one end, lie one or more OLTP databases. These databases manage corporate data in a standard relational infrastructure that permit easy updates to minor components. However, for read-heavy, non-transactional analytics workloads, this infrastructure is insufficient. Conversely, datacubes[1] excel at analytics workloads, but do not support transactional relational updates. Corporations that require heavy-duty analysis of relational data often resort to periodically extracting the contents of their relational database to re-construct or update their analytics datastructures, a process that can take hours, or even days.

Periodic scraping is inefficient. Not only does each run duplicate effort expended by prior runs, the time spent re-constructing (or even only updating) an entire datacube en-

¹The name “SpreadDB” is anonymized for double-blind reviewing

sure that analysts are working with stale versions of the data.

In this paper, we present SpreadDB, a system that constructs and maintains an in-memory OLAP database in near-real time. SpreadDB is developed based on three observations about current corporate datacube infrastructures:

1. Disk drive sizes are dramatically larger than they were a few years ago, but disk read latency remains nearly constant[]; To minimize latency, disk-based corporate datacubes and other analytics structures are stored on vastly more (nearly empty) disks than are required.[]
2. RAM is cheap, plentiful, and low-latency; Server-class machines commonly have 32 GB of ram apiece[]. Terrabyte ram DIMMs are becoming available[]
3. Aggregation is a highly parallelizable task.

Rather than maintaining a datacube partitioned across an array of disks, SpreadDB constructs the entire datacube in-memory across a network of machines. The minimal functionality required from each node in this network not only minimizes the cost of those nodes, but allows them to be constructed without the use of error-prone, temperature-sensitive disk drives.

Because all state is stored in-memory rather than on-disk, SpreadDB is amenable to the use of non-linear datastructures, for example datacubes. Moreover, because data is partitioned, the task of projecting and aggregating is easily farmed out to the hosting nodes. Using a set of novel compilation and optimization techniques, SpreadDB translates datacube construction queries into an entire data warehouse infrastructure.

SpreadDB’s compiler produces an aggressively optimized execution plan that translates updates to the individual relations of an underlying relational database updates into delta operations on the warehouse. Additionally, the compiler generates a *distribution plan* that guides data partitioning across nodes within the warehouse to maximize communication efficiency. The result is a warehouse infrastructure that not only answers OLAP queries quickly and efficiently, but that can use live data to answer those queries.

EXAMPLE 1. *As a running example in this paper, we will use a datacube loosely based on TPC-H[3] Query 12:*

The remainder of this paper is organized as follows. Section 2 discusses prior work in the field of OLAP query processing. In Section 3, we provide an overview of SpreadDB’s

Schema

```
create table customers(cid int, nid int);
create table orders(
  oid int, o_cid int, opriority int, spriority int,
  foreign key(o_cid) references customers(oid)
);
create table lineitems(
  Loid int, lateship int, latecommit bool, shipmode bool,
  foreign key(Loid) references orders(oid)
);
```

Query

```
select count(*),
  nid, cid, oid, opriority, spriority,
  lateship, latecommit, shipmode
from customers, orders, lineitems
where o_cid=cid and Loid=oid
group by cube
  nid, cid, oid, opriority, spriority,
  lateship, latecommit, shipmode;
```

Figure 1: An example query that constructs a warehouse for analyzing customer behavior with respect to shipping. Given a table of customers, orders, and lineitems in those orders, the query builds a datacube over the full join of those tables.

online infrastructure and discuss how data is managed within that infrastructure. Section 3 describes the offline compilation process used to manage and optimize SpreadDB. Section 5 presents a series of experiments that demonstrate the viability and scalability of in-memory data warehouses. Finally, Section 6 discusses a set of optimizations provoked by our experiments. The paper concludes with Section 7

2. RELATED WORK

3. ARCHITECTURE

SpreadDB consists of three components: a runtime, a *Query Layer*, and a compiler. A diagram of this architecture is presented in Figure 2. The warehousing runtime accepts relation updates at a coordinator node called the *Switch*, and distributes their data throughout an array of data warehouse

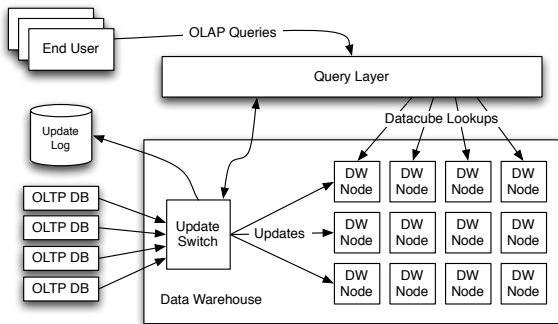


Figure 2: SpreadDB's architecture.

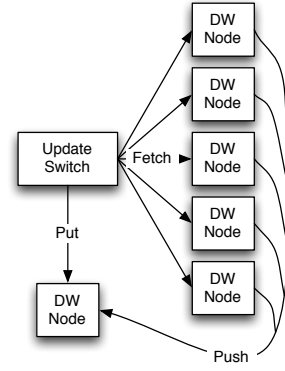


Figure 3: Information flow during one map update.

storage nodes, or *DW Nodes*. Though this paper focuses on a runtime with a centralized switch, we discuss a distributed switch implementation in Section 6.1.

Adjacent to the runtime is the *Query Layer*, a component that acts as an intermediary between the end user and the dw nodes. The query layer accepts roll-up and drill-down queries, translates them into the corresponding set of data warehouse lookups, and executes those queries on the warehouse.

SpreadDB's final component is a compiler that guides the behavior of the other three components. The SpreadDB compiler takes schemas for one or more input relations and an arbitrary SQL query as input. The query is decomposed into a set of subqueries, each representing a join over some subset of the input relations. Each subquery is materialized into a *map*, that is partitioned across the DW nodes.

Based on these subqueries, the compiler generates a series of *update rules* that translate changes in the input relations to map deltas. These update rules take the form:

$$+R(x_1, x_2, x_3, \dots) : \text{Map } K[y_1, y_2, \dots] += \text{Expression}$$

That is, when the tuple (x_1, x_2, x_3, \dots) is inserted into input relation R , adjust the target $\text{Map } K[y_1, y_2, \dots]$ by the indicated arithmetic expression over constants, variables (x_n) , or the contents of other maps.

It is possible for an update rule to modify an entire cross section of a map. Such updates are expressed via variables in the update rule not bound to one of the input parameters. We refer to these as loop variables. A loop variable occurs exactly twice in an update rule, once as a key in the target map, and once as a key for a map in the expression. Thus an update rule with loop variables behaves as if it were a set of parallel updates; every update corresponds to one possible valuation of all loop variables selected from the domains of the map they index into.

When a tuple is added to or removed from an input relation, the Switch composes a set of messages parametrized on the tuple's contents and dispatches them into the data warehouse as a sequence of map reads and writes. Nodes being read from accept and process the requests and forward the responses to the appropriate writers. We refer to this entire sequence of operations as an update.

3.1 Anatomy of an Update

A map update begins at the Switch when an input table

```

1:  $Msgs \leftarrow \emptyset$ 
2: for all Update  $U \in \text{Trigger}$  do
3:   Region  $Reg = \pi_{U.target\_loop\_vars}(index(U.target\_map))$ 
4:   for all Partition  $\{P | P \in U.target\_map \wedge (P \cap Reg \neq \emptyset)\}$  do
5:      $Reads = \emptyset$ 
6:     for all Ref  $R \in get\_map\_refs(U.expression)$  do
7:        $RReg \leftarrow \pi_{R.loop\_vars}P$ 
8:        $RPart \leftarrow \{RP | RP \in R.map \wedge (RP \cap RReg \neq \emptyset)\}$ 
9:        $Reads \leftarrow Reads \cup \{(ReadP.node, R)\}$ 
10:    end for
11:     $Reads \leftarrow group\_by\_node(Reads)$ 
12:     $Msgs \leftarrow put(U, P, Reads.size)$ 
13:    for all (Node  $N, \{Ref\ R\}) \in Reads$  do
14:      // PUSH results to  $P.node$ 
15:       $Msgs \leftarrow fetch(N, \{R\}, P.node)$ 
16:    end for
17:  end for
18: end for

```

Figure 4: The Switch’s trigger dispatch algorithm

is modified.

Each input table is associated with one or more triggers, each requiring a write to some range of map values and zero or more reads from different maps. For each trigger, the Switch identifies and sends a PUT message to each DW Node managing a map being written to. Additionally, the Switch identifies all DW Nodes managing maps that will be read from and sends a FETCH message to them, requesting the desired map values. Finally, if required (i.e., if SpreadDB is being used standalone, without an underlying database), the switch can log the update to disk for persistence.

The nodes receiving a FETCH request perform the appropriate reads and send their responses to the PUT nodes in a PUSH message. Upon receipt of all necessary PUSH messages, the PUT nodes compute the update expression and modify the affected maps. An illustration of the complete message-passing process is presented in Figure 3.

3.1.1 Trigger Dispatch

Maps are partitioned along dimensional axes when they grow beyond the capacity of a single node, akin to the partitioning done in grid files[2]. To streamline the dispatch of messages to component nodes, the switch pre-generates a spatial index for each template, similar to the grid file directory. Every entry in the spatial index contains a set of PUT and FETCH messages. When the template is triggered, the tuple’s values are used to index into the spatial index and the corresponding messages are parametrized and sent. The algorithm for generating the spatial index is presented in Figure 4.

Looping updates require the Switch to match corresponding read and write partitions. The correspondence is obtained by identifying intersections between partitions of the target map that are affected by the update, and those of each map in the update expression. This is equivalent to a join over components of the spatial index stored at the Switch. Loop-free updates are a special case of this, where only one partition is required from each map. The trigger dispatch algorithm is shown in Figure 4.

3.1.2 Get Collation

```

1: Given: Update  $U$ 
2: for all Ref  $R \in U.expression$  do
3:    $Inputs \leftarrow Inputs \cup (R, \emptyset)$ 
4: end for
5: for all (Ref  $InR$ , Value  $V$ )  $\in \cup(msgPUSH)$  do
6:   for all ( $R, Table$ )  $\in Inputs$  do
7:     if  $check\_match(InR, R)$  then
8:        $Table \leftarrow Table \cup (InR.keys, V)$ 
9:     end if
10:  end for
11: end for
12: for all ( $K, \{V\}$ )  $\in JOIN(Keys, Val) \in Inputs$  do
13:    $Target = bind\_vars(U.target \leftarrow Keys)$ 
14:    $apply(Target, bind\_vars(U.expression \leftarrow \{V\}))$ 
15: end for

```

Figure 5: The DW Node’s collation algorithm

FETCH responses, or PUSH messages for an update are sent to the node managing the partition being updated. Having received the number of FETCH messages sent by the switch with the PUT message, the destination node can buffer PUSH messages until all have arrived. At this point, if the update is loop-free, the destination node simply uses the contents of the PUSH messages to evaluate the update expression.

If the update requires a loop, the destination node must do some processing. The node first generates a set of tables, one for each map reference in the update expression. Arriving map values populate tables that correspond to any map reference matching the value’s keys, where loop variables act as wildcards. When all values have been received, the destination node computes the natural join of all of the generated tables, effectively producing one update value for every assigned value in the domain of all of the loop variables. The loop-free update is a special case of this where each generated table contains only one row. The collation algorithm is shown in Figure 5

3.2 Consistency Model

The materialize

SpreadDB’s update delta-encoding uses materialized subqueries already stored within the warehouse to perform updates. This dependency requires that all reads see a consistent snapshot of the warehouse at the completion of the previous update. Thus, all reads and writes must be executed as if there were a total ordering over all updates. As the clearinghouse for updates, the Switch presents an ideal point for generating this ordering.

Every update is assigned a version number by the switch as it is issued. The As they are

Every update is assigned a version number by the switch. The maps that each update modifies and reads are known. The switch maintains a list of modification times for each partition in the warehouse. With each request to the DW Nodes, the switch attaches a version

first looks up the most recent modification time for each get required by the update, and subsequently updates the modification times for each put required by the update.

When applying gets, each warehouse node first ensures that it has successfully applied all requests prior to the indi-

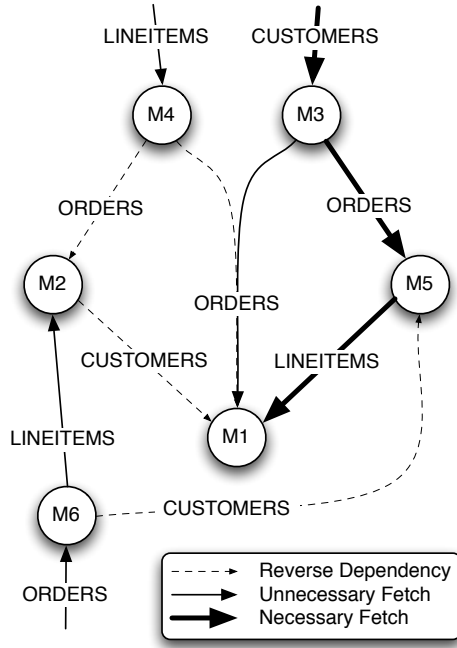


Figure 6: Data-flow graph for the example query. Light and dashed edges can be optimized out, given cascading delete foreign key constraints.

cated version number. A short history of prior incarnations of each partition are also maintained in case a get arrives after a subsequent put. Each put also includes the version number of the last fully completed update. This allows the target node to discard old incarnations of the partition.

Some trickery will be required to get this table into the middleware. Perhaps the switch can periodically broadcast it over?

4. COMPILATION

4.1 Update Rules

- Translating SQL into Update Rules
- Update DAGs/Data flow graphs
- Exploiting Foreign Keys
- Cascading Map Rejection

4.2 Layout

How do we partition data (the layout)? Where do various bits of code get executed live? What kind of runtime analytics do we need to collect to manage the data layout on the fly?

5. EXPERIMENTS

6. OPTIMIZATIONS

6.1 Distributing the Switch

The role of the switch is to provide a synchronization mechanism for the updates. Specifically, the delta encoding approach used by the update rules requires that all update

rules be applied to a consistent snapshot of the maps. The current implementation of SpreadDB performs this synchronization at a single node. However, limiting the switch to only one node creates a scaling bottleneck.

Despite the cloud computing mantra of rejecting consistency, this application requires it. Complex locking protocols have poor scaling performance, so a simpler, lock-free protocol is required. We achieve this goal by introducing the notion of *pipeline scheduling*. Pipeline scheduling exploits the acyclicity of compiler-generated data-flow graphs to allow nodes to correctly interleave and process update rules with only limited network overhead and processing latency.

6.2 Pipeline Scheduling

Loose clock synchronization between nodes is assumed, and allows time to be partitioned into a sequence of numbered ticks, each lasting on the order of seconds. When a set of updates is triggered, a switch node sends tentative put requests to all participating nodes. These nodes respond with their current tick counter, and the switch forwards the maximum returned tick to all nodes.

The updates are considered to have been posted at the maximum returned tick. However processing is deferred for a number of ticks equal to the depth of the map being updated. The result is a data-flow process resembling a parallelized CPU pipeline.

During a given tick, all updates scheduled for processing are evaluated. Once all updates scheduled for the tick have completed, the node responds to FETCH requests for the tick with PUSH messages. Recall that each edge in the data flow graph represents a FETCH request for a specific update. The difference in depth between the edge's ends is the number of ticks in advance of the PUT that the response is sent.

For example, in Figure 6 without removing any edges, $Depth(M1) = 2$, $Depth(M2) = 1$, and $Depth(M4) = 0$. Updates to map $M1$ scheduled for processing during tick 4 would receive data from map $M2$ during tick 3, and from map $M1$ during tick 2.

7. CONCLUSIONS

8. REFERENCES

- [1] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery*, 1(1):29–53, March 1997.
- [2] J. Nievergelt, H. Hinterberger, and K. C. Sevcik. The grid file: An adaptable, symmetric multikey file structure. *ACM Trans. Database Syst.*, 9(1):38–71, 1984.
- [3] Transaction Processing Performance Council. *TPC Benchmark H (Decision Support)*, revision 2.6.0 edition, 2006. <http://www.tpc.org/tpch/spec/tpch2.6.0.pdf>.