

Measures in SQL

Julian Hyde
Google Inc.
San Francisco, CA, USA
julianhyde@google.com

John Fremlin
Google Inc.
New York, NY, USA
fremlin@google.com

ABSTRACT

SQL has attained widespread adoption, but Business Intelligence tools still use their own higher level languages based upon a multidimensional paradigm. Composable calculations are what is missing from SQL, and we propose a new kind of column, called a measure, that attaches a calculation to a table. Like regular tables, tables with measures are composable and closed when used in queries.

SQL-with-measures has the power, conciseness and reusability of multidimensional languages but retains SQL semantics. Measure invocations can be expanded in place to simple, clear SQL.

To define the evaluation semantics for measures, we introduce context-sensitive expressions (a way to evaluate multidimensional expressions that is consistent with existing SQL semantics), a concept called evaluation context, and several operations for setting and modifying the evaluation context.

CCS CONCEPTS

• Information systems → Relational database query languages; Data analytics; Online analytical processing.

KEYWORDS

data management, query processing, business intelligence

ACM Reference Format:

Julian Hyde and John Fremlin. 2024. Measures in SQL. In *Companion of the 2024 International Conference on Management of Data (SIGMOD-Companion '24)*, June 9–15, 2024, Santiago, AA, Chile. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3626246.3653374>

1 INTRODUCTION

About thirty years ago the first Business Intelligence (BI) tools were introduced. They had a semantic model based on the multidimensional model, and good support for data exploration and visualizations. Since then, the SQL language has expanded immeasurably in its capabilities, adding support for XML, JSON, geospatial, temporal, text and nested data. An increasing proportion of business data resides in powerful cloud SQL engines. But today's BI tools continue to use semantic models based on the multidimensional model. Why?

Semantic models serve several purposes. They provide the building blocks from which users can build queries (using some language,

perhaps graphical, perhaps textual). They guide users in the construction of queries, and aid creation in visualizations and reports. But we believe that their core strength is the ability to express calculations in a concise manner, and to compose and reuse those calculations.

In this paper, we show that the relational model imposes repetition of filter expressions: changing the date range of a query requires updating many WHERE clauses. Therefore the challenge is how to extend the data model offered by SQL, in ways that do not change the semantics of currently valid SQL expressions or confound SQL users' expectations. Incorporating ideas from software engineering, we extend SQL's fundamental data type, the table, with a new type of column, called a *measure*, attaching a *context-sensitive expression* (CSE) to a table. We show that tables with measures have a similar closure property to regular tables.

SQL with measures can be expanded into traditional SQL. Therefore, the path to integrating measures into existing systems is relatively straightforward.

1.1 Contributions

Encapsulation. Measures define calculations close to the data. When a measure is used, it maintains its relationship to its table.

Clarity of query plan. By eliminating the need for self-joins and other forms of repetition in many queries, measures make it easier for the optimizer to choose more efficient algorithms.

Easier target for generative AI. Generative AI algorithms find it hard to correctly generate SQL queries that have repeated subqueries, especially if those subqueries need to be consistent. Measures enable more concise queries that are easier for AI to generate.

Modeling. Measures allow you to define calculations in views, and CSEs allow you to compose calculations into richer measures. SQL can therefore take over work that was previously done in a BI tool (semantic layer and multidimensional query language).

Abstraction. You can use a view without having knowledge of the formulas in that view, or access to the tables referenced by the view.

All of the above are delivered while retaining SQL's closure properties, security, and governance. Our extensions are backwards compatible: queries that do not use measures have the same semantics as regular SQL.

These extensions have been implemented in Apache Calcite [2] and in Looker's Open SQL interface [7] (described further in subsection 5.6).

2 RELATED WORK

Adding measures to SQL requires us to bring together two theories, and classes of database, long considered to be incompatible. It is worth reviewing their parallel histories and the path to convergence.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMOD-Companion '24, June 9–15, 2024, Santiago, AA, Chile

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0422-2/24/06

<https://doi.org/10.1145/3626246.3653374>

The relational model [3] gave rise to relational databases during the 1970s, and the dimensional model to multidimensional databases in the 1990s. Vendors of the latter, assisted by E.F. Codd [4], were at pains to point out how relational databases' record-oriented storage was fundamentally unsuitable for OLAP [5].

Multidimensional databases had no textual query language and were generally inseparable from their user interface (which was provided by the same vendor); in early attempts to standardize access to multidimensional databases (MDAPI [16] and JOLAP [13]) programmers would construct queries by calling an API.

It was difficult to imagine unifying relational and multidimensional databases when they differed in the fundamentals: whether there should be a query language; the data model (relations, rows and columns versus cubes, dimensions, hierarchies, and measures); and the algebra (relational operators select, project, join, union versus dimensional operators slice, dice, drill, pivot [15, 18]).

Things began to change in the late 1990s. Kimball [14] introduced patterns to model complex business analytics. In particular, semi-additive and non-additive measures [10] were patterns for measures more complex than mere aggregate functions. The SQL CUBE operator [9] showed that it was possible to represent various levels of subtotals in a query result without adding the complexity of hierarchies to the data model. Analytic functions (OVER) [20] allowed running totals and calculations of mixed grain, in some cases allowing the elimination of self-joins [22]. FILTER, WITHIN GROUP and WITHIN DISTINCT clauses [11] provided finer control over the values going into an aggregate function.

MDX was (at last!) a textual language for dimensional queries [19]. Unfortunately, its designers chose a syntax that was superficially similar to SQL, and therefore many failed to grasp its radically different semantics. Among those features were an evaluation context consisting of one member for each of the current cube's dimensions, and the ability to define calculated measures and members using context-sensitive expressions. As a standard language, there were multiple implementations of MDX, including Microsoft Analysis Services, Mondrian [1], SAP BW, and SAS. Some of these implementations were backed by relational databases (a technique called ROLAP [17]), and dimensional languages came to be seen as a semantic layer on top of the relational model.

The semantic layer's main contribution was not cubes. (Data sets with axes, hierarchies and cells are harder for downstream tools to consume than relations.) It was the ability to define, just once, the calculations central to the business, and to associate columns with presentation metadata such as value formats and default sort order. For example, Tableau's Level of Detail (LOD) expression language allows users to control the grain at which aggregations occur; Looker's centralized model makes governance easier and makes calculations consistent.

But these semantic layers' languages were not SQL; to benefit from a semantic layer, users had to use its less-expressive, vendor-specific query language. In the next section, we describe how to extend SQL so that it can serve as the semantic layer.

3 MEASURES

In this section we describe the new concepts and their SQL syntax. We illustrate with examples; semantics are deferred to section 4.

3.1 Tables are broken

Tables are SQL's fundamental data model. Tables are implemented in several ways, including base tables, views, and query specifications, but any table you use in a query will have the same behavior. If you have a query that uses a view, and you substitute a base table that has the same rows as the view, the query will give the same results. Furthermore, the model is closed: every SQL query returns a table.

Tables are unable to provide reusable calculations. Suppose we have an Orders table that contains several orders for each product name and customer (table 2), and an expert SQL user has written a query (listing 1) to compute the profit margin for each product.

custName	custAge
Alice	23
Bob	41
Celia	17

Table 1: Customers table

prodName	custName	orderDate	revenue	cost
Happy	Alice	2023/11/28	6	4
Acme	Bob	2023/11/27	5	2
Happy	Alice	2024/11/28	7	4
Whizz	Celia	2023/11/25	3	1
Happy	Bob	2022/11/27	4	1

Table 2: Orders table

```
SELECT prodName,
COUNT(*) AS c,
(SUM(revenue) - SUM(cost)) / SUM(revenue)
AS profitMargin
FROM Orders
GROUP BY prodName;
```

Listing 1: Summarizing Orders by product name

We now wish to create a SQL view that will allow less-expert users to perform similar queries without typing out the formula for profit margin. Listing 2 creates the view SummarizedOrders and attempts to use its profitMargin column in a query to compute the profit margin for each product.

```
CREATE VIEW SummarizedOrders AS
SELECT prodName, orderDate,
(SUM(revenue) - SUM(cost)) / SUM(revenue)
AS profitMargin
FROM Orders
GROUP BY prodName, orderDate;

SELECT prodName, AVG(profitMargin)
FROM SummarizedOrders
GROUP BY prodName;
```

Listing 2: SummarizedOrders view

The query does not return the desired result; the desired result would weigh each order equally, but actual result is an average over each (prodName, orderDate) combination. There is no correct query in valid SQL; any correct query must read all rows

in Orders, but the rules of relational algebra do not allow the SummarizedOrders view to return any more information to a query than its (summarized) rows.

3.2 Measures and the AGGREGATE aggregate function

To solve the problem, we introduce measures. Informally, a measure is a column defined by a formula, and when that measure is used, the formula is ‘copy-pasted’ into the invocation. (More formally, as we shall see later, a measure behaves as a context-sensitive expression, taking its evaluation context from the clause in which it is used.) This means each use of a measure can be expanded into an traditional SQL query by explicitly, repetitively spelling out the contextual filters.

Listing 3 defines a view, EnhancedOrders, that contains a measure, and uses it in a query.

```
CREATE VIEW EnhancedOrders AS
  SELECT orderDate, prodName,
    (SUM(revenue) - SUM(cost)) / SUM(revenue)
    AS MEASURE profitMargin
  FROM Orders;

SELECT prodName, AGGREGATE(profitMargin)
FROM EnhancedOrders
GROUP BY prodName;
```

Listing 3: EnhancedOrders view

There are a few things to note:

- The AS MEASURE syntax indicates that profitMargin is to be a measure, not a regular column.
- The EnhancedOrders view does not contain a GROUP BY clause, and therefore returns the same number of rows as the Orders table.
- The measure formula contains aggregate functions, which would not be valid if this were a normal query. Measures need to be aggregatable — that is, valid with any possible GROUP BY clause in the query that uses it — and therefore their formulas often contain aggregate functions.

The query uses the profitMargin measure and evaluates it in the context of the current group row, aggregating over all rows with the current value of prodName. Users of the EnhancedOrders view do not need to know the formula for profitMargin, nor need access to the underlying Orders table or its revenue and cost columns, which meets our goal of providing an abstraction.

3.3 Measures are not really aggregate functions

The AGGREGATE function is present for largely cosmetic reasons. SQL users know that a column that is not in the GROUP BY clause must be wrapped in an aggregate function when used in the SELECT clause, so the AGGREGATE function makes such users (and tools that generate SQL) more comfortable. As an aggregate function, AGGREGATE conveniently converts any query into an aggregate query.

But framing measures as aggregate functions sells them short. They are in fact evaluated very differently from aggregate functions. Consider the following query (listing 4).

```
SELECT prodName, AGGREGATE(profitMargin),
  COUNT(*)
FROM EnhancedOrders
GROUP BY prodName;

prodName  profitMargin  count
=====  =====
Acme      0.60             1
Happy     0.47             3
Whizz     0.67             1
```

Listing 4: Evaluating a query

What happens in the SELECT clause as the query is about to emit the row for ‘Happy’? GROUP BY has assembled a group of 3 rows for which prodName equals ‘Happy’, and the COUNT(*) aggregate function is evaluated in the usual way over these rows, emitting the value 3.

The measure does not operate on the group rows (except indirectly). Its only argument is the evaluation context, which consists of the predicate¹

prodName = 'Happy'.

The effect is as if the query has been expanded as follows (listing 5):

```
SELECT prodName,
  (SELECT (SUM(i.revenue) - SUM(i.cost)) / SUM(i.revenue)
   FROM Orders AS i
   WHERE i.prodName = o.prodName),
  COUNT(*)
FROM Orders AS o
GROUP BY prodName;
```

Listing 5: Query after expansion of measure

The measure has been replaced by a scalar subquery. The subquery is over Orders, the base table of the view in which the measure was defined, and uses the same formula. To the subquery has been added a WHERE clause that expresses the evaluation context, and therefore the formula will be evaluated over the precise subset of rows in Orders.

In the next section, we shall define measures in terms of context-sensitive expressions.

3.4 Context-Sensitive Expressions

You might regard a measure as simply ‘a column that knows how to aggregate itself,’ and indeed many measures are just that. But the goal is reusable calculations, which means that the client query does not know the measure’s formula, and the measure may use data that is not accessible to the client.

So, we define the behavior of measures in terms of a new concept: the *context-sensitive expression*. Some definitions:

- A **context-sensitive expression** (CSE) is an expression whose value is determined by an evaluation context.
- An **evaluation context** is a predicate whose terms are one or more columns from the same table.
- This set of columns is the **dimensionality** of the CSE; we sometimes informally refer to these columns as **dimension columns** even though they are regular columns.

¹We have simplified a little; if prodName allowed null values, the predicate would use IS NOT DISTINCT FROM, rather than =, in order to handle null values correctly.

- A **measure** is a special kind of column that becomes a CSE when used in a query. Its dimensionality is the set of non-measure columns in its table.
- If a query references a table that has a measure, then any use of that measure in an expression has an **implicit evaluation context**. This context depends on the values of the measure's dimension columns and on the call site (which query clause, and whether there are joins or filters).
- The **data type** of a CSE is t MEASURE, for some data type t ; for example INTEGER MEASURE).
- The **evaluation operator** EVAL evaluates a CSE in the evaluation context of the call site; if the expression has type t MEASURE, the value has type t .
- The **context transformation operator** AT modifies the evaluation context.

Applying these concepts to the query in listing 4:

- The measure in the query is profitMargin, and its dimensionality is the column set {prodName, orderDate}.
- profitMargin has type DOUBLE MEASURE, and therefore AGGREGATE(o.profitMargin) has type DOUBLE.
- profitMargin is a measure, and therefore a reference to it is CSE.
- AGGREGATE(o.profitMargin) expands to EVAL(o.profitMargin AT (VISIBLE)).²
- The call site is the SELECT clause of an aggregate query, and therefore the evaluation context is a predicate that restricts to the rows matching the current group key, prodName = o.prodName. Per the requirements of an evaluation context, it is in terms of one of profitMargin's dimension columns, prodName. (The right-hand side of the equality, o.prodName, is a correlation variable that is effectively constant when the predicate is invoked.)
- Substituting the measure with a scalar subquery and a predicate that expresses the evaluation context yields the expanded query in listing 5, as expected.

CSEs and aggregate functions have fundamentally different evaluation models:

- Aggregate functions, like relational algebra, are *bottom-up*. The result of the calculation depends on the input rows, and the sequence of operators applied to them.
- CSEs are *top-down*. The result of the calculations is determined by the evaluation context.

The top-down evaluation model has a number of advantages.

- Whereas aggregate functions can only be used in call sites where there is a set of rows to aggregate over, such as the SELECT or HAVING clause of a GROUP BY query, measures and CSEs can be evaluated at any call site.
- If you wish to evaluate a calculation in different contexts (say to compute profit growth between last year and this year, or to compare profit for a particular product with that for all products), top-down is more concise. In bottom-up, each calculation requires a separate pass over the input rows. In practice, this results in queries that have similar repeated

subqueries and self-joins to combine the results of those subqueries on their common keys.

- Top-down makes it easier to manage the grain of a calculation (daily versus monthly, per-order versus per-customer). A measure is locked to the grain of its defining table, and joining another table does not introduce double-counting the way it often does for bottom-up calculations.

3.5 Modifying the evaluation context

In the previous section we saw that CSEs are evaluated in an evaluation context that depends on the call site. We now introduce the AT operator, which allows you to modify the evaluation context. Syntax is as follows:

cse AT (modifiers)

where *cse* is a context-sensitive expression and *modifiers* is a list of context modifiers as shown in table 3. If there are multiple modifiers, they take effect in sequence; *cse AT (modifier₁ modifier₂)* is equivalent to *(cse AT (modifier₂)) AT (modifier₁)*.

Syntax	Effect
ALL	Sets the evaluation context to TRUE
ALL <i>dimension</i> [<i>dimension...</i>]	Removes any <i>dimension</i> terms from the evaluation context
SET <i>dimension</i> = <i>expression</i>	Adds a <i>dimension</i> = <i>expression</i> term to the context (replacing any occurrence of CURRENT <i>dimension</i> with the current value of <i>dimension</i>), removing any existing <i>dimension</i> terms
VISIBLE	Adds terms to the evaluation context for the current query's WHERE clause and join conditions (if present), to ensure that measures are calculated over only the rows returned by the query
WHERE <i>predicate</i>	Sets the evaluation context to <i>predicate</i>

Table 3: Context modifiers

ALL. The ALL modifier allows you to compute a grand total. For example, the following query (listing 6) shows each product's revenue and its proportion of the total revenue:

```
SELECT prodName, sumRevenue,
       sumRevenue / sumRevenue AT (ALL prodName)
       AS proportionOfTotalRevenue
FROM (
  SELECT *, SUM(revenue) AS MEASURE sumRevenue
  FROM Orders) AS o
GROUP BY prodName;
```

Listing 6: Query with proportion of total revenue

When the query is emitting a row, the evaluation context for the top-level sumRevenue will be prodName = o.prodName, but due to the AT operator, the evaluation context for the sumRevenue measure inside the sumRevenue AT (ALL prodName) expression will be TRUE. The measure sumRevenue will be evaluated by iterating over the orders of a particular product, whereas sumRevenue AT (ALL prodName) will be evaluated by iterating over all orders.

²The AT operator and its VISIBLE modifier will be explained in subsection 3.5.

ALL with no arguments removes all filters, even filters not associated with a particular dimension, and therefore the measure will be evaluated over its entire source table.

SET. The SET modifier allows you to change the value of one dimension. The following query (listing 7) uses SET with the orderYear dimension to show profit margins in 2024 and 2023 for products sold in 2024:

```
SELECT prodName, orderYear,
       profitMargin,
       profitMargin AT (SET orderYear = CURRENT orderYear - 1)
       AS profitMarginLastYear
FROM (
  SELECT *,
         (SUM(revenue) - SUM(cost)) / SUM(revenue)
         AS MEASURE profitMargin,
         YEAR(orderDate) AS orderYear
  FROM Orders
)
WHERE orderYear = 2024
GROUP BY prodName, orderYear;
```

Listing 7: Comparing profit margins in 2023 and 2024

This query is doing something novel for SQL: it is evaluating the profitMargin measure over data that has already been removed from the query by the WHERE clause.

The CURRENT qualifier applied to a dimension returns the null value if the dimension has not been constrained to a single value by a SET modifier or WHERE clause in the enclosing evaluation context.

If the argument to SET (or ALL) is an expression, such as DAY-OFWEEK(orderDate), it is treated as an *ad hoc* dimension. *Ad hoc* dimensions do not greatly complicate the semantics for evaluating measures. All filters in the evaluation context, whether on dimensions, or on expressions involving dimensions, are combined into a single predicate, and the measure value is only determined only by values returned by the predicate, not the structure of the expressions that built that predicate.

VISIBLE. The VISIBLE modifier adds terms to the evaluation context so that the measure only includes rows allowed by the current WHERE clause³. Consider the following query (listing 8), which computes the count and sum of revenue for orders not made by Bob, grouped by product.

```
SELECT o.prodName,
       COUNT(*) AS c,
       AGGREGATE(o.sumRevenue) AS rAgg,
       o.sumRevenue AT (VISIBLE) AS rViz,
       o.sumRevenue AS r
FROM (SELECT *, SUM(revenue) AS MEASURE sumRevenue
      FROM Orders) AS o
WHERE o.custName <> 'Bob'
GROUP BY ROLLUP(o.prodName);
```

prodName	c	rAgg	rViz	r
Happy	2	13	13	17
Whizz	1	3	3	3
	3	16	16	25

Listing 8: Query with visible totals

Do we wish the grand total (the last row, with empty prodName) to include purchases by Bob, excluded by the WHERE clause? There are cases where each would make sense, and the VISIBLE modifier makes it possible to choose. The r column, which uses the

³And join conditions, as we shall see in subsection 3.6

default evaluation context ignoring the WHERE clause, includes all customers; rViz, which uses the VISIBLE modifier, includes only orders not made by Bob.

COUNT and AGGREGATE (columns c and rAgg) total only the visible rows, as is customary for SQL aggregate functions. This is why we remarked earlier that AGGREGATE(*m*) expands to EVAL(*m* AT (VISIBLE)) for any measure *m*.

Advanced context modifiers. We do not regard the list of modifiers allowed by the AT operator as complete or final. For instance, there is a compelling argument for ‘named filters’ that can be added by a UI control and removed or overridden in the evaluation context by the SQL runtime, but we have not included them in this paper. The reason is simple: when a measure is evaluated, it cares only about the predicate — do I include this row in the total, or not? — and not about the structure of the evaluation context that created the predicate.

We look forward to useful context modifiers devised by others, and we believe that they will not change the fundamentals of how measures are evaluated.

3.6 Measures and joins

It’s worth discussing how measures work in join queries, because people’s desired semantics are complicated, and because the natural semantics of measures is different — we believe in a good way — from people’s expectations of SQL, namely aggregate functions.

Consider a query that joins a table with measures (EnhancedCustomers) to another table (Orders).

```
WITH EnhancedCustomers AS (
  SELECT *, AVG(custAge) AS MEASURE avgAge
  FROM Customers)
SELECT o.prodName,
       COUNT(*) AS orderCount,
       AVG(c.custAge) AS weightedAvgAge,
       c.avgAge AS avgAge,
       c.avgAge AT (VISIBLE) AS visibleAvgAge
FROM Orders AS o
JOIN EnhancedCustomers AS c USING (custName)
WHERE c.custAge >= 18
GROUP BY o.prodName;
```

Listing 9: Joining measures

The join is one-to-many. A given customer may match zero, one or many orders. The query semantics do not depend on the SQL system knowing which primary keys and foreign keys exist. That would arguably contradict the data independence principles of SQL.

How many rows are returned? What are the values of prodName and orderCount? These are straightforward questions to answer, because measures do not affect the basic operations of SQL, such as the number of rows in a relation. A row is returned for each product that has at least one order to a customer 18 or older, and the count is the number of orders.

The weightedAvgAge column computes the average customer age in the traditional SQL way. It joins orders to customers, removes child customers under 18, and for all joined rows with the same product computes a weighted average of the ages. If one product has one order, and another has two orders from the same customer, the second contributes twice as much to the average as the first.

Which average is correct — the weighted average, the visible average (containing customers only 18 or older), or the unweighted

average — depends, of course, on what you want the number for, but it is useful that there is a concise syntax for each.

4 SEMANTICS

In the previous section, we introduced several new SQL concepts: measures, context-sensitive expressions, and operations that modify the evaluation context. We now define their semantics.

In our data model for analytic SQL, which adds measures to tables, it is important to separate how measures are defined from how they are used. A measure *may* be defined using the AS MEASURE construct, or it may be defined in some other way, but any query that uses that measure should never be able to tell.

To keep the semantics separate, we therefore proceed as follows. First we define the evaluation context, and how it is perceived by expressions. Then we define how a table interacts with the query optimizer to convert measure references into expressions. Lastly, we define the AT operator.

4.1 Lambdas

In order to simplify the explanation of semantics, we use a functional extension to the SQL language, as follows.

A note on safety. Adding function values, also known as closures or lambdas, to SQL would make the language Turing complete, and therefore make it difficult to reason about query termination. This proposal does not step into those stormy waters. First, these extensions are expanded for the query optimizer. We do not make them accessible to the SQL user. The use of closures here is just for clarity of exposition, particularly to clarify which definition is meant when a name is defined in different scopes. Second, the closures that we introduce during the planning process are gone before planning is complete. There are no function values at runtime.

- A **closure** represents a function expression. Its type is

FUNCTION(*A*) RETURNS *R*,

where FUNCTION is a type constructor, *A* is the argument type and *R* is the result type.

- A **lambda** (denoted $\rightarrow$) is a SQL operator that denotes a closure. For example,

$(x : \text{INTEGER}) \rightarrow \text{MOD}(x, 2) = 0$

is a function expression that returns whether its integer argument is even; its type is FUNCTION(INTEGER) RETURNS BOOLEAN.

- **APPLY** is a SQL operator that applies a closure to an argument. For example,

$\text{APPLY}((x : \text{INTEGER}) \rightarrow \text{MOD}(x, 2) = 0, 3)$

returns FALSE, because 3 is not even.

4.2 Semantics of context-sensitive expressions

Having defined lambdas, we outline a process to rewrite measures.

- For every measure *M* of value type *V* that belongs to a table whose row type (excluding measures) is *R*, the system defines an **auxiliary function** that has name compute_M^4 and type FUNCTION(rowPredicate: FUNCTION(*R*) RETURNS BOOLEAN)

RETURNS *V*. The auxiliary function must be pure and deterministic but may contain a reference to the table.

- At any point in the query where *M* is accessible, the system is able to generate a **row predicate** of type FUNCTION(*R*) RETURNS BOOLEAN. The row predicate reflects the evaluation context of the measure.
- If an expression occurs within a call to AT, the evaluation context is modified by applying the modifiers in succession.
- From an evaluation context for *M* can be generated a **row predicate** of type FUNCTION(*R*) RETURNS BOOLEAN
- At any point in the query where *M* is referenced in an expression, the compiler replaces the measure reference with a call to its auxiliary function; the argument is the row predicate and the return value has type *V*, as required.

Here is an example that follows the above rules. We have the following query (listing 10) that computes the ratio of this year's revenue to last year's revenue, for each product.

```
CREATE VIEW OrdersWithRevenue AS
SELECT *, SUM(revenue) AS MEASURE sumRevenue
FROM Orders;

SELECT prodName, YEAR(orderDate) AS orderYear,
       sumRevenue / sumRevenue AT
       (SET orderYear = CURRENT orderYear - 1) AS ratio
FROM OrdersWithRevenue
GROUP BY prodName, YEAR(orderDate);
```

Listing 10: Year over year revenue by product

The measure *M* is sumRevenue, and the row type *R* is the type OrdersRow consisting of the non-measure columns of the Orders view. Listing 11 shows the definition of a type for *R*, and the query with the two references to sumRevenue replaced by calls to the auxiliary function computeSumRevenue. Each call has a row predicate that reflects the evaluation context at its call site. The first call has the evaluation context of output from the GROUP BY; in the second call, the year in the filter context is set to the year before the current one.

```
-- Row definition
CREATE TYPE OrdersRow AS ROW (prodName: VARCHAR,
                              custName: VARCHAR, orderDate: DATE,
                              revenue: INTEGER, cost: INTEGER);

-- Auxiliary computation for sumRevenue
CREATE FUNCTION computeSumRevenue(
    rowPredicate: FUNCTION(r: OrdersRow)
    RETURNS BOOLEAN) AS
SELECT SUM(o.revenue)
FROM Orders AS o
WHERE APPLY(rowPredicate, o);

-- After expansion of sumRevenue occurrences
SELECT o.prodName, YEAR(o.orderDate) AS orderYear,
       computeSumRevenue(
           r -> r.prodName = o.prodName AND
           YEAR(r.orderDate) = YEAR(o.orderDate))
       / computeSumRevenue(
           r -> r.prodName = o.prodName AND
           YEAR(r.orderDate) = YEAR(o.orderDate) - 1)
AS ratio
FROM Orders AS o
GROUP BY prodName, YEAR(orderDate);
```

Listing 11: Expansion of query comparing average revenue

⁴Or a variation of that name that is unique within the namespace

5 DISCUSSION

5.1 Self-joins and window aggregates

There is a fascinating correspondence between measure expressions, window aggregates, and self-joins.

The correspondence between window aggregates and self-joins (expressed in the form of correlated subqueries) was first noted in [22], whose WinMagic algorithm rewrites certain kinds of subquery to window aggregates. The four queries in listing 12 are equivalent, and all find orders whose revenue is higher than the average for their product. WinMagic provides an algorithm to rewrite query 1 (correlated subquery) to query 3 (window aggregates); queries 2 and 4 are equivalent queries using self-join and measures.

```
-- Query 1: correlated subquery
SELECT o.prodName, o.orderDate
FROM Orders AS o
WHERE o.revenue >
  (SELECT AVG(revenue)
   FROM Orders AS o1
   WHERE o1.prodName = o.prodName);

-- Query 2: self-join
SELECT o.prodName, o.orderDate
FROM Orders AS o
LEFT JOIN
  (SELECT prodName, AVG(revenue) AS avgRevenue
   FROM Orders
   GROUP BY prodName) AS o2
ON o.prodName = o2.prodName
WHERE o.revenue > o2.avgRevenue;

-- Query 3: window aggregate
SELECT o.prodName, o.orderDate
FROM
  (SELECT prodName, revenue, orderDate,
         AVG(revenue) OVER (PARTITION BY prodName
                           AS avgRevenue
                           FROM Orders) AS o
   FROM Orders) AS o
WHERE o.revenue > o.avgRevenue;

-- Query 4: measures
SELECT o.prodName, o.orderDate
FROM
  (SELECT prodName, orderDate, revenue,
         AVG(revenue) AS MEASURE avgRevenue
   FROM Orders) AS o
WHERE o.revenue >
  o.avgRevenue AT (WHERE prodName = o.prodName);
```

Listing 12: Four equivalent queries to find orders with more revenue than average for their product

Observe that queries 3 and 4 have very similar structure. This is because the OVER operator (window aggregation) and AT operator (measures) have the same function: to evaluating a calculation over a collection of rows meeting some criterion. AT is more powerful than OVER; it can evaluate arbitrary predicates where OVER's PARTITIONBY can evaluate only = predicates; and it can query rows that have been removed by a WHERE clause.

Why is the WinMagic rewrite beneficial? Observe that Orders appears twice in queries 1 and 2 but only once in 3 and 4. This suggests to the optimizer an execution strategy that you might call 'localized self-join'. The engine scans order records grouped by product; when it has finished a product, and knows the average revenue of that product, it rewinds to the beginning of the product and emits orders whose revenue is greater than the average.

This strategy, of small loops probing into intermediate results cached in memory, is characteristic of in-memory OLAP engines. We believe it is worth investigating whether this strategy is also beneficial in SQL engines.

Aside from the runtime benefits, the queries with less repetition are easier to optimize, because optimizers have difficulty identifying common sub-trees in relational algebra.

5.2 Hierarchies

We chose not to explicitly support hierarchies. Hierarchies are a major part of dimensional systems, but they complicate the language and are largely used for user interface concerns (for example, suggesting fields to drill down on). For our purposes, it is sufficient to be able to treat any expression on a dimension (for example, YEAR(orderDate)) as an *ad hoc* dimension.

That said, when I set the year dimension, I should not have to explicitly clear the month dimension. In order to achieve that effect, we hope (in a future version of this language) to allow dimensions to be 'linked' for purposes of their ALL and SET behavior.

5.3 Wide tables

Business Intelligence tools typically have a 'cube' or 'business view' concept that contains measures from a fact table and columns from several dimension tables. This is attractive to end-users because they do not need to specify joins. Without measures, 'wide tables' composed as join views were not advisable because denormalization would introduce inconsistencies such as double-counting. But with measures, calculations maintain their own consistency, and wide tables are a recommended practice.

Wide tables can also contain measures with complex behaviors:

- A **semi-additive measure** rolls up using different aggregate functions on different dimensions but can sometimes be summed; for example, an *items on hand* (inventory) measure rolls up using LAST_VALUE on the time dimension and SUM on other dimensions;
- A **non-additive measure** never aggregates by summing, typically a calculation based on other measures; for example, *return rate* is the ratio of product units sold to product units returned.
- Other custom measures might use a different formula for different levels of a hierarchy; for example, the *revenue* measure might have a different formula at a business unit level than at a country level. The SQL GROUPING_ID function can be used to identify the level.

5.4 Composability

Measures are composable in several ways.

First, as we have mentioned, the query language is closed. A query can reference tables with (or without) measures, and returns a table with (or without) measures. Queries can therefore be nested to arbitrary depth, as in regular SQL. Views with measures can be created upon relations (such as a traditional relational database, or a directory of CSV files) that do not have measures.

Second, measures can reference measures in the same query. Measures defined using the AS MEASURE syntax can reference by name other measures defined in the same SELECT.

(We do not, in the current language, allow recursive or mutually recursive measures. We believe that they are useful, but there are implementation hurdles. Termination is one concern, although spreadsheet formulas manage perfectly well without provable termination. A greater concern is that recursive measures cannot be implemented using a static rewrite, and will require some form of unbounded state, such as a call stack.)

Third, a measure can reference a measure or measures from an input table, and thus a measure seems to be propagated effortlessly through a stack of nested queries. But the semantics are defined one step at a time. Each query is evaluating a context-sensitive expression, in its own evaluation context, and defining a new measure whose dimensionality is determined by the columns that it projects, and it is that measure that is consumed by its enclosing query.

5.5 Security

SQL’s security model is simple and robust: if I own tables that contain sensitive information, I can write a query that accesses those tables, publish that query as a view, and grant access to the view but not the underlying tables. People can access the data that I allow them to, and the optimizer will ensure that those queries have efficient plans.

Do views with measures offer the same robust security model? The answer is yes. This may be surprising, given that measures return much richer values than regular columns, so let’s justify that assertion.

A regular SQL view, without measures, returns a fixed amount of information; this is easy to see because if I replace the view with a base table with the same contents, every possible query will return the same results.

Now consider a view that has regular columns a and b , hidden columns c and d that are not projected by the view, and measures m and n . Queries that only use a and b are straightforward; they map to the relational core. But what of queries that also use the measures? They too are bounded. Each measure does not return a single value, of course, but it returns a map that can be read by providing a predicate. If I ask for the value of measure m with the predicate $a = 0$ and $b < 10$, it returns 6; if I ask for the value of m with the predicate $a = 1$, it returns 12, and so forth.

Furthermore, the predicate can only be in terms of the dimension columns a and b , not in terms of the hidden columns c and d . If two rows in the underlying table(s) cannot be distinguished based on their a and b values, then I cannot construct a predicate to separate them.

To use an analogy, if regular column values are like pixels of a discrete image, then measures are like holograms; their data has more dimensions, but is still finite.

A view with measures thus allows me to create an interface that limits which questions can be asked of the underlying data.

5.6 Looker’s Open SQL Interface

Looker[8] is a BI platform that was acquired by Google in 2019 and is now part of Google Cloud. Using Looker’s LookML™ language, analysts define objects called “Explores”, which are a form of the wide tables described in subsection 5.3. These are the starting point for data exploration via pivot tables, charts, and dashboards.

Looker also serves as a semantic layer for third-party visualization tools such as Google Sheets, Microsoft Power BI, Tableau, and ThoughtSpot. Those tools query the Explores, benefiting from the joins, measures and other calculations, and presentation and navigation information encapsulated in them. Organizations choose to use a semantic layer so that Explores are defined just once, in one place, as opposed to many redundant and inconsistent definitions in the visualization layer.

In Looker’s Open SQL Interface[7], each Looker Explore appears as a SQL table, the measures in that Explore appear as measure columns, and the dimensions in that Explore appear as regular columns. The SQL Interface accepts SQL queries that adhere to GoogleSQL syntax, and supports most of the BigQuery operators.

Before the SQL Interface was introduced, building a connector from a third-party tool was complicated, because expressions in the tool’s expression language had to be translated into Looker’s expression language. Connectors built using the SQL Interface are much simpler, and are similar to the tools’ existing connectors to conventional SQL databases. When generating SQL, tools can use measures defined in Looker (e.g. `AGGREGATE(profitMargin)`) or can define their own measures using aggregate functions on top of regular columns (e.g. `SUM(revenue)`).

The implementation uses Apache Calcite’s SQL parser, query planner, and SQL function library.

5.7 Natural Language to SQL

For applications such as natural-language-to-query translation, including those powered by Large Language Models (LLMs) and Generative AI, SQL-with-measures is an attractive target language, for three reasons.

First, it manages complexity. Like humans, generative AI has difficulty correctly generating large expressions, especially when consistency is required between regions of those expressions that are widely separated. If the target language is regular SQL, the generated queries are large, deeply nested, and have many joins, including complex self-joins. In SQL-with-measures, the joins and calculations can be encapsulated in a view, and context-sensitive expressions eliminate the need for self-joins, and therefore the generated query is more concise and less complex.

Second, current query-generation systems use a multidimensional semantic layer — for example, Analyza [6] uses a catalog containing “additional information about the type of the column (e.g. is it a metric, dimension, etc.), data formats (e.g. should the number be formatted as a dollar amount), and date range defaults” — and measures allow us to encapsulate that semantic layer as SQL views.

Last, the corpus of queries in SQL is larger than in any other query language, and therefore training LLMs is much easier.

Early indications are that the generated queries are smaller and more accurate — and easier to understand. More research is needed in this area.

6 FUTURE WORK

6.1 Formal semantics

In this paper, we have presented an informal semantics. It would be useful if a future publication described a formal semantics for

measures, context-sensitive expressions and the evaluation context. Perhaps these could be extensions to relational algebra.

The semantics of the AT operator should be clarified. It seems reasonable to allow expressions within SET, for example `profitMargin AT (SET YEAR(orderDate) = CURRENT YEAR(orderDate) - 1)`.

The CURRENT operator should return a valid value if the evaluation context implies a single value for all possible rows; for example, if the query has `GROUP BY FLOOR(orderDate TO MONTH)` all rows in a given group will have the same month and therefore the same year. To allow the SQL semantic analyzer to safely make that deduction, we need new rules for deducing functional dependencies among expressions, perhaps a notion similar to Calcite's nested time frames [12].

6.2 Generating queries from natural language

As mentioned in subsection 5.7, research should ascertain whether SQL-with-measures is an effective target language for AI-powered query generation.

6.3 Operators for managing grain

Measures have the useful property that they preserve grain in the presence of joins (preventing double-counting), but we need more operators for managing grain.

For example, an *items on hand* semi-additive measure might take the count of each product on the warehouse shelf on the last day of the time period, and then sum over all products and warehouses. A *rank change* non-additive measure might rank each product by revenue in a given region and time period, and then compute the difference with the rank in the previous time period. Such measures perform multiple aggregation steps, each step using a different aggregate function and occurring in a particular order.

A promising candidate is the PER clause for aggregate functions, proposed as a generalization of Calcite's WITHIN DISTINCT clause [11].

6.4 Implementation strategies

Strategies to implement queries with measures and context-sensitive expressions require further study.

One strategy is to rewrite queries in terms of simpler operations. Our algorithm in subsection 4.2, which rewrites a measure reference as a correlated scalar subquery, is general-purpose but not very efficient. In simple cases (such as a query with GROUP BY and no JOIN) it may be valid to inline the measure definition. In cases with joins, a WITHIN DISTINCT clause may be introduced to preserve the measure's grain. The correspondences noted in subsection 5.1 suggests that some queries can be rewritten to window aggregates, especially if window aggregates are generalized to computations to access "lost" rows.

As we remarked earlier, recursive measures cannot be solved using a static rewrite, and may require a new physical algorithm. That algorithm may also be applicable to other cases.

6.5 Forecasts and time series

Forecasting and time series analysis are similar domains. Time series analysis often involves interpolation, such as changing the

temporal grain of a measure (resampling) to match other measures, or to fill gaps where no measurement is available; forecasting generally extrapolates, creating estimates of a measure in the future based on past values of that measure and related measures.

Both make extensive use of statistical techniques; for example, autoregressive integrated moving average (ARIMA) can detect and compensate for periodicity. Measures can simplify things for users: an expert defines the calculations, encapsulates them in a model (view) as measures, and the user can use the model without worrying about the complexity.

A challenge to be solved is that both techniques create new values for dimensions (for example, recording a revenue of zero on a holiday, when the business is closed, or generating a revenue forecast for a future year, for which there are not yet any orders). We will need to devise a query syntax for synthesizing rows. At the same time, we can answer the important question, "How can I evaluate a measure on a table that has no rows?"

6.6 Log files and sequential processing

Much modern data processing, especially during load and transformation phases, takes place on log files that have a nested structure. Records are processed in sequence, often in a single pass, but with a processing context that includes the current record, sibling records that occur within the same parent (such as the group of records for the same browser session), the parent record, and perhaps other data values computed from various "ancestor" records. Measures might allow such calculations to be expressed declaratively.

On the related topic of sequence data, measures may be helpful in organizing the complex rules for identifying logical business events as part of the data model. Their relationship with SQL's existing MATCH_RECOGNIZE clause [21] should be investigated.

7 SUMMARY

Measures are a natural extension to the relational data model. They allow calculations, including aggregate functions, to be encapsulated in the definition of a table. These calculations offer context-dependent views of the table; not a single static image but one that varies based on the viewer, like a hologram.

The evaluation context of a measure is established in its definition and can optionally be adjusted when it is used, by making changes to just the expression that invokes the measure. This locality of reference allows queries to be written concisely, allows queries to be composed reliably, and brings modularity to relational systems using SQL.

Recent explorations with LLMs remind us how challenging were those non-local transformations that we previously required of human SQL authors. Measures make these repetitive filters and self-joins invisible, and we hope that they improve the lot of humans and machines alike.

8 ACKNOWLEDGMENTS

This work would not have been possible without many design discussions and much patient, constructive feedback. The authors would like to thank their Google colleagues Adam Wilson, Alexey Leonov-Vendrovskiy, Bengu Li, David Wilhite, Goetz Graefe, Jeff

Shute, Lloyd Tabb, Marieke Gueye, Matthew Brown, Mosha Pasumansky, Riccardo Muti, Romit Kudtarkar, and Serhiy Tykhan-sky.

REFERENCES

- [1] William D Back, Nicholas Goodman, and Julian Hyde. 2013. *Mondrian in Action: Open source business analytics*. Manning Publications Company.
- [2] Edmon Begoli, Jesús Camacho-Rodríguez, Julian Hyde, Michael J Mior, and Daniel Lemire. 2018. Apache Calcite: A Foundational Framework for Optimized Query Processing Over Heterogeneous Data Sources. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, 221–230. <https://doi.org/10.1145/3183713.3190662>
- [3] E. F. Codd. 1970. A relational model of data for large shared data banks. *Commun. ACM* 13, 6 (jun 1970), 377–387. <https://doi.org/10.1145/362384.362685>
- [4] Edgar F Codd. 1993. Beyond decision support. *Computerworld* (1993).
- [5] George Colliat. 1996. OLAP, relational, and multidimensional database systems. *ACM Sigmod Record* 25, 3 (1996), 64–69.
- [6] Kedar Dhamdhere, Kevin S. McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring Data with Conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) (IUI '17). Association for Computing Machinery, New York, NY, USA, 493–504. <https://doi.org/10.1145/3025171.3025227>
- [7] Google. 2023. Looker Open SQL interface. <https://cloud.google.com/looker/docs/sql-interface>. [Online; accessed 01-Apr-2024].
- [8] Google. 2024. Looker. <https://cloud.google.com/looker>. [Online; accessed 12-Apr-2024].
- [9] Jim Gray, Surajit Chaudhuri, Adam Bosworth, Andrew Layman, Don Reichart, Murali Venkatrao, Frank Pellow, and Hamid Pirahesh. 1997. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data mining and knowledge discovery* 1 (1997), 29–53.
- [10] John Horner, Il-Yeol Song, and Peter P. Chen. 2004. An analysis of additivity in OLAP systems. In *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP* (Washington, DC, USA) (DOLAP '04). Association for Computing Machinery, New York, NY, USA, 83–91. <https://doi.org/10.1145/1031763.1031779>
- [11] Julian Hyde. 2021. *WITHIN DISTINCT clause for aggregate functions*. Feature request CALCITE-4483. Apache Calcite. <https://issues.apache.org/jira/browse/CALCITE-4483>
- [12] Julian Hyde. 2022. *Custom time frames*. Feature request CALCITE-5155. Apache Calcite. <https://issues.apache.org/jira/browse/CALCITE-5155>
- [13] JSR-69 2003. *Java™ OLAP Interface (JOLAP), final draft*. Technical Report. JSR-69 Expert Group. <https://jcp.org/aboutJava/communityprocess/first/jsr069/index.html>
- [14] Ralph Kimball and Margy Ross. 2002. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (2nd ed.). John Wiley & Sons, Inc., USA.
- [15] Bart Kuijpers and Alejandro Vaisman. 2017. An algebra for OLAP. *Intelligent Data Analysis* 21, 5 (2017), 1267–1300.
- [16] MDAPI-2.0 1998. *MDAPI™ the OLAP Application Program Interface Version 2.0*. Technical Report. The OLAP Council.
- [17] Konstantinos Morfonios, Stratis Konakas, Yannis Ioannidis, and Nikolaos Kotsis. 2007. ROLAP implementations of the data cube. *ACM Comput. Surv.* 39, 4 (nov 2007), 12–es. <https://doi.org/10.1145/1287620.1287623>
- [18] Oscar Romero and Alberto Abelló. 2007. On the Need of a Reference Algebra for OLAP. In *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 99–110.
- [19] Mark Whitehorn, Robert Zare, and Mosha Pasumansky. 2004. Fast Track to MDX. <https://api.semanticscholar.org/CorpusID:61077971>
- [20] Fred Zemke, Krishna Kulkarni, Andy Witkowski, and Bob Lyle. 1999. Introduction to OLAP functions. *Change proposal. ANS-NCTS H2-99-14 (April)* (1999).
- [21] Fred Zemke, Andrew Witkowski, Mitch Cherniak, and Latha Colby. 2007. *Pattern matching in sequences of rows*. Change proposal for ISO 9075-1. ANSI INCITS.
- [22] Calisto Zuzarte, Hamid Pirahesh, Wenbin Ma, Qi Cheng, Linqi Liu, and Kwai Wong. 2003. WinMagic: Subquery elimination using window aggregation. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, California, USA, June 9-12, 2003*. ACM, 652–656. <https://doi.org/10.1145/872757.872840>