Temporal Data in SQL Server

Dejan Sarka
SAY THANK YOU TO OUR SPONSORS!
Sponsor Raffle!!!

• Each sponsor stamp will opt you into their raffle prize and mailings
• Collect 9+ sponsor stamps on your Badge to be eligible for the Xbox
• Hand entire Badge/ribbon back into registration desk at end of day
• We will draw Badges for prizes at 5pm in Cromwell (if you are drawn and do not have the pre-requisite stamp/s..... You lose!)
Social

• Make sure you tweet on #spscambridge or #sqlsatcambridge
• During the event we have Giant Jenga, Sack races and Conker Fights!
• After event, join us for a post event SharePint/ SQLPint from our bar
• Don’t forget to thank Sponsors, Volunteers and Speakers!
• The event will close at 6.30pm
Introduction

- Dejan Sarka
  - dsarka@solidq.com, dsarka@siol.net, @DejanSarka
- MCT, Data Platform MVP
- 30 years of data modeling, data mining and data quality
- 15 books, writing more
- 15+ courses
Agenda

- Background and references
- Temporal databases
  - Allen’s interval operators
- Optimizing problems
- SQL Server temporal tables
  - Creating
  - Querying
  - Altering and dropping
- Temporal tables as DW dimensions
Theoretical Background


SQL Server Solutions


- Dejan Sarka: Various methods for optimizing temporal queries, Pluralsight course *Working with Temporal Data in SQL Server* (http://goo.gl/XdeRbl)

Temporal Databases

- Many relational databases show current state only
  - Constraints are valid in present, past and future time
  - E.g., supplier A is under a contract
- Temporal databases hold *time-stamped* propositions
  - *Since* predicate means ever since and not immediately before
  - E.g., supplier A is under a contract since a specific time point A
  - *During* (or *from* ... *to*) predicate means throughout and not immediately before or immediately after
  - E.g., supplier A is under a contract from time point A to time point B
- *Timeline* consists of discrete, indivisible time quanta
  - *Time quantum* is the smallest represented unit
  - Can represent time quanta with integers
Semi-Temporal Problems

- Example: suppliers, supplied parts

<table>
<thead>
<tr>
<th>S#</th>
<th>Since</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>d04</td>
</tr>
<tr>
<td>S2</td>
<td>d07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S#</th>
<th>P#</th>
<th>Since</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>P1</td>
<td>d04</td>
</tr>
<tr>
<td>S1</td>
<td>P2</td>
<td>d05</td>
</tr>
<tr>
<td>S2</td>
<td>P1</td>
<td>d08</td>
</tr>
<tr>
<td>S2</td>
<td>P2</td>
<td>d09</td>
</tr>
</tbody>
</table>

- No problem with PKs and FKs
- Need additional constraint: no supplier can supply a part before the supplier is under contract
## Full-Temporal Problems (1)

<table>
<thead>
<tr>
<th>S#</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>d04</td>
<td>d10</td>
</tr>
<tr>
<td>S2</td>
<td>d02</td>
<td>d04</td>
</tr>
<tr>
<td>S2</td>
<td>d07</td>
<td>d10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S#</th>
<th>P#</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>P1</td>
<td>d04</td>
<td>d10</td>
</tr>
<tr>
<td>S1</td>
<td>P2</td>
<td>d05</td>
<td>d09</td>
</tr>
<tr>
<td>S2</td>
<td>P1</td>
<td>d08</td>
<td>d10</td>
</tr>
<tr>
<td>S2</td>
<td>P2</td>
<td>d09</td>
<td>d10</td>
</tr>
<tr>
<td>S2</td>
<td>P2</td>
<td>d03</td>
<td>d03</td>
</tr>
</tbody>
</table>

- Two candidate keys in each table (BCNF)
- *To* should not be less than *From*
- Cannot end contract on one day and begin on the next day
- No abutting
Full-Temporal Problems (2)

- No supplier can be under two distinct contracts on the same day
  - No overlapping
- No supplies if not under contract
  - \( SP(S\#, \text{From}) \) is not a FK!
  - Inclusion dependency:

\[
SP(S\#, \text{From,To}) \subseteq S(S\#, \text{From,To})
\]
Valid and Transaction Times

- Bi-temporal databases support valid and transaction times
  - Valid times are human times
  - Transaction times are database times
  - Valid times are updatable
  - Transaction times are not updatable

- In model, we should not need transaction times
  - Should be implemented in a RDBMS
  - Support for valid times would also be appreciated
<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equals</td>
<td>((i_1 = i_2))</td>
<td>((b_1 = b_2) \text{ AND } (e_1 = e_2))</td>
</tr>
<tr>
<td>Before</td>
<td>((i_1 \text{ before } i_2))</td>
<td>((e_1 &lt; b_2))</td>
</tr>
<tr>
<td>After</td>
<td>((i_1 \text{ after } i_2))</td>
<td>((i_2 \text{ before } i_1))</td>
</tr>
<tr>
<td>Includes</td>
<td>((i_1 \supseteq i_2))</td>
<td>((b_1 \leq b_2) \text{ AND } (e_1 \geq e_2))</td>
</tr>
<tr>
<td>Properly includes</td>
<td>((i_1 \supset i_2))</td>
<td>((i_1 \supseteq i_2) \text{ AND } (i_1 \neq i_2))</td>
</tr>
<tr>
<td>Meets</td>
<td>((i_1 \text{ meets } i_2))</td>
<td>((b_2 = e_1 + 1) \text{ OR } (b_1 = e_2 + 1))</td>
</tr>
<tr>
<td>Overlaps</td>
<td>((i_1 \text{ overlaps } i_2))</td>
<td>((b_1 \leq e_2) \text{ AND } (b_2 \leq e_1))</td>
</tr>
<tr>
<td>Merges</td>
<td>((i_1 \text{ merges } i_2))</td>
<td>((i_1 \text{ overlaps } i_2) \text{ OR } (i_1 \text{ meets } i_2))</td>
</tr>
<tr>
<td>Begins</td>
<td>((i_1 \text{ begins } i_2))</td>
<td>((b_1 = b_2) \text{ AND } (e_1 \leq e_2))</td>
</tr>
<tr>
<td>Ends</td>
<td>((i_1 \text{ ends } i_2))</td>
<td>((e_1 = e_2) \text{ AND } (b_1 \geq b_2))</td>
</tr>
</tbody>
</table>
### Allen’s Interval Algebra Operators

<table>
<thead>
<tr>
<th>Name</th>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union</td>
<td>((i_1 \text{ union } i_2))</td>
<td>((\text{Min}(b_1, b_2) : \text{Max}(e_1, e_2)), \text{ when } (i_1 \text{ merges } i_2); \text{ NULL otherwise})</td>
</tr>
<tr>
<td>Intersect</td>
<td>((i_1 \text{ intersect } i_2))</td>
<td>((\text{Max}(b_1, b_2) : \text{Min}(e_1, e_2)), \text{ when } (i_1 \text{ overlaps } i_2); \text{ NULL otherwise})</td>
</tr>
<tr>
<td>Minus</td>
<td>((i_1 \text{ minus } i_2))</td>
<td>((b_1 : \text{Min}(b_2 - 1, e_1)), \text{ when } (b_1 &lt; b_2) \text{ AND } (e_1 \leq e_2); \text{ NULL otherwise})</td>
</tr>
</tbody>
</table>

- | \(\text{Max}(e_2 + 1, b_1) : e_1\), when \((b_1 \geq b_2) \text{ AND } (e_1 > e_2); \text{ NULL otherwise}\) |
Allen’s Interval Algebra Operators

Union

Interval 1
Interval 2
Result
Interval 1
Interval 2
Result
Interval 1
Interval 2
Result
Interval 2

Minus

Interval 1
Interval 2
Result
Interval 1
Interval 2
Result
Interval 1
Interval 2
Result
Interval 2

Intersect

Interval 1
Interval 2
Result
Interval 1
Interval 2
Result
Interval 1
Interval 2
Result
Interval 2
NULL
Classical T-SQL Solution

- Intervals represented with two columns – begin and end \([b, e]\)
- Representation with integers
- Find all intervals \([b, e]\) that overlap with a given interval \([@b, @e]\)
- Classic predicate: WHERE \(b \leq @e \land @b \leq e\)

\[(i_1 \leftarrow \text{overlaps} \rightarrow i_2) \equiv (b_1 \leq e_2 \land b_2 \leq e_1)\]
Optimizing Problem

- Optimal indexes exist
  - IDX1 – key b, included e
  - IDX2 – key e, included b

- Problem: two range predicates (WHERE \( b \leq @e \) AND \( @b \leq e \))
  - Only one index is used
  - E.g., use IDX1, seek for b, and then use e as the residual predicate to filter while scanning the rows after the seek
  - Such a seek is very efficient if \( @e \) is high
  - Otherwise, IDX2 could be used
  - Indexes are efficient only if they eliminate most of the rows before the scan of the residual rows – if the \([@b, @e]\) is close to the beginning or to the end of the timeline
  - Inefficient around the middle of the timeline
  - Must prevent parameter sniffing
Enhanced T-SQL Solution (1)

- A simple idea: try to use a single index to eliminate non-matching intervals from both sides
- Eliminating using the [lower] index
  - Eliminating intervals from the right side: just eliminate all intervals where the beginning is at least one unit bigger (more to the right) of the end of the given interval
  - In order to use the same index for eliminating from the left, you need to use the beginning of the intervals in the table in the WHERE clause of the query
  - Go to the left side away from the beginning of the given interval at least for the length of the longest interval in the table
  - The intervals that begin before the beginning of the given interval minus the length of the longest interval cannot overlap with the given interval
Enhanced T-SQL Solution (2)
Enhanced T-SQL Solution (3)

- Can get extremely efficient queries
- The solution is simple
- However, the performance drops substantially as soon as you have one very long interval in your table
- Therefore, the solution efficiency depends on the interval length distribution
  - The more uniform distribution, the more efficient solution
SQL Server Temporal Tables

- System-versioned tables store a full history of data changes
  - A validity period, is stored in two datetime2 columns – SysStartTime and SysEndTime
- All versioning is automatic, no change in the code needed
  - Can be added to existing tables
- History is stored in an associated table
  - Historical table can be named, or take system name
Considerations and Limitations (1)

- The SysStartTime and SysEndTime columns must use the datetime2 data type.
- The table must have a primary key; the historical table cannot use constraints.
- To name the history table, specify the schema and table names.
- History table compression defaults to PAGE.
- The history table must reside in the same database as the current table.
Considerations and Limitations (2)

- System-versioned tables are not compatible with FILETABLE or FILESTREAM
- Columns with a BLOB data type, such as varchar(max) or image, can result in high storage requirements
- INSERT and UPDATE statements cannot reference the SysStartTime or SysEndTime columns
- You cannot directly modify data in the history table
- You cannot truncate a system-versioned table
- Merge replication is not supported
Creating a Temporal Table

```sql
CREATE TABLE dbo.Employee
(
    EmployeeID int NOT NULL PRIMARY KEY CLUSTERED,
    ManagerID int NULL,
    FirstName varchar(50) NOT NULL,
    LastName varchar(50) NOT NULL,
    SysStartTime datetime2
        GENERATED ALWAYS AS ROW START NOT NULL,
    SysEndTime datetime2
        GENERATED ALWAYS AS ROW END NOT NULL,
    PERIOD FOR SYSTEM_TIME (SysStartTime, SysEndTime)
)
WITH (SYSTEM_VERSIONING =
    ON (HISTORY_TABLE = dbo.EmployeeHistory));
```
Querying Temporal Tables

- System-versioned tables can be queried using the new FOR SYSTEM_TIME clause in the FROM part of a query
- Use one of the following four sub-clauses:
  - AS OF <date_time>
  - FROM <start_date_time> TO <end_date_time>
  - BETWEEN <start_date_time> AND <end_date_time>
  - CONTAINED IN (<start_date_time>, <end_date_time>)
  - ALL to return everything
Altering Temporal Tables

- When you use ALTER TABLE to add a new nullable column, change data type or remove an existing column, system will automatically perform the action against both current and history tables.

- When you add a non-nullable column, you need to provide a default constraint as well.
  - You should take in account that this operation is an offline operation in all editions of SQL Server except in the Enterprise Edition.

- You can also use ALTER TABLE statement to add or remove the attribute HIDDEN to period columns or to remove it.
Altering Temporal Tables Limitations

- Adding an IDENTITY or computed column
- Adding a ROWGUIDCOL column or changing an existing column to it
- Adding a SPARSE column or changing an existing column to it, when history table is compressed
- If you need to perform schema changes to a temporal table not supported in the ALTER statement, you have to set its SYSTEM_VERSIONING attribute to false, to convert the tables to non-temporal tables, perform the changes and then convert back to temporal table
Dropping Temporal Tables

- You cannot drop a system-versioned temporal table
  - Both current and history tables are protected until the attribute SYSTEM_VERSIONING of the current table is set to ON
  - When you set it to OFF, both tables automatically become non-temporal tables and are fully independent of each other
  - Therefore, you can perform all operations against them that are allowed to you according to your permissions
- You can also drop period if you want to convert a temporal table to non-temporal one
Temporal Tables as Dimensions

- System-versioned tables can be out of the box solution for Type 2 SCD problem
  - Can simplify the ETL process
- However, SSAS does not support temporal tables yet
  - You are limited to query the data warehouse directly, with T-SQL queries
- System versioning works on row level
  - Might need column-level versioning
Especially for you...

- Only for attendees: 40% discount on e-book!
  - Discount Code: **ACBUBLV40**
  - Start date: September 9\textsuperscript{th} 2017
  - Expiry date: September 30\textsuperscript{th} 2017
  - Create a login on the Packt site [www.packtpub.com](http://www.packtpub.com) and add the book to the cart
  - Click “View Cart”
  - “Do you have a promo code?” field enter (code provided above)
  - Click the “Apply” button to apply the discount
Wait! There’s more!

- Only for attendees: 50% discount on e-book!
  - Discount Code: AUSQLSSC50
  - Start date: September 9th 2017
  - Expiry date: September 30th 2017

- Questions?
- Visit SQL Saturday Slovenia, Ljubljana, December 9th!