Constrained Data Base with Slow Streams

Stream Reasoning can be described as Reasoning Upon Rapidly Changing Information[1https://streamreasoning.org/][E. Della Valle, & F. Heintz, 2018]

Linear temporal logic or linear-time temporal logic[1][2] (LTL) is a modal temporal logic with modalities referring to time. In **LTL**, one can encode formulae about the future of paths, e.g., a condition will eventually be true, a condition will be true until another fact becomes true, etc. [2https://en.wikipedia.org/wiki/Linear_temporal_logic]

- We will not revise extensively neither Stream Reasoning nor LTL here, please consider reading more about to further understand why **Constrained Data Base with Slow Streams** is different.

Stream Reasoning tries to put constraints on streams in *real time*, over a window of time (memory) so as to decide whether arriving data fits a logical rule or not. Stream Reasoning applies reason on arriving data and past data with a window memory.

LTL can be an expression language of Stream Reasoning. [3https://www.diva-portal.org/smash/get/diva2:1043981/FULLTEXT02.pdf]

Context

- If we think of a database with a stream data source, the problem of data consistency can be present. Generally, even if a database business and functional specifications consists of hundreds or thousands of logical rules (ex: table1.column1 > table1.column2 …), generally we only care about already arrived data, but not future arriving data, as long as data consistency is maintained in the database, we can ignore rejects, log them, or historize them, but we have no control over neither:
    - **"anticipated" future rule changes nor**
    - **data changes in the future, even if we have knowledge of future changes.**

  Note: these two cases are common with real time applications (databases also) with multiple data sources and frequent changes from data providers.

- Data consistency in *real time databases* can be achieved with SQL TRANSACTION command, or beforehand in the ETL step (using other programming languages).
- Classical constrained data intends to control changes over domain definition of data
- **Constrained Data-Base with "Slow Streams" intends to control by anticipation changes of data schema (changing format again) in the future and changes of business rules in the future.**
- **Specifying a framework to force by imposed structures and rules acting on stream data using Linear Time Logic combined with Constraint Logic Programming can make applications in these context more robust to change.**

⇨ **When to consider Constrained Data-Base with Slow Streams**

In the case, an application has not full control of

- Schema of its sourcing feeds
- Definition domain (values of columns)
- Changes are slow (humanly predictable); In contrast with constant unpredictable mutations of data feeds by and from another system.

If we are more concerned with a "consistent database", rather than a "consistent stream" (time window of arriving data); Also if we have at least a minor vision on the future arriving data (we will talk about this later) we can use:

- **LTL language**
- **Constraint Logic Programming language**
  [SQL: https://arxiv.org/pdf/1907.10914.pdf ,
  OCL: https://en.wikipedia.org/wiki/Object_Constraint_Language]
- **Event-Condition-Action**

All three In order to :

- Expect new changes of data definition in the future.
- Specify in advance known changes of rules in the future.
- Historize events and actions.

⇨ **Why Slow Streams**

Talking about Streams in academia, we find generally all subjects address performance problems, solutions to rapid (changing or not) of input data to a system. We find reactive programming for example based on asynchronous calls and decoupling of functionalities of a system, all benefiting parallel computing. We are not addressing this issue, slow streams here emphasize that is not the subject. Rather we focus on maintaining a system robust on rapid changes of business specifications

Normally data does not change at the same rate of its arrival speed, otherwise It would be complete random and non-useful data; Except when we are talking about unstructured data that is meant to be so to some extent. To have a consistence database with changing sourcing data with some knowledge of upcoming new consistency rules or changes. Using **LTL, Constraint Logic Programming**, and an **event-condition-action** language, we can express near future "known changes" by consistency rules. With the described pattern and definitions, we hopefully could have a consistent database by pushing and changing rules in a defined **LTL-CONSTRAINT-EVENT-ACTION** layer in contrast with traditional ETL/Databases solutions.

Examples

A run can be described as a set of rules and actions.

Actions can be BACK-UP, IGNORE, HISTORISE, …

Rules can target columns in different tables.

Current Runs rules can be mixed.

```
run1 : @cond WHEN tableA.ID == tableB.ID AND tableA.column1 > tableB.column2
        BACK-UP
       FLAG tableA.rule1
AFTER run1 : LOG ('WARN')

run2 : @cond WHEN tableA.column1 > 0
        DO NOT BACK-UP
       FLAG tableA.rule2
AFTER run2 : LOG ('ERROR')

run3 : @cond WHEN exists(tableC.column1)
        BACK-UP tableC.column1
       FLAG tableA.rule3
AFTER run3 : LOG ('ERROR')
```

ALWAYSE (run2 , run1) UNTIL run3@cond

Can omit @cond and expect “run” as a condition when it is an operand in a LTL expression.

LTL-C (LTL-Constraint) variant

As described above, LTL is meant to express a *fact*, however the expression highlighted in green, is different, because it is bound to the *event-action* concept, it is meant to impose itself as true in the system, by applying predicate tests and actions, The system *obeys* the LTL formulae choosing from possible actions so as to satisfy the abstract LTL formulae *as soon as possible*. In other words, the system is always searching for *a sub-domain* that satisfies the *LTL formulae*. It is a form of constraint optimization paradigm expressed over LTL.

Analogy to Predicate logic:

First-order logic uses quantified variables over non-logical objects like *5+5 == 10*

LTL-C is a higher order logic where constraints expressed in a constraint language are the propositional variables.

Example

ALWAYSE (run2 , run1) UNTIL run3@cond run3

In this example: run1, run2, and run3 are all self-conditioned runs, they are execution only when there conditions are met. For instance: run1 could be run only if run1@cond is satisfied.

The query means the system always run the conditioned runs run2 and run1 while checking also every time whether run3@cond is satisfied, if it is the case, then run3 is executed instead of run1 and run2.

Discussion

Read more about the idea (in its early stage) in my posts on the internet:

https://stackoverflow.com/q/58199623/1951298

In the solution, I talked about TDDA which is an open source library which only describes constraints on columns in the same table at that time. At that time, I thought it would be a sufficient solution as I did not picture the abstraction described above.

The second limit, is that constraints are expressed only over arriving data and cannot be extended to future changes like LTL would do.

LTL-C Memory engine (not yet specified)

The proposed LTL-EVENT-ACTION solution will need memory. It caches run “runs” to be able to handle future events. (Like “until”, “sometimes”, “eventually” ..).