A Mathematical Database to Process Time Series

Cyrille Ponchateau, LIAS/ENSMA, Futuroscope Chasseneuil, France
Ladjel Bellatreche, LIAS/ENSMA, Futuroscope Chasseneuil, France
Carlos Ordonez, University of Houston, Houston, USA
Mickael Baron, LIAS/ENSMA, Futuroscope Chasseneuil, France

ABSTRACT

In scientific research, the results of an experiment commonly take the form of a time series, in which such time series consists of measurements collected from a sensor over time. After time series are stored, mathematical models are derived using numerical methods. Even though there exist plenty of tools to store and analyze time series data, there is scarce research aimed at storing and querying derived models, which are the most important mechanism for a scientist to understand data. In this article, the authors propose to help scientists with a flexible database structure to persist and manage mathematical models with a mathematical models store, with extended features, to handle time series. In this article, the authors introduce the concept of a mathematical models store enriched with numerical processing methods to allow queries based on raw time series data. Then they introduce a prototype, that is an implementation of such a data store with PostgreSQL.

KEYWORDS

Binary Parser Tree, Comparison Series-Equations, Differential Equations Store, ER Schema, Mathematical Models Database, Microservices, Numerical Processing

INTRODUCTION

The evolution of computer technology in terms of computing and storage capacity, now allows the processing of large data sets containing complex numerical data, a typical task for mathematical software. Such data sets are commonly found in experimental sciences (Shatay & Zdonik, 1996), from medicine (medical imaging (Bagnall, Ratanamahatana, Keogh, Lonardi, & Janacek, 2006), electrocardiogram (Lang, Morse, & Patel, 2010; Fu, 2011) and medical surveillance (Esling & Agón, 2012)) to physics (particle tracking (Lang et al., 2010)). But also in other domains such as finance (weekly sales total (Fu, 2011), stock price movements (Lee, Kwon, & Lee, 2003)).

This work, in particular, was initiated by an automatic control team. In their domain (as in other experimental sciences domains), the researchers design experiments to study physical systems. For instance, when a voltage is applied at the terminals of an electrical motor, it starts rotating and the rotating speed depends on the aforementioned voltage. Then, the aim of the automatic control

DOI: 10.4018/IJDWM.2018070101
researcher is to mathematically describe the dependence between voltage and rotating speed. Thereafter, the motor is tested with different voltages and the rotating speed evolution is measured by a sensor. The measurements are taken at a fixed rate, generating a series of chronological values, represented as a time series. The next step consists in analyzing the time series, using numerical software such as Matlab, Octave, R, and so on. Finally, the analysis provides a mathematical model (a differential equation), describing how the motor rotating speed behaves according to the voltage variations.

From a practical standpoint, when models are produced, in general, researchers have no standard structure or format to store them. Therefore, they end up being stored in different formats in many files in a disorganized manner. For instance, models are sometimes embedded in Matlab, R or Python scripts. Moreover, they can be stored on text files, spreadsheets, word processing files and so on. As a result, finding a particular model requires a long and cumbersome search among different files and directories. In order to develop a solution that provides both standardization and organization to avoid manual search and retrieval of models, the authors propose a system with two major features: (1) storing model equations and numeric data in a database, (2) adding a numerical processing software layer connected to the DBMS.

Currently, there exist mature technologies that allow storage of massive time series (ToKuDB (Namiot, 2015; Bartholomew, 2014), Vertica (Lamb et al., 2012), OpenTSDB (Dunning & Friedman, 2015), etc). However, just storing the time series is not enough to satisfy analysts’ needs. Indeed, the goal of experimental scientists is the derivation of mathematical models computed to describe in an abstract manner the observations they make (i.e. fitting a model to the time series data). This fact is especially important when the aforementioned model is not a simple mathematical formula, but a concrete and detailed set of mathematical expressions explaining how the observed system behaves and why. The authors defend the idea that such information is more crucial to persist than the observed time series itself. Also, since mathematics provides general tools to solve diverse problems, one model can be applied on different systems. Keeping this motivation in mind, the authors’ goal is to provide a standard solution to represent, store and exchange models, and take profit of this organized storage of models to help the user retrieve a model based on his raw experimental numeric data.

In consequence, the authors propose to extend the ER model with cube dimensions, to define a multidimensional database schema that allows experimental scientists to explore the database with queries. This hybrid ER extension (ER model combined with mathematical expressions) shares similarities with the extended ER model proposed in (Ordonez, Maabout, Matusevich, & Cabrera, 2014). In this work, the authors extend the ER model to represent a mathematical equation, which can be seen as a transformation of multiple input variables mixing arithmetic operators and calculus functions as primitives. The work by (Ordonez et al., 2014) takes a step further by deriving the ER model automatically from an SQL script (a sequence of queries), whereas in the authors’ prototype an equation is manually parsed and integrated into the hybrid ER model. In the future, the authors plan to add symbolic capabilities into system.

**SCIENTIFIC CONTEXT OF STUDY**

The authors’ laboratory conducts interdisciplinary research on computer science and automation. Consequently, this case study is taken from the automation domain. However, the authors emphasize that the problems and solutions addressed in this paper have broad applicability in the experimental and data sciences. As explained above, mathematical model storage is generally highly disorganized, leading to significant difficulties to retrieve and interpret them. Furthermore, it may require additional manual effort to extract relevant information from files, especially script files, where the information is obfuscated by code (i.e., a complicated program). In order to bring a solution that provides both standardization and organization as well as avoiding manual search and retrieval of models, a database coupled with a software layer of numerical processing is a promising approach.
A Hierarchical Security Model for Multimedia Big Data
[www.igi-global.com/chapter/a-hierarchical-security-model-for-multimedia-big-data/150178?camid=4v1a](www.igi-global.com/chapter/a-hierarchical-security-model-for-multimedia-big-data/150178?camid=4v1a)

Multidimensional Model Design using Data Mining: A Rapid Prototyping Methodology
[www.igi-global.com/article/multidimensional-model-design-using-data-mining/173704?camid=4v1a](www.igi-global.com/article/multidimensional-model-design-using-data-mining/173704?camid=4v1a)