Production Analytic Platform—It's a Matter of Time

NOVEMBER 2017

A ThoughtPoint by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

The Production Analytic Platform marries predictive analytics to production values and goals. Handling the complexities of analyzing time dependent data, especially from the Internet of Things, is central to this objective. The Teradata Database embeds extensive temporal and time-series storage and analytic function to enable success.

Analytics of IoT data is all about time

The analysis, representation, storage, and management of data as a function of time is vital as the Internet of Things (IoT) delivers ever more data. The Production Analytic Platform plays a key role in providing this functionality.

Time waits for no man, according to an old proverb. In the modern world, the Internet of Things waits for nobody. In the early days of data warehousing, an end-of-week snapshot was good enough for management; overnight was state of the art. Today, many decisions driven by the IoT and other real-time sources are worthless if delayed at all. Furthermore, such data often makes sense only if analyzed in *time series* that provide the context of what preceded and follows it. Traditional data types and decisions based on historical periods of time—weekly, monthly, etc.—also continue to be important.

These trends—increased data timeliness in a growing set of decisions as well as a broader range of timescales of data for analysis—amplify the demands on technology to handle time fully and efficiently. From database storage to application design, the issue of time becomes central.

Timeliness and other temporal considerations were important in delineating operational and informational environments. However, in the current business and technical world, such a simple binary distinction does not suffice. The Production Analytic Platform is designed to offer a broader range of possibilities. Two key aspects are explored here: (1) temporal database structure, which specifies how time is represented and manipulated in the data stores, and is a foundation for (2) time series support, which details how changes in data over time can be meaningfully captured, managed and analyzed.

SPONSORED BY

www.teradata.com TERADATA.

Prime time for analytics

A Production Analytic Platform provides the ideal locus for time series analytics because of its combination of a wide range of powerful analytic function and its high performance and highly reliable operational characteristics.

If you were a banker, you'd see sums of money flowing in and out of your accounts. Each movement is a *transaction*: an event of legal significance that happens at a moment in time. Although you are quite happy to track all these transactions, managing the business requires, in addition, a different view, such as the total balance in your accounts as of any given time or period. This *status* view, in contrast to a transaction, has a duration—a beginning and end date and time—and is central to business computing in every industry.

The data generated by Internet of Things records events, measures and messages[1] from the real world that are similar in structure to financial transactions. All these types of data are examples of time series data. Our focus here is on IoT data, because its size and speed give rise to particular challenges as well as offering new business opportunities.

Time series become serious business

A Boeing 777 starts up its engines: two *events* are recorded a few minutes apart. A stream of events follows: ailerons set, brakes released, and so on. The captain speaks to the control tower: a *message*. Meanwhile, on-board computers are monitoring engine temperatures, fuel flows, voltages and more on a sub-second basis: *measures* by the million. From now until the flight ends, time series are unfolding measure by measure, event by event, message by message, creating a comprehensive record of the flight.

This data, terabytes per flight, offer airlines and aircraft manufacturers, airports and aviation authorities enormous business opportunities. Preventative maintenance avoids costly flight cancellations with attendant airport disruption. New analyses reduce fuel consumption. Safety standards are improved. Similar opportunities for cost saving and profit generation are evident in almost every industry. The Internet of Things is revered; actually, it's the time series data it produces that is the real boon.

Time series data is nothing more than a time-ordered set of data records, each of which carries the date and time (usually) of creation. While common in scientific computing, times series data has been relatively rare in business computing until the advent of the IoT. Each record consists of a timestamp, an identifier of its recording device, and a payload of data of interest, often little more than strings of numbers or text. In the case of measures, observations are made at regular intervals; events and messages are recorded when they happen. Even with regular observations, data records may be lost in transmission. As a result, the vast majority of IoT data is classified as irregular time series.

Each sensor creates its own time series of identically coded measurements. A thermal sensor provides a single temperature in each record, for example. Controllers aggregate data from multiple different types of sensors into a single record, meaning that the payload may consist of a string of values of different types, such as temperature, velocity, location, etc., called multivariate time series. Because the content can change over time, the payload is often stored in flexible formats, such as JSON, CSV, and so on.

Analysis of time series data begins with the basic observation of trends. Is the temperature of this engine part increasing? When might it reach a critical value? On its own, this may be interesting, but real analytics only begins when multiple time series can be compared. Is there a simultaneous rise in vibration in an adjacent component? Can it be correlated with a change in speed or rapid ascent? What about weather conditions?

Due to record timing problems, missing values, and so on, most time series analytics is based on grouping records into time buckets or intervals, the starting point and size of which is determined by differing analytical needs.

It is the management and manipulation of multiple time series in flexible bucketing schemes that challenges data scientists. Within a single time series, simpler analytical functions—such as the mean of a sliding window—can be computed manually, paying careful attention to missing values and other issues. However, when data consists of multiple time series, each with different time intervals and missing values, the challenges rapidly escalate. Specialized, time-series aware, in-built function is the only viable approach.

Flat file solutions run rapidly out of control and power as data volumes increase. NoSQL data stores are a better solution in terms of scalability and performance, but often suffer from their weak metadata management and limited attention to operational reliability and maintainability needs.

The Production Analytic Platform to the rescue

The Production Analytic Platform offers the best of all worlds by combining the power and management of a relational database with the ability to store non-relational data, and manipulate and analyze it with a combination SQL and advanced analytic functions.

The Teradata Database now offers this foundational function with the recent introduction of support for time series data. This includes the ability to load and store data in specialized tables with a primary time index and a set of time-aware SQL functions operating within and across time buckets. Existing function that supports non-relational data, such as JSON, with full SQL support allows for multivariate time series.

In most instances, time series analytics is a high-skill exercise, requiring data scientists with knowledge of specialized languages and/or techniques. By embedding the function in the Teradata Database, access to the function is simplified and broadened. Straightforward SQL statements allow a broad range of "ordinary" business people to easily modify analytic parameters or set and change the size of the time buckets. Analyses can be iterated easily and quickly, leading to faster decisions and more relevant actions.

The space-time context

> For business, two aspects of time are of interest. Time series data from the IoT captures the dynamics of the changing world. Status data from the operational environment provides the context in business terms.

But there's more! At the top of the last section, I mentioned that the second type of time-aware data, status data, was central to business computing. With the hype around the IoT and the focus on time series data, we must keep this traditional data in mind because it is the primary source of context for time series data in the business.

Consider again our Boeing 777 airplane. Analysis of its flight data predicts that an auxiliary power unit is likely to fail within the next 50 hours of flight. It's easy to see that it should be replaced before it fails, but it's also not a safety issue, so the decision for the airline is when and where should the maintenance be performed. This depends on understanding the plane's schedule, the location and availability of the part, the cost of using a standby aircraft if necessary, and a host of other business considerations.

The data required for this decision process resides in the traditional operational systems and data warehouse of the airline—all status data with an important temporal aspect: the time periods during which all the above data is valid. The underlying technology is known as a temporal database and is a key component of the Production Analytic Platform.

Temporal databases (as the) rock

Data modelling and database design start from a view of the world in some mythical moment of "now" and ensure that the current business status is fully described and its interrelationships accurately recorded. In the real world, of course, to this is added business transactions that change the status. Furthermore, mistaken entries are made and must be corrected. Technical glitches corrupt data and are rectified later. All these events result in complex application design and extensions to the data model and database design. Recording time for events and transactions is simple. Status data is more challenging.

Tom Johnston provides an in-depth explanation of the true philosophical and technical complexity of this status data[2]. Since the early '90s database designers have proposed that time in status data is best represented in a ***bitemporal*** model, which adds two timestamps to each database record: ***valid time***, during which a fact is true in reality and ***transaction time***, during which the database record is accepted as correct. Indeed, Johnston argues that further timestamps may be needed in certain circumstances. However, even bitemporal support has been slow in its implementation, arriving only in the current decade, and limited to a subset of the main databases.

With its focus on data warehousing, the Teradata Database became the first mainstream database to add support for bitemporal data in 2010 and has since been upgraded with enhanced temporal analytic features such as derived periods and sequenced views. The correct handling the temporal aspects of status data becomes even more important in the Production Analytic Platform because the enormous volumes and variety of events to be processed demands a solid foundation of status data and its temporal context.

Conclusion

A key consideration for the Production Analytic Platform is its support for both time series data and the time-enabled status data from the operational environment and data warehouse that provides the business context.

Time series data is the basis for advanced analytics of the output of the Internet of Things. Bitemporal status data describes the fundamentals of the business, from logistics to financials, and thus offers the context in which business decisions and action can be taken based on time series analytics. The ability to fully support both these aspects of time—time series data and bitemporal data—in a single environment is a central requirement in the design and implementation of a Production Analytic Platform.

*Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** (http://bit.ly/BunI-TP2) was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI, BACollaborative, and more, Barry is based in Cape Town, South Africa and operates worldwide.

Brand and product names mentioned in this paper are trademarks or registered trademarks of Teradata and other companies.

[1] Devlin, B., *"Business unIntelligence"*, (2013), Technics Publications, New Jersey, http://bit.ly/BunI-TP2

[2] Johnston, T., *"Bitemporal Data: Theory and Practice"*, (2014), Morgan Kaufmann, Waltham, MA, http://bit.ly/2gmpdDR

SPONSORED BY

www.teradata.com

TERADATA.