Time-Series and Temporal databases and analytics

MarketUpdate

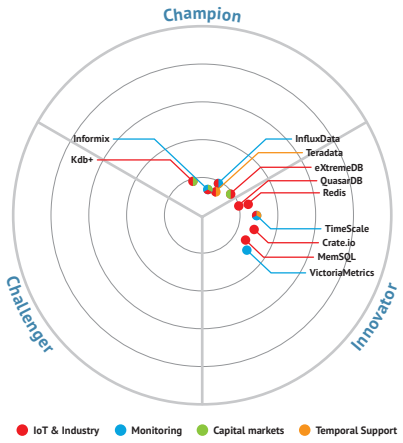
Market basics

Time-series databases represent the fastest growing database sector over the last two years, according to *www.db-engines.com*. However, we should first say that there is a distinction between time-series databases and databases that support time-series. As one might expect, the latter tend to have more general capabilities, with the former being more pure play. However, some products, targeted specifically at time-series use cases, which we would describe as time-series databases, have been built on relational roots and so inherit the capabilities of the underlying technology. As with other database markets where solutions are either native or not, the former may have theoretical advantages but it is the practical implementation that matters.

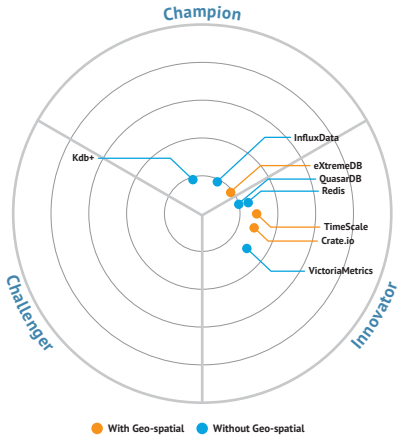
Apart from the database technologies themselves, time-series solutions are, or may be, targeted at any of three distinct markets. The first of these is about collecting metrics and monitoring your operational environment, with popular use cases include creating dashboards to monitor your network or other infrastructure, or to support DevOps or online gaming. The second is related to stock ticks in capital markets, and the third is more analytics or hybrid analytic/operational based with use cases focused on the Internet of Things (IoT) and industrial applications. While in theory any time-series database can do any of these, there are structural differences. For example, stock ticks are irregular while sensor readings are usually regular. In any case, in practice our research suggests that some – not all – products are much more focused on one area rather than another. In particular, pure-play time-series databases tend to lack any advanced geo-spatial capabilities, which are frequently required for IoT use cases, while only a minority of vendors have any focus on capital markets. In order to clarify these on our Bullseye diagrams we have colour-coded products to ensure that we are comparing like with like.

Another significant consideration is whether you just want a time-series database or whether you want a database that supports time-series but also has additional functionality such as transaction processing or OLAP (online analytic processing). This isn't just a case of time-series databases versus databases that support time-series because TimeScale, as an example, falls into the former category but, because it is based on PostgreSQL, it can do all the relational things that one would expect.

In order to clarify the various options, in this Market Update we have opted to present two Bullseye diagrams rather than just one. The first positions all the products we are comparing in this report and colour codes the various offerings by use case. However, the second considers only those products that you might choose if you had a purely time-series application in mind, with colour coding indicating the presence or absence



Bullseye 1: The whole landscape with target use cases



Bullseye 2: Likely candidates for time-series only applications

Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score. It is important to note that colour coded products have been scored relative to other products with the same colour coding.

of significant geo-spatial capabilities. Note that this doesn't mean that you couldn't use one of the omitted products – all of which have geo-spatial functionality – for this purpose but that we view it as less likely that you would license any of these products for just a single use case.

Also in this Market Update we consider a further characteristic, which is support for temporal data (the difference between this and time-series is discussed later). The only companies that we are aware of that have both time-series and temporal support are Teradata and TimeScale, the latter thanks to its PostgreSQL underpinnings. This support for temporal processing is also shown through colour coding. By way of interest, this paper does include a write-up on one further database (FaunaDB) that supports temporal data, but this is not shown in either of our Bullseyes. One further point that is worth noting is the distinction between operational and analytic processing. Some companies (Teradata is an example) purely focus on analytics, while others offer (hybrid) operational as well as analytic capabilities.

Finally, we also discuss time-series analytics in this paper, though without any comparative product analyses. Indeed, there is a paucity of products that provide analytics against historic time-series data as would be stored in a database. However, we have identified a few such products, and these are discussed in detail, though there is no accompanying Bullseye.

Market trends

Interest in time-series databases has increased significantly over the last few years, though it represents only a small sector within the overall database market (less than 1%). As there is no such thing as a temporal database (only databases that have temporal capabilities) there are no comparable figures for that market.

In practice, time-series capabilities have been around for a long time, initially to support capital market requirements. For example, Informix incorporated Illustra's time-series functionality after it acquired that company in 1996. So, this is a technology that is not new. What is new are a) databases designed specifically to ingest, index and store time-series data and b) new use cases (often IoT) that can be well served by time-series databases. In the case of the former, there is a distinction in the market between products designed from the ground up to support time-series as opposed to those that have existing relational or NoSQL-based underpinnings.

As far as use cases are concerned the initial interest in new time-series databases has come from operational monitoring of various, typically operational, environments. This would include log data analysis, network monitoring, the performance of Linux servers, DevOps progress, application performance and so on. This sort of environment forms the core constituency of users for many pure-play time-series vendors. However, there is growing interest amongst this community, in using their capabilities to support industrial and IoT use cases. However, most – not all – of these pure-plays lack the geo-spatial capabilities that are necessary for many IoT use cases, so they are currently limited to environments where this is not important. We expect this to change over time but, for the moment, competitive vendors that have time-series capabilities but are not necessarily time-series databases per se, will often have an advantage where they also have strong geo-spatial capabilities.

Vendors

There are several pure-play, open source time-series monitoring tools available on the market: Graphite, OpenTSDB (which runs on top of Hadoop) and Prometheus. The first of these is described as a *"data logging and graphing tool for time-series data"* (though the graphing is often replaced by Grafana – see later), and OpenTSDB is similar. Prometheus, on the other hand, is described as a *"time-series database and monitoring system"*. Since this is a Market Update focused on databases we review Prometheus but not Graphite or OpenTSDB. However, Prometheus as a database is quite limited, only providing local storage, and there are various providers of more sophisticated storage engines for Prometheus. For example, you can use either TimeScale or InfluxDB in this way. However, these are reviewed in their own right in this report so we have chosen to review VictoriaMetrics in conjunction with Prometheus, as an exemplar of this approach.

From a trend perspective the main one is simply that more vendors are introducing time-series capability. A notable example is Redis; MemSQL, which has an architecture that is well suited to supporting time-series, is planning to add specialised time-series capabilities – for example, time buckets – that it does not currently support; and Amazon Timestream is in preview. While we have been provided with details with respect to Timestream the company was unable or unwilling to give us a release date for this product, so it is not included in this report. Another major trend, as already mentioned, is for vendors (Crate.io, InfluxData, QuasarDB and so on) to add geo-spatial capabilities, while a different focus for some

suppliers is on providing improved compression capabilities. TimeScale, which historically been limited to a single node version, has announced a multi-node edition.

Temporal databases

Temporal data is a different kettle of fish from time-series data. Where the latter consists of a series of either regular (sensor data) or irregular (stock ticks) readings, temporal values have a start time and (at least theoretically) an end time, either or both of which may be applied retrospectively. For analytic purposes this means being able to make *"point-in-time"* queries against your database. Thus you might use temporal data to analyse emergency vehicle response times. In addition, you will often need support for both *"valid times"* and *"transaction times"*. This is best explained by example. Suppose you moved to address A on the 1st of January 2006 and moved to address B on the 1st of January 2009 but forgot to inform your local tax authority about the move until the 1st of January 2010. The *"valid time"* that you were at address A was from 2006 to 2009 but the *"transaction time"* (when the local tax authority thought you were at address A, even if you weren't) was from 2006 to 2010. For obvious reasons, the tax authorities will need to understand both valid and transaction times. Some temporal use cases like this – for example, if a product is moved from one product group to another – can be handled using slowly changing dimensions, but many cannot.

Historically, SQL did not support temporal processing and as far back as 1994 a specification of TSQL2 was published, which has been adopted by some vendors, most notably Teradata. As far as ANSI standard SQL is concerned, temporal processing functions were not agreed until SQL 2011 and there are a number of vendors that support this, including (but not limited to): IBM Db2, MariaDB, MarkLogic, Microsoft SQL Server, Oracle and PostgreSQL. In theory, you could also use any append-only database as a temporal database. However, many temporal use cases involve personal data, in which case you would need special capabilities to handle such things as the right to be forgotten. In practice, therefore, you need some sort of special capabilities in order to satisfactorily support temporal capabilities. We include a description of FaunaDB in this report precisely because it is an append-only database that does have such special capabilities.

As far as this report is concerned both Teradata and TimeScale support temporal as well as time-series data but, of these two, only Teradata has any sort of focus in this area.

Analytics

Of course, ingesting, storing and retrieving time-series data efficiently is one thing, but analysing it is quite another. There are several relevant scenarios:

- monitoring and dashboarding using appropriate visualisations,
- analytics on streamed (time-series) data,
- machine learning (time-series forecasting) and
- analytics against historic time-series and/or temporal data, which may/may not be combined with real-time data.

There are lots of tools that support visualisation. Grafana is a notable open source example but you can also use Tableau, Qlik and many of the other usual suspects. We do not claim to be visualisation experts, so no analysis of these tools is included here. The same applies to streaming analytics, where all the products support time window functions – see ***https://www.bloorresearch.com/technology/streaming-analytics-platforms/***. Again, we would expect all machine learning libraries to support time-series forecasting.

By contrast, there is a distinct paucity of analytic products that have any special features built into them to support time-series analyses. Tableau, for example, told us that this was on their roadmap for 2020. During our research for this report we specifically asked database vendors what analytics tools could be used with their data and we got very few answers. Two that can be so used are TIBCO Spotfire and Trendalyze, reviews of both which are included here as is Interana, which focuses on behavioural analytics based on time-series data. Also notable are Trend Miner from Software AG and Seeq, both of which offer time-series based analytics specifically to support process manufacturing.

One trend we have noticed is towards shape (or pattern) detection and recognition. For example, IBM Informix has facilities to detect patterns, though it needs to know what patterns to look for). Similarly, IBM SPSS has also introduced the ability to find time-series patterns to support, for example, segmentation. Trendalyze focuses specifically on this area and in this context, it is interesting that DARPA (Defense Advanced Research Project Agency) has recently stated that it believes that the next wave of AI – going beyond the current statistical learning approaches – will be in what it calls *"contextual learning"*, which is precisely where shape recognition comes into play.

Metrics

We have evaluated the databases in this report against the following criteria, which are arranged in two groupings:

- **Time-series architecture.**
  By this we mean the capabilities provided for ingesting time-series on the one hand and storing and retrieving it on the other. In terms of ingestion this will typically mean support for loading terabytes of data per day and needs to support both regular and irregular time-series. In some environments you may need to support very short time intervals (sometimes referred to as Hertz) where you are getting tens of thousands of readings per second. From a storage point of view – often via arrays – compression is an important consideration, as is the ability to refrain from recording time stamps for regular time-series. We would like to see more companies support static capabilities whereby you don't store additional information when readings remain static, or within defined limits, for specified periods of time. For retrieval, specialised indexes will be appropriate. A database optimiser tuned to support time-series should understand these indexes.

- **General architecture.**
  This includes all those elements not specific to time-series. In particular, ease of maintenance and administration, automated tuning capabilities, high availability and so on.

- **Time-series processing.**
  This is essentially about the way that you manipulate time-series data. This includes language support and we take the view that using SQL is an advantage compared to having to learn a new language. This category also includes the provision of pre-defined statistical and other functions for analysing time-series data, including such things as support for time buckets, windows (rolling, static or moving), and facilities to write user defined functions or routines.

- **Tools.**
  This includes both third-party integration and the provision of complementary tools by the vendor itself. We regard this as particularly important both from an ingestion and processing perspective. As far as the former is concerned, we would like to see support for Kafka, Flink, Edgent, NiFi and other technologies that provide streaming services into the database environment. This is important even where the database vendor can provide its own streaming capabilities. A similar argument applies to both visualisation – which some companies provide – and also to analytics tools. We would prefer to see native connectors rather than ODBC/JDBC and support for machine learning libraries as well as products such as Jupyter Notebooks.

- **Performance and scalability.**
  This should be self-evident. In the latter case we are talking about the ability to scale up and out as opposed to scaling down to be embedded into sensors and other devices (see below).

The second group of metrics are significant, depending on the use case, but are not present in all products. Scores have been omitted on the individual product write-ups where these are not present or are minimal. For Bullseye positioning, multiplying factors have been applied so that products without a particular capability are not disadvantaged.

- **Geo-spatial capability.**
  As we have mentioned, not all products have significant capability in this respect with support ranging for barely being able to tore latitude and longitude to products that implement multiple (as many as 20) different co-ordinate systems. Products with little or no geo-spatial capability have not been scored. As a brief summary of requirements, you will want to support spatial joins (or the equivalent in non-relational environments), the ability to associate different mapping layers while keeping them physically separate, and to store both vector and raster data. The most common form of vector data is in so-called *"shape files"*, of which there are various types, including extensions that support elevation information. Vector data is normally stored as an array (not dissimilar to time-series) and indexed (ditto) along with extensions to SQL (or whatever language is used) to support spatial analytics. Support for GeoJSON is increasingly common. Spatiotemporal queries require the ability to combine geo-spatial and time-series data within a single query or search. Typically, you will require lossless compression or lossy (designed to cater for the fact that human vision will fill in gaps) compression, in order to reduce storage requirements and improve performance. Note that this means having one type of compression for spatial data and another for time-series data.

- **Embedding.**
  In the context of IoT, a small footprint, allowing the database to be embedded in edge and gateway devices. In this context, a database that supports a *"fire and forget"* approach is mandatory. Alternatively, there are products that support agents that can be deployed at the edge, rather than the database per se. It is also worth commenting further on the database optimiser. Most modern databases use a cost-based optimiser. However, this has the downside that you need to collect statistics, which puts an overhead on performance. There is an argument in favour of a rules-based optimiser if you are going to embed a database at the edge.

- **Other functionality.**
  This is the extent to which the product supports other capabilities beyond time-series. Temporal capability is relevant but so is support for conventional transaction processing and analytics. Note that scores here are not necessarily directly comparable with one another. For example, Teradata is a data warehousing vendor while MemSQL is general-purpose database: you wouldn't expect them to have the same set of features and capabilities.

For products that do not appear in either of our Bullseye diagrams (that is, FaunaDB and various analytic offerings) different metrics are used, each of which is dependent on the particular product, and what we would expect, and like to see, from a product of this ilk. The names of the metrics used should give a good indication of what we are measuring. The exceptions that may require explanation are the metrics applied to analytics products for Shape Discovery and Shape Matching. The former refers to the discovery of shapes (patterns) of interest and the latter to the identification of shapes that match those patterns of interest.

For the products that we are directly comparing, the metric scores are shown on the next page.

Conclusion

The market for time-series is both mature and immature. The initial development of time-series capabilities was in the early 1990s – Kx Systems was founded in 1993 – but this was for irregular time-series for stock ticks. It wasn't until a decade later – McObject was founded in 2001 – that companies started to develop support for the sort of regular time-series typified by sensor data. And it wasn't until the current decade (and the latter part of it that) that pure-play time-series databases have emerged for monitoring infrastructure and DevOps environments. Temporal support is a similarly recent development. More mature products can be expected to have all of the scalability, resilience and high availability that mission-critical applications will require. While this is less likely to be the case in more immature environments, it is arguable that monitoring type environments are not typically mission-critical and therefore do not need to meet the same standards.

Selecting a product to support time-series capability is therefore not a trivial process. It obviously depends on the use case but in the fastest growing area – IoT – it also depends on the extent to which you need to combine geo-spatial with time-series data, and/or whether you want to embed your selected product into edge or gateway devices, or whether this is more of a central processing function, or both. Further, there is the question of whether any solution is going to be dedicated specifically to this use case or whether you would like a platform where time-series is only one of several use cases that you wish to resolve.

Vendors always look to expand their product's functionality and we expect a convergence of capabilities across the various products we have reviewed in this Market Update, over the next few years. There will remain distinctions between companies targeting different broad categories of use case but otherwise the market will become more homogenous. However, this has not happened yet and there remain significant differences between the various offerings in this space. We trust that this Market Update will help you to determine which is the most useful solution for your use case.

About the author

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

MarketUpdate

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

tel: **+44 (0)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**