Bloor

InDetail

InDetail Paper by Bloor
Author **Philip Howard**
Publish date **March 2016**

Kx and the Internet of Things/ Big Data

Kdb+ has proved itself in what is arguably the most demanding big data market: financial trading and risk management. The technology is well-suited to a variety of other environments including preventative maintenance (asset management), smarter cities, the connected car and other such environments where the collection and analysis of time-series based data is important.

Author **Philip Howard**

Executive summary

Kx technology consists of the kdb+ database, the q language and the CEP (complex event processing) engine that are associated with it. In addition, there are also a number of tooling products designed for non-specialist users that have been built using Kx technology. In this paper we are primarily concerned with the kdb+ database. This is a column-based relational database with extensive in-memory capabilities, developed and marketed by Kx Systems. Like all such products, it is especially powerful when it comes to supporting queries and analytics. However, unlike other products in this domain, kdb+ is particularly good (both in terms of performance and functionality) at processing, manipulating and analysing data (especially numeric data) in real-time, alongside the analysis of historical data. Moreover, it has extensive capabilities for supporting time-series data. For these reasons Kx Systems has historically targeted the financial industry for trading analytics and black box trading based on real-time and historic data, as well as real-time risk assessment; applications which are particularly demanding in their performance requirements. The company has had significant success in this market with over a hundred major financial institutions and hedge funds deploying its technology. In this paper, however, we want to explore the use of kdb+ for big data and Internet of Things based use cases.

Fast facts

Kdb+ is a column-based, hybrid in-memory database with stream processing capabilities, primarily designed for analytic workloads. In so far as in-memory capability is concerned, we refer to it as "hybrid" because it uses in-memory processing as much as it can but recognise that in some cases it may be impracticable to load all relevant (typically historic) data into memory and that you therefore need to employ techniques that will not only leverage memory-based processing but optimise performance when not all the data can fit into memory. This is similar to the approach taken by IBM with its DB2 BLU Acceleration, as an example. The product's stream processing capabilities (which means that you can analyse very large quantities of information in-flight, in real-time) arise from the fact that kdb+ is tightly integrated with the product's development language q. This is a vector (array) processing language that is much more efficient than SQL and which can be used to develop analytic applications as well as for query purposes.

Key findings

In the opinion of Bloor Research, the following represent the key facts of which prospective users should be aware:

- Data is stored in sequentially ordered columns. Note that other column-based databases are not typically ordered in this fashion so you would expect performance for operations such as sorts to be much faster when using kdb+. As is usual with column-based databases it is not necessary to define indexes because columns are self-indexing (and this [is] even more true in this case, thanks to the sequential ordering). However, you may define indexes if you wish to.
- Kdb+ supports compression at comparable levels to other vendors. That is, up to around 10x depending on the type of data.
- In-memory processing is genuinely in-memory: it is not just a cache. However, the system has been designed so that both pure in-memory and hybrid in-memory/disk based processing are optimised.
- The support for time series in kdb+ is especially relevant to many Internet of Things applications such as smart metering, preventative maintenance and other environments where large quantities of data need to be collected at regular intervals and then analysed. Note that native support for time series is extremely rare across database products.
- We are pleased to hear that Kx has recently implemented geospatial capabilities as many Internet of Things applications are both time and location based.

> **"...unlike other products in this domain, kdb+ is particularly good (both in terms of performance and functionality) at processing, manipulating and analysing data (especially numeric data) in real-time, alongside the analysis of historical data."**

Today, Bloor Research believes that the company is well placed to exploit the capabilities of kdb+ in other big data markets and, especially, with respect to the Internet of Things.

- The q language is significantly more efficient than other languages that you might use (both procedural and declarative) for analysis purposes. It allows very complex business logic to be developed quickly. However, its use does imply a learning curve. Alternatively, there are specific C# and JavaScript interfaces that are provided as a part of kdb+. There are interfacing capabilities for a host of other languages and environments, including (but not limited to) Python, R, Matlab, Excel and Mathematica.
- Because programming is typically in q rather than SQL most business intelligence tools will not run natively with kdb+. Kx Systems' parent, First Derivatives, has built a suite of business intelligence tools including a visualisation tool on top of kdb+. In addition, kdb+ has an ODBC 3 driver which works with Tableau and Excel.
- Kdb+ is scalable, does not require a large investment in hardware, is relatively simple to maintain, performs extremely well and is highly available.

The bottom line

Kx Systems was founded more than 20 years ago, although the precursor to kdb+ (kdb) was not introduced until 1998. In the eighteen years since then the company and its product have earned a well-deserved reputation in the financial services market. That, in itself, speaks volumes: hedge funds and other financial institutions have been processing "big data" since before the term came into common parlance and Kx Systems has been a leading provider to that market. Today, Bloor Research believes that the company is well placed to exploit the capabilities of kdb+ in other big data markets and, especially, with respect to the Internet of Things. In fact, Kx Systems has already started to make inroads in this area, as we will discuss (by means of case studies) in due course. Our view is that Kx Systems has, perhaps serendipitously, built a platform that makes it ideal for many big data analytic requirements.

The product

Kdb+ runs on Solaris, Windows and Linux platforms and is currently in version 3.3 (version 3.4 will be released in Q2 2016). In addition to the standard kdb+, which is 64-bit, there is also a free 32-bit version.

The standard version has a parallel architecture for deployment across multiple partitions (and, notably, you can implement different indexes on different partitions) and has a distributed capability so that it can scale across a clustered environment. This architecture helps to support the product's failover capabilities, which are supplemented by full logging facilities and automatic recovery. Load balancing and replication are both provided. The product includes its own Web Server, which is used to support web-based queries that utilise either a standard URL link or which may derive from applications such as Microsoft Excel. The results are returned to the user's browser either in HTML or XML format, or they can be exported to an Excel spreadsheet. Unicode is supported.

Historically, there was an optional product called kdb+tick that was specifically designed for ingesting and analysing (stream processing) stock tick information. However, this has now been folded into kdb+. There is also a fast loading capability.

The solution supports commodity servers and storage devices in any topology (local, remote, SAN, NAS, flash, SSD, HDD), providing flexibility in cost-effectively adding capacity when required.

The standard version has a parallel architecture ... and has a distributed capability so that it can scale across a clustered environment.

Architecture

There are really two parts to kdb+: the q language and the database. We will discuss each in turn.

The q language

The origins of q lie in a programming language called APL. This was a language used from the early days of computing, primarily to process numerical data. In 1988, while he was at Morgan Stanley, Arthur Whitney developed A+ as a replacement for APL in order to support non-mainframe environments. However, while this crunched numbers more efficiently than APL, it was a proprietary development on behalf of Morgan Stanley. So, in 1993, Arthur co-founded Kx Systems and developed a further and more powerful replacement language, called k. This was extremely efficient but equally terse and it was very difficult for non-experts to read or understand so, during the first decade of this century Arthur developed a more verbose version of k, called q. While you can still use k almost all Kdb+ customers use q which, while still implying a learning curve, is much easier to use and understand: for example, it has common functions familiar from SQL such as SELECT statements and WHERE clauses as well as updates, deletes and so forth. Where it goes beyond other environments is in its mathematical functions, such as variances, and in the fact that developed applications are tightly integrated with the data and database.

More technically q is a vector programming language (that is, it addresses vectors [arrays] rather than tables) that uses memory mapped files for numeric processing. This has important implications, not only in its own right (because you get better performance) but also because Intel is increasingly adding vector processing capabilities to its processors. As Kx Systems is a partner of Intel, the former aims to leverage each new piece of relevant functionality that Intel introduces, so that performance will continue to improve.

There are a number of IDEs (integrated development environments) that can be used in conjunction with kdb+. Because of the conciseness of the code, programs tend to consist of relatively few lines, which makes debugging relatively simple. Code is interpreted but the interpreter takes up not much more than 100K and this makes it easy for the environment to support many simultaneous processes. This small footprint not only means that installation is very fast, it also reduces the risks and costs associated with upgrades and maintenance.

> **q has functions familiar from SQL such as SELECT statements and WHERE clauses, as well as updates, deletes and so forth.**

The database

As has already been stated, data can be stored in compressed format in a columnar database. Enough has been written (by both Bloor Research and others) about the advantages of column-based databases for analytics that we do not need to reiterate those arguments here. However, it is worth commenting on the approach kdb+ takes to in-memory processing. There are, essentially, three methods of implementing so-called in-memory databases. One is essentially to take a cache-based strategy and call it in-memory, which it isn't. The second is to assume that all relevant data can fit into memory. The problem with this is that it becomes expensive if the datasets to be analysed are anything other than relatively small. And it becomes prohibitively expensive as datasets become very large. Another problem is what happens if there is a system failure and you have to recover the system. The third option is to implement a hybrid in-memory/disk architecture that is optimised to use whatever resources are available which, in our view is a better solution. In particular, stream processing often requires that historic data is needed for context purposes and, given that a year's worth of tick data (for example) comprises something like 6 or 7TB then you really do need hybrid capability. As far as availability is concerned kdb+ stores memory images on disk to facilitate rapid recovery.

A further feature of kdb+ that is worth mentioning is the fact that it supports user-defined datatypes as well as built-in datatypes. In the latter case, time series support (which can be accurate down to nanoseconds) is implemented by means of specific datatypes for elements such as dates. Similarly, a relational table is a base datatype.

Use cases

So far we have merely described kdb+ in terms of its technology, and we have suggested that it would be suitable for use outside of financial services and in other industries where big data and the Internet of Things are significant issues. In fact, Kx Systems has already started to make headway in this market and it will be useful to discuss uses in general. Some of the following are live applications, some of them are planned by existing users and some of them are theoretical possibilities but they all highlight the potential that kdb+ offers. There are, of course, endless possibilities when it comes to analysing time-sensitive machine generated data but the following will give a flavour of environments for which kdb+ is suitable.

Utilities

Historically, utility companies have collected meter readings via readers calling at the door or, more recently, by customers posting their own readings online. Such readings are typically processed in a transactional database of some kind, which also supports query and analytic processing of this data. However, with smart meters collecting data very much more frequently, the amount of data to be processed in such queries is beyond the scope of traditional databases, especially when low latency queries are required. One company in this space has used kdb+ for a data mart to meet these query requirements, which include on-demand, batch, and ad-hoc analysis and aggregation queries. The kdb+ database is updated in real-time via change data capture.

Smart devices

This is an area which other Kx customers are investigating, specifically in the healthcare sector, to capture medical sensor data (telemedicine) and data in hospitals in areas such as intensive care and neonatal units. In particular, companies think that new smart "fit" wristbands and similar devices, that become part of the Internet of Things, are perfect for using kdb+ to give doctors the ability to perhaps predict events such as heart attacks before they occur.

Asset management

There are two aspects of asset management for which kdb+ might be suitable. The first is preventative maintenance and the second is real-time monitoring. These are closely allied but distinct. As an example, a compressor on an oil rig typically has around 120 sensors that are constantly being monitored. Immediate alerts need to be raised if, for example, vibration exceeds a certain tolerance level. Typically, this is currently done via banks of red lights but real-time monitoring can not only raise alerts but also provide additional contextual information to engineers based on previous history. Preventative maintenance takes this same facility a step further by taking historic information, analysing it and predicting when problems are likely to occur so that pre-emptive action can be taken.

Risk analysis

Another application where kdb+ can be (and is) deployed is to support epidemiological studies, a typical scenario being post-market risk studies. That is, analysing how a particular drug is being used by different populations, its efficacy, whether any unexpected side-effects are present, and so on. Information to be analysed is mainly derived data generated by insurance companies from patient claims, as well as prescription data from reporting pharmacies. The results form part of a feedback loop with R&D.

Quality management

This is another [area] being explored by one of Kx's customers in manufacturing. In this environment there is a lot of time-series sensor data that could be quickly manipulated by Kdb+. The company wants to correlate this manufacturing sensor data with the quality of its products.

"...with smart meters collecting data very much more frequently, the amount of data to be processed in such queries is beyond the scope of traditional databases..."

Performance and competition

It is worth commenting on some of the conversations that have been held with Kx customers in order to compile this report. In particular, when it comes to "big data" most organisations will think of Hadoop or other NoSQL environments in the first instance. Kdb+ users, on the other hand, have concluded that while Hadoop is good for batch-based processing and for environments where queries are pre-determined it does not offer the real-time and ad hoc performance that kdb+ can provide. Moreover, Hadoop requires considerably more system resources: one user commented that after their proof of concept they estimated that they would require between 10 and 50 times more servers to support Hadoop than kdb+. A similar comparison with an appliance-based solution required sixteen times as many cores. The amount of programming and management overhead associated with Hadoop was also a major concern.

In terms of performance, one company, which has a database that is measured in hundreds of billions of rows, stated that queries which used to take weeks to run on their previous system now run in *"seconds or less"*. In so far as learning q is concerned the comment we were given was that "*q is easier to learn than some of the open source technologies and libraries*". Moreover, we were told that it is much more efficient than SQL: *"by orders of magnitude"*.

On a slightly different topic we want to briefly mention STAC. This is an independent benchmarking group that is funded by some 200 leading financial (and other) institutions. It benchmarks both hardware and software by keeping one fixed and changing the other. Vendors frequently use kdb+ as its fixed software because that is the fastest database product it has been able to find (or that has applied).

Kdb+ users have concluded that Hadoop does not offer the real-time and ad hoc performance that kdb+ can provide.

The vendor

Kx was founded in Palo Alto, California in 1993 and kdb was first introduced in 1998 with kdb-tick following in 2001. These were both 32-bit products which have subsequently been replaced by the 64-bit versions discussed in this report.

Kx Systems, Inc. is a subsidiary of First Derivatives Inc. plc, which acquired a majority shareholding in Kx in October 2014. First Derivatives, which has been a long-time partner of Kx, and which previously had a minority shareholding in the company, is based in Northern Ireland and is publicly listed on the AIM market. Over 20 years old, it employs over 1,500 people worldwide and has operations in London, New York, Stockholm, Singapore, Hong Kong, Tokyo, Sydney, Toronto, Philadelphia, Dublin, Belfast, Zurich and Palo Alto.

The company's historical approach to marketing has been both through direct marketing and via channels. It has a number of partners around the world that provide product sales, training and installation as well as first line support. Many of these partners have extended Kx's capabilities by offering specialised financial capabilities with things like graphical user interfaces for business intelligence and extended complex event processing. In addition to these partners the company also has OEM partners. For example, 1010Data, a cloud-based data warehousing solution for financial service companies, is based on Kx technology.

However, while this has been Kx's approach to financial markets it has been directly addressing big data opportunities outside the financial sector. As this becomes more mature we expect the company to follow a similar channel model.

Kx web address: ***www.kx.com***

“It has a number of partners around the world that provide product sales, training and installation as well as first line support.”

Summary

Kdb+ has proved itself in what is arguably the most demanding big data market: financial trading and risk management. The company is now targeting other environments. It has already achieved initial success in smart metering and pharmaceuticals/healthcare and, in our view, the technology is well-suited to a variety of other environments including preventative maintenance (asset management), smarter cities, the connected car and other such environments where the collection and analysis of time-series based data is important.

We would have to say that Kx has been fortunate. We do not believe that it anticipated the Internet of Things. What it did was to focus on offering the best possible performance for its chosen (financial) market. It just happens that the (big data) issues faced by that market are exactly analogous to many Internet of Things environments. Not only is Kdb+ suitable for implementation for many of these use cases but, more importantly, the product has significant technical advantages in these areas. We therefore expect a significant expansion of the company's user base within the Internet of Things domain.

> **We ... expect a significant expansion of the company's user base within the Internet of Things domain.**

FURTHER INFORMATION
Further information is available from
www.BloorResearch.com/update/xxxx

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.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter 'noise' and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent 25 years distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

Copyright and disclaimer

2nd Floor
145–157 St John Street
LONDON EC1V 4PY
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**