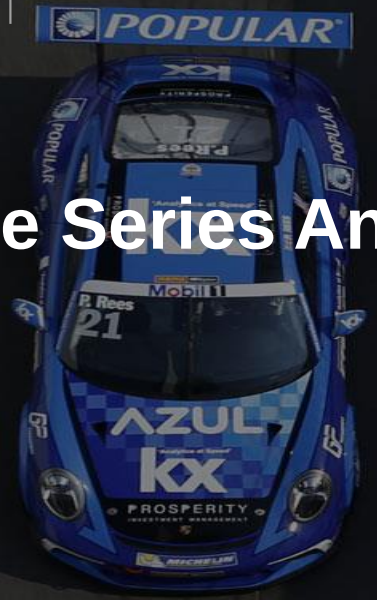
kx | it's about time



Time Series Analytics for Big Fast Data

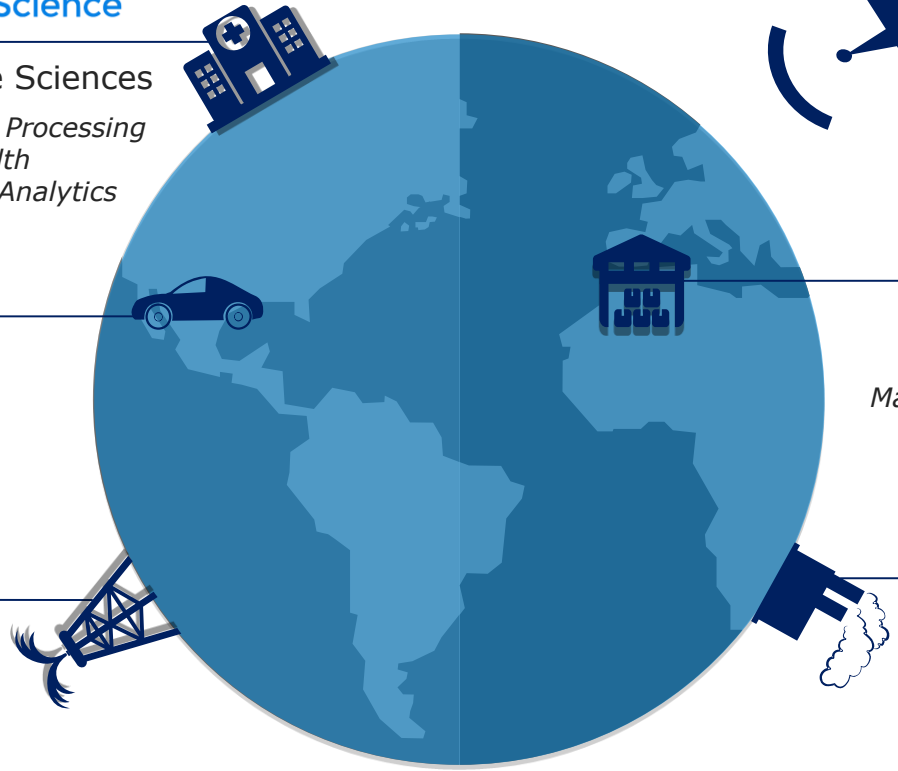
Jonathan Guy
June 2018

About Kx

- Founded in 1993
- Subsidiary of First Derivatives Plc (LON:FDP)
- 2,000+ employees
- Global presence

Customers & Partners

Other Verticals



Health & Life Sciences

Genomics Data Processing
Connected Health
Patient Record Analytics

Space and Telco

Earth Observation
Geospatial Data Analytics
Anomaly Detection

Automotive

Telemetry Analytics
Anomaly Detection
Performance Analytics

Retail

Retail Analytics
Marketing Optimization
Customer Journey

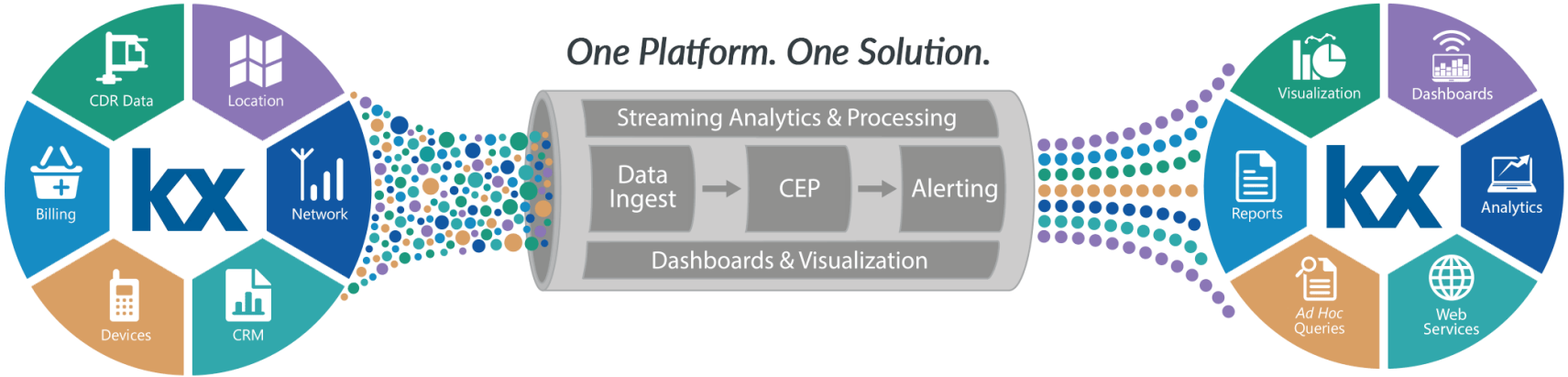
Utilities

Sensor Analytics
Smart Meters Data
Predictive Analytics

Manufacturing

Edge Computing
Multivariate Analysis
Fault Detection

The Kx System



Limitless New & Existing Sources

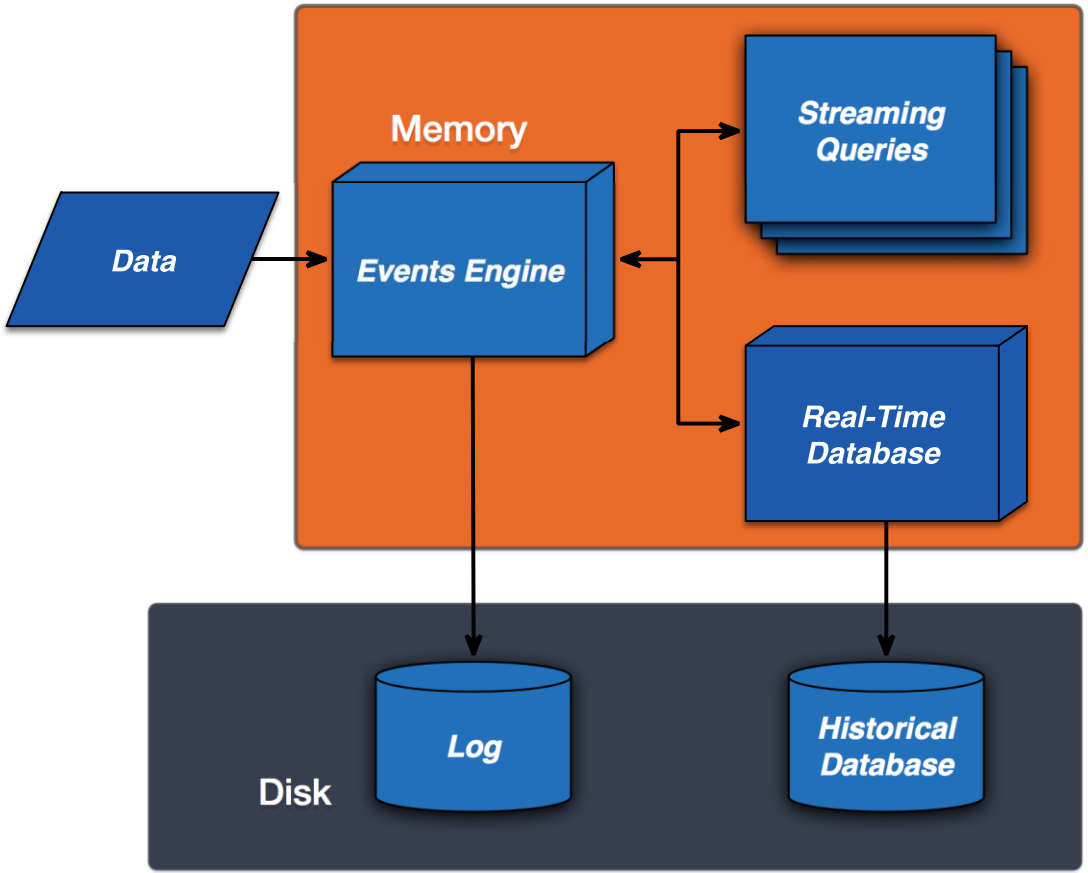
- Network Data
- Real-time & Historical Data
- Fully Managed Data Quality

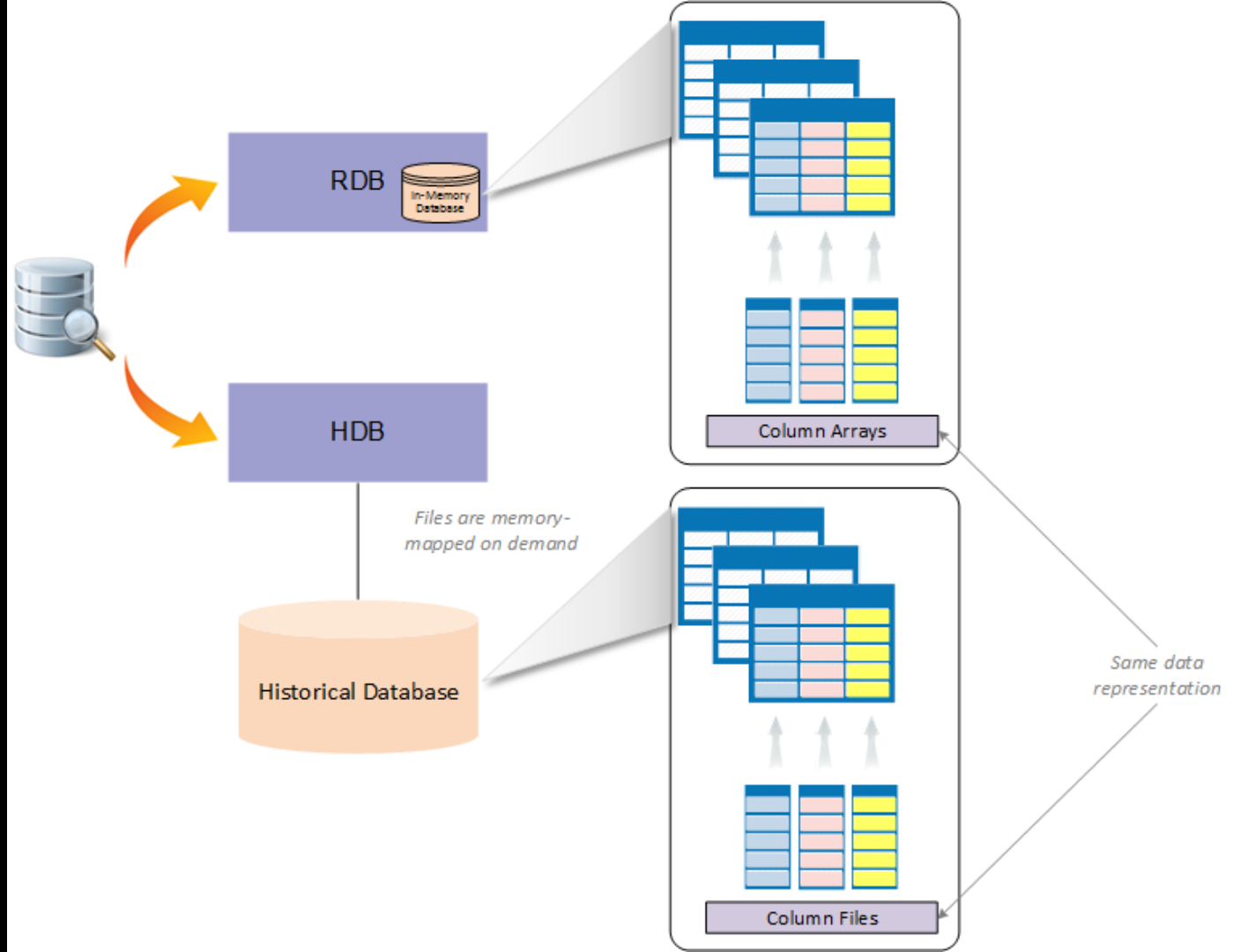
Real-time Actionable Insights

- Extreme Performance
- In-memory, Relational & Columnar
- Powerful Query Language

- Unified columnar database & programming system
- Lambda architecture
  - Streaming
  - Real-Time (In-Memory)
  - Historical (On-Disk)
- In-database analytics
- Support for joins
- 500 kB binary

Typical kdb+ Architecture




RDB
In-Memory Database
HDB
Historical Database
Files are memory-mapped on demand
Column Arrays
Column Files
Same data representation

Data Types

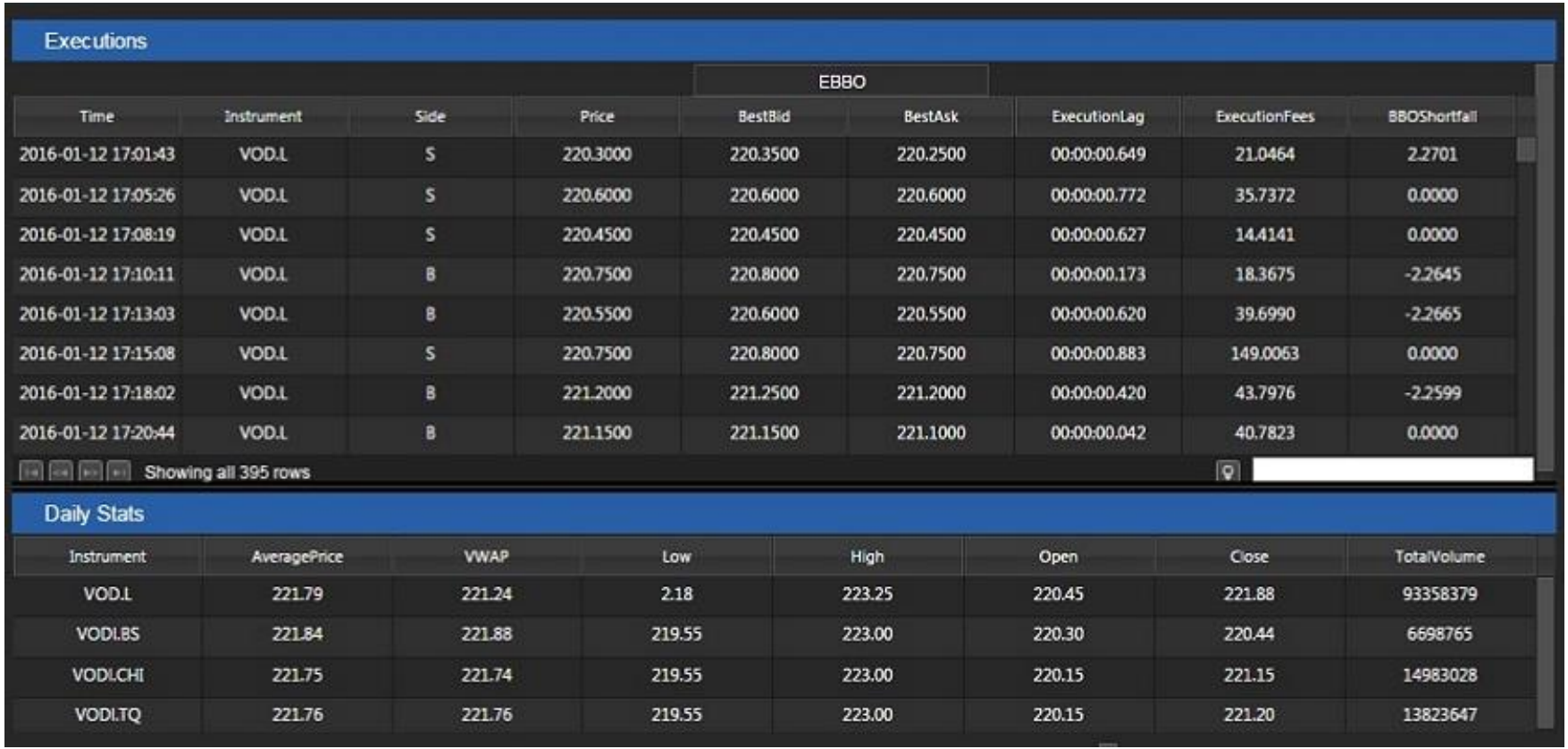
Database Data Types

- Boolean
- Byte
- Integer (short/int/long)
- Floating Point (real/float)
- Character
- Symbol (interned character)
- GUID
- Enumeration
- Dictionaries
- (Keyed) Tables
- Functions

Time Series Data Types

- Date
- Time
- Minute
- Second
- Month
- Datetime
- Timespan (ns)
- Timestamp (ns)

Example Database



Executions

EBBO

Time	Instrument	Side	Price	BestBid	BestAsk	ExecutionLag	ExecutionFees	BBOShortfall
2016-01-12 17:01:43	VOD.L	S	220.3000	220.3500	220.2500	00:00:00.649	21.0464	2.2701
2016-01-12 17:05:26	VOD.L	S	220.6000	220.6000	220.6000	00:00:00.772	35.7372	0.0000
2016-01-12 17:08:19	VOD.L	S	220.4500	220.4500	220.4500	00:00:00.627	14.4141	0.0000
2016-01-12 17:10:11	VOD.L	B	220.7500	220.8000	220.7500	00:00:00.173	18.3675	-2.2645
2016-01-12 17:13:03	VOD.L	B	220.5500	220.6000	220.5500	00:00:00.620	39.6990	-2.2665
2016-01-12 17:15:08	VOD.L	S	220.7500	220.8000	220.7500	00:00:00.883	149.0063	0.0000
2016-01-12 17:18:02	VOD.L	B	221.2000	221.2500	221.2000	00:00:00.420	43.7976	-2.2599
2016-01-12 17:20:44	VOD.L	B	221.1500	221.1500	221.1000	00:00:00.042	40.7823	0.0000

Showing all 395 rows

Daily Stats

Instrument	AveragePrice	VWAP	Low	High	Open	Close	TotalVolume
VOD.L	221.79	221.24	2.18	223.25	220.45	221.88	93358379
VODI.BS	221.84	221.88	219.55	223.00	220.30	220.44	6698765
VODI.CHI	221.75	221.74	219.55	223.00	220.15	221.15	14983028
VODI.TQ	221.76	221.76	219.55	223.00	220.15	221.20	13823647

q Programming Language

- Functional
- Array/Vector
- Query
- Interpreted

q – a language for time series

- Time Series Functions/Joins
  - xbar
  - Bi-Temporal
    - aj (As-of join)
    - wj (Window join)
- Temporal Arithmetic

OMG I CAN'T BREATHE 😂😭

Parallelism

- Vertical Scaling
  - Multi-threading
  - No serialization
  - Automatically distributes queries across CPU cores
- Horizontal Scaling
  - Multi-processing
  - Automatically distributes queries across machines

Compression

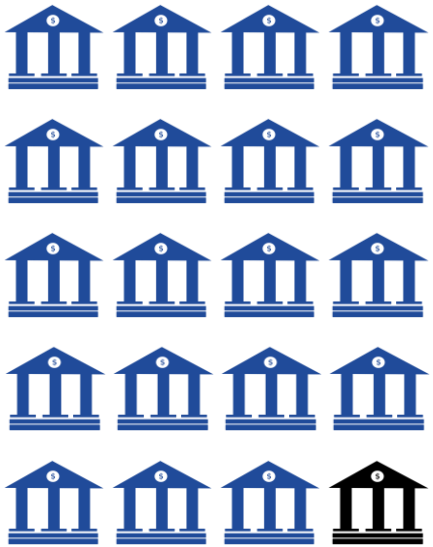
- WebSocket compression
- In-flight compression between hosts
- On-disk compression
  - kdb+ algorithm
  - gzip
  - Google Snappy
  - lzh4c

Kx Performance Snippets

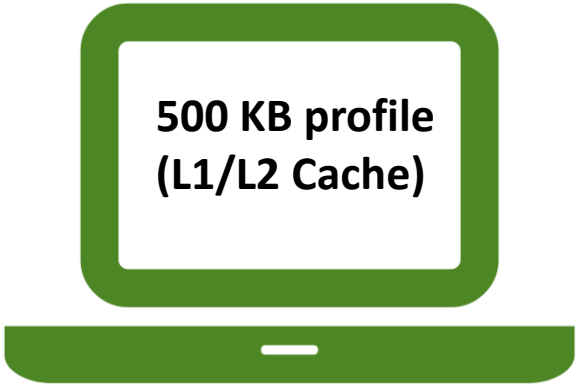
Process & store **4.5 million** events/second/core

Ingest data at **10 million** records/ second/core

Search in-memory tables at **4 billion** records/ second/core



Trusted by 19/20 World's Top Investment Banks

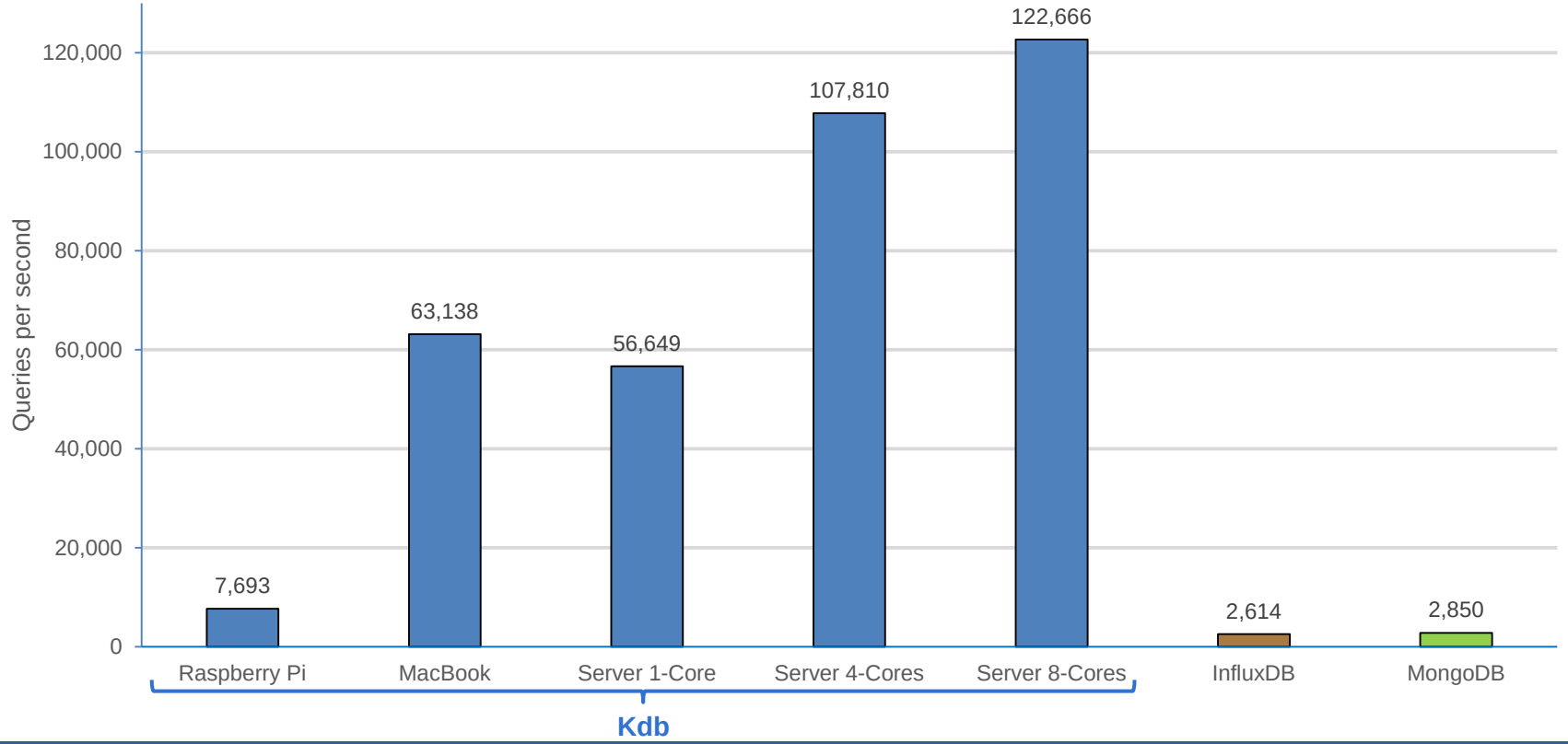


Streaming

1.6 TB

of Data Daily

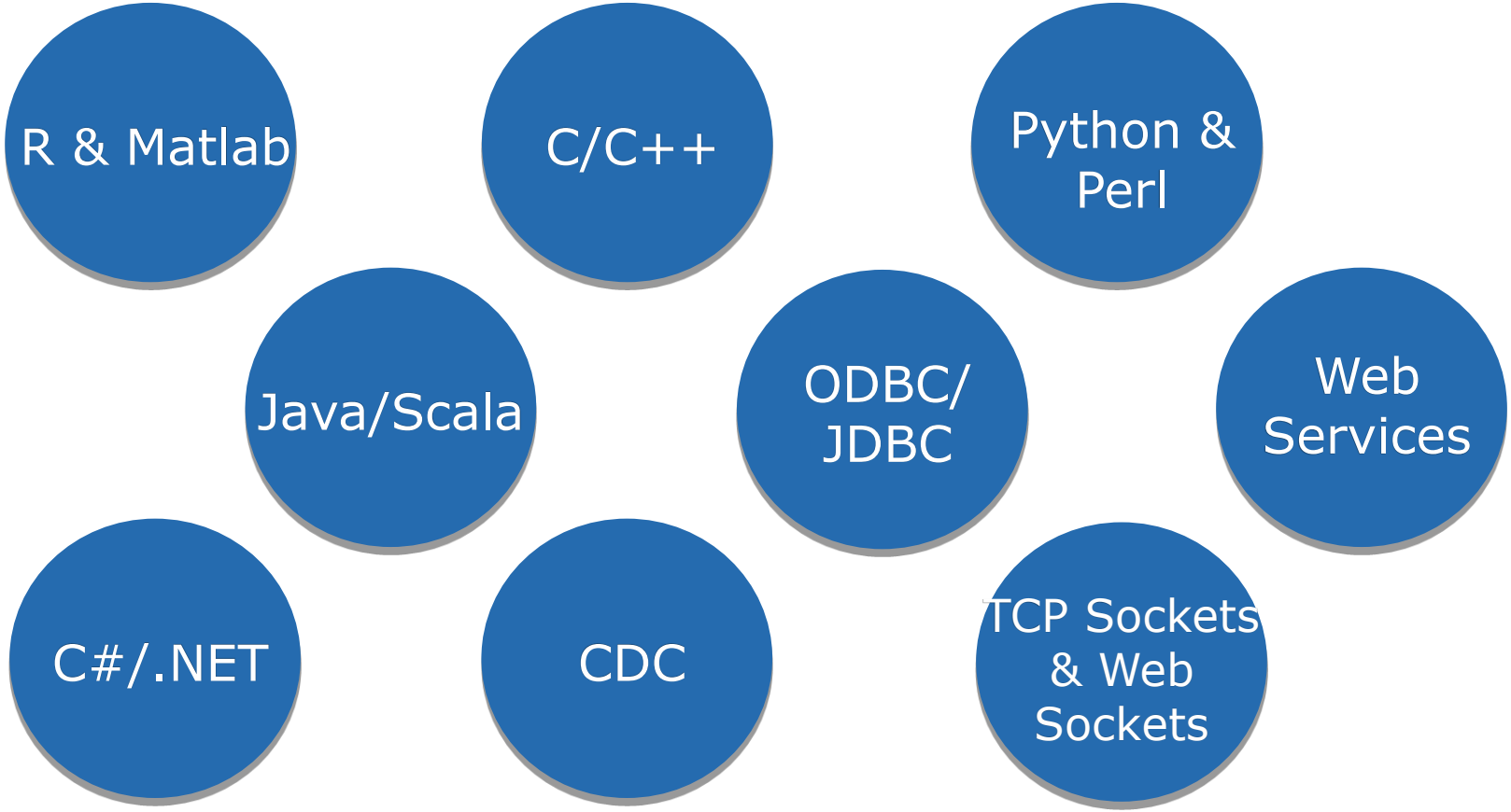
Query Rate: Kdb+ vs InfluxDB vs MongoDB

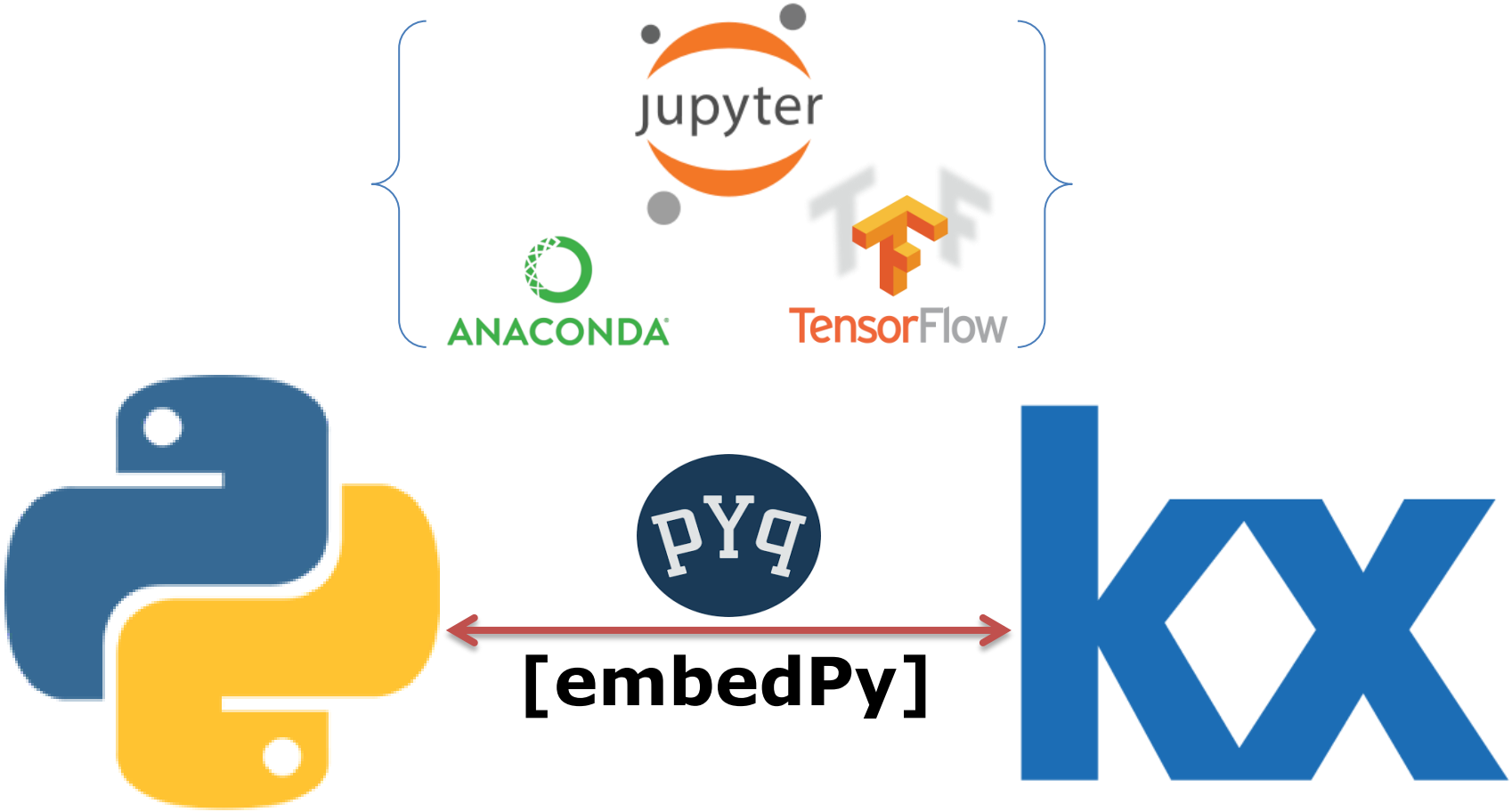


On similar servers, kdb+ is over **40 times faster** than InfluxDB and MongoDB

I DREAMED
I WAS KAFKA
memes.com

Enterprise Interfaces




jupyter
ANACONDA
TensorFlow
pyq
[embedPy]
kx

Kdb+ and Python Integration

PyQ

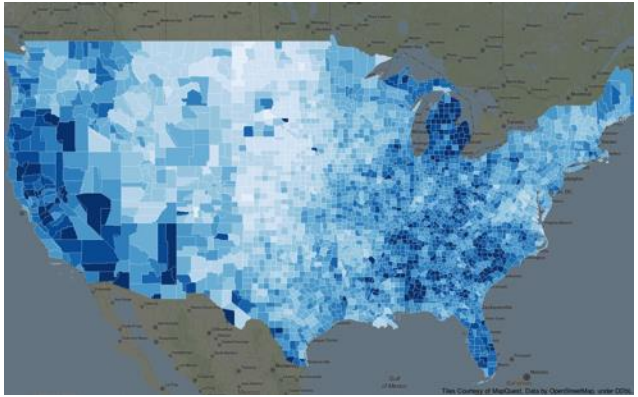
- In PyQ, Python and q objects live in the same memory space and share the same data
- This is achieved by bringing the Python and q interpreters into the same process, so that code written in either of the languages operates on the same data

embedPy

- The interface allows execution of Python code directly in a q console or from a script. embedPy, embeds python functions/data/modules directly into a q process

jupyterQ

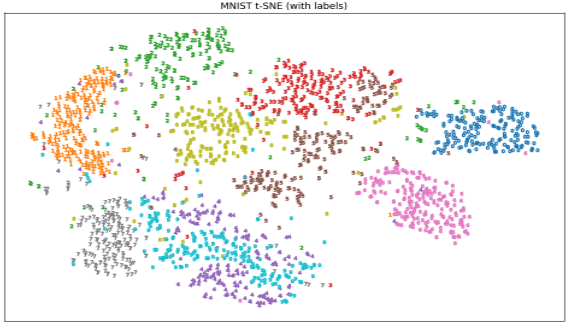
- jupyterQ enables users to utilise the power of Jupyter Notebooks for visual inspection and interpretation of Q and Python. Documents are both human-readable documents containing the analysis description and the results (figures, tables, etc.) as well as executable documents which can be run to perform data analysis



Visualise t-SNE (checking with labels)

```
In [16]: subplots:plt[`subplots;*][]

         fig:subplots[`$"__getitem__";*][0]
         fig[`set_figheight;*;8];
         fig[`set_figwidth;*;12];

         axarr:subplots[`$"__getitem__";*][1]
         axarr[`set_title;*;"MNIST t-SNE (with labels)"];
         {[x]where testLabelsSample=x;axarr[`scatter;*;resTSNE[1;0];resTSNE[1;1];`c pykw 25;`c pykw "C",string x;`marker pykw "$",string
         [x],"$"]}each distinct testLabelsSample;
         plt[`:;()]each`xticks`yticks;
         plt[`show;*][];
```

Out[16]:



PyQ

- Python/q **share the same memory space.**
- Makes q more accessible to others.
- Code can be developed in either python or q.
- Both languages can operate on the same data.

Same
Working
Environment

Results

PyQ - Basic Example

```
>>> q()
```

```
q)trade:([]date:();sym:();qty:())
```

```
q)\
```

```
>>> q.insert('trade', (date(2006,10,6), 'IBM', 200))
k(',0')
>>> q.insert('trade', (date(2006,10,6), 'MSFT', 100))
k(',1')
```

```
>>> q.trade.show()
date       sym  qty
------------------
2006.10.06 IBM  200
2006.10.06 MSFT 100
```

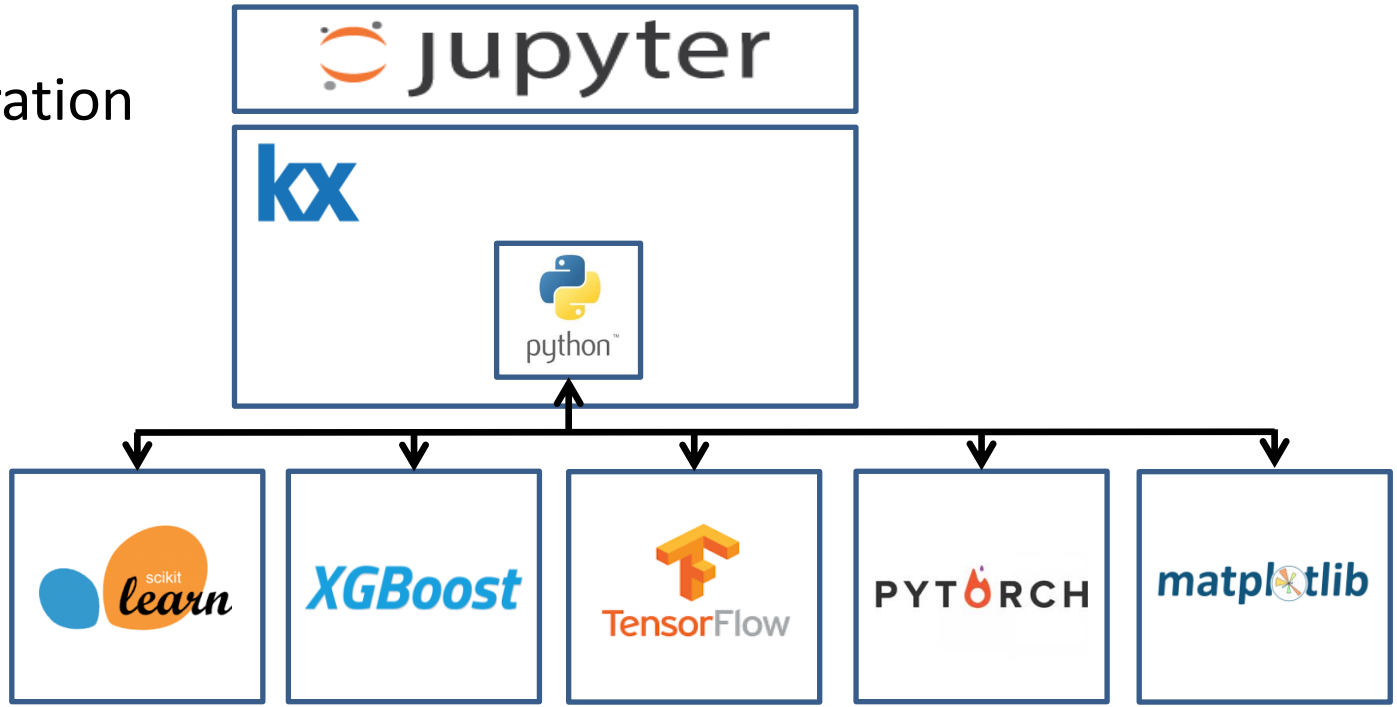
Define a function in q:

```
q)f:{[s;d]select from trade where sym=s,date=d}
```

Call the `q` function from Python and pretty-print the result:

```
>>> x = q.f('IBM', date(2006,10,6))
>>> x.show()
date       sym qty
------------------
2006.10.06 IBM 200
```

- Python ***inside*** q
- Access to world of Machine Learning
- Visualizations
- Jupyter integration



Case Study: embedPy and Machine Learning – The Mission

Precision Medicine & Genetic Testing are relatively untapped resources of massive potential – Machine Learning can change this.

The Challenge:

- Cancer tumours can have thousands of genetic mutations.
- Distinguishing the mutations that contribute to tumour growth is a real challenge!
- **The mutations are categorized from one to nine by expert oncologists.**
- Very Laborious.
- The ability to automate this process could revolutionize the field of oncology.
- **The application of Kx Technology could change lives.**

Case Study: embedPy and Machine Learning – The Data

The data is presented as massive text files of clinical records.
This is very laborious and time consuming work…

```
ID   | 1290i
Text| `We report a 10-year-old girl presenting with severe neonatal hypertrophic cardiomyopathy (HCM), feeding difficulties, mi
ldly abnormal facial features, and progressive skeletal muscle symptoms but with normal cognitive development. Targeted oligonu
cleotide-selective sequencing of 101 cardiomyopathy genes revealed the genetic diagnosis, and the mutation was verified by Sang
er sequencing in the patient and her parents. To offer insights into the potential mechanism of patient mutation, protein struc
tural analysis was performed using the resolved structure of human activated HRAS protein with bound GTP analogue (PDB id 5P21)
in Discovery Studio 4.5 (Dassault Systèmes Biovia, San Diego, CA). The patient with hypertrophic cardiomyopathy and normal cogn
itive development was diag                                                    syndrome affecting a high
ly conserved amino acid, t                                                    ysis rendering HRAS unresp
onsive to GTPase activati                                                     likely affects binding of
guanidine-nucleotide-excha                                                    inactivation. Patients wit
h the HRAS mutation c.173(                                                    th other mutations causin
g Costello syndrome. We ex                                                    ng the first case in Euro
pe and the fourth case in                                                     hanism of the mildly activ
ating p.T58I mutation.
```

The training and testing database have 3321 and 5668 samples respectively.

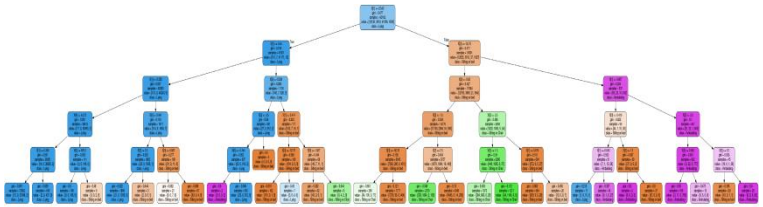
The power of kdb+ and embedPy can change this!

Case Study: embedPy and Machine Learning – The Method

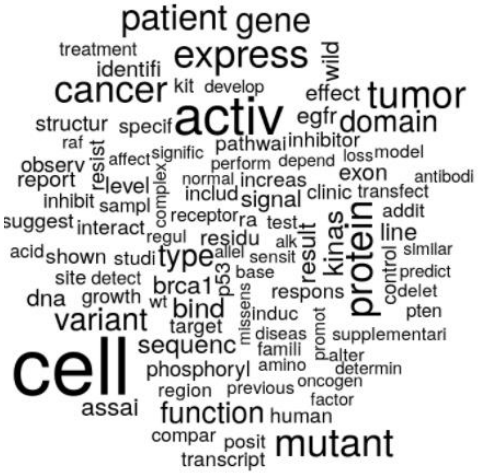
3. The Algorithm - Random Forest Classifier

It is important to note before diving into the Random Forest classifier that there are a plethora of methods to produce accurate results. Gradient Boosting Machine (GBM), Generalized Linear Model (GLM) & Deep Learning (DL) have all been used to great effect in this challenge thus far. An equally accurate algorithm that is easy to follow is the random forest.

[28]:
```
//The line below imports from the python code required to create a pipeline, allowing us to perform multiple tasks sequentially
pipeline:.p.import[`sklearn.pipeline;`:Pipeline]

//countvectorizer is imported from python again here
countvectorizer:.p.import[`sklearn.feature_extraction.text;`:CountVectorizer]

//The Random Forest Classifier code is imported (once again from the SKlearn Toolkit) here
randomforest:.p.import[`sklearn.ensemble;`:RandomForestClassifier]

//Define rf to be the Random Forest without any parameters yet
rf:randomforest[]

//Define the pipeline
classifier:pipeline[((`vect,countvectorizer[]`);(`clf,randomforest[]`))]
```



The **text** holds the *key* to the classification problem

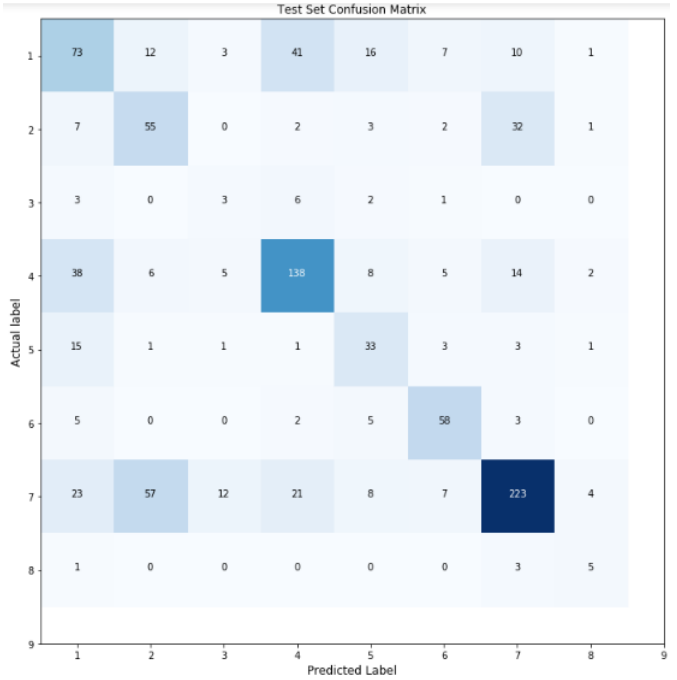


Case Study: embedPy and Machine Learning – The Result

Using Kx Technology, a Random Forrest ML algorithm produced:

- Predictions for *every record* in **30 seconds.**
- Had an overall accuracy of 63%

ID	Gene	Variation	Pred	Text
0	ACSL4	R570S	7	2. This mutation resulted in a myeloproliferat..
1	NAGLU	P521L	1	Abstract The Large Tumor Suppressor 1 (LATS1) ..
2	PAH	L333F	7	Vascular endothelial growth factor receptor (V..
3	ING1	A148D	7	Inflammatory myofibroblastic tumor (IMT) is a ..
4	TMEM216	G77A	1	Abstract Retinoblastoma is a pediatric retinal..
5	CD40LG	A123E	4	The accurate determination of perfluoroalkyl s..
6	KLF11	T220M	7	Aberrations in the mTOR (mechanistic target of..
7	SGCB	T151R	4	Oncogenic mutations in the monomeric Casitas B..
8	CLCF1	R197L	2	NPM1 gene at chromosome 5q35 is involved in re..
9	SDHAF1	R55P	7	Introduction The epidermal growth factor recep..
10	SPTLC2	I504F	7	The protein kinase B-RAF is mutated in approxi..
11	SUMF1	A348P	7	Genes of the Raf family encode kinases that ar..
12	TET2	Y1902A	6	TET proteins oxidize 5-methylcytosine (5mC) on..
13	G6PD	D312H	7	The p110δ subunit of phosphoinositide 3-kinas..
14	SNCB	P123H	7	Abstract Purpose: Positron emission tomograph..



Additional Resources

Resources - Benchmarks

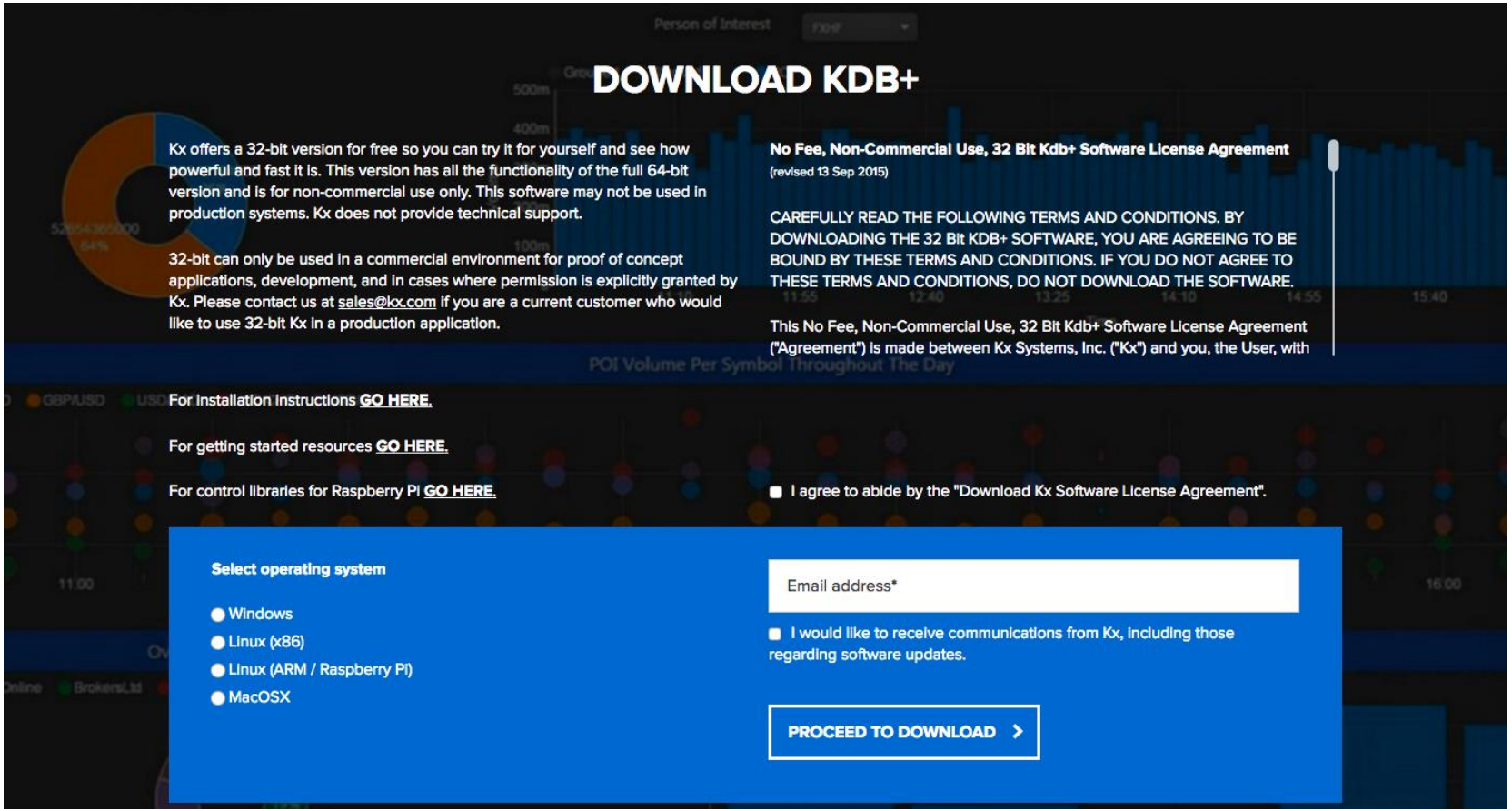
TM

1.1 Billion Taxi Rides on kdb+/q & 4 Xeon Phi CPUs

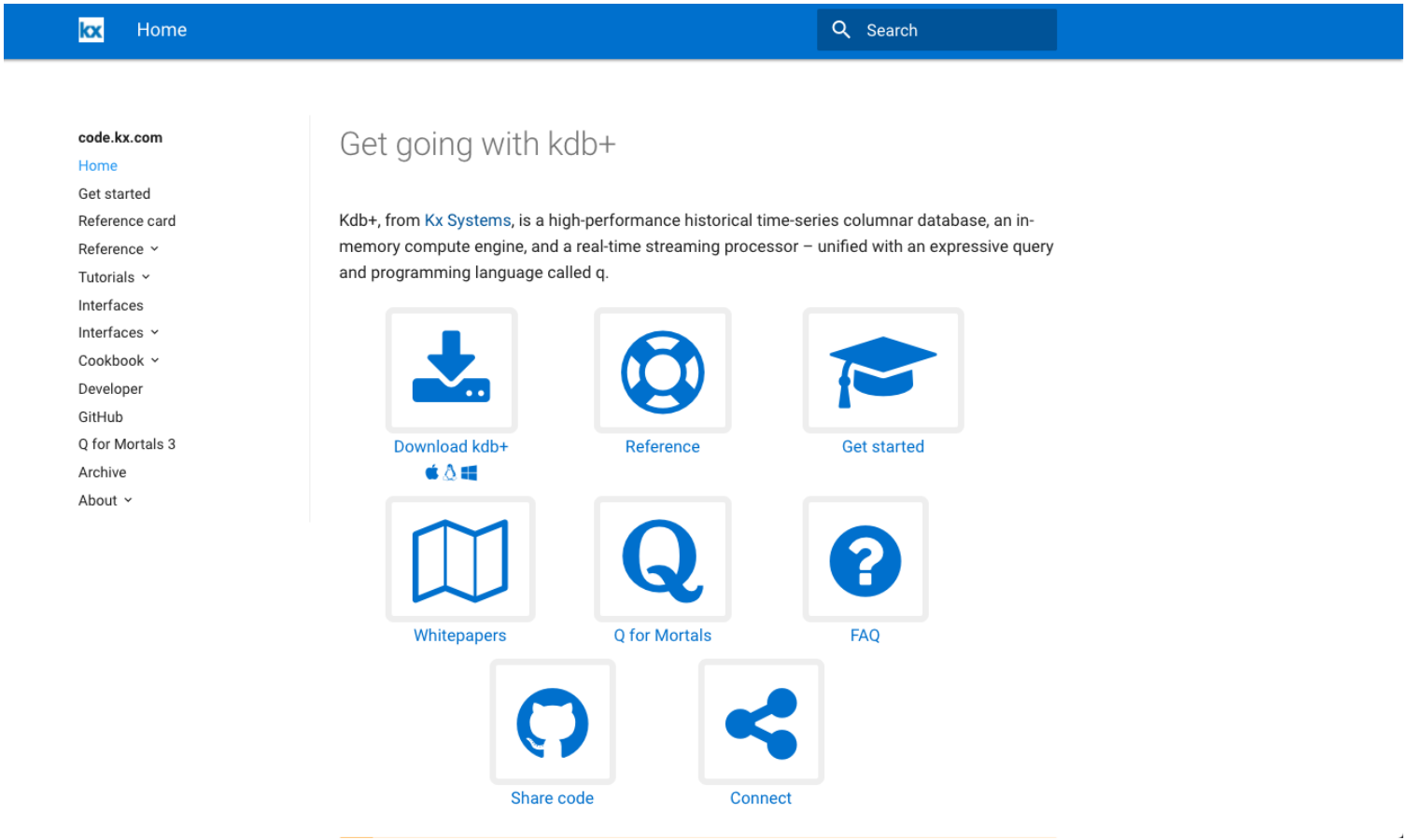
Q is a programming language with a built-in, column-oriented, in-memory and on-disk database called kdb+. Q both includes and extends SQL. Q is native to the database engine so, unlike most databases, there is no shipping of data between the client and the server.

Q and kdb+ are shipped as a single binary that has a small memory footprint and is capable of running in the L2 and L3 caches of modern CPUs making the system very performant. Its tables can be stored on a local disk or distributed but will nonetheless appear as a single table. Table data can be partitioned and segmented into memory mapped files which helps remove I/O bottlenecks.

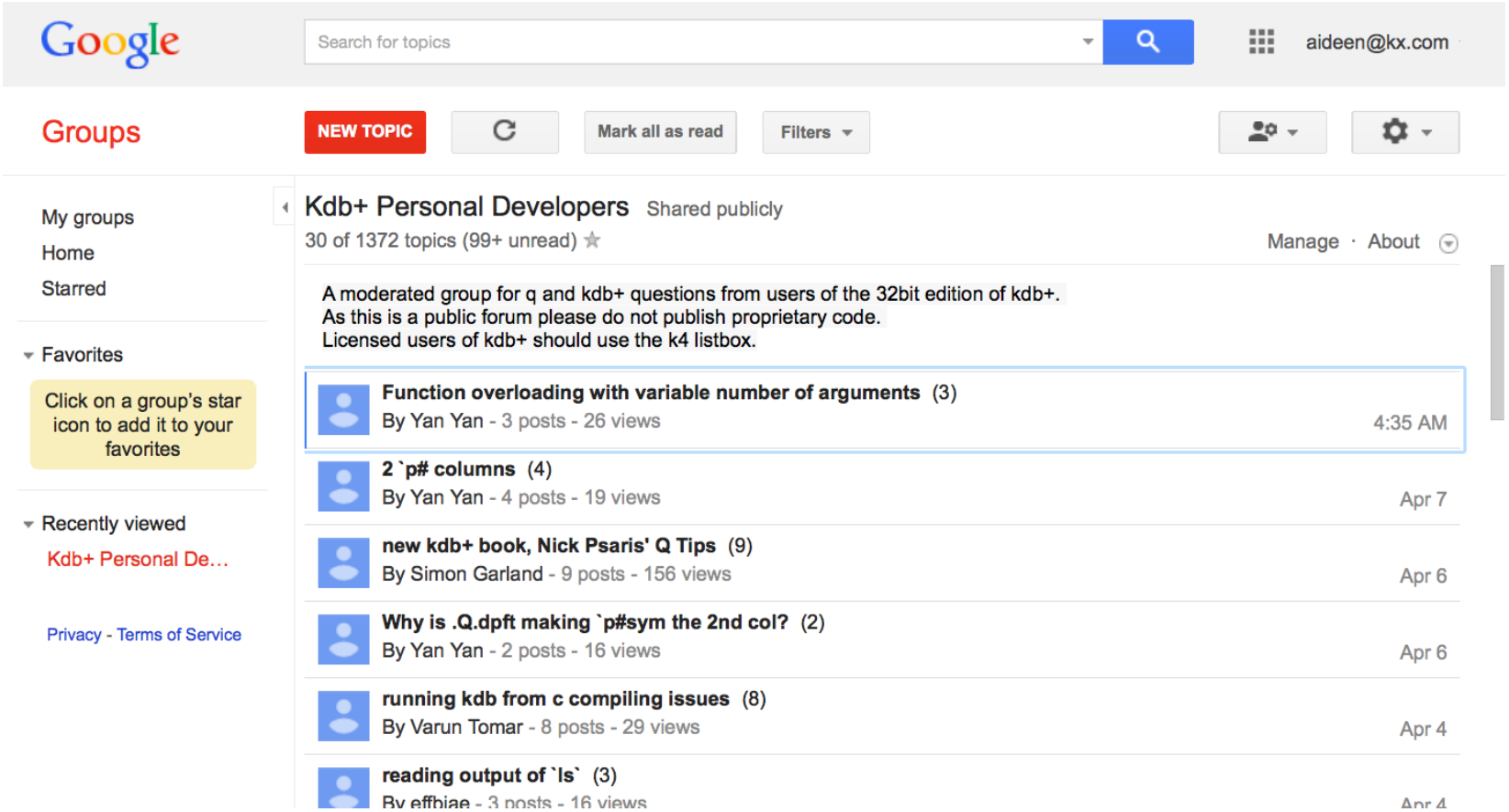
Resources - Free version



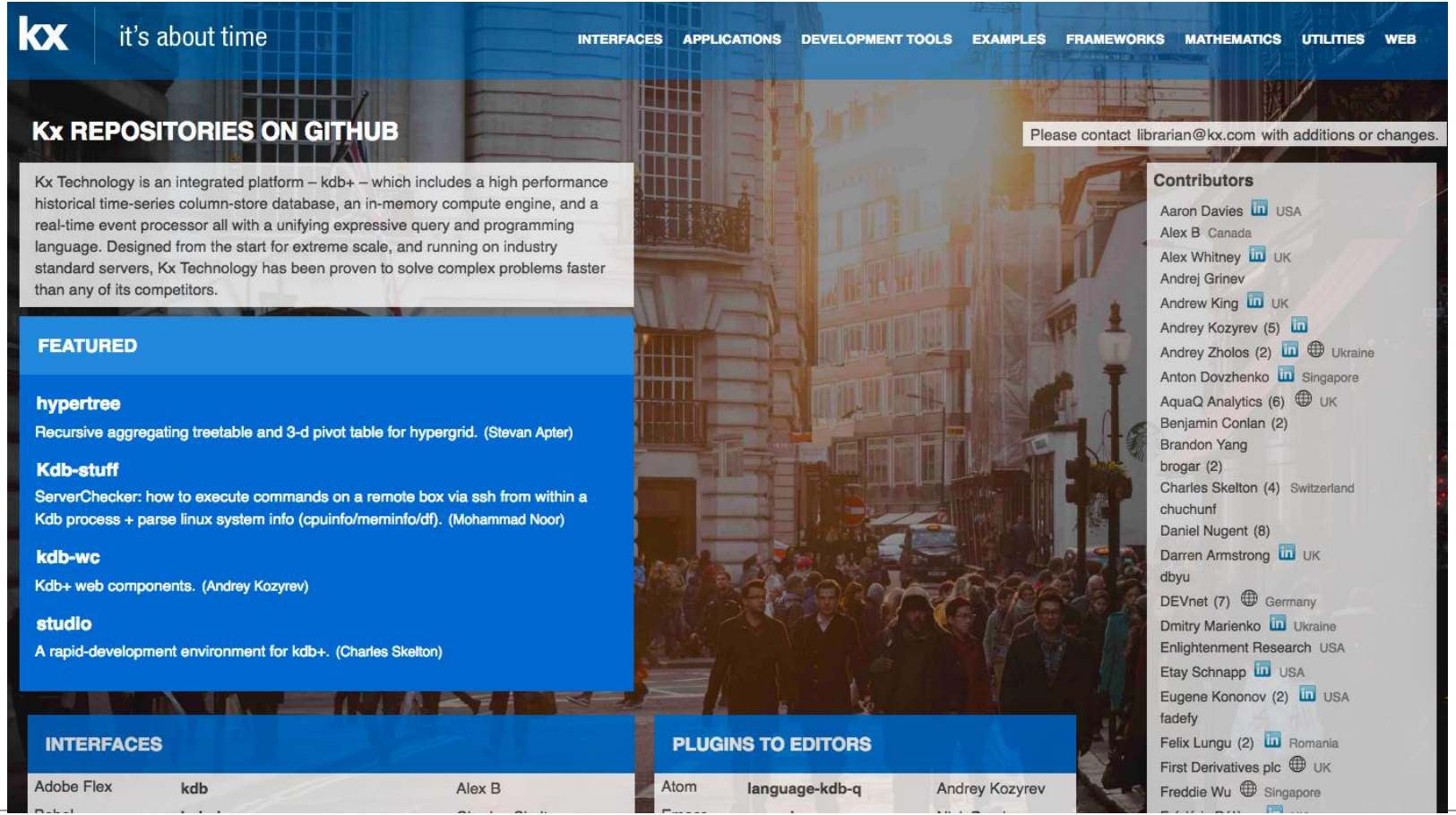
Resources – Wiki site



Resources – Google Group



Resources – Github Repository



Resources – Meetups

Kx

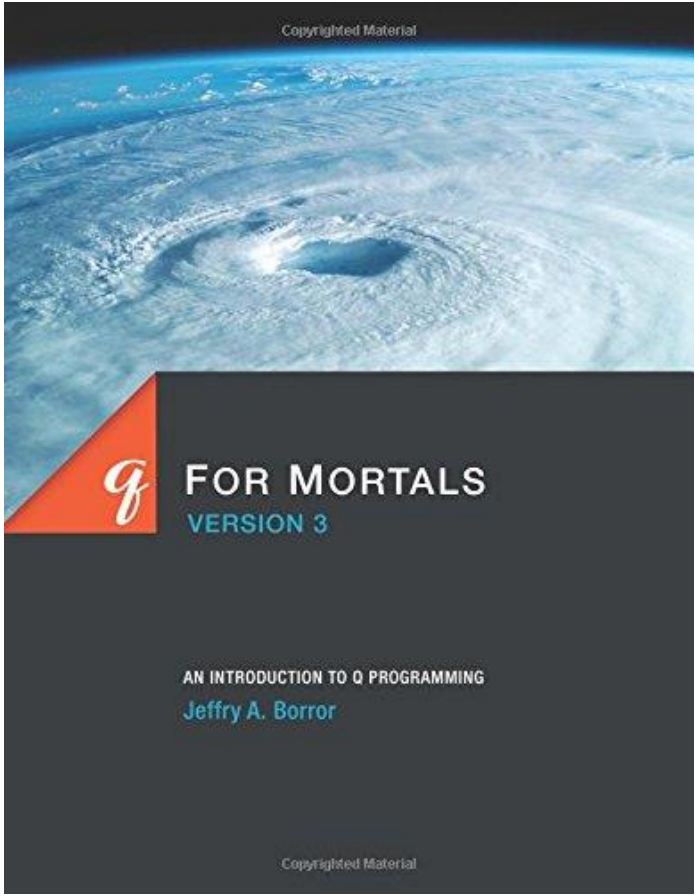
Find out what's happening in Kx Meetup groups around the world and start meeting up with the ones near you.

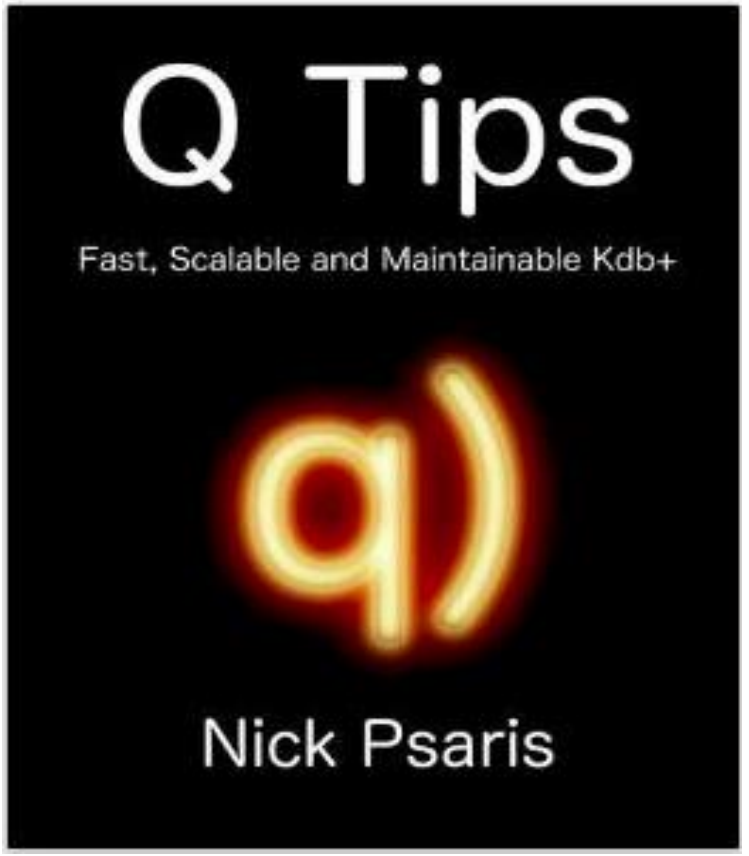
12,833 members | **131** Meetups

Join Kx Meetups

Related topics: kdb+ · Big Data · Data Visualization · Big Data Analytics · Time Series Database · Data Analytics · Internet of Things · Predictive Analytics · NoSQL · Data Science

Resources – Books





Kx TECHNOLOGY FUND

Learn more at www.kx.com/kx-technology-fund/

Recruitment

- Opportunities to travel and live internationally
- Paid for city accommodation
- Competitive salary and expenses package
- Health and travel insurance programme
- Generous retirement savings programme

Interested in joining us or know someone who might be?

Talk to me at the end or visit

www.firstderivatives.com/careers

Thank You

Jonathan Guy

Email: jguy@kx.com

https://kx.com/

https://kx.com/download/

Kx® and kdb+ are registered trademarks of Kx Systems, Inc., a subsidiary of First Derivatives plc