influxdata®

AN INFLUXDATA TECHNICAL PAPER

Why Time Series Matters for Metrics, Real-Time Analytics and Sensor Data

PAUL DIX, CTO AND FOUNDER, INFLUXDATA
REVISION 5, JULY 2021

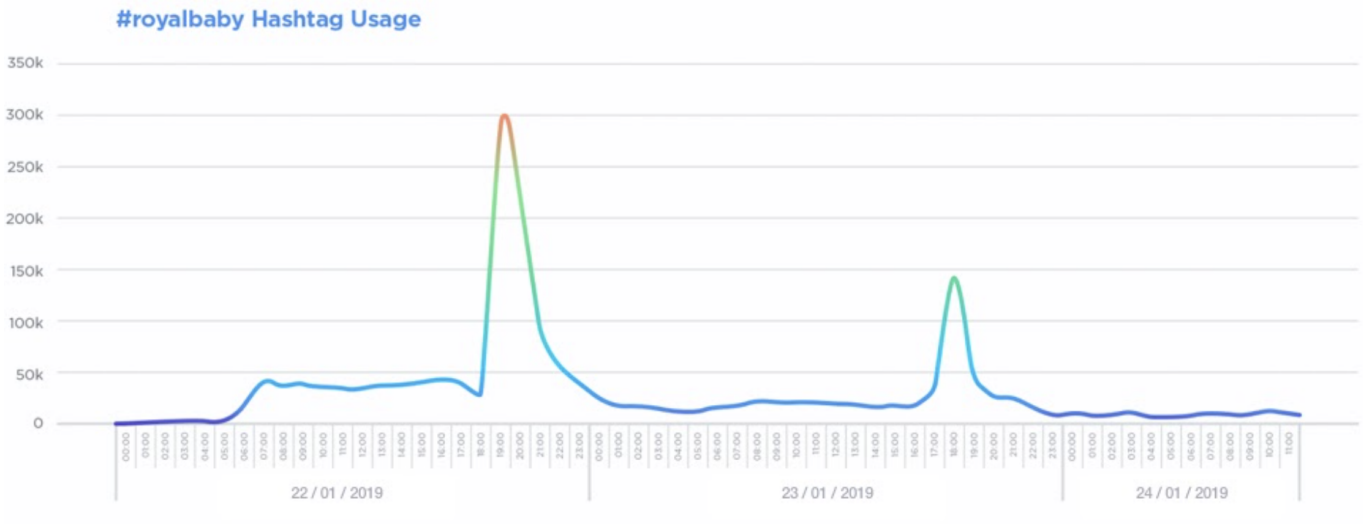
Introduction

Time series data has historically been associated with applications in finance. However, as developers and businesses move to instrument more of their servers, applications, network infrastructure and the physical world, time series is becoming the de facto standard for how to think about storing, retrieving and mining this data for real-time and historical insight. This paper will:

- Define what time series data is (and what it isn't)
- Explain how the time series data domain differs from more traditional data workloads like OLTP or full-text search
- Examine what makes the InfluxData platform different from other proposed solutions

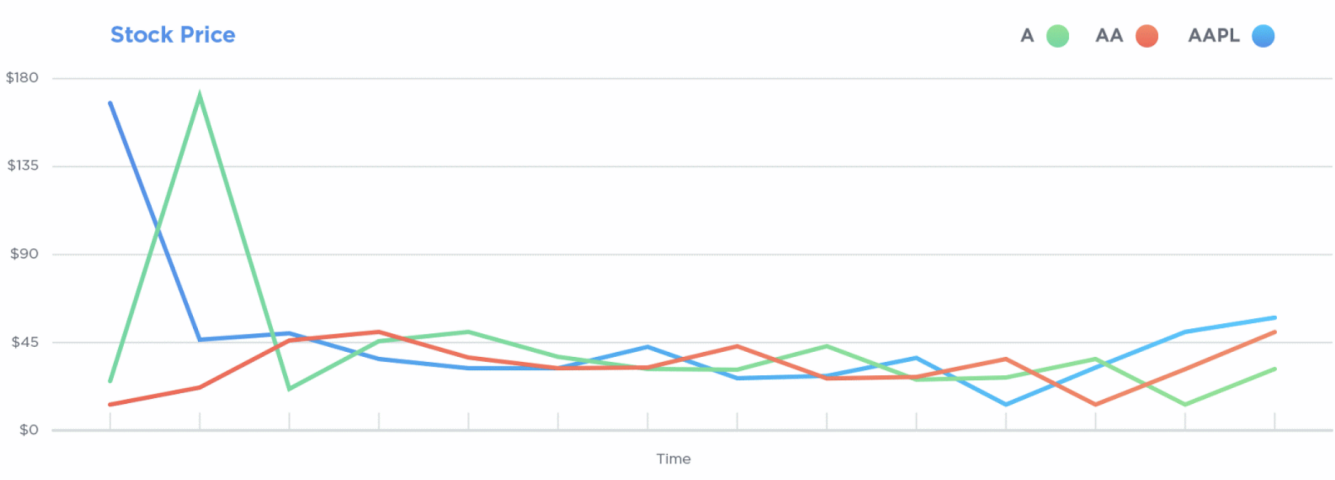
What is time series data?

Time series are simply measurements or events that are tracked, monitored, downsampled and aggregated over time. This could be server metrics, application performance monitoring, network data, sensor data, events, clicks, trades in a market and many other types of analytical data. The key difference that separates time series data from regular data is that you're always asking questions about it over time. A simple way to determine if the dataset you are working with is time series or not is to see if one of your axes is time. Below are a few examples of time series data plotted on graphs:



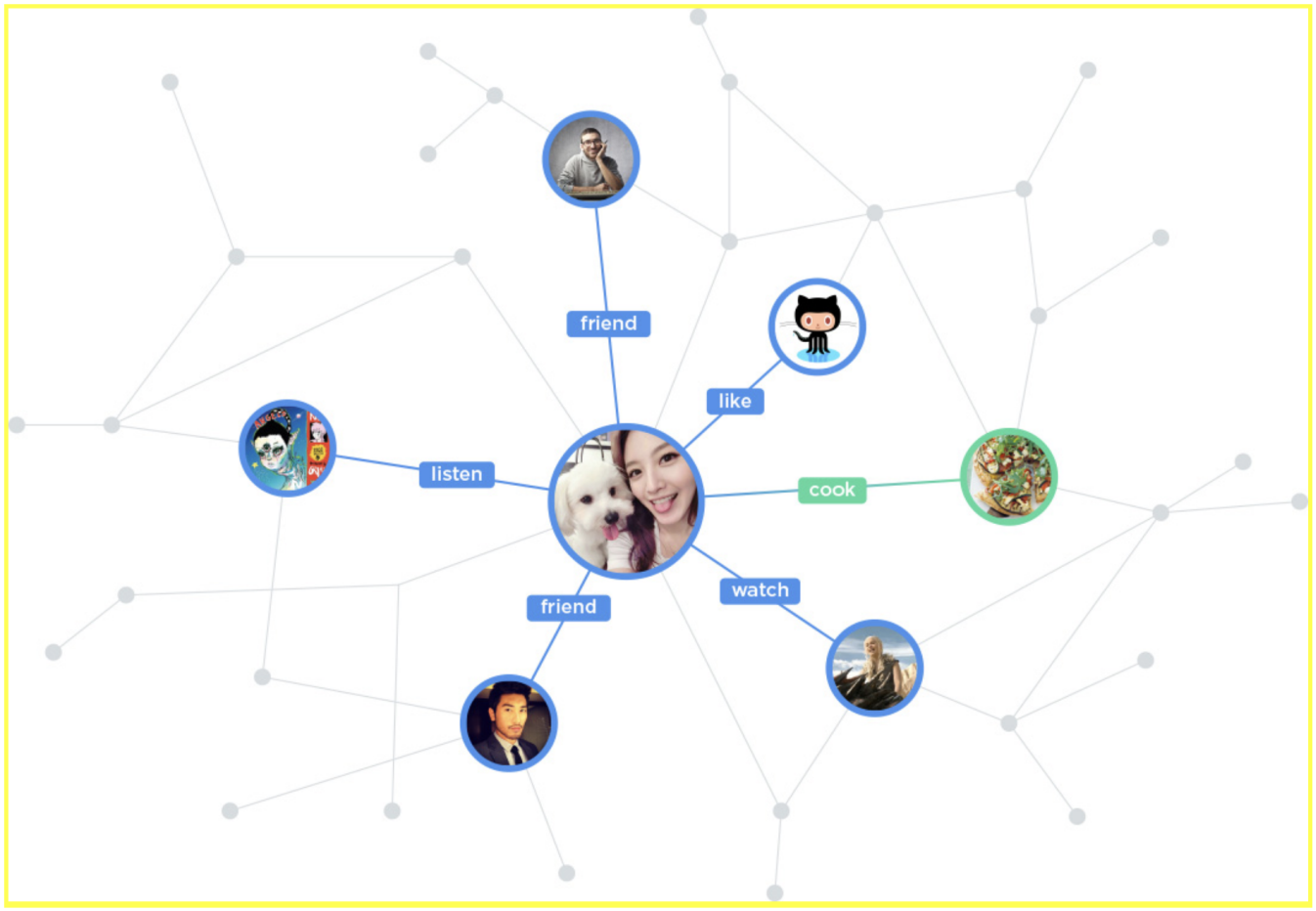
Time Series Example: Hashtag Frequency Over Time

Time-stamped data: Data collected at different points in time

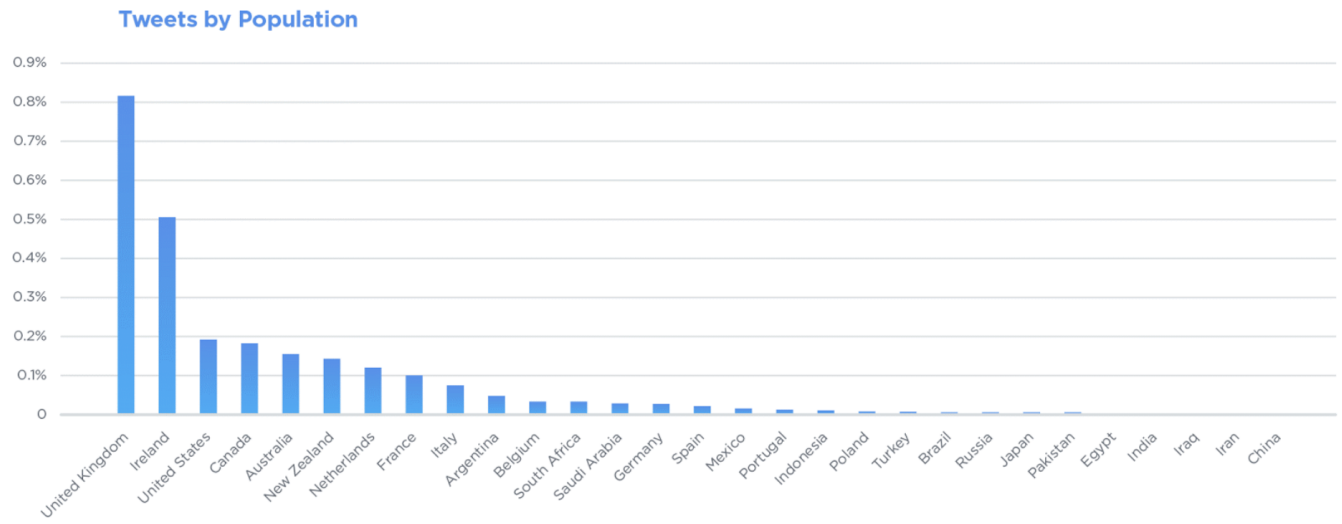


Time series example: stock ticker prices over time

Now, a few examples that are NOT time series data, plotted on graphs:



NOT a time series example: graph of relationships



NOT a time series example: tweets by population

Since time is a constituent of everything that is observable, time series data is everywhere. As our world gets increasingly instrumented, sensors and systems are constantly emitting a relentless stream of time series data. Such data has applications across various industries, such as tracking weather data; changes in application performance; medical device data; network logs; and many others. Time series data is used in time series analysis and time series forecasting to detect and predict patterns — essentially looking at change over time.

Time series data is immutable; i.e., since time series data comes in time order, it is almost always recorded in a new entry, and as such, should be append-only (appended to the existing data).

Time series data comes in two forms: regular and irregular. Regular time series consist of measurements gathered from software or hardware sensors at regular intervals of time (every 10 seconds, for example) and are often referred to as metrics. Irregular time series are events driven either by users or other external events.

Summarizations of irregular time series become regular themselves. For example, summarizing the average response time for requests in an application over one minute intervals or showing the average trade price of Apple stock every 10 minutes over the course of a day.

The InfluxData stack is optimized for both use cases, which is a significant differentiator from other solutions like Graphite, RRD, OpenTSDB, or Prometheus. Many services and time series databases support only the regular time series metrics use case. InfluxDB lets users collect from multiple and diverse sources, store, query, process and visualize raw high-precision data in addition to the aggregated and downsampled data. This makes InfluxDB a viable choice for applications in science and sensors that require storing raw data.

The InfluxDB platform organizes time series in a structured format. At the top level is a measurement name, followed by a set of key/value pairs called tags that describe the metadata, followed by key/value pairs of the actual values called fields. Field values in InfluxDB can be boolean, int64, float64 or strings. Finally, there is a timestamp for the set of values. All data is queried by the measurement, tags, and field along with the time range.

This structure makes it easy for developers to build tools around the APIs that InfluxData provides. Unlike relational or document databases, InfluxDB organizes time series data into a set structure. This structure is also what sets InfluxDB apart from other solutions. The richness of the time series data structures that can be represented open InfluxDB up to more time series and metrics use cases, while also widening the performance lead that InfluxDB has over other solutions. With the right schema and setup, a single InfluxDB server can handle over 2,000,000 writes per second, something the competition is unable to match.

Horizontal use case

In recent years, time series has become a common use case across many industries and a database category of its own. Metrics, events and other time-based data are being generated at an exponential rate, as there is a growing requirement for analyzing today's complex environments. The InfluxData platform provides a comprehensive set of tools and services to collect and accumulate metrics and events data, analyze the data, and act on the data via powerful visualizations and notifications.

Whether the data comes from humans, sensors or machines, InfluxData empowers developers to build next-generation monitoring, analytics, and IoT applications faster, easier, and to scale delivering real business value quickly.

InfluxData has customers and users that span three primary use cases: DevOps monitoring, real-time analytics, and IoT monitoring. Anyone who has sensors, servers, VMs, containers, applications, users or events to track could benefit from using InfluxData.

DevOps monitoring

Real-time visibility across all systems, applications and infrastructure

For our users in the DevOps monitoring and metrics space, most are in medium to large organizations. Some of these users are building custom monitoring solutions from scratch, deploying the InfluxDB platform to first track their servers, VMs, containers, data stores, and network hardware and later as a generalized metrics platform for application developers within the organization. Others are using InfluxData to supplement commercial APM products to instrument aspects of the InfluxDB platform for which no probes or agents exist, or to stitch together metrics from multiple monitoring solutions. And in both cases, users are not only gathering metrics to determine a baseline about the health of their systems but also using InfluxDB as the place to accumulate their log events. Having access to log data is secondary and an important contextual source to help further triage and resolve issues.

Real-time analytics

Real-time process and performance data to make decisions on the fly

We see organizations of all sizes working with real-time analytics. Some are building applications that will face their users with InfluxDB as the underlying database while others are instrumenting business, social or development metrics in real-time for internal consumption. We frequently see users start with the DevOps custom monitoring and metrics use case, who then move into real-time analytics once the platform is deployed. InfluxDB eventually becomes the central store for all time series, sensor and analytics data.

Internet of Things (IoT)

Insights from sensor data to enable automation, predictive maintenance and innovation

We have also found that there are a number of IoT use cases. We've seen users in industrial settings like factories, oil and gas as well as renewable energy plants, agriculture, smart homes, roads and infrastructure. There are also users in consumer grade IoT-like wearables, consumer devices and trackers.

Solutions built on the InfluxData platform

For organizations around the world, in nearly every industry, InfluxDB has become the system of insight for unified metrics and events – enabling the most demanding SLAs and providing a foundation for solutions such as Application Performance Monitoring (APM), Google Cloud Monitoring, Industrial IoT, Kubernetes Monitoring, Metrics as a Service, Network Performance Monitoring, and Stream Processing. Here's a brief overview of each.

InfluxDB for application performance monitoring

In a digital economy where complexity is a given and a responsive application is a requirement, visibility into your entire application has become a necessity for enterprises. APM can be performed using InfluxData's platform InfluxDB. InfluxDB is a purpose-built time series database, real-time analytics engine and visualization pane. It is a central platform where all metrics, events, logs and tracing data can be integrated and centrally monitored. InfluxDB also comes built-in with Flux: a scripting and query language for complex operations across measurements. Learn more about performing APM using InfluxDB.

The InfluxDB Google Cloud monitoring solution

The presence of InfluxDB Cloud on GCP means that customers have ready-to-use access to the industry's leading time series and data analytics platform for real-time decision making. Google Cloud Monitoring with InfluxDB provides visibility into the performance, uptime, and overall health of your Google Cloud-powered applications, cost-effectively and at cloud scale. With InfluxDB Cloud on GCP, customers can address a wide range of use cases. InfluxDB Cloud as a part of the Google Cloud marketplace brings the unification of procurement, billing and support along with the entire Google Cloud menu of services.

InfluxDB for Industrial IoT monitoring

The industrial world has a long history of modernizing processes in order to keep production running efficiently and safely while minimizing downtime. Yet many are locked in established data historian solutions that are costly and lack the methods needed to provide innovation and interoperability. In contrast, InfluxDB — the open source time series database — inherently provides diverse design perspectives not available from a single software vendor. It provides the freedom to integrate with other solutions and allows you to adapt the code to fit your ever-changing system requirements. This is why many industrial enterprises around the world are choosing InfluxDB for IIoT monitoring.

InfluxDB for Kubernetes monitoring

Kubernetes orchestration provides built-in fault tolerance, automating scaling and maintenance for a desired cluster state. However, visibility must come with the necessary granularity and information for fast identification of the source of trouble. Monitoring and accountability are what make automation reliable. InfluxDB helps to identify and resolve problems before they affect critical processes, and most importantly, offers ways to implement Kubernetes monitoring that accommodates developers' need for instrumentation without overloading IT operations.

InfluxDB for Metrics as a Service

Metrics as a Service is a concept that combines centralization resource efficiency, overhead offload and maximum value extraction from the data collected organization-wide. All data converging to one platform, available in one pane, allows much richer visualization and analytics across multiple data sources and data types. InfluxDB has become the preferred platform for centralized monitoring of metrics, events and key business indicators, for metrics as a service.

InfluxDB for network performance monitoring

When network speed degrades or connectivity fails, the data flow sustaining applications and IT operations will struggle or halt along with the network. Networks — the lifeline of IT infrastructure — are dynamic environments. They require monitoring to deliver consistent, predictable network performance. Dealing with so much monitoring data, it can be easy to be consumed by it. However, there is a way to effectively manage your IT infrastructure, by centralizing, analyzing, and automating it. InfluxData enables you to do this with its network monitoring tools. Its collection agent Telegraf, with 200+ plugins, supports protocols such as ICMP/Ping, SNMP, NetFlow, SFlow, and Syslog. InfluxDB, for its part, contains a powerful query engine for processing multiple data sources in real-time. To gain the necessary visibility in the status, performance and responsiveness of all network devices in their enterprise, cloud or hybrid application environments, enterprises are deploying the InfluxData platform for network performance monitoring.

InfluxDB for stream processing

Stream processing is the processing of data in motion. It unifies applications and analytics by processing data as it arrives, in real-time, and detects conditions within a short period of time from when data is received. The key strength of stream processing is that it can provide insights faster, often within milliseconds to seconds. Stream processing naturally fits with time series data, as most continuous data series are time series data. And time series data needs a purpose-built database to ingest, store and process it. This is exactly what InfluxDB is. And this is why, given its high-write throughput and the scalability it allows, InfluxDB suits stream processing.

The time series workload

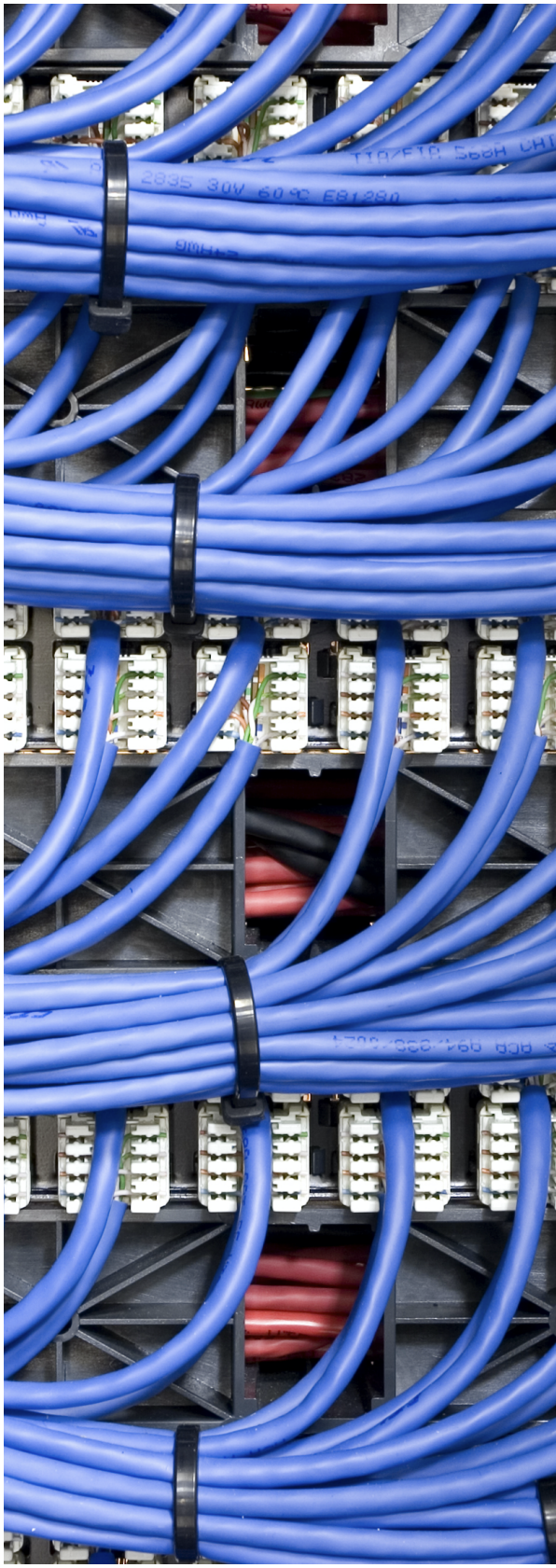
Time series data has a few properties that make it very different from other data workloads. Data lifecycle management, summarization and large range scans of many records are what separate time series from other database use cases.

With time series, it's common to request a summary from a larger period of time. This requires going over a range of data points to perform some computation like a percentile to get a summary of the underlying series to the user. This kind of workload is very difficult to optimize for a distributed key value store. InfluxDB is optimized for exactly this use case, giving millisecond level query times over months of data.

With time series, it's also common to keep high-precision data around for a short period of time. This data is aggregated and downsampled into longer-term, trend data. This means that for every data point that goes into the database, it will have to be deleted after its period of time is up.

This kind of data lifecycle management is difficult for application developers to implement on top of the typical database. They must devise schemes for cheaply evicting large sets of data and constantly summarizing that data at scale.

InfluxDB is designed as a time series database with solutions built-in for summarization and data lifecycle management at a large scale. These come out of the box with no application level code required from the developer. To learn more, see the continuous queries and retention policies documentation.



What makes InfluxData different?

We often get asked questions about what makes InfluxData different from other technology solutions. Generally these can be organized into three categories: applications, databases, and stream processing systems.

In the next few sections we'll look at these and Elasticsearch specifically since there is more overlap with their ELK stack than pure databases. So, let's start with a look at the InfluxDB platform and InfluxDB's data model, which is a key differentiator.

The InfluxDB Platform

InfluxDB is the essential time series toolkit — offering everything you need in a time series data platform in a single binary: a multi-tenanted time series database, UI and dashboarding tools, background processing and monitoring agent. All this makes deployment and setup a breeze and easier to secure.

influxdb™ — **InfluxDB** is the database and storage tier.

telegraf™ — **Telegraf** is InfluxDB's native plugin-driven metrics collection agent and has 200+ plugins that integrate with other products.

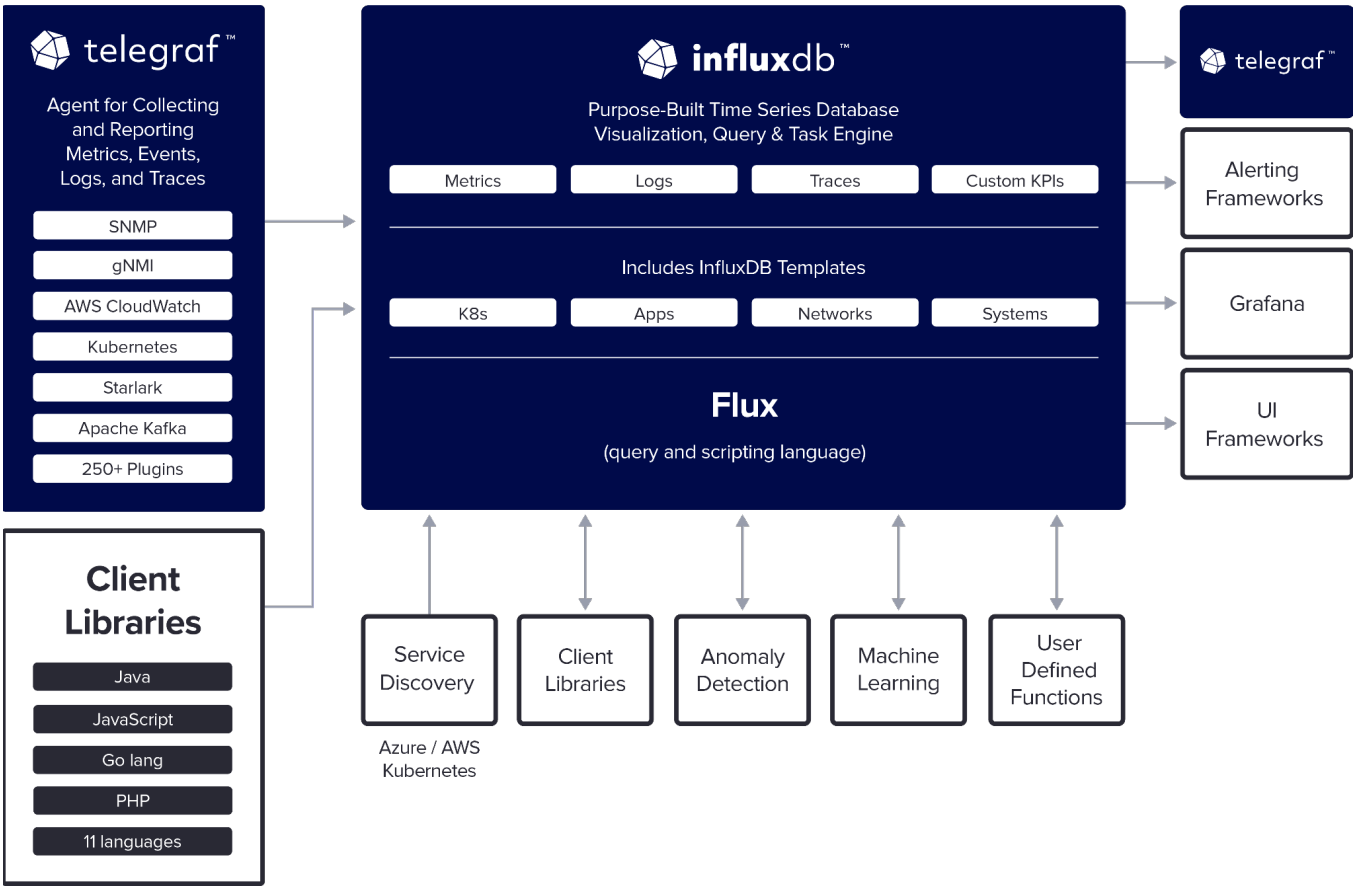
influxdb™ — **InfluxDB client libraries** allow you to ingest and query data in your favorite programming language.

flux™ — **Flux** is a fourth-generation programming language, built into InfluxDB, that is designed for data scripting, ETL, monitoring and alerting. As Flux is a functional language, you can structure queries and separate common logic into functions and libraries that are easily shared and help speed development. Flux can also be used to enrich time series data with other SQL data stores (Postgres, Microsoft SQL Server, SQLite, SAP Hana) along with cloud-based data stores (Google Bigtable, Amazon Athena, and Snowflake). Enriching time series data provides context that can provide further insights into your data.

InfluxDB is optimized for developer productivity. Everything in InfluxDB — ingest, query, storage and visualization — is accessible in a unified API. This enables faster time to awesome for developers because everything in the platform can now be programmatically accessed and controlled. This is combined with a powerful set of client libraries across 11 languages (like Go, Java, PHP and Python). A set of InfluxDB command line tools helps developers develop in a way that is most familiar to them.

InfluxDB features a best-in-class UI that includes a Data Explorer, dashboarding tools, and a script editor. Use the Data Explorer to quickly browse through the metric and event data you collected and apply common transformations. The Dashboarding tool comes with a handy list of visualizations that help you to see insights from your data faster. And finally, use the script editor to quickly learn Flux with easily accessible examples, auto-completion and real-time syntax checking.

InfluxDB Templates — a new set of tools that includes a packager and a set of pre-made monitoring solutions — allow you to share your monitoring expertise with coworkers and other community members around the world. The InfluxDB Templates gallery features available templates covering some of the most popular tools, applications, and protocols. These templates can also be checked in as code, fitting in with your continuous integration and deployment pipelines to make deploying (and more importantly rolling back) changes painless.



To further support developers, in November 2017, InfluxData announced Flux, the industry's first purpose-built functional query language for time series data — perfect for querying, analyzing, and acting on time series data. Flux takes the power of InfluxQL and the functionality of TICKscript and combines them into a single, unified syntax. It is powerful, flexible, and provides support for key features like joins, math across measurements and histograms which help you gain even stronger insights into your time series data.

The InfluxDB data model

The InfluxDB data model is quite different from other time series solutions like Graphite, RRD, or OpenTSDB. InfluxDB has a line protocol for sending time series data which takes the following form:

```
<measurement name>,<tag set> <field set> <timestamp>
```

The measurement name is a string, the tag set is a collection of key/value pairs where all values are strings, and the field set is a collection of key/value pairs where the values can be int64, float64, bool, or string. The measurement name and tag sets are kept in an inverted index which make lookups for specific series very fast.

For example, if we have CPU metrics:

```
cpu,host=serverA,region=uswest idle=23,user=42,system=12 1549063516
```

Timestamps in InfluxDB can be by second, millisecond, microsecond, or nanosecond precision. The micro and nanosecond scales make InfluxDB a good choice for use cases in finance and scientific computing where other solutions would be excluded. Compression is variable depending on the level of precision the user needs.

On disk, the data is organized in a columnar style format where contiguous blocks of time are set for the measurement, tagset, fieldset. So, each field is organized sequentially on disk for blocks of time, which make calculating aggregates on a single field a very fast operation.

There is no limit to the number of tags and fields that can be used. Other time series solutions don't support multiple fields, which can make their network protocols bloated when transmitting data with shared tag sets. Most other time series solutions only support float64 values, which means the user is unable to encode additional metadata along with the time series data.

Even OpenTSDB and Kairos, which support tags (unlike Graphite and RRD) have limitations on the number of tags that can be used. At around five to six tags, the user will start seeing hot spots within their cluster of HBase or Cassandra machines. InfluxDB, on the other hand, doesn't have this limitation.

The InfluxDB data model is purpose-built for time series, specifically. It pushes the developer in the right direction to get good performance out of the database by indexing tags and keeping fields unindexed. It's flexible in that many data types are supported and the user can have many fields and tags.

InfluxDB vs. APM and logging apps

We often get asked how InfluxData is different from applications like Datadog, SumoLogic, Splunk, New Relic, and other metrics and monitoring systems. First and foremost, InfluxData is purpose-built as a time series data platform. It exists so that developers can build their applications on top of the platform.

The above-mentioned applications have their user interfaces and business logic built-in. They're meant to be off-the-shelf solutions that give developers whatever they need for the given problem. Most developers wouldn't build their custom applications on top of these bundled solutions.

InfluxData is a platform for developers to build upon. It's meant to be customized for the unique business logic of each organization it's deployed by. This makes it an ideal choice for larger organizations that are looking to develop solutions designed specifically for their needs or for application developers creating solutions for customer-facing products.

InfluxDB vs. other databases

InfluxDB is often compared to other databases. However, when doing a comparison, the entirety of the InfluxDB platform should be taken into account. There are multiple types of databases that get brought up for comparison. Mostly these are distributed databases like Cassandra or more time-series-focused databases like Graphite or RRD.

When comparing InfluxData with Cassandra or HBase, there are some stark differences. First, those databases require a significant investment in developer time and code to recreate the functionality provided out of the box by InfluxDB. Specifically, developers will need to write code to shard the data across the cluster, aggregate and downsample functions, data eviction and lifecycle management, and summarization. Finally, they'll have to create an API to write and query their new service.

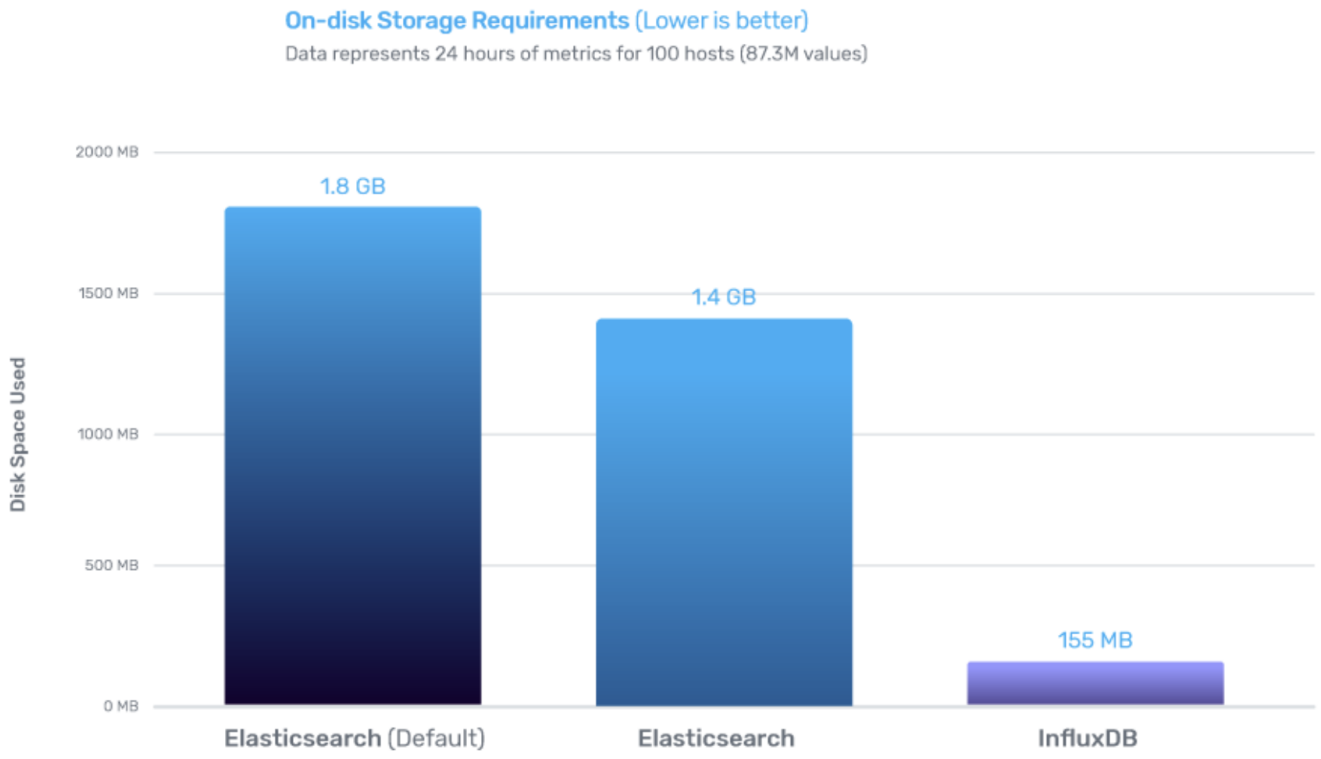
When the rest of the InfluxDB platform is brought into the picture, developers using Cassandra or HBase have even more ground to make up. They'll need to write tools for data collection, introduce a real-time processing system and write code for monitoring and alerting. Finally, they'll need to write a visualization engine to display the time series data to the user. While some of these tasks are handled with other time series databases, there are a few key differences between the other solutions and InfluxDB. First, other time series solutions like Graphite or OpenTSDB are designed with only regular time series data in mind and don't have the ability to store raw high-precision data and downsample it on the fly.

While with other time series databases, the developer must summarize their data before they put it into the database, InfluxDB lets the developer seamlessly transition from raw time series data into summarizations.

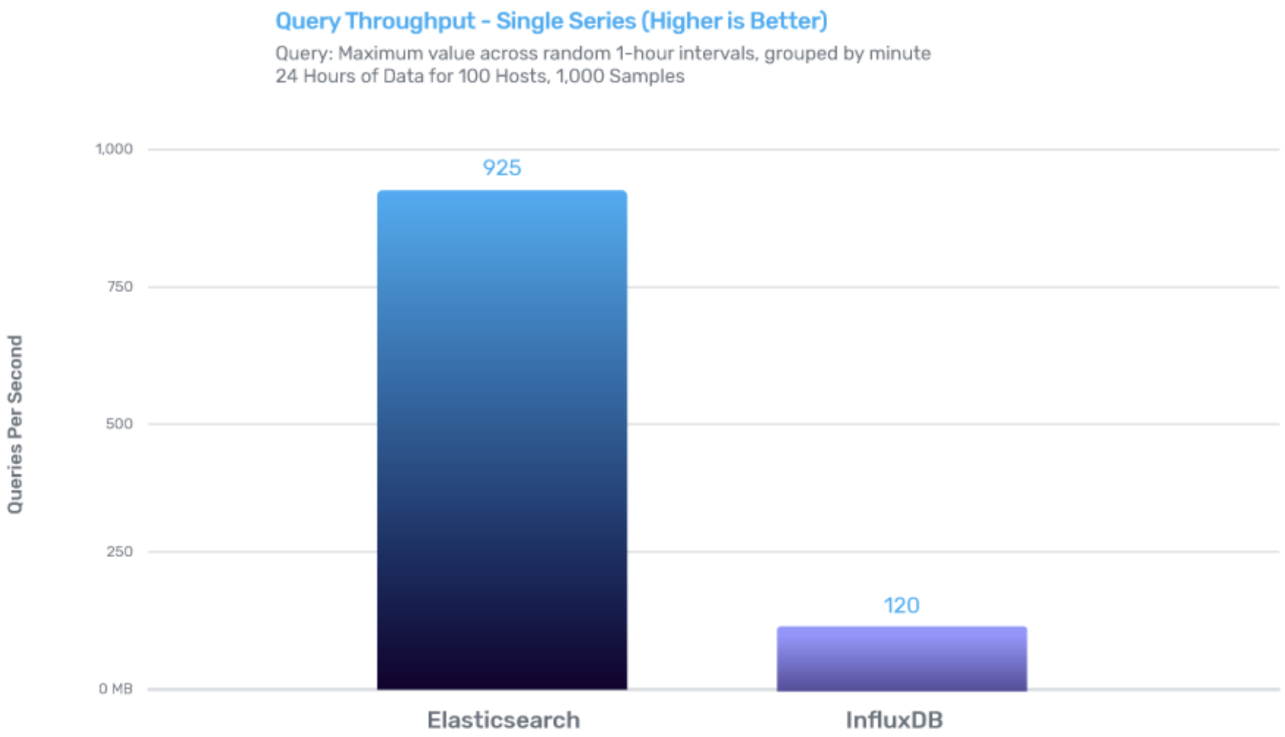
InfluxDB also has key advantages for developers over Amazon Timestream. Amazon Web Services (AWS) recently entered the market with Timestream, a hosted time series database. Public details about Timestream's technical capabilities are limited, but based on the AWS model, there are likely several significant differences between it and InfluxData's offerings — enough to appeal to distinct developers and use cases. Among them:

- **Open source** — InfluxData is first and foremost an open source company. It is committed to sharing ideas and information openly, collaborating on solutions and providing full transparency to drive innovation. InfluxData's products are continuously improved by an energized group of developers that help to make them more reliable, secure and awesome. The power of the open source community to drive innovation is unsurpassed by any proprietary software solutions.
- **Hybrid cloud and on-premises support** — Distributing assets on one or across multiple cloud-hosting environments is often the best choice for companies. Key among many advantages is avoiding vendor lock-in, which can limit the ability to customize systems and negotiate better rates, ultimately making it very difficult — and expensive — to change cloud providers to meet ever-evolving business and technical needs. Every business has specific functionalities, and a hybrid cloud system offers the flexibility to choose services that best fit their needs, whether to support GDPR regulatory requirements or teams that are spread across multiple providers. From an operations perspective, a multi-cloud system increases efficiency and provides another layer of security to ensure that there is no downtime.

Finally, on-disk size is 9x larger on Elasticsearch than InfluxDB if you need to query the raw data later.



InfluxDB demonstrates approximately 7.7x faster query performance than Elasticsearch.



For more details on how InfluxDB stacks up against Elasticsearch plus details on the benchmarks, download the Benchmarking InfluxDB vs. Elasticsearch for Time-Series Data, Metrics & Management technical paper.

InfluxDB vs. stream processing solutions

Stream processing solutions like Spark, Storm, Kafka and others are often used for real-time analytics and other time series processing workloads. With InfluxData's Kapacitor, there is some overlap with these solutions. However, these are mostly complementary tools that users have already included in their InfluxDB architectures.

With that being said, we can look at how Kapacitor differs from these other general-purpose stream processing tools. First, Kapacitor has the structure and schema of InfluxDB as a core part of its processing engine. It has a domain-specific language built in to help developers do common tasks on InfluxDB time series.

For developers using a general-purpose processing engine, they'll need to write code to decode, process, and re-encode time series data. Kapacitor comes with these tools out of the box letting the developer think more about what they want to monitor and less about the tooling needed to get the job done.

Additionally, the development of Flux, InfluxData's new scripting and query language, has made InfluxDB's time series analytics capabilities all the more powerful.

InfluxData *is* time series data

Fueled by the massive growth of connected devices (i.e., IoT) and rapidly increasing instrumentation requirements of next-generation software, time series technology has become more popular. Since launching InfluxDB, an open source time series platform in 2013, we have seen millions of downloads, built an expanding list of enterprise customers, and fostered a growing community that is always finding new ways to deploy and build on our platform. InfluxData also offers two paid editions of InfluxDB: InfluxDB Cloud (managed database as a service) and InfluxDB Enterprise (subscription that turns any InfluxData instance into a production-ready cluster that can run anywhere).

InfluxData has a narrow focus for what has rapidly become a horizontal use case. Our narrow focus means that the entire stack can have optimizations for performance and developer productivity that other general-purpose solutions can't match, such as high compression, super-fast engines and powerful query language designed to best work with flow-based models. At the same time, InfluxData provides a platform that is broadly customizable, making it a perfect choice for developers who want greater control over their tooling than what out-of-the-box applications and solutions provide.

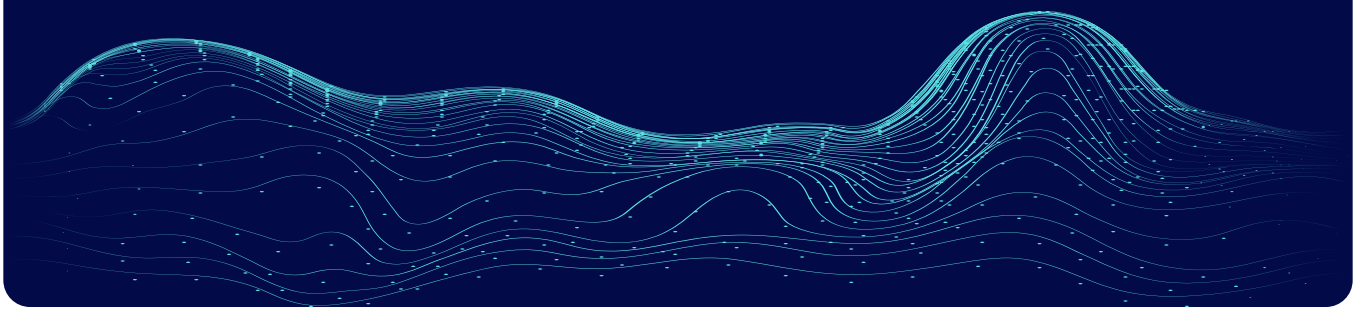
The level of performance that a single InfluxDB server can provide outstrips what users can do on a 10-server deployment of HBase, Cassandra, or Elastic. InfluxDB is highly optimized for this use case, and as a result, can deliver anywhere from 10x to 1000x performance compared to other solutions.

With time series as the key ingredient for custom DevOps monitoring and metrics, real-time analytics, and sensor data and the Internet of Things, InfluxData continues to be at the forefront of what is sure to be the next wave of data platforms after NoSQL.

According to DB-Engines, over the last 24 months, time series has been the fastest growing database category. InfluxData's InfluxDB is the overwhelming leader among time series database management systems, according to DB-Engines' results since January 2016.

About InfluxData

InfluxData is the creator of InfluxDB, the leading time series platform. We empower developers and organizations, such as Cisco, IBM, Lego, Siemens, and Tesla, to build transformative IoT, analytics and monitoring applications. Our technology is purpose-built to handle the massive volumes of time-stamped data produced by sensors, applications and computer infrastructure. Easy to start and scale, InfluxDB gives developers time to focus on the features and functionalities that give their apps a competitive edge. InfluxData is headquartered in San Francisco, with a workforce distributed throughout the U.S. and across Europe. For more information, visit influxdata.com and follow us @InfluxDB.



548 Market St. PMB 77953, San Francisco, CA 94104
influxdata.com Twitter: @InfluxDB