TECHNICAL PAPER

Architecting for IoT: The Need for an IoT Data Platform

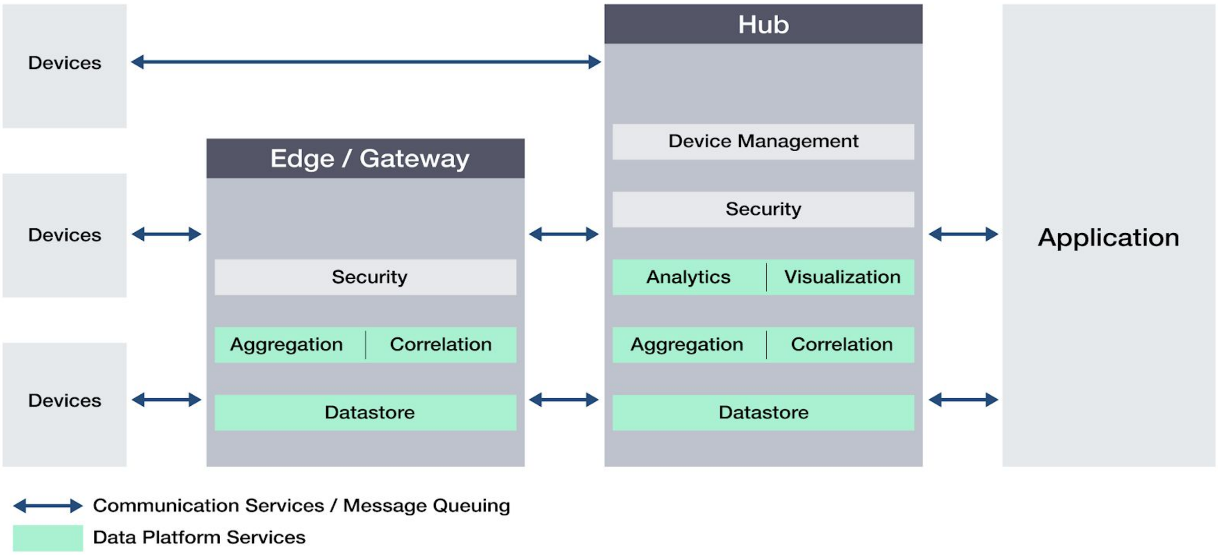
May 2017

Architecting for IoT - choosing the right IoT data platform

We are witnessing the instrumentation of every available surface in the material world — streets, cars, factories, power grids, ice caps, satellites, clothing, phones, microwaves, milk containers, planets, human bodies. Industry experts estimate that there will be more than 30 billion connected IoT devices by 2020. These devices will generate more data than we have ever experienced. This data is streaming in real-time and it will force companies to determine which IoT platform architecture will be resilient, scalable, and extensible enough to handle these new workloads.

The purpose of IoT projects is to gather data from sensors or devices in order to gain real-time insights, accelerate decision-making, perform automated tasks, and create value by enabling organizations to become data-driven. So the ability to ingest large amounts of data without interruption is a key differentiator in IoT data platforms. Pushing data services further to the edge is the current direction of IoT Data Platforms and one that will continue as a defense against rising network costs and as edge devices become more powerful.

Distributed IoT platform



Although the below is not an exhaustive list, we see three key architectural components in a typical IoT architecture:

Devices: Sometimes referred to as “Things” or “end points,” these can be software or hardware devices that generate data or measurements, from the object they are monitoring and/or controlling. Some examples are temperature sensors, voltage sensors, humidity sensors, and machine rotation or vibration sensors. Sensors can be co-located together as in the case of a wind turbine or a car while others form part of another device (like the Smart Watch with biometric data). Some will even have control features like industrial valves.

Edge/gateways: Gateways are usually deployed closer to the edge (some might actually be part of a sensor collection and are deployed with the sensor) and allow for the collection of data, the selective transfer of data to the Hub, and provide some command-and-control functionality back to an automated device, for example. This area is further expanding into what is called edge computing. This allows for critical decisions to be made as close as possible to where the data is created without it having to always go back to the Hub.

Hub: The Hub is conceptually the IoT solution's central processing area. Hubs may be deployed in a distributed fashion but conceptually provide the ability to process data streams from the gateways (or directly from the sensors), to store, monitor, and analyze data, and provide the basis for new applications to enable competitive differentiation.

To support this architecture, an IoT platform needs to provide a variety of services:

Communication services: The necessary services that provide device connectivity, message/event queuing and transportation services across Wi-Fi, cellular or other wireless protocols, and fixed connections.

Security services: A set of security services that provide verification, encryption and authentication services to ensure that devices (and software on the devices) are secure and tamper-proof; also usually required are additional services that provide security to the communication services.

Device management services: Services that support devices' provisioning and lifecycle management.

Data services: Key set of data services that support collection, aggregation, storage, visualization and analytics of the sensor data. The next section of this whitepaper is to explore this set of services in more depth.

IoT data characteristics

Data from devices and sensors, collectively known as IoT data, has unique characteristics that must be addressed by the data services:

IoT data is time series data: IoT is synonymous with time-stamped data, or time series data, since the purpose of any sensor is to measure change over time. To gain insight and act, systems need to evaluate and analyze the data based on timeframes and ranges. Time is not an afterthought for IoT data — it is a constituent of data. A data platform for IoT therefore needs to be optimized for time series data.

IoT data is streaming data: IoT is synonymous with streaming data. Data is created by a myriad of sensors, and each sensor emits a relentless stream of data. A data platform for IoT needs to be optimized for streaming data and have an abundance of actionable analytics to find the signal from all the noise.

IoT data is real-time: Sensors capture and emit data in real time. Greater business value is derived if the data can also be ingested, analyzed, interpreted and acted upon in real time. A data platform for IoT needs to be designed to handle real-time ingestion and real-time streaming analytics to ensure greater business value.

IoT data platform services

The new IoT workloads, and IoT data characteristics — more data points, more data sources, more monitoring, more controls — demand a paradigmatic shift in how we approach building systems that can support these unique characteristics. There is the need for a modern IoT Data Platform which provides a comprehensive set of tools and services that are optimized for IoT data and provide:

Data aggregation services: Lightweight services that can be deployed in a distributed manner to collect, normalize, correlate, and aggregate metrics and events from sensors and other important data sources across a wide range of IoT data protocols.

Data storage services: Data storage services that support high write loads, the storage and real-time retrieval of large sets of time series data. In addition, the database needs to provide time series functions that support query and analytics.

Streaming analytics and visualization services: A set of services that provide visualization and dashboarding services, real-time pattern detection services, a set of notification, control and action services to automate the entire system.

InfluxData - the modern IoT data platform

InfluxData delivers a modern open source IoT Data Platform built from the ground up to support organizations that are looking at building solutions to take advantage of IoT data. Specifically, InfluxData provides the following services:

Data storage services: At the heart of the InfluxData offering is InfluxDB, an open source Time Series Database that supports high write loads, large data set storage, and conserves space through high-efficiency compression, downsampling, automatically expiring and deleting unwanted data as well as backup and restore. InfluxDB also makes it easy to analyze data by providing an easy-to-use, powerful data query language. According to DB-Engines, InfluxDB is the leading Time Series Database.

Data Aggregation Services: InfluxData provides a comprehensive set of tools and services to get metrics and events data from sensors, devices, systems, and machines. InfluxData's collection services are built from the open source Telegraf project. With an ever-growing set of plugins, Telegraf enables the collection of IoT data across multiple protocols popular with IoT devices. In addition, InfluxData provides services to normalize, correlate, and aggregate this data by using services from the open source Kapacitor project.

Streaming Analytics Services: InfluxData uses the services from the open source project Kapacitor as a native data processing engine. It can process both streaming data or batch data from the data store. Kapacitor lets you plug in your own custom logic, User-Defined Functions (UDFs), or machine learning libraries, to process the data and create alerts with dynamic thresholds, perform pattern matching, compute statistical anomalies, etc. that can then trigger UDFs to form the basis of your IoT control plane.

Visualization Services: InfluxData allows for the graphical real-time visualization of data with the open source project Chronograf, and performs ad hoc exploration of your data. Chronograf includes support for templates and a library of intelligent, pre-configured dashboards for common data sets. With a variety of configurable User Interface components, dashboards can provide fast, visually impactful real-time access to data trends for rapid analysis.

In a nutshell, InfluxData provides a comprehensive set of open source IoT data services that can be deployed from the hub to the gateway and the edge.

When you take a step back and consider what you are doing with all your devices connected in an IoT platform, you realize that it really is all about data. IoT is expected to generate large amounts of data with nanosecond precision, from diverse locations, with the consequent necessity for quick aggregation of the data, and an increase in the need to index, store, process and react to this data effectively.

By adopting InfluxData as your IoT Data Platform, you join the ranks of an impressive, rapidly growing global user base. InfluxData's products are already in use at companies in various industries including Tesla, Siemens, BBOXX, Spiio, tado°, LineMetrics, and many others.

In the next section we provide additional insight and some lessons learned from some of our clients on their journey to adopting an IoT Data Platform.

Learning from our clients: choosing the right IoT data platform

At InfluxData, we have many customers that have adopted an IoT Data Platform, but some have taken a meandering path on this journey. From these lessons learned, we hope to help provide a more direct roadmap to choosing the correct IoT data platform by skipping the risks and costs of adopting unsuitable databases and gaining a clear perspective of what constitutes the correct data platform for IoT.

Lesson 1: Don’t start with the wrong data platform architecture

The typical path for clients often starts with data technologies they know:

Starting with MySQL: MySQL is a popular open source relational database management system. But database technologies are written to solve a particular use case. Relational technologies such as MySQL and Oracle are designed for keeping references to interesting data (like Customers have Orders, Orders have Order-lines). But relational technologies do a terrible job with IoT data — streaming, time-stamped real-time data. Relational databases provide poor compression, poor time performance, poor query, and poor scalability for IoT data.

Try HBase or Cassandra: The next common stop — presenting another similar uphill battle — is usually HBase (an open source, non-relational, distributed database modeled after Google's Bigtable and written in Java) or Cassandra. These databases are columnar or key-value databases and are designed for arbitrary amounts of “big data,” but they too are not designed to handle the data characteristics of IoT data. For instance, neither model supports time as a key constituent; they don’t compress time-stamped data for better resource utilization; they can’t handle millions of writes per second that streaming IoT data requires; they have no built-in functions to perform time-dependent queries (for example, has this sensor’s reading broken the 14-day moving average more than twice this week); and they can’t automatically downsample the time precision of data for older data to free up resources. All these capabilities could be added with some engineering resources, but why waste precious resources on adding these capabilities when you can use a platform that already has these capabilities built-in? Read how InfluxDB outperformed Cassandra for IoT data in write throughput and on-disk compression by a significant margin.

Try Elasticsearch: Clients then realize they need a different type of database and try Elasticsearch — an open-source distributed, full-text search engine suitable for enterprise workloads which also happens to have a datastore that can be used for time series data. However, Elasticsearch’s flexibility comes at a price: any particular use case needs to be modeled to correctly utilize the primitives Elasticsearch provides, and using it optimally requires knowing how the internal mechanisms work and requires a much steeper learning curve. InfluxDB is purpose-built for time series and is therefore easy to setup and use for an IoT use case. In addition, our benchmark comparisons showed that InfluxDB outperformed Elasticsearch for IoT data across three tests — for data ingest performance, on-disk storage requirements, and mean query response time — by a significant margin.

Try MongoDB: At this point in the journey, some clients try MongoDB — an open-source, document-oriented database, colloquially known as a NoSQL database, written in C and C++. Though it’s not a Time Series Database, its creators often promote its use for IoT workloads. But MongoDB is a NoSQL database which is more suited for use as a document store or key-value pair store. Neither of these use cases support IoT data well. Our benchmark comparison indicated that InfluxDB outperformed MongoDB for IoT data in data ingestion and on-disk storage by a very wide margin.

Adopt a Time Series Database: Eventually, customers try a Time Series Database (TSDB), because IoT Data is Time Series data. A modern TSDB is built specifically for handling metrics and events or measurements that are time-stamped. A TSDB is optimized for measuring change over time. Properties that make time series data very different than other data workloads are data lifecycle management, summarization and large range scans of many records. By staying in a native format of time series data, a Time Series Platform provides much greater efficiency. In terms of analytics in general, there are specific kinds of analytics that are looking at real-time change over time — for deriving insights from these changes, a Time Series Database is critical. One indication of the future's arrival today is the rise of Time Series DBMS as the fastest growing category among the variety of DBMS platforms over the last 12 months.

Lesson 2: Don't settle for first-generation Time Series Databases

Once clients decide to test Time Series Databases, they often try first-generation ones, yet Time Series Databases differ in the types of time series data they can handle. Time series data comes in two forms: traditional regular (metrics) and irregular (events) — both are present in IoT data and some may argue that most sensor data is irregular or event-driven data. Time Series Databases such as Graphite, RRD, or OpenTSDB support only regular time series metrics. In contrast, modern time series data platforms, like the InfluxData stack, are optimized for both regular and irregular time series data.

Further, modern Time Series Databases, like InfluxDB, have evolved their data model over first-generation time series solutions and impose no limits to the number of tags and fields that can be used and allow timestamp precision in nanoseconds. This is important as sub-millisecond operations become more common. See our benchmark comparison showing how InfluxDB outperforms OpenTSDB.

Lesson 3: Use an IoT Data PLATFORM, not just a database

Some clients start with just the database, which does a great job supporting high write loads, large data set storage, time series functions, and optimal data compression. However, they soon come to the conclusion that they need more than just the database. They need the ability to perform correlation, aggregation, and pattern detection of the streaming data before it gets to the database (streaming analytics). They need the ability to visualize real-time data and set up notification, control and action services to automate the entire system. In short, they need a Time Series Platform to support their IoT architecture.

Representative customer case studies

Some representative InfluxData customer case studies are cited below as examples, to different extents, of the potential delays and resource investment involved if you don't start with InfluxDB for your time series use case.

- BBOXX - This solar energy product provider transitioned from its outsourced monitoring solution to real-time data with InfluxDB Cloud to continuously monitor their geographically dispersed 85,000 solar rooftop units.
- Spiio - This IoT company, focused on plant analytics, started from MySQL for a proof of concept, considered generic/branded IoT platforms, proceeded to try other Time Series Databases, then adopted InfluxDB.
- tado° - This IoT-based home climate control company started with MySQL for a proof of concept, then progressed to next-gen MySQL, then settled on InfluxData.

We would love to hear your path to an IoT Data Platform.

About InfluxData

InfluxData is the creator of InfluxDB, the open source time series database. Our technology is purpose-built to handle the massive volumes of time-stamped data produced by IoT devices, applications, networks, containers and computers. We are on a mission to help developers and organizations, such as Cisco, IBM, PayPal, and Tesla, store and analyze real-time data, empowering them to build transformative monitoring, analytics, and IoT applications quicker and to scale. InfluxData is headquartered in San Francisco with a workforce distributed throughout the U.S. and across Europe.

Learn more.

InfluxDB documentation, downloads & guides

Download InfluxDB
Get documentation
Additional case studies
Join the InfluxDB community

799 Market Street
San Francisco, CA 94103
(415) 295-1901
www.InfluxData.com
Twitter: @InfluxDB
Facebook: @InfluxDB