/ WHITE PAPER

Time-series data in manufacturing

Definition, potential and technologies

CRATE.IO

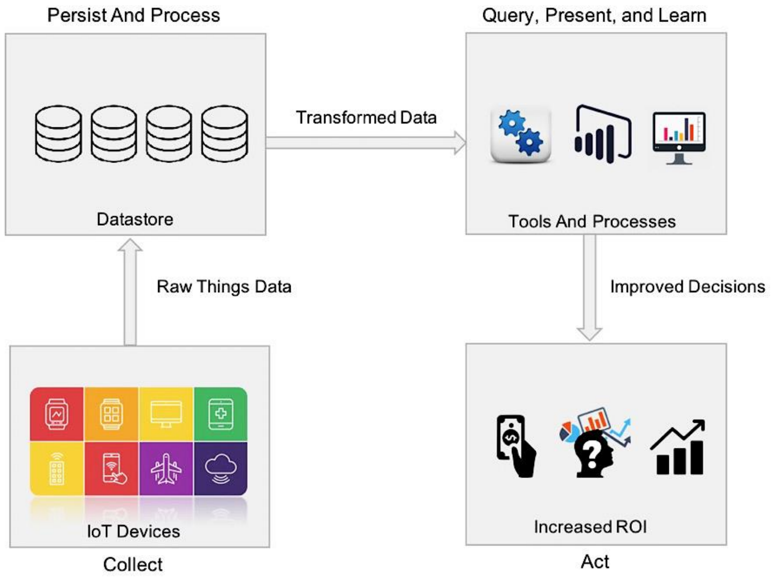
Outline

- The data boom
- What are time series?
- Why are time-series data so relevant right now?
- Time-series data in manufacturing
- Choosing the right technology for industrial time series
  - Operational historians
  - Traditional RDBMS
  - NoSQL databases
  - Time series databases
  - NewSQL databases
- Summary

The data boom

The technological advances of the last decade have allowed the digitalization of all sorts of day-to-day objects. Appliances, vehicles, and machines are now continuously measuring, collecting, storing, and analyzing huge amounts of information; the amount of data we are collecting is higher than ever, and the insights we are gaining are incredibly valuable. This data tells a very detailed story about all sorts of interesting things, including how your business is working and what your customers want—and that's a story that we all want to hear.

In order to access that information, huge amounts of data must be continuously ingested and processed. The great advantages in data technology of the past decade are making it possible: we have now access to effective tools for data processing and analysis, and to affordable storage due to cloud computing. However, the data volume keeps growing, the IoT applications get more and more sophisticated, and the challenge continues.



The basic IoT workflow: 1) Raw data is collected by IoT devices; 2) Data is stored and processed; c) The information is presented; c) This information is used for decision-making

Any major analyst has projections concerning the impact that IoT will have in the future economy, and their results show it will be huge. According to Statista, there will be over 70 billion IoT devices connected by the year 2025. McKinsey predicted that 11 Trillion USD will be generated per year due to the expansion of the IoT. Something is clear: the era of big data is here to stay.

What are time series?

The IoT revolution demands to generate data where it was not possible before, information that can then be used to improve systems and processes. Data opens the door to all sorts of valuable insights, but not only that—it allows us to use those insights to take action. If we can know what a user needs, we can deliver it exactly. If we have access to information showing why a business is not working, we can improve it. It's all about knowing the whole story—and it happens that time-series data are very good storytellers.

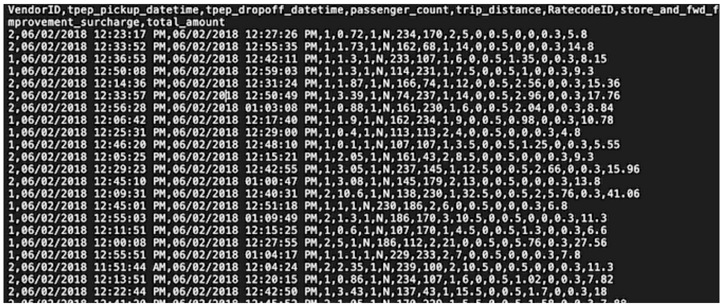
When setting up an IoT application, the goal is to be able to read the environment as accurately as possible and to communicate the results in a clear way to the system, user, manager, marketer, or whoever is interested in that information. In order to do that, it is especially valuable to set a data source that records the state of a particular metric overtime. By periodically recording the state of a metric (or many), we can build a pretty accurate image of how any system is working, in real-time and also historically.

For example, by continuously recording the temperature in a room, we can automatically know how well the thermostat is working. We can also identify when there's a problem, even if we are not in the room. By relating the temperature-time record to other data (i.e. who touched the thermostat at what time) we could easily identify the reason why that room is suddenly too hot. The problem can be corrected fast… And we even know who to blame.

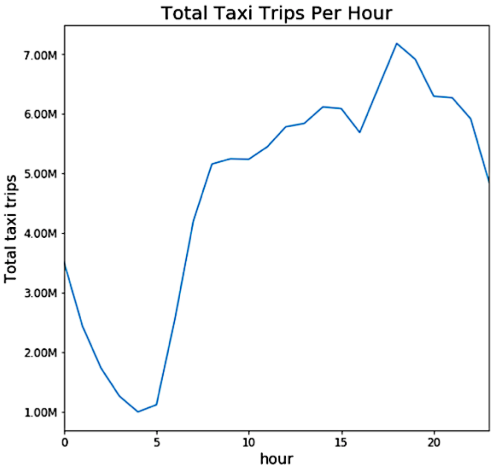
In this example, we could still check the temperature of the room (and figure out who messed up with our thermostat) otherwise. However, if instead of one sensor we needed to monitor thousands of them, things start getting complicated. If we also want to analyze the data from all these sensors as a whole, discovering patterns and relations between them, the degree of complexity increases even more. And if we want to consider the influence of external factors in the analysis of the sensor data, the degree of complexity is maximum—but the insights we gain get extremely interesting.

Why are time-series data so relevant right now?

This type of dataset is, obviously, not new. We are all surrounded by sensors that measure things overtime, and we have been for a while. The difference is that now we can use time-series data not to only see the latest state of simple systems, but to study the overtime evolution of complex operations. Since this implies the handling of massive volumes of data, before this was something out of our technological scope—but we live now in the data era. The potential of data has become too obvious to ignore, and the real-time analysis and storage of huge datasets is now a reality.



```
VendorID,tpep_pickup_datetime,tpep_dropoff_datetime,passenger_count,trip_distance,RatecodeID,store_and_fwd_f
mprovement_surcharge,total_amount
2,06/02/2018 12:23:17 PM,06/02/2018 12:27:26 PM,1,0.72,1,N,234,170,2,5,0,0.5,0,0,0.3,5.8
2,06/02/2018 12:33:52 PM,06/02/2018 12:55:35 PM,1,1.73,1,N,162,68,1,14,0,0.5,0,0,0.3,14.8
1,06/02/2018 12:36:53 PM,06/02/2018 12:42:11 PM,1,1.3,1,N,233,107,1,6,0,0.5,1.35,0,0.3,8.15
1,06/02/2018 12:50:08 PM,06/02/2018 12:59:03 PM,1,1.3,1,N,114,231,1,7.5,0,0.5,1,0,0.3,9.3
2,06/02/2018 12:14:36 PM,06/02/2018 12:31:24 PM,1,1.87,1,N,166,74,1,12,0,0.5,2.56,0,0.3,15.36
2,06/02/2018 12:33:57 PM,06/02/2018 12:50:49 PM,1,3.39,1,N,74,237,1,14,0,0.5,2.96,0,0.3,17.76
2,06/02/2018 12:56:28 PM,06/02/2018 01:03:08 PM,1,0.88,1,N,161,230,1,6,0,0.5,2.04,0,0.3,8.84
1,06/02/2018 12:06:42 PM,06/02/2018 12:17:40 PM,1,1.9,1,N,162,234,1,9,0,0.5,0.98,0,0.3,10.78
1,06/02/2018 12:25:31 PM,06/02/2018 12:29:00 PM,1,0.4,1,N,113,113,2,4,0,0.5,0,0,0.3,4.8
1,06/02/2018 12:46:20 PM,06/02/2018 12:48:10 PM,1,0.1,1,N,107,107,1,3.5,0,0.5,1.25,0,0.3,5.55
2,06/02/2018 12:05:25 PM,06/02/2018 12:15:21 PM,1,2.05,1,N,161,43,2,8.5,0,0.5,0,0,0.3,9.3
2,06/02/2018 12:29:23 PM,06/02/2018 12:42:55 PM,1,3.05,1,N,237,145,1,12.5,0,0.5,2.66,0,0.3,15.96
2,06/02/2018 12:45:10 PM,06/02/2018 01:00:47 PM,1,3.08,1,N,145,179,2,13,0,0.5,0,0,0.3,13.8
1,06/02/2018 12:09:31 PM,06/02/2018 12:40:31 PM,2,10.6,1,N,138,230,1,32.5,0,0.5,2,5.76,0.3,41.06
1,06/02/2018 12:45:01 PM,06/02/2018 12:51:18 PM,1,1,1,N,230,186,2,6,0,0.5,0,0,0.3,6.8
1,06/02/2018 12:55:03 PM,06/02/2018 01:09:49 PM,2,1.3,1,N,186,170,3,10.5,0,0.5,0,0,0.3,11.3
1,06/02/2018 12:11:51 PM,06/02/2018 12:15:25 PM,1,0.6,1,N,107,170,1,4.5,0,0.5,1.3,0,0.3,6.6
1,06/02/2018 12:00:08 PM,06/02/2018 12:27:55 PM,2,5,1,N,186,112,2,21,0,0.5,0,5.76,0.3,27.56
1,06/02/2018 12:55:51 PM,06/02/2018 01:04:17 PM,1,1.1,1,N,229,233,2,7,0,0.5,0,0,0.3,7.8
2,06/02/2018 11:51:44 AM,06/02/2018 12:04:24 PM,2,2.35,1,N,239,100,2,10.5,0,0.5,0,0,0.3,11.3
2,06/02/2018 12:13:51 PM,06/02/2018 12:20:15 PM,1,0.86,1,N,234,107,1,6,0,0.5,1.02,0,0.3,7.82
2,06/02/2018 12:22:44 PM,06/02/2018 12:42:50 PM,1,3.43,1,N,137,43,1,15.5,0,0.5,1.7,0,0.3,18
```



Both these images represent time-series data. In the first one, the data is untreated; after being processed, the information can be visually displayed in a chart as in the one in this figure

However, the challenges are still considerable. IoT applications imply unprecedented volumes of time-series data and a high query complexity. Huge amounts of different types of data are continuously being collected by different devices, sensors, machines and systems; this data needs to stored, processed and analyzed fast. And in the particular case of industrial IoT applications, the difficulties increase even more.

Time-series data in manufacturing

Time series are relevant for a lot of sectors, but they are especially important for the manufacturing industry. Factory floors have been automated for decades; huge amounts of data are already being collected by many sensors, which are regularly measuring temperatures, pressures, velocities, volumes, and all kinds of different metrics. This data is rarely being used for more than the simple monitoring of individual systems.

But what if this information was used to have a complete view of all the operations, allowing to monitor every aspect of the business? What if the data could be compared and analyzed as a whole, identifying what needs to be corrected, and implementing the necessary changes? The overall efficiency of the manufacturing process would be improved, and unnecessary costs would significantly be reduced.

This is the future that industrial IoT offers to the industry. However, to actually set up the technology required to process such volume of time-series data is a particularly difficult challenge. Manufacturers have been working with their time-series data for decades, usually using operational historians—software recording instrument readings by time. Many vendors in the sector of automation offer operational historians, which are usually built for specific niche industries. However, historians are having trouble keeping up with the current data needs the sector. Besides, historians were built for the traditional operation of a factory floor: they were never designed for being used for data analysis or correlation.

New times, new solutions. The modern industry requires more simple, holistic solutions to use their data in all its potential. Industrial IoT levels up the complexity of common IoT: data volumes are huge, and they come in very different formats—the number of signals and sensors in a factory floor is very high. Besides, it is key to process the data in real-time, since the goal is to achieve an integrated visualization of the operations in order to identifying and correct failures. And the market can change quickly: the technology must stay relatively flexible, to adapt to the changes of the industry.

Scalability, performance, flexibility, and convenience: these aspects are key for the successful handling of time-series data in IIoT applications. On top of that, cost-efficiency is a crucial element for manufacturers. The industry operates with small economic margins, and as the business grows, the technology costs can also grow substantially—something to be avoided. In a sector where price matters, it is important to consider technologies easily scalable, not requiring the use of high-end hardware or excessive maintenance.

Choosing the right technology for industrial time series

When it’s time to decide which tools to use for handling industrial time-series data, the needs and priorities of the use case must always come first. In what respects to data management, there is never a one-for-all solution: in order to make the best decision, it is crucial to analyze which technology is more suitable for providing what the use case needs.

The main technologies available for working with time-series data can be grouped into five main categories: operational historians, traditional RDBMS, NoSQL databases, time series specialized databases, and NewSQL databases.

Operational historians

We mentioned operational historians before. They are the solution traditionally used to manage operational time-series data in the industry, and they still offer numerous benefits. However, as it was discussed in the previous section, historians are not built to accommodate the current needs of the sector. The field of data analysis has boomed since operational historians were conceptualized; they are not able to fully implement modern techniques as machine learning and predictive maintenance. Besides, historians were built for the traditional manufacturing flow, where every process is treated separately. They are difficult to integrate into other web applications and tools, they are expensive to maintain, and difficult to scale and integrate.

Traditional RDBMS

Pros

- SQL accessibility
- Easy integration
- High data availability

Cons

- Difficult to scale
- Performance degrades with high data volumes

Some of the benefits associated with relational databases are tremendously valuable for the industry. The use of SQL allows an easy accessibility and integration, and aspects like replication, sharding and snapshots help to maximize data availability.

But unfortunately, traditional RDBMS are not optimized for working with the massive data volumes that IIoT implies. Traditional databases are designed to scale vertically, and to upgrade the infrastructure every time that it is required to increase capacity is not a practical option for the industry.

Another interesting aspect to consider is that traditional RDBMS provide ACID semantics. This is a very desirable feature for many use cases, but not necessarily for working with industrial time series; in fact, it can be quite the opposite. In order to be able to handle requirements of IIoT applications, it is a must to increase performance as much as possible—and ACID transitions are, computationally speaking, very expensive. In ACID-compliant databases, the strong consistency is provided with the cost of handling all the writes with a single, master node, which substantially limits the scaling performance of the database.

NoSQL databases

NoSQL databases were born to fill the scalability void and the lack of flexibility of traditional relational databases. They offer valuable advances for the processing of industrial time-series, as efficient scaling and distributed architectures.

They have something else in common—the abandonment of SQL. This not only implies the loss of the powerful capabilities of SQL for the processing of structured data, but the need for specialized engineers to manage the database. This is related to another downside of using NoSQL databases for industrial IoT: the possibility to be locked in. To use technology as adaptable as possible will solve many future compatibility problems.

Pros

- Scalability
- Distribution
- Flexibility

Cons

- Complex to operate
- Difficult integration
- Expensive

In addition, NoSQL databases come with high costs, in terms of database operation and maintenance. And from the perspective of industrial IoT applications, they have a common downside with RDBMS: ACID compliance. Their architecture is often not optimized for IIoT use cases, and their performance for this use case might not be as good as expected.

Time series databases

The interest raised by time-series data during the recent years has originated a new type of databases designed exclusively for the handling of time series. They offer a very good combination of capabilities for working with this type of workload, combining functions of the traditional operational historians with those of RDMS and NoSQL databases.

Pros

- Built for time series

Cons

- Performance degrades for complex aggregations and high volumes
- Need for another database

However, from the point of view of industrial IoT applications their features are not exactly ideal. Actually, to be able to manage IIoT use cases exclusively with a time series database is quite rare: they usually need to work in combination with another database. Besides, industrial IoT use cases usually imply an intense parallel usage, but the performance of time series databases is often optimized for single-node operations. The ability of the database to provide real-time responses is a crucial need of the industry, even when the volume grows, and the complexity of the queries increases.

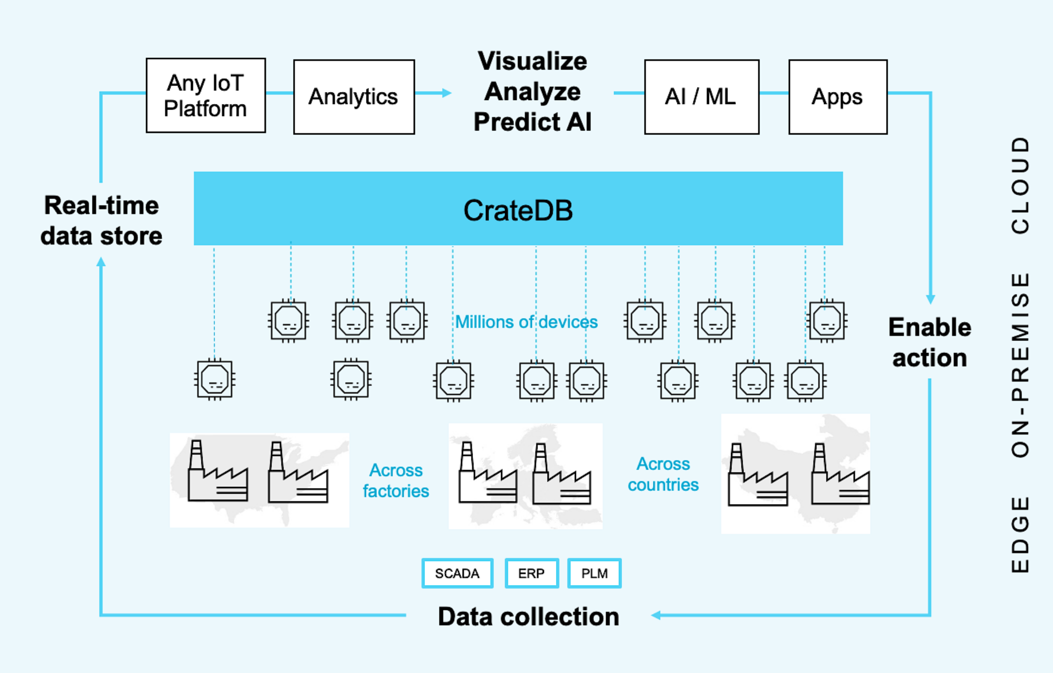
IIoT systems don’t just need to visualize data streams—they must perform analysis, run highly concurrent loads, and perform reads, writes and ad hoc queries at the same time. The right database should be able to provide such performance, while staying flexible and efficient. Time-series specialized databases are not able to provide this combination of properties.

NewSQL databases

There is one last option. What if we could recover the advantages of the traditional relational databases, but with the scalability and flexibility of NoSQL? What if we add time series specific features as well?

A new generation of relational databases was born to make that true. CrateDB, a database purpose-built to fill the needs of industrial IoT applications, fits in this category. A NoSQL foundation gives CrateDB a fully distributed functionality, with linear scalability performance; at the same time, a full SQL access keeps the database simple. It has dynamic schemas, integration is easy, and it is built with an architecture that prioritizes the active use of resources.

CrateDB is optimized to offer fast responses even with high concurrency. It provides real-time capabilities even while ingesting huge data volumes, from thousands of different sensors. And it comes with a very versatile data model, allowing to process with a single database the time series data and the unstructured data that is also a part of industrial operations.



Summary

Due to the IoT boom, a lot of interest is focused on time-series data. Time series are a very good way to gain insights about how a system, application, process or business is really working, allowing to identify what and how to improve.

In the manufacturing industry, high volumes of valuable time series are already being produced; however, to actively use this data for implementing IoT protocols is being difficult. When it's time to evaluate which technology to use, it is important to consider all the needs of the IIoT use case; most existing technologies cannot cope with the volume and complexity of industrial time series, and they don't offer enough flexibility or cost-efficiency.

To use solutions specialized in industrial applications will allow manufacturers to get the most of their data. For that purpose, CrateDB was designed with the needs of the industry in mind. Horizontal scalability, real-time performance, a full SQL access and an easy integration—CrateDB offers the features required for successfully handling high volumes of industrial time series data, without giving up efficiency and simplicity.

Curious about CrateDB? Give it a try!