AN INFLUXDATA CASE STUDY

InfluxDB at CERN and Its Experiments

Adam Wegrzynek
Senior Engineer, CERN

JUNE 2018

Company in brief

CERN, the world's largest particle physics center, consists of 16,868 members and collaborators, covers over 6,250,000 m2 of land, spans 700 buildings, and conducts various experiments. Founded in 1954, the CERN laboratory sits astride the Franco-Swiss border near Geneva. It was one of Europe's first joint ventures and now has 22 member states. At CERN, physicists and engineers are probing the fundamental structure of the universe.

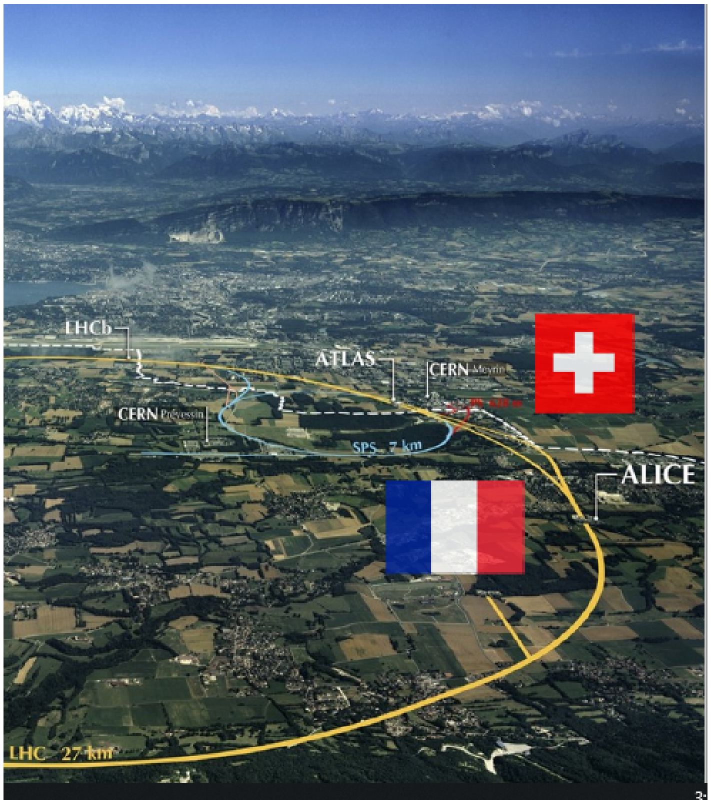
The instruments used at CERN are particle accelerators and detectors. Accelerators boost beams of particles to high energies before the beams are made to collide with each other or with stationary targets. Detectors observe and record the results of these collisions. The particles are made to collide together at close to the speed of light. The process gives physicists clues about how particles interact, and provides insights into the fundamental laws of nature.

The latest addition to CERN's accelerator complex, the Large Hadron Collider is the world's largest and most powerful particle accelerator. ALICE (A Large Ion Collider Experiment) is a heavy-ion detector on the LHC ring. It is designed to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma forms.

Case overview

CERN wanted to upgrade the data monitoring system of one of its Large Hadron Collider experiments called ALICE (A Large Ion Collider Experiment) to ensure the experiment's high efficiency. They needed to constantly monitor their 2000 nodes processing data at 3.4 TB/s which leads to an incredible **600 kHz metric rate**. These metrics are collected and aggregated by Flume and Spark, and CERN chose InfluxDB to store these metrics.

> *"There are many projects at CERN that are using InfluxDB. The largest installation is with the monitoring of one of our data centers where we are using 31 instances of InfluxDB, ingesting over 1.6 TB of metrics a day. "*
>
> ***Adam Wegrzynek,*** *senior engineer*



InfluxDB used for critical monitoring of accelerator systems, experiments and data centers at CERN

LHC and ALICE at CERN

The CERN Accelerator Complex hosts a number of projects, among them the Large Hadron Collider (LHC):

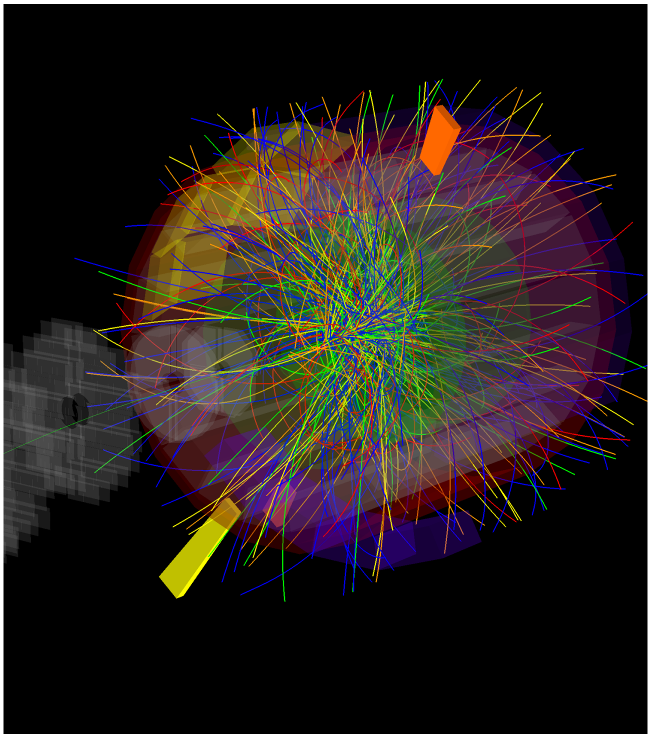
- **Large** due to its size (a ring approximately 27 km in circumference), superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way
- **Hadron** because it accelerates protons or ions, which are hadron
- **Collider** because these particles form two high-energy particle beams travelling, at close to the speed of light, in opposite directions, and which collide in the four interaction points where the major experiments are located

Located 100 meters underground, LHC is the largest cryogenic system in the world and one of the coldest places on Earth, operating at -271.3 °C. Such a cold temperature is required to operate the magnets that keep the protons on course.

One of the LHC experiments, ALICE is a detector specialized in measuring and analyzing lead-ion collisions. It studies quark-gluon plasma (a state of matter thought to have formed just after the big bang) as it expands and cools, observing how it progressively gives rise to the particles that constitute the matter of our universe today.

Collision data processing

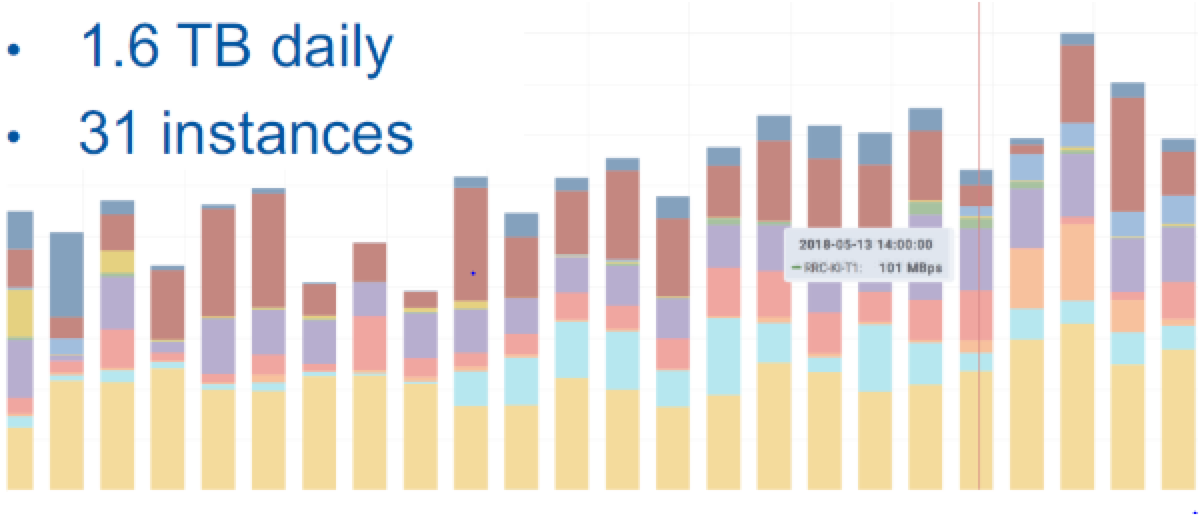
When the particles are about to collide, a special hardware, called Trigger, tells all the other subsystems to start collecting data. During the collision, the particle shower is created, and detectors generate the signal, which is pushed via special electronic devices into the computing firm. This data is then processed and compressed. Many different algorithms are applied and it is eventually sent to storage.The reconstructed data is stored and analyzed in the Worldwide LHC Computing Grid (WLCG), which spans over 170 computing centers in 42 different countries, linking up national and international grid infrastructures. It runs around 2 million tasks a day using 750,000 cores. It already stores around 800 petabytes of data.



InfluxDB use cases in CERN IT

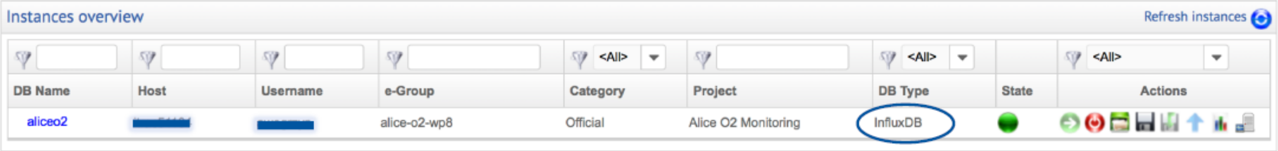
The IT department at CERN is responsible for Tier 0, the core of the WLCG. The Tier 0 core consists of two interconnected data centers, providing 20% of the computing power. InfluxDB is used to monitor this Tier 0, which is quite a large center. For this reason, 31 instances of InfluxDB are used, and 1.6 terabytes of metrics a day are written to InfluxDB.

WLCG Tier 0: InfluxDB



The IT department at CERN also has a service system called DB On Demand, a custom-made deployment and cluster management system. It has a simple web self-service interface through which users can request any database they want, among them InfluxDB. CERN already has 90 projects using this service, collectively writing an impressive 1.5 million points per second.

InfluxDB in CERN's DB On Demand Service System



				<All>		<All>		<All>
DB Name	Host	Username	e-Group	Category	Project	DB Type	State	Actions
aliceo2	[redacted]	[redacted]	alice-o2-wp8	Official	Alice O2 Monitoring	InfluxDB	●	

InfluxDB in the ALICE experiment

> *"We write everything through InfluxDB, which is our main hardware's system."*

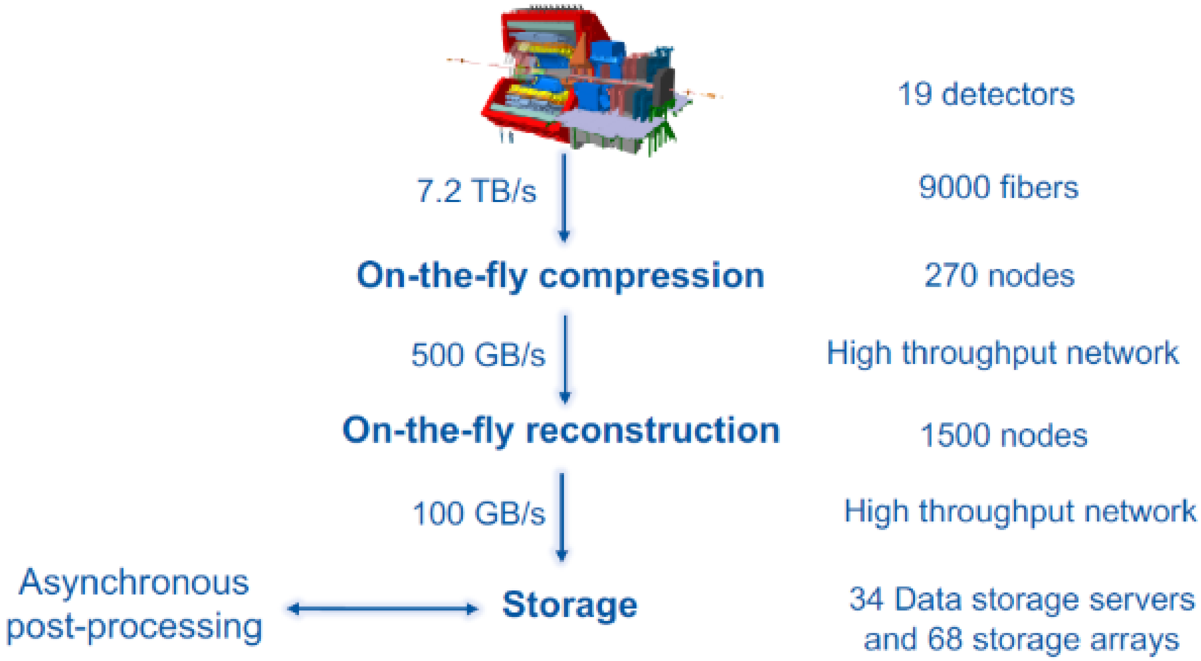
Why InfluxDB?

The new monitoring system of ALICE consists of 19 detectors specialized in heavy ions. The heavy ions allow to form quark-gluon plasma. These deep collisions generate a temperature 100,000 times hotter than the Sun.

The upgraded monitoring system of ALICE will continuously read out about around 7.2 terabytes a day per second, over 9,000 fibers. Then it applies first-level compressing, decreasing it to 500 GB per second and then, eventually, GPU-accelerated reconstruction that brings down the data to 100 GB. This is written for storage and eventually to the WLCG.

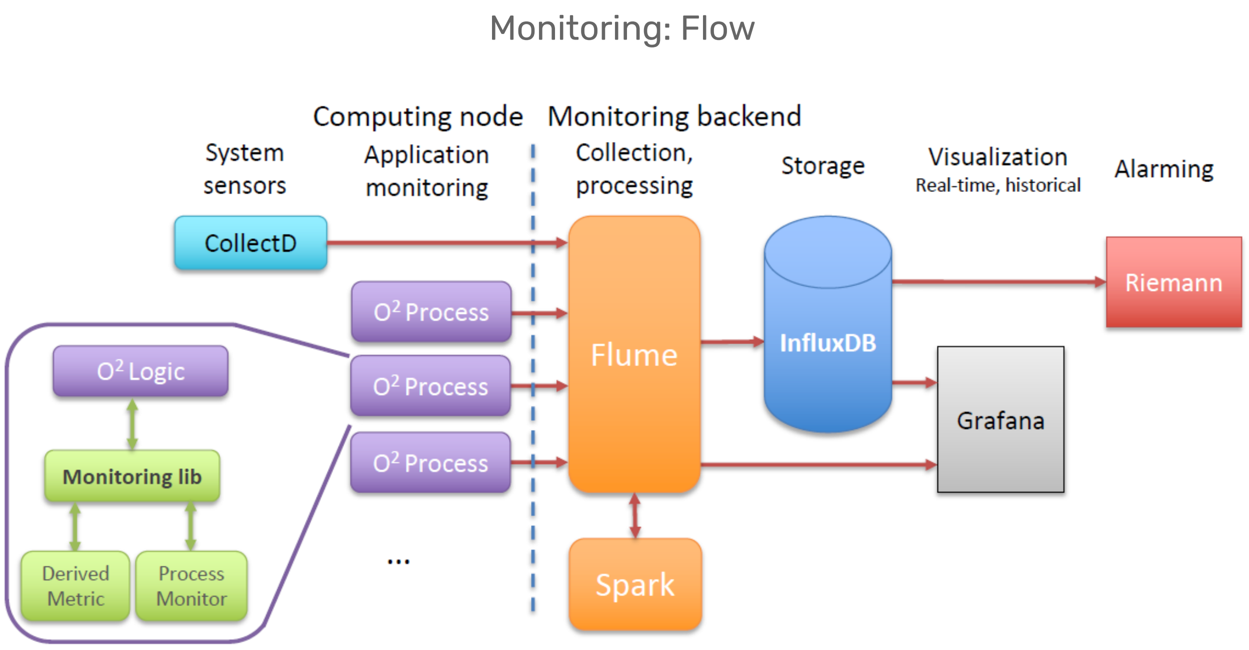
Such a system requires extensive monitoring. The team estimated around 600,000 metrics per second arriving asynchronously from 100,000 different sources, because this is the number of processes they expect to have in their firm. They also didn't want to introduce any high latency into the system because they have many critical values they want to ship to the shift crew in the control center.

Monitoring System Upgrade



To achieve their monitoring system upgrade, they considered tools such as MonALISA, Zabbix, StatsD, collectD, Prometheus, InfluxDB, PostgreSQL, Apache Flume, Apache Kafka, Apache Spark, Riemann, and Grafana. After evaluating many tools, they chose InfluxDB as their time series database because it met their monitoring system requirements in terms of latency, scalability, and write throughput.

Technical architecture



The monitoring flow for ALICE's new data monitoring system is as follows:

- Collectd is used to gather performance metrics and monitor CERN's custom hardware.
- This data is shipped to Flume over UDP protocol.
- Flume routes data to Spark for aggregation and data is re-injected back to Flume.
- Some data is shipped directly to Grafana without going through InfluxDB, and other data goes to Riemann to trigger alarms and notifications for the operations team.
- Everything is written through InfluxDB, their main hardware's system.

The custom monitoring library

In ALICE's new monitoring system, processing devices implement the compression and reconstruction logic and are linked against ALICE's custom C++ monitoring library. Written by Adam Wegrzynek, purely in C++, this library allows writing user-defined metrics, to various backends. It supports many monitoring systems and can monitor the process it's running within. It governs how much CPU, memory and network each process is using, for the 100,000 processes it's running. It also performs initial aggregation to calculate some direct metrics like rate, average, deltas, etc. So central monitoring can be offloaded from basic processing. The library can also monitor custom metrics. It is open source and free to use.

Test configuration for InfluxDB

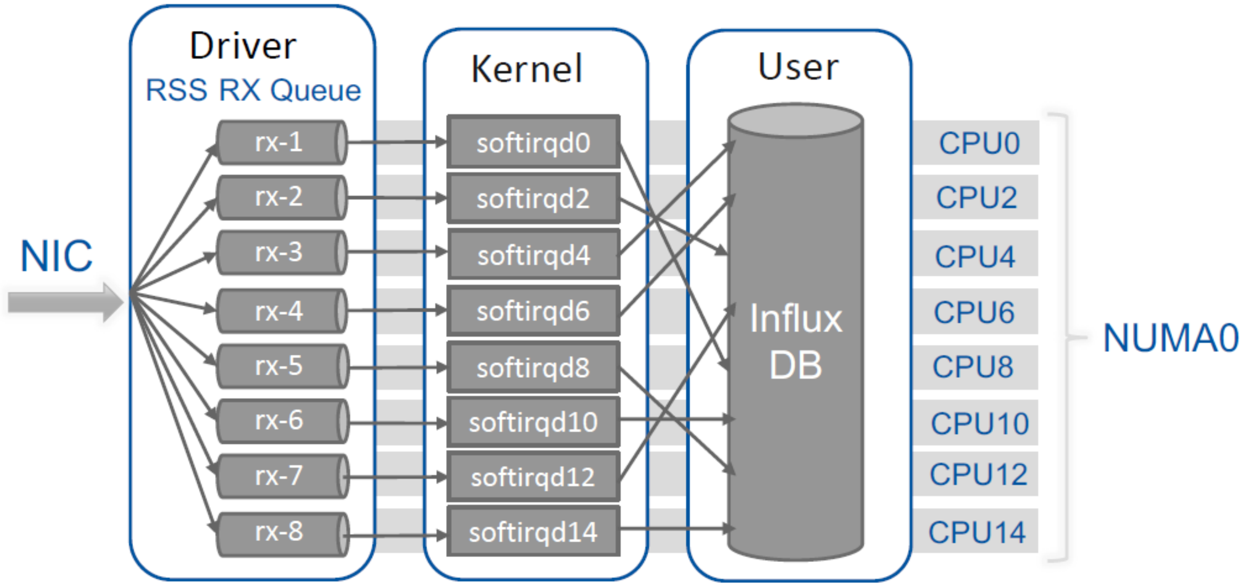
To find out if their system is able to handle the data load, CERN have a test configuration node with one InfluxDB database and the following hardware specs:

CPU: 2x Intel Xeon E5-2670 v3
Memory: 64 GB DDR4
Network: Mellanox MT27520 40 GbE1
Storage: QLogic FastLinQ QL4121 25 GbE2
2x Intel S3610 800GB SSD (RAID 0)
OS: CERN CentOS 7.4 3.10.0-693.21.1
InfluxDB: 1.5.2

Internals of the Receive Stack

Some optimizations were performed, for testing purposes, to get the best values possible.

InfluxDB: Optimization



These optimizations involved the following:

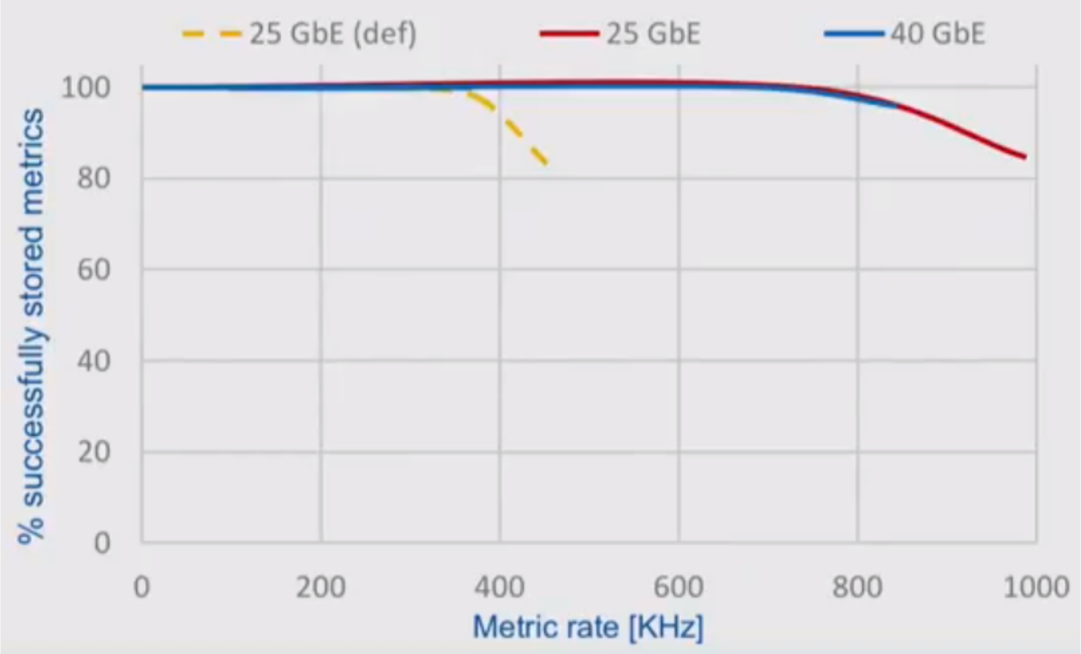
- Linux Scaling Governor (a power-saving mode that scales CPU frequency)
- Receive Side Scaling - RSS (which allows using multiple cues)
- RX Flow Hash (to add more variables like port numbers to be able to use all the cues)
- Socket receive queue memory

- Irqbalance
- NIC, IRQs and InfluxDB at same NUMA node
- LRO/GRO (Large Receive Offloads/Generic Receive Offloads)
- SO_REUSEPORT? (an option that allows having the same port number for multiple sockets)

Visualizing performance

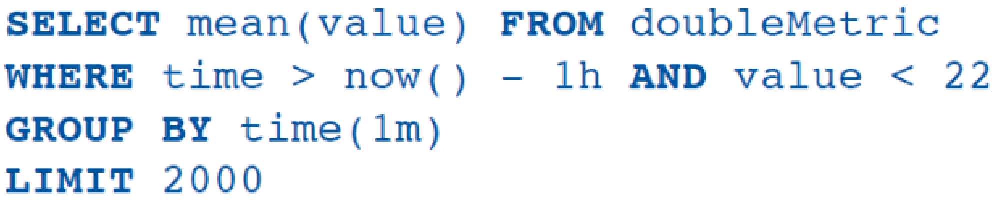
The below plot shows the percentage of successfully stored metrics in a function of the metric rate in kilohertz. On the yellow curve, at around 500 kilohertz, metrics begin to drop. From this graph, the CERN team can find out up to which point they can use the system, what the drop will be, and learn the consequences of going higher. Upon visualizing performance, the team found that InfluxDB behaves quite well.

InfluxDB: Writing



They also executed a simple query to estimate how much time it takes to read the data when there is no major tasks running in the machine. The request completes at around 4 milliseconds, which was sufficient since that meant it could perform 550 requests per second.

InfluxDB: Reading



```
SELECT mean(value) FROM doubleMetric
WHERE time > now() - 1h AND value < 22
GROUP BY time(1m)
LIMIT 2000
```

To improve the reading time when the system was performing reading and writing simultaneously, they are in the process of partitioning the system and using multiple InfluxDB instances. This would enable them to read and write all the data they want. Since they use Flume, it can choose where to write and push data.

What's next for ALICE?

Future steps planned for the system include:

- Partitioning
- Alarming
- Grafana real-time data source
- Custom hardware sensors

Summary

At CERN, several major projects use InfluxDB, and CERN's DB On Demand service promotes the usage of InfluxDB among CERN users. InfluxDB was chosen for ALICE's new monitoring system due to being a purpose-built time series database that allows for high throughput ingest, compression and real-time querying of that same data in line with the system's monitoring requirements.

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.