AN INFLUXDATA CASE STUDY

Edge Computing for IoT Using InfluxDB and Kapacitor

External Contributors

Anil Joshi

CEO, AnalyticsPlus, Inc

Pankaj Bhagra

Co-Founder and Software Architect, Nebbiolo Technologies

November 2019

Company in brief

AnalyticsPlus is a leading Chicago-based advanced analytics and predictive modeling company with strong domain knowledge in healthcare and other Industries. In the manufacturing sector, it has created unique IP driven streaming analytics, anomaly detection, and edge computing in the Internet of Things (IoT) space. Its mission is to help clients generate operational efficiency and derive a competitive advantage through advanced analytics. Its core technical experience includes working with Data Analytics, Big Data, Hadoop, Spark, R, SAS, IBM Netezza, IBM SPSS, Machine Learning, and classical hypothesis driven statistical analysis. It helps clients discover business insights from their own data; helps them grow revenues, reduce cost through automation, and improve their current offerings for increased client satisfaction. AnalyticsPlus is an OEM partner of InfluxData.

Nebbiolo Technologies, based in California, was founded in 2015 by Cisco executives that defined a new computing paradigm — "Fog Computing", known today as "Edge Computing". Nebbiolo's products apply these technologies at scale in the Industrial Automation domain. Nebbiolo provides an edge software platform for real-time, distributed computing. Platform capabilities include Edge Application Virtualization and Workload Consolidation; Edge streaming analytics; centralized management and orchestration for compute, storage, networking, analytics and security; and device-to-cloud, patented industrial security.

Case overview

Nebbiolo Technologies™ and AnalyticsPlus wanted to meet customers' need to perform analytics at the edge, in the data center or in the cloud — a requirement in today's distributed landscape. The two companies partnered to build an IoT Edge Computing Platform that brings the flexibility of virtualized computation, network and storage resources to the edge, as an integrated solution combined with ML and AI libraries into their fogOS middleware. At the heart of the solution is the open-source time series database, InfluxDB, its native data processing framework Kapacitor, and its metrics collection agent Telegraf. Grafana is used for data visualization.

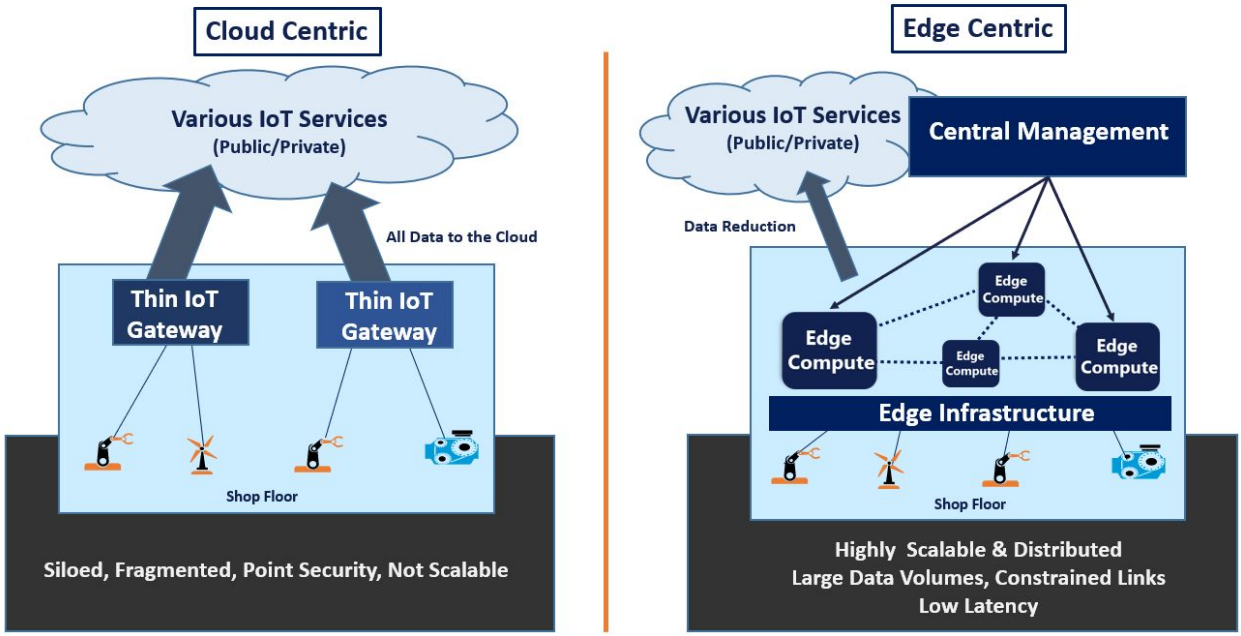
The point-and-click edge computing platform helps customers in 3 specific verticals (manufacturing, healthcare, and finance) to unlock the power of high-frequency data in real-time to become data-driven organizations.

> *"Some of our customers are doing random sampling of the parts they're producing, one part a day. For quality control, they're not looking into each and every part. So data-driven insight is a key, fundamental driving force in the industrial world."*
>
> ***Pankaj Bhagra,** co-founder and software architect, Nebbiolo Technologies*

The business problem

Having the flexibility to perform analytics at the edge, data center or cloud is needed in today's distributed landscape. By 2022, more than 50% of enterprise data will be created and processed outside the data center or cloud, up from less than 10% in 2019, according to a Gartner report.

While many companies are now well-versed in cloud-based infrastructure, edge computing has also become a business necessity for many industries since it has advantages over cloud computing.



The need for edge/fog computing

Edge computing vs. cloud computing

As machines and devices become instrumented, the more you analyze and process the data next to them, the less expensive and secure the process, and the more scalable the environment:

- Edge computing reduces latency (response immediacy). Real-time, data-driven decision-making often requires computing on the side of the device or the machinery where the data is being collected. As the IoT expands and more devices have sensors, edge computing will become more prevalent.
- In the cloud framework, all the data streams into the cloud, incurring data processing and storage costs. On the edge, because processing is distributed on the device side, not all data needs to be used because you can analyze it and use only anomalous data to take needed action. Distributed data processing reduces data load while maintaining latency.
- In edge computing, you can install as many nodes as you want (such as on a factory floor or in a hospital system with different types of machines) and collect all the data coming through these machines, process data locally next to the machine, and take required action quickly based on the insights generated.

To develop the analytics functionalities of their existing platform, AnalyticsPlus partnered with Nebbiolo Technologies — a pioneer in edge computing platforms — whose vision is to apply the Fog Computing Paradigm to Industrial Automation and other related IoT verticals. The partnership aims at giving the edge a cloud-like infrastructure that enables applications to migrate flexibly between the cloud and edge, and therefore free clients to run any type of analytics. Through this partnership, Nebbiolo and AnalyticsPlus sought to build a holistic platform for deploying applications and running analytics.

The technical problem

The platform had to meet high-volume, high-velocity, real-time data processing needs and provide streaming analytics based on machine learning (ML) and artificial intelligence (AI). This is because:

- As data volumes exponentially grow, ML and AI methods are much more efficient and insightful than the traditional methods of statistical analysis.
- Compared to the batch analytics traditionally prevalent in most verticals for the last 50 to 60 years, streaming analytics have become the industry focus,
- Since machines are getting smarter and fitted with sensors, the platform had to integrate analytics and data aggregation capabilities to collect data from various sensors and sources and generate insights about business problems in real time.

The companies' teams (consisting of Master-and PhD-level data scientists, data architecture and software engineers) were well-positioned to build this platform given their extensive experience in:

- **Algorithms** — traditional hypothesis-driven analytics as well as machine learning AI-based analytics
- **Data Exposure** — high-velocity robotic data, EMR & EHR, consumer data, insurance and debt data, SNOMED
- **Relevant technologies** — Big Data tools, TICK Stack, business intelligence tools, multiple programming languages (such as R, Python, PySpark, SQL, HQL and HIVE) and data science packages

Yet one of the biggest challenges of performing real-time data analytics in an industrial environment is massive high-velocity data streaming from sensors and machines. Since IoT data is time series data, what they needed was a purpose-built time series platform with high ingestion and built-in functions to store, process and evict time-stamped data.

The solution

> *"Fast ingestion and configurable retention policies were absolutely needed, and InfluxDB just does the best in this world."*
>
> ***Pankaj Bhagra***

Why InfluxDB?

Nebbiolo and AnalyticsPlus chose InfluxData's open source time series platform, the TICK Stack (using three of its components Telegraf, InfluxDB and Kapacitor) because they needed a mechanism for collecting, storing, and processing data locally, rather than pushing it to data lakes.

On the edge, industrial users are mostly interested in one week's or month's worth of data rather than historical data dating back years, and only the important data is replicated to the big data lakes. To meet that need and keep stored data volumes under control and at the desired granularity, configurable retention policies and auto-purging (downsampling) capability were critical — and a main reason why InfluxDB was selected. InfluxDB and Kapacitor are used to provide the foundation for

performing analytics either at the edge, in the data center or on the cloud. Telegraf, the TICK Stack's open-source, plugin-driven server agent, was chosen for metrics collection.

Below is a summary of the platform's technical requirements and how TICK Stack met them.

	Requirement per Node	How TICK or modified TICK fits the bill
1	Fast data ingestion – up to 100k rec/sec/node	Best in class
2	Configurable retention policies	Works like a charm
3	Query response. Used for visualization. 10s of queries with response under few sec.	Reasonable if the queries are well-behaved. Purge the long queries.
4	Native visualization with aggregation and filters	Grafana with SSO
5	Streaming analytics support	Kapacitor with UDFs
6	Metrics collection	Telegraf with extension for Linux and Windows
7	High availability and distribution	Clustering built-in. No single point of failure (SPOF).
8	Security, authentication, role-based access control (RBAC)	Using Encryption, Auth, RBAC, certificates. Hooks from TICK allowed stitching it seamlessly with their SSO.
9	Open and Extensible	UDF, graph to TICK, SSO Integration, Grafana
10	Lower footprint as each node has TICK stack	Best in class

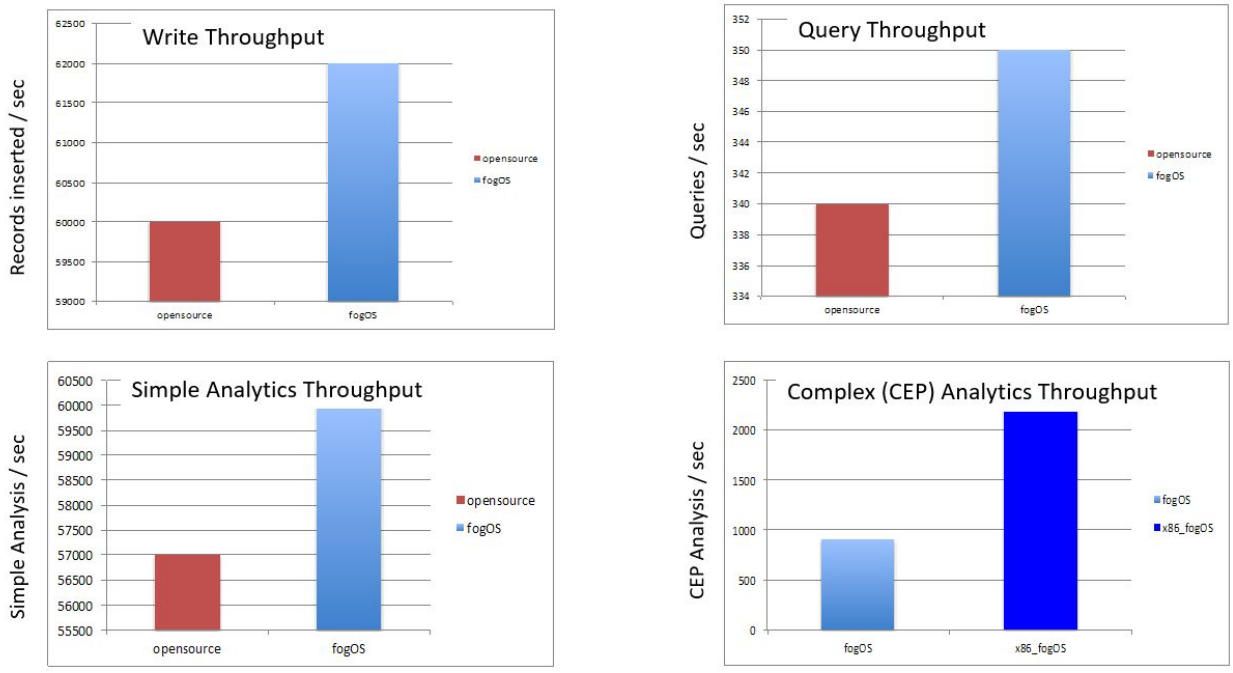
Testing and tuning InfluxDB performance for the platform

They had to tune the nodes in the TICK Stack to ensure they are rightly configured and optimized for the platform, and that they're able to meet or beat the open source numbers. Below are highlights of their benchmarking test, based on an i-5 class of node, 4 HTC/4G RAM:

- High data ingestion rate of 100k rec/sec for simple analytics and 5K rec/sec for complex analytics with ML
- 3x improvement of SciKit, NumPy and other ML libraries on Intel CPU acceleration with custom Python build
- Data distributed and managed across hundreds and thousands of nodes
- Highly optimized for edge and security in layers with role-based access control (RBAC) and single sign-on (SSO)
- Ability to store 1M+ series and 4B+ records per fogNode
    - Their customer wanted to query data through multiple tags (new tags are generated as new parts are produced).
    - This resulted in series explosion.
    - Adopting InfluxDV 1.6 (the latest at the imte) with the TSI engine saved the day by storing the indexes much more efficiently.

InfluxDB benchmark tests were done for various platforms (Advantech, Dell, Kontron, Siemens). For example, in the Advantech MIC-7900 — a standard industrial PC that comes with four-core Intel i5 series processor, 8 GB RAM and 256 GB disk — the write throughput hit 62K records per second, as shown below.

Benchmarking – Advantech MIC-7900



A significant amount of work has also gone into tuning the machine learning libraries since finely tuned libraries are necessary for real-time insights.

Platform capabilities overview

Powered by the InfluxDB platform, the IoT Edge Computing Platform brings the flexibility of virtualized computation, network and storage resources to the edge, as an integrated solution with centralized management combined with a rich set of machine learning and artificial intelligence libraries as part of the fogOS middleware. The distributed analytics platform allows users to configure and deploy data pipelines for various purposes.

- **Edge platform for compute, network, storage, security & analytics** — works for any kind of streaming data
- **Central management for distributed apps** — distributed compute nodes can be managed through a centralized management system
- **Edge virtualization for industrial apps** — enables running VMs and containers at the edge
- **IoT Security** — features that add to the security of the system
- **Edge analytics deployment at scale** — enables real-time decision making on the edge
- **Programless AI/ML pipelines** — small learning curve that does not require users to be data scientists or statisticians to run analytics to solve business problems

Platform standout features

The platform's distributed, centrally managed federation of fogNodes enables time-sensitive and secure 360° communication, device and data management, analytics and application — all positioned between endpoints and clouds.

- Cloud at the Edge bridges the intelligence gap between IoT devices and Cloud infrastructure.

- The federation of fogNodes is distributed and centrally managed.

- Defense is provided in-depth for IT/OT bridge with granular network and container-based segmentation.

- Real-time analysis and control using Kapacitor and ready-to-use user-defined functions (UDFs) provide utmost flexibility for any kind of analytics. User-defined functions were developed (such as quality assurance, quality checks and machine learning algorithms including neural network and anomaly detection algorithms). Algorithms are chosen to run on a given set of data depending on the situation at hand.

- fogSM provides a cloud-based centralized management platform that enables Zero-Touch deployment and provisioning of devices at the edge.

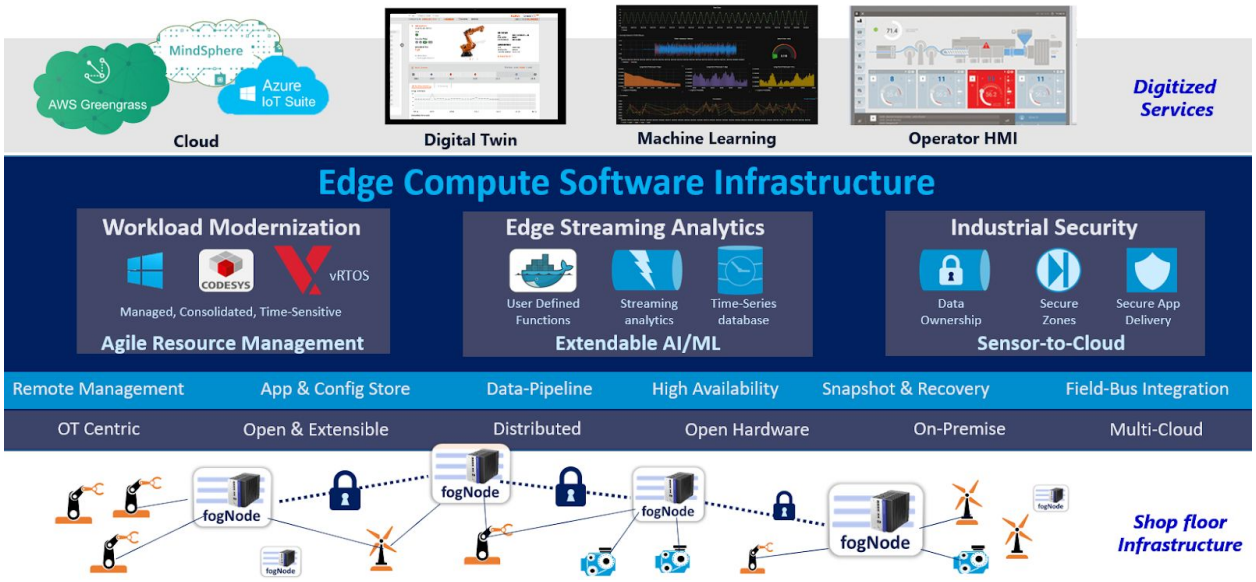
- Intuitive, drag-and-drop UI allows building and deploying an analytics pipeline in minutes without data science expertise. For example, factory floor supervisors can use the UI to easily

extract insights about the machines that they're supervising. The complexity and depth of analytics live underneath, and get filtered in, the drag-and-drop UI.

Technical architecture

> *"We have definitely needed streaming analytics capability, and Kapacitor and User Defined Function (UDFs) did the trick for us."*
>
> ***Pankaj Bhagra***

Nebbiolo's Edge Compute Software Infrastructure: Providing Agility, Insights and OT/IT Security



This modular and converged software platform for the Industrial Edge modernizes any industrial compute (X86 and ARM) with virtualization and containerization and provides central management for compute, storage, networking, security and analytics (on-prem/on-cloud). It has three pillars:

1. **Workload modernization/edge virtualization (managed, consolidated, time-sensitive)** — workloads handled include legacy, real-time, and modern workloads for IoT.
2. **Industrial security (zoning, Intrusion Prevention System - IPS, data ownership)** — encompasses data ownership, secure zones, and secure app delivery. Data is ingested with a unique ID maintained throughout the system, and only the right users with role-based access can access the data. Users can also choose to encrypt the data with their own specified keys. As the data leaves the platform, it can be encrypted out only by the user keys.
3. **Edge streaming analytics (programless, real-time, extendable AI/ML)** — allows deploying data pipelines from the cloud, in a programless fashion, at the edge to run statistical analysis or AI/ML. Streaming analytics that run at the edge employ a Kapacitor user-defined function and store data locally on the nodes.

Each of the platform's nodes is deployed at the edge and has its own InfluxDB instance, its own streaming analytics infrastructure, and its own capability to visualize the data locally and choose what data to send to the adjacent node or to the cloud. That whole infrastructure runs on each node and is managed from the central entity, fogSM.

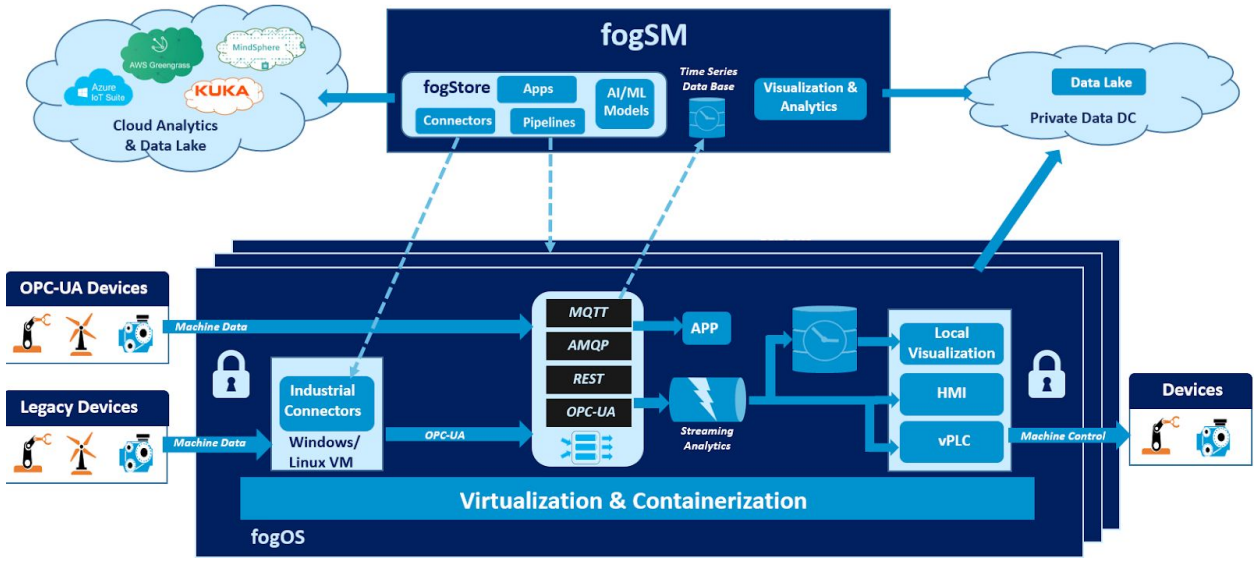
Nebbiolo used Kapacitor's user-defined functions to add the above capabilities. In their open API, you can extend the user-defined functions. You can bring your own UDF, apply it to the pipeline, and manage the pipeline from fogSM.

As shown above, key platform highlights include:

- The platform can be remotely managed.
- App & Config Store allows users to onboard and push their applications securely and in a scalable fashion to multiple nodes.
- Data pipelines are composed in a simple user-defined canvas, configured in fogSM and then pushed to the edge.
- The system is highly available with no single point of failure. If one node in the cluster fails, an adjacent node picks up the workloads. The clusters at the edges and other distributed nodes are fully resilient and highly available.
- Snapshot & recovery capability is available for applications which are running.
- Field-bus integration allows data ingestion through integrations.
- The system is fully open and extensible, distributed, runs on any hardware, on-prem or on cloud, and has the ability to do the cloud federation with a multi-cloud.

The platform provides full, holistic functionality from data ingestion to data cleaning, storage and visualization. fogSM is the central management station, and the "Virtualization & Containerization" is the execution agent, as shown below. The illustration shows applications that run on top of the platform's infrastructure and offers a high-level view of how and where the platform deploys the TICK Stack: on the cloud, on each edge, and replicated across the clusters.

Infrastructure Overview: Applications



The data flow is as follows:

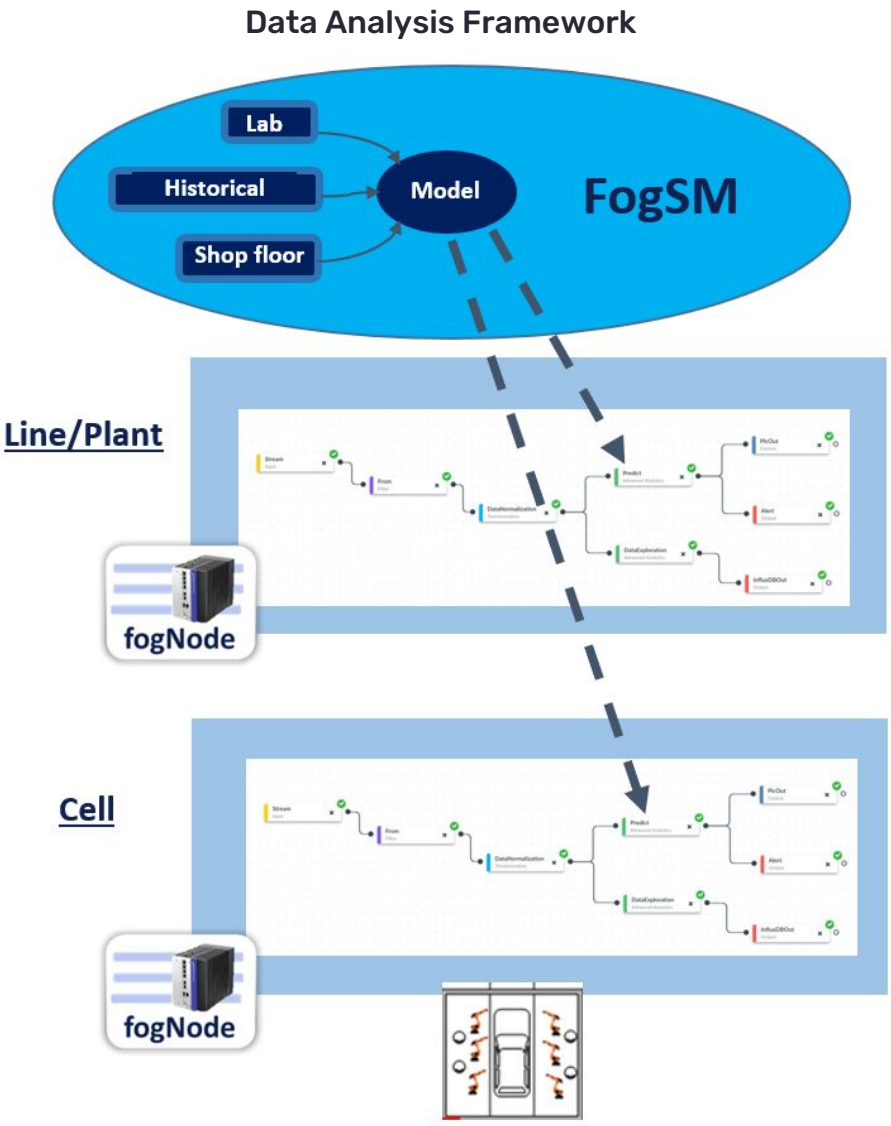
1. Industrial connectors provide data ingestion capability from devices (since only 10%-20% of devices talk to standard protocols such as OPC UA or MQTT, and the rest of devices have legacy protocols).
2. Streaming analysis is where Kapacitor is deployed and where streaming data is analyzed, processed, and used to drive insights.
3. On each node, the data broker runs to provide publish-subscribe semantics where a topic-based subscription can be set. Multiple devices can send the data to the data broker bus, where various consumer applications can consume the data based on their topics of interest.
4. Once you subscribe to topics, you specify what you want to transform the data: clean it, do a statistical sample or perform machine learning with it.
5. Based on subscribed topics, users build data pipelines — which specify how to analyze, prepare, and visualize data. Then, the data goes to the streaming analytics infrastructure.
6. The inference engine runs on the topics of interest and is where the streaming analytics is managed in the form of the user configuring data pipelines. Data pipelines are built in the fogSM and pushed down to the edges.

7. The entire data is stored in InfluxDB locally and can be locally visualized, which enables actions or insights to be driven to the machines.

Data analysis framework

As shown below, the model is built by looking at the historical data, lab data, and continuous shop floor data. Using that data, the models are continuously built. The more label sets available, the richer and more accurate the model.

Once the model has been built, it is pushed down to the inference engine which runs on the edge. As the hierarchy illustration shows, in an industrial pyramid, there is the cell, where the real production gets done. At the level above that —the line/plant level — there is another layer of compute.

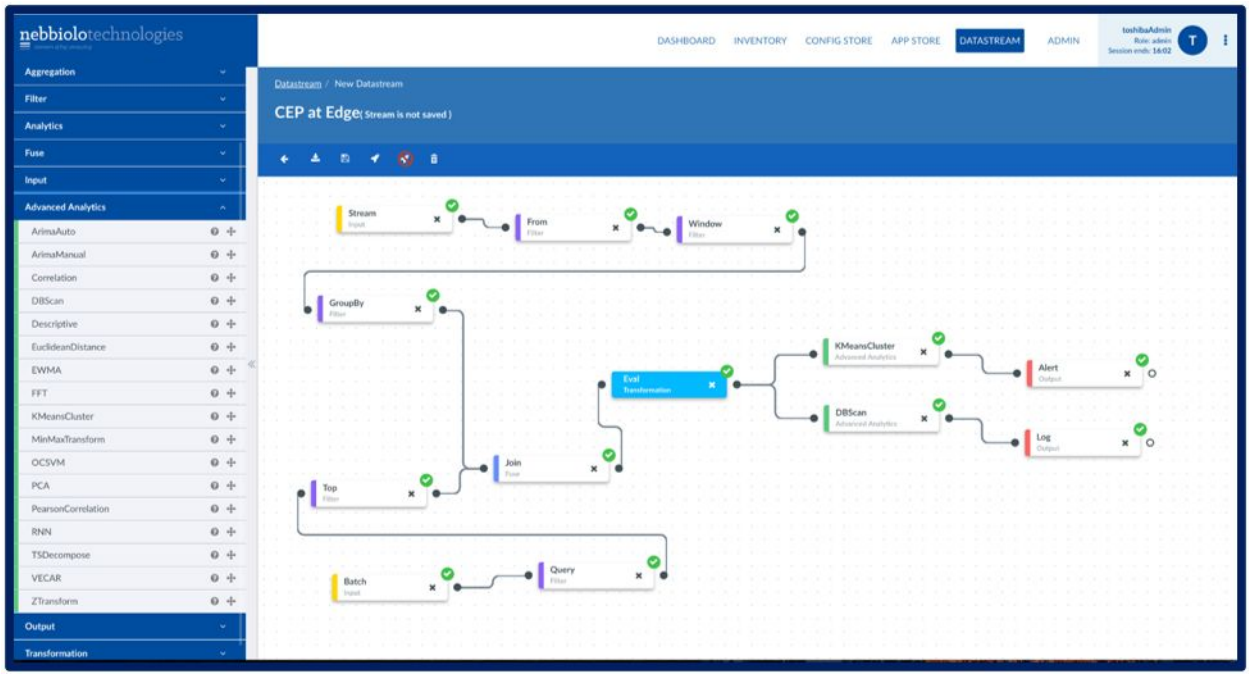


You can push your model either at the plant level, at the aggregated level, or all the way to the lowest level of your model where you want to run it. The data pipelines are primarily Kapacitor functions with added user-defined functions. So you start producing a data stream and specify what to do with it (clean, analyze, predict). All of this runs in a distributed fashion, in which you manage and view the results from fogSM.

Analytics pipeline

The data pipeline for analytics and visualization is configurable and extendable. It supports complex event processing like sensor fusion and anomaly detection, without the need to write a single line of code.

Analytics Pipeline



The above screenshot shows standard Kapacitor and InfluxDB nodes. Here is how the analytics pipeline is built:

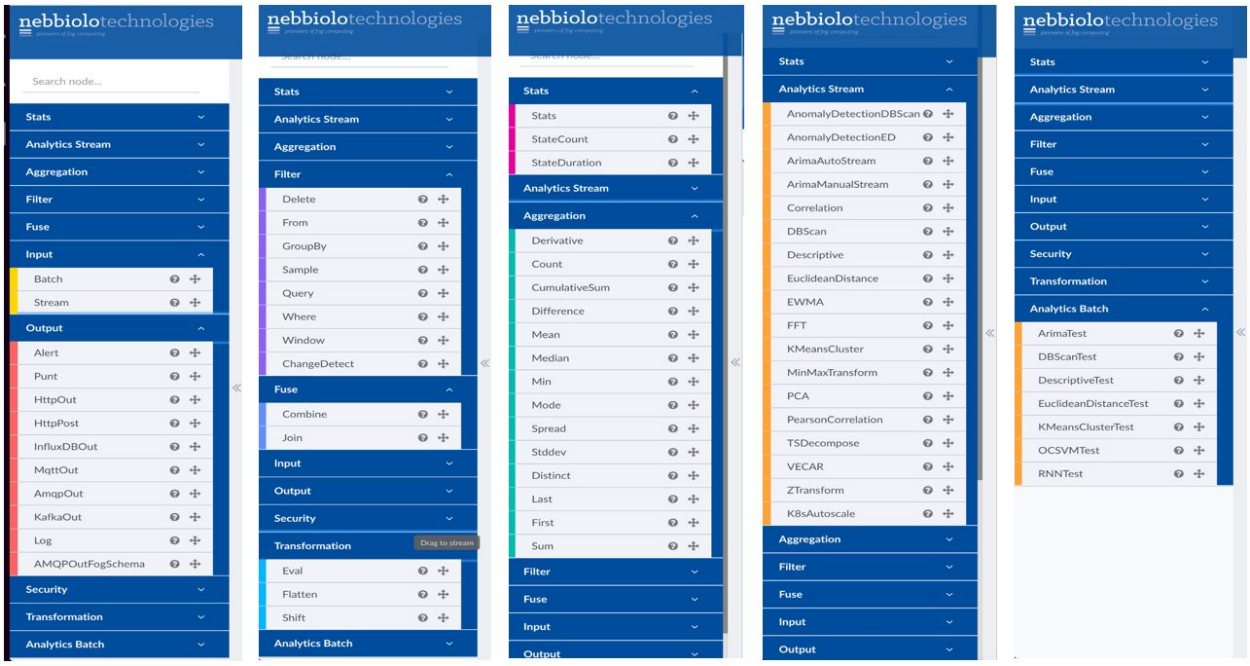
- The stream processing is followed by writing the TICKscript.
- Then, a topic selection is made, and the vendor nodes and additional nodes are used for filtering transformations.
- You can join multiple streams and then do a lambda transformation.
- Through the EvalNode, the TICKscript allows you to have a directed graph. You can split your graph, so you can take the same stream and split it into two parts and apply different functions on different parts of your tree traversals.

- Many platform users have multiple models. Since they're unsure during the data cleaning phase which model will yield the best result, they expose the data to multiple models running in parallel and examine the accuracy of each, then determine which model to adopt.
- Nebbiolo and AnalyticsPlus have written wrappers (shell scripts that embed a system command or utility) on top of TICKscript, to simplify the process for non-data-science experts and provide them with toolsets which they call the “lego” set, to enable drag-and-drop functionality.
- Next, in the path for input, you select whether you want to start with a stream or a batch.

The left menu in the above screenshot shows various transformation functions and user-defined functions.

Below is another screenshot displaying the data pipeline’s built-in function menus. Extensions have been added to accommodate the user community’s need to use MQTT, AMQP and Kafka. One interesting node is the PuntNode where the packet is punt out of Kapacitor to an additional docker. This extends the pipeline to any function, avoiding the need to write the function inside Kapacitor.

Data Pipeline: Built-in Functions



For a video demo of how this IoT Edge Computing Platform works, watch the AnalyticsPlus & Nebbiolo Technologies webinar.

Results

> *"There are use cases in quality control and predictive maintenance...where you want to prevent any kind of anomalous behavior that could develop in any particular machine, so that the damage is much less than it would have been if there was no kind of analytics happening in real time."*
>
> ***Anil Joshi,** CEO, AnalyticsPlus, Inc*

The IoT Edge Computing Platform built by AnalyticsPlus and Nebbiolo Technologies is a complete platform for industrial automation. It meets Industry 4.0's need for a cloud-inspired infrastructure closer to the end devices with real-time, scalable, safe and secure capabilities. The platform brings cloud-like functionality closer to data-producing sources and merges real-time capabilities to integrate IT and OT in a single platform. It simplifies and accelerates real-time analytics from concept to deployment; is open and extensible with a small footprint; reduces cloud data transport costs; enables real-time decision making at the edge; modernizes security; and drives overall ROI. The platform can be used in various verticals. Here are some common industry use cases.

Industrial automation

Edge computing can help create machines that sense, detect, and learn things without having to be programmed. For example, if the sun shining through a window hits a machine for part of the day, the machine will eventually be able to tell that the temperature change doesn't mean that something is wrong.

Predictive maintenance

Edge computing can help detect machines that are in danger of breaking, and find the right fix before they do. Alerts for what's happening with a machine are best done close to that machine.

Healthcare

Edge computing offers exciting new possibilities for delivering patient care. With IoT devices capable of delivering vast amounts of patient-generated health data (PGHD), healthcare providers could potentially have access to critical information about their patients in real time rather than interfacing with slow and incomplete databases. Edge computing could make a significant impact on the delivery of healthcare services to hard-to-reach rural areas. Medical devices themselves could also be made to gather and process data throughout the course of diagnosis or treatment.

Blockchain

Distributed ledger technology like blockchain requires decentralized computing models. Each node in a blockchain is a compute unit, so blockchain isn't a centralized ledger; it's a distributed ledger. Therefore, it's edge. Today, there is almost 40 to 60 billion dollar of fraud and abuse in healthcare claims filing. But if you can create a blockchain where there's a linkage that goes through the patient, now, any kind of fake claims that are coming from a hospital or a clinic to the insurance company could all be avoided because you need a signature from the patient or the approval of the patient, and this all could be done locally.

Retail

Several retail chains are creating more immersive in-store environments with technologies like augmented reality (AR) to attract additional shoppers. This requires lower latency, which is where edge computing capabilities come in. An AR application example is one that allows you to see how an outfit will look on you through augmented reality, without having to go to the dressing room.

Connected homes & offices

Many people use Amazon Alexa or Google Assistant to complete tasks like turning on lights on-command, or changing the temperature. However, right now those tasks tend to take a few seconds to occur. With edge computing, it will be possible for them to happen in near real-time.

Software-defined networking

Software-defined networking technologies, some of which will power the move to 5G, require local processing to determine the best route to send data at each point of the journey. Each node and network can make a decision about the quality of service it needs to give a particular piece of information that comes to it, and then route that in a different way. It might jump protocols, and it might go from wi-fi to cellular or back again all over.

Autonomous vehicles

Self-driving cars need to be able to learn things without having to connect back to the cloud to process data. Machine learning techniques such as reinforcement learning don't rely on training large models with big data sets — instead, you can run inferences directly in the car, which is essentially edge computing.

Industrial IoT

Below are two sample industrial IoT use cases, involving data management and streaming analytics for inline quality control. Through these use cases, the platform is driving real-time data-driven insights at the edge. The general objectives common to both use cases are:

- Predictive and inline detection of failures
- Data-driven sampling of parts to be inspected

IIoT use case 1: A major German manufacturer's welding and riveting shop

The shop faced the problem of having to do random sampling of the car parts they're producing. Instead, they wanted to sample each part and use the analytics and data-driven models to predict which part is going to fail. Because the quality control was not done on every part, they were overproducing the welds by 20% more than necessary, which impacts production cost and time as well as vehicle weight.

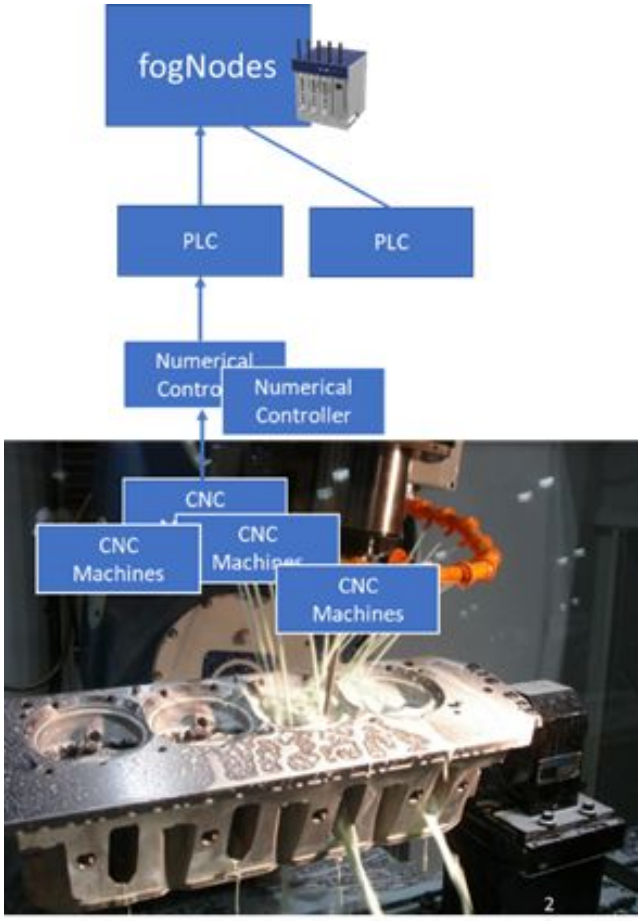
Through the platform, the customer is now able to perform inspection and analytics on each part, by gathering the sensor data and doing analysis at the edge, thereby shaving off a significant portion of the overcommitting of the welds. Given the amount of data being produced, this couldn't have been done otherwise for every part produced across multiple factory lines, and would have even been hard to do in the cloud. They needed immediate response time so that if continuous parts being produced are identified as having bad quality, they could change the control loop (which is 18 milliseconds) and fix the weld head to produce subsequent parts with a better quality. Doing your analysis, identifying the events, determining the control event, which fixes the subsequent part, all within 18 milliseconds absolutely requires real-time streaming analytics. The platform's streaming analytics successfully identify these bad parts and impact the quality of the subsequent parts within that tight loop.

Welding & riveting shop (Germany)

The platform enabled the shop to achieve the following:

- From manually testing 1 car/day to inline predictive testing of all 20,000 welds for all cars for quality
- Continuous Improvement of predictive AI/ML models in the cloud and flexible deployment at the edge

IIoT use case 2: A major Alabama-based car engine head manufacturer



Engine head manufacturing (USA)

The platform is used to generate timely application of insights for quality control of the engine heads produced. Instead of random sampling, the platform identifies which part is going to fail, looks at each data sensor, and predicts using various models. This is where the AnalyticsPlus and Nebbiolo data science teams figure out which part has a high predictability of failure. This architecture can replicate itself for similar use cases and even in adjacent verticals.

The platform enabled the manufacturer to achieve the following:

- From 1% manual inspection of engine parts to 100% automated inline testing
- Data & ML-Models across 5 lines & 25 CNC machines aggregated to predict failure across all lines

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