influxdata

How Houghton Mifflin Harcourt gets real-time views into their AWS spend with InfluxData

AN INFLUXDATA CASE STUDY

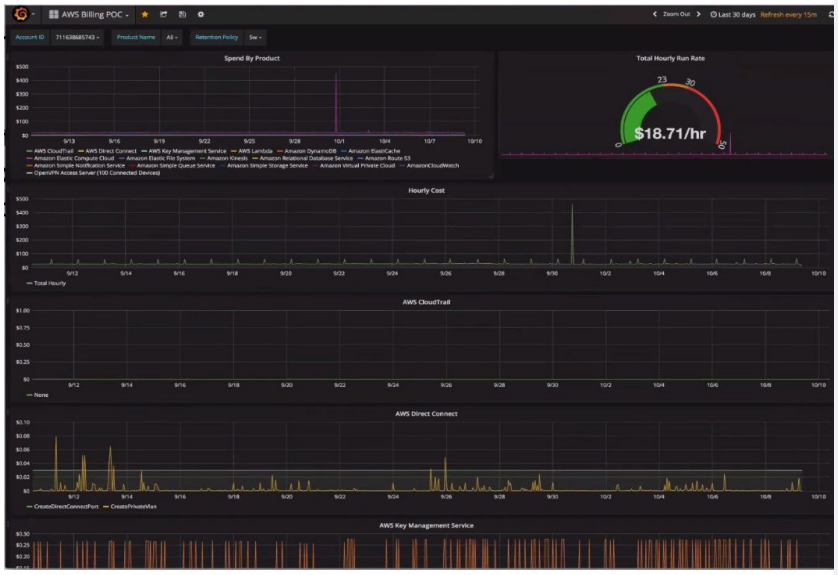
Robert Allen
Director of Engineering, Houghton Mifflin Harcourt

October 2017

Overview

Houghton Mifflin Harcourt (HMH) faced three business challenges: gaining real-time visibility into their AWS spend that covers several accounts across multiple business units and many developer teams, optimizing its DevOps monitoring, and gathering and tracking KPIs to measure the company's online educational business.

Since all these challenges involve time series data, the HMH Bedrock Platform Technical Services team chose InfluxEnterprise. They take full advantage of InfluxEnterprise's unique set of features—such as flexible retention policies and control over downsampling without sacrificing performance—and also deploy InfluxData's metrics collection agent Telegraf, to optimize spend and performance and inform decisionmaking.



About Houghton Mifflin Harcourt

Headquartered in Boston, Houghton Mifflin Harcourt (HMH) is a nearly 200-year-old publishing company that has been "changing people's lives by fostering passionate, curious learners." The company primarily focuses on K-12 education, both learning materials and assessment materials, and publishes textbooks, instructional technology materials, assessments, reference works, and fiction and nonfiction for both young readers and adults. Some of its most famous books include Lord of the Rings, Tolkien's works, and Life of Pi.

As a global learning company, HMH specializes in pre-K–12 education content, services and cutting edge technology solutions for today's changing landscape. Available through multiple media, the company's content meets the needs of students, teachers, parents and lifelong learners, no matter where and how they learn.

> *"It used to be, without having this level of insight into where your expenses were at, it was sort of just a black box."*
> *Robert Allen, Director Engineering*

The Business Problem

Houghton Mifflin Harcourt creates engaging, dynamic and effective educational content and experiences from early childhood to K-12 and beyond the classroom, serving more than 50 million students in more than 150 countries.

Problem: Tracking AWS cost across accounts and business units
The company needed real-time cost visibility into its many accounts, each of which manages its own services. Each month, they ingest about 23 million line items from AWS (the AWS Cost and Usage report contains line items for each unique combination of AWS product, usage type, and operation that an AWS account uses). Each of these line items is converted into points and is at one-hour granularity, so every line item represents the billing for one hour.

Problem: Optimizing infrastructure cost for each individual product
Engineers were incentivized to just get the infrastructure working with no true hard regard as to what the cost was. They lacked that closed loop of feedback into what their infrastructure decisions are costing.

Problem: Measuring performance of online educational business
The HMH team had been collecting most of their KPI information using Elasticsearch, such as how many students were rostered for a particular school district. That metrics collection was done in a very mechanical fashion using log parsing. They needed a more dynamic and automated way to track those metrics.

> *"Some tools are built for time series data, some are actually just indexes. For me, Elasticsearch is an index. It's an index first, database maybe second. That's where I'm at with that."*

Solution

Why Telegraf?

The HMH team found Telegraf to be very clean and easy to configure, and it offered much-needed control around the automation that goes into it. Features they found attractive include the ability to:

- Consume metrics for all aspects of the infrastructure
- Easily develop custom plugins and applications for metrics capture

- Capture once and ingest by multiple services
- Operate on one-to-many metric sources

> *“We change things a lot, but we don't generally change them on a whim, and there has to be a pretty compelling reason to do it. Telegraf gave us plenty of reasons that it was compelling to make the move. It's been a boon for us, really, because making the switch from what we were doing to Telegraf was generally a very painless process.”*

Why InfluxDB?

The HMH team chose InfluxDB because they:

- Tried other Time Series Databases, such as Druid, and found them lacking
- Wanted persistent data storage
- Did not want to sacrifice resolution of data for storage or performance
- Wanted developers to be able to adopt the solution easily with minimal disruption
- Demanded minimal operational overhead
- Demanded tagging/labels for robust/dimensional data

InfluxData features that attracted HMH included:

- InfluxQL (a SQL-like query language)
- Easy to use and operate
- Robust and performant tagging
- Retention policies fundamental to InfluxDB
- Continuous Queries / Kapacitor that provide first-class workflows
- Suite of tools in one platform working to a common goal

One feature HMH really found attractive from InfluxDB was the ability to support individual databases for each of their development teams. They currently have approximately 25 of what they call “roles” developed—each being a functional team or group of individuals working to solve a particular product or a platform service problem within the company. Each of those is able to maintain their own databases and retention policies with InfluxDB:

- Platform metrics are stored in a common database
- Retention policies allow segregation of data responsibilities
- Annotation events from curl calls and others, for deployment, are pushed into InfluxDB and then used for annotations with Grafana

HMH also chose InfluxData over Prometheus for DevOps Monitoring, given InfluxData's InfluxQL (SQL-like query language), its time series storage properties, and the fact that they found it improving over time with every release.

Why Kapacitor?

Kapacitor brought several advantages that helped solve recurrent data challenges:

- Teams maintain their own workflows
- Templating of various workflows DRY (Don't Repeat Yourself)
- The ability to write User Defined Functions
- Alerting for PagerDuty, Slack, Webhooks, etc.
- Advanced downsampling and transformation workflows

Technical Architecture

AWS Programmatic Billing provides very detailed data (a form of multi-billing that comes out in a massive CSV file). By nature, that data is rolled up at one-hour intervals. Users can be as detailed or broad in their dimensions using custom tags or other forms of reporting files. The HMH team found that InfluxDB v1.3 handles large cardinality extremely well when TSI is enabled.

Monitoring AWS Billing with InfluxDB

Two Retention Policies

Two retention policies are set for two distinct data types:

- The 5-week retention policy stores non-invoiced data and is used for day-to-day monitoring and for investigative purposes.
- The Unlimited retention policy stores all invoiced data and never expires.

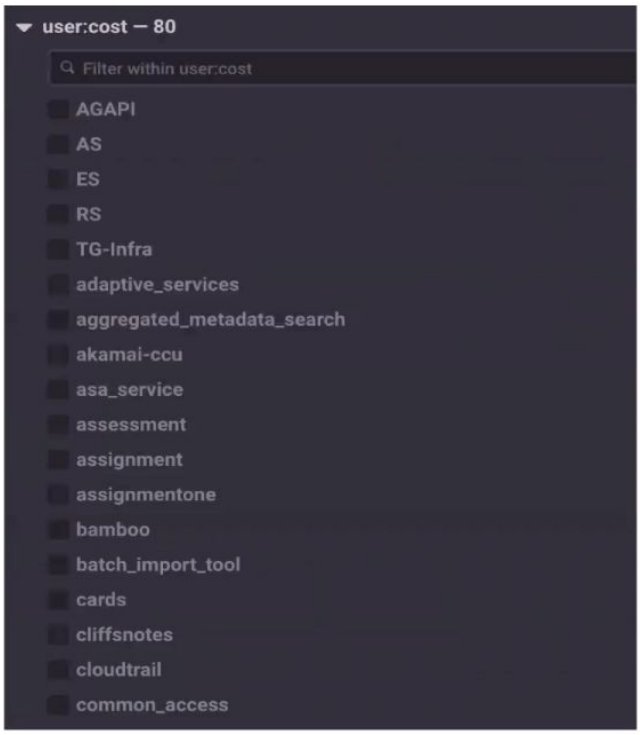
Retention Policies: 5 weeks and unlimited

```
> show retention policies
name    duration shardGroupDuration replicaN default
----    -------- ------------------ -------- -------
autogen 0s       168h0m0s           2        true
all     0s       24h0m0s            2        false
5w      840h0m0s 24h0m0s            2        false
>
```

HMH currently has approximately 4 years of data for all of the accounts resulting in 23 million line items/month.

Project Cost Allocation

Custom tags enable reporting on cost analysis at the application/stage level in detail:



A tag is a label that the user or AWS assigns to an AWS resource. After cost allocation tags are activated, AWS uses them to organize resource costs on the cost allocation report, to facilitate categorizing, and to track AWS costs.

Custom tags are used for cost allocation of products. This provides a near real-time way, within 24 hours, to monitor expenses or costs with infrastructure for each individual product, which has helped create awareness about spend throughout engineering at HMH.

Collection of KPI Data
Telegraf is used to capture the telemetry's KPI data, which is stored in InfluxDB as events, counters, and metrics—because the HMH team now has the flexibility of high cardinality, they are able to collect these metrics and roll them up using time series data.

Results

> *"I really feel that there's going to be a lot of different ways that we're able to use time series data in areas that we haven't even thought of yet."*

With the real-time visibility, time to value, and control that InfluxData has provided, Houghton Mifflin Harcourt's Bedrock Platform Technical Services team is now able to better align performance with the fiduciary aspects of infrastructure operation.

The resource ID's (captured by AWS) are used as tags in InfluxData to perform data drilldowns to discover all the different ways that the company's products are consuming AWS services and also to view cost changes to identify deviations from what they expect versus what they know. They can now view Spend by Product, Hourly Cost, and also break down their Run Rate by certain products.

Telegraf proved to be the perfect fit for their container-based environment, as it accommodates very specific ways to collect metrics and can run in the container with a given process and monitor it locally. Telegraf supported their goal of enabling hundreds of engineers to coexist and develop without adversely impacting the work of their peers, and to gain more control over what metrics they collect and how they collect them.