



BLOGS

An Inside Look at FINRA's Massive Data & Analytics Architecture



SEP 28, 2021

by **Okera Marketing Team****Learn more**

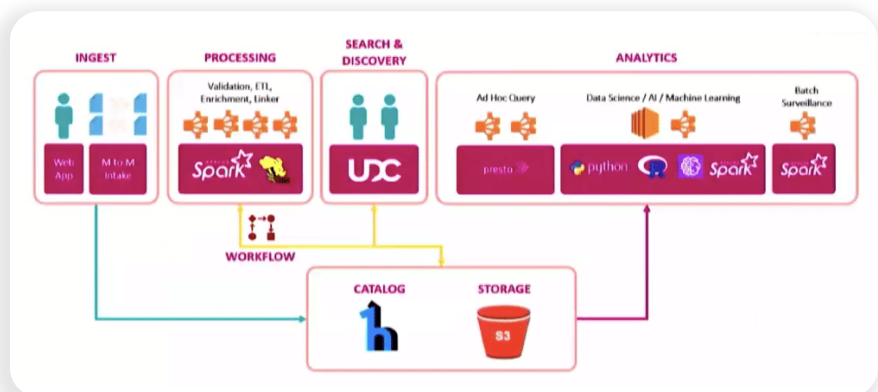
If you're in the financial services industry, you know [FINRA](#), the Financial Industry Regulatory Authority. It's the quasi-governmental regulatory body that keeps our financial markets safe. FINRA monitors an incredible 445,000,000,000 (that's billion!) market events per day to detect and prevent fraud, abuse, and insider trading in the US financial markets. As you might imagine, an enormous amount of data is involved in supporting their mission to protect investors and market integrity.

At the 2021 [MIT CDOIQ Symposium](#), Aaron Carreras, Vice President of Data Management and Transparency Services Technology at FINRA, and Nate Weisz, Senior Director of Data Management at FINRA, shared an in-depth look at the impressive data and analytics architecture their team has

6:24

The team generously agreed to let us share the story of how they built and manage an elastic infrastructure that supports, up to 150,000 compute nodes processing over 200 petabytes of storage.

FINRA's Data & Analytics Infrastructure



Like most enterprises, FINRA started its analytics journey using traditional relational databases. But rapid growth in data volume and the corresponding costs associated led the data and technology teams to re-evaluate their infrastructure strategy.

The team ran into four main challenges using on-premises infrastructure:

Rapid Data Growth and Infrastructure Cost:

Building for rapid data growth – 20-30% year over year – on-premises was difficult and costly. Aaron shared, “We constantly had to reassess where market was going... The choices were either buy additional hardware or try to scale up again. It was taking away from our ability to support our core mission.”

Data Governance and Data Management: The team often had to shuffle data between different analytics platforms, which created challenges around data governance. It wasn't easy to keep track of data location or to determine which data could be moved to a different storage appliance to allow the necessary performance to complete the jobs at hand.

Benefits of Moving to a Cloud Data Lake

“When we decided it was time for us to take a different path, the choice was obvious. We needed to move to the cloud,” explained Aaron. They decided on Amazon S3 for all of their data, and he put their reasons succinctly, “It was **one location, scalable, durable, performant, and cost-effective.**” By using a global provider like AWS, their data was also **cross-regionally replicated** and **redundant**. This means that they could continue with business as usual, even if one cloud region failed.

Another feature inherent to cloud data architecture that FINRA saw being critical as it evolved was the

ability to separate storage from compute. Aaron explained, "We wanted the ability to run transient nodes, have elastic compute, pay market pricing, be engine agnostic, and evolve the ecosystem as we saw new technologies evolving – without impacting our production systems." Nate demonstrates a real-world example of how they took advantage of this to make migration much more agile and low-risk, so watch the full session below if you're interested in this topic.

Last but certainly not least, Aaron shared that, "Pretty soon after getting into the cloud, our security group started to say that **the security posture could very well be better in the cloud than we had in our data center.**" As an oversight organization, FINRA's security has to be equivalent or better than the industry it's regulating. Data security is fundamental to Amazon S3 storage, with 99.999999999% (11 nines) durability.

(Data authorization is another story, and we'll talk about that a bit later. Spoiler alert: FINRA is using [Okera!](#))

Add a Data Catalog for Metadata and Data Governance

6:24

The great thing about data lakes is that you can pull in any kind of data, in any format – structured, semi-structured, unstructured. But there's also risk associated with this benefit. If you're not careful, your data lake can easily turn into a data swamp with no active management or data governance throughout the data lifecycle.

FINRA's solution to this potential hazard was implementing a data catalog alongside their data lake to provide contextual metadata. Like everything else, their data catalog and metadata efforts have evolved.

For the first couple of years, any data brought into the data lake automatically registered the format, the columns, the data types, S3 location – the basics that a data consumer needs to use that data. Since then, they've made their mission to build a more user-friendly UI and have added semantic metadata, such as business-friendly table and column names and descriptions.

So what can FINRA manage with its infrastructure?
A lot.

**Up To 450 Billion Market Events & 100s of TB
Day**

6:24

A market event is everything that makes a complex order lifecycle: the quotes, orders, trades, cancels, pricing, and reference data in the equities and options markets. FINRA validates all of this data through quality checks, requiring massive joins and transformations from ETL and huge scans and aggregation from batch analytics. And central to how it all comes together is event notifications (which Nate discusses in more detail in the session).

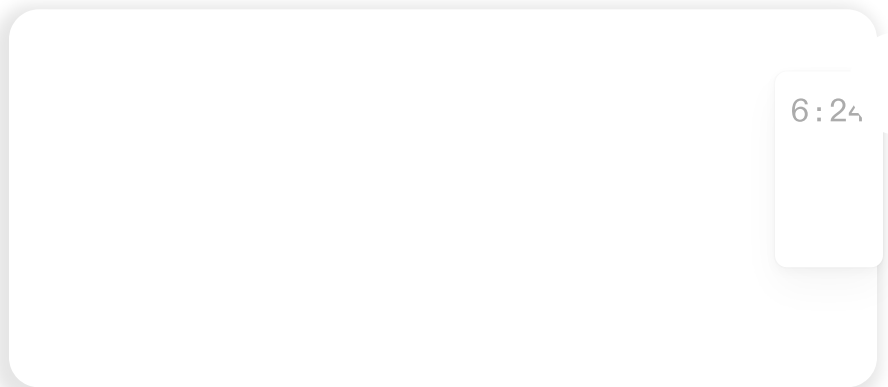
Up to 150,000 Compute Nodes per Day

6:24

Using the cloud allows the team to quickly spin up more computer power as needed. This is hugely different from the pre-cloud data lake when the team had to predict future hardware capacity requirements—and risk getting it wrong. Today, FINRA uses a lot of transient clusters to achieve this astonishing degree of elasticity and variability. Many of those clusters are sized variably according to the market volume. For example, suppose there's an ETL job doing the transformation. Upstream data is processed and arrives; a notification triggers the transformation job. The first thing the job does is look upstream at the number of rows and adjust the size of the cluster needed for the transformation.

Aaron also goes into an illuminating discussion of the enormous cost savings achieved by using Amazon's spot market—"30, 40, 50 percent of the price of the on-demand pricing"—so be sure to watch the session if you're interested in learning about that.

Over 200 Petabytes of Data Storage



S3 made scaling data storage straightforward and cost-effective, which comes in handy when you're storing 200 petabytes of data. Since deep archiving comes at such a major drop in cost, there is a strong economic incentive to move less frequently used data there. But since it's offline and requires an operational restore before it can be read, data users were hesitant to let any of their data be stored there. (What if they needed it! Would it get lost?) After testing large restores and bringing back several hundred terabytes reliably – “much, much more reliably than tape,” reported Nate — FINRA now has many petabytes of analytics data in deep storage, resulting in significant cost saving.

What's New and Next for FINRA's D&A Architecture?

As we teased earlier, FINRA is working on centralizing data authorization with Okera. “Analytics has kind of grown organically because we want people to use the best of breed and keep

up with the industry. Unfortunately, the data authorizations and entitlements are kind of done in that analytics layer, which is...inconsistent and it could be better managed centrally, we believe. So we're working with Okera."

Okera's software will help FINRA bring entitlement and data access controls across all of FINRA's data platforms. Nate touched on their enthusiasm for Okera's fine-grained access controls for engines like Spark and Presto, providing robust column and row-level access controls simply and easily.

The team continues to make considerable investments in advanced analytics. "It's going to be the next transformation for us... We've gone from going to the cloud and building a data lake. We're going to significantly increase the amount of AI and ML, and NLP we use in our regulatory work. To bring up a quote from my favorite sci-fi author, William Gibson, 'The future is here. It's just not equally distributed.' And that, I think, is a bit of where we are with our advanced analytics overall. We're doing really cool stuff. There are pockets of wild success, but we're making investments to bring the latest and greatest technology to all corners of FINRA's latest-gen analytic platforms. We're going to up-skill the people involved. And this is not just a data lake thing; this is FINRA wide."

Watch the session now:

6:24

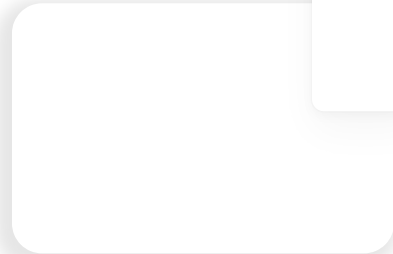
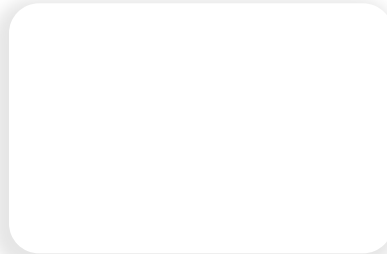
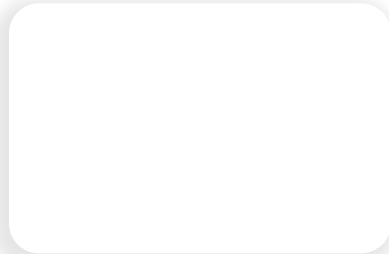
To learn how Okera helps companies gain full visibility into how sensitive data is used and how to standardize and simplify fine-grained access control across the enterprise, sign up for a [free demo](#) or [contact us](#) to learn more.

Share on Social



Recent Posts

6:24



BLOGS

A New Chapter in Our
Journey: Okera is
Joining the Databricks
Family

BLOGS

Distributed
Stewardship Now
Easier with Role-based
Templates

BLOGS

Easy Onboard
Experience for
Dynamic Data Access
Management

[Privacy Policy](#) | [Terms of Use](#)

©2023 Okera, All Rights Reserved. | Site by Go Fish.