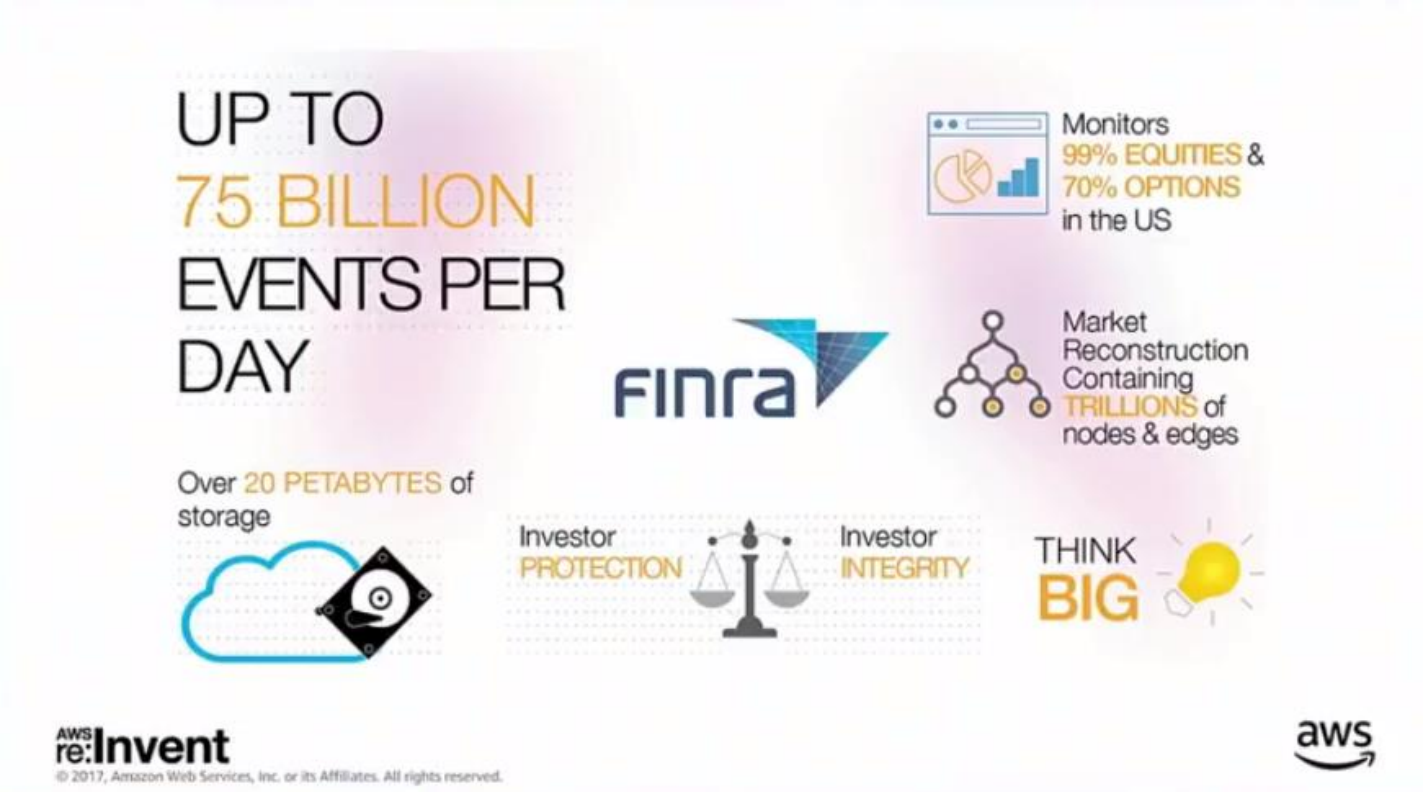
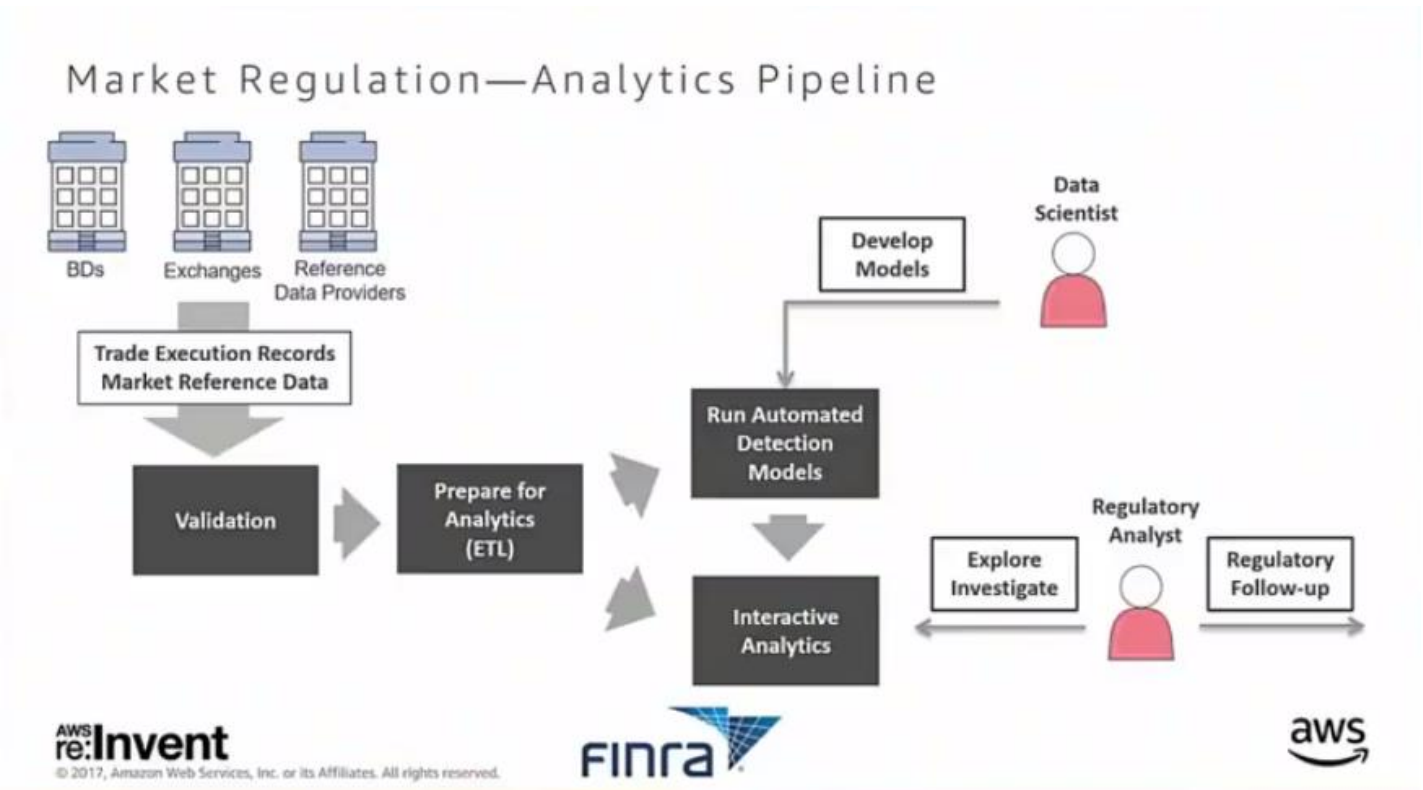
This talk will be about how FINRA created and handles its Big Data Lake in the cloud for its analytics use case being built by the Enterprise Data Platforms Group at FINRA.

FINRA is the financial markets regulatory authority that tries to keep fraud out of trading markets and create a fair-trading environment. FINRA gets up to 75Billion records a day from the various trading platforms, FINRA then does different analyses on the data to detect fraud issues





This is what the data lifecycle looks like. All the data comes in different forms like via FTP, direct to the cloud. We then have to make sure that the data is clean and can be trusted by data validation checks for the fields and entries. Next, the validated data is then put into standardized formats that we can then run surveillance against. We have above 250 patterns that are run in an automated way to detect fraud, we also have data scientists building new models for detecting new sophisticated frauds.

FINRA's Data Challenge

Keeping track of 40M+ tables can be a challenge...

What data do we have?

Where is the data used?

How many versions of this data exist?

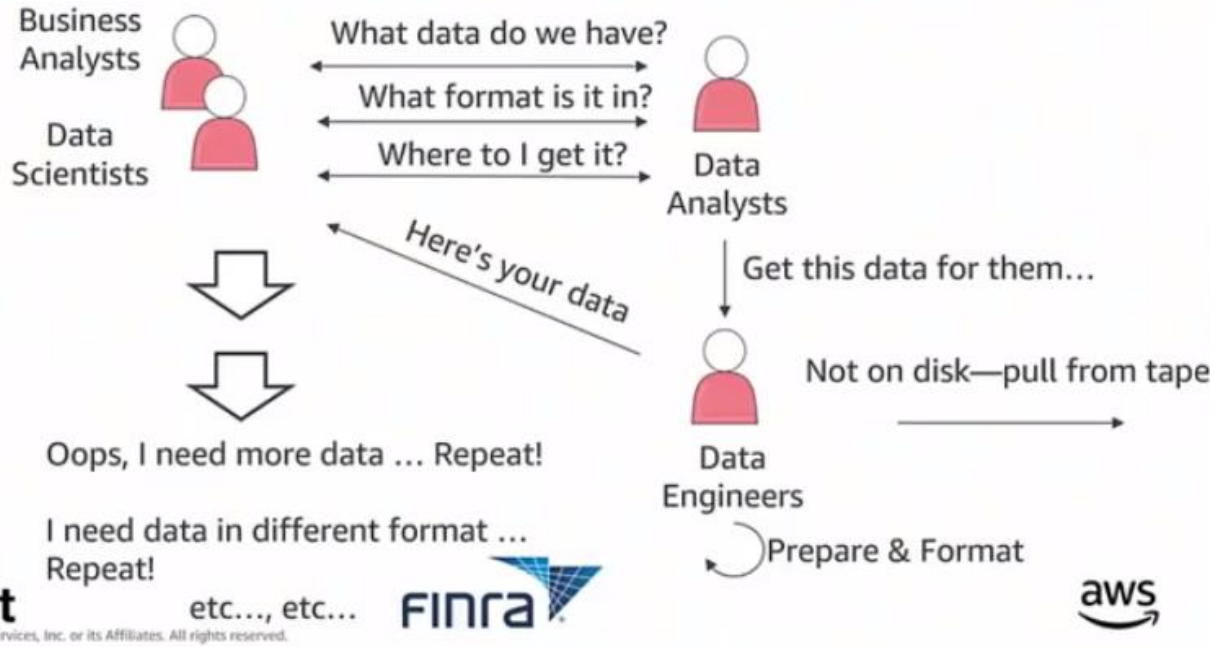
What is the source of this data?

What is the retention policy?

FINRA

aws

Data availability and analytics is complex



FINRA

aws

Infrastructure can be limiting & costly

Does not scale well as volumes and workloads increase

Duplication of effort in data management (data lifecycle, retention, versioning)

Data sync issues—manual effort to keep data in sync

Challenges to run analytics across fragmented data

Costly system maintenance and upgrades

FINRA

aws

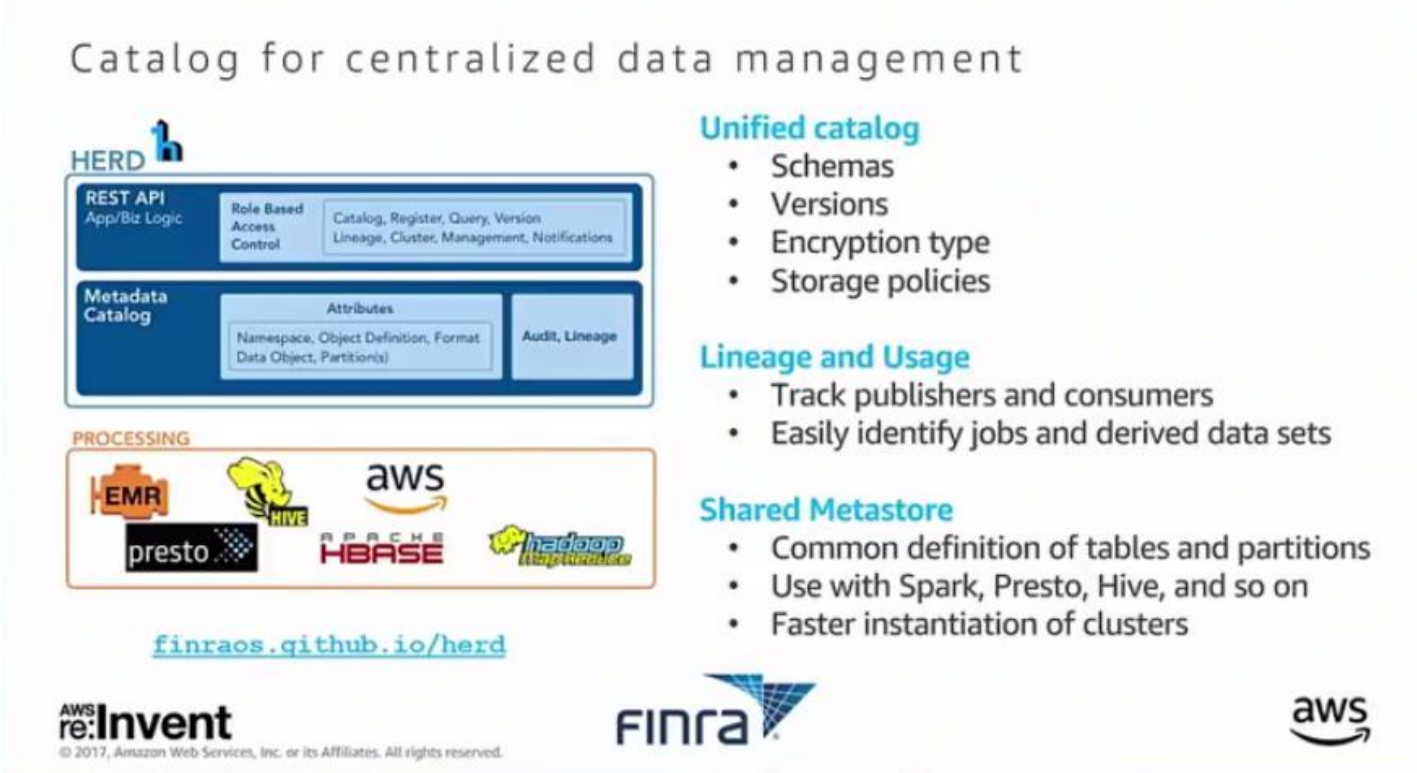
FINRA's Managed Data Lake in the cloud

Key principles of our big data architecture

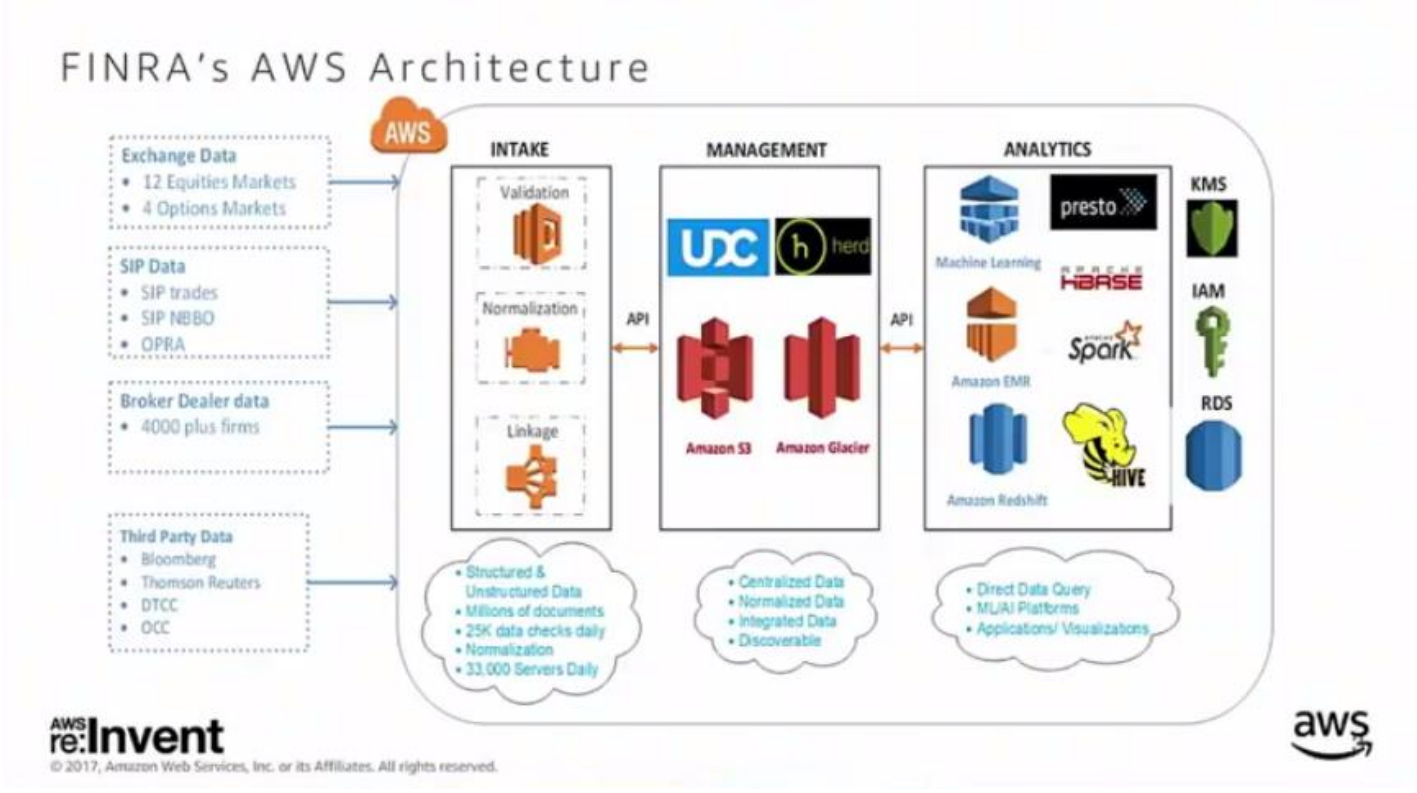
- Separate storage and compute
- Register and track all data in our data catalog
- Keep all versions of each data set
- Protect the data—encrypt at rest and in transit
- Partition data for extra performance
- Backup to another region for business continuity
- Optimize storage and processing costs

FINRA

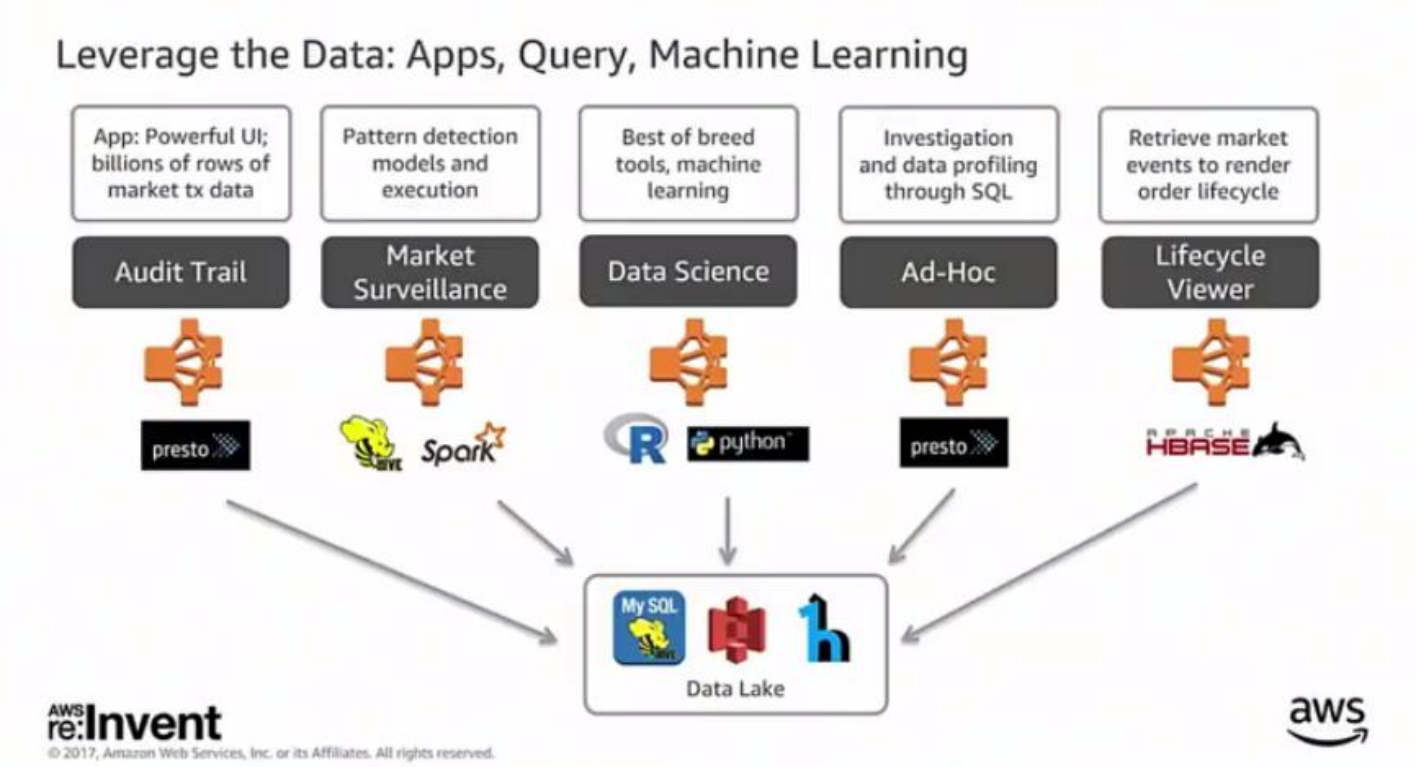
aws



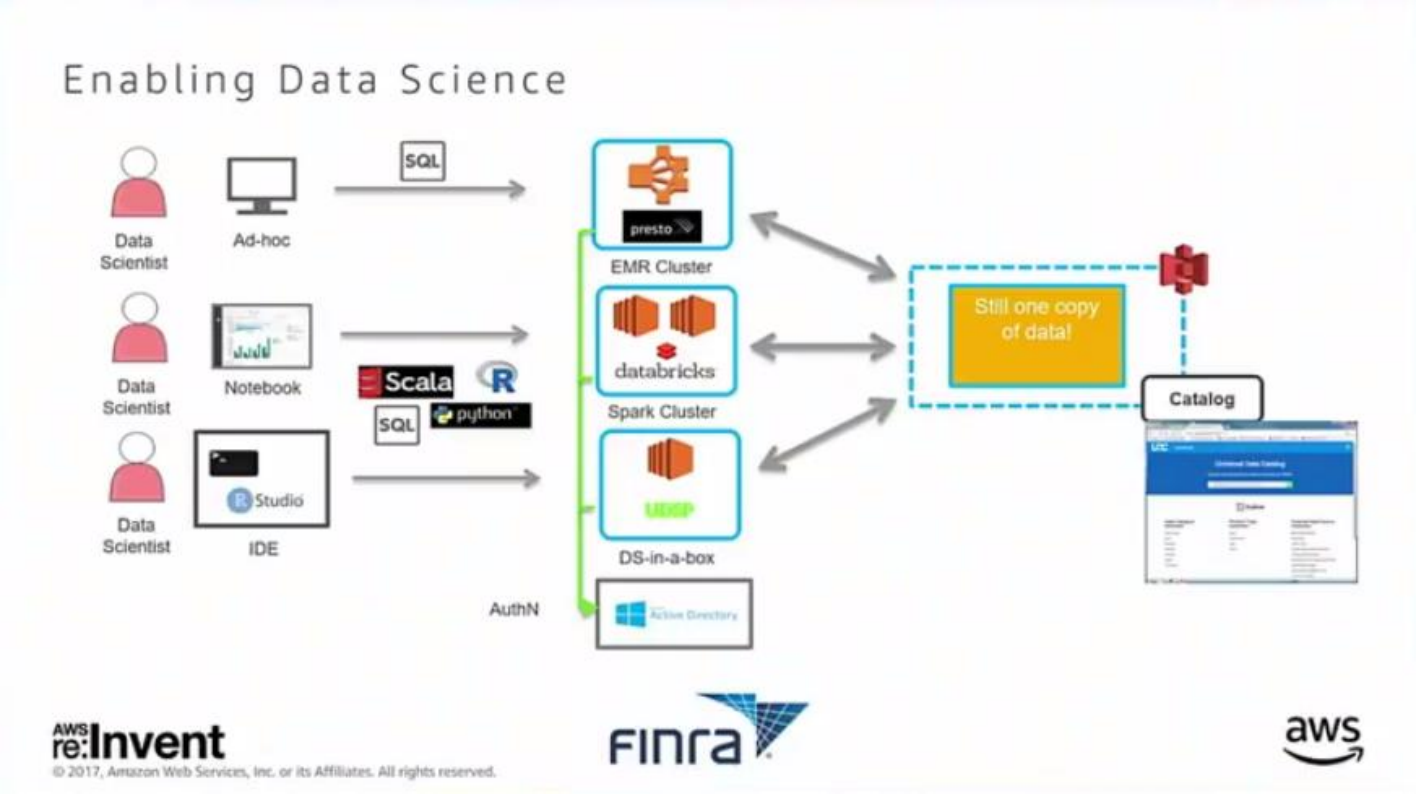
This is the thing that sits between the storage and processing systems we have. The Unified Catalog is mostly about metadata.

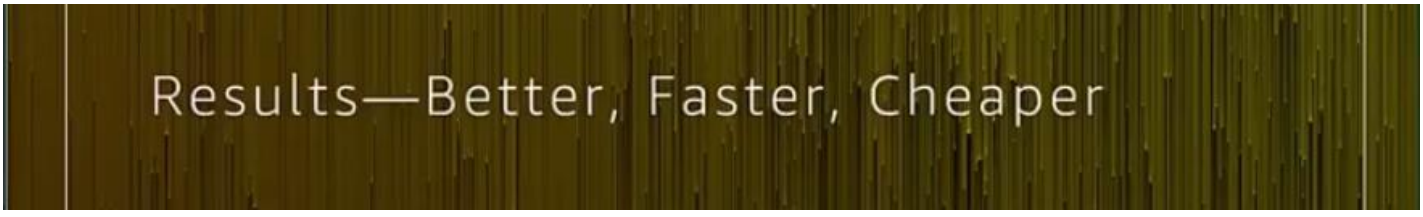


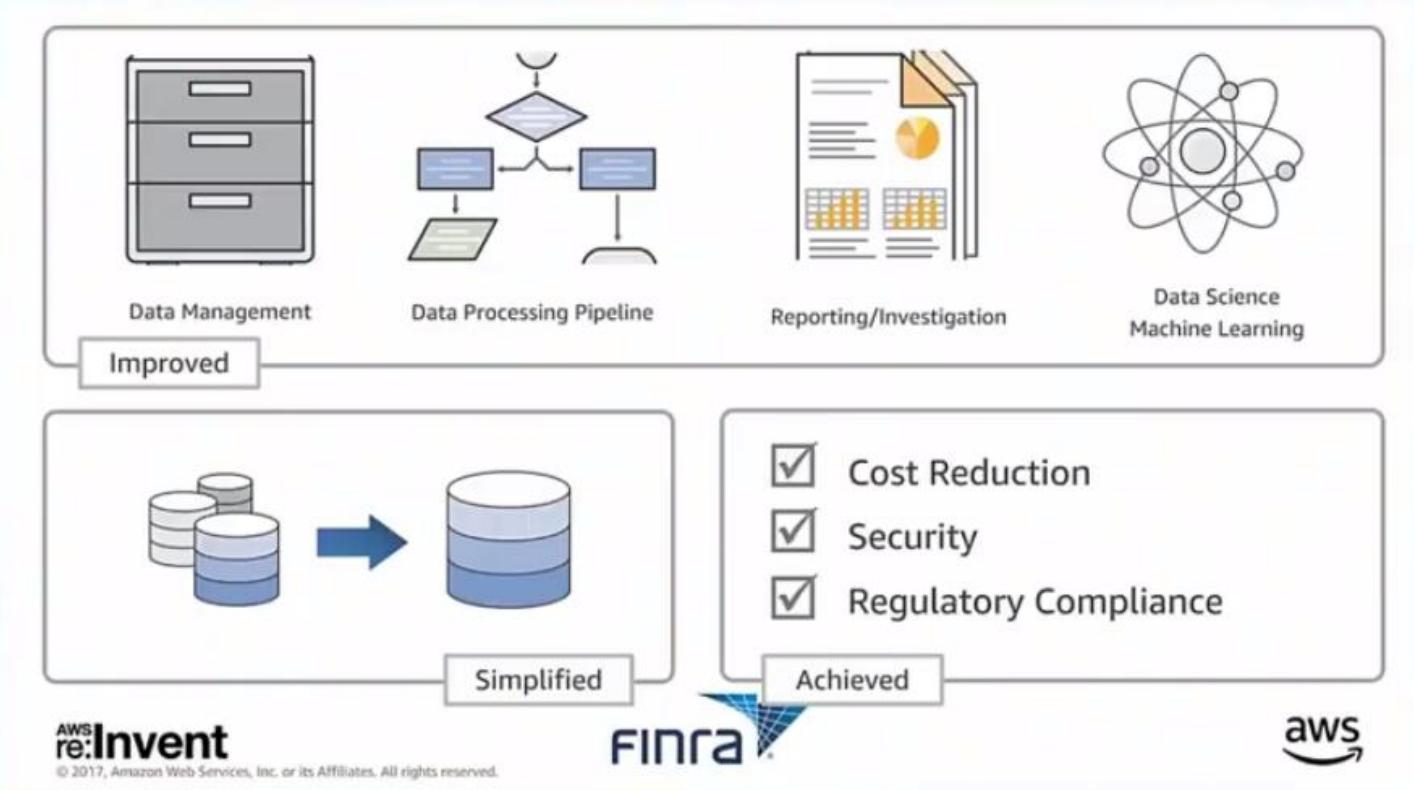
When data comes in to our data ingestion platform, the first thing we do is to register the data source in our data management platform and it is copied to our S3 data lake. A notification process then run validation, normalize the data, and transform it into needed form before storing it again in S3, registered in our Herd catalog, and made available to internal and outside users.

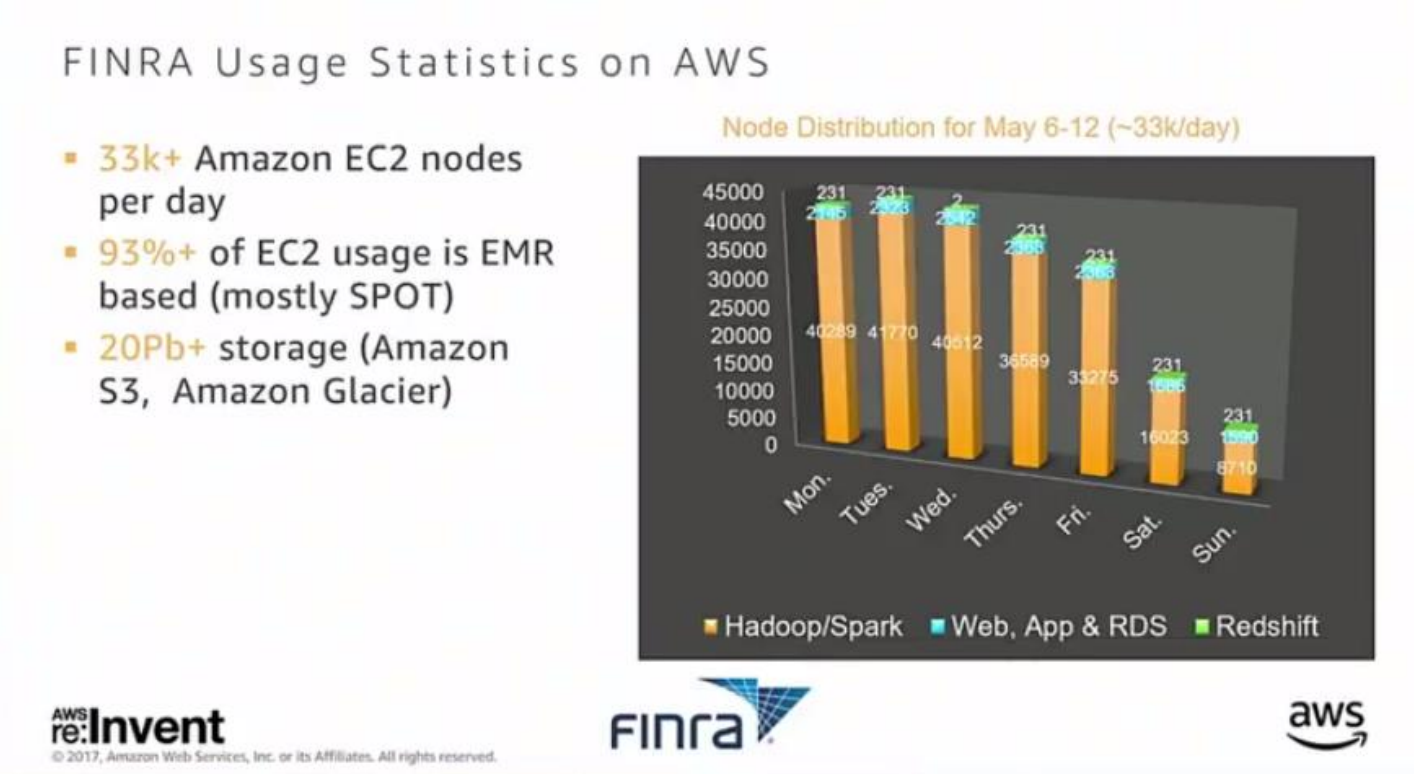


All our data is in a single S3 data lake as one source of truth.



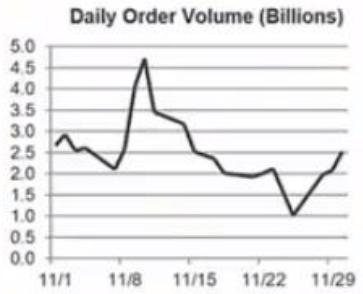


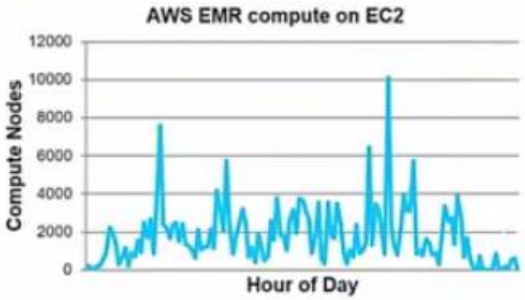




We don’t have permanent EC2 nodes or clusters, we simply bring them up when needed and shut them down when the job is done.

Achieve Dynamic processing





20k – 25k EC2 nodes per day
Over 50k nodes on peak day

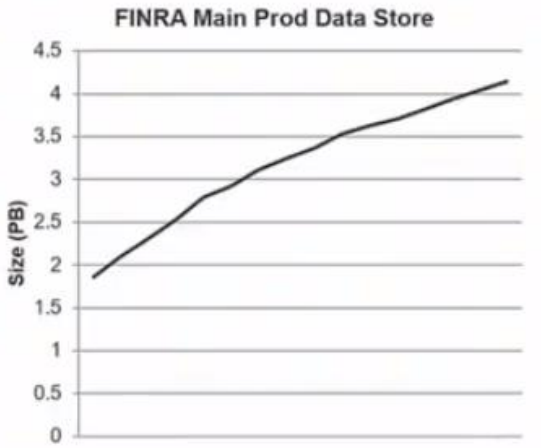
93% of EC2 is on EMR
Avg EC2 node: 3 cores

Avg EC2 uptime: 3 hours
96% of EC2 nodes live < 24 hrs

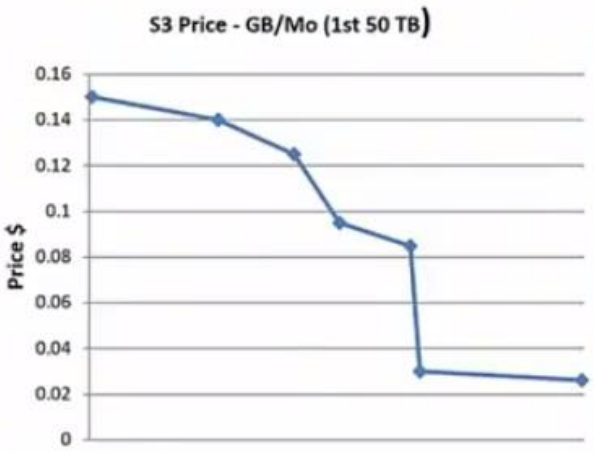
FINRA

aws

Data growth while saving $



Incrementally grow data store by petabytes—no operations work



Periodic price reductions—no operations work

FINRA

aws

Achieving Interactive Query Speed

Query	Table size (rows)	Output size (rows)	ORC	TXT/BZ2
select count(*) from TABLE_1 where trade_date = cast('2016-08-09' as date)	2469171608	1	4s	1m56s
select col1, count(*) from TABLE_1 where col2 = cast('2016-08-09' as date) group by col1 order by col1	2469171608	12	3s	1m51s
select col1, count(*) from TABLE_1 where col2 = cast('2016-08-09' as date) group by col1 order by col1	2469171608	8364	5s	2m5s
select * from TABLE_1 where col2 = cast('2016-08-10' as date) and col3='I' and col4='CR' and col5 between 100000.0 and 103000.0	2469171608	760	10s	2m3s

Key points:

Use ORC (Or Parquet) for performant query

Test Config:
Presto 0.167.0.6t (Teradata) On EMR
Data on S3 (external tables)
Cluster size: 60 worker node x r4.4xlarge

FINRA

aws

We run Presto for our Hadoop SQL jobs along with ORC and Parquet compressed data formats.

Benefits We've Seen

Analytics

- ☑ Analysts can now interactively analyze 1000x more market events (billons vs million rows)
- ☑ Querying order route detail went from 10s of minutes to seconds
- ☑ Quicker turnaround to provide data for
- ☑ Machine Learning model development is easier

Agility

- ☑ Easily reprocess data … used to take weeks to find capacity now can be done in day/days
- ☑ Cloud makes it very easy to share (even large) data sets with third parties in Cloud
- ☑ Can perform model (pattern) reruns in days not weeks

Resiliency

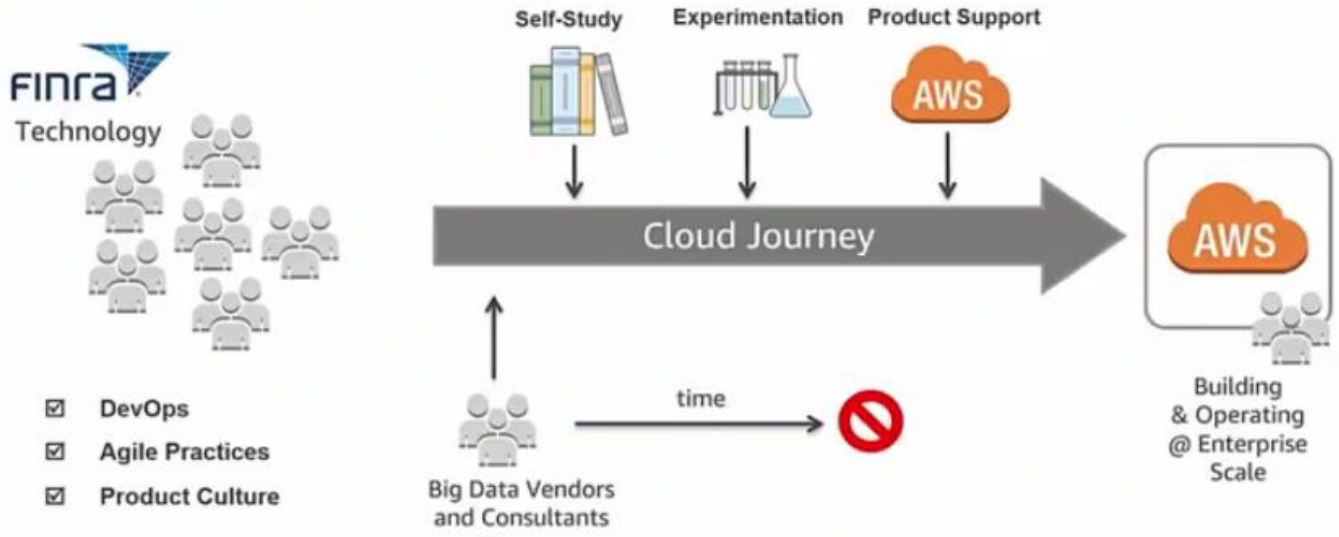
- ☑ Market volume changes no longer disruptive events
- ☑ Improved system uptime vs in-house

At TCO 30% less expensive than with our data center

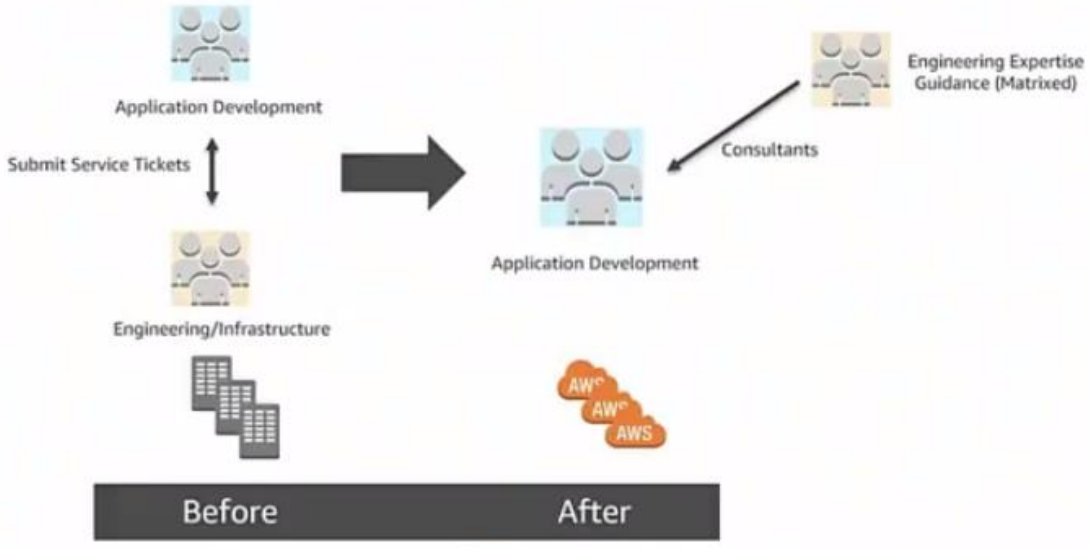
FINRA

aws

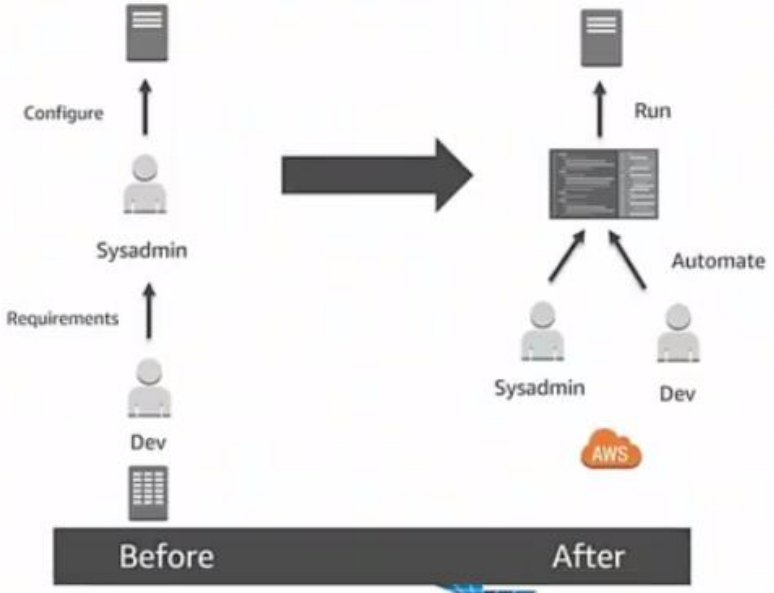
Gaining skills for Cloud—FINRA case



Culture change—integrate sys eng & app dev teams



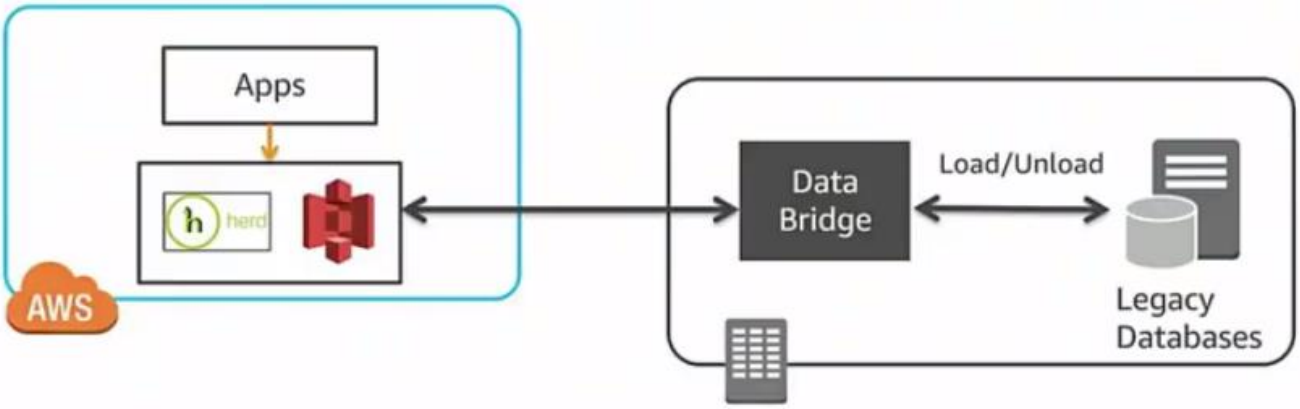
Culture changes—infrastructure as code



FINRA

aws

Continuity of service across complex data ecosystem



Multiple producers and consumers, could not migrate at same time

Leveraged integration with Data Catalog—in use for 2½-year migration

FINRA

aws

Lessons Learned & What Next

Lessons Learned

Disrupt legacy practices	Embrace the chance, you must accomplish this to be successful.
Automate, automate, automate	Infrastructure as code makes this not only possible but critical for efficiency.
Transform people & processes	Cloud is new way of thinking, leverage this opportunity to transform your existing team and process.
Go all-in on agile	This includes having QC testers, system engineers & applications engineers (automation specialists) matrixed into the development teams.
Continuous improvement	The cloud provides with you with never ending opportunities to further improve your compute environment (as it relates to costs, performance, resiliency and security). You are never done building in the cloud.
Capex to Opex financial model	Moving to the cloud means moving from a Capex to OpEx financial model. Also, in the new world order, make your finances fungible between labor and non-labor.

What's Next for FINRA

Spot Fleet for Amazon EMR	Better manage our Amazon EMR cost
Amazon Athena	Serverless query
Broader adoption of AWS Lambda	Already in use to perform ½ trillion validations per day Expanding usage to APIs and microservices
Machine learning and AI	The next frontier for regulatory analytics
AWS Step Functions	Leverage to orchestrate business processing for applications
Multi-region support	For some business critical applications, we are looking at enhancing to be multi-region capable
AWS Glue Integration	Integration of AWS Glue with Herd data catalog and ETL workloads

Continue to monitor what's coming from AWS, new technologies create new opportunities


To lean more: technology.finra.org/
AWS re:Invent
Thank you!
Tim.Griesbach@finra.org
Ranganathan.Rajagopal@finra.org
AWS re:Invent
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
aws