FINRA's analytics platform unlocks the value in capital markets data by accelerating trade analytics and providing a foundation for ***machine learning*** at scale. The platform enables FINRA's analysts to perform discovery on petabytes of trade data to identify instances of potential fraud, market manipulation, and insider trading. By centralizing all data in ***S3***, FINRA's architecture offers improved agility, scalability, and cost effectiveness. Analytics services such as Amazon EMR and Amazon Redshift have freed FINRA's data scientists from the constraints of desktop tools, allowing them to apply machine learning techniques to develop and test new surveillance patterns. All of this is done while meeting FINRA's security and compliance responsibilities as a financial regulator. At the end of this session, you'll have an understanding of how to apply FINRA's architecture to trade analytics and other financial services use cases, including meeting regulatory requirements such as the Consolidated Audit Trail (CAT) reporting.

Data lake

Central Storage
Secure, cost-effective
storage in
Amazon S3

aws

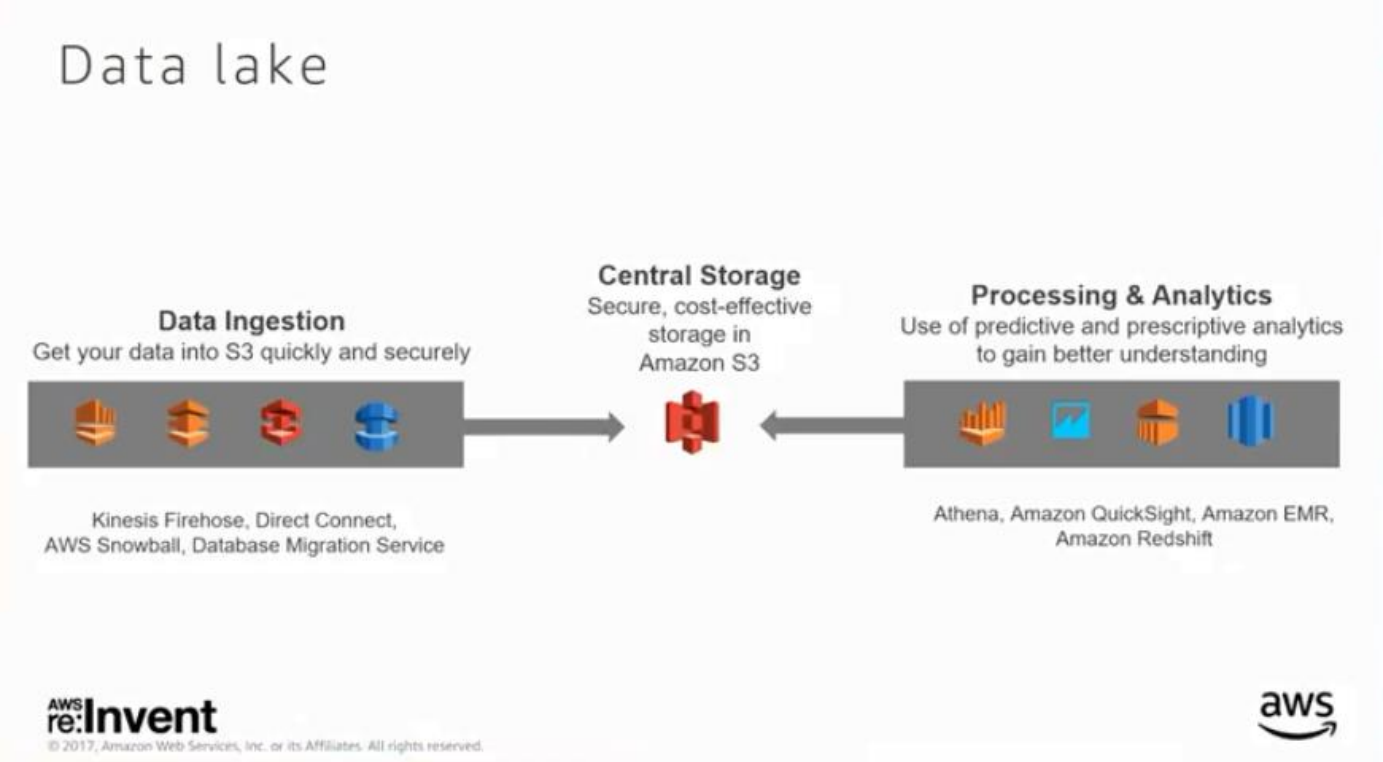
Data lake

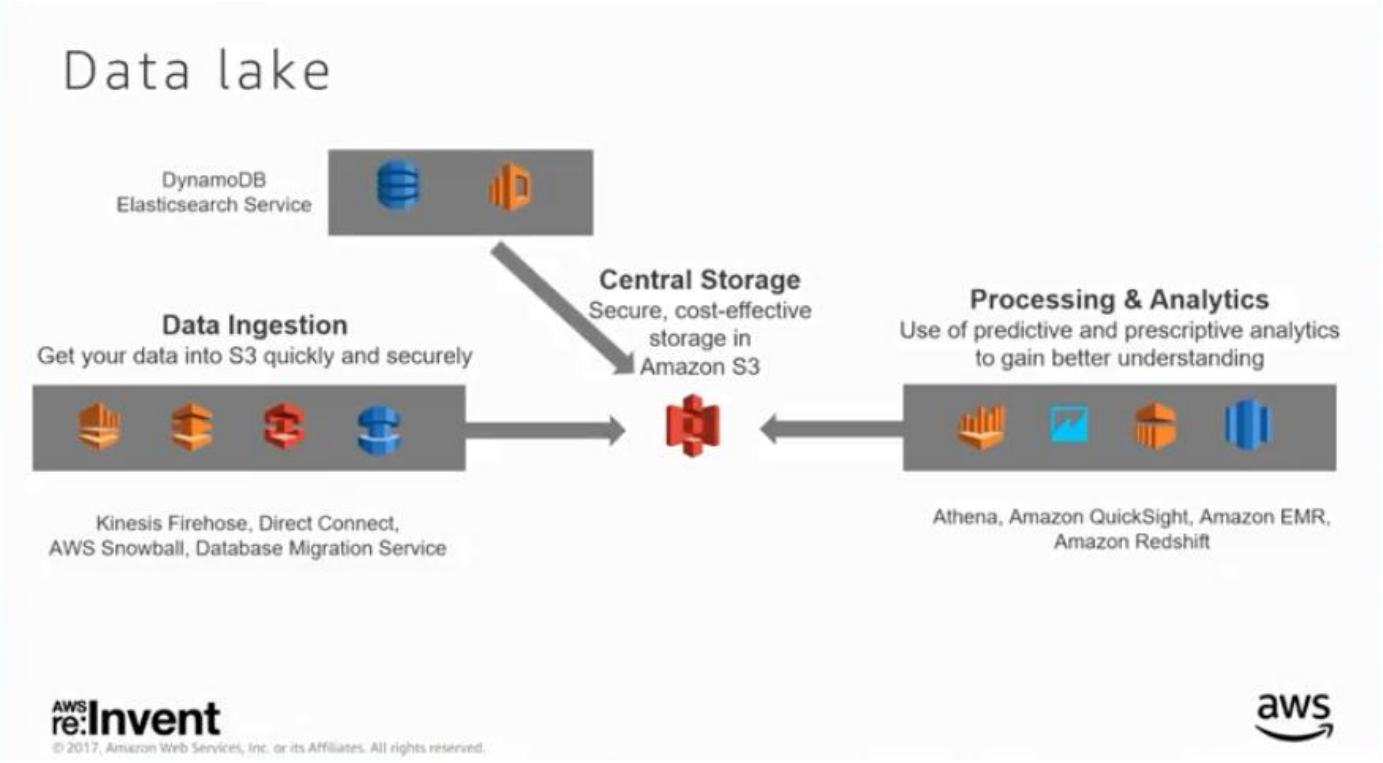
Data Ingestion
Get your data into S3 quickly and securely

Kinesis Firehose, Direct Connect,
AWS Snowball, Database Migration Service

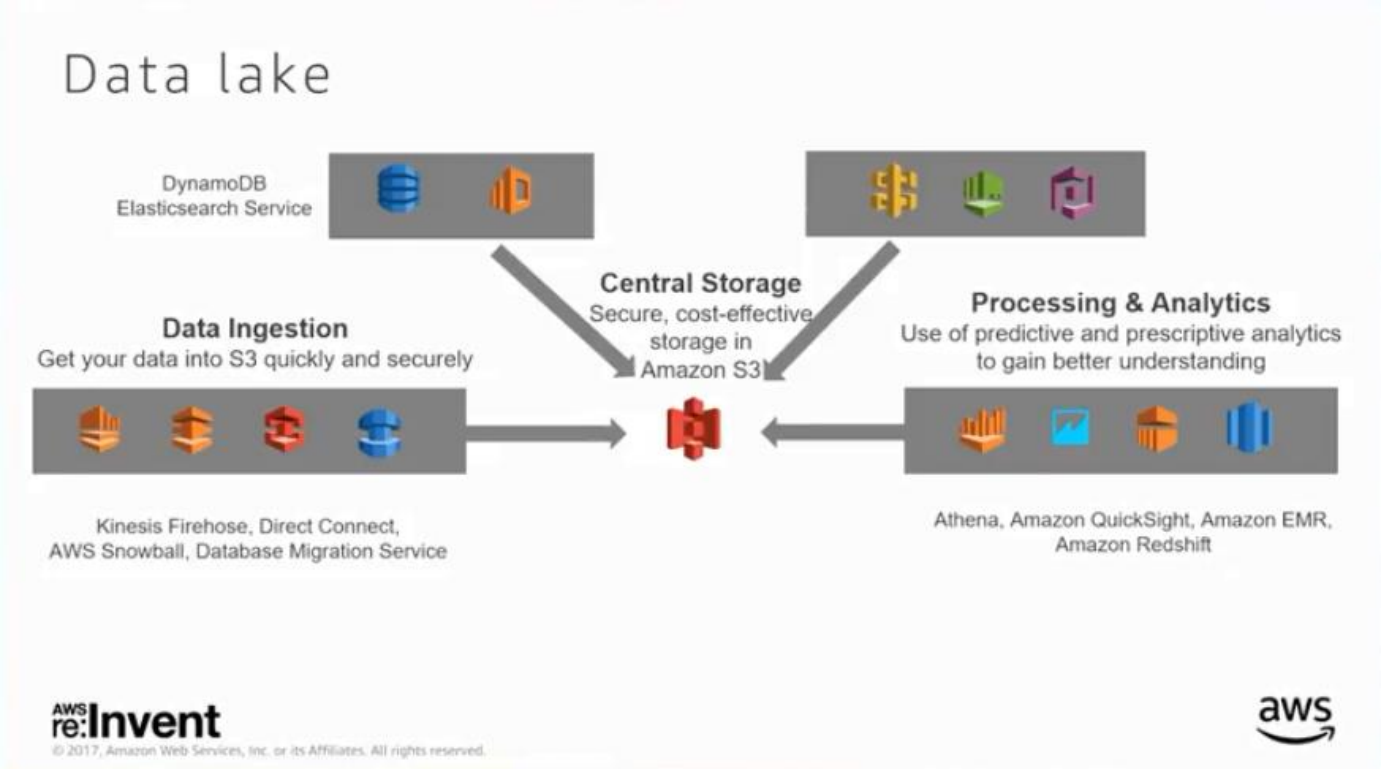
Central Storage
Secure, cost-effective
storage in
Amazon S3

aws

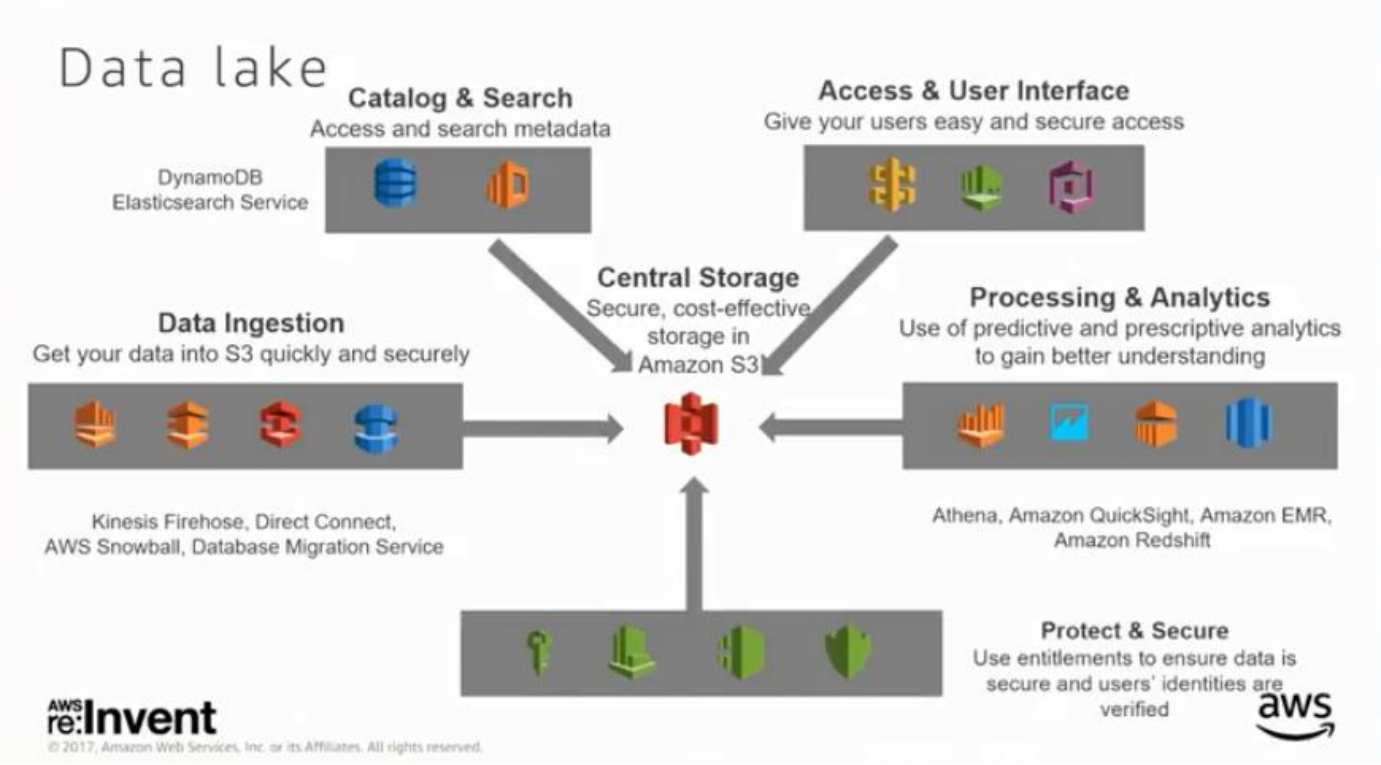




You can then use DynamoDB to catalogue your data and store metadata about the data, this allows you to be able to quickly know your data location and access the data for faster queries



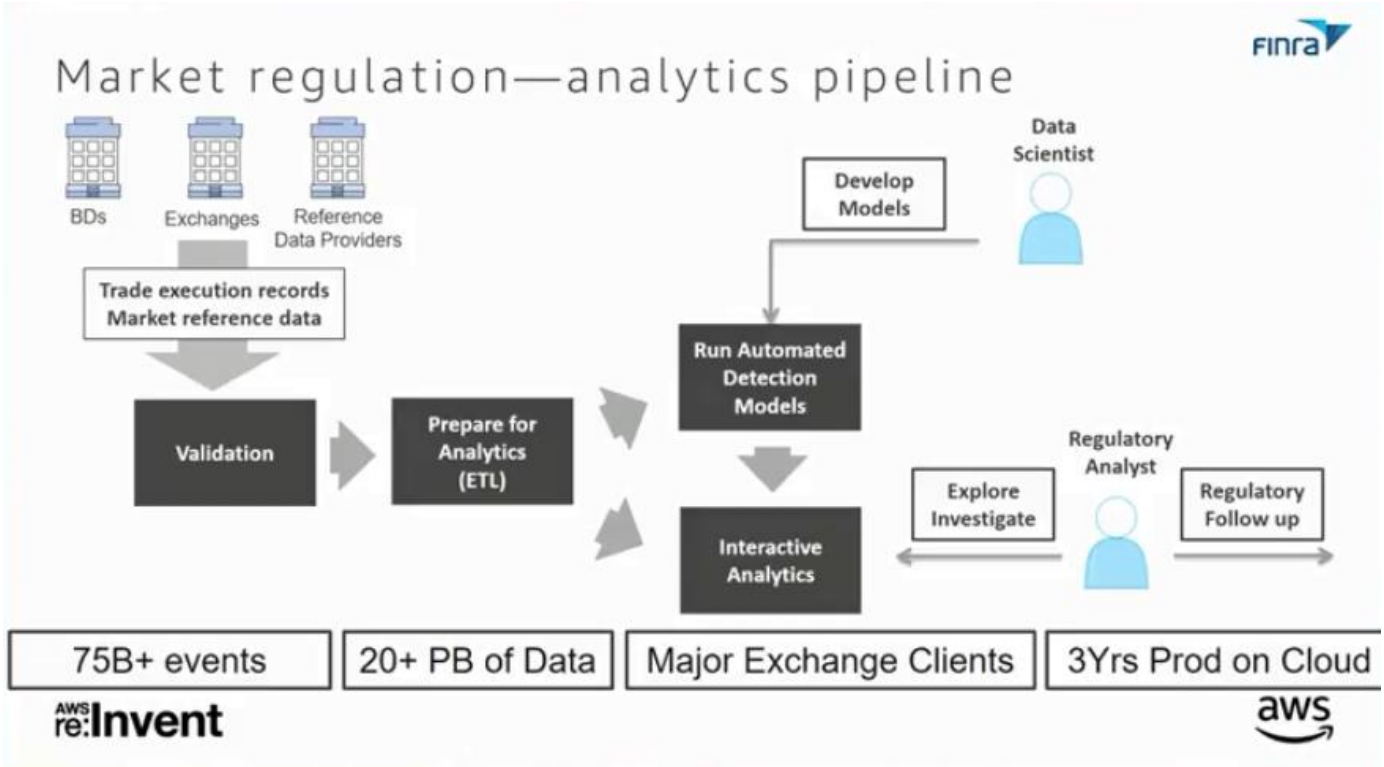
You can then use services like Cognito and API Gateway for providing access to the data in S3.



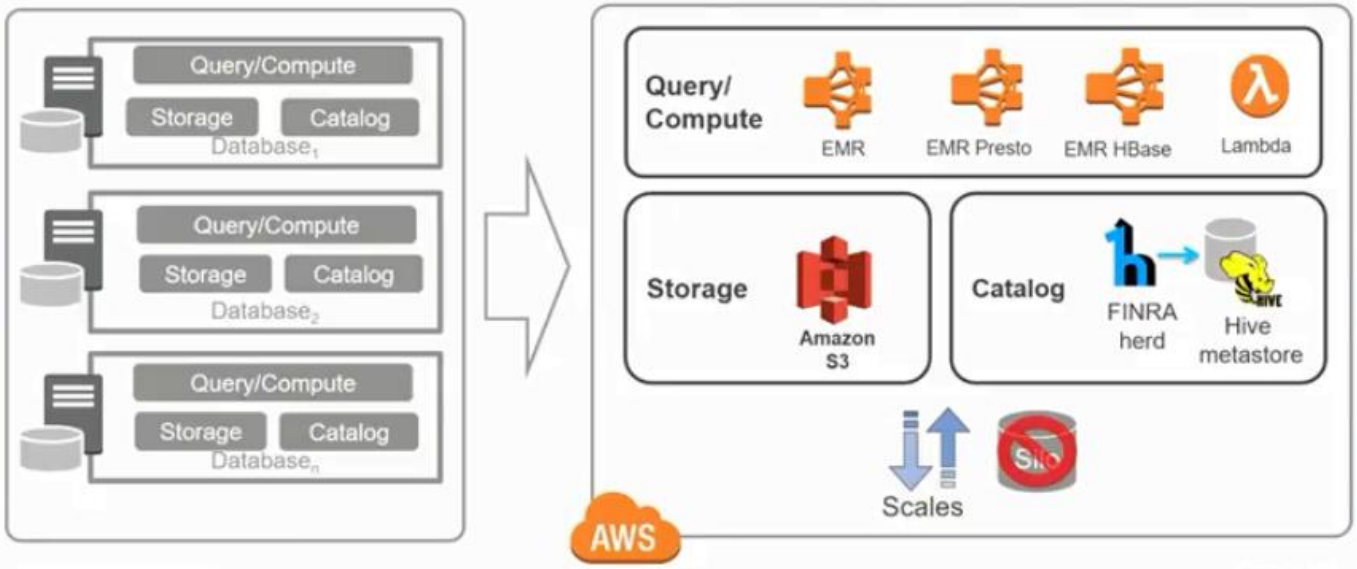
You can then secure and protect your data with services like KMS for encryption, using your own encryption modules, and IAM for identity and access management.

FINRA's Data Lake

Surveilling markets with FINRA's multi-petabyte enterprise-grade data lake



Cloud journey—data puddles to data lake



Herd catalog—for centralized data management

Unified catalog

- Schemas
- Versions
- Encryption type
- Storage policies

Lineage and Usage

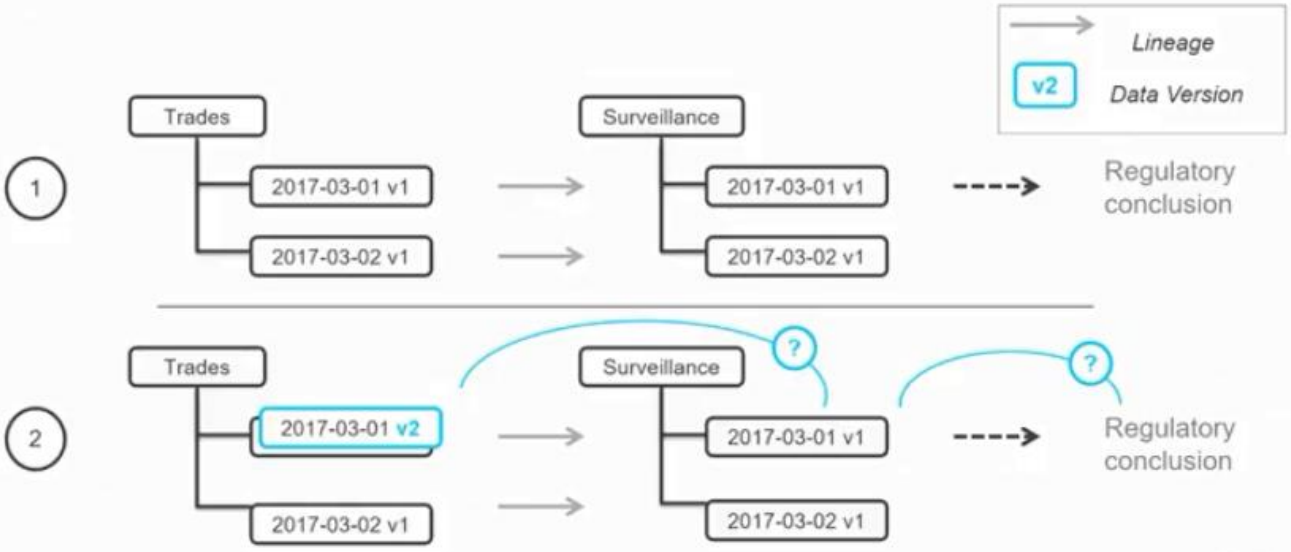
- Track publishers and consumers
- Easily identify jobs and derived data sets

Shared Metastore

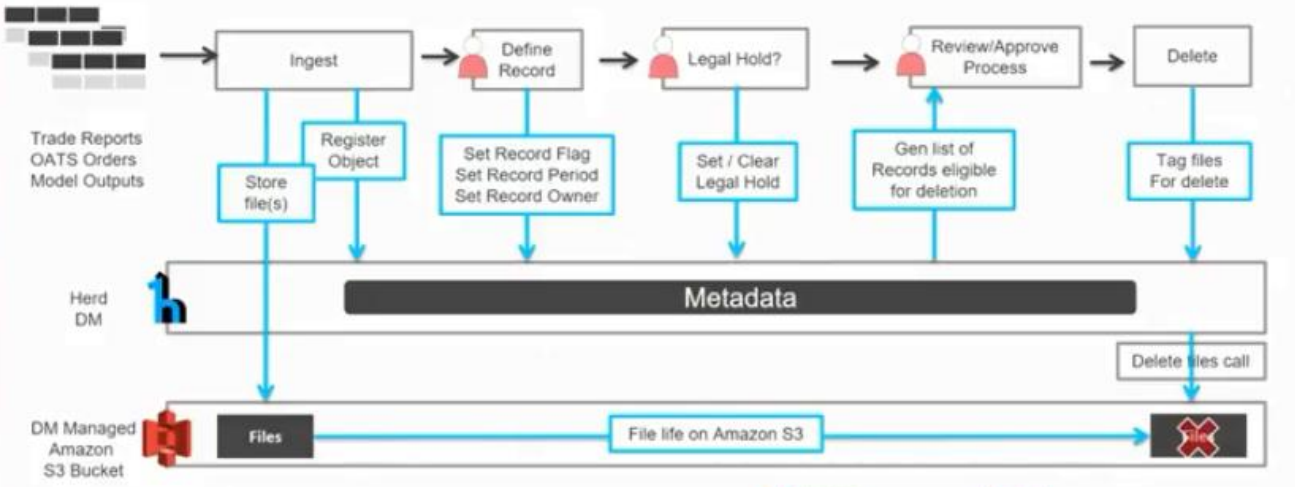
- Common definition of tables and partitions
- Use with Spark, Presto, Hive, etc.
- Faster instantiation of clusters

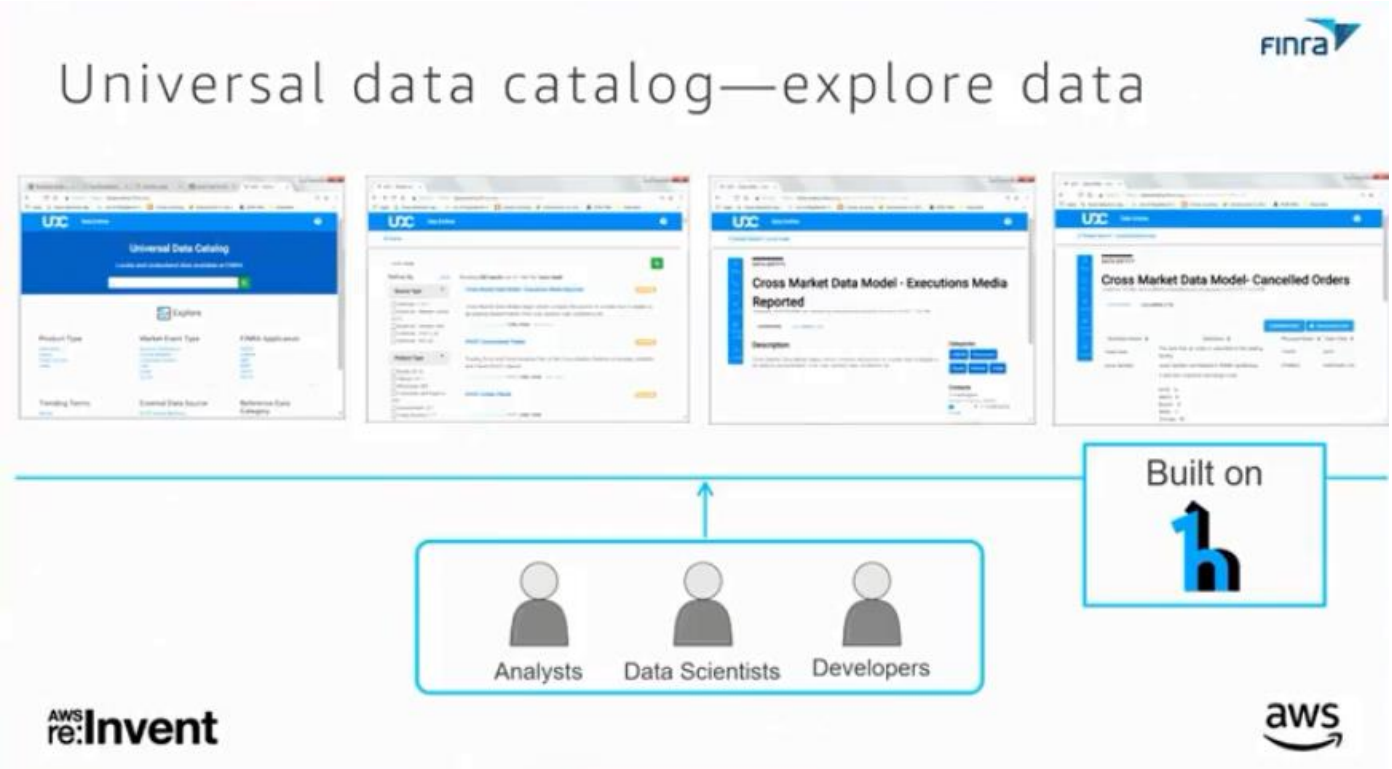
http://finraos.github.io/herd

Example—lineage and data versioning

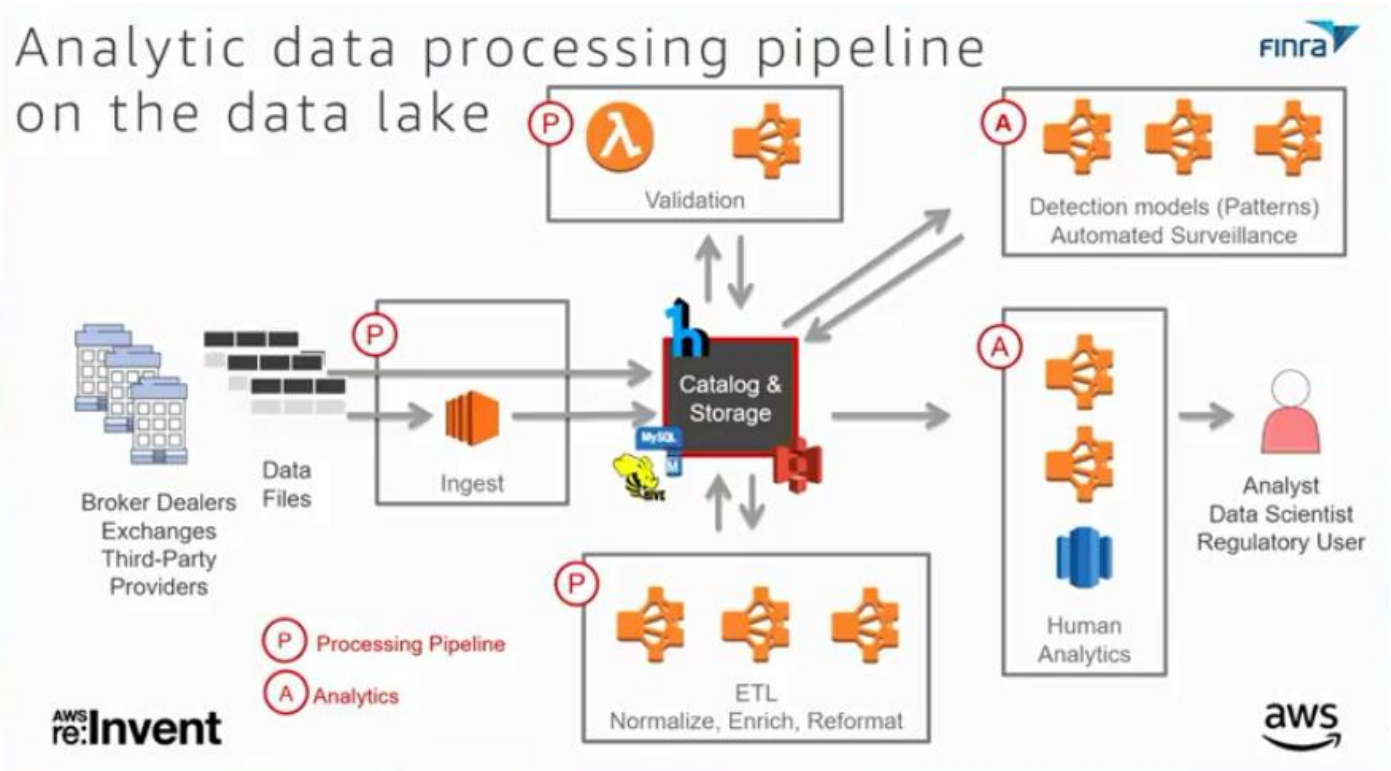


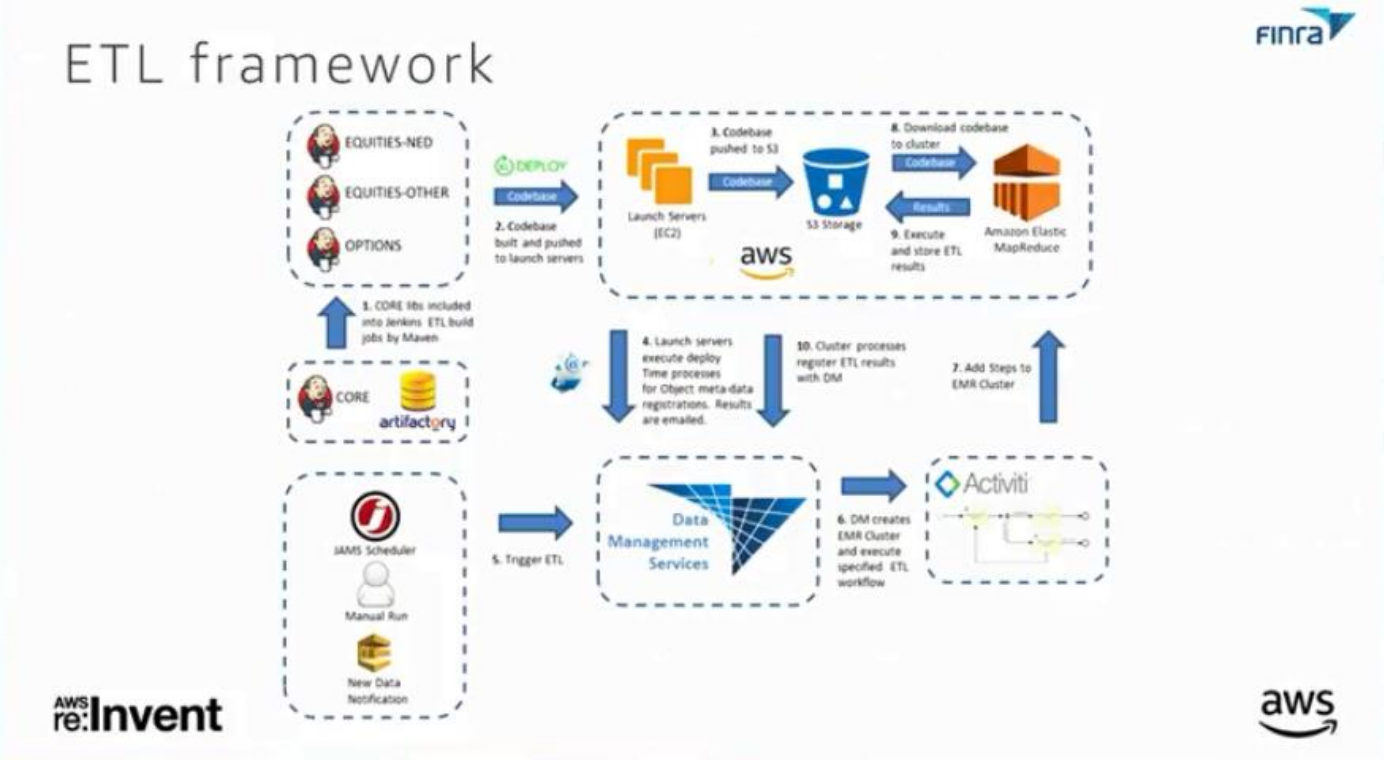
Herd—foundation for records management



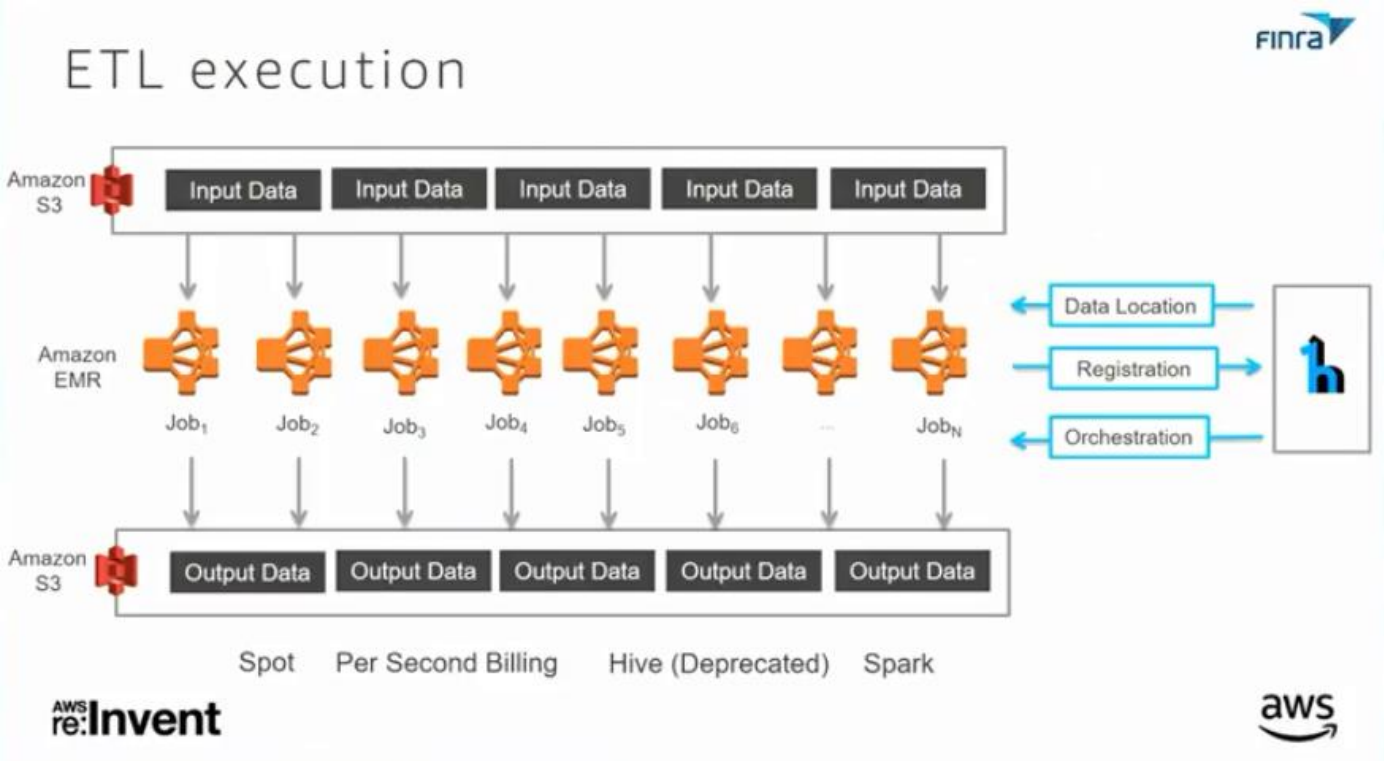


We have also created a UI app sitting on top of the data catalog that allows our analysts and data scientists to discover and use the data sets available in S3 for their work. They can search through both the business and technical metadata to do business related searches for the data they need.

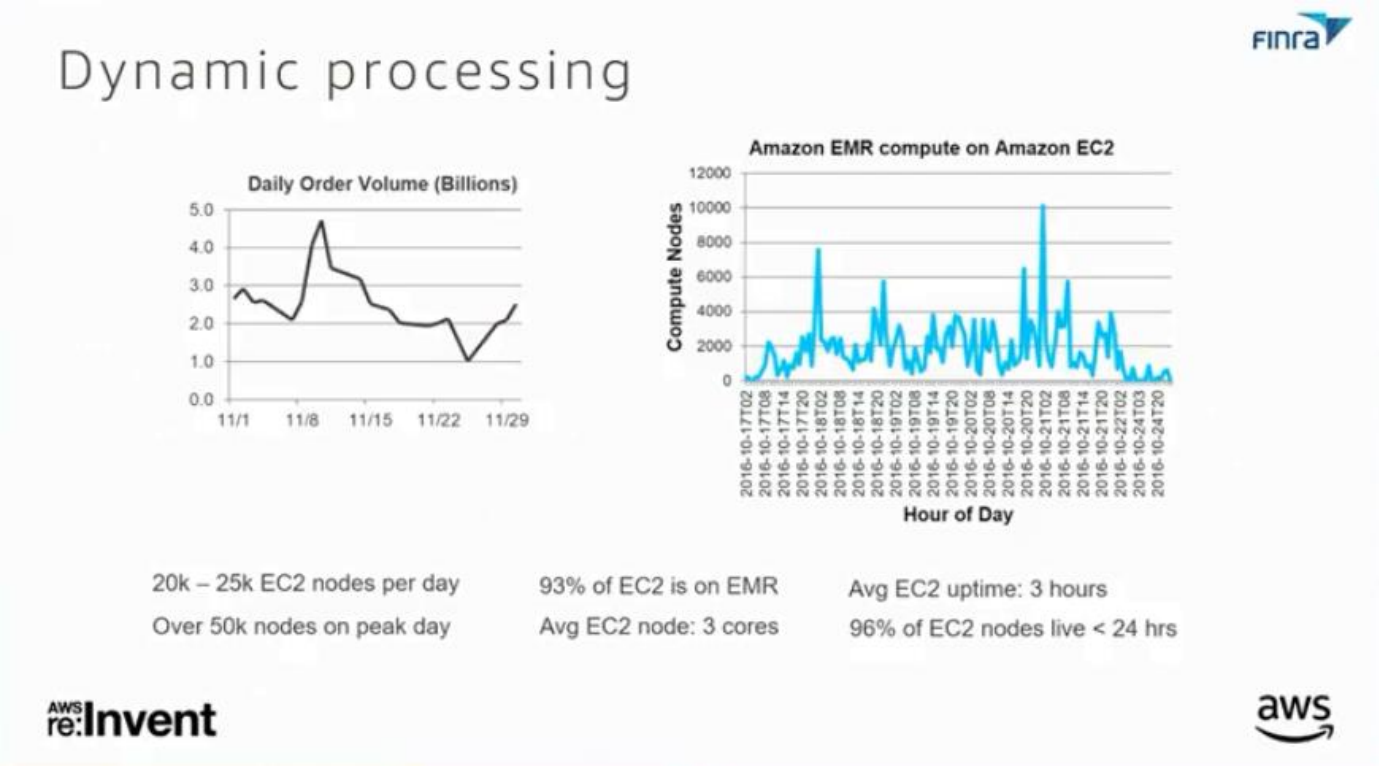




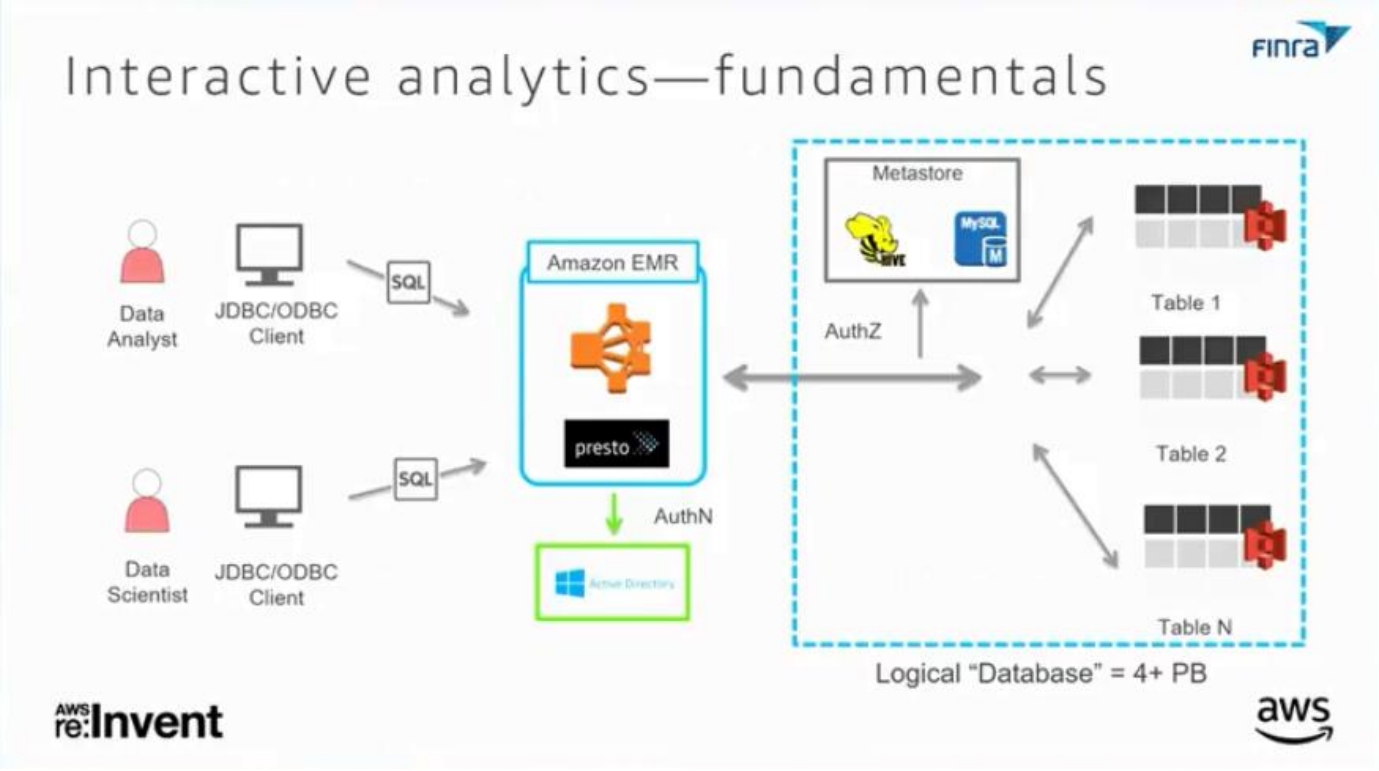
The in-house ETL Framework is a tool that lets our developers write ETL jobs using SQL and then deploy that as a pipeline that gets triggered based on a schedule or based on an event coming our of our data catalog like new data getting stored in S3.



We have a separate EMR cluster for each process that we run. So, when we trigger an ETL job, we spin up an EMR cluster, as part of the bootstrap process for that cluster it will go out and interrogate the catalog to find the information it needs to do its processing, go out to S3 or simply query against data in S3 and grab that data, run its workflow, store results back in S3, and terminate the workflow and cluster. The majority of the jobs we run are batch workloads that run on spot. Majority of our jobs are written in Spark.



This is a very dynamic environment with EMR several clusters being spun up and down as jobs are needed to be run on spot instances.



For interactive analytics, we run Presto and Spark on top of EMR clusters. They let you define data files in S3 as external tables and that lets you query against the data directly while keeping it in S3 instead of loading it up into your instance memory. We do this by populating an OSS catalog called the Hive Metastore for defining all the available data objects and all the tools that Presto and Spark can use. This now allows our analysts and data scientists to use think clients having JDBC/ODBC drivers to run exploratory ad-hoc queries on the S3 data securely and using data encryption.

Achieving interactive query

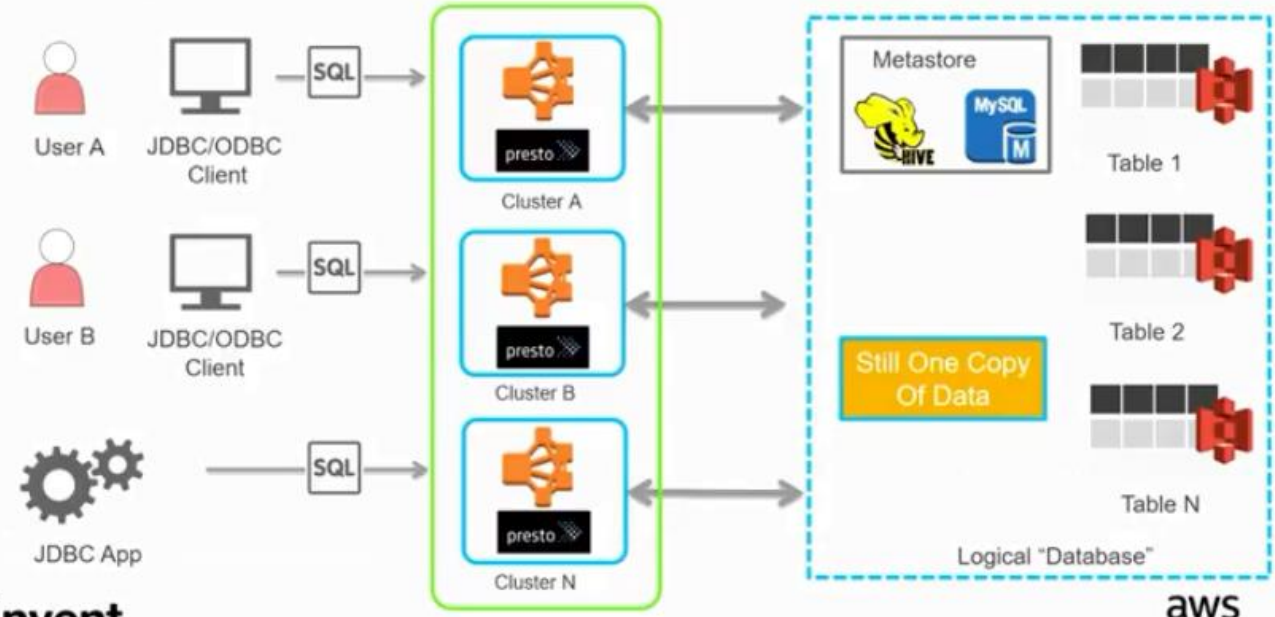
Query	Table size (rows)	Output size (rows)	ORC	TXT/BZ2
select count(*) from TABLE_1 where trade_date = cast('2016-08-09' as date)	2469171608	1	4s	1m56s
select col1, count(*) from TABLE_1 where col2 = cast('2016-08-09' as date) group by col1 order by col1	2469171608	12	3s	1m51s
select col1, count(*) from 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

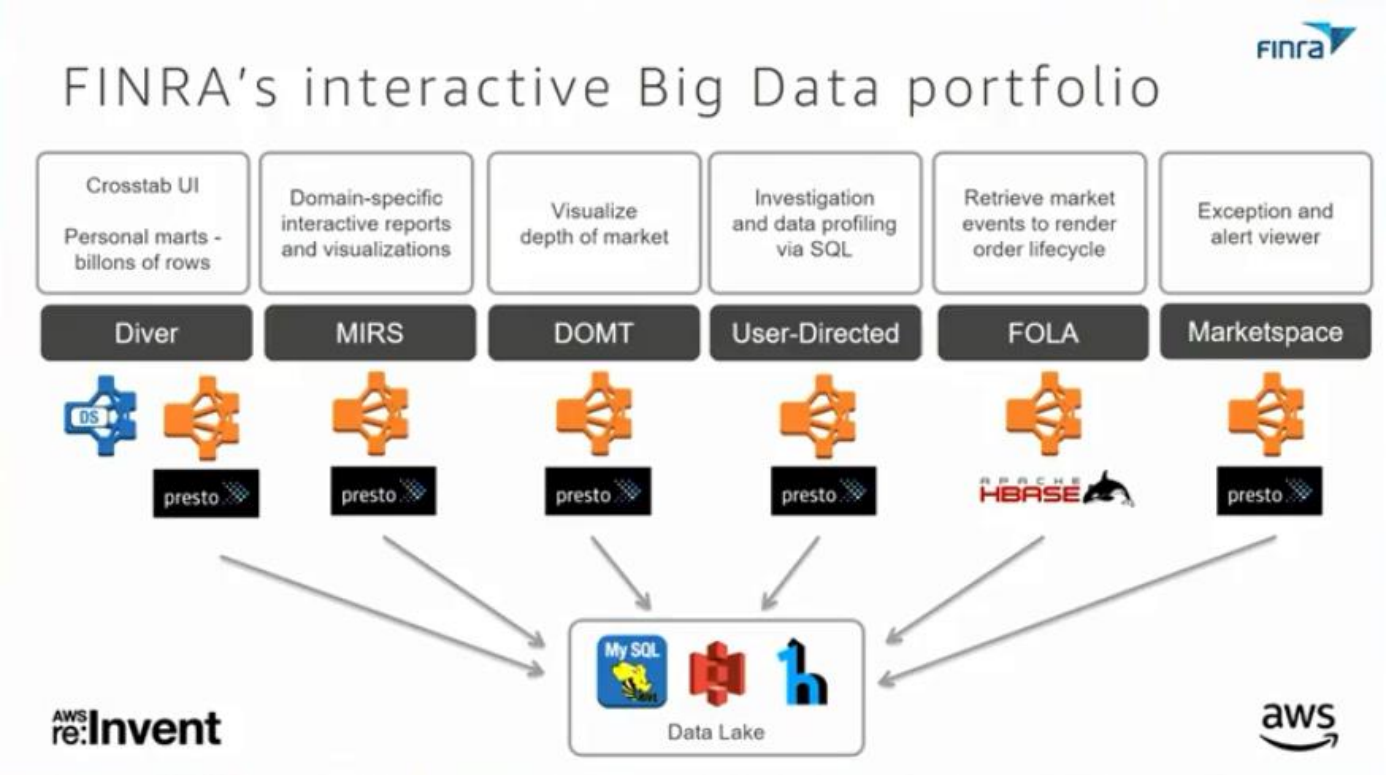
Key points:

Use ORC (Or Parquet) for performant query

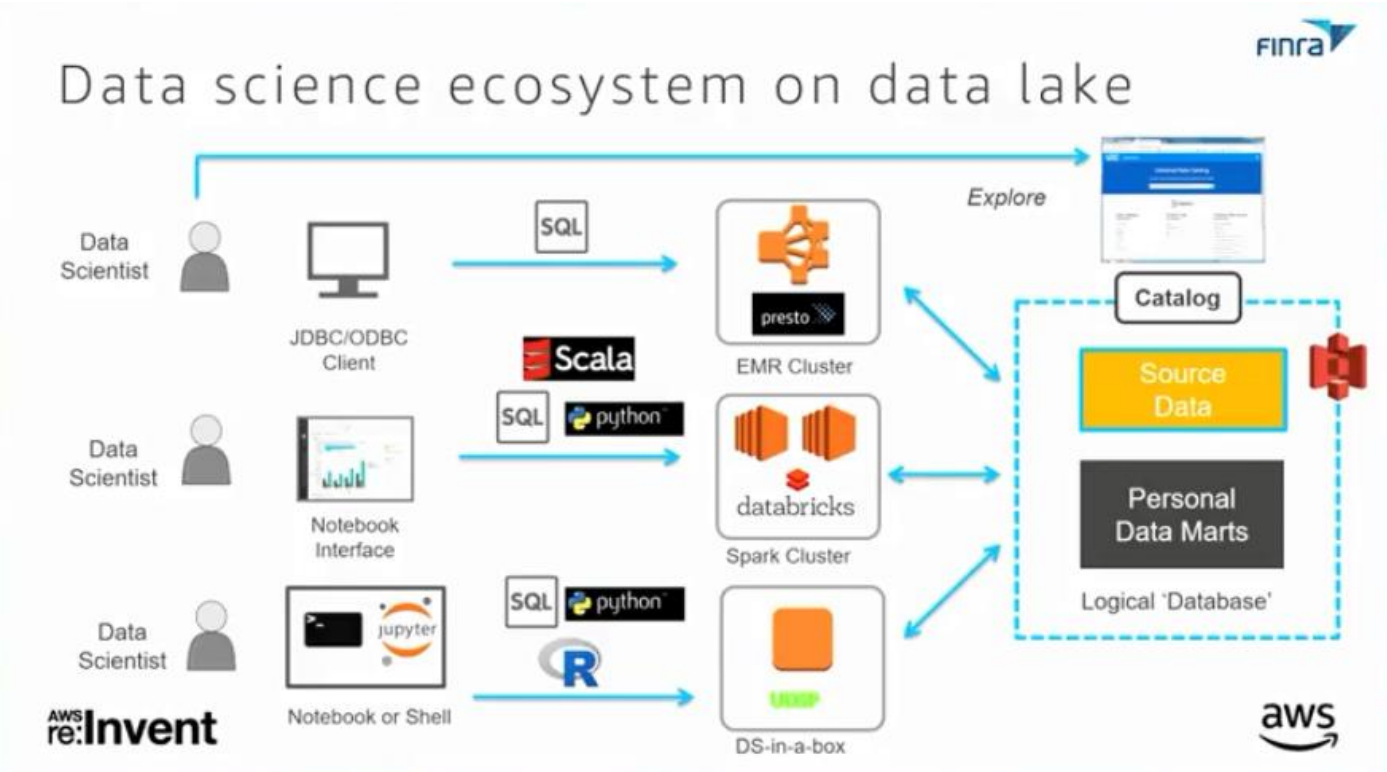
Test Config:
Presto 0.167.0.6t (Teradata) On EMR
Data on S3 (external tables)
Cluster size: 60 worker node x r4.4xlarge

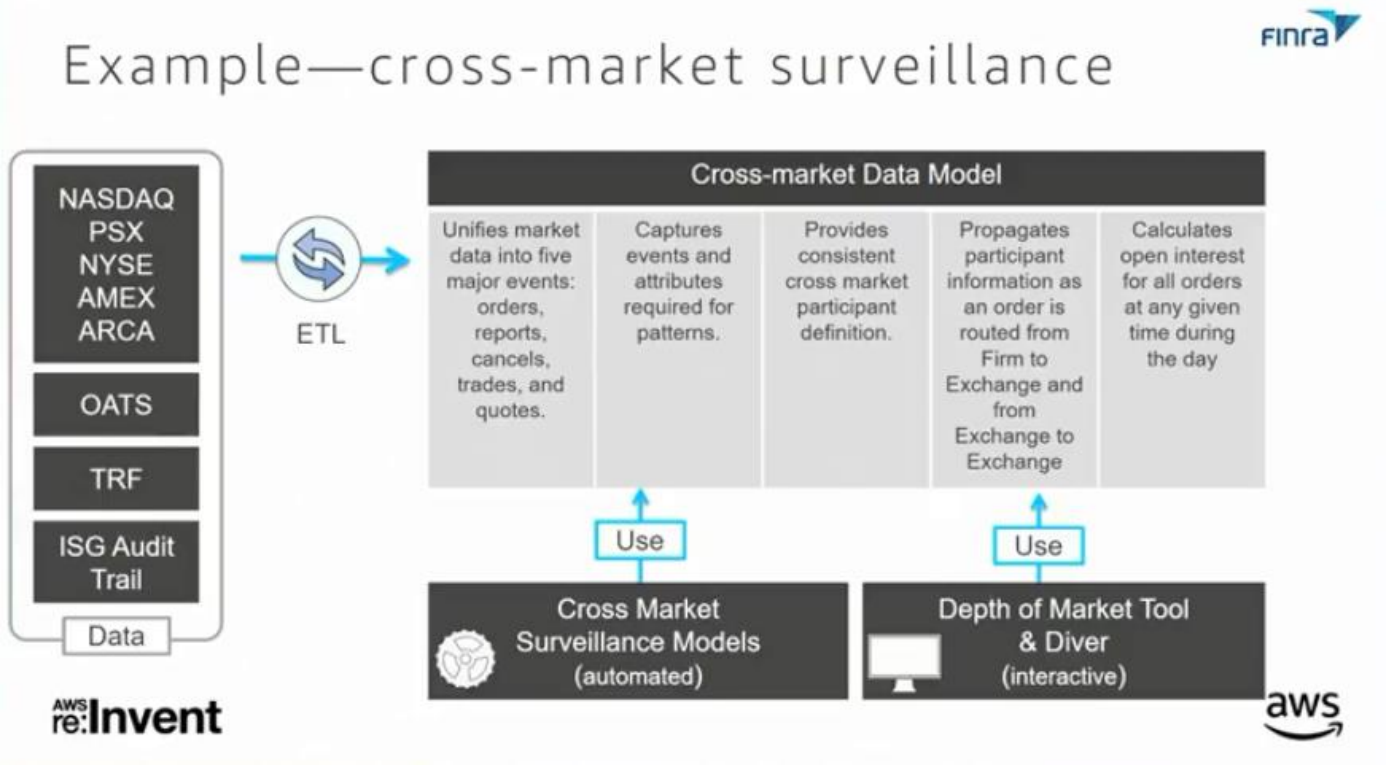
Scaling out interactive query

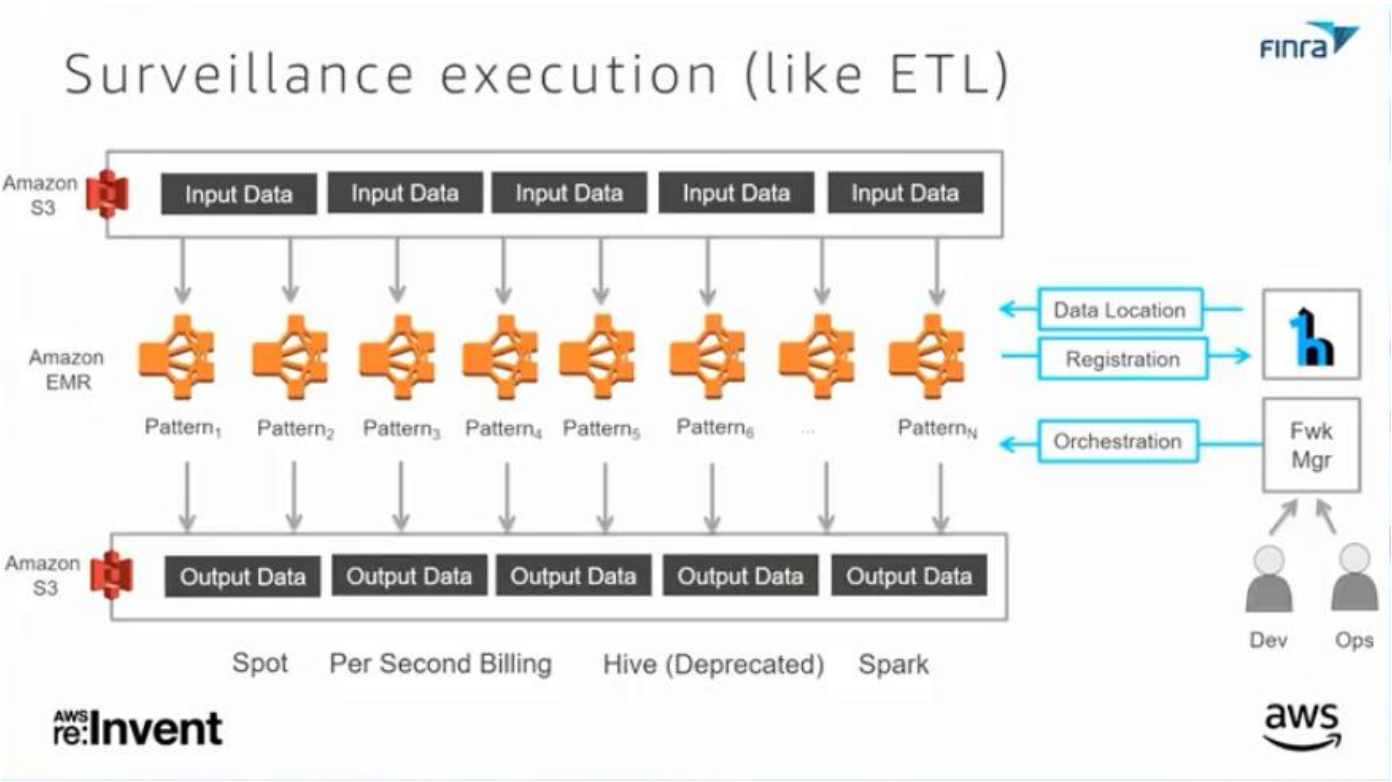




EMR also has the ability to use HBase while keeping the data in S3.





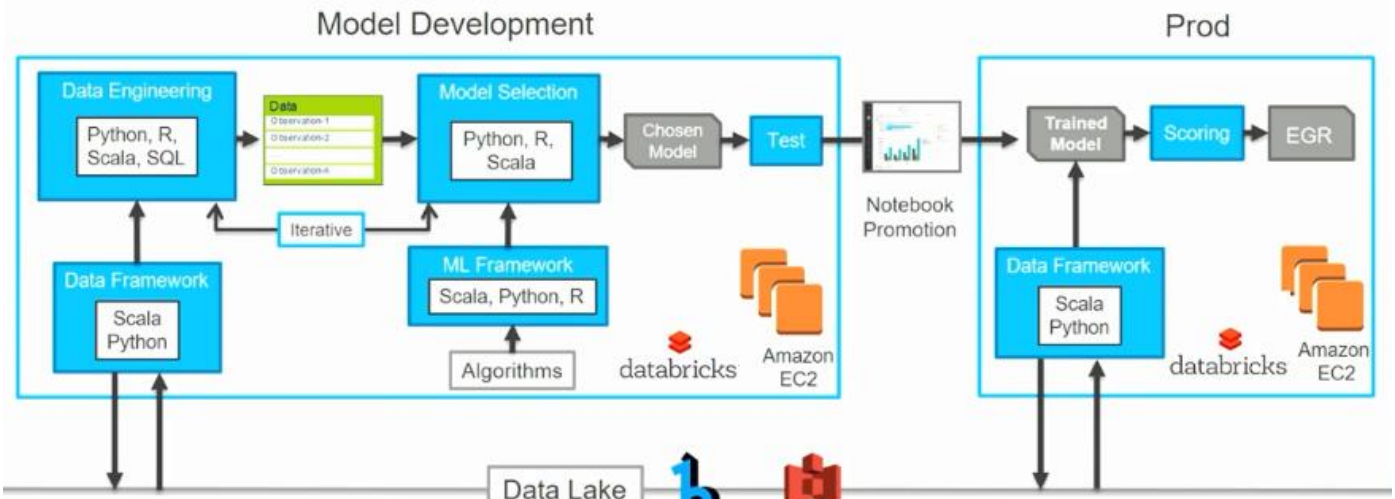


Every detection model run spins up its EMR cluster, runs its processing directly on data in S3, stores its data outputs back into S3 and then shuts the cluster down.

Surveillance evolution

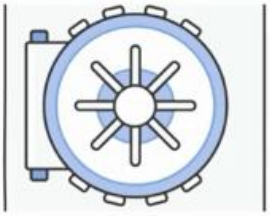
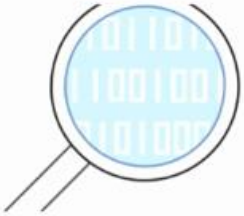
	Before Cloud	Cloud v1	Cloud v2
Execution Engine	Relational DB	Hive, Spark	Spark
Language	SQL	SQL (HiveQL, Spark SQL)	Scala, Python, R, SQL, Java
Production Logic	SQL w/ some scripting	SQL w/ some scripting	ML model (H2O, MLlib)
Data Catalog	N/A	Catalog provides schema/ location	Create dataframes Catalog provides schema/ location
Data Framework	N/A	N/A	Data manipulated as dataframe API for common manipulations

FINRA's dynamic surveillance platform



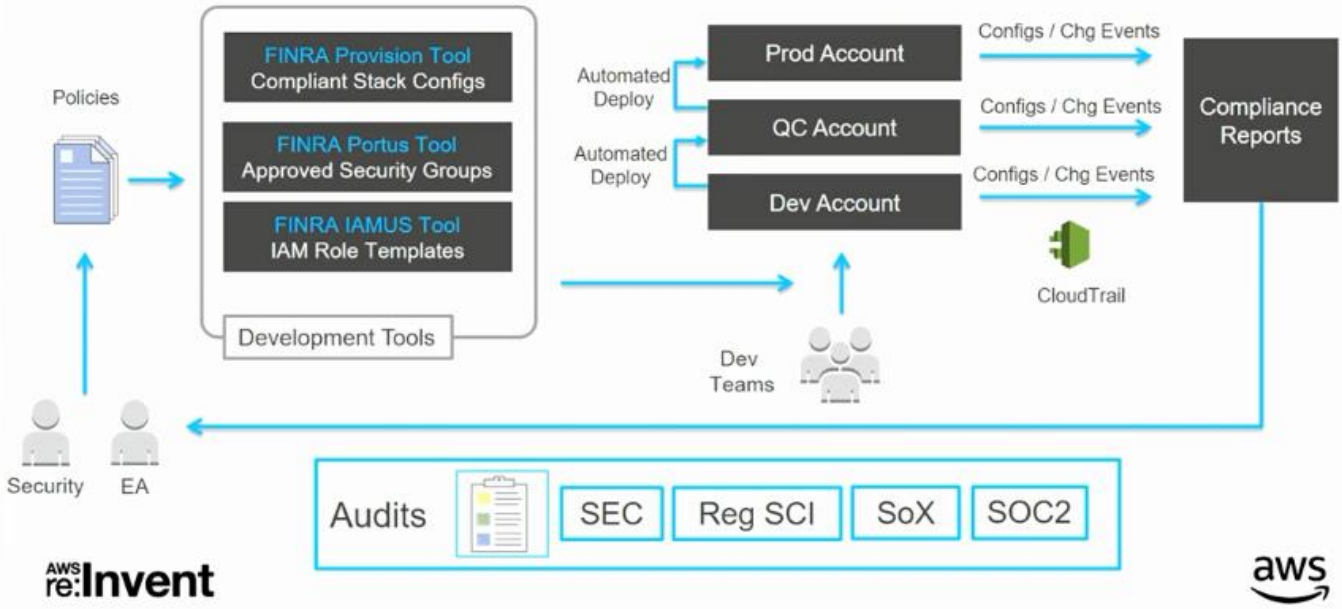
Security

FINRA

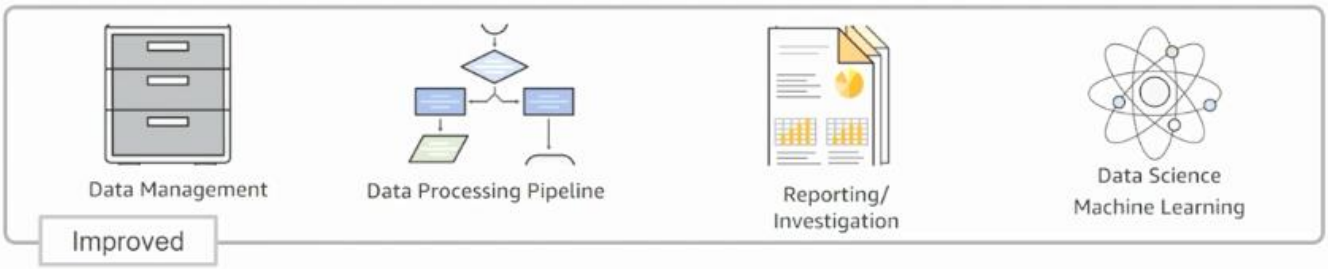
Isolation	Encryption	AuthN/AuthZ	Monitoring
			
VPC isolation Security Groups VPC Endpoints SDLC Isolation (Accts)	AWS KMS EMR Security Configs S3 SSE S3 KMS EBS KMS	Role-based access IAM ADFS Federation Temporary token access AD LDAP Integration (Apps)	AWS CloudTrail Splunk Nagios

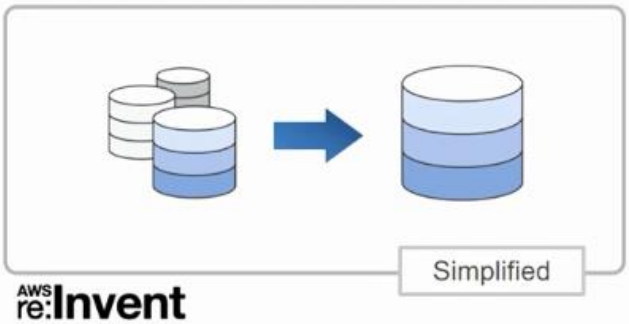
Compliance—consistency, transparency

FINRA



Benefits of a data lake implementation





- [x] Cost Reduction
- [x] Security
- [x] Regulatory Compliance

Achieved

FINRA Presentations re:Invent 2017

FSV307 – Capital Markets Discovery: How FINRA Runs Trade Analytics and Surveillance on AWS
The FINRA analytics platform unlocks the value in capital markets data by accelerating trade analytics and providing a foundation for machine learning at scale. Monday, Nov 27, 10:45 a.m. – 11:45 a.m. Venetian, Level 5, Palazzo P

SID326 – AWS Security State of the Union
Steve Schmidt, chief information security officer of AWS, addresses the current state of security in the cloud. As part of this presentation, John Brady (CISO of FINRA) shares the FINRA journey to the cloud. Wednesday, Nov 29, 12:15 p.m. – 1:15 p.m. MGM, Level 3, Premier Ballroom 316

ABD310 – How FINRA Secures Its Big Data and Data Science Platform on AWS
Learn how FINRA secures its Amazon S3 Data Lake and its data science platform on Amazon EMR and Amazon Redshift, while empowering data scientists with tools they need to be effective. Wednesday, Nov 29, 11:30 a.m. – 12:30 p.m. Aria, Level 3, Juniper 3

ENT328 – FINRA's Managed Data Lake: Next-Gen Analytics in the Cloud
The Financial Impact Regulatory Authority (FINRA) Technology Group has changed its customers' relationships with data by creating a managed data lake Thursday, Nov 30, 1 p.m. – 2 p.m. MGM, Level 3, Premier Ballroom 319

DEV335 – Manage Infrastructure Securely at Scale and Eliminate Operational Risks
Managing AWS and hybrid environments securely and safely while having actionable insights is an operational priority and business driver for all customers. Thursday, Nov 30, 4 p.m. – 5 p.m. Venetian, Level 2, Venetian E


AWS re:Invent
Thank you!
AWS re:Invent
© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.
aws