
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ANT209

Build interactive analytics applications

Raj Devnath (he/him)

Product Manager
Amazon Athena, AWS

Ravi Ramanujam (he/him)

Principal Engineer
FINRA

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Interactive analytics with Amazon Athena for Apache Spark

Key use cases

Getting started

FINRA use case

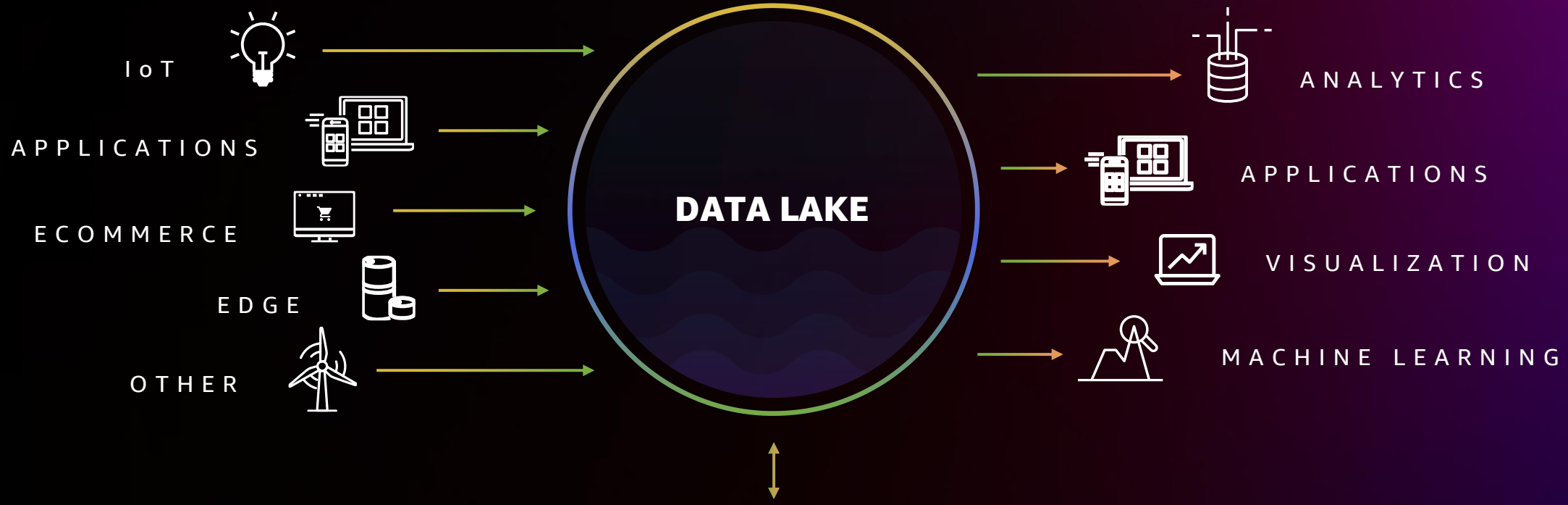
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

NOW AVAILABLE

Amazon Athena for Apache Spark

Get started with interactive analytics on Apache Spark in under a second

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Interactive queries using SQL

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SQL-based query



SIMPLE

Use SQL to extract and aggregate data

- “Select”, “filter” queries
- Aggregate: sum, count, etc.

Easy to learn and use



OPTIMIZED

Optimized performance:
Bring code to the data

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges with SQL

- Range and functionality
  - Difficult for complex data analysis, data science, or machine learning tasks
- Lengthy and hard to read
- Using ecosystem to import functions

```
WITH year_total AS (
  SELECT
    c_customer_id customer_id,
    c_first_name customer_first_name,
    c_last_name customer_last_name,
    c_preferred_cust_flag customer_preferred_cust_flag,
    c_birth_country customer_birth_country,
    c_login customer_login,
    c_email_address customer_email_address,
    d_year dyear,
    sum(((ss_ext_list_price - ss_ext_wholesale_cost - ss_ext_discount_amt) +
      ss_ext_sales_price) / 2) year_total,
    's' sale_type
  FROM customer, store_sales, date_dim
  WHERE c_customer_sk = ss_customer_sk AND ss_sold_date_sk = d_date_sk
  GROUP BY c_customer_id,
    c_first_name,
    c_last_name,
    c_preferred_cust_flag,
    c_birth_country,
    c_login,
    c_email_address,
    d_year
  UNION ALL
  SELECT
    c_customer_id customer_id,
    c_first_name customer_first_name,
    c_last_name customer_last_name,
    c_preferred_cust_flag customer_preferred_cust_flag,
    c_birth_country customer_birth_country,
    c_login customer_login,
    c_email_address customer_email_address,
    d_year dyear,
    sum((((cs_ext_list_price - cs_ext_wholesale_cost - cs_ext_discount_amt) +
      cs_ext_sales_price) / 2)) year_total,
    'c' sale_type
  FROM customer, catalog_sales, date_dim
  WHERE c_customer_sk = cs_bill_customer_sk AND cs_sold_date_sk = d_date_sk
```

SQL Source: TCP-DS benchmarks

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using Python for analytics

- General-purpose programming language
  - Support for variables, classes, loops
- Easily extensible through open-source third-party packages
  - 300K+ ecosystem of libraries*
- Write unit tests for any part of your data processing
- Debugging is easier

**Source: https://pypi.org/project/all-packages/*

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges with Python



Less optimized for data query and aggregation

- Load into local memory
- Latency



Need distributed processing of data sets

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

How do we bridge this?

- SQL and Python are complementary
- SQL: Query + join + aggregate, run fast on a data lake
- Python with wide ecosystem for exploratory analysis
- Open source big data framework such as Apache Spark
  - Allow both SQL and Python
  - Distributed computing libraries
  - Run efficiently over a large cluster



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

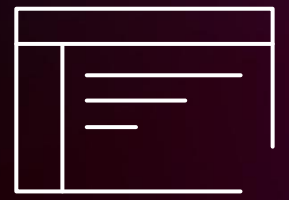
Challenges with Apache Spark



HARDWARE

Set up hardware/clusters
- Slow start
- Delayed productivity

Manage, provision, tune infrastructure



SOFTWARE

Many settings to configure

Software versions to manage

Need Spark expertise

Need ease of use, with a serverless, managed Spark experience

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Athena

Serverless, interactive analytics service



SIMPLE,
INSTANT START



INTERACTIVE,
ADVANCED
ANALYTICS



OPEN AND
FLEXIBLE



COST
EFFECTIVE

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introducing Amazon Athena for Apache Spark

Get the instant, interactive experience of Amazon Athena

with

Popular open source, big data, distributed Apache Spark framework

using

The expressive power and wide ecosystem of Python

AMAZON ATHENA



+

APACHE SPARK

+

python

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Spend more time on insights, not waiting for results

Improve your productivity with interactive applications

Instant on, instant auto-scale



Applications start under 1s

Compute scales instantly to match demand removing need for warm pool

Optimized Spark runtime



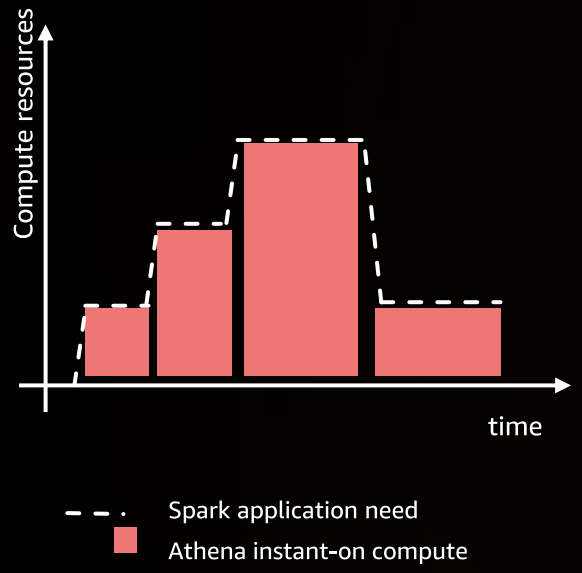
Faster than open source Spark

Get your results faster

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Instant start Spark infrastructure



- Athena provides instant start compute
  - Using Firecracker technology: same as AWS Lambda
  - Sub second start-up
  - Up to 75x faster than other serverless solutions
- Auto-scaling:
  - Compute provisioned matches application need
  - No need for warm pool of cluster

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Harness Apache PySpark for complex & powerful analytics

Use Python for programming



Expressive language for complex questions

Large ecosystem of libraries. Athena provides many Python libraries out of the box for convenience

Simple notebook interface



Use simplified, no-setup notebooks in Athena console

Use notebooks to query data, chain calculations and visualize results

Example notebooks to get started quickly

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Focus on your business application

Serverless



Zero setup cost

Athena manages, provisions, tunes, and scales the infrastructure as needed by the application

Fully managed Spark



Athena handles Spark configurations needed to run applications optimally

Athena manages Spark versions

Run Spark applications without worrying about Spark settings

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Work with all your data where it lives

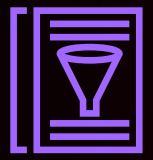
Supports open data formats



Supports Parquet, CSV, and other standard data formats

Point to S3 and start querying

Work seamlessly with AWS Glue



Use data discovered and categorized by AWS Glue

AWS Glue databases and tables are available for query in Athena for Apache Spark out of the box

Power your business applications



Use Athena APIs to interface your business application to your data

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Athena for Apache Spark Pricing

Minimize your idle charges

$0.35/DPU-hour

Pay only for compute used

Pay for drivers and worker nodes used in the Spark applications

Notebook server node is provided free of charge

Run Spark applications cost effectively

- Instant-on infrastructure. No warm pools needed
- Automatic scaling: Match capacity to workload
- Minimize idle fees: Workers turned off during idle

Cost controls

- Account level DPU limits: 160 DPUs
- Session limit: 60 DPUs

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use case details

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

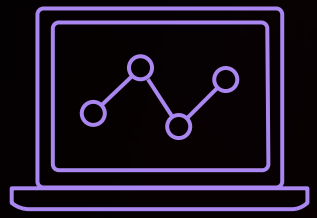
Key use cases: interactive analytics

Data exploration



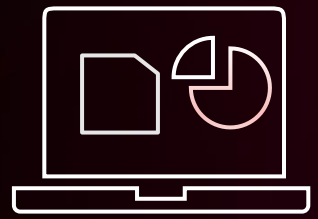
Use Athena and AWS Glue to explore data sets and work with the data

Data lake queries across multiple datasets



Bring together data from various data sets and data formats together to generate insights

Interface to business applications, BI tools



Build SaaS applications that use Athena for Apache Spark to interactively work with data

Data science exploration and feature engineering

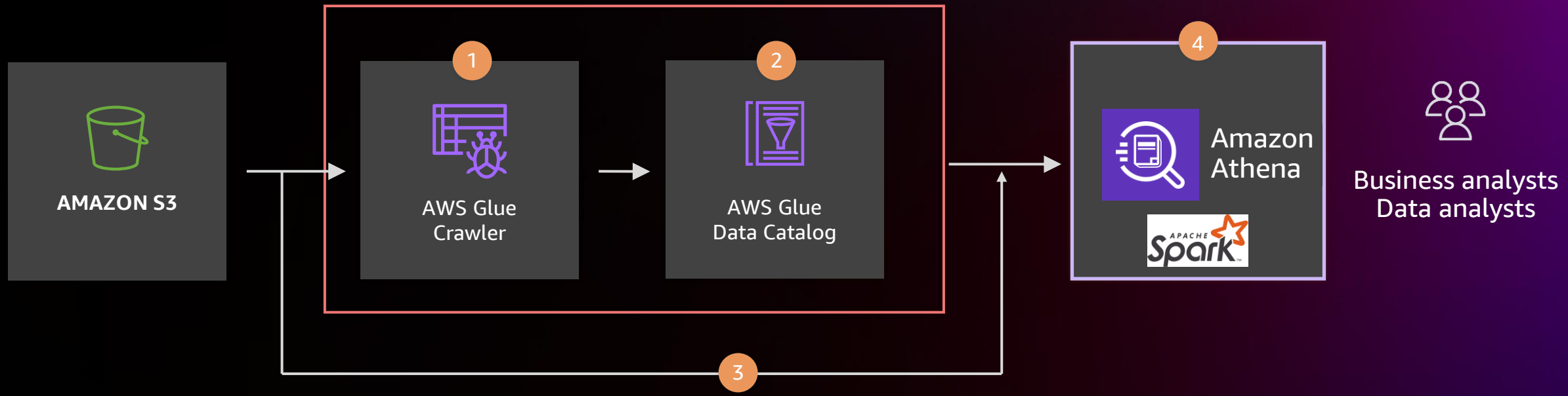


Explore data stores to generate sample datasets, interactive feature generation

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

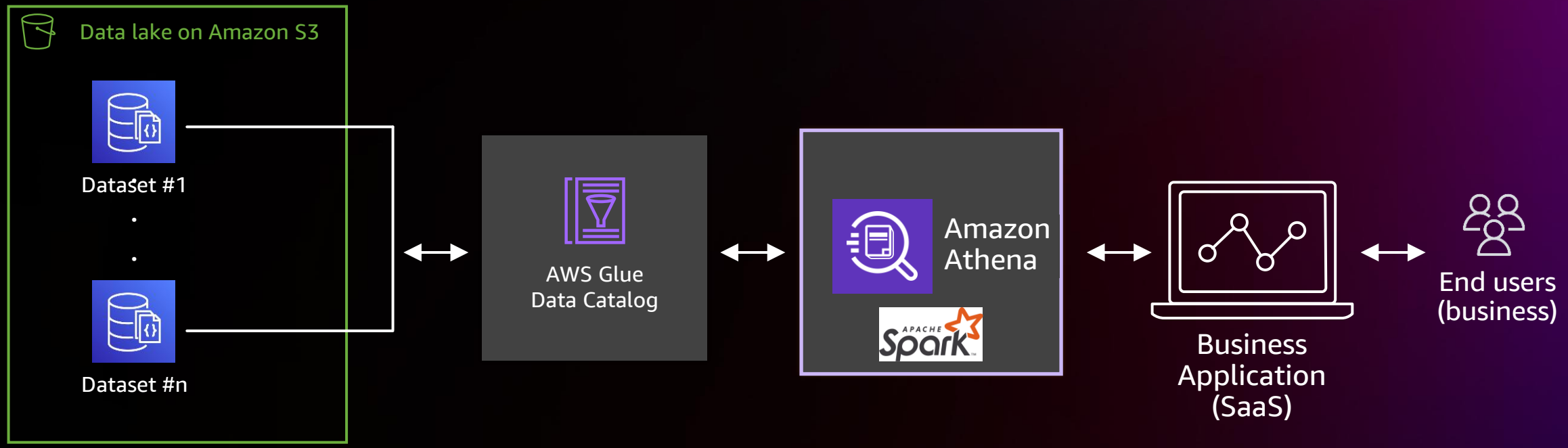
Use case: Interactive data exploration



1. Crawlers scan your data sets and populate the data catalog
2. The data catalog serves as a central metadata repository
3. Alternately, directly infer the schema using Apache Spark
4. Query and visualize your data using notebooks in Amazon Athena for analytics

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use case: Interface to business applications, BI tools



1. Business users log into business applications
2. Business users write interactive queries in business applications
3. Business applications use Athena for Apache Spark APIs to query and join the data from various sources
4. Business application presents the results to the business end users

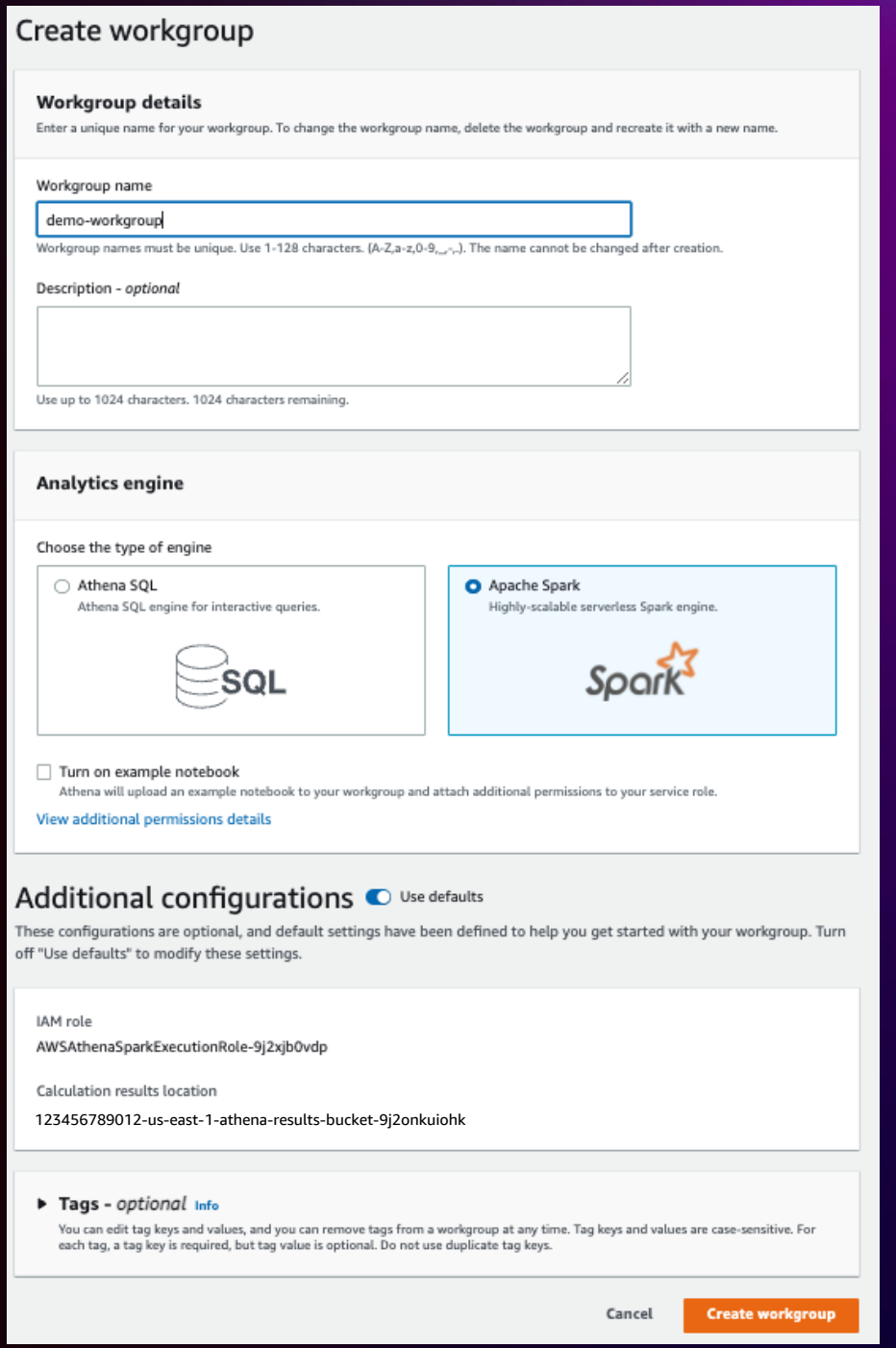
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Getting started

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

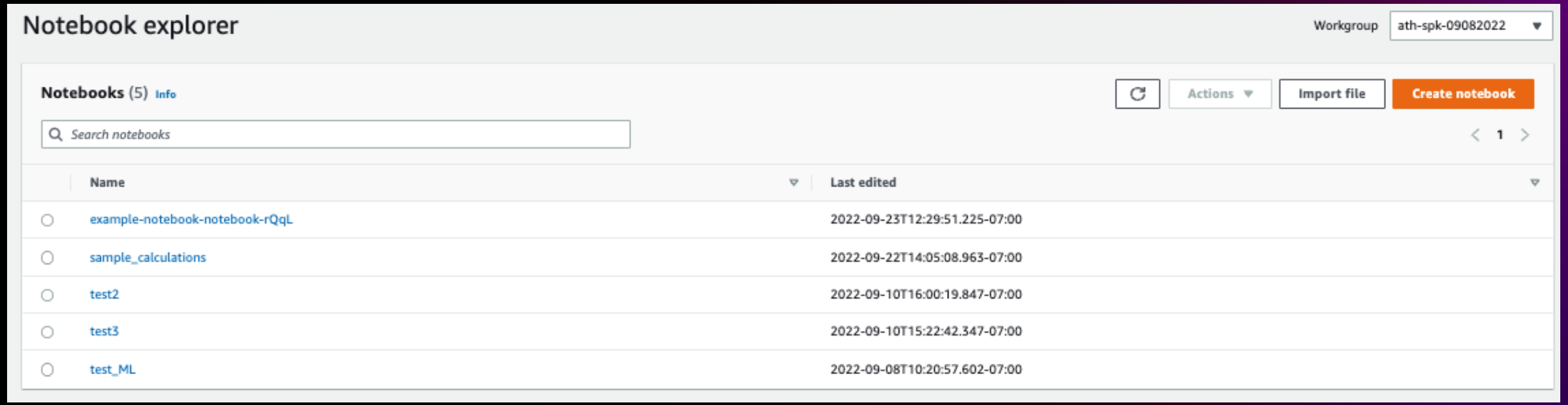
Create workgroup

- Separates workloads between engines
  - Cost, security boundary
- Simplified workgroup creation
  - Smart defaults for IAM roles, S3 calculation results
- Notebook and S3 bucket encryption
  - Using AWS key or customer KMS key



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Explore notebooks

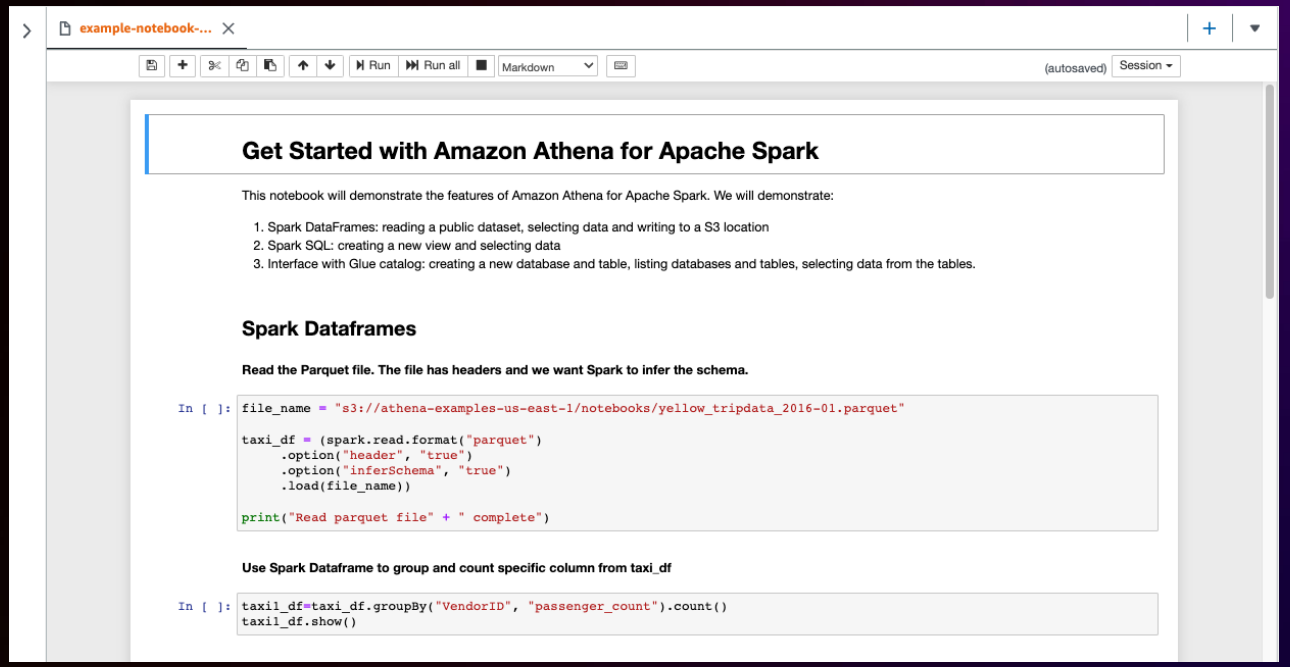


- Import notebooks from external sources
- Open, rename existing notebooks
- Export notebooks to share with other users

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use notebooks to build PySpark applications

- Start with example notebook or build your own
- Simplified Jupyter-based notebook interface
- Tabbed interface for multiple notebooks
- New Spark session for every notebook
- Customize session parameters
- Calculation results are stored in S3
- Session and calculation detail page



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

PySpark details

Apache Spark

- Spark version 3.2
- DataFrame, Spark SQL
- Distributed PySpark Pandas API

Python

- Version 3.9
- Add custom libraries: sc.addPyFile()
- Out of box convenience libraries

Category	Examples
Math, ML	NumPy, Pandas, SciPy, scikit-learn, mpmath
AWS	boto3, botocore
Visualization	Matplotlib, Seaborn, Plotly
Python utils	Requests, urllib3, pyparsing, python-dateutil

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Athena for Apache Spark at FINRA

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Who We Are



FINRA processes and analyzes **630k+ brokers** looking for fraud, abuse, and insider trading.

Named **#1 Best Place to Work in IT** 10th appearance on the annual list

FINRA regulates **3,400 securities firms,** across US Equities & Options market transactions

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Who We Are

FINRA

FINRA processes at peak volume

600+ billion

market events per day

24 Exchanges

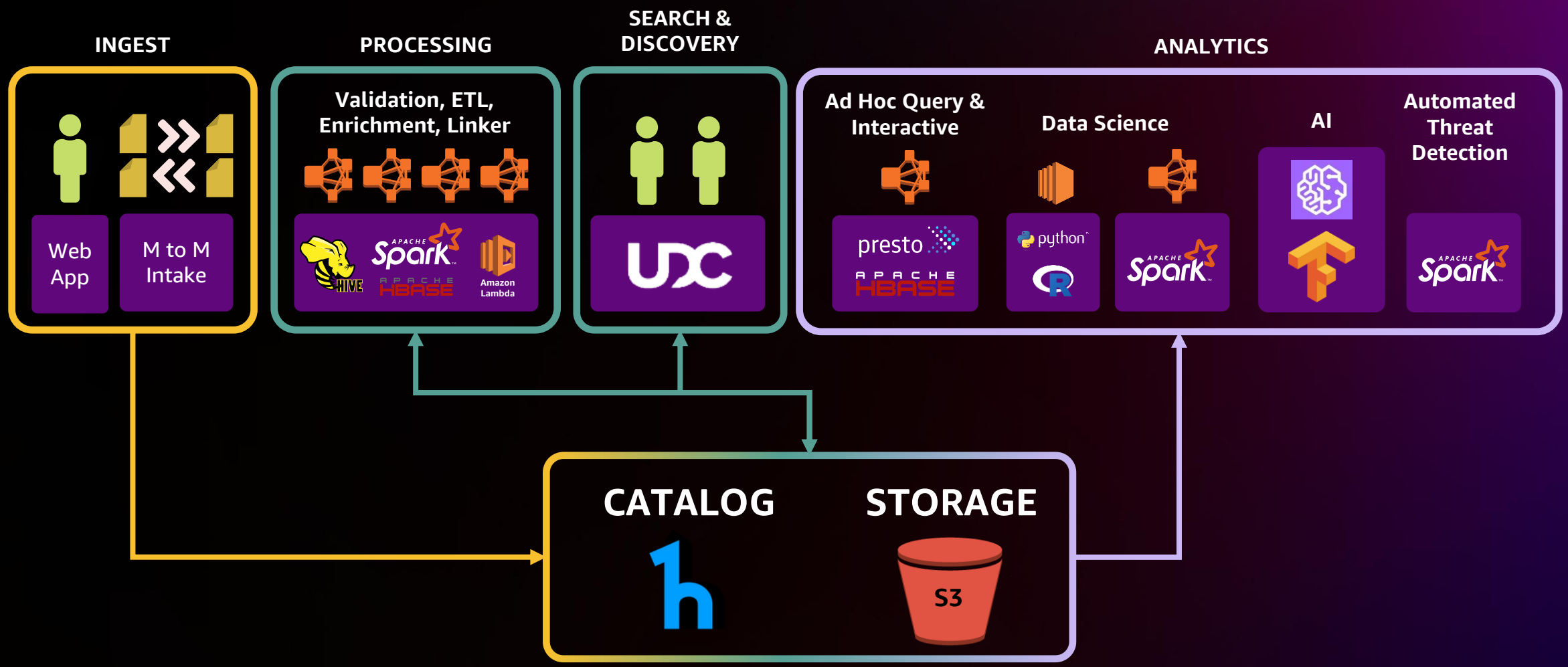
FINRA runs up to **300K+ compute nodes** per day

FINRA manages approximately **500+ Petabytes** of storage

aws

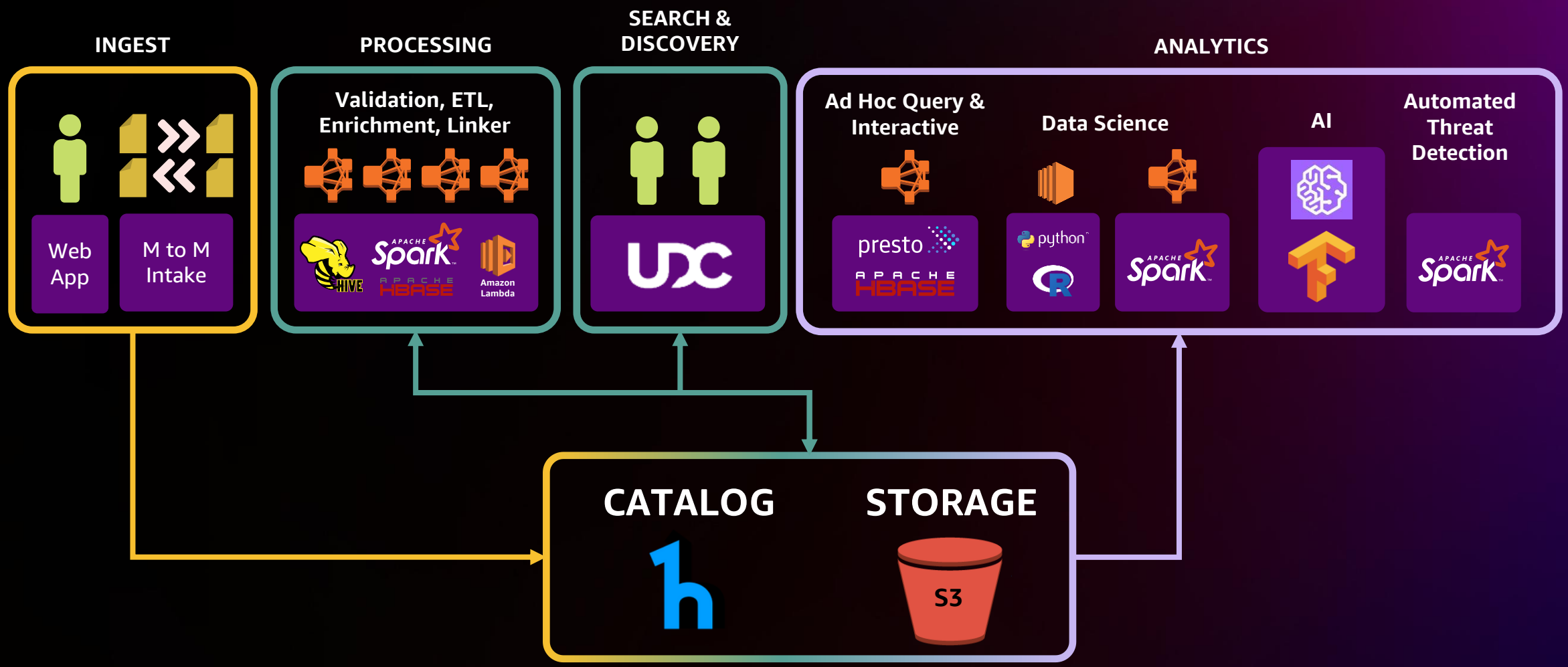
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our Data Processing Ecosystem



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our Data Processing Ecosystem

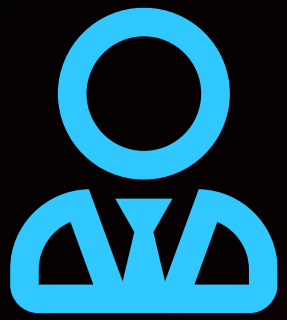


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Analysis: Personas

FINRA

I need to be agile with changing Market Dynamics



Business Users

I would like to understand the data well

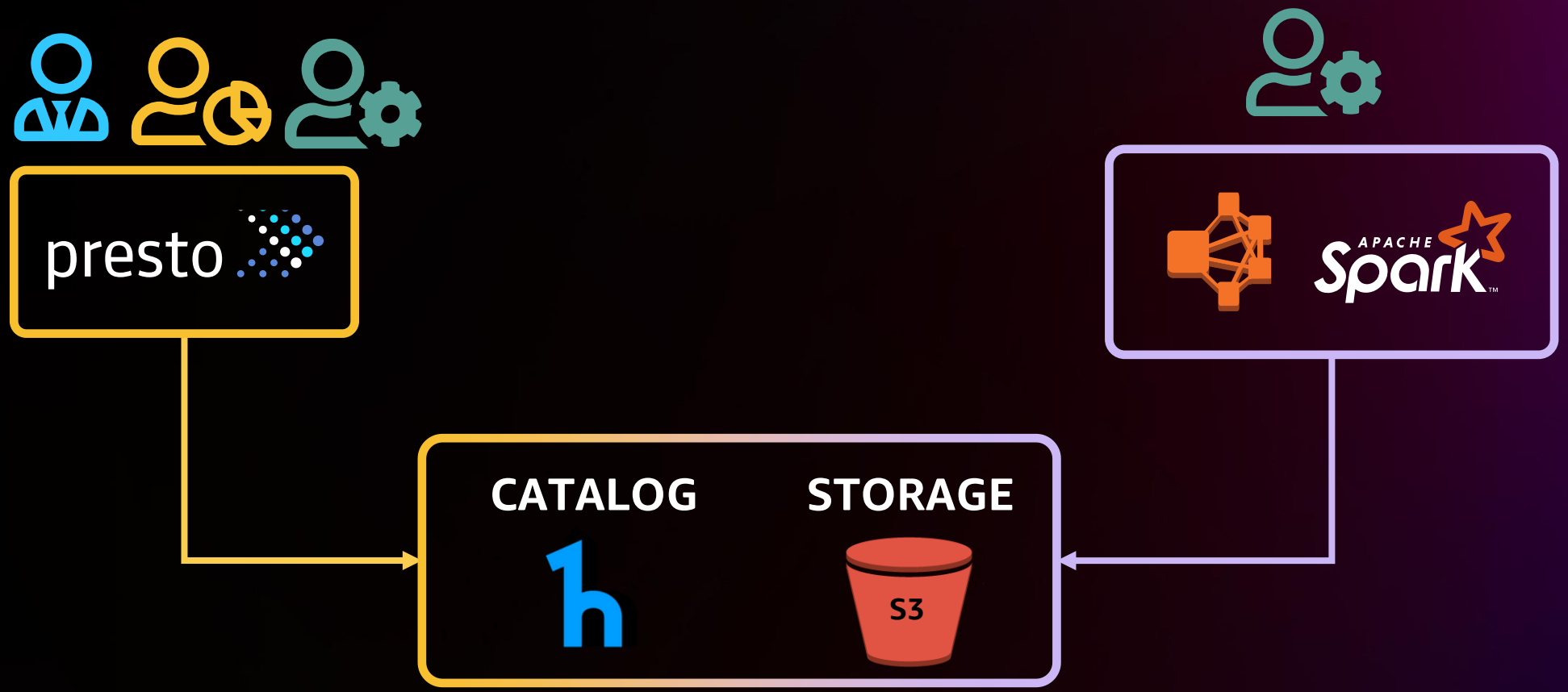


Analysts

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Analysis: Current State



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Analysis: Business Users & Analysts

Be Nimble



Existing Metrics can't be leveraged

Empower

Spark-based analysis needs engineering assistance

Presto-based sharing involves spreadsheets, no option for viz

Prototyping needs deeper insights on larger datasets

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Analysis: Business Users & Analysts

Be Nimble



Existing Metrics can't be leveraged

Empower

Spark-based analysis needs engineering assistance

Presto-based sharing involves spreadsheets, no option for viz

Prototyping needs deeper insights on larger datasets

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Spark on Athena: Our work



Data Partition

12 TB

47 Billion rows

343 Columns

Parquet on S3



Libraries



Bring Your Own Library



Catalog



AWS Glue Catalog

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our Findings



Instant Startup

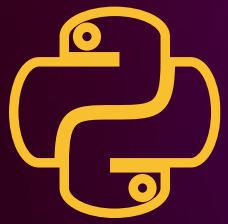
No cluster startup time



Dynamic Scaling

EMR: Configure Cluster (Instance Type, Count)

Spark on Athena: No configuration needed, Scale Dynamically



Python Support

Expressive language helps with complex data analysis across large dataset(s)

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our Findings



Data Sharing

Share Notebooks to users within Workgroup

Narrate data using visualization



Cost Optimization

Pay for what you use

Only Driver node is up when notebook is not actively running



Security

Athena: Internode Encryption

S3: KMS

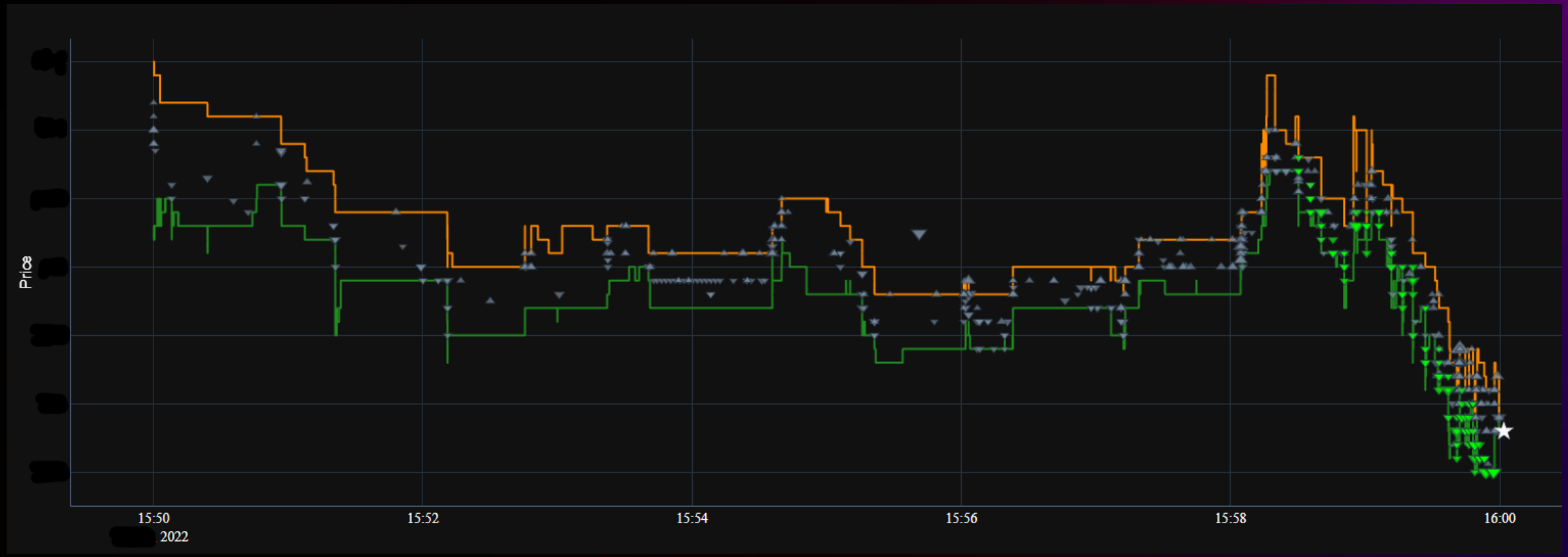
IAM roles mapped to Athena Workgroup

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Analysis: Storytelling

Marking the close



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Interactive Analysis: Business Users & Analysts



Faster Analysis

Instant startup of Spark Notebook



Better Viz

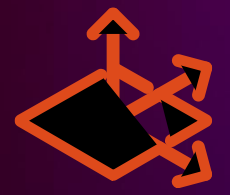
Users able to see data in a better presentable format



Empowered Users

Low friction to Adoption

No need for Spark or Cluster configuration



Deeper Insights

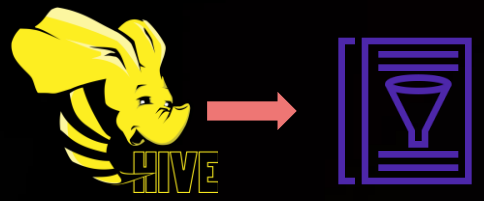
Deep Insights to data

Better Prototyping

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Next Steps



Data Catalog

Use AWS Glue catalog



Federated Access

Fine-grained data access controls for users



Git Integration

Code version control

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Learn more about Amazon Athena for Spark at

aws.amazon.com/athena/spark

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Raj Devnath
rajdevn@amazon.com

Ravi Ramanujam
ravi.ramanujam@finra.org



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.