

The background features a dark blue gradient with large, overlapping, semi-transparent shapes in shades of purple and magenta. Two thin, light blue lines cross the scene diagonally. The text is positioned on the left side of the image.

AWS re:Invent

DECEMBER 2 - 6, 2024 | LAS VEGAS, NV

ANT352

AI-powered data integration and governance with Amazon Q Developer

Vipin Mohan

Principal Product Manager
Amazon Web Services

Mohit Saxena

Senior Software Dev Manager,
AWS Glue & Amazon EMR
Amazon Web Services



Agenda

- 01 Challenges builders are facing
- 02 What is Amazon Q Developer?
- 03 What is SageMaker Unified Studio?
- 04 Amazon Q in SageMaker Unified Studio
- 05 Amazon Q & generative AI capabilities for Apache Spark in SageMaker Data Processing
- 06 Summary

Let's do a quick poll . . .

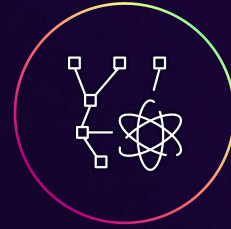
WHICH HAT(S) DO YOU WEAR IN YOUR ORGANIZATION?



**Data engineer/
Data analyst**



**Data steward/
governance**



**Data scientist/
ML scientist**



App developer

What challenges are builders facing?



Onboarding

- Find accurate and specific technical guidance from documentation, best practices, resources
- Understand complex schemas, comprehend structure and “relationships of diverse datasets”



Governance

- Access to relevant data, quickly and securely
- Admin controls and operations
- Protecting sensitive data

What challenges are builders facing?



Code development

- Generate code, identify, and mitigate code issues
- Modernize and refactor old code and dependencies



Troubleshooting

- Difficulty identifying root causes
- Manual debugging

Amazon Q Developer



Reimagines the experience across the entire software development lifecycle (SDLC)

Helps developers and IT professionals build and manage secure, scalable, and highly available applications

Helps you write, debug, test, optimize, and upgrade your code faster

Converses with you to explore new AWS capabilities, learn unfamiliar technologies, and architect solutions

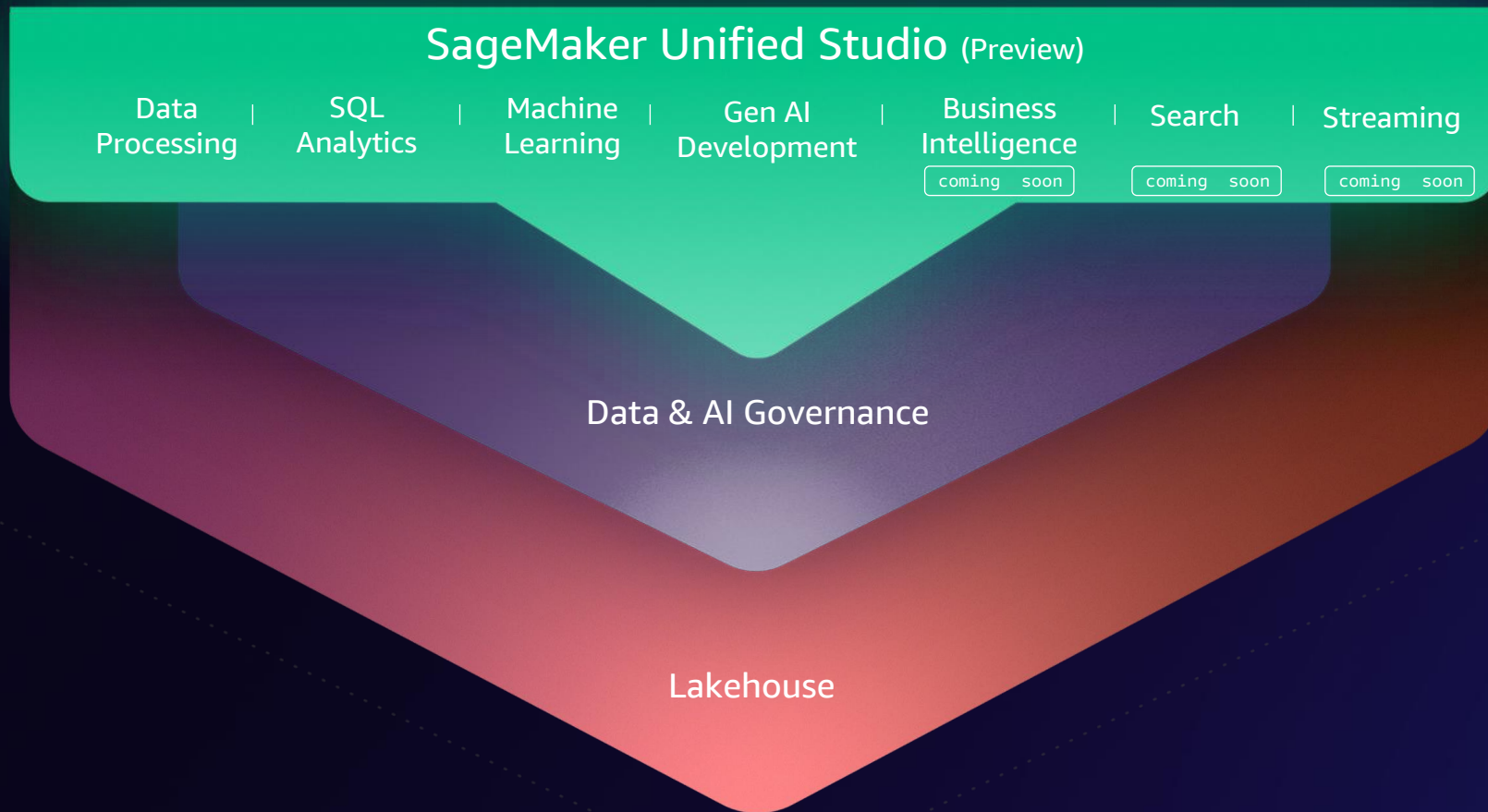
Amazon Q is built with security and privacy in mind from the start, making it easier for organizations to use generative AI safely

Amazon Q Developer in SageMaker Unified Studio



The next generation of Amazon SageMaker
is the center for **Data**, **Analytics & AI**

Amazon SageMaker



NEW

[Preview]

SageMaker Unified Studio

SageMaker Unified Studio

All of your data and tools for analytics in a single data and AI development environment

Use **best-in-class** tools, no matter the job

Train, customize, and deploy AI models at **scale**

Rapidly build custom generative AI applications

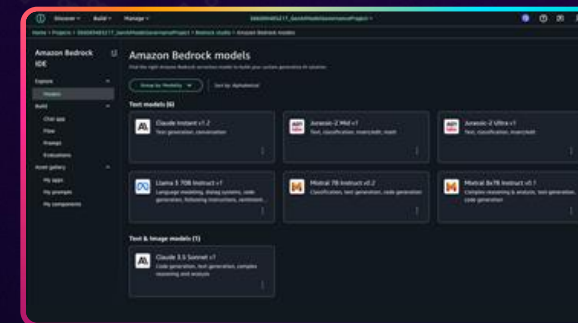
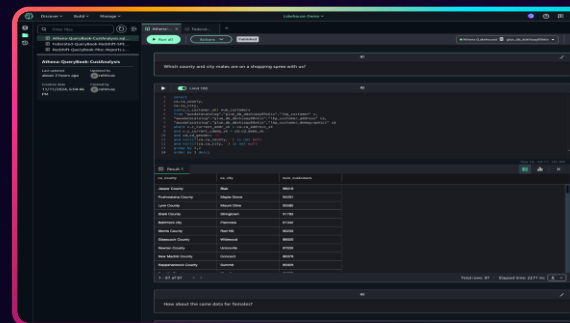
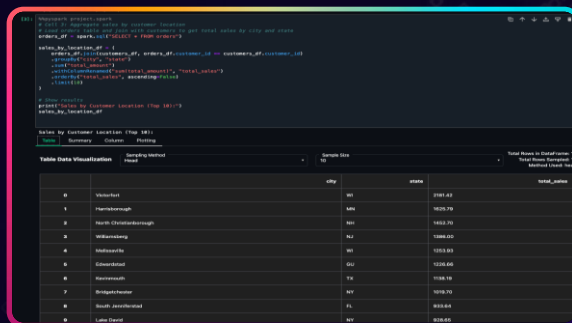
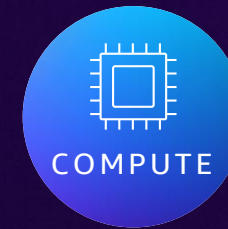
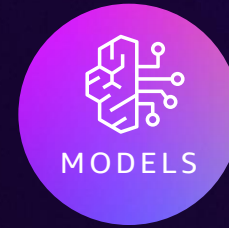
Accelerate your data journey with Amazon Q Developer



Embedded data and AI governance

Seamless access to all data, all resources, and all tools

NEW
[Preview]

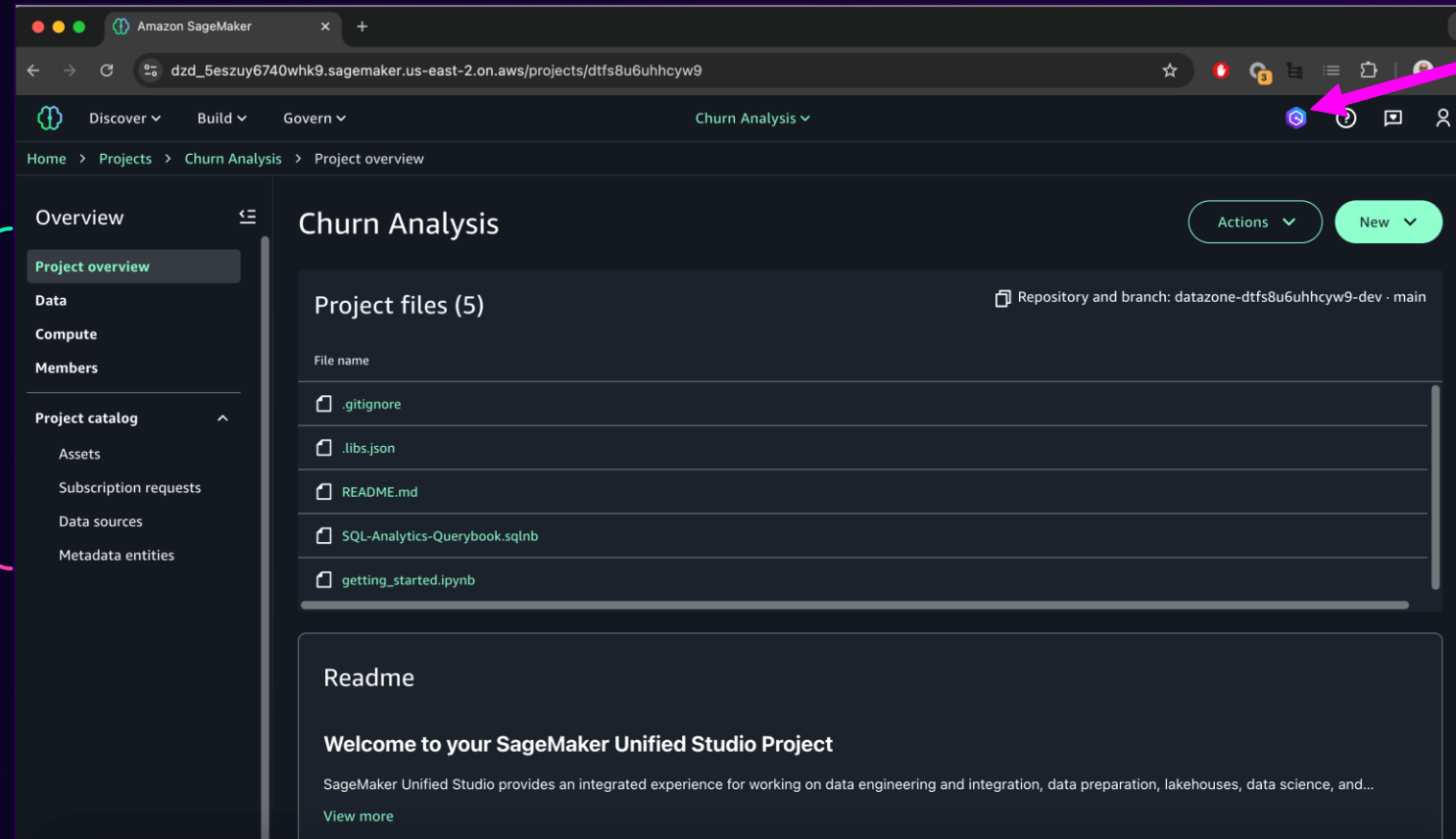


SageMaker Unified Studio

TOOLS, DATA, AND CODE UNIFIED IN ONE PLACE

NEW

[Preview]



Gen AI assisted experience powered by Amazon Q

See the code, data, ML, and compute for your projects in one view

Work with notebooks, queries, models, and more from familiar AWS tools



Amazon Q in SageMaker Unified Studio

NEW

[Preview]



Q & A

Ask questions about SageMaker, understand best practices, quickly get info about your AWS resources, and more!



Data discovery

Search and discover data from your BDC using natural language



Code authoring

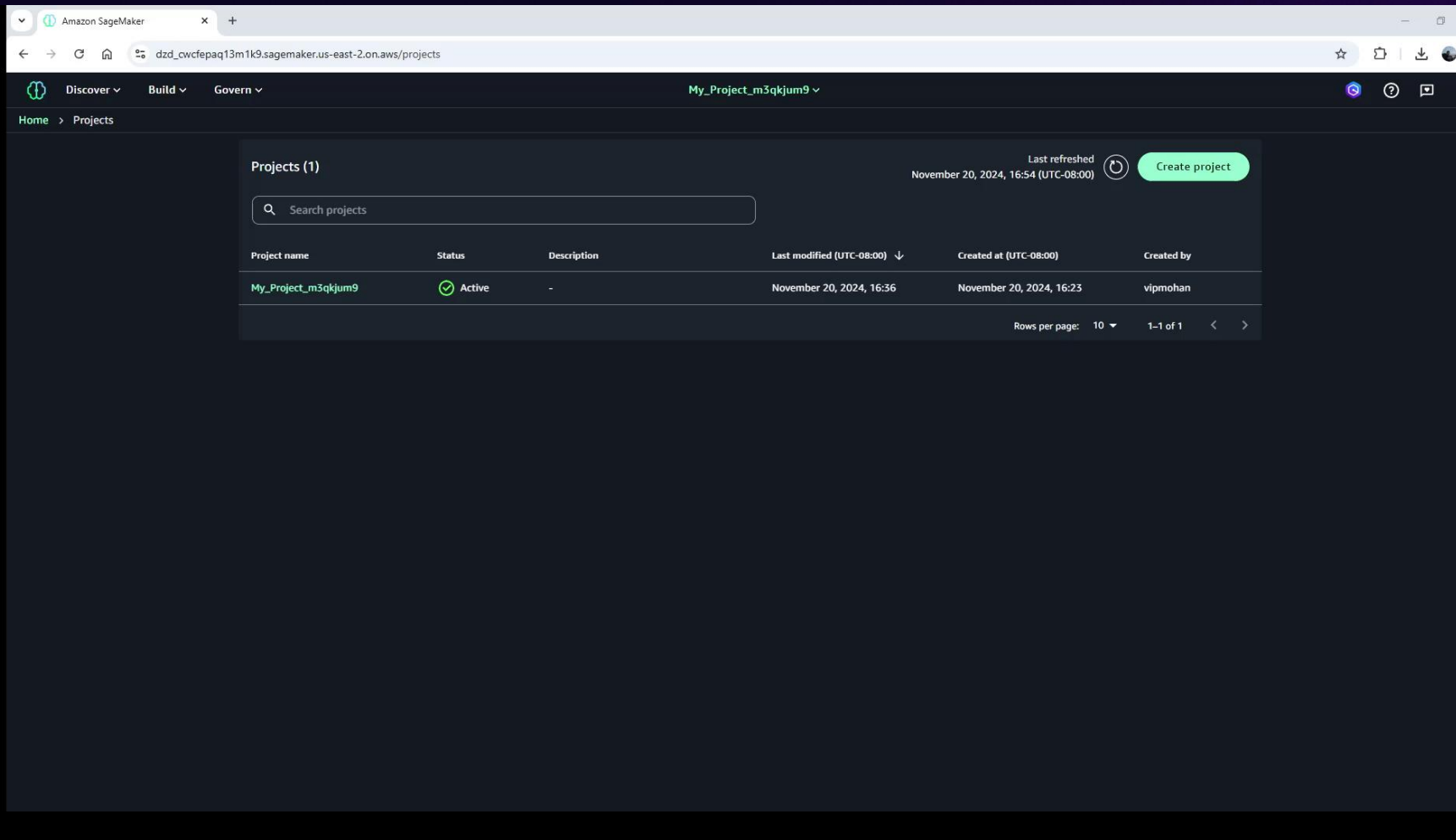
Generate code snippets or entire functions, identify and fix errors, upgrade apps to newer versions



SQL querying

Simplify query authoring and increase productivity. Express queries in natural language and receive SQL code recommendations.

An instant guide – Get answers quickly & easily NEW [Preview]



Step-by-step assistance, making it easier for you to learn about SageMaker platform and ramp up quickly

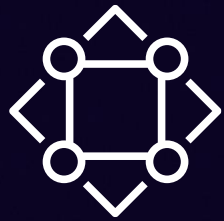
Get guided answers based on extensive AWS knowledge and Well-Architected best practices

Governance – Search and Discover data

NEW

[Preview]

DISCOVER, GOVERN, AND COLLABORATE ON DATA AND AI SECURELY



Discover

Automate data discovery and cataloging with ML and gen AI

SageMaker Catalog

Gen AI powered business metadata generation

Leverage natural language and semantic search to discover most relevant data assets

Curation of business catalog at asset and column level

Business context and recommended analysis for datasets with generative AI

Remove time from manual entry of data attributes in the data catalog

- Overview
- Project overview
- Data
- Compute
- Members
- Project catalog
 - Assets
 - Subscription requests
 - Data sources
 - Metadata entities

Data sources

[What's this?](#)

CREATE DATA ASSET CREATE DATA SOURCE

Use this section to add new or manage existing data sources for publishing data assets in Amazon DataZone.

Find sources

SOURCE TYPE: ALL STATUS: ALL

Type	Source name	Connection	Schedule	Source status	Last run	Actions
	590183921665-catalog_0746-default-datasource	project.iam	MTWTFSS	Ready Enabled	✓ Nov 19, 2024, 08:06:06 PM	⋮
	590183921665-AwsDataCatalog-glue_db_0746_5vf82rnba1porr-default-datasource	project.iam	MTWTFSS	Ready Enabled	✓ Nov 20, 2024, 10:03:03 AM	⋮
	Tooling-default-sagemaker-modelpackagegroup-datasource	project.iam	MTWTFSS	Running Enabled	○ Nov 19, 2024, 07:52:07 PM	⋮

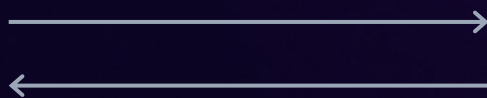
< 1 >

How does this work?



Write a prompt to search for assets in the Business Data Catalog

SageMaker Unified Studio sends prompt to Amazon Q



Receive a list of semantic search results



Amazon Q

Secure and safe - Uses permissions associated with the user

Searches metadata in SageMaker Catalog associated with your domain

Generative SQL in SageMaker Unified Studio

NEW

[Preview]

The screenshot displays the Amazon SageMaker Unified Studio interface. The browser address bar shows the URL: `dzd_4v9rmx6it4bip3.sagemaker-gamma.us-east-1.on.aws/projects/6okrr3566ktpnr`. The interface includes a navigation menu on the left with sections like 'Overview', 'Data', 'Compute', 'Members', 'Project catalog', 'Assets', 'Subscription requests', 'Data sources', and 'Metadata entities'. The main content area shows the 'Project overview' for 'My_Project_m3nwwnda'. It features a 'Project files (5)' section with a list of files: `.gitignore`, `.libs.json`, `README.md`, `getting_started.ipynb`, and `query_book_demo_v1.sqlnb`. Below this is a 'Readme' section with the text: 'Welcome to your SageMaker Unified Studio Project' and 'SageMaker Unified Studio provides an integrated experience for working on data engineering and integration, data preparation, lakehouses, data science, and generative AI initiatives...'. At the bottom, there is a 'Project details' section with a table of project information.

Project name	Project ID	Last modified	Amazon S3 location
My_Project_m3nwwnda	6okrr3566ktpnr	November 18, 2024, 19:54 (UTC-08:00)	s3://amazon-sagemaker-590183921665-us-east-1-

Increase productivity

Simplify query authoring. Ask questions in plain English and get SQL code suggestions

Onboard quickly

Generate custom SQL code without needing to understand your organization's complex database

Securely access your data

Experience consistent data governance compliance and user permissions



Amazon Q & generative AI capabilities for Apache Spark in SageMaker Data Processing



SageMaker Data Processing with Apache Spark

ANALYZE, PREPARE, AND INTEGRATE DATA FOR ANALYTICS AND AI

SageMaker Data Processing

Performance optimized runtime for Apache Spark



Best price performance for big data analytics

Apache Spark on AWS Glue and Amazon EMR

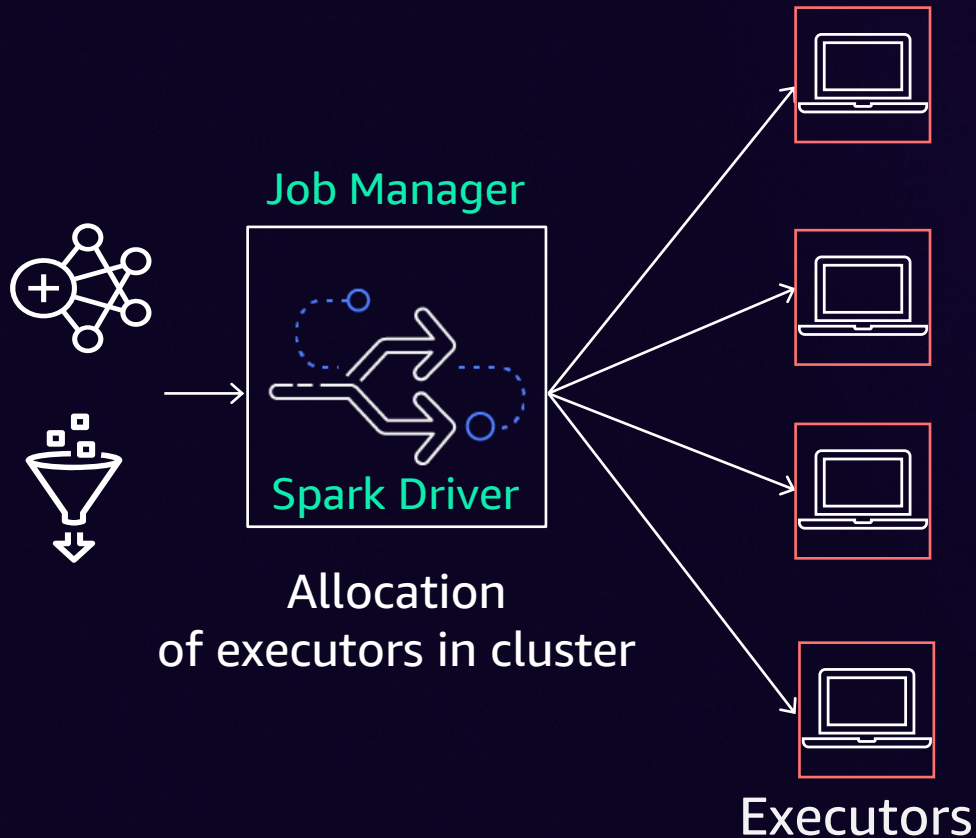
Discover, prepare, and integrate all your data at any scale using AWS Glue
Peta Byte scale data analytics using Amazon EMR

All-in-one serverless data integration service

Flexibility to run big data workloads on EC2, EKS, and in serverless mode

Fine grained access control with Amazon SageMaker Lakehouse

Apache Spark execution model



Massive distributed parallelism

Apache Spark runs **data parallel** jobs

Jobs are divided into **stages**

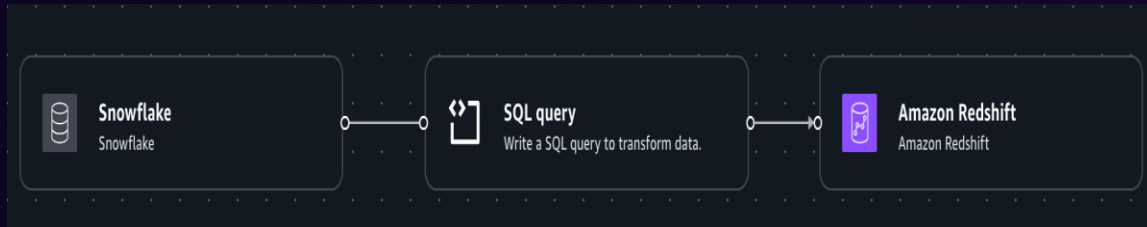
Data is divided into **partitions (shards)** that are processed concurrently

1 stage x 1 partition = 1 **task**

Job manager schedules tasks on **executors**

Spark applications can allocate **1000s of executors**

Apache Spark programming interface



ETL DAG

- Highly expressive language
 - Mix of **SQL** and **Python**
- Extensive **configuration** options
- **Lazy evaluation** of operations

```
import sys
from awsglue.transforms import *
from pyspark.context import SparkContext
from pyspark.sql import SparkSession
from connectivity.adapter import CatalogConnectionHelper

sc = SparkContext.getOrCreate()
spark = SparkSession.builder.getOrCreate()

# Snowflake connection options
ConnectionV2DataSource_dsource1_additional_options = {
    "dbtable": "airport_tbl",
    "sfDatabase": "airports_db",
    "sfSchema": "public",
    "autopushdown": "on",
}

# Snowflake source DataFrame
ConnectionV2DataSource_dsource1 = CatalogConnectionHelper(spark).read(
    "snowflake", "sf_connection_name",
    ConnectionV2DataSource_dsource1_additional_options
)

# SparkSQL transformation to project columns
ConnectionV2DataSource_dsource1.createOrReplaceTempView("airport")
query = "SELECT name, country, build_time FROM airport"
ProjectedColumnsTransformDF = spark.sql(query)

# Redshift connection options
ConnectionV2DataSink_dsink1_additional_options = {
    "dbtable": "redshift_db",
    "tempdir": "s3://redshift_tmp_bucket/staging_location",
}

# Redshift data sink
CatalogConnectionHelper(spark).write(
    ProjectedColumnsTransformDF,
    "redshift",
    "redshift_connection_name",
    ConnectionV2DataSink_dsink1_additional_options,
)
```

PySpark application code

Core use cases of gen AI for data processing with Spark

Build



Troubleshoot



Modernize &
upgrade

Build data processing and data integration **visual flows** with English

In-line code recommendations for **developers** in Notebooks



Days → Minutes
Mean time to resolution

Instant **root cause analysis** for Spark jobs with actionable recommendations

Stay current, stay competitive
Effortless upgrade to latest Spark releases

Modernize Spark Jobs without the manual overhead

NEW

[GENERALLY AVAILABLE]

Amazon Q for data integration with Spark

INTEGRATE DATA WITH NATURAL LANGUAGE

Enables data engineers and developers to
build data integration Spark jobs **faster**



Build Spark data processing flows faster



English to Visual DAG and Spark Code

Advanced models and compiler techniques



Context awareness

Connection and job level context awareness



Multi-turn chat

Iterate with Amazon Q until you get your desired pipeline

In 2024, Amazon Q has helped tens of thousands of users globally with 100K+ questions for data integration

Browser tabs: Isengard, AWS Glue | us-east-1, Amazon SageMaker, Isengard

URL: https://dzd_6mqvlfzo2th7r.sagemaker.us-east-1.on.aws/projects/4k2eu5pc09ywx3/data

Navigation: Discover, Build, Govern, My_Project_nov24test3

Breadcrumbs: Home > Projects > My_Project_nov24test3 > Data

Overview

Project overview

- Data
- Compute
- Members

Project catalog

- Assets
- Subscription requests
- Data sources
- Metadata entities

Lakehouse

- AwsDataCatalog
 - glue_db_4fthqih3vvk1if
 - canvas-sample-housing
 - event
 - events
 - housing
 - housing2
 - housing3
 - housing4
 - sf_pcd_complete_new_feet_0
 - venue
 - galcoolrms

Redshift

S3

event

Actions

Table details

Description	Last updated	Creation date
Table created from uploading data in SageMaker Unified Studio	November 26, 2024, 10:37 (UTC-08:00)	November 26, 2024, 10:37 (UTC-08:00)
Created by	Publish status	Add business metadata
arn:aws:sts::356769412531:assumed-role/datazone_usr_role_4k2eu5pc09ywx3_5oyri522ckqip/d4189468-7031-70fd-3717-37486ba0bfa5@5oyri522ckqip	Not published yet	add

Columns

Sample data

Results (50)

find rows

dateid	eventid	catid	starttime	eventname	venueid
1851	1	8	2008-01-25 14:30:00	Gotterdammerung	305
2114	2	8	2008-10-15 20:00:00	Boris Godunov	306
1935	3	8	2008-04-19 14:30:00	Salome	302
2090	4	8	2008-09-21 14:30:00	La Cenerentola (Cinderella)	309
1982	5	8	2008-06-05 19:00:00	Il Trovatore	302

NEW

PREVIEW IN AWS GLUE

Generative AI troubleshooting for Apache Spark in AWS Glue

AUTOMATE SPARK TROUBLESHOOTING

Detect root cause of Spark job failures automatically

Isolate line of code with AI-driven insights

Recommend actionable solutions for complex Spark failure patterns



Spark manual troubleshooting flow

Recent job runs (1) Info Last updated (UTC) March 13, 2024 at 00:23:36 Table View Card View

Filter job runs by property < 1 >

2024/03/12 21:52:23, Rewind job bookmark

Error Category: UNCLASSIFIED_SPARK_ERROR; org.apache.spark.SparkException : Job aborted due to stage failure: Total size of serialized results of 7 tasks (1160.2 MiB) is bigger than spark.driver.maxResultSize (1024.0 MiB)

- Distributed logs, metrics, and Spark UI for debugging
- Connecting the dots require hours to days

```
glue-etl-scala-job
Last modified on 3/12/2024, 2:52:21 PM Actions Save Run

Script Job details Runs Data quality - updated Schedules Version Control

Script info
1 import com.amazonaws.services.glue.ChoiceOption
2 import com.amazonaws.services.glue.GlueContext
3 import com.amazonaws.services.glue.MappingSpec
4 import com.amazonaws.services.glue.ResolveSpec
5 import com.amazonaws.services.glue.errors.CallSite
6 import com.amazonaws.services.glue.util.GlueArgParser
7 import com.amazonaws.services.glue.util.Job
8 import com.amazonaws.services.glue.util.JsonOptions
9 import org.apache.spark.SparkContext
10 import scala.collection.JavaConverters._
11 import org.apache.logging.log4j.Logger
12 import org.apache.logging.log4j.LogManager
13
14 object GlueApp {
15   val LOG:Logger = LogManager.getLogger(this.getClass)
16   def main(sysArgs: Array[String]) {
17     val sparkContext: SparkContext = new SparkContext()
18     val glueContext: GlueContext = new GlueContext(sparkContext)
19     val spark = glueContext.getSparkSession
20
21     val df = sparkSession.read.option("delimiter", "|").csv("s3://my-demo-bucket/my-demo-prefix/store_sales/")
22     df.show(1)
23     df.printSchema()
24
25     df.createOrReplaceTempView("t1")
26     val df2 = sparkSession.read.option("delimiter", "|").csv("s3://redshift-downloads/TPC-DS/1TB/store_sales/") //sqlContext.sql("select * from t1")
27     df2.createOrReplaceTempView("t2")
28
29     val res = sqlContext.sql("SELECT COUNT(*) , t1_c2 FROM t1 GROUP BY t1_c2")
30     df.union(df2).createOrReplaceTempView("res")
31     val res = sqlContext.sql("SELECT COUNT(*) , res_c2 FROM res GROUP BY res_c2")
32
33     println("the result is")
34     val frame = spark.read.csv("s3://aws-glue-temporary-782104008917-us-west-2-encrypted/lxiaobin/data/csv-250MB/3TB/store_sales/")
35     frame.collect()
36   }
37 }
```

Check Glue Error Logs

2024-03-12T21:56:44.397Z 2024-03-12 21:56:44,396 ERROR [main] glue.ProcessLauncher (Logging.scala:logError(77)): Exception in Us...

Spin up Spark History Server and view the the Spark UI Logs

Spark 3.3.0-bran-1 Jobs Stages Storage Environment Executors S3C / Data frame native-spark-glue-etl-scala-job-j... application U...

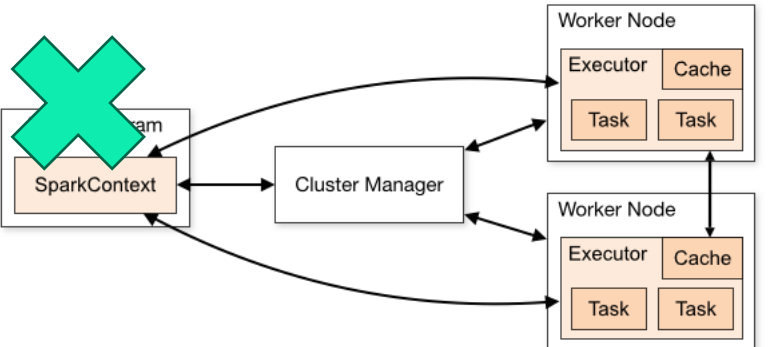
Spark Jobs [?] Users: spark Total Uptime: 33 min Scheduling Mode: FIFO Completed Jobs: 2 Failed jobs: 1

Event Timeline

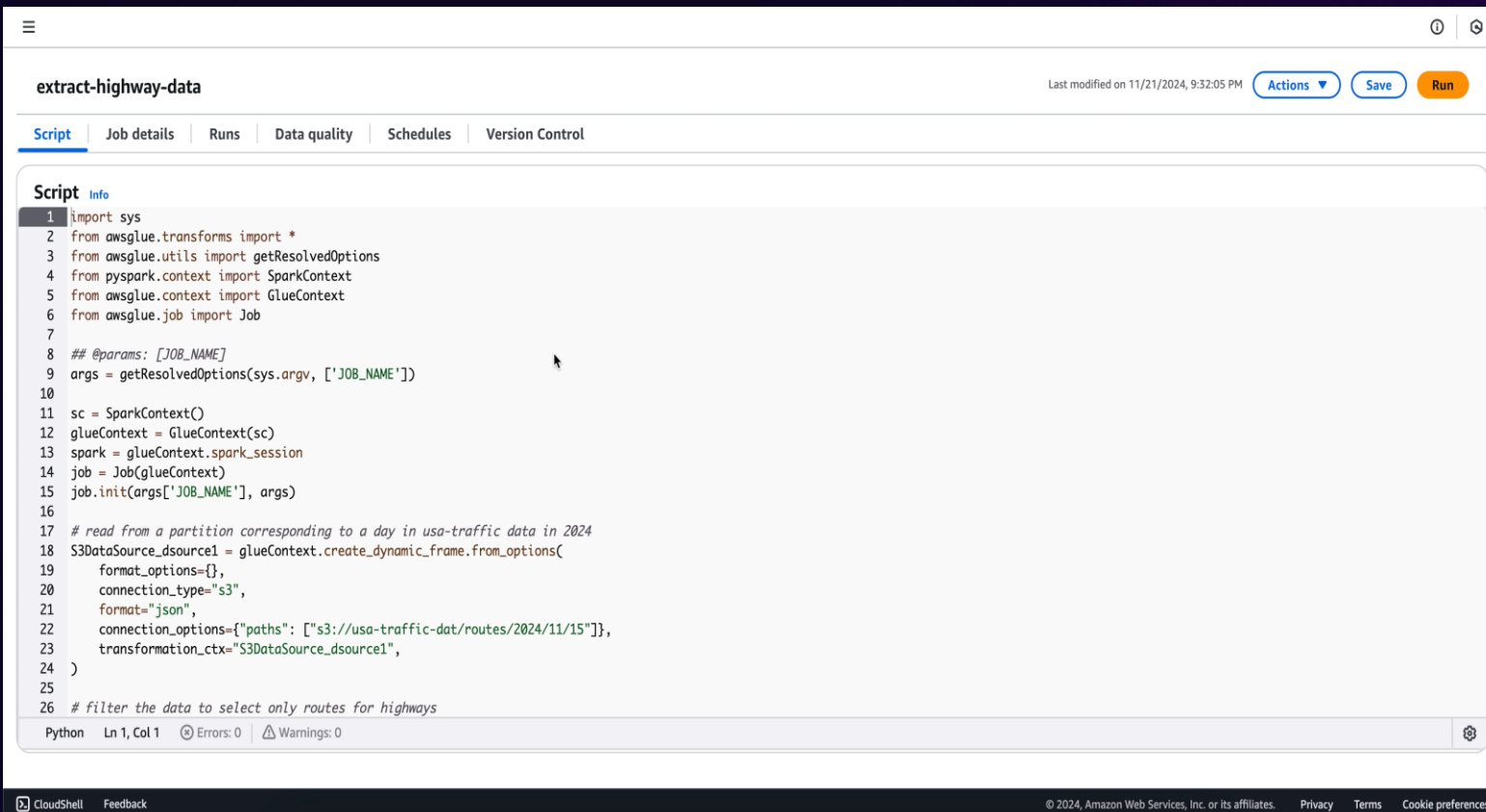
- Executors
- Jobs

Reach out to AWS Support / Research on internet

What does "org.apache.spark.SparkException : Job aborted due to stage failure: Total size of serialized results of 7 tasks ..." mean??



Gen AI Spark troubleshooting workflow



```
extract-highway-data Last modified on 11/21/2024, 9:32:05 PM Actions Save Run  
Script | Job details | Runs | Data quality | Schedules | Version Control  
Script Info  
1 import sys  
2 from awsglue.transforms import *  
3 from awsglue.utils import getResolvedOptions  
4 from pyspark.context import SparkContext  
5 from awsglue.context import GlueContext  
6 from awsglue.job import Job  
7  
8 ## @params: [JOB_NAME]  
9 args = getResolvedOptions(sys.argv, ['JOB_NAME'])  
10  
11 sc = SparkContext()  
12 glueContext = GlueContext(sc)  
13 spark = glueContext.spark_session  
14 job = Job(glueContext)  
15 job.init(args['JOB_NAME'], args)  
16  
17 # read from a partition corresponding to a day in usa-traffic data in 2024  
18 S3DataSource_dsource1 = glueContext.create_dynamic_frame.from_options(  
19     format_options={},  
20     connection_type="s3",  
21     format="json",  
22     connection_options={"paths": ["s3://usa-traffic-dat/routes/2024/11/15"]},  
23     transformation_ctx="S3DataSource_dsource1",  
24 )  
25  
26 # filter the data to select only routes for highways  
Python Ln 1, Col 1 Errors: 0 | Warnings: 0
```

- **Hard-to-diagnose issues:** Spark out-of-memory and out-of-disk errors
- **Common issues:** Resource setup or configuration errors

NEW

PREVIEW IN AWS GLUE

Generative AI upgrades for Apache Spark in AWS Glue

AUTOMATE SPARK VERSION UPGRADES

Modernize and **upgrade** Spark applications faster with generative AI

Automate manual work to identify & update Spark code & configurations

Validate upgraded Spark applications automatically



Gen AI Spark upgrade workflow

The screenshot displays the AWS Glue Studio interface. The left sidebar shows navigation options for AWS Glue, including ETL jobs, Data Catalog, and Legacy pages. The main content area is titled 'AWS Glue Studio' and features three options for creating a job: 'Visual ETL' (Author in a visual interface focused on data flow), 'Notebook' (Author using an interactive code notebook), and 'Script editor' (Author code with a script editor). Below these options is a section for 'Example jobs' with a 'Create example job' button. The 'Your jobs (39)' section contains a table of existing jobs.

Job name	Type	Created by	Last modified	AWS Glue version
books-review-analytics	Glue ETL	Script	11/21/2024, 3:13:51 PM	2.0
book-reviews-daily-analytics-job	Glue ETL	Script	11/21/2024, 3:11:33 PM	2.0
Tests4S3ToGDCToS3-copy	Glue ETL	Script	11/21/2024, 3:10:40 PM	2.0
Tests4Rag	Glue ETL	Script	11/21/2024, 8:33:42 AM	2.0
Tests4S3ToGDCToS3	Glue ETL	Script	11/20/2024, 6:31:08 PM	2.0
Tests4GDCToS3ToGDC	Glue ETL	Script	11/20/2024, 5:05:47 PM	2.0
Tests2S3ToGDC	Glue ETL	Script	11/15/2024, 5:58:51 PM	2.0
Tests6S3ToS3	Glue ETL	Script	11/15/2024, 4:38:18 AM	2.0
Tests3S3ToS3	Glue ETL	Script	11/15/2024, 4:38:17 AM	2.0
Tests3GDCToS3ToGDC	Glue ETL	Script	11/15/2024, 4:38:16 AM	2.0

Submit Spark applications for **upgrade analysis**

Specify **target** Spark release

Review upgraded **code** and **configurations** and apply changes

Summary of core use cases



Drive developer productivity

Gen AI-driven catalog search & discovery, inline code recommendations for SQL & PySpark, and visual ETL flows with contextual analytics



Reduce time for operations with AWS expertise

Instant root cause analysis for complex errors, receive expert guidance drawn from years of AWS best practices



Retire tech debt at scale

Modernize projects and upgrade Spark applications with single-click AI-driven plan, validation, and analysis

Thank you!

Vipin Mohan

[linkedin.com/in/vipinmohan](https://www.linkedin.com/in/vipinmohan)

Mohit Saxena

[linkedin.com/in/mohitsax](https://www.linkedin.com/in/mohitsax)



Please complete the session survey in the mobile app