aws SUMMIT

SAN FRANCISCO | APRIL 20–21, 2022

ANA313

Interactive data integration with AWS Glue

Saroj Yadav
Software Development Manager, AWS Glue
AWS

Zach Mitchell
Sr. Big Data Architect, AWS Glue | AWS Lake Formation
AWS

What to expect

01

Interactive data integration

Why interactive data integration is important and hard to do

02

Key AWS Glue innovations

How AWS Glue Interactive Sessions and AWS Glue Studio Notebooks help give a better interactive experience

03

Build interactive applications

How to unlock new paradigms for data integration and Spark development

Why customers do data integration

Meaningful insights

Increased collaboration

Faster decisions

Break data silos

AWS Glue

FAST, OPEN, AND SCALABLE DATA INTEGRATION SERVICE

- Integrate data faster
- Scale with your growth
- No servers or capacity to manage

Understanding your data assets . . . interactively

EXPONENTIAL DATA GROWTH

DATA FROM NEW SOURCES

INCREASINGLY DIVERSE DATA

Data interactivity for different personas

NO OR LOW CODE

DATA ENGINEERS

SPARK DEVELOPERS

New kind of interactive data integration

REAL-TIME/
SLA SENSITIVE

COMPLEX

COST-SENSITIVE

Current interactive Spark development

Complex

Hard to get started

Lengthy

Time to first query takes 10–30 minutes

Wasted cost

Over-provisioning

No cost control mechanisms

Current interactive data integration process

Slow and lengthy

Poor feedback mechanism

Works only for certain users

aws Glue

[interactive_sessions: ⇩]:

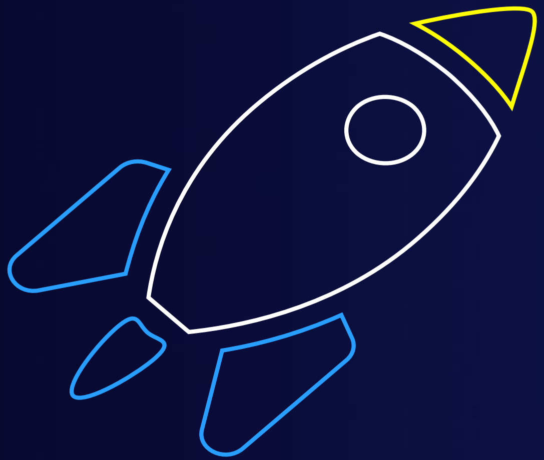
AWS Glue Interactive Sessions

Rapid Spark development

Pay as you go

Development tool of your choice

AWS Glue Interactive Sessions is rapid



Rapid creation of serverless Spark

Execute first code in seconds

Configure and install packages in seconds

Dedicated resources for every session

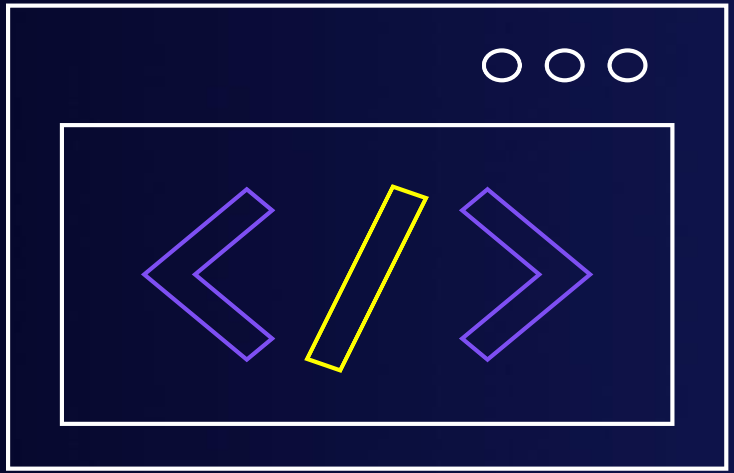
AWS Glue Interactive Sessions is frugal

Pay as you use

Inactive timeout & cleanup

Per second billing with 1-minute minimum

AWS Glue Interactive Sessions is easy to use

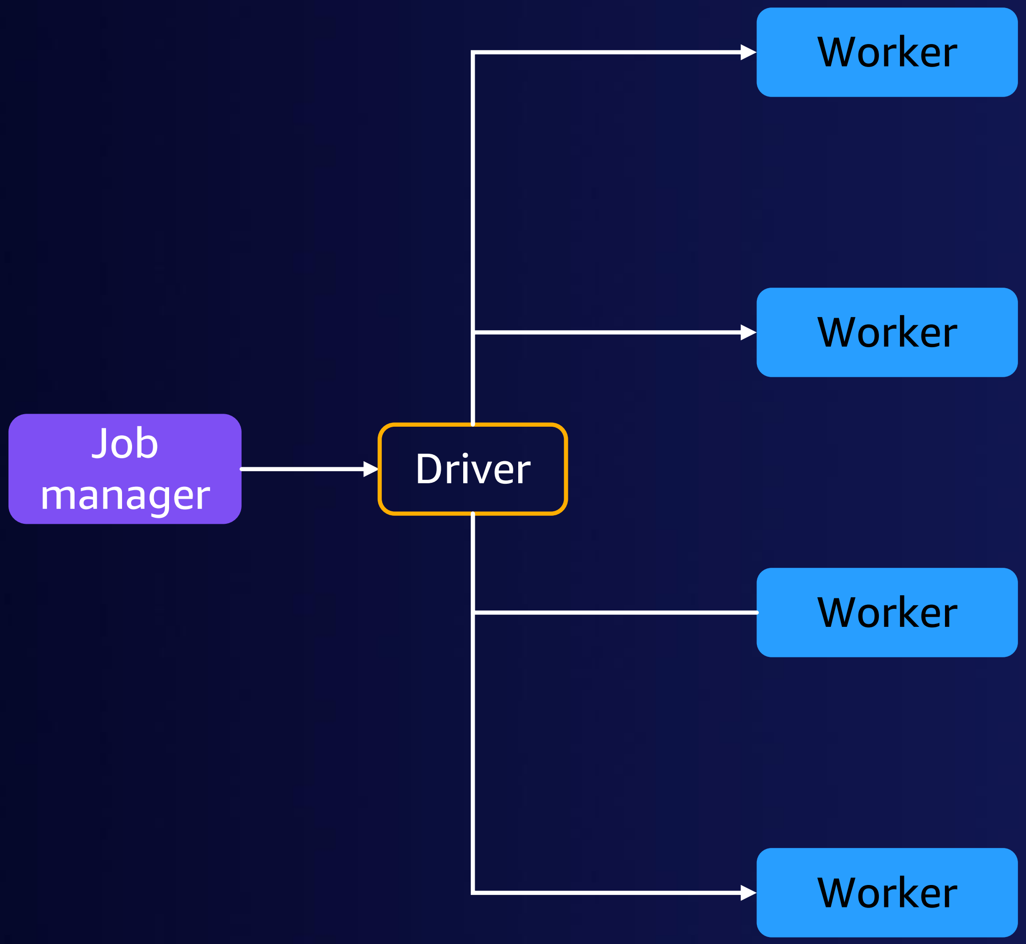


No complex setup needed

Live serverless data integration

Works with you favorite IDE

Same serverless execution engine as Glue jobs



Jobs are divided into stages

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

Job manager schedules task on worker

Glue’s execution scales to 1000s of worker

You pay for compute used

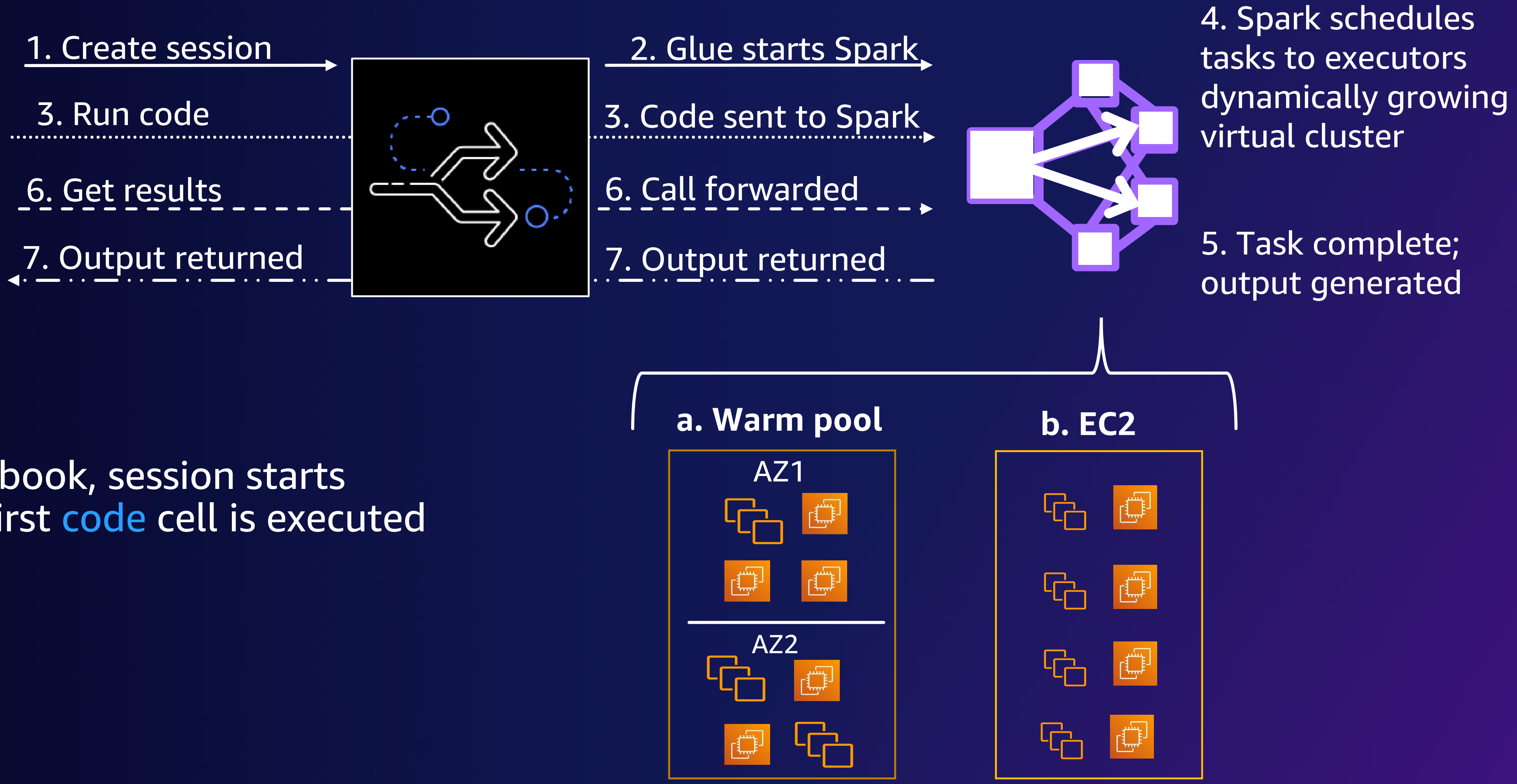
AWS Glue Interactive Sessions is RESTful

Integrate your own applications using APIs

aws Glue

`[interactive_sessions: ⇩ ]:`

Built on AWS Glue Execution Engine



1. Create session

3. Run code

6. Get results

7. Output returned

2. Glue starts Spark

3. Code sent to Spark

6. Call forwarded

7. Output returned

4. Spark schedules tasks to executors dynamically growing virtual cluster

5. Task complete; output generated

In notebook, session starts when first code cell is executed

AWS Glue Interactive Sessions API

`create_session()`	Start a session with your parameters
`run_statement()`	Submit Spark code to your session
`get_statement()`	Get returned output of submitted code
`stop_session()`	Stop the session

AWS Glue Interactive Sessions for Jupyter

Open-source Jupyter Kernel

Choose your favorite IDE

- Jupyter Notebooks & Lab
- Visual Studio Code
- IntelliJ | PyCharm | DataSpell

Configure with Magics

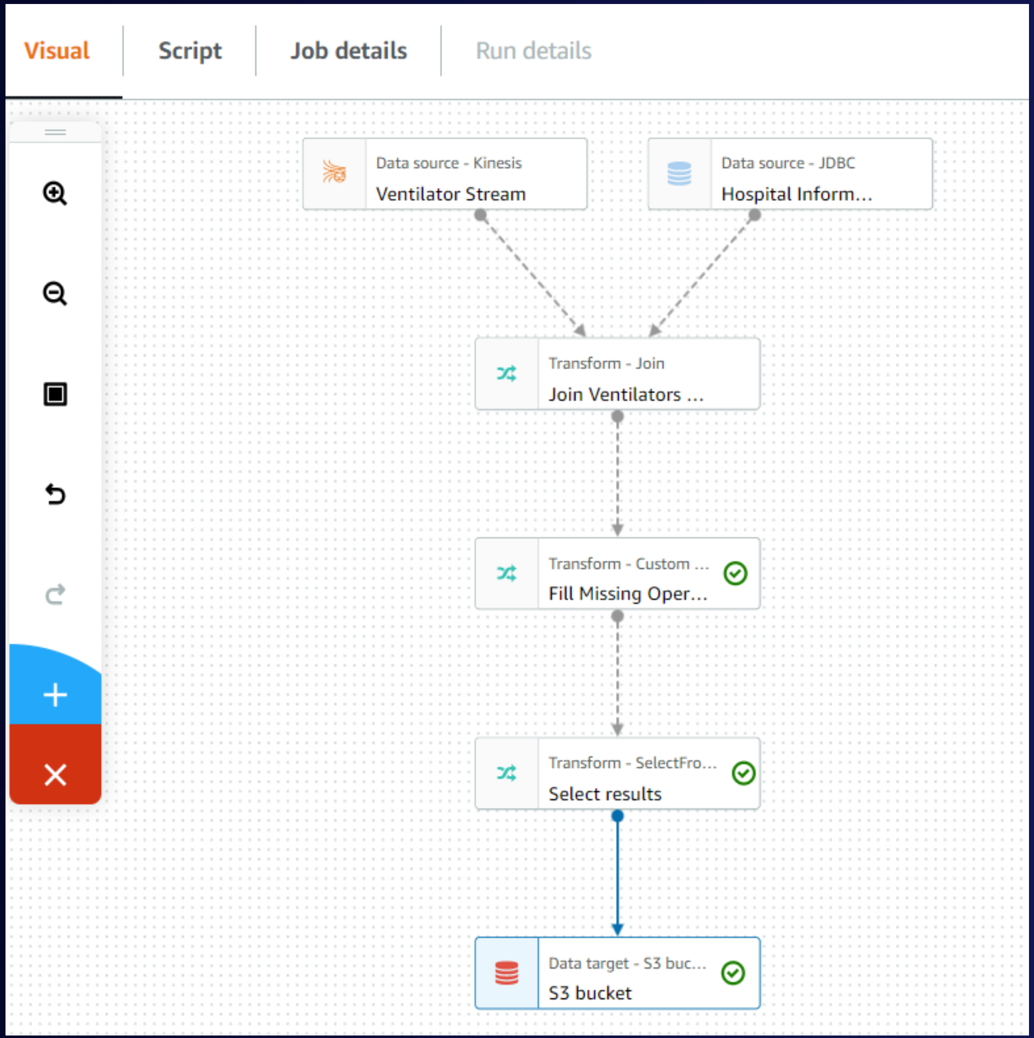
# Config stays with code

```
[ ]: %streaming
     %worker_type G.2X
     %connection my-vpc
     %%additional_python_modules awswrangler, s3://mybucket/mycustom.whl
     %%configure
       {
         "glue_version": "3.0",
         "number_of_workers": "20",
         "--conf": "spark.serializer=org.apache.spark.serializer.KryoSerializer”
        }
```

```
[ ]: dyf = glue_context.create_dynamic_frame.from_catalog()
```

AWS Glue Studio: Visual ETL interface

MAKES IT EASY TO AUTHOR, RUN, AND MONITOR AWS GLUE ETL JOBS



Author AWS Glue jobs visually without coding

Monitor 1000s of jobs through a single pane of glass

Distributed processing without the learning curve

Advanced transforms though code snippets

AWS Glue Studio benefits

Visual ETL in AWS Glue

Rapid development

Author ETL jobs faster and with less debugging

Monitor jobs in one place

See all your job runs through a single pane of glass

Streaming and batch ETL

Use one service to process data from all your sources

Perform complex transformations

Use visual transforms or write code snippets to clean and prepare your data for analysis

Process structured and semi-structured data

Use Glue studio to handle structured and semi-structured data like IoT logs

AWS Glue Studio Notebooks

Serverless

Free

1-click job creation
& scheduling

AWS Glue Interactive Sessions applications

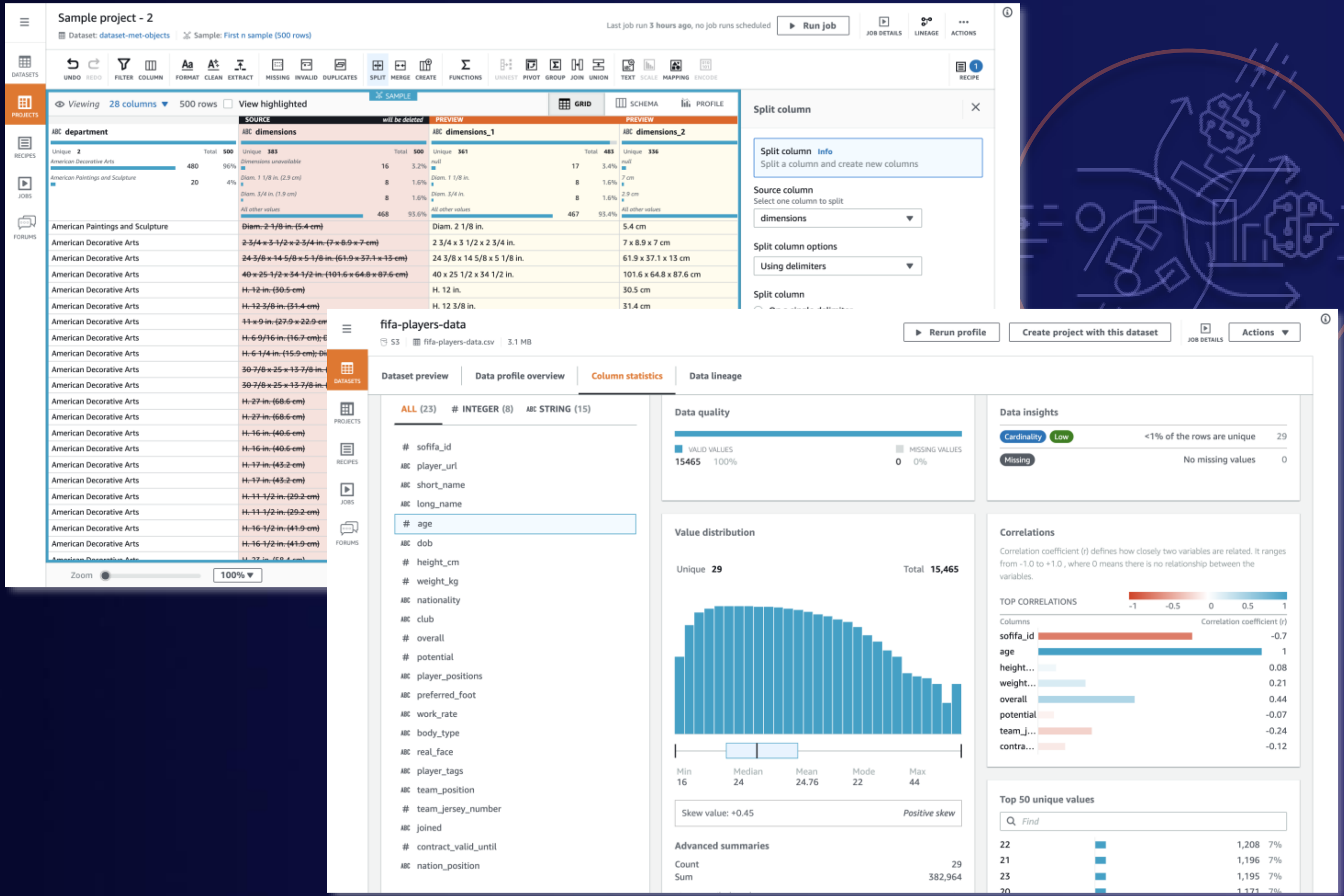
Direct API access to Glue Spark

Including Glue Streaming

Extend applications into Spark

AWS Glue DataBrew

POWERED BY AWS GLUE INTERACTIVE SESSIONS



Sample project - 2

Dataset: dataset-met-objects | Sample: First n sample (500 rows)

Last job run 3 hours ago, no job runs scheduled

Viewing 28 columns ▼ 500 rows □ View highlighted

	SOURCE will be deleted	PREVIEW	PREVIEW
ABC department	ABC dimensions	ABC dimensions_1	ABC dimensions_2
Unique 2 Total 500 American Decorative Arts 480 96% American Paintings and Sculpture 20 4%	Unique 383 Total 500 Dimensions unavailable 16 3.2% Diam. 1 1/8 in. (2.9 cm) 8 1.6% Diam. 3/4 in. (1.9 cm) 8 1.6% All other values 468 93.6%	Unique 361 Total 483 null 17 3.4% Diam. 1 1/8 in. 8 1.6% Diam. 3/4 in. 8 1.6% All other values 467 93.4%	Unique 336 null 7 cm 2.9 cm All other values
American Paintings and Sculpture	~~Diam. 2 1/8 in. (5.4 cm)~~	Diam. 2 1/8 in.	5.4 cm
American Decorative Arts	~~2 3/4 x 3 1/2 x 2 3/4 in. (7 x 8.9 x 7 cm)~~	2 3/4 x 3 1/2 x 2 3/4 in.	7 x 8.9 x 7 cm
American Decorative Arts	~~24 3/8 x 14 5/8 x 5 1/8 in. (61.9 x 37.1 x 13 cm)~~	24 3/8 x 14 5/8 x 5 1/8 in.	61.9 x 37.1 x 13 cm
American Decorative Arts	~~40 x 25 1/2 x 34 1/2 in. (101.6 x 64.8 x 87.6 cm)~~	40 x 25 1/2 x 34 1/2 in.	101.6 x 64.8 x 87.6 cm
American Decorative Arts	~~H. 12 in. (30.5 cm)~~	H. 12 in.	30.5 cm
American Decorative Arts	~~H. 12 3/8 in. (31.4 cm)~~	H. 12 3/8 in.	31.4 cm
American Decorative Arts	~~11 x 9 in. (27.9 x 22.9 cm~~		
American Decorative Arts	~~H. 6 9/16 in. (16.7 cm); D~~		
American Decorative Arts	~~H. 6 1/4 in. (15.9 cm); Di~~		
American Decorative Arts	~~30 7/8 x 25 x 13 7/8 in. (~~		
American Decorative Arts	~~30 7/8 x 25 x 13 7/8 in. (~~		
American Decorative Arts	~~H. 27 in. (68.6 cm)~~		
American Decorative Arts	~~H. 27 in. (68.6 cm)~~		
American Decorative Arts	~~H. 16 in. (40.6 cm)~~		
American Decorative Arts	~~H. 16 in. (40.6 cm)~~		
American Decorative Arts	~~H. 17 in. (43.2 cm)~~		
American Decorative Arts	~~H. 17 in. (43.2 cm)~~		
American Decorative Arts	~~H. 11 1/2 in. (29.2 cm)~~		
American Decorative Arts	~~H. 11 1/2 in. (29.2 cm)~~		
American Decorative Arts	~~H. 16 1/2 in. (41.9 cm)~~		
American Decorative Arts	~~H. 16 1/2 in. (41.9 cm)~~		
American Decorative Arts	~~H. 23 in. (58.4 cm)~~		

Split column

Split column Info

Split a column and create new columns

Source column

Select one column to split

dimensions ▼

Split column options

Using delimiters ▼

Split column

fifa-players-data

S3 | fifa-players-data.csv | 3.1 MB

Dataset preview | Data profile overview | Column statistics | Data lineage

ALL (23) # INTEGER (8) ABC STRING (15)

- # sofifa_id
- ABC player_url
- ABC short_name
- ABC long_name
- # age
- ABC dob
- # height_cm
- # weight_kg
- ABC nationality
- ABC club
- # overall
- # potential
- ABC player_positions
- ABC preferred_foot
- ABC work_rate
- ABC body_type
- ABC real_face
- ABC player_tags
- ABC team_position
- # team_jersey_number
- ABC joined
- ABC contract_valid_until
- ABC nation_position

Data quality

VALID VALUES 15465 100%

MISSING VALUES 0 0%

Value distribution

Unique 29 Total 15,465

Min	Median	Mean	Mode	Max
16	24	24.76	22	44

Skew value: +0.45 Positive skew

Advanced summaries

Count	29
Sum	382,964

Data insights

Cardinality Low	<1% of the rows are unique	29
Missing	No missing values	0

Correlations

Correlation coefficient (r) defines how closely two variables are related. It ranges from -1.0 to +1.0 , where 0 means there is no relationship between the variables.

TOP CORRELATIONS -1 -0.5 0 0.5 1

Columns	Correlation coefficient (r)
sofifa_id	-0.7
age	1
height...	0.08
weight...	0.21
overall	0.44
potential	-0.07
team_j...	-0.24
contra...	-0.12

Top 50 unique values

Find

22	1,208	7%
21	1,196	7%
23	1,195	7%
20	1,171	7%

Chose from 350+ transformations to prepare data

Identify dataset quality with 50+ statistics

Share, reuse, and schedule transforms across users and teams

Powered by Interactive Sessions

Summary

- Build faster, easier, and more cost-effectively
- End-to-end data integration
- Interact with Apache Spark using your favorite IDE
- Build interactive applications

aws Glue

[interactive_sessions: ⇩]:

What would you build?

Thank you!

Saroj Yadav

saroyada@amazon.com
https://www.linkedin.com/in/sarojyadav/

Zach Mitchell

mitczach@amazon.com
https://www.linkedin.com/in/zachary-mitchell-ab882853/