CommunityDays.it

C06 – serverless Analytics con Amazon Web Services

Francesco Diaz

Amazon Web Services

@francedit

The AWS analytics portfolio

DATA, VISUALIZATION, ENGAGEMENT, AND MACHINE LEARNING

- Data Exchange
- QuickSight
- Pinpoint
- SageMaker
- Comprehend
- Lex
- Polly
- Rekognition
- Translate

+ many more

ANALYTICS

- Redshift
- EMR (Spark and Hadoop)
- AWS Glue (Spark and Python)
- Athena
- Elasticsearch Service
- Kinesis Data Analytics

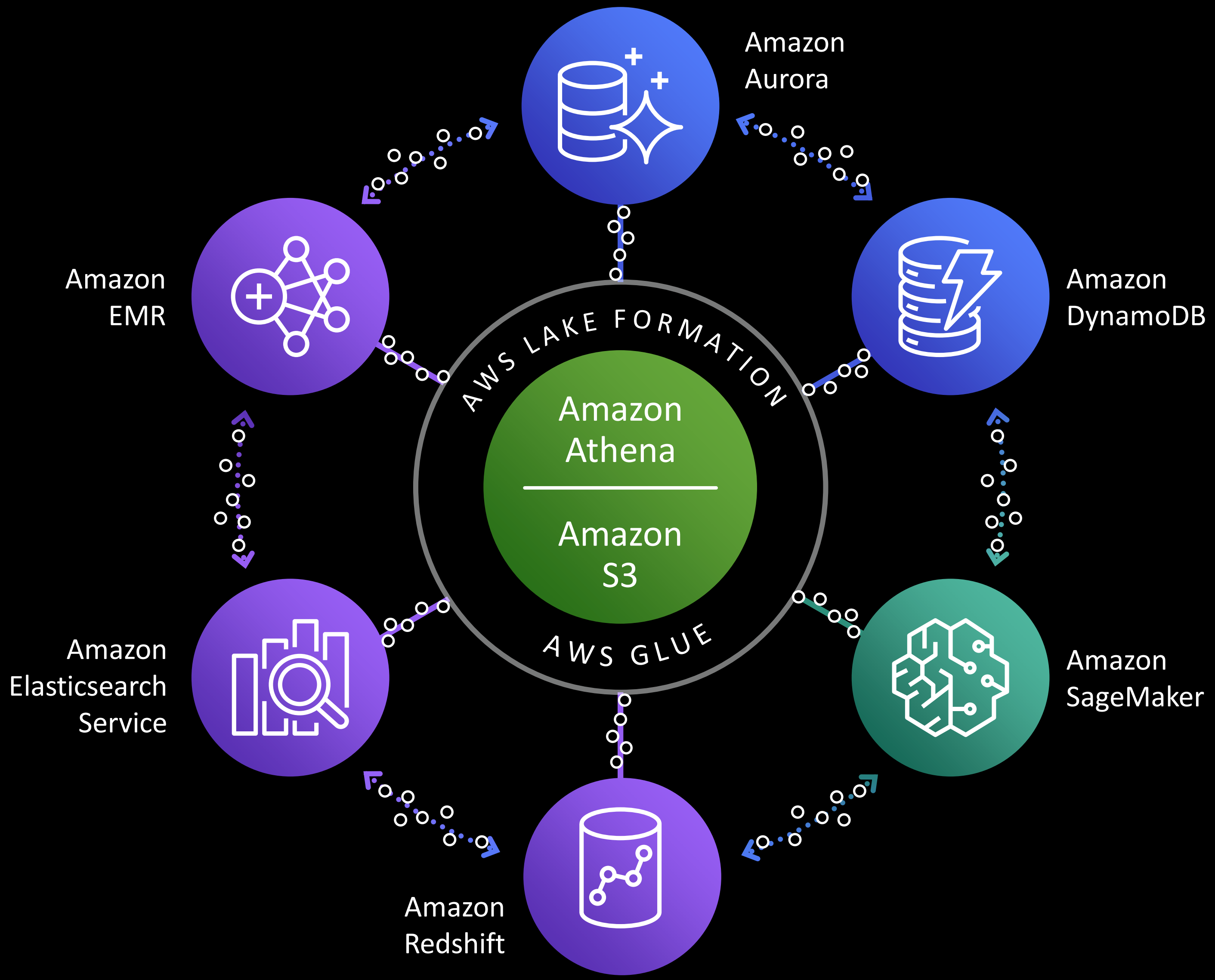
DATA LAKE INFRASTRUCTURE AND MANAGEMENT

- S3/Glacier
- Lake Formation
- AWS Glue

DATA MOVEMENT

Database Migration Service | Snowball | Snowmobile | Kinesis Data Streams | Kinesis Data Firehose | Managed Streaming for Apache Kafka

Lake House architecture on AWS



SCALABLE DATA LAKES

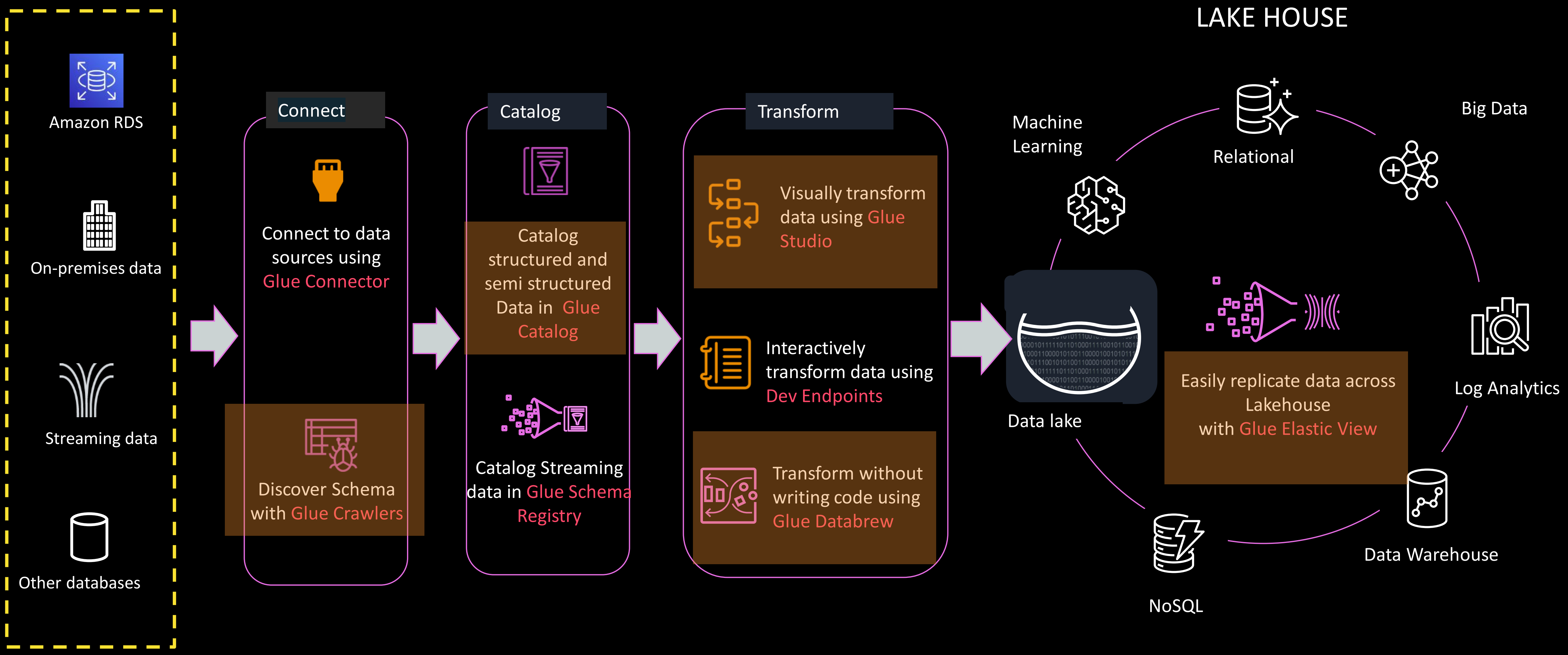
PURPOSE-BUILT DATA SERVICES

SEAMLESS DATA MOVEMENT

UNIFIED GOVERNANCE

PERFORMANT AND COST-EFFECTIVE

AWS Glue ecosystem



Amazon RDS

On-premises data

Streaming data

Other databases

Connect

Connect to data sources using Glue Connector

Discover Schema with Glue Crawlers

Catalog

Catalog structured and semi structured Data in Glue Catalog

Catalog Streaming data in Glue Schema Registry

Transform

Visually transform data using Glue Studio

Interactively transform data using Dev Endpoints

Transform without writing code using Glue Databrew

LAKE HOUSE

Data lake

Machine Learning

Relational

Big Data

Log Analytics

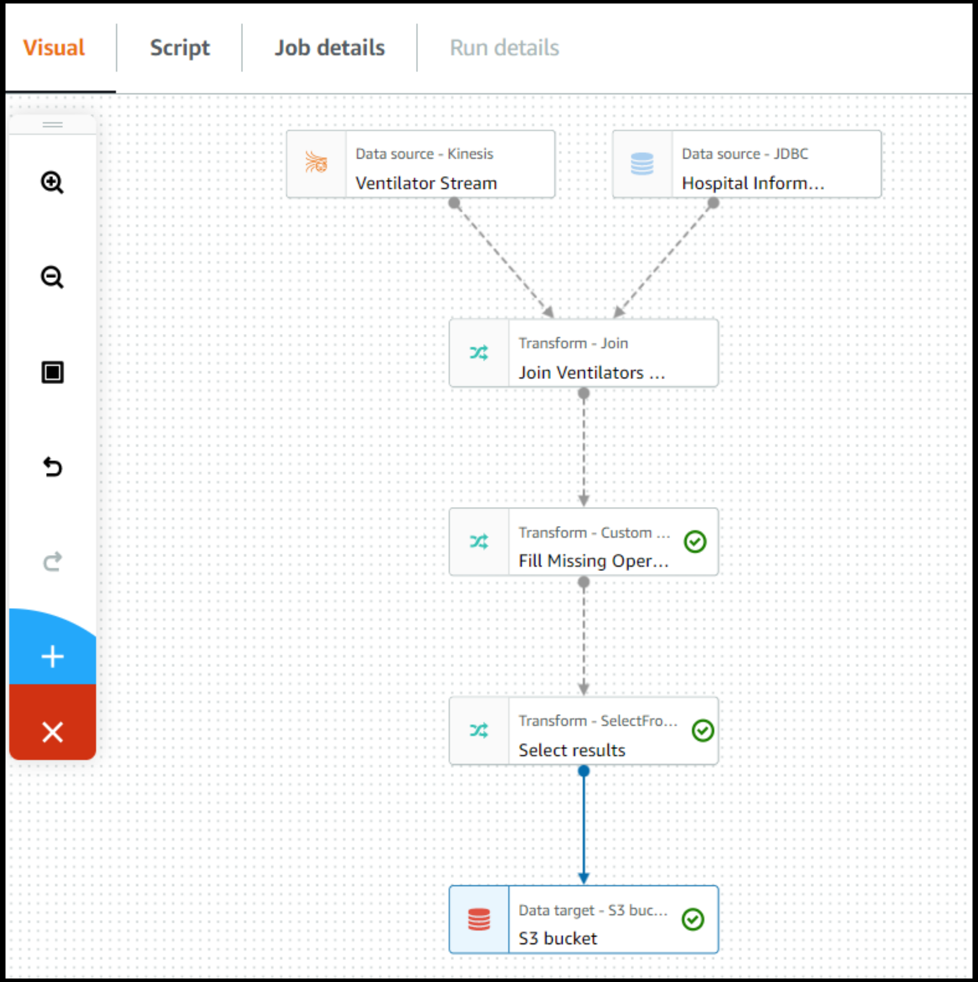
Data Warehouse

NoSQL

Easily replicate data across Lakehouse with Glue Elastic View

AWS Glue Studio

Easily author, run, and monitor AWS Glue ETL jobs



Author AWS Glue jobs visually without coding

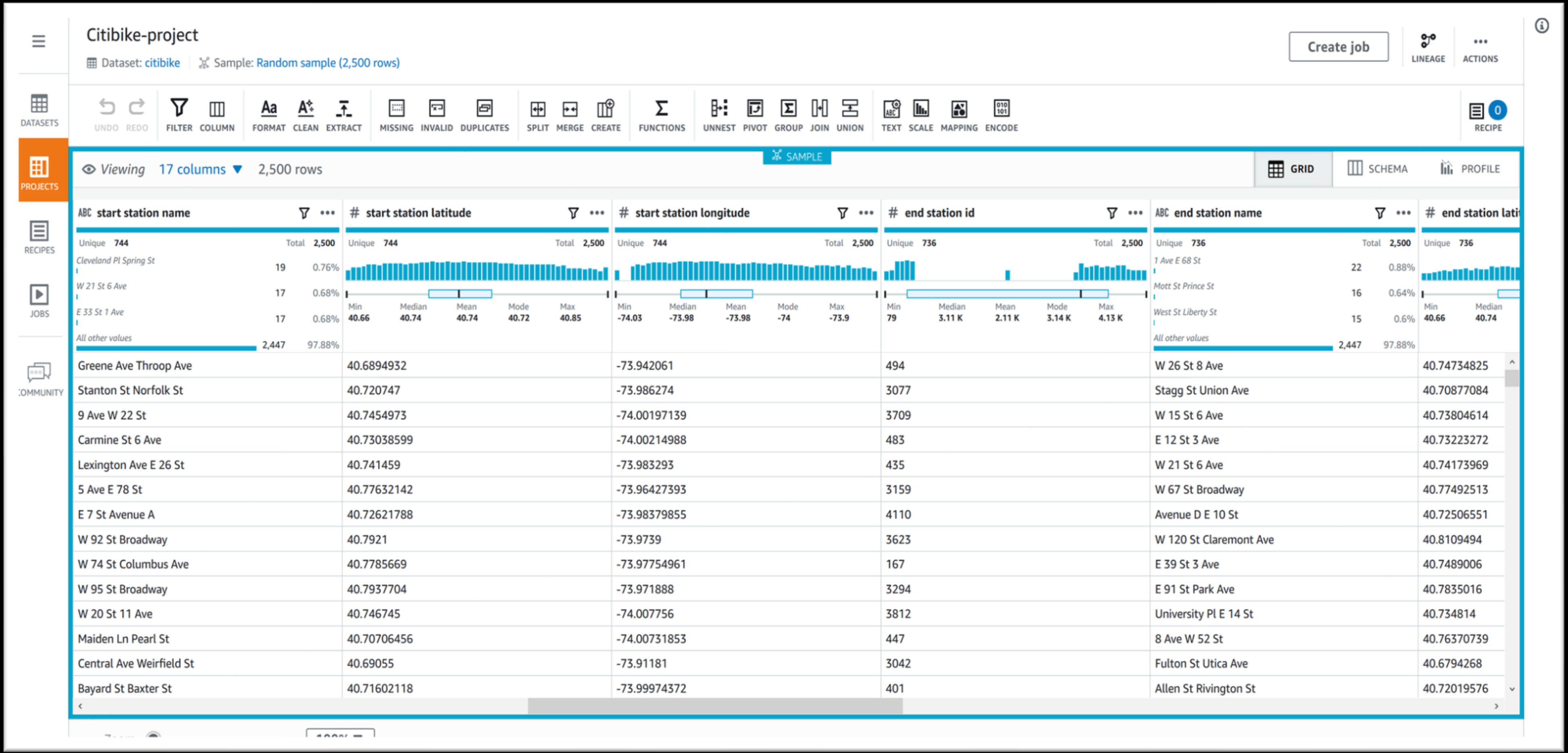
Monitor 1,000s of jobs through a single pane of glass

Distributed processing without the learning curve

Advanced transforms though code snippets

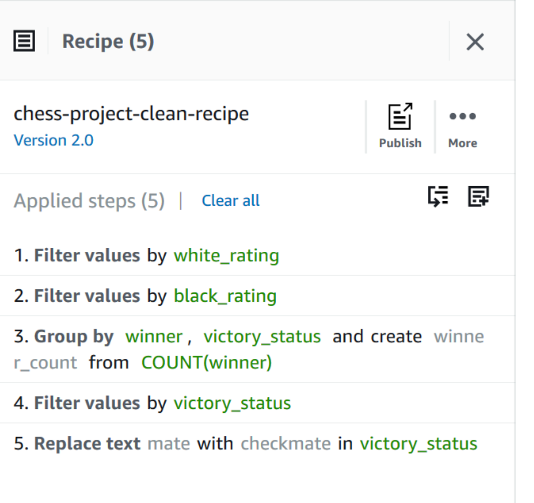
AWS Glue DataBrew – Visual Data Preparation

Built-in transformations



ABC start station name	# start station latitude	# start station longitude	# end station id	ABC end station name	# end station lati
Greene Ave Throop Ave	40.6894932	-73.942061	494	W 26 St 8 Ave	40.74734825
Stanton St Norfolk St	40.720747	-73.986274	3077	Stagg St Union Ave	40.70877084
9 Ave W 22 St	40.7454973	-74.00197139	3709	W 15 St 6 Ave	40.73804614
Carmine St 6 Ave	40.73038599	-74.00214988	483	E 12 St 3 Ave	40.73223272
Lexington Ave E 26 St	40.741459	-73.983293	435	W 21 St 6 Ave	40.74173969
5 Ave E 78 St	40.77632142	-73.96427393	3159	W 67 St Broadway	40.77492513
E 7 St Avenue A	40.72621788	-73.98379855	4110	Avenue D E 10 St	40.72506551
W 92 St Broadway	40.7921	-73.9739	3623	W 120 St Claremont Ave	40.8109494
W 74 St Columbus Ave	40.7785669	-73.97754961	167	E 39 St 3 Ave	40.7489006
W 95 St Broadway	40.7937704	-73.971888	3294	E 91 St Park Ave	40.7835016
W 20 St 11 Ave	40.746745	-74.007756	3812	University Pl E 14 St	40.734814
Maiden Ln Pearl St	40.70706456	-74.00731853	447	8 Ave W 52 St	40.76370739
Central Ave Weirfield St	40.69055	-73.91181	3042	Fulton St Utica Ave	40.6794268
Bayard St Baxter St	40.71602118	-73.99974372	401	Allen St Rivington St	40.72019576

Recipe steps



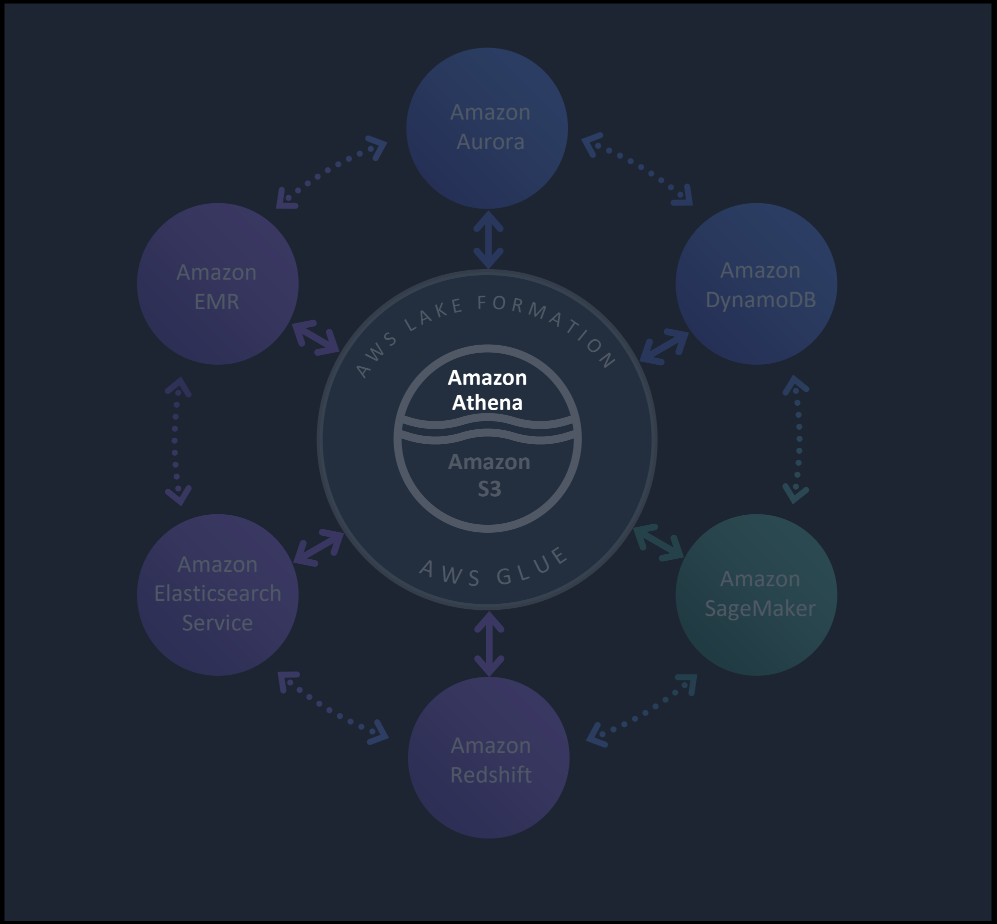
chess-project-clean-recipe
Version 2.0

Applied steps (5)

1. **Filter values** by white_rating
2. **Filter values** by black_rating
3. **Group by** winner , victory_status and create winner_count from COUNT(winner)
4. **Filter values** by victory_status
5. **Replace text** mate with checkmate in victory_status

Amazon Athena

Query data in S3 using SQL



Serverless

Quickly query S3 data without managing infrastructure, and pay only for the queries you run

Open and standard

Use ANSI SQL for querying with support for Parquet, CSV, JSON, Avro and other standard data formats

Fast interactive performance

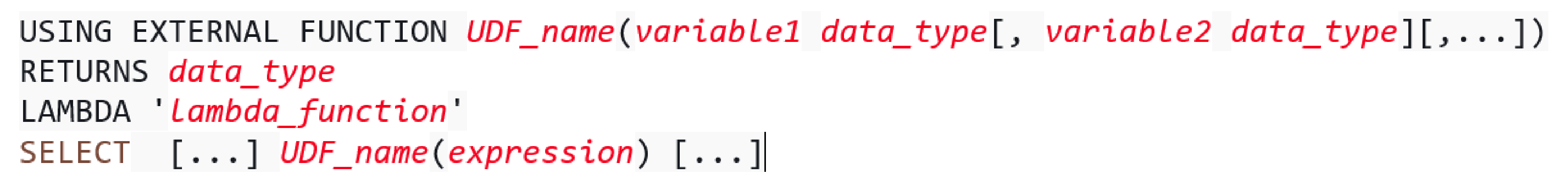
Parallel execution to deliver most results within seconds, with no cluster management required

Cost effective

Pay only for queries run and save 30–90% by compressing, partitioning, and converting your data into columnar formats

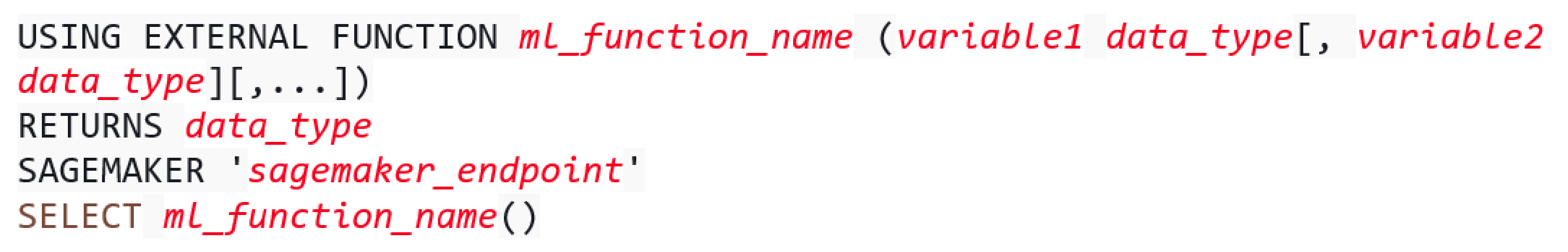
Amazon Athena

- User Defined Functions (UDF)



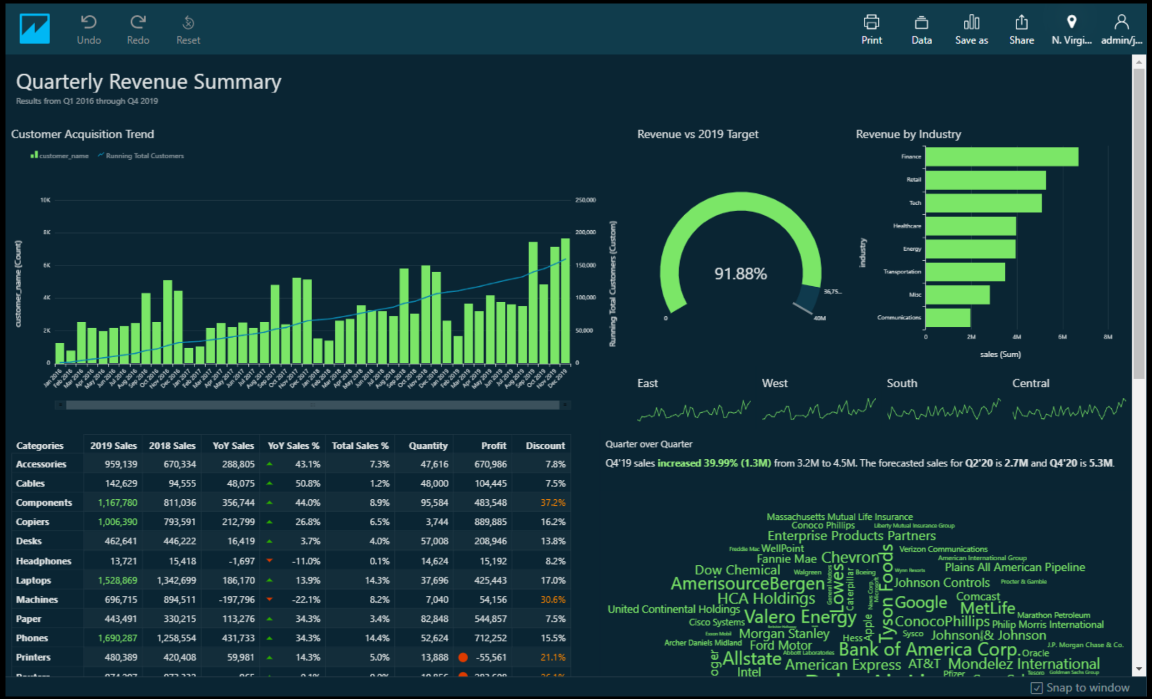
```
USING EXTERNAL FUNCTION UDF_name(variable1 data_type[, variable2 data_type][,...])
RETURNS data_type
LAMBDA 'lambda_function'
SELECT  [...] UDF_name(expression) [...]
```

- Amazon Athena Federated Query - support for running SQL queries across relational, non-relational, object, and custom data sources
    - data source connectors on AWS Lambda
- Using Machine Learning with Amazon Athena



```
USING EXTERNAL FUNCTION ml_function_name (variable1 data_type[, variable2
data_type][,...])
RETURNS data_type
SAGEMAKER 'sagemaker_endpoint'
SELECT ml_function_name()
```

Amazon QuickSight



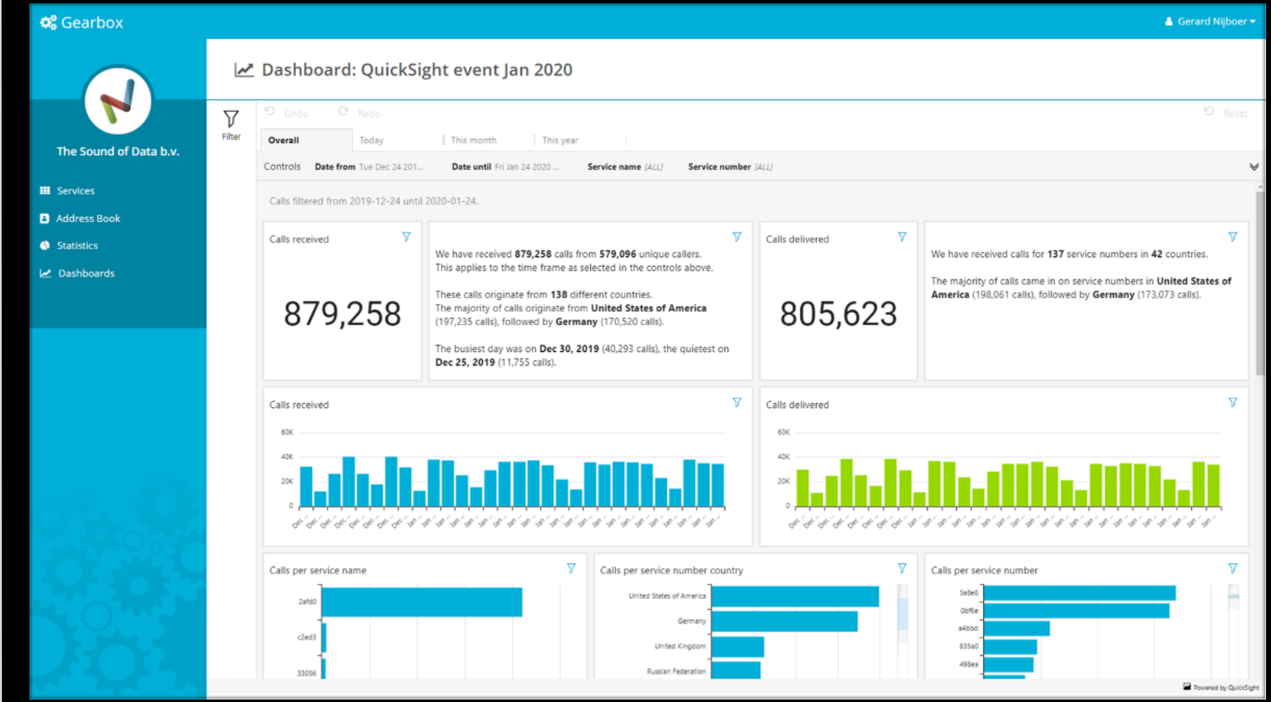
Quarterly Revenue Summary

Results from Q1 2016 through Q4 2019

Categories	2019 Sales	2018 Sales	YoY Sales	YoY Sales %	Total Sales %	Quantity	Profit	Discount
Accessories	959,139	670,334	288,805	▲ 43.1%	7.3%	47,616	670,986	7.8%
Cables	142,629	94,555	48,075	▲ 50.8%	1.2%	48,000	104,445	7.5%
Components	1,167,780	811,036	356,744	▲ 44.0%	8.9%	95,584	483,548	37.2%
Copiers	1,006,390	793,591	212,799	▲ 26.8%	6.5%	3,744	889,885	16.2%
Desks	462,641	446,222	16,419	▲ 3.7%	4.0%	57,008	208,946	13.8%
Headphones	13,721	15,418	-1,697	▼ -11.0%	0.1%	14,624	15,192	8.2%
Laptops	1,528,869	1,342,699	186,170	▲ 13.9%	14.3%	37,696	425,443	17.0%
Machines	696,715	894,511	-197,796	▼ -22.1%	8.2%	7,040	54,156	30.6%
Paper	443,491	330,215	113,276	▲ 34.3%	3.4%	82,848	544,857	7.5%
Phones	1,690,287	1,258,554	431,733	▲ 34.3%	14.4%	52,624	712,252	15.5%
Printers	480,389	420,408	59,981	▲ 14.3%	5.0%	13,888	● -55,561	21.1%

Quarter over Quarter

Q4'19 sales **increased 39.99% (1.3M)** from 3.2M to 4.5M. The forecasted sales for **Q2'20** is **2.7M** and **Q4'20** is **5.3M**.



Dashboard: QuickSight event Jan 2020

Controls **Date from** Tue Dec 24 201... **Date until** Fri Jan 24 2020 ... **Service name** *[ALL]* **Service number** *[ALL]*

Calls filtered from 2019-12-24 until 2020-01-24.

Calls received 879,258

We have received **879,258** calls from **579,096** unique callers.
This applies to the time frame as selected in the controls above.

These calls originate from **138** different countries.
The majority of calls originate from **United States of America** (197,235 calls), followed by **Germany** (170,520 calls).

The busiest day was on **Dec 30, 2019** (40,293 calls), the quietest on **Dec 25, 2019** (11,755 calls).

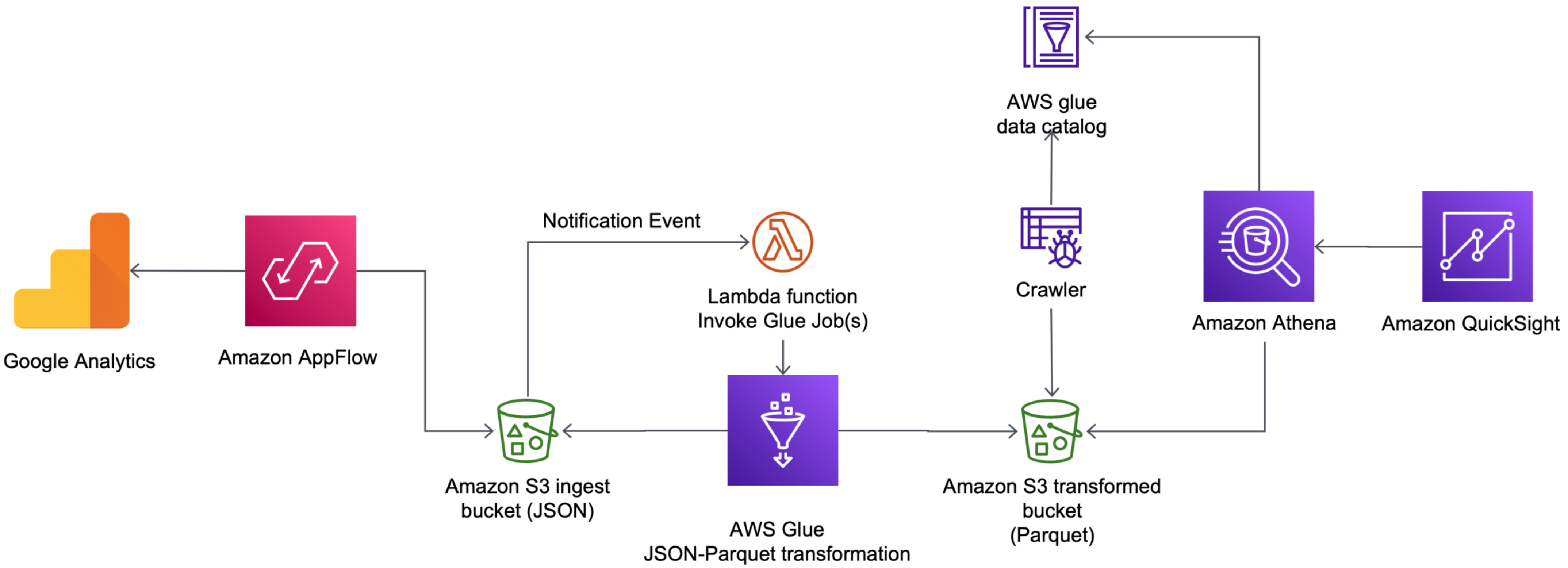
Calls delivered 805,623

We have received calls for **137** service numbers in **42** countries.

The majority of calls came in on service numbers in **United States of America** (198,061 calls), followed by **Germany** (173,073 calls).

Corporate & self-service BI

External BI / Embedded BI


AWS glue
data catalog
Notification Event
Google Analytics
Amazon AppFlow
Lambda function
Invoke Glue Job(s)
Crawler
Amazon Athena
Amazon QuickSight
Amazon S3 ingest
bucket (JSON)
AWS Glue
JSON-Parquet transformation
Amazon S3 transformed
bucket
(Parquet)

Demo

References

- Cloud Formation template and Workshop: https://appflow-immersionday.workshop.aws/en/googleanalytics.html
- Using Athena with JDBC: https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html
- Create a Glue crawler: https://docs.aws.amazon.com/glue/latest/ug/tutorial-add-crawler.html
- Getting started with Glue Studio: https://docs.aws.amazon.com/glue/latest/ug/tutorial-create-job.html
- Amazon QuickSight workshop: https://learnquicksight.workshop.aws/en/

Grazie!

- Il materiale sarà online nei prossimi giorni su http://www.communitydays.it