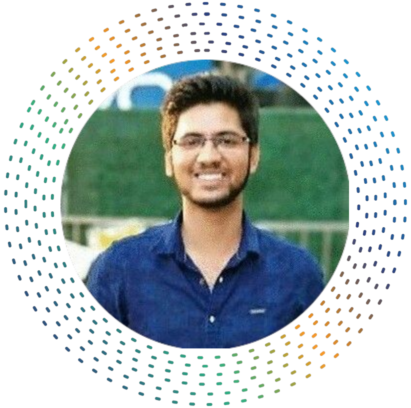
Data Engineering Workshop

Rajas Walavalkar & Sanchit Jain
8th Dec, Thursday
2:00 PM to 4:00 PM

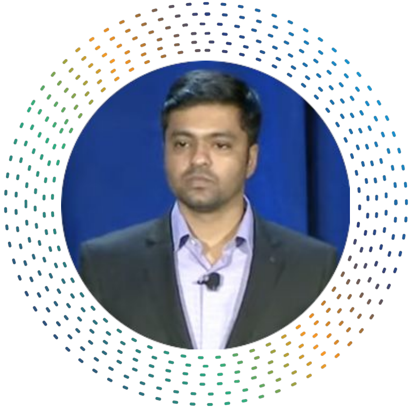
Speakers



Rajas Walavalkar

Solution Architect - AWS at Quantiphi
AWS Community Builder

FOLLOW ME

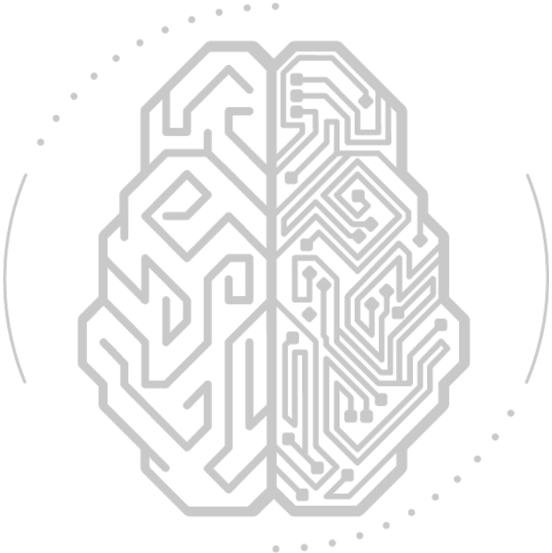


Sanchit Jain

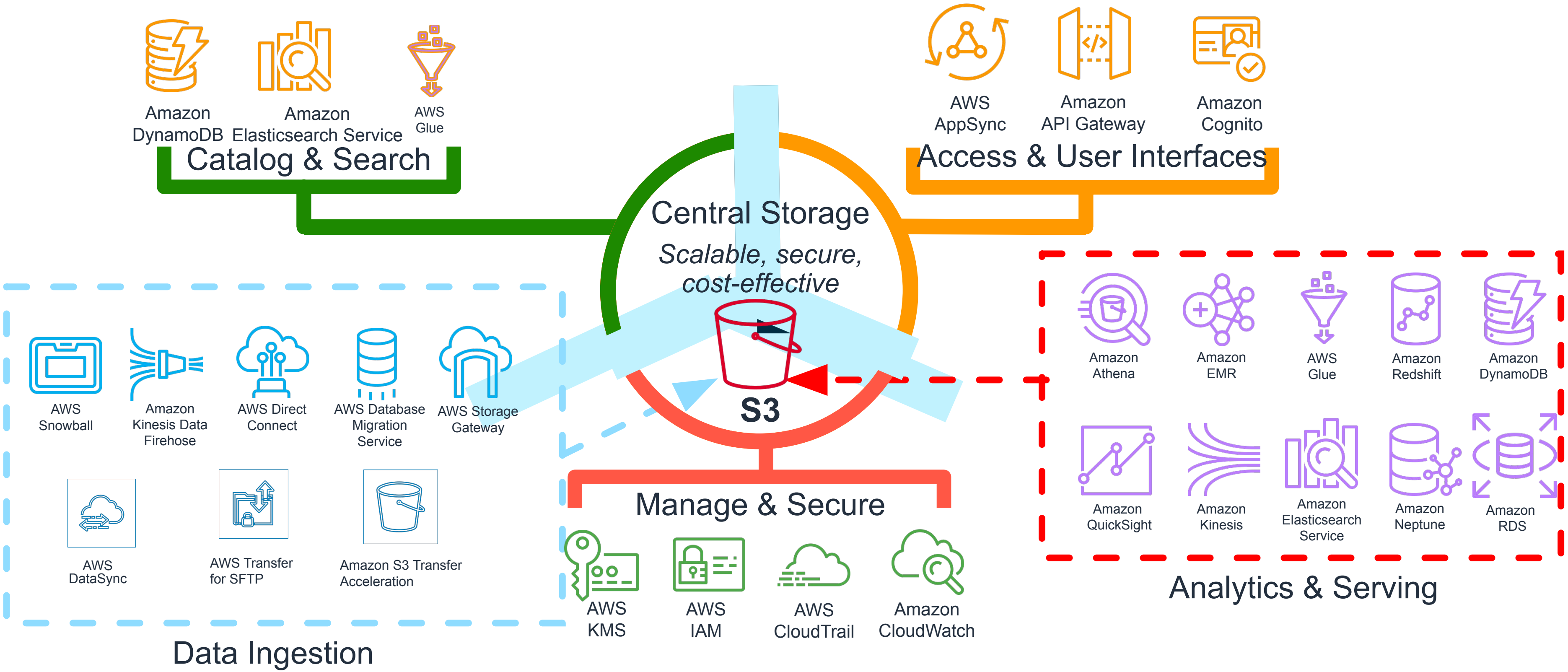
DNA & Cloud Practice Lead - AWS at Quantiphi
AWS Hero & AWS Ambassador

FOLLOW ME

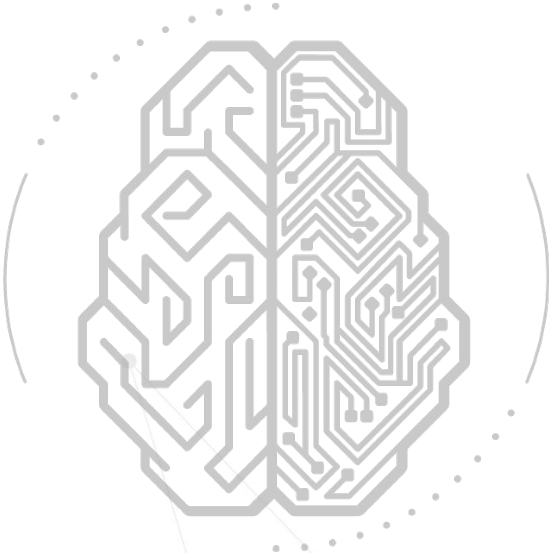
What we will learn?



Session's Focus – Query The Data Lake



AWS Simple Storage Service - S3



Amazon S3 Introduction

- Amazon Simple Storage Service (Amazon S3) is storage for the Internet. We can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web.
- Amazon S3 is one of the main building blocks of AWS - It's advertised as ”infinitely scaling” storage. Many websites use Amazon S3 as a backbone, and even many AWS services use Amazon S3 as an integration as well
- S3 Buckets
  - Buckets must have a globally unique name (across all regions all accounts)
  - Buckets are defined at the region level
  - S3 looks like a global service but buckets are created in a region
- S3 Objects
  - Objects (files) have a Key, and the key is the FULL path: s3://my-bucket/my_folder1/another_folder/my_file.txt
- S3 Durability:
  - High durability (99.999999999%, 11 9's) of objects across multiple AZ
  - If you store 10,000,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000 years
  - Same for all storage classes
- Availability
  - S3 standard has 99.99% availability, which means it will not be available 53 minutes a year
  - Varies depending on storage class

Amazon S3 Features

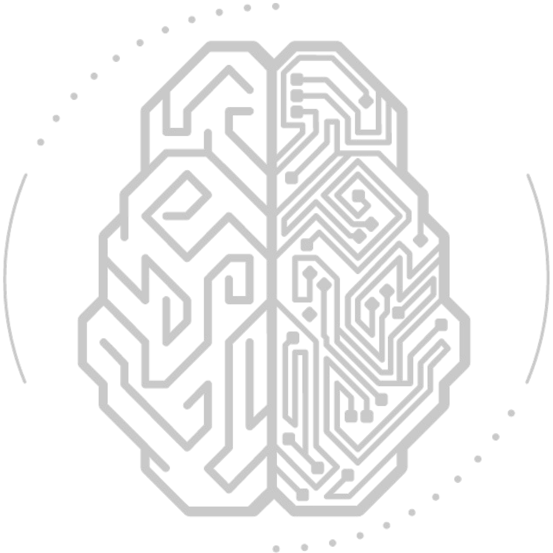
- Amazon S3 - Static Websites
  - S3 can host static websites and have them accessible on the www

- Amazon S3 - Versioning
  - You can version your files in Amazon S3
  - It is enabled at the bucket level
  - It is best practice to version your buckets
    - Protect against unintended deletes (ability to restore a version)
    - Easy roll back to previous version

- S3 Access Logs
  - Very helpful to come down to the root cause of an issue, or audit usage, view suspicious patterns, etc
  - Any request made to S3, from any account, authorized or denied, will be logged into another S3 bucket

- S3 Replication
  - Must enable versioning in source and destination, and Buckets can be in different accounts
  - Replication is Asynchronous in nature - Copying is asynchronous

AWS Glue - Features & Benefits



AWS Glue Overview

AWS Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development. AWS Glue provides all the capabilities needed for data integration so that you can start analyzing your data and putting it to use in minutes instead of months.

FEATURES

Serverless Spark

There is no infrastructure to maintain. Allocate needed compute power and run jobs. Job starts in few seconds and can run at petabyte scale

Data Integration for every user

Development environments catered to different skill sets

Cost Effective

All-in-one pricing model includes infrastructure and is 55% cheaper than other cloud data integration options

Handles complex workloads

Glue connects to 60+ data sources, processes petabytes of data in real-time, batch and event driven modes

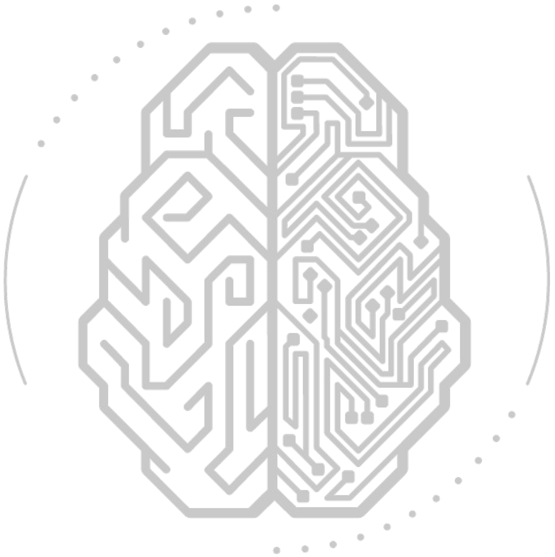
No Lock-In

Develop data integration pipelines in open source SparkSQL, PySpark and Scala

More Power

AWS Glue automates much of the effort spent in building, maintaining, and running ETL jobs

AWS Glue - Components



AWS Glue: Components

Data Catalog

- Hive metastore compatible with enhanced functionality
- Crawlers automatically extract metadata and create tables
- Integrated with Athena, Amazon Redshift Spectrum

Hive metastore compatible with enhancedfunctionality
Crawlers automatically extract metadata and create tables
Integrated with Athena, Amazon Redshift Spectrum

Job Authoring

- Auto-generates ETL code
- Builds on open frameworks—Python and Spark
- Developer-centric—editing, debugging, sharing

Auto-generates ETL code
Builds on open frameworks—Python and Spark
Developer-centric—editing, debugging, sharing

Job Execution

- Runs jobs on a serverless Spark platform
- Provides flexible scheduling
- Handles dependency resolution, monitoring, and alerting

Runs jobs on a serverless Spark platform
Provides flexible scheduling
Handles dependency resolution, monitoring, and alerting

AWS Glue Data Catalog

Manage table metadata through a Hive metastore API or Hive SQL. Supported by tools like Hive, Presto, Spark, etc. AWS added a few extensions:

- **Search** over metadata for data discovery
- **Connection info—**JDBC URLs, credentials
- **Classification** for identifying and parsing files
- **Versioning** of table metadata as schemas evolve and other metadata are updated

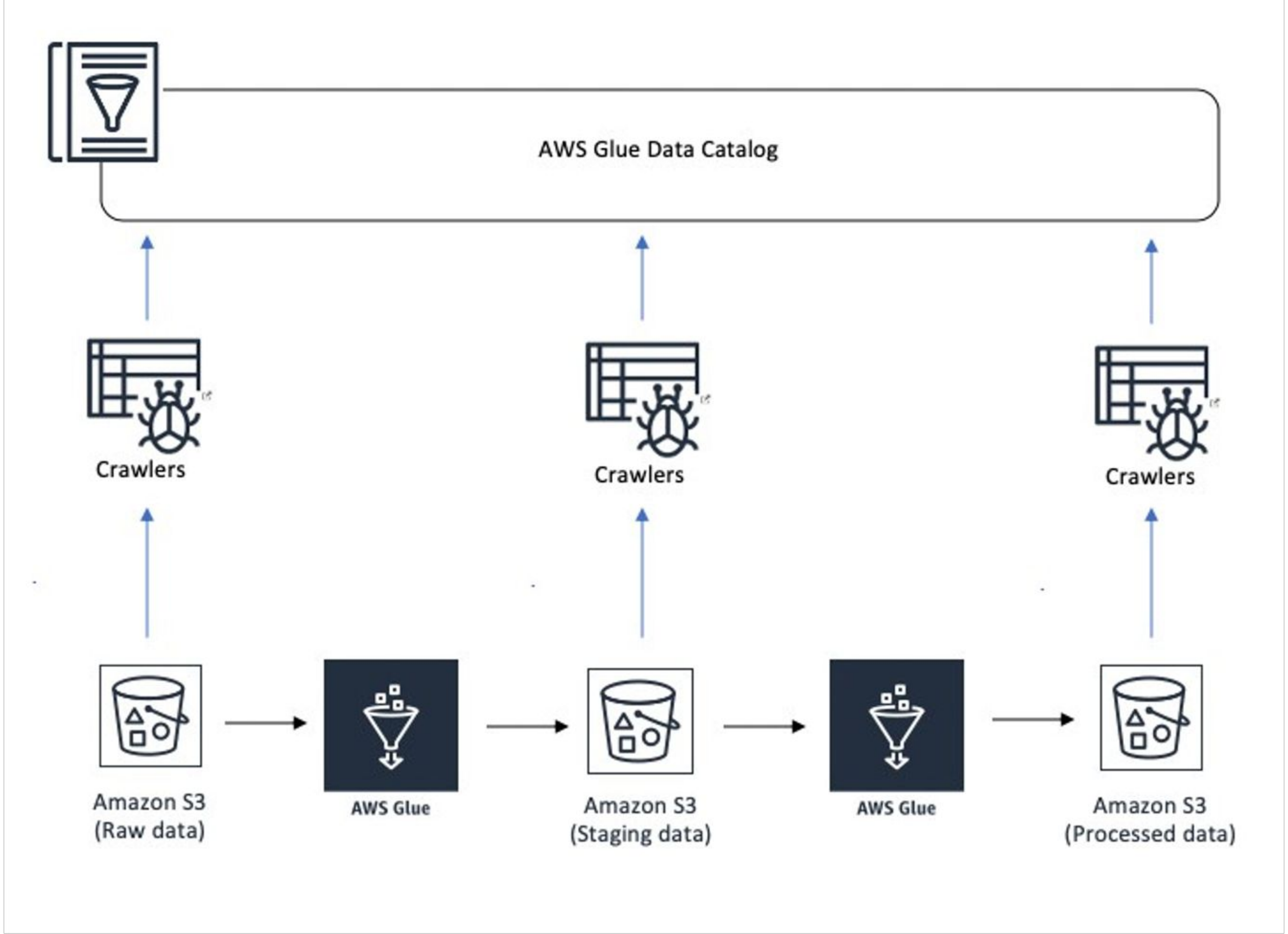
Populate using Hive DDL, bulk import, or automatically through **crawlers**

AWS Glue Data Catalog: Crawlers

Crawlers automatically build your Data Catalog and keep it in sync

- Automatically discover new data, extract schema definitions
  - Detect schema changes and version tables
  - Detect Hive style partitions on Amazon S3
- Built-in classifiers for popular types; custom classifiers using Grok expressions
- Run ad hoc or on a schedule; serverless—only pay when crawler runs

AWS Glue In Action



What is AWS Glue DataBrew ?

AWS Glue DataBrew is a visual data preparation tool that makes it easy for data analysts and data scientists to prepare data with an interactive, point-and-click visual interface without writing code.

CAPABILITIES OF GLUE DATABREW

- Profile
- Clean and Normalize
- Map Data Lineage
- Automate

NEED FOR DATABREW

"Upto 80% of data analysis time is spent on preparing data"

Time Consuming

- Multi-step process to extract, clean, normalize & load data at scale
- The right tools for the right persona must be integrated

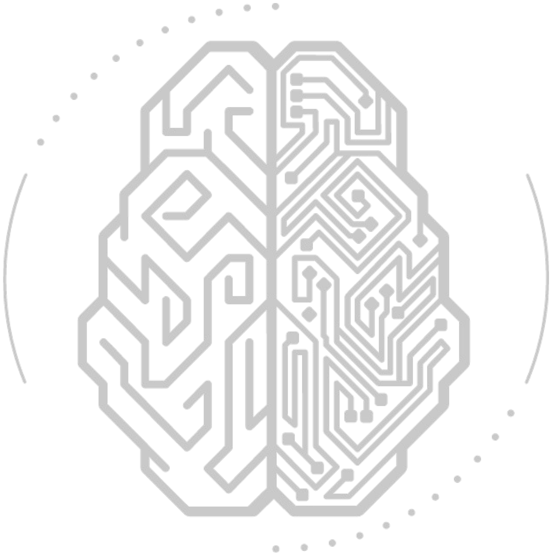
Expensive

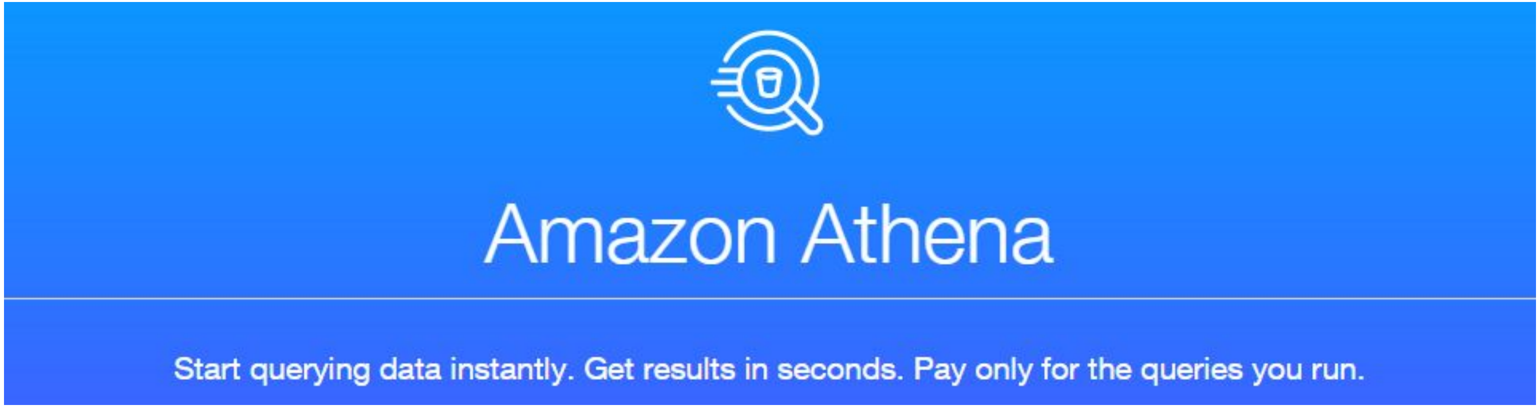
- Costly user licenses & siloed tools that cause rework
- Often requires moving large amount of data into silos

Manual

- Needs a lot of code-based heavy lifting to work at scale
- Hard to operationalize & build repeatable workflows

Querying the Data Lake with Amazon Athena





Amazon Athena

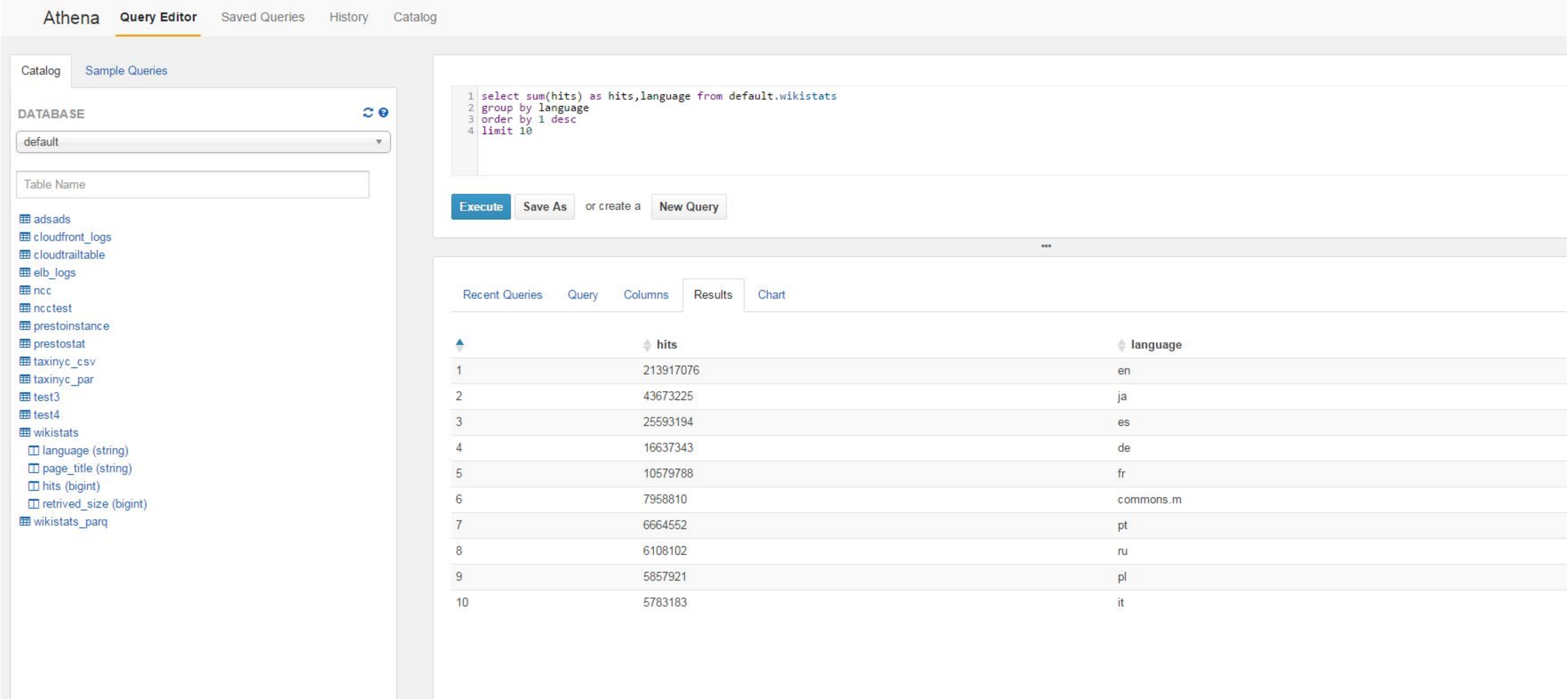
Start querying data instantly. Get results in seconds. Pay only for the queries you run.

An interactive query service that makes it easy to analyze data directly from Amazon S3 using Standard SQL

Amazon Athena

- Query data in your Amazon S3 based data lake
- Analyze infrastructure, operation, and application logs
- Interactive analytics using popular BI tools
- Self-service data exploration for data scientists
- Embed analytics capabilities into your applications

What does it look like?

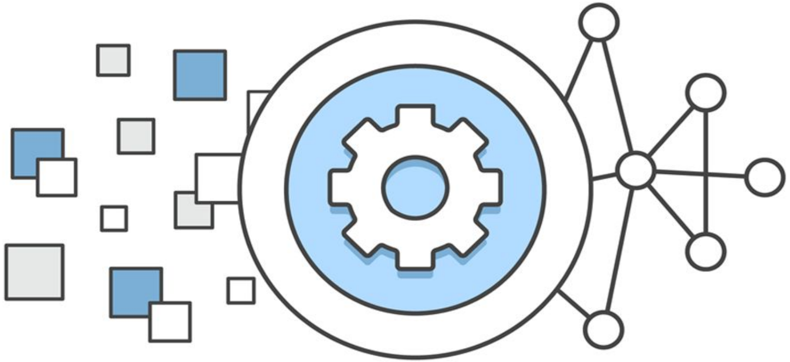


```sql
select sum(hits) as hits,language from default.wikistats
group by language
order by 1 desc
limit 10
```

	hits	language
1	213917076	en
2	43673225	ja
3	25593194	es
4	16637343	de
5	10579788	fr
6	7958810	commons.m
7	6664552	pt
8	6108102	ru
9	5857921	pl
10	5783183	it

Athena is Serverless

- No Infrastructure or administration
- Zero Spin up time
- Transparent upgrades



Familiar Technologies Under the Covers

Used for SQL Queries

In-memory distributed query engine

ANSI-SQL compatible with extensions

Used for DDL functionality

Complex data types

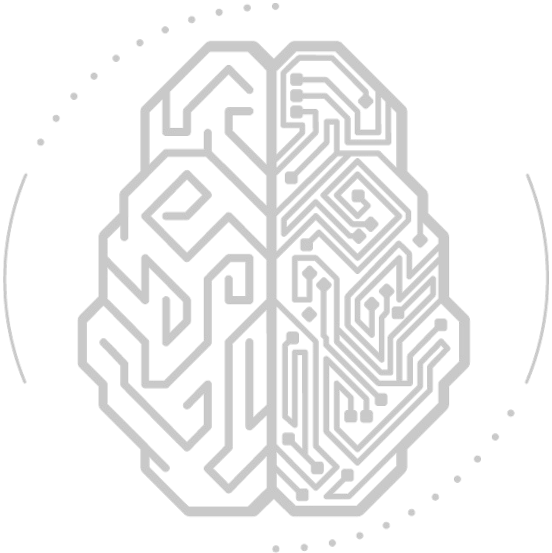
Multitude of formats

Supports data partitioning

Amazon Athena is Cost Effective

- Pay per query
- $5 per TB scanned from S3
- DDL Queries and failed queries are free
- Save by using compression, columnar formats, partitions

Athena Workgroups



Athena Workgroups

Athena Workgroups are used to isolate queries between different teams, workloads or applications, and to set on amount of data each query or the entire workgroup can process

Workload Isolation

Query Metrics

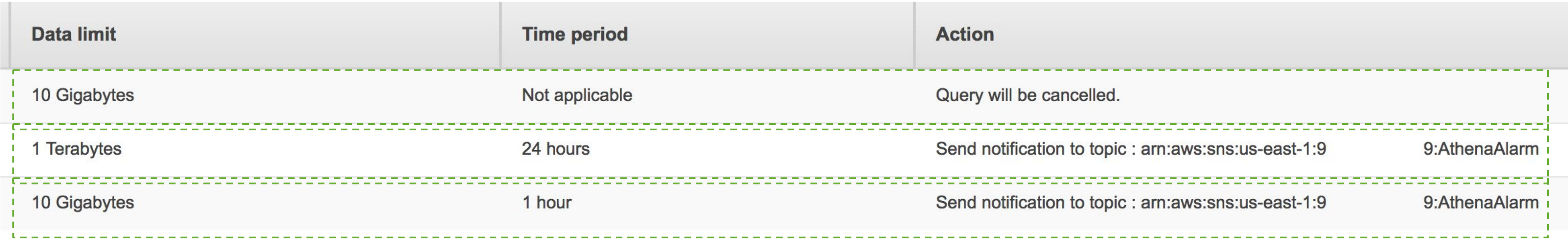
Cost Controls

Workgroups – Cost Controls

- Per query data scanned threshold; exceeding, will cancel query

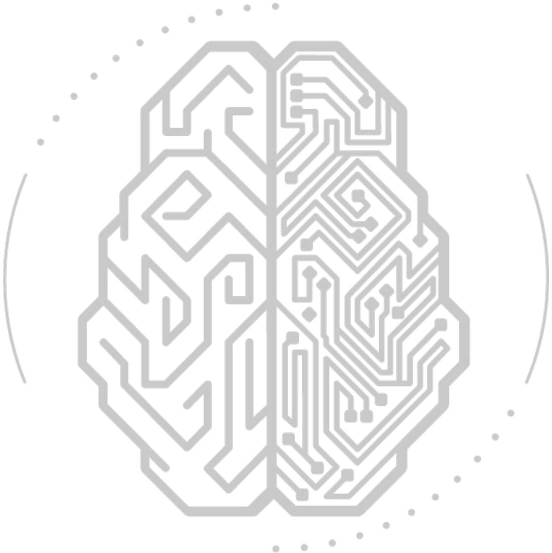
- Trigger alarms to notify of increasing usage and cost
- Disable Workgroup when all queries exceed a maximum threshold

Any Athena metric



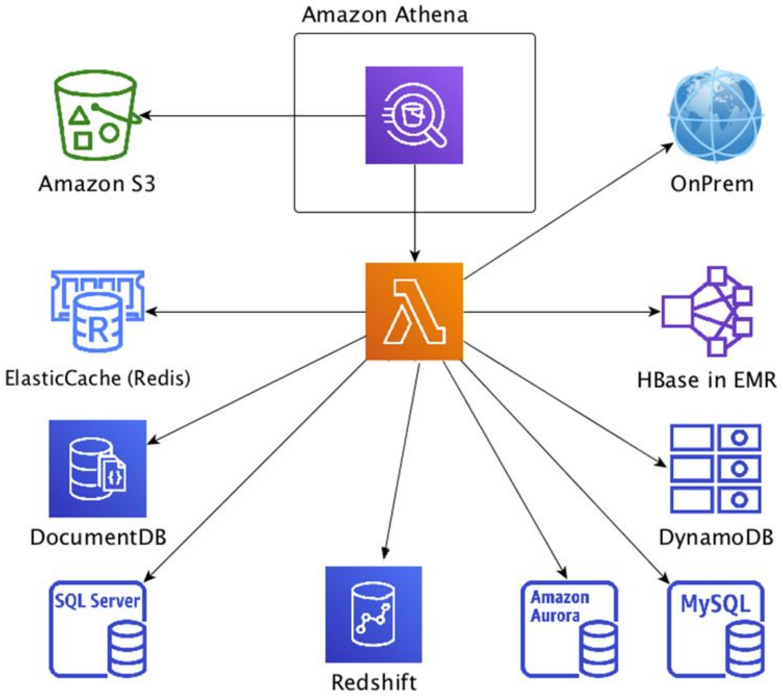
Data limit	Time period	Action
10 Gigabytes	Not applicable	Query will be cancelled.
1 Terabytes	24 hours	Send notification to topic : arn:aws:sns:us-east-1:9 9:AthenaAlarm
10 Gigabytes	1 hour	Send notification to topic : arn:aws:sns:us-east-1:9 9:AthenaAlarm

Athena federated query

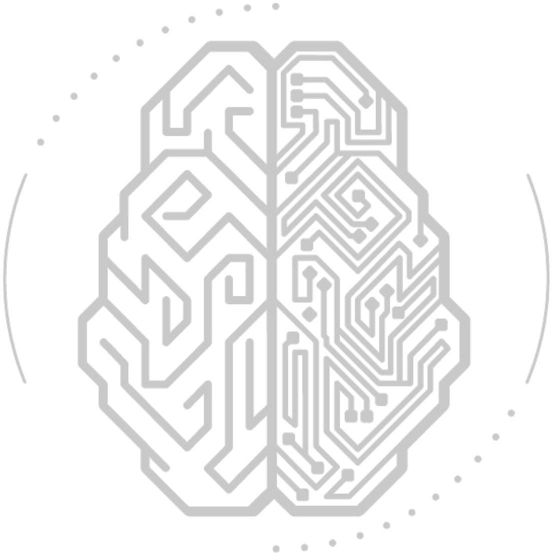


Athena federated query

- Run query across relational, non-relational, object, or custom data sources
- Run query across On-Premises or cloud data sources
- Can be used for ad-hoc investigations, or complex pipelines, or applications



Visualizing the Data Lake using Amazon QuickSight



Why Amazon QuickSight?

Cloud native = No servers = Auto-Scale

No servers or software to manage, maintain, deploy. Start with 10s of users and scale to 10s of 1000s

Fully integrated with AWS

Build end-to-end analytics in AWS. Secure private VPC access, fine-grained access control, ML integrations

Secure and global

End-to-end encryption. Native High Availability. 10 Global regions. HIPAA, PCI, ISO, SOC and FedRamp eligibility

Easy to develop and maintain

Design with Amazon QuickSight, integrate with APIs. Secure data with row-level security and authenticate seamlessly via single sign-on

Fast, consistent performance

Fast, predictable performance every time. Concurrent users or increased interactions do not slow down the system

ML insights

Contextual, relevant insights with ML-powered anomaly detection, forecasting, alerts and customizable narratives

Insights for everyone

Provide access to all users, pay only for usage. No upfront costs, no charges for inactive users

Customize and embed

Embed in applications and enable analytics in hours, not months or years. Use themes to match application/corporate branding

Connect to your data, wherever it is

QuickSight is natively integrated with AWS data sources, as well as on-premises and hosted databases and third party business applications

On-premises

Securely connect to on-premise databases and flat files like Excel and CSV

- Excel
- CSV
- Teradata
- MySQL
- SQL Server
- PostgreSQL

MySQL®

In the cloud

Connect to hosted database, big data formats, and secure VPCs

- Presto
- Spark
- SQL Server
- Postgre SQL
- MariaDB
- Snowflake
- IoT Analytics

aws

Applications

Connect directly to third party business applications

- Salesforce
- Square
- Adobe Analytics
- Jira
- ServiceNow
- Twitter
- Github

salesforce

Adobe Analytics

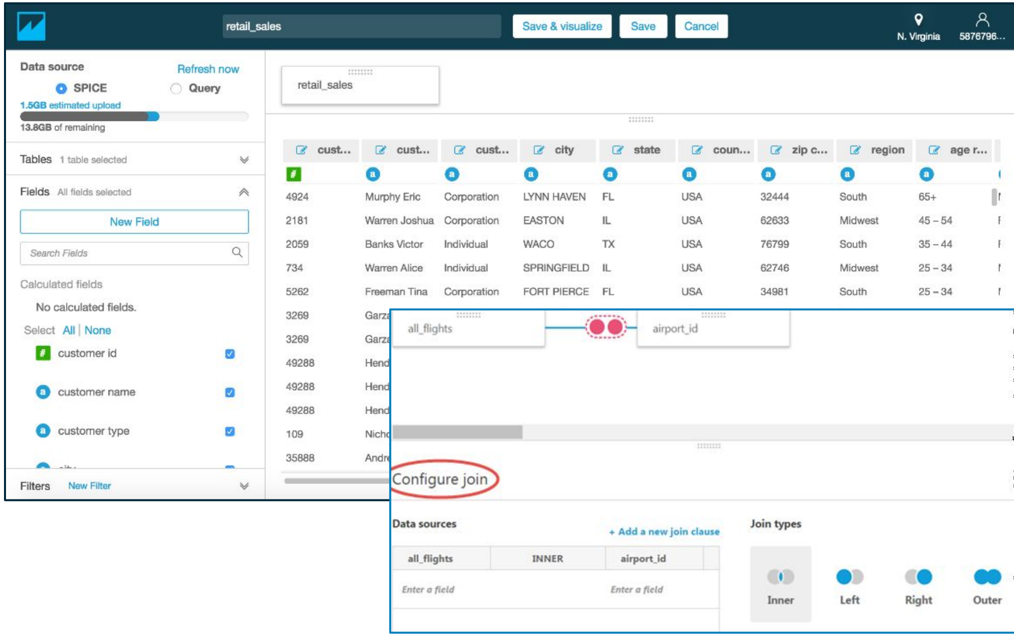
Square

servicenow

Data Prep

Optional step when creating data sets:

- Preview data
- Rename, remove fields, change data types
- Create new calculated fields
- Filter rows
- Issue direct query or ingest to SPICE
- Push down custom SQL queries
- Join across all data sources supported by QuickSight including file-to-file,file-to-database, and database-to-database joins



Data source
● SPICE ○ Query
1.5GB estimated upload
13.8GB of remaining

Tables 1 table selected

Fields All fields selected

Calculated fields
No calculated fields.

customer id ☑
customer name ☑
customer type ☑

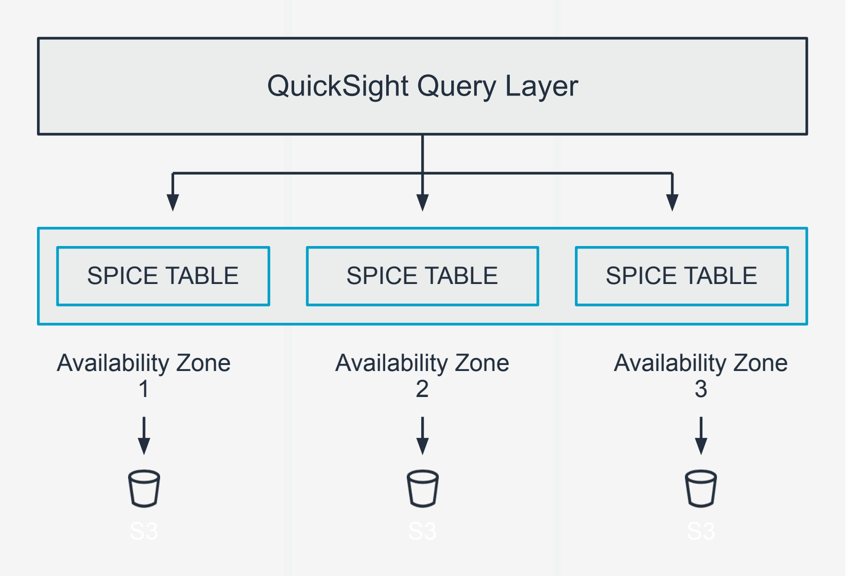
cust...	cust...	cust...	city	state	coun...	zip c...	region	age r...
4924	Murphy Eric	Corporation	LYNN HAVEN	FL	USA	32444	South	65+
2181	Warren Joshua	Corporation	EASTON	IL	USA	62633	Midwest	45 – 54
2059	Banks Victor	Individual	WACO	TX	USA	76799	South	35 – 44
734	Warren Alice	Individual	SPRINGFIELD	IL	USA	62746	Midwest	25 – 34
5262	Freeman Tina	Corporation	FORT PIERCE	FL	USA	34981	South	25 – 34
3269	Garza							
3269	Garza							
49288	Hend							
49288	Hend							
49288	Hend							
109	Nicho							
35888	Andre							

Data sources

all_flights	INNER	airport_id
Enter a field		Enter a field

SPICE

QuickSight is powered by SPICE, a super-fast calculation engine that delivers performance and scale, regardless of how many users are active.



Up to 10X faster (millisecond latency)

Fault-tolerant, self-healing

Support for high concurrency

Backed up in S3 (Write Ahead Log)

Instant failover with zero impact

User Types / User Roles

Admin

Manage Users

Manage SPICE Capacity

Manage VPC Connections

Manage Account Settings

QS Admin

Sometimes separate from Business Users, sometimes the same

Usually has AWS Console

Author

Create Data Sets

Create Analyses

Create Dashboards

Analyst

Sometimes in IT, sometimes Business Users

‘Data Analyst’

‘Data Engineer’

‘BI Engineer’

Reader

Consume Dashboards

Business User

Anyone

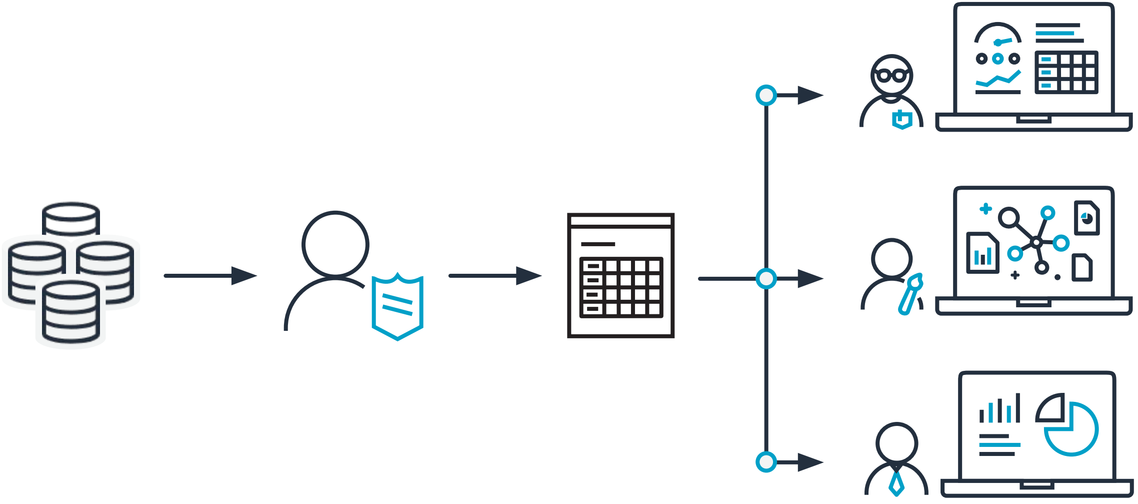
Can be internal or external users (customers/partners3rd

Data governance

Create managed datasets that give power users and authors the flexibility to perform self-serve analytics on data that you control.

Create datasets that:

- Can be shared with any user
- Automatically refresh
- Have row level security
- Users cannot modify
- Dynamically update with changes



Differentiate with natural language and ML

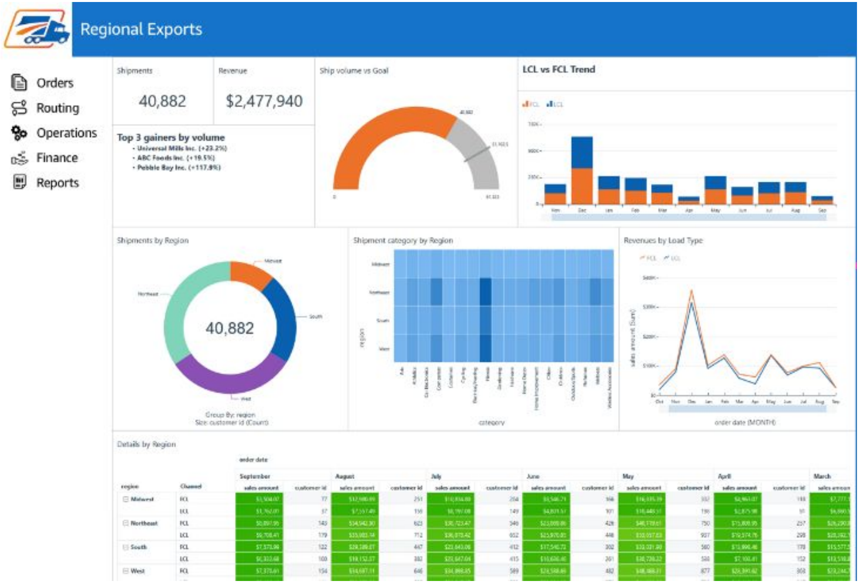
Auto narratives
Summarize your business metrics in plain language

Forecasting
Machine learning forecasting with point and click simplicity

Anomaly detection
Discover unexpected trends and outliers against millions of business metrics

ML predictions
Visualize and build predictive dashboards with Amazon SageMaker models

Embed Amazon QuickSight Dashboards



Shipments

40,882

Revenue

$2,477,940

Top 3 gainers by volume

- Universal Mills Inc. (+23.2%)
- ABC Foods Inc. (+19.5%)
- Pebble Bay Inc. (+117.8%)

Details by Region

		order date												
		September		August		July		June		May		April		March
region	Channel	sales amount	customer id	sales amount	customer id	sales amount	customer id	sales amount	customer id	sales amount	customer id	sales amount	customer id	sales amoun
⊟ Midwest	FCL	$3,504.07	77	$12,980.89	251	$10,834.88	254	$9,546.71	166	$16,035.39	332	$4,961.07	118	$7,777.
	LCL	$1,762.01	37	$7,557.49	156	$8,197.08	149	$4,801.51	101	$10,448.51	198	$2,875.98	91	$6,880.
⊟ Northeast	FCL	$5,897.95	143	$34,542.30	623	$30,723.47	346	$23,008.86	428	$40,179.61	750	$15,808.95	257	$26,290.
	LCL	$9,708.41	179	$35,983.14	712	$36,073.42	652	$25,970.05	448	$33,037.83	937	$19,374.76	298	$20,382.
⊟ South	FCL	$7,373.98	122	$28,289.07	447	$22,843.08	412	$17,540.72	302	$33,031.80	560	$13,996.48	178	$15,577.
	LCL	$6,333.68	100	$19,152.57	382	$23,647.04	415	$10,636.40	261	$38,728.22	586	$7,100.41	152	$13,518.
⊟ West	FCL	$7,375.61	154	$54,687.11	646	$34,888.85	589	$29,598.69	482	$48,488.31	877	$28,391.62	868	$23,244.

Fully interactive with drill down, filtering, & external links

Personalized views with row-level security

No servers to manage, no long-term commitments

Pay for usage with pay-per-session reader pricing

Seamless authentication

Amazon QuickSight Examples

THANK YOU