



Working With the Data Lake

Subtitle

Team or presenters name

Date



Session's Focus – Query the The Data Lake

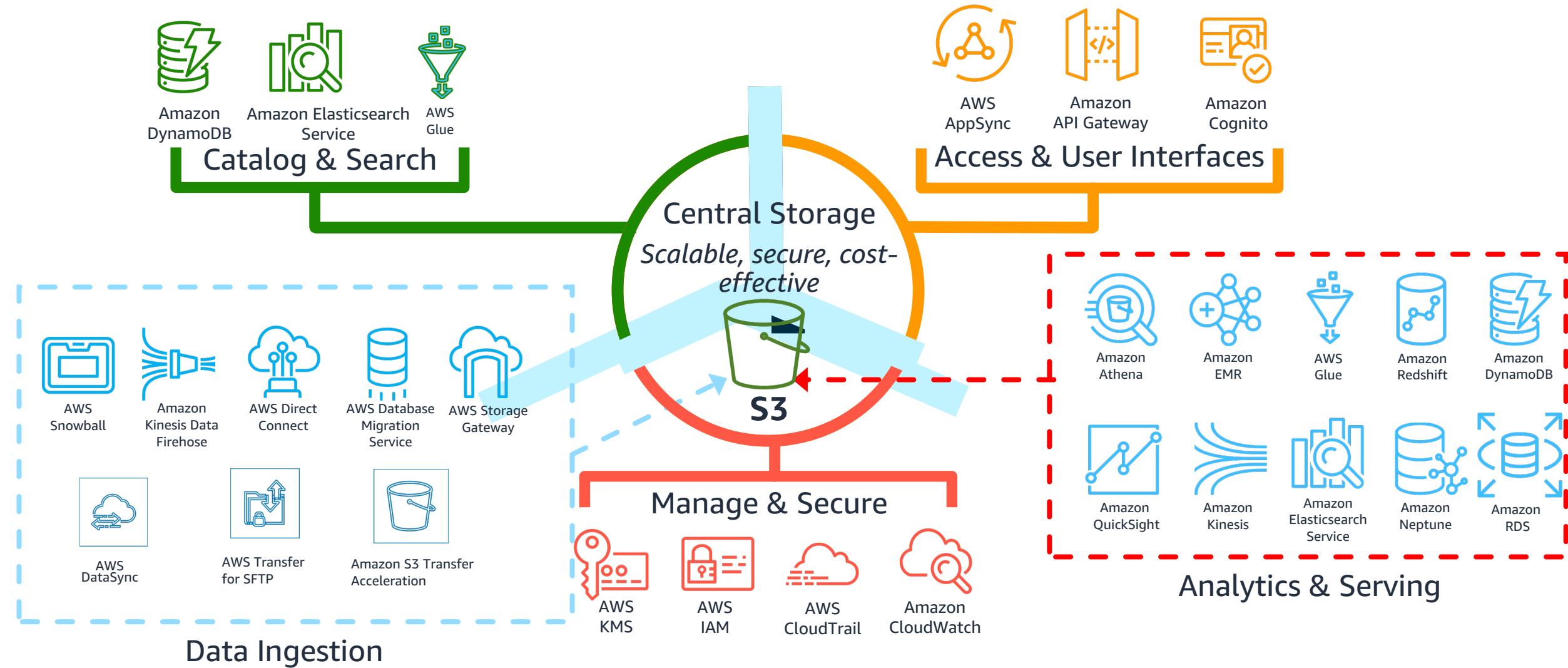


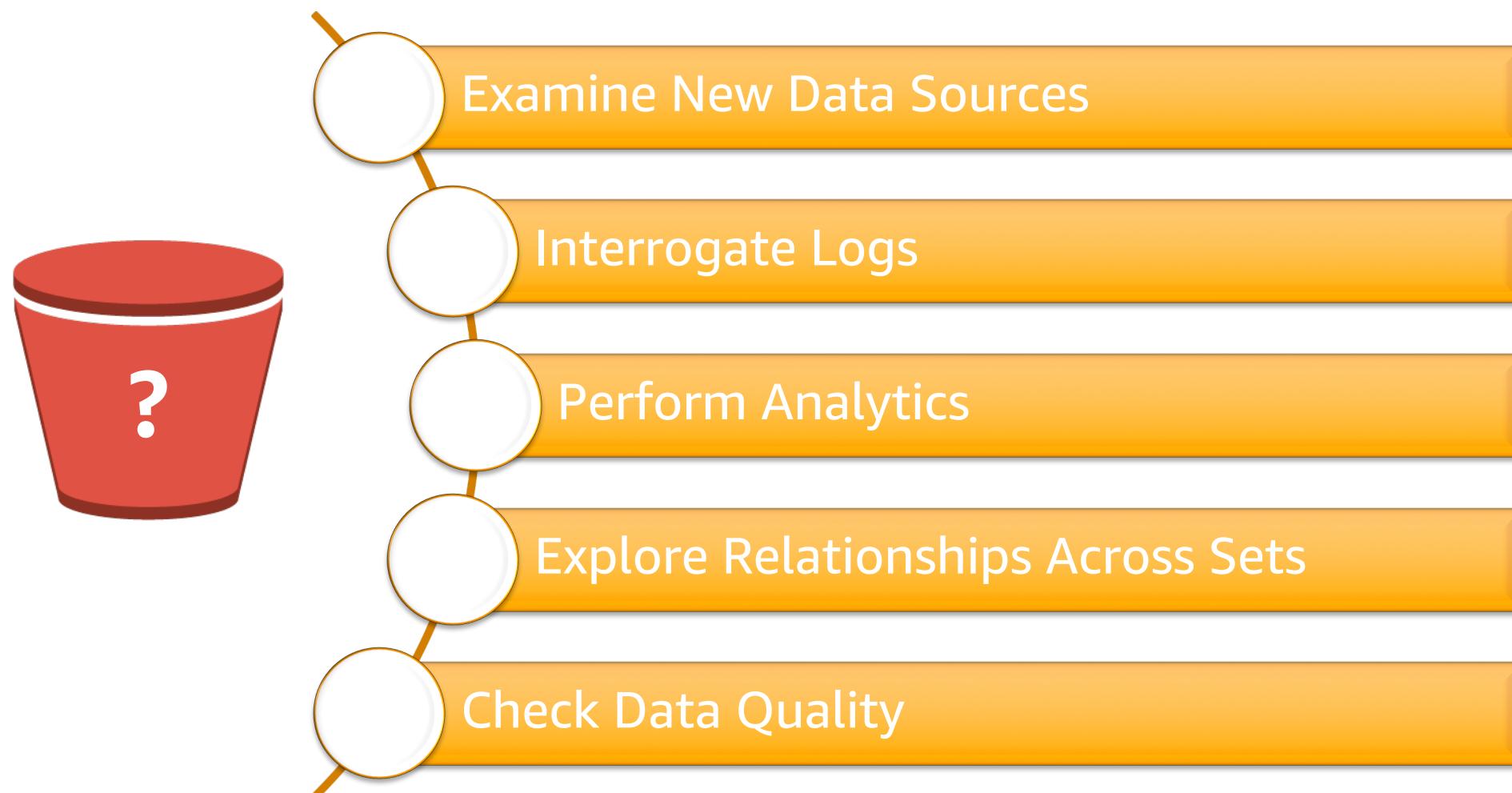
Table of contents

1. Querying the Data Lake with Amazon Athena
2. Visualizing the Data Lake with Amazon QuickSight

Querying the Data Lake

Working with Amazon Athena

When would you query the Data Lake?





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?

Athena **Query Editor** Saved Queries History Catalog

Catalog Sample Queries

DATABASE

default

Table Name

- adsads
- cloudfront_logs
- cloudtrailtable
- elb_logs
- ncc
- ncctest
- prestoinstance
- prestostat
- taxinyc_csv
- taxinyc_par
- test3
- test4
- wikistats
- language (string)
- page_title (string)
- hits (bigint)
- retrieved_size (bigint)
- wikistats_parq

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

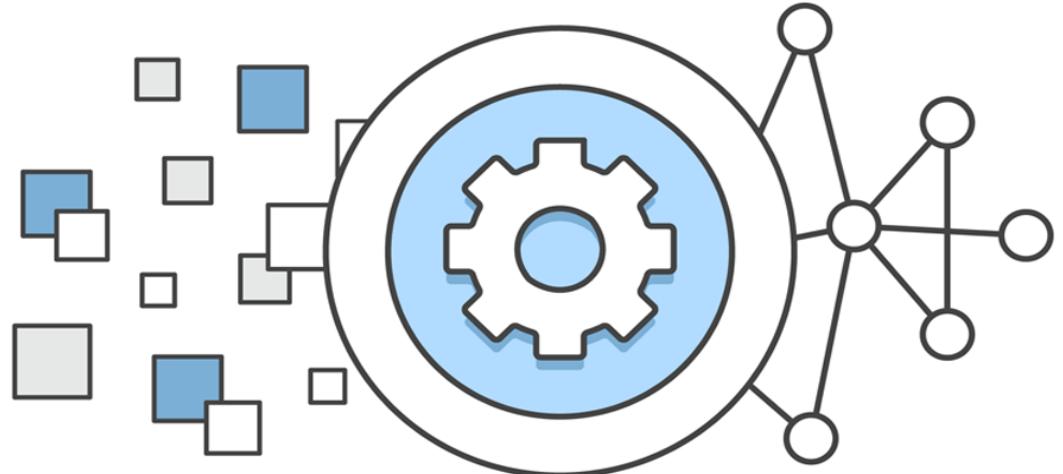
Execute Save As or create a New Query

Recent Queries Query Columns **Results** Chart

	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



Use ANSI SQL

- Support for complex joins, nested queries & window functions
- Support for complex data types (arrays, structs)
- Support for partitioning of data by any key

```
1  WITH q21_tmp1_cached AS
2  (SELECT l_orderkey,
3   count(DISTINCT l_suppkey) AS count_suppkey,
4   max(l_suppkey) AS max_suppkey
5   FROM lineitem_parq
6   WHERE l_orderkey IS NOT NULL
7   GROUP BY l_orderkey),
8   q21_tmp2_cached AS
9  (SELECT l_orderkey,
10   count(DISTINCT l_suppkey) count_suppkey,
11   max(l_suppkey) AS max_suppkey
12   FROM lineitem_parq
13   WHERE l_receiptdate > l_commitdate
14   AND l_orderkey IS NOT NULL
15   GROUP BY l_orderkey)
16  SELECT s_name,
17   count(1) AS numwait
18  FROM
19  (SELECT s_name
20   FROM
21  (SELECT s_name,
22   t2.l_orderkey,
23   l_suppkey,
24   count_suppkey,
25   max_suppkey
26   FROM q21_tmp2_cached t2
27   RIGHT OUTER JOIN
28  (SELECT s_name,
29   l_orderkey,
30   l_suppkey
31   FROM
32  (SELECT s_name,
33   t1.l_orderkey,
34   l_suppkey,
35   count_suppkey,
36   max_suppkey
37   FROM q21_tmp1_cached t1
38   JOIN
39  (SELECT s_name,
40   l_orderkey,
41   l_suppkey
42   FROM orders_parq o
43   JOIN
44  (SELECT s_name,
45   l_orderkey,
46   l_suppkey
47   FROM nation_parq n
48   JOIN supplier s ON s.s_nationkey = n.n_nationkey
49   AND n.n_name = 'SAUDI ARABIA'
50   JOIN lineitem_parq l ON s.s_suppkey = l.l_suppkey
51   WHERE l.l_receiptdate > l.l_commitdate
52   AND l.l_orderkey IS NOT NULL) l1 ON o.o_orderkey = l1.l_orderkey
53   AND o.o_orderstatus = 'F') l2 ON t2.l_orderkey = t1.l_orderkey) a
54   WHERE (count_suppkey > 1)
55   OR ((count_suppkey=1)
56   AND (l_suppkey <> max_suppkey)) b
57   WHERE (count_suppkey IS NULL)
58   OR ((count_suppkey=1)
59   AND (l_suppkey = max_suppkey)) c
60  GROUP BY s_name
61  ORDER BY numwait DESC,
62  s_name LIMIT 100;
```

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

Workload Isolation

Query Metrics

Cost Controls

Workgroups – Workload Isolation

Workgroup name* Use 1 - 128 characters. (A-Z,a-z,0-9,_,-,.)

Description Use up to 1024 characters.

Query result location Select Enter a path to an S3 bucket or prefix.

Encrypt query results Encrypt results stored in S3

Encryption type ⓘ

Encryption key ⓘ Create KMS key

Metrics Publish query metrics to AWS CloudWatch ⓘ

Override user settings ⓘ

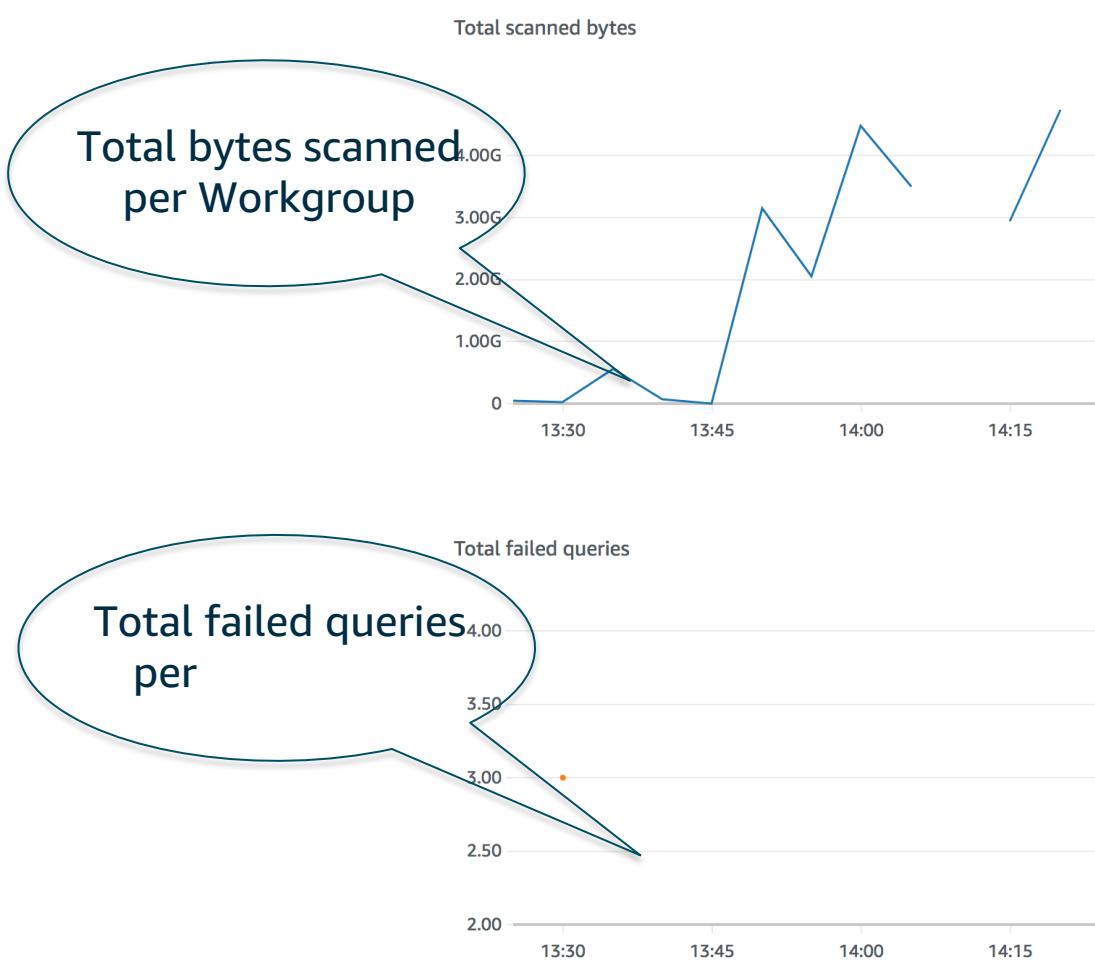
Unique query output location per Workgroup

Encrypt results with unique AWS KMS key per Workgroup

aggregated metrics per Workgroup to AWS CloudWatch

Use Workgroup settings eliminating need to configure individual users

Workgroups – Metric Reporting



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

Workgroups – Usage Notifications

Define a hierarchy of alarms to be alerted as usage

Screenshot of the AWS CloudWatch Metrics Insights interface showing usage notifications for a Workgroup.

The top navigation bar includes "Create Alarm", "Add to Dashboard", "Actions", and filtering options ("State is ALARM"). A search bar and a "Hide all AutoScaling alarms" checkbox are also present.

The main table displays two alarms:

State	Name	Threshold
ALARM	2GBAthenaAlarm	ProcessedBytes >= 2 for 1 datapoints within 10 minutes
ALARM	1GBAthenaAlarm	ProcessedBytes >= 1 for 2 datapoints within 10 minutes

Details for the selected alarm (1GBAthenaAlarm):

State Details: State changed to ALARM at 2018/11/20. Reason: Threshold Crossed: 2 datapoints [5.043427761923077E9 (20/11/18 14:23:00), 3.10345876475E9 (20/11/18 14:18:00)] were greater than or equal to the threshold (1.0).

Description:

- Threshold:** ProcessedBytes >= 1 for 2 datapoints within 10 minutes
- Actions:** In ALARM:
 - Send message to topic "AthenaAlarm" (royon@amazon.com)

Namespace: AWS/Athena
Metric Name: ProcessedBytes
Dimensions: WorkGroup = primary
Statistic: Average
Period: 5 minutes

Treat missing data missing
as:

Percentiles with evaluate
low samples:

A line chart titled "1GBAthenaAlarm" shows the "ProcessedBytes" metric over time. The Y-axis represents Bytes with major ticks at 0, 2.25G, and 4.49G. The X-axis shows time from 12:00 to 14:00. The chart shows a sharp spike from approximately 2.25G to 4.49G around 13:45, followed by a drop and a smaller spike. A red box highlights this spike, and a red exclamation mark is in the top right corner of the chart area. A tooltip below the chart states: "ProcessedBytes >= 1 for 2 datapoints within 10 min...".

[View in metrics](#)

Athena support for interface endpoint (PrivateLink)

Submit queries securely

- No internet gateway required in your VPC
- Secure communication between your VPC and Athena APIs
- Set VPC endpoint policies
 - Example endpoint policy

```
{  
    "Statement": [{  
        "Principal": "*",
        "Effect": "Allow",
        "Action": [  
            "athena:StartQueryExecution",
            "athena:RunQuery",
            "athena:GetQueryExecution",
            "athena:GetQueryResults",
            "athena:CancelQueryExecution",
            "athena>ListWorkGroups",
            "athena:GetWorkGroup",
            "athena:TagResource"  

        ],
        "Resource": [  

            "arn:aws:athena:us-west-1:AWSAccountID:workgroup/workgroupA"  

        ]
    }]  
}
```

Athena support for INSERT INTO

Inserts new rows into a destination table based on a SELECT query statement that runs on a source table, or based on a set of VALUES provided as part of the statement

- Supported Format
 - Avro, JSON, ORC, Parquet, Text file
- Example

```
INSERT INTO cities
VALUES (1, 'Lansing', 'MI', 'Si quaeris peninsulae amoenam circumspice')
```

```
INSERT INTO cities
VALUES (1, 'Lansing', 'MI', 'Si quaeris peninsulae amoenam circumspice'),
(3, 'Boise', 'ID', 'Esto perpetua')
```

Visualizing the Data Lake

Working with 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



Fast, consistent performance

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



Fully integrated with AWS

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



ML insights

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



Secure and global

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



Insights for everyone

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



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



Customize and embed

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

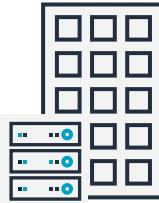


Connect to your data, wherever it is

QuickSight is natively integrated with AWS data sources, as well as on-premise 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



In the cloud

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



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



Applications

Connect directly to third party business applications

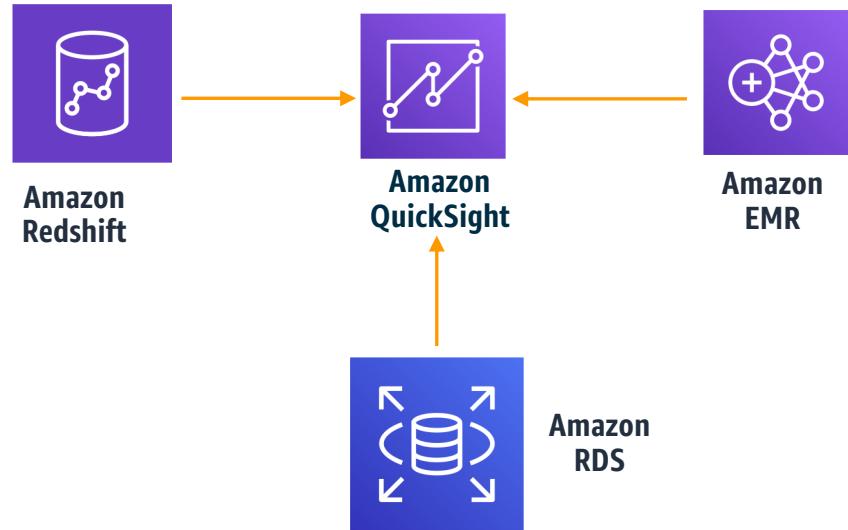


- Salesforce
- Square
- Adobe Analytics
- Jira
- ServiceNow
- Twitter
- Github



service**now**

AWS + Amazon QuickSight



Private VPC connectivity = no public routing of data

Direct query to data sources or SPICE for fast access



Native fine-grained permissions for users with AWS Identity and Access Management (IAM)

Cost allocation by business unit or team

Create dashboards in minutes

Pay for what you use

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

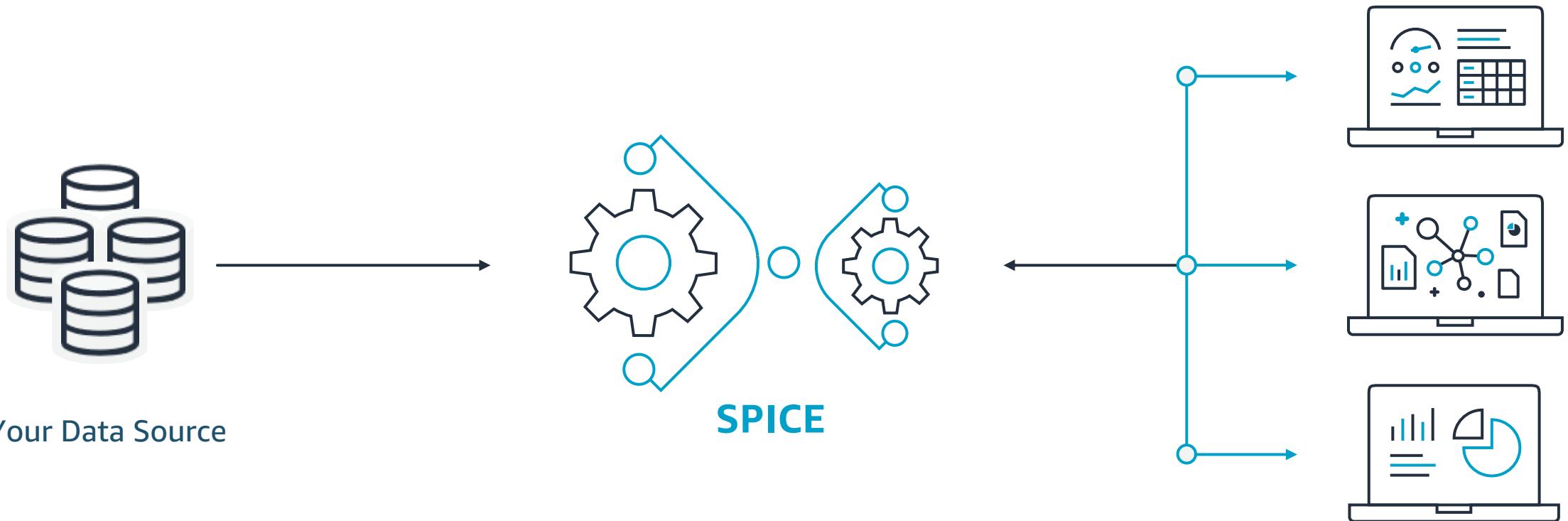
Visually join tables from the same relational DB

Push down custom SQL queries

The screenshot shows the AWS Data Prep interface. On the left, the 'Data source' section indicates a SPICE source with 1.5GB estimated upload remaining. The 'Tables' section shows 1 table selected. The 'Fields' section lists all fields selected, including 'customer id', 'customer name', and 'customer type'. The 'Calculated fields' section is empty. The main area displays a preview of the 'retail_sales' dataset with columns like cust..., city, state, coun..., zip c..., region, and age r... containing various data points. A visual join configuration is shown between the 'all_flights' table (customer id) and the 'airport_id' table. A red circle highlights the 'Configure join' button. At the bottom, the 'Data sources' section shows 'all_flights' and 'INNER' joined with 'airport_id', and the 'Join types' section includes options for Inner, Left, Right, and Outer joins.

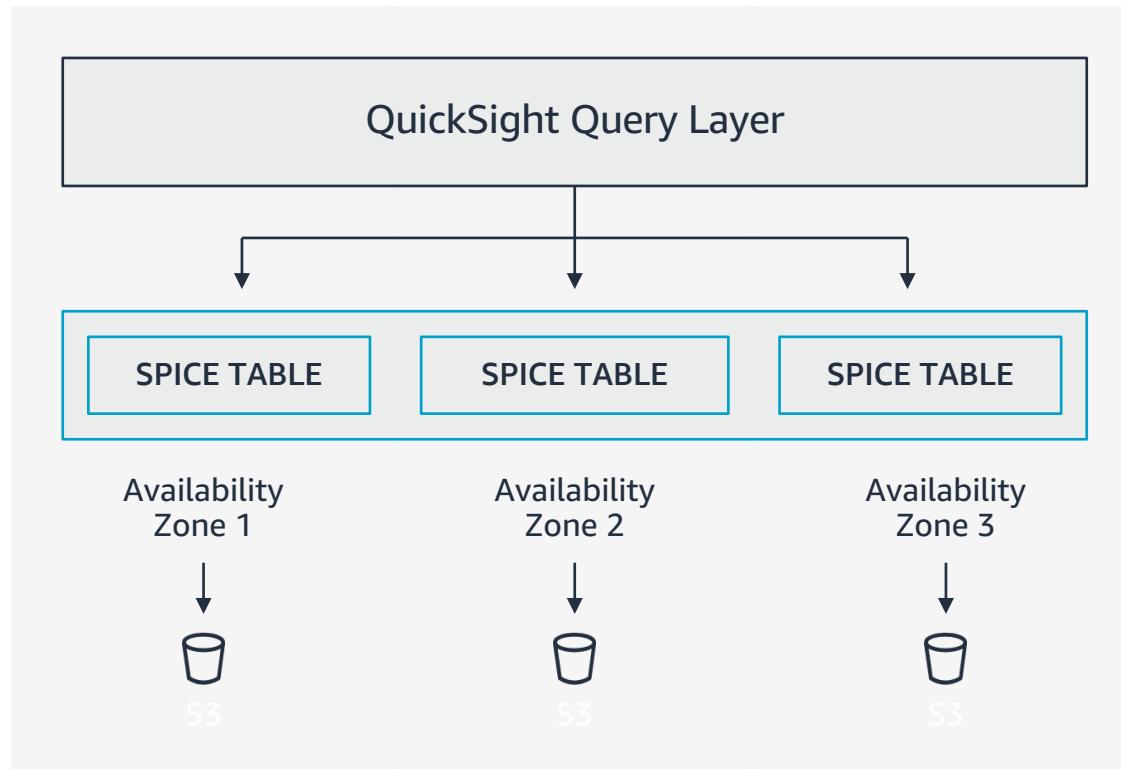
SPICE

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



SPICE

SPICE not only provides your users with instant response times, but automatically scales with user activity, protecting your underlying data sources saving you time and money.



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



Author

Create Data Sets
Create Analyses
Create Dashboards



Reader

Consume Dashboards

QS Admin

Sometimes separate from Business Users,
sometimes the same
Usually has AWS Console

Analyst

Sometimes in IT, sometimes Business Users
'Data Analyst'
'Data Engineer'
'BI Engineer'

Business User

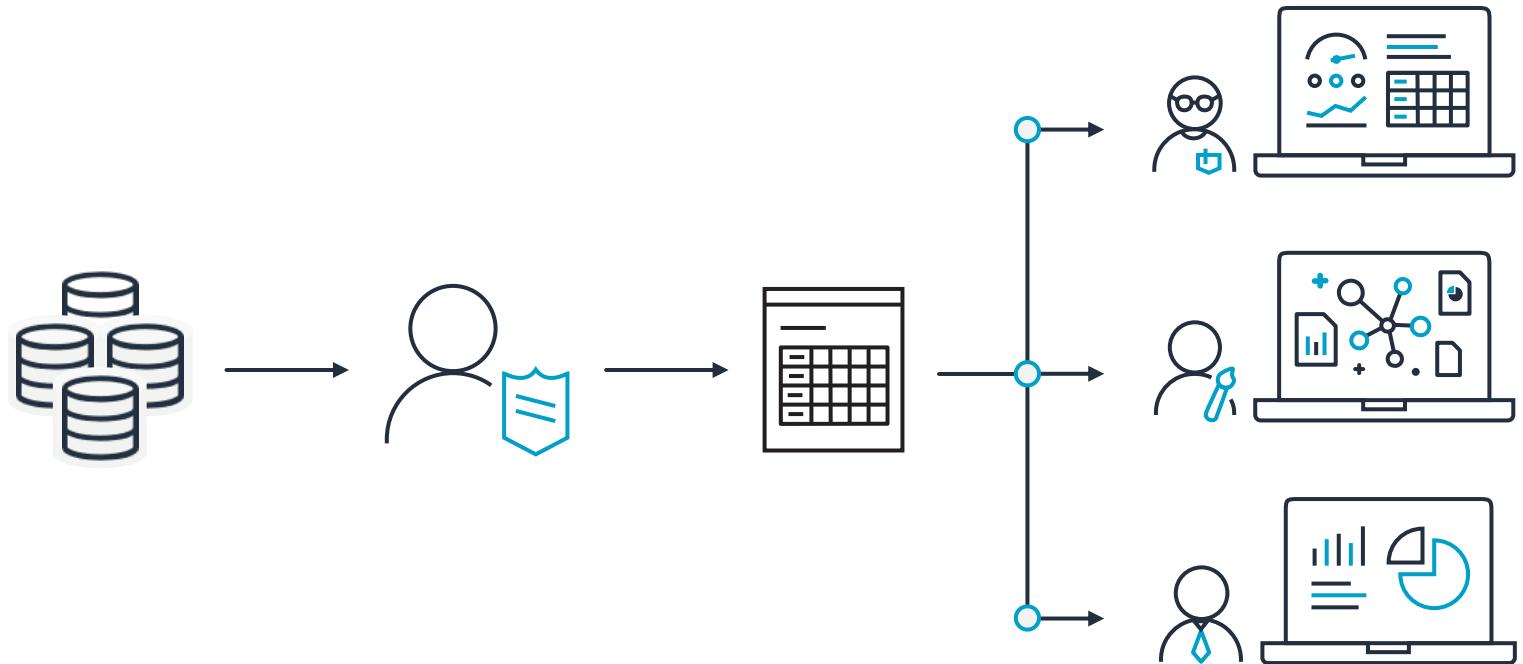
Anyone
Can be internal or external users
(customers/partners/3rd)

Data governance

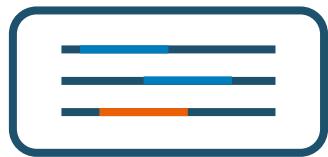
Create managed datasets that give power users and authors the flexibility to perform self-serve analytics on data that you control.

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

Anomaly Detection

Contributors

Daily sales for Region: APAC | Product: Office Supplies increased 20% from \$10K to \$12K on Oct 31, 2018 compared to the day before and was 15% higher than expected.

Last updated on Nov 13th 2018

Top Contributors Configure

Sort by Contribution percentage

Segment	% contr.	DoD change
SMB	53%	30% ***
Unicorn	17%	7% ***
Enterprise	13%	5% ***

Channel	% contr.	DoD change
Online	81%	12% ***

Region: APAC | Product: Office Supplies

Daily revenue for Segment: SMB increased from \$3K to \$ 4.1K (+25%). It contributed to 53% of the total increase for Region: APAC | Product: Office Supplies, higher than the expected 30% contribution.

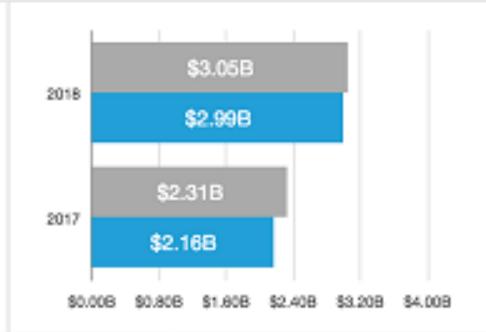
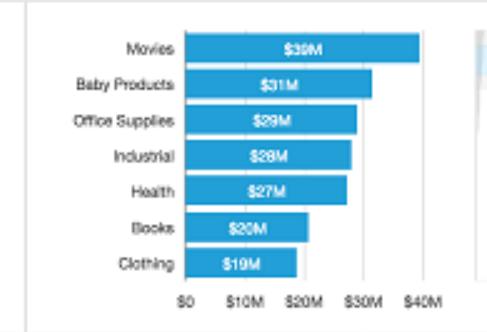
Region: North America | Product: Outdoor

©

Forecasting



Natural Language Narratives

<p>Daily Revenue</p> <p>Daily revenue decreased -0.51% (-\$57,032.99) on Nov 18, 2018, from \$11.19M to \$11.14M compared to the previous day and is -1.78% (-\$202,111.70) below goal of \$11.34M. We are \$2789.67K (0.334%) above 30-day average revenue of \$8.35M. We're operating at an run rate of \$4.06B.</p>	<p>YTD Revenue</p> <p>Year-to-date revenue increased by 61.95% (\$1.14B) from \$1.85B to \$2.99B compared to the same period last year and is -1.81% (\$55.03M) below plan of \$3.05B. We are at 98.19% achievement of YTD goal and 84.61% achievement for annual goal.</p>	<p>Callouts By Product and Country</p> <p>Daily revenue for Baby Products Russia on Nov 22, 2018 was lower than expected at \$1.170.58.</p>	<p>Daily Revenue Forecast</p> <p>Daily revenue is predicted to reach \$11.96M by end of the year. We expect to exit the year with an annualized run rate of \$4.37B. Total revenue for 2018 is predicted to reach \$3.47B, \$63.59M (-1.80%) below annual target of \$3.54B.</p>																																												
																																															
<p>Top / Bottom Movers by Product</p> <p>Top daily revenue increase by products are:</p> <ul style="list-style-type: none"> Electronics increased by \$537.84 (7.61%) from \$7,066.98 to \$7,604.82. Clothing increased by \$484.86 (0.06%) from \$769,561.89 to \$770,046.75. Industrial increased by \$427.49 (0.04%) from \$1,078,710.02 to \$1,079,137.51. Home Services increased by \$114.68 (0.21%) from \$55,338.04 to \$55,452.72. Music increased by \$87.40 (0.67%) from \$12,997.83 to \$13,085.23. 	<p>Top / Bottom to Plan Variance by Product</p> <p>Top products above plan for today are:</p> <ul style="list-style-type: none"> Movies is \$147,437.42 above goal. Financial Services is \$98,111.46 above goal. Clothing is \$42,220.37 above goal. Computers is \$38,003.06 above goal. Outdoors is \$27,404.67 above goal. <p>Top products below plan for today are:</p> <ul style="list-style-type: none"> Digital is -\$236,951.70 below goal. Health is -\$147,834.66 below goal. 	<p>Revenue by Product Category</p> <table border="1"> <thead> <tr> <th>Product Ca...</th> <th>Nov 18, 2018</th> <th>Nov 17, 2018</th> <th>N</th> </tr> <tr> <th></th> <th>Revenue</th> <th>Revenue</th> <th></th> </tr> </thead> <tbody> <tr> <td>Arts</td> <td>\$4,988.40</td> <td>\$4,988.18</td> <td></td> </tr> <tr> <td>Automotive</td> <td>\$52,309.00</td> <td>\$52,493.34</td> <td></td> </tr> <tr> <td>Baby Product</td> <td>\$1,354,243.11</td> <td>\$1,368,901.33</td> <td></td> </tr> <tr> <td>Beauty</td> <td>\$8,114.71</td> <td>\$8,116.06</td> <td></td> </tr> <tr> <td>Books</td> <td>\$1,330,700.80</td> <td>\$1,331,300.71</td> <td></td> </tr> <tr> <td>Business</td> <td>\$38,736.02</td> <td>\$41,916.59</td> <td></td> </tr> <tr> <td>Clothing</td> <td>\$770,046.75</td> <td>\$769,561.89</td> <td></td> </tr> <tr> <td>Collectibles</td> <td>\$709.44</td> <td>\$803.63</td> <td></td> </tr> <tr> <td>Computers</td> <td>\$539,996.53</td> <td>\$540,576.75</td> <td></td> </tr> </tbody> </table>	Product Ca...	Nov 18, 2018	Nov 17, 2018	N		Revenue	Revenue		Arts	\$4,988.40	\$4,988.18		Automotive	\$52,309.00	\$52,493.34		Baby Product	\$1,354,243.11	\$1,368,901.33		Beauty	\$8,114.71	\$8,116.06		Books	\$1,330,700.80	\$1,331,300.71		Business	\$38,736.02	\$41,916.59		Clothing	\$770,046.75	\$769,561.89		Collectibles	\$709.44	\$803.63		Computers	\$539,996.53	\$540,576.75		
Product Ca...	Nov 18, 2018	Nov 17, 2018	N																																												
	Revenue	Revenue																																													
Arts	\$4,988.40	\$4,988.18																																													
Automotive	\$52,309.00	\$52,493.34																																													
Baby Product	\$1,354,243.11	\$1,368,901.33																																													
Beauty	\$8,114.71	\$8,116.06																																													
Books	\$1,330,700.80	\$1,331,300.71																																													
Business	\$38,736.02	\$41,916.59																																													
Clothing	\$770,046.75	\$769,561.89																																													
Collectibles	\$709.44	\$803.63																																													
Computers	\$539,996.53	\$540,576.75																																													

ABCO Daily Sales Report

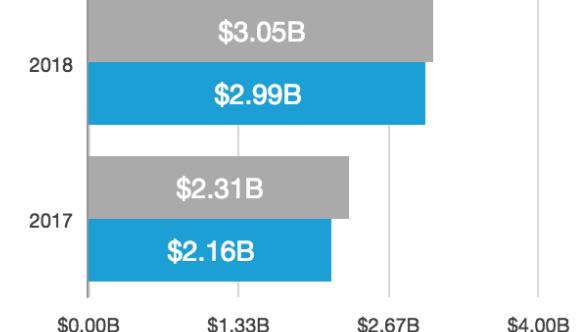
Daily Revenue

Daily revenue **decreased -0.51%**
(-\$57,032.99) on Nov 18, 2018, from \$11.19M to \$11.14M compared to the previous day and is **-1.78% (-\$202,111.70) below** goal of \$11.34M. We are **\$2789.67K (0.334%) above** 30-day average revenue of \$8.35M. We're operating at an **run rate of \$4.06B**.



YTD Revenue

Year-to-date revenue **increased by 61.95%** (**\$1.14B**) from \$1.85B to \$2.99B compared to the same period last year and is **-1.81% (\$55.03M) below** plan of \$3.05B. We are at **98.19%** achievement of YTD goal and **84.61%** achievement for annual goal.



Callouts By Product and Country

Daily revenue for **Baby Products | Russia** on Nov 22, 2018 was **lower than expected** at \$1 170 58

Callouts By Customers

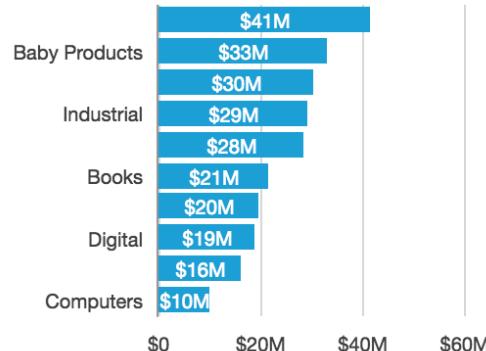
Daily revenue for **MULTIDEL INC.** on Nov 22, 2018 was **higher than expected** at \$60 433 05

Daily Revenue Forecast

Daily revenue is predicted to reach **\$12.02M** by end of the year. We expect to exit the year with an annualized run rate of **\$4.39B**. Total revenue for 2018 is predicted to reach \$3.47B, **\$63.06M (-1.78%) below** annual target of \$3.54B.

Top / Bottom Movers by Product

Top daily revenue **increase** by products are:



Top / Bottom to Plan Variance by Product

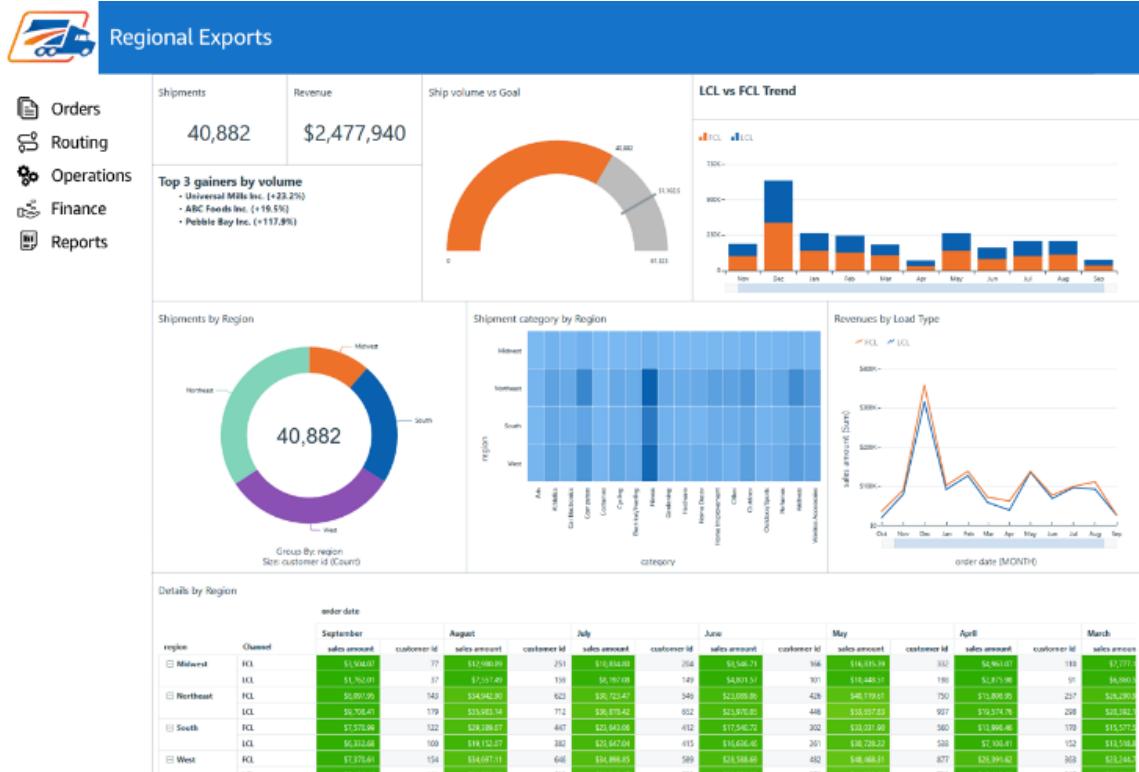
Top products **above** plan for today are:

Revenue by Product Category

Nov 18, 2018

Nov 17, 2018

Embed Amazon QuickSight Dashboards



Fully interactive with drill down, filtering, & external links

Personalized views with row-level security

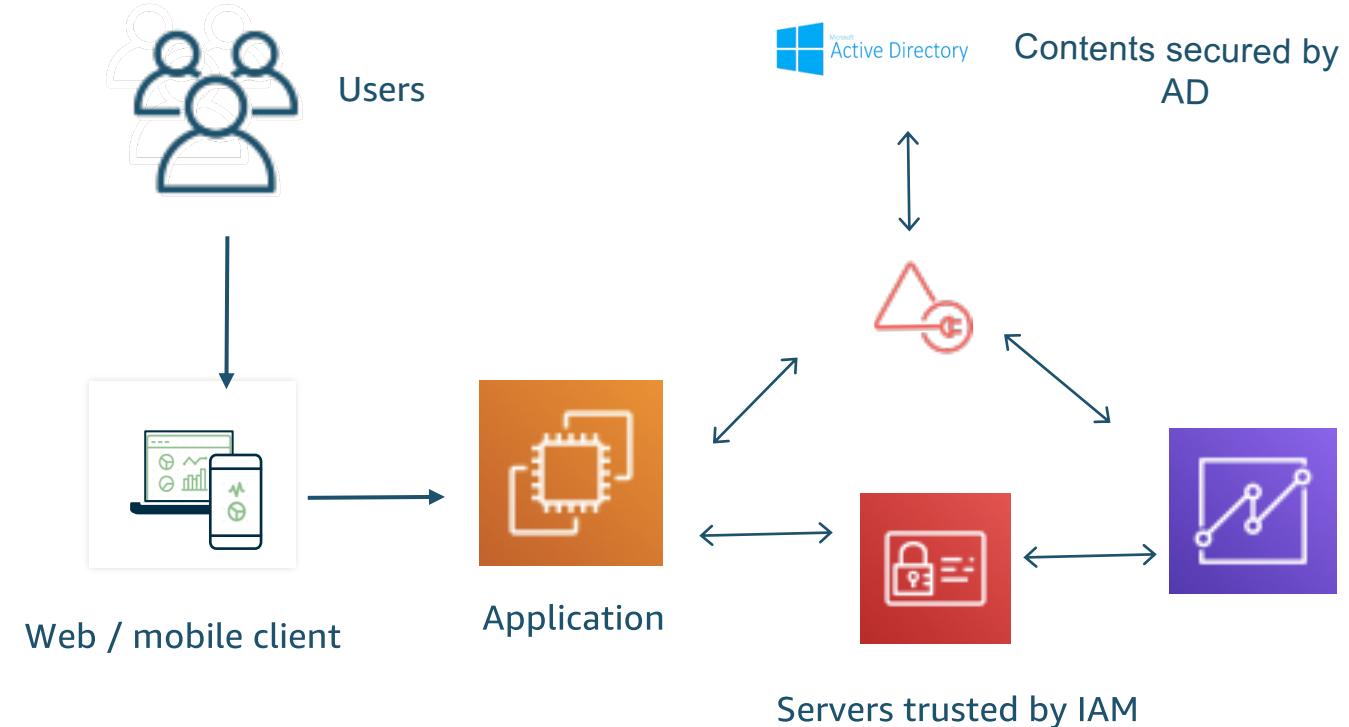
No servers to manage, no long-term commitments

Pay for usage with pay-per-session reader pricing

Seamless authentication

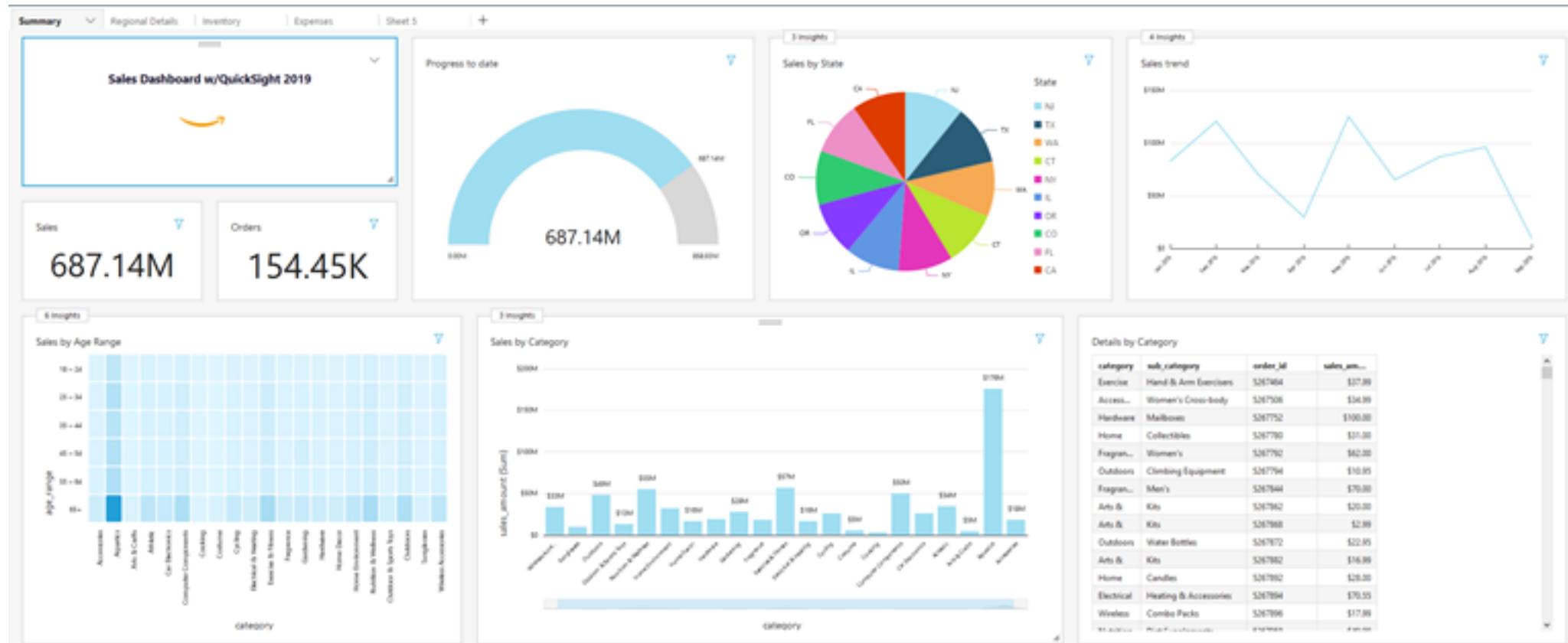
Why Amazon QuickSight for embedded dashboards

- Create dashboard
- Configure permission
- Generate signed URL
- Embed URL



What's new? Theming

You can now create a collection of themes, and apply a theme to an analysis and all its dashboards



New APIs

Data source APIs

Create/manage data source connections w/o sharing credentials

Audit data sources, manage ownership and access

Dataset APIs

Create customized datasets for users or groups

Allows isolation of data, additional row-level security possible

SPICE ingestion APIs

Trigger SPICE ingestion of data when ETL/data load is complete

Easily review history and trace SPICE ingestion details

Fine grained access control APIs

Manage user/group access to Amazon S3/Athena data via IAM policies

Template APIs

Create templates from analysis

Metadata for dashboard with placeholder for datasets

Templates accessible via API only

Can be copied/referenced across accounts

Dashboard APIs

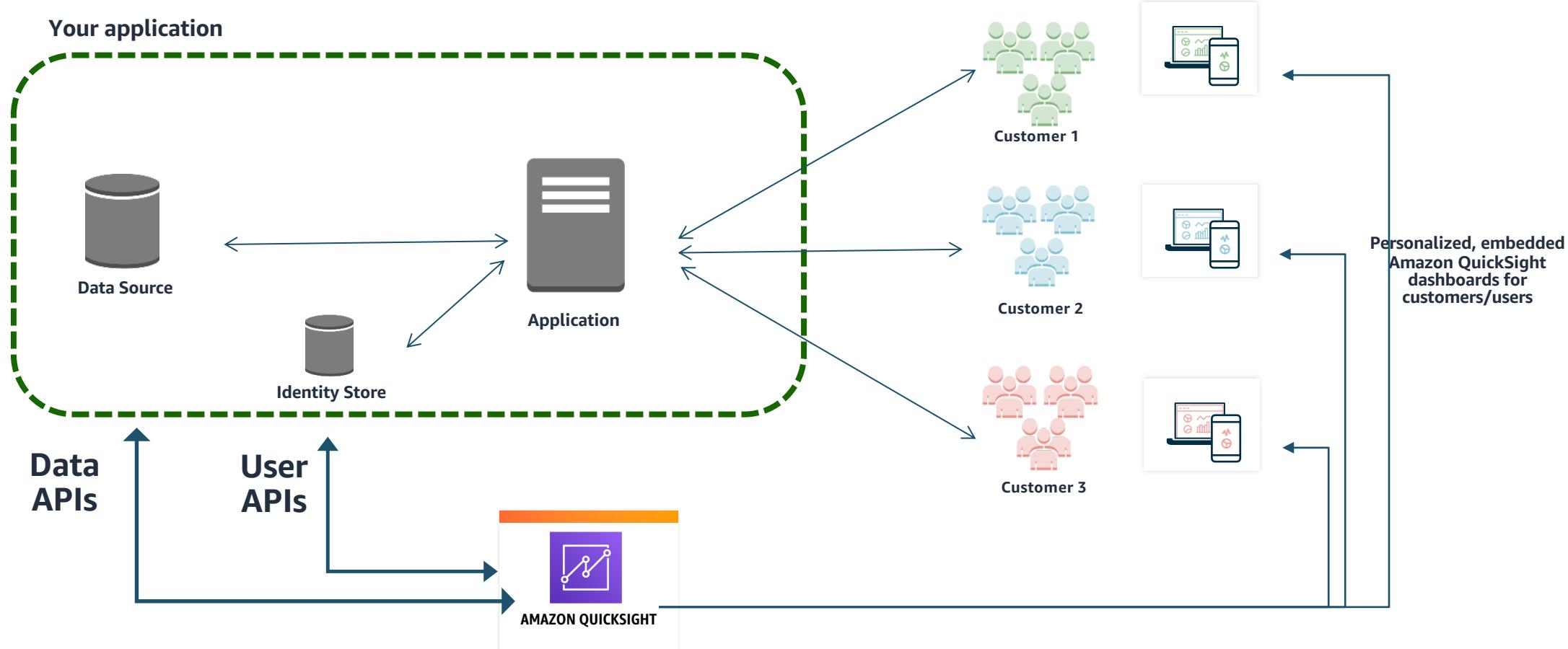
Easily create dashboards per user/group w/different datasets

Move dashboards across dev environments

Version dashboards for easy rollbacks

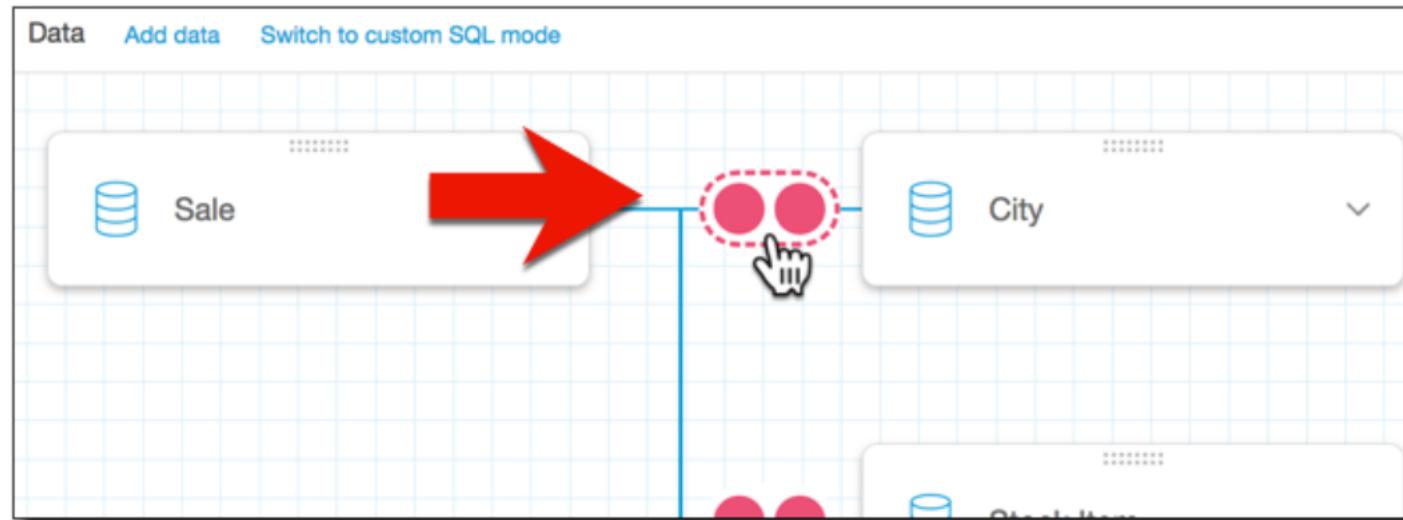
Audit dashboards, list by users, manage ownership/sharing

Using APIs for your embedded deployments

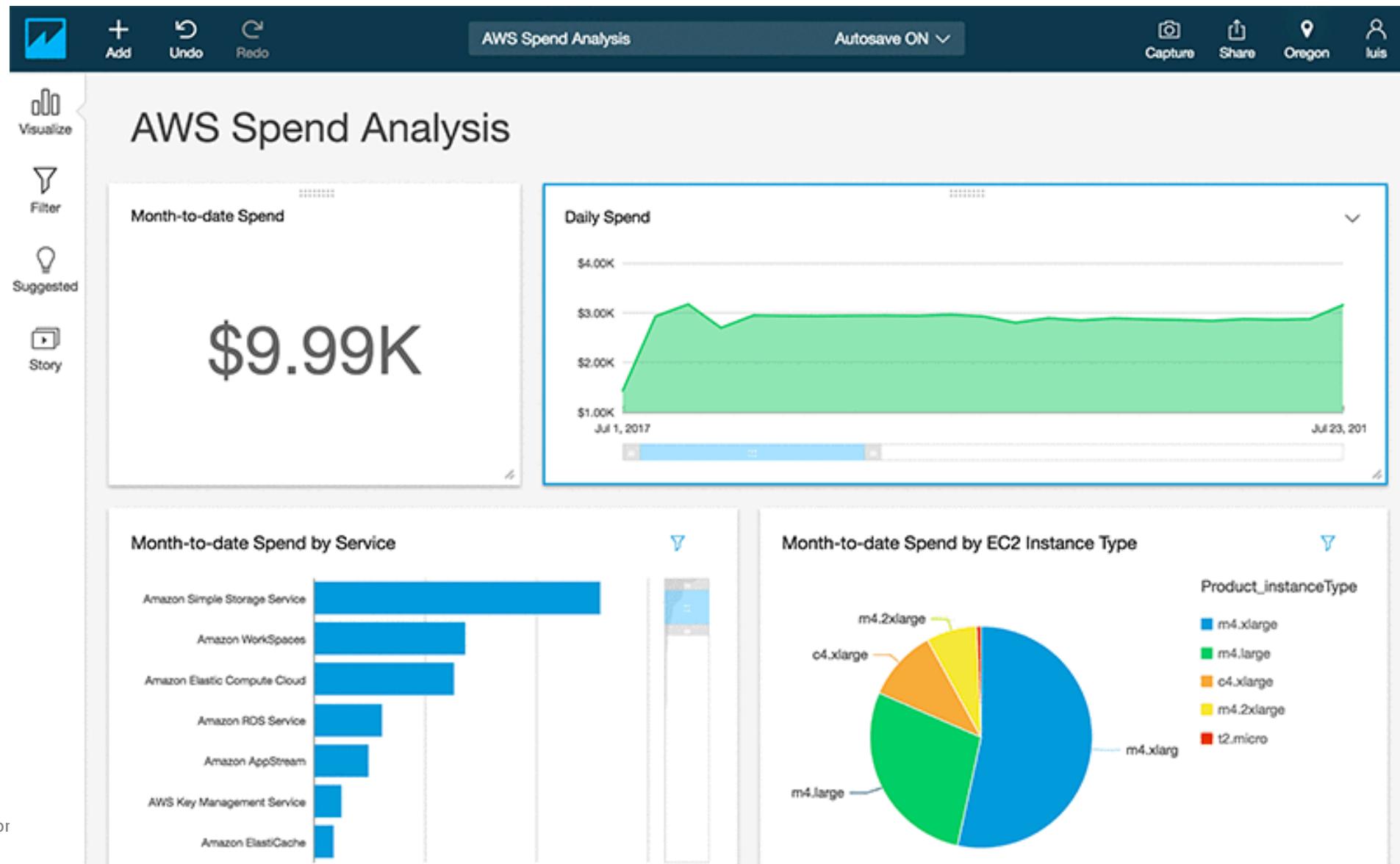


QuickSight Support for Cross Data Source Join

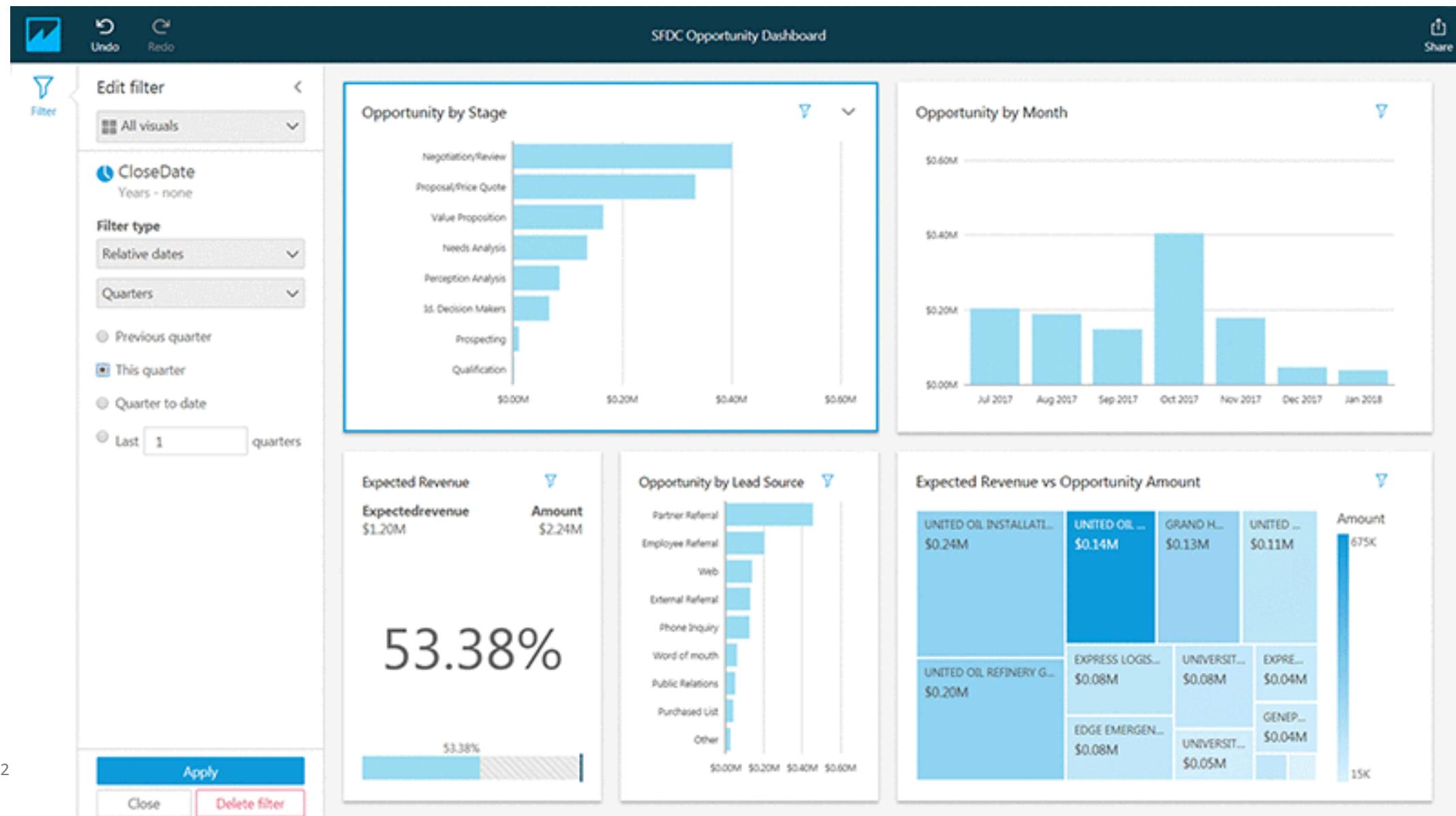
- Join across all data sources supported by QuickSight including file-to-file, file-to-database, and database-to-database joins



Amazon QuickSight: Examples – AWS Cost & Usage Reporting



Amazon QuickSight: Examples – Salesforce Analytics



Q & A