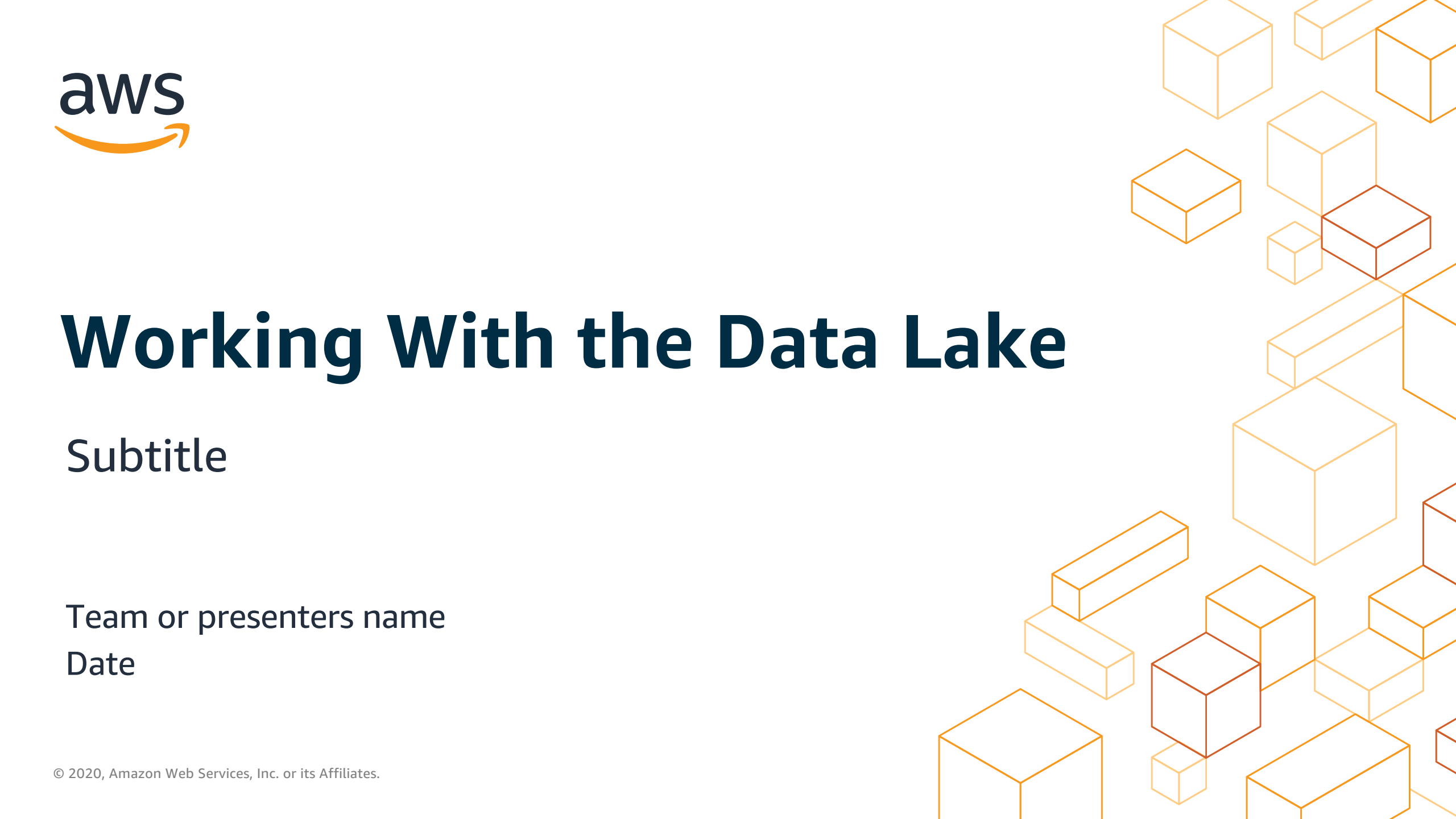


Working With the Data Lake

Subtitle

Team or presenters name
Date

Session's Focus – Query the The Data Lake

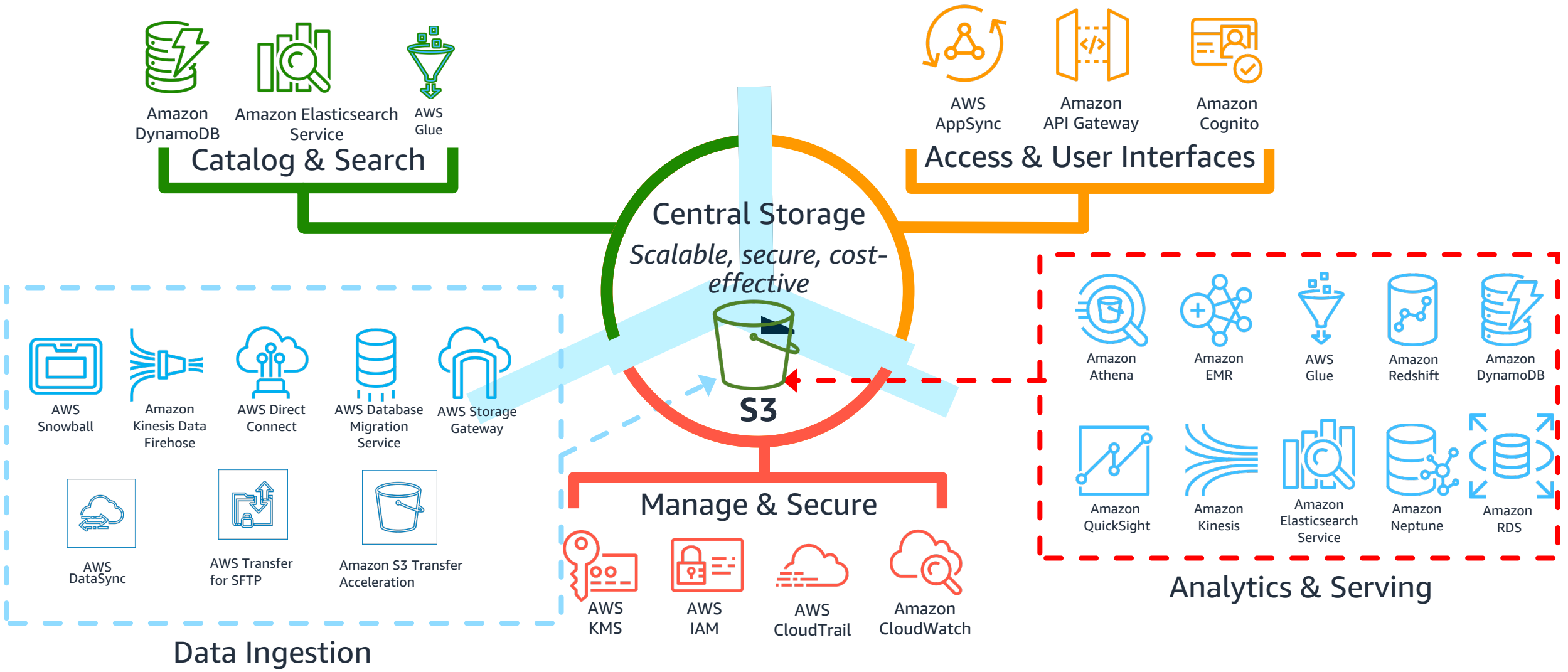


Table of contents

aws

Querying the Data Lake

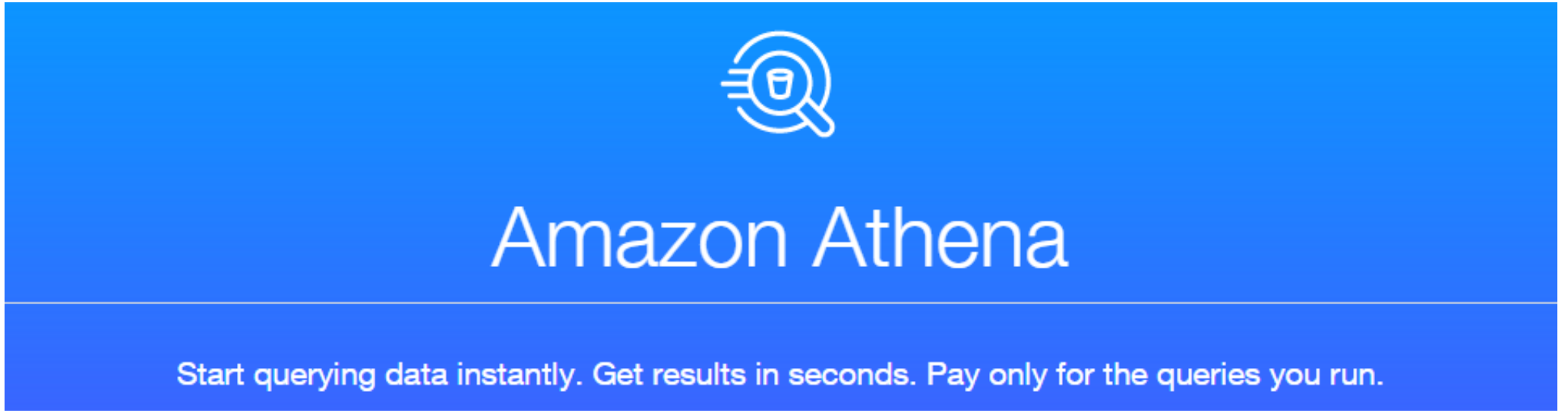
Working with Amazon Athena

aws

When would you query the Data Lake?

- Examine New Data Sources
- Interrogate Logs
- Perform Analytics
- Explore Relationships Across Sets
- Check Data Quality

aws



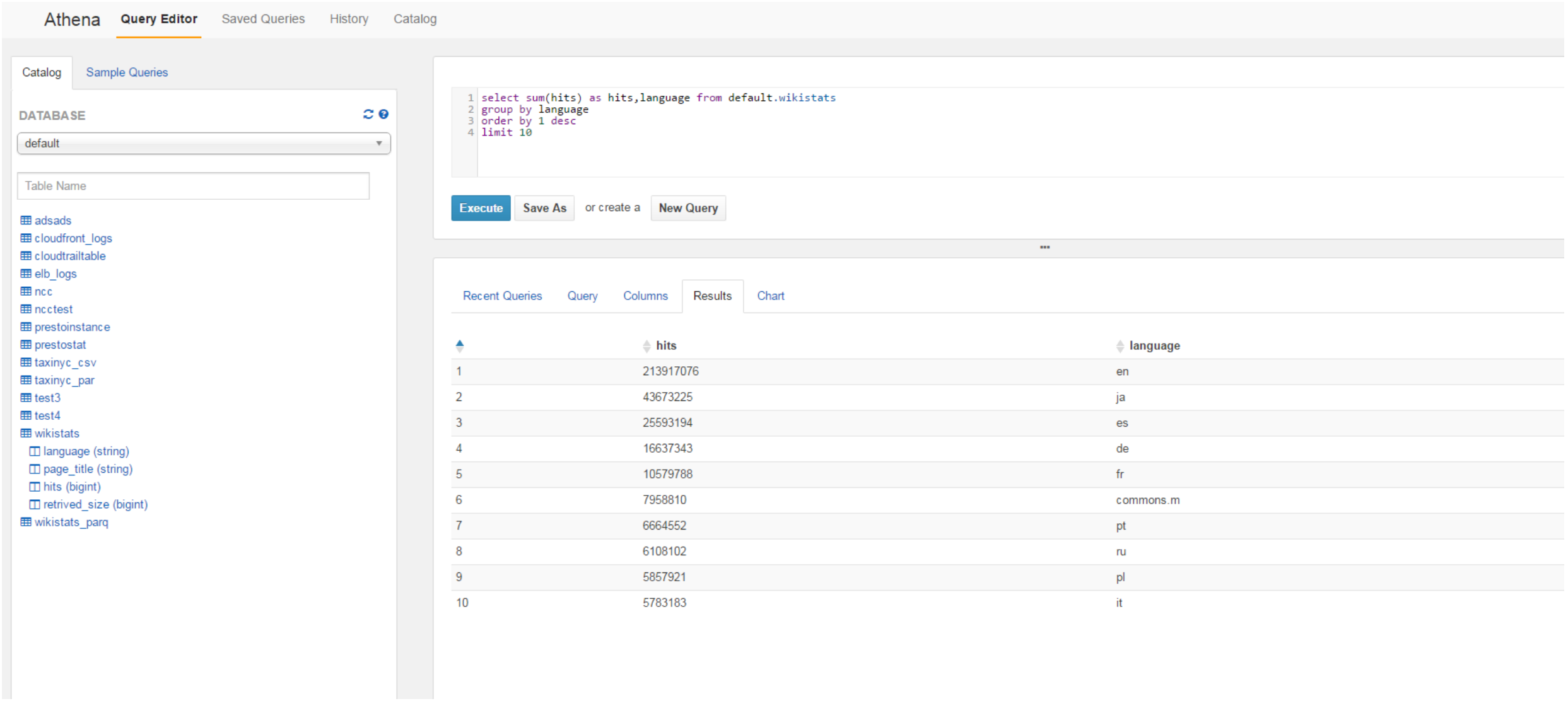
An interactive query service that makes it easy to analyze data directly from Amazon S3 using Standard SQL

aws

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

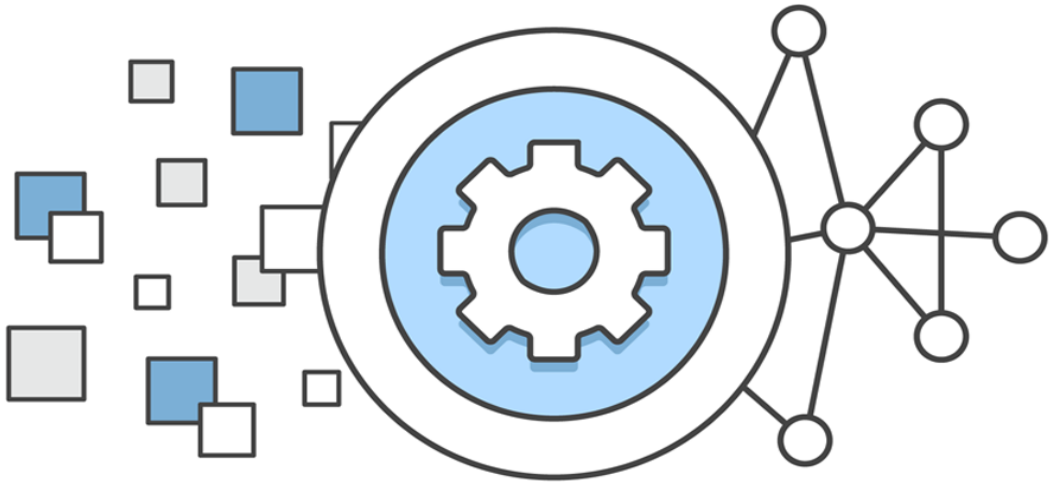
What does it look like?



aws

Athena is Serverless

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



aws

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

```
WITH q21_tmp1_cached AS
 (SELECT l_orderkey,
         count(DISTINCT l_suppkey) AS count_suppkey,
         max(l_suppkey) AS max_suppkey
  FROM lineitem_parq
  WHERE l_orderkey IS NOT NULL
  GROUP BY l_orderkey),
     q21_tmp2_cached AS
 (SELECT l_orderkey,
         count(DISTINCT l_suppkey) count_suppkey,
                                   max(l_suppkey) AS max_suppkey
  FROM lineitem_parq
  WHERE l_receiptdate > l_commitdate
    AND l_orderkey IS NOT NULL
  GROUP BY l_orderkey)
SELECT s_name,
       count(1) AS numwait
FROM
  (SELECT s_name
   FROM
     (SELECT s_name,
             t2.l_orderkey,
             l_suppkey,
             count_suppkey,
             max_suppkey
      FROM q21_tmp2_cached t2
      RIGHT OUTER JOIN
        (SELECT s_name,
                l_orderkey,
                l_suppkey
         FROM
           (SELECT s_name,
                   t1.l_orderkey,
                   l_suppkey,
                   count_suppkey,
                   max_suppkey
            FROM q21_tmp1_cached t1
            JOIN
              (SELECT s_name,
                      l_orderkey,
                      l_suppkey
               FROM orders_parq o
               JOIN
                 (SELECT s_name,
                         l_orderkey,
                         l_suppkey
                  FROM nation_parq n
                  JOIN supplier s ON s.s_nationkey = n.n_nationkey
                  AND n.n_name = 'SAUDI ARABIA'
                  JOIN lineitem_parq l ON s.s_suppkey = l.l_suppkey
                  WHERE l.l_receiptdate > l.l_commitdate
                    AND l.l_orderkey IS NOT NULL) l1 ON o.o_orderkey = l1.l_orderkey
               AND o.o_orderstatus = 'F') l2 ON l2.l_orderkey = t1.l_orderkey) a
         WHERE (count_suppkey > 1)
           OR ((count_suppkey=1)
               AND (l_suppkey <> max_suppkey))) l3 ON l3.l_orderkey = t2.l_orderkey) b
   WHERE (count_suppkey IS NULL)
     OR ((count_suppkey=1)
         AND (l_suppkey = max_suppkey))) c
GROUP BY s_name
ORDER BY numwait DESC,
         s_name LIMIT 100;
```


aws

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

aws

Athena Workgroups

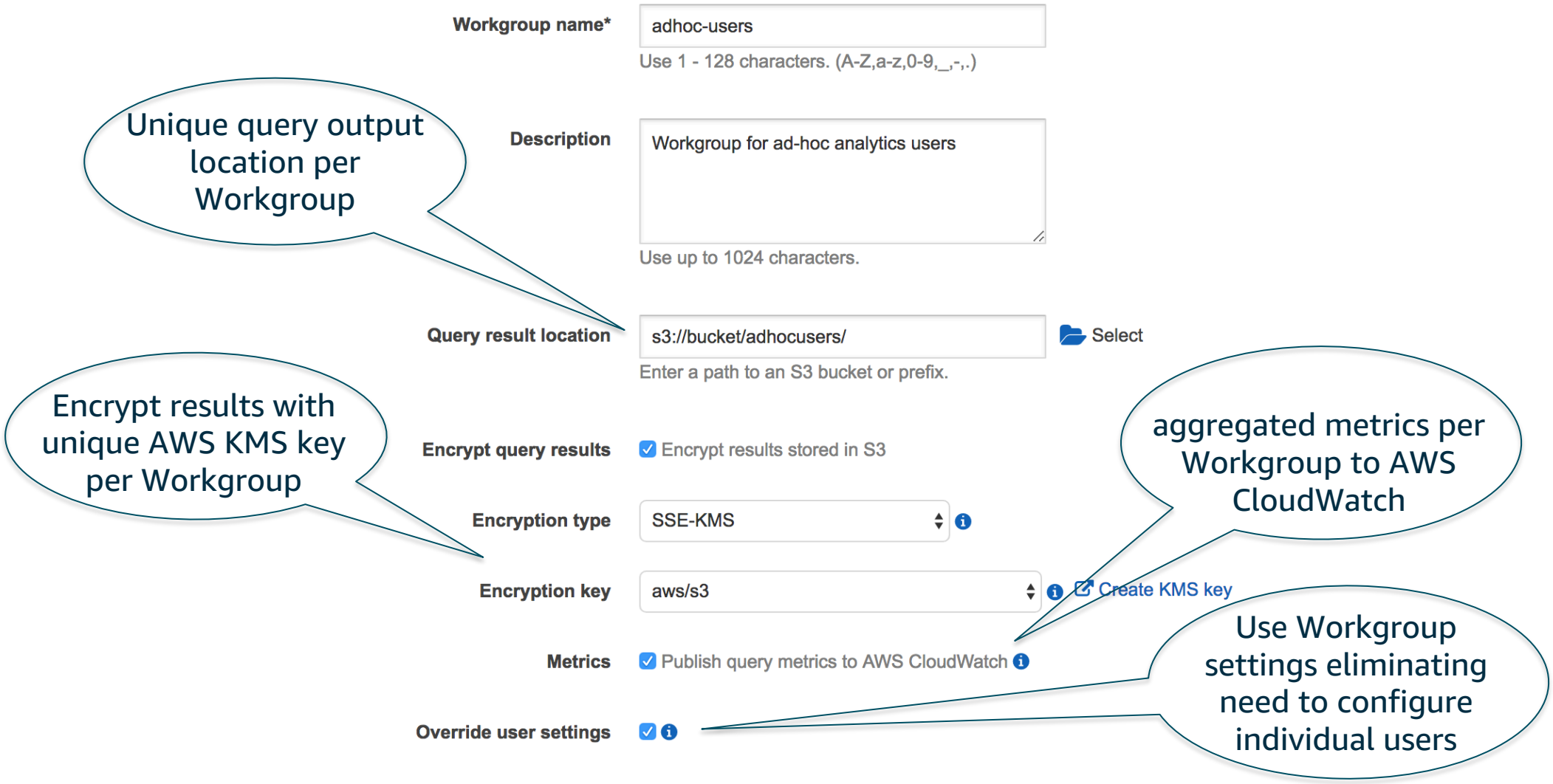
Athena Workgroups are used to isolate queries between different teams, workloads or applications, and to set on amount of data each query or the entire workgroup can process

Workload Isolation

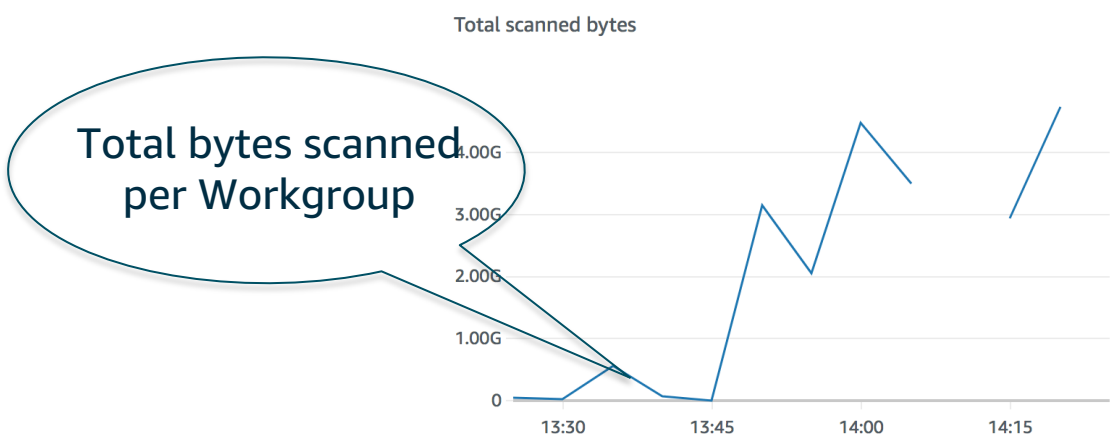
Query Metrics

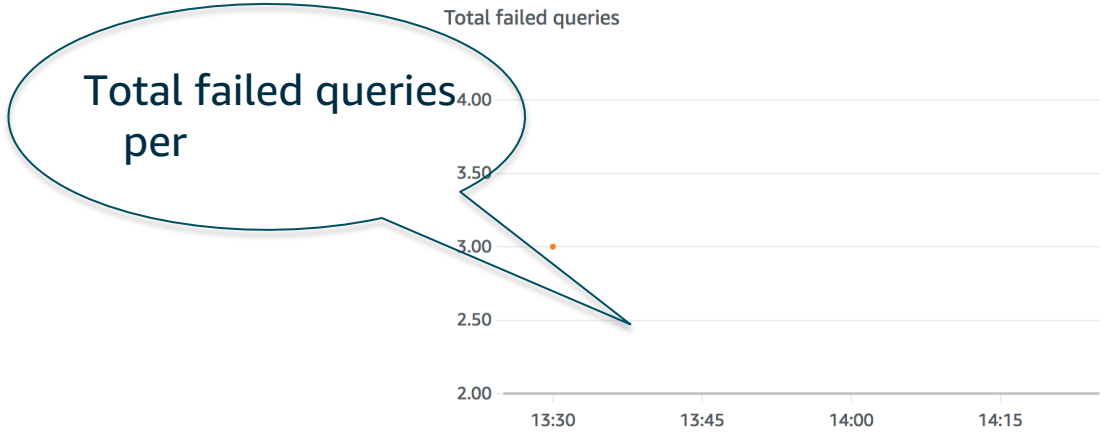
Cost Controls

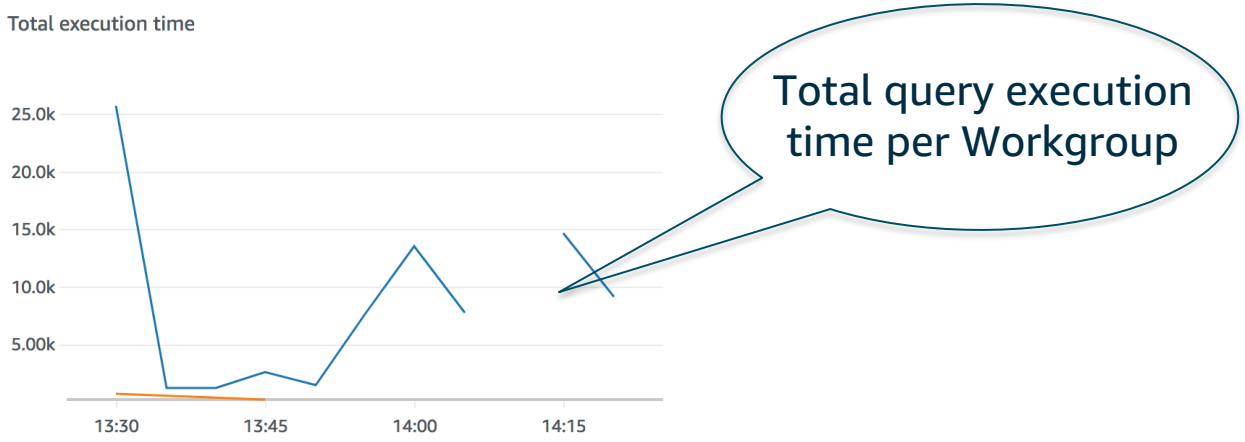
Workgroups – Workload Isolation



Workgroups – Metric Reporting







Workgroups – Cost Controls

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

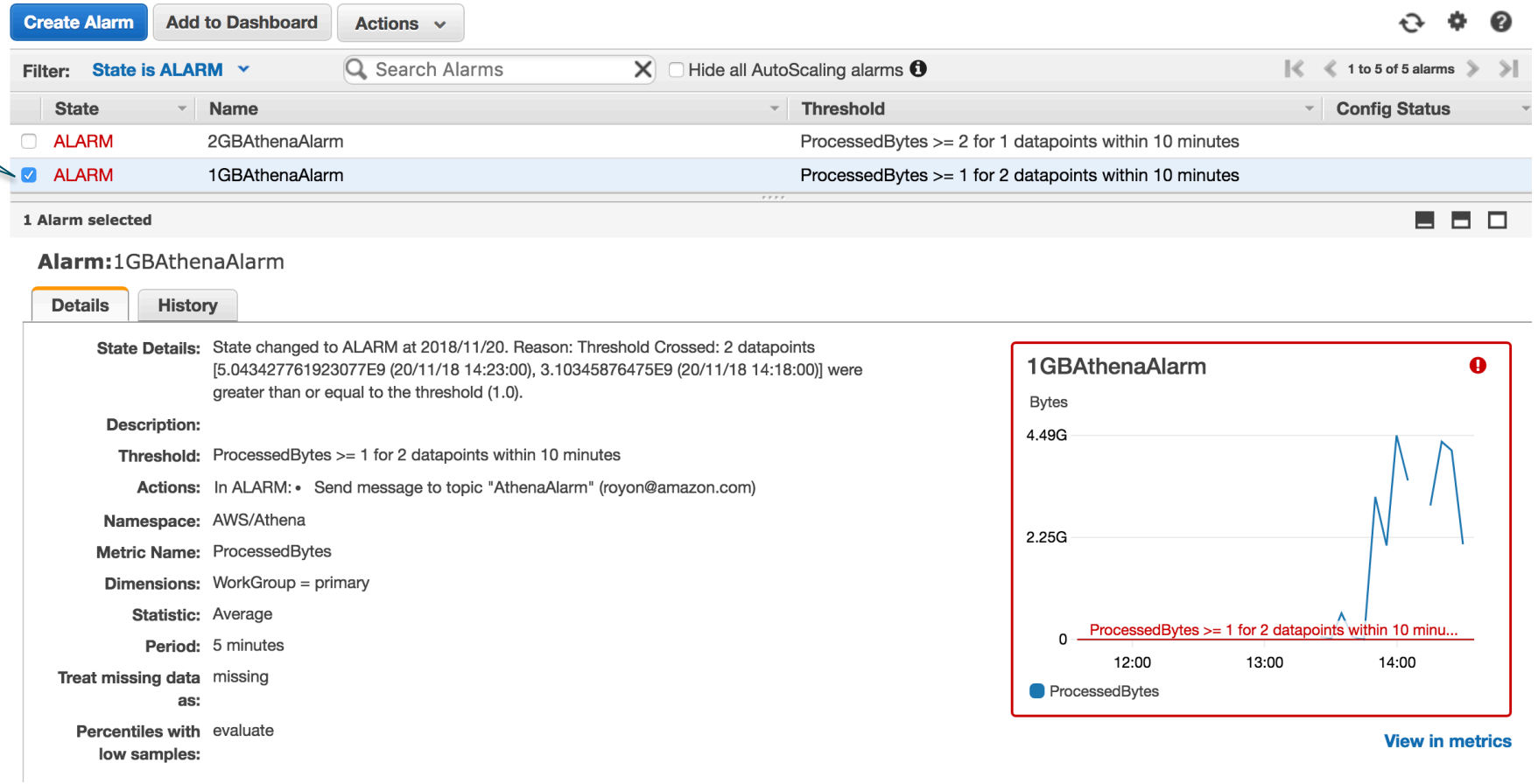
- Trigger alarms to notify of increasing usage and cost
- Disable Workgroup when all queries exceed a maximum threshold

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



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"
    ]
  }]
}
```


aws

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 peninsulam amoenam circumspice')
```

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

Visualizing the Data Lake

Working with Amazon QuickSight

aws

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

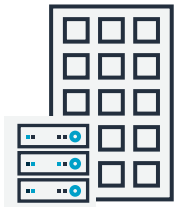
Embed in applications and enable analytics in hours, not months or years. Use themes to match application/corporate branding

Connect to your data, wherever it is

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

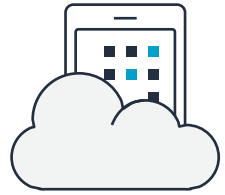
In the cloud

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

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

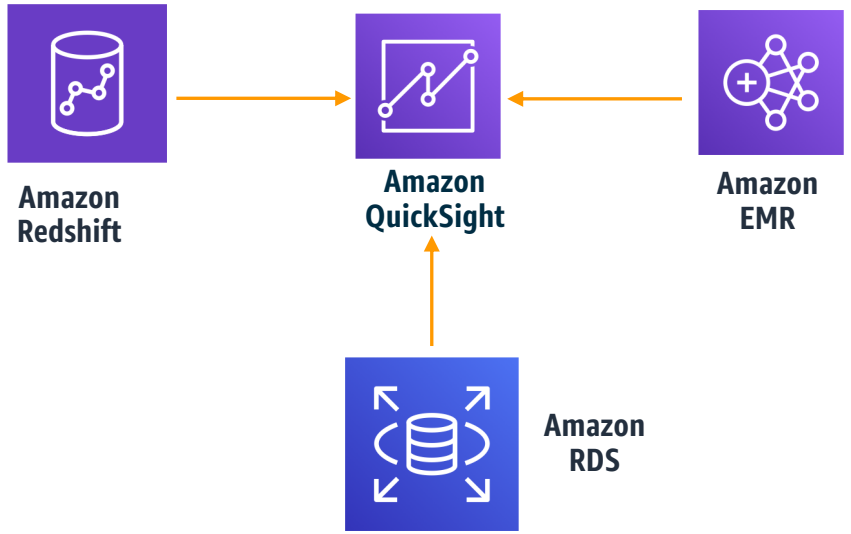
Applications

Connect directly to third party business applications



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

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

aws

Data Prep

Optional step when creating data sets:

Preview data

Rename, remove fields, change data types

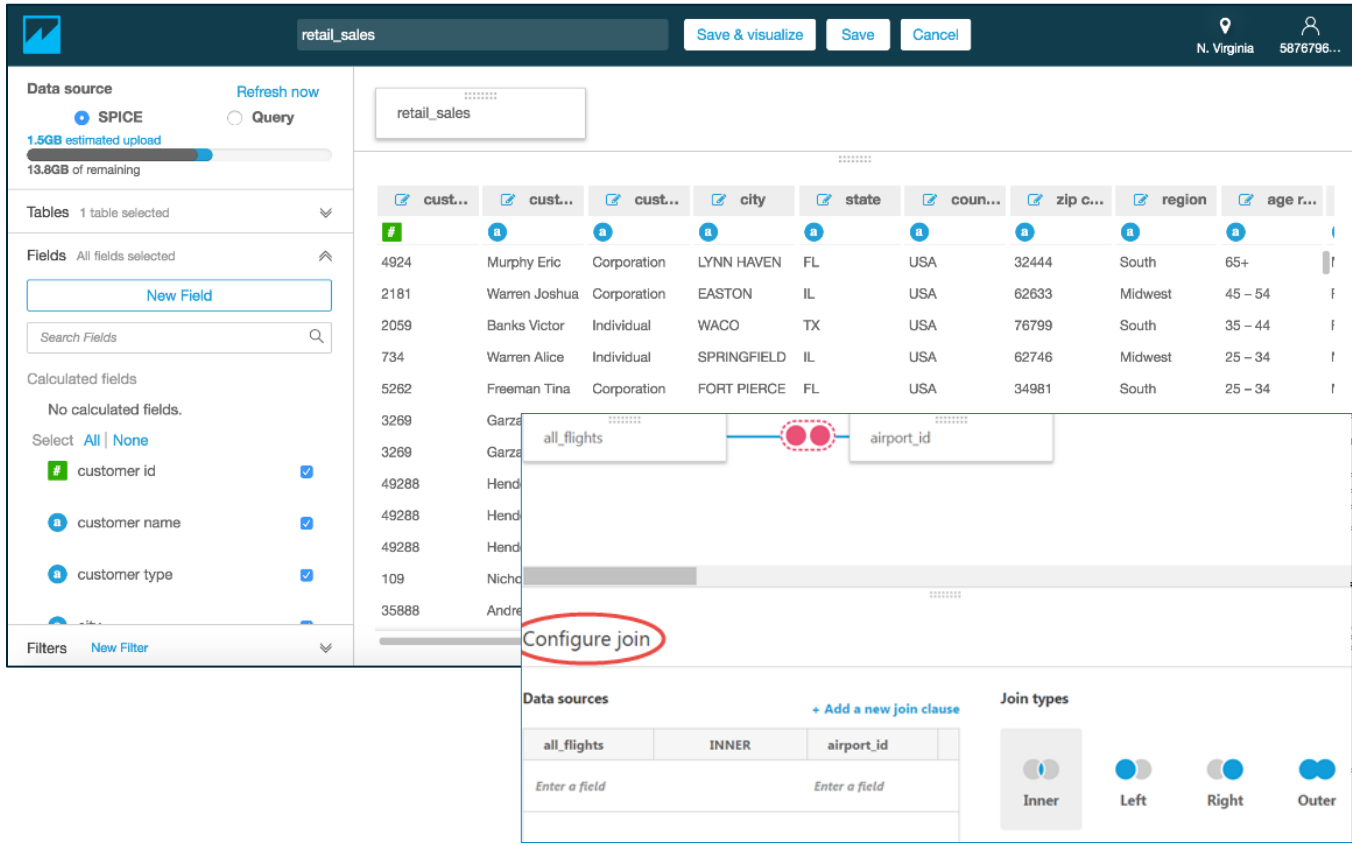
Create new calculated fields

Filter rows

Issue direct query or ingest to SPICE

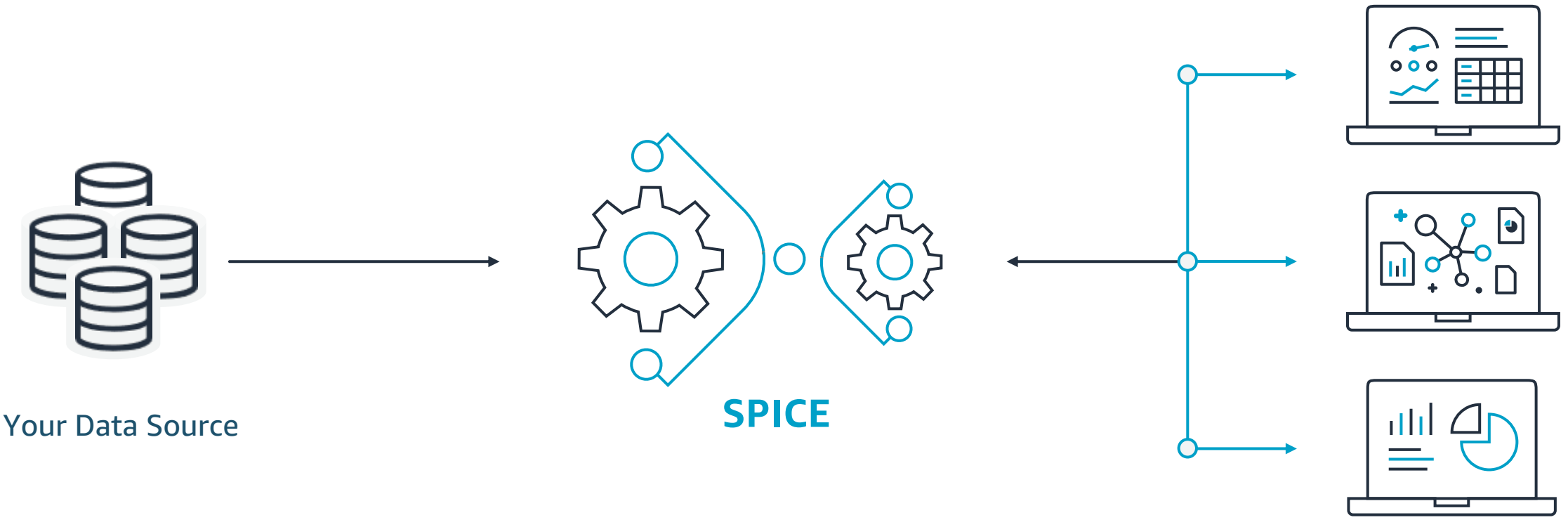
Visually join tables from the same relational DB

Push down custom SQL queries



SPICE

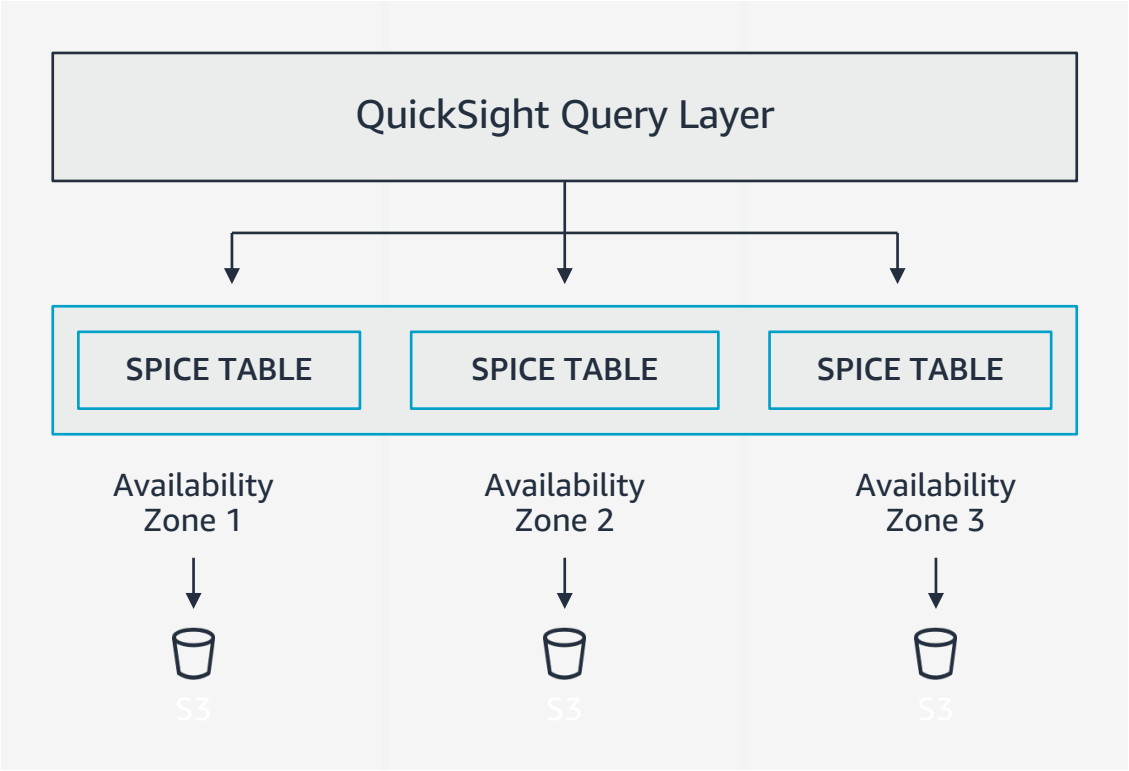
QuickSight is powered by SPICE, a super-fast calculation engine that delivers performance and scale, regardless of how many users are active.



aws

SPICE

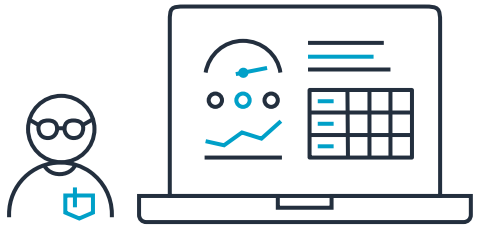
SPICE not only provides your users with instant response times, but automatically scales with user activity, protecting your underlying data sources saving you time and money.



- 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

aws

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/partners3rd

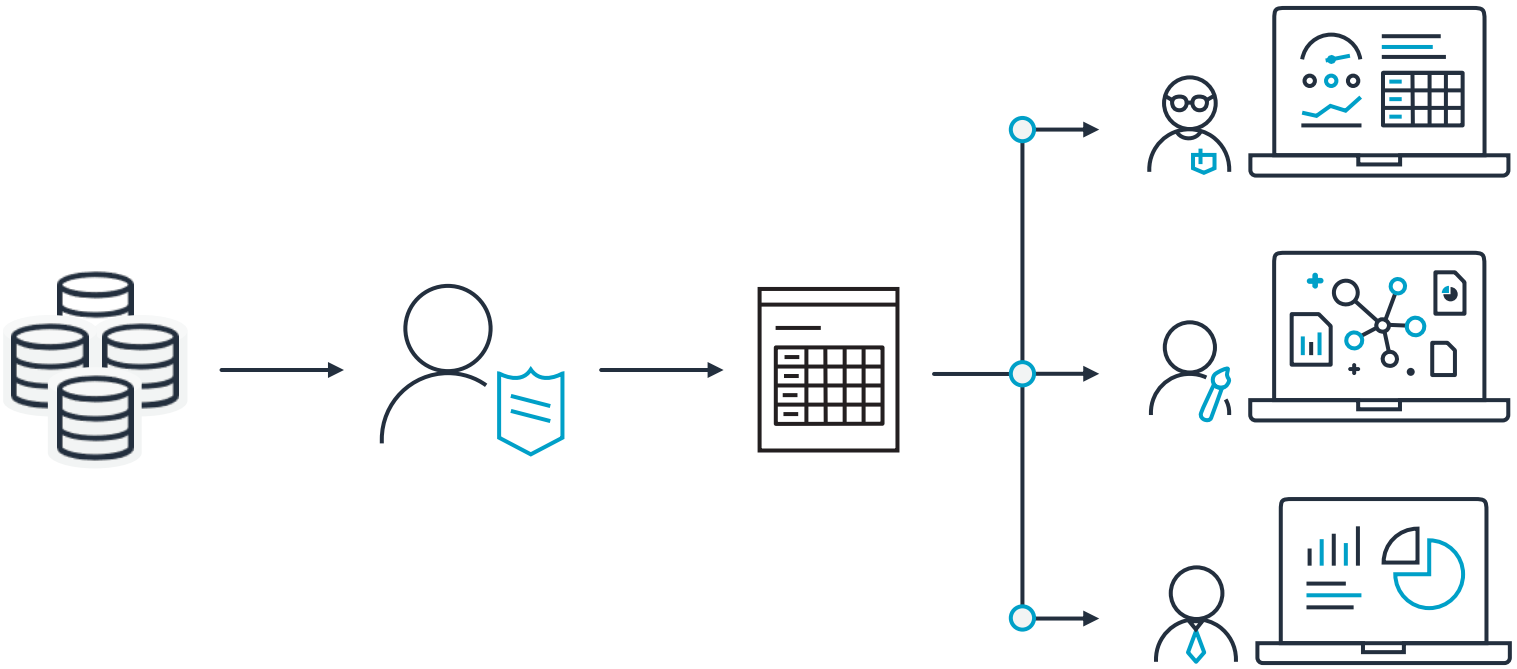
aws

Data governance

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

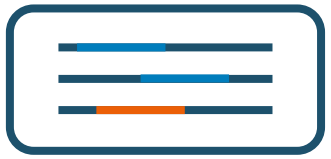
Create datasets that:

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



aws

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

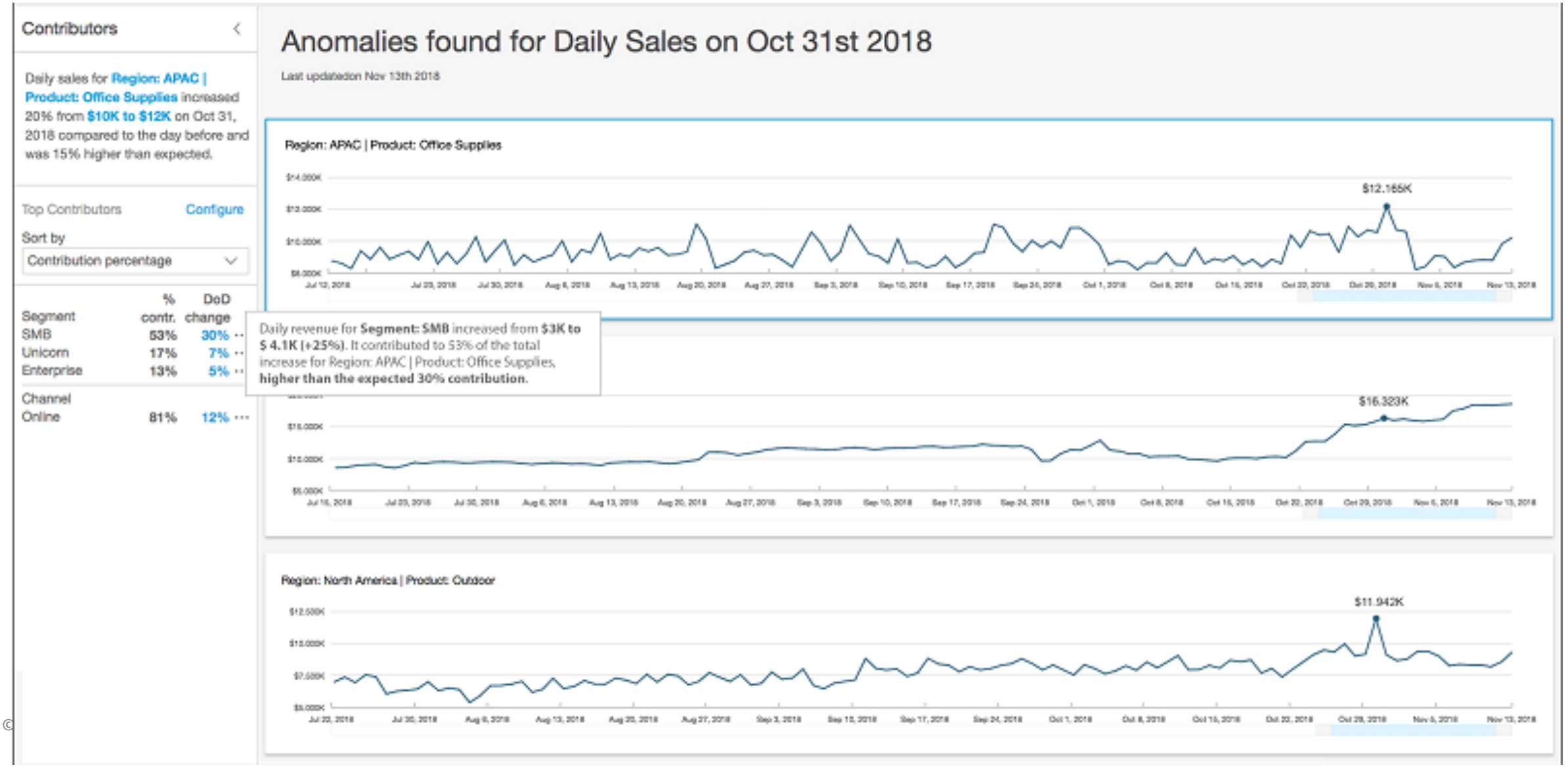
Discover unexpected trends and outliers against millions of business metrics

ML predictions

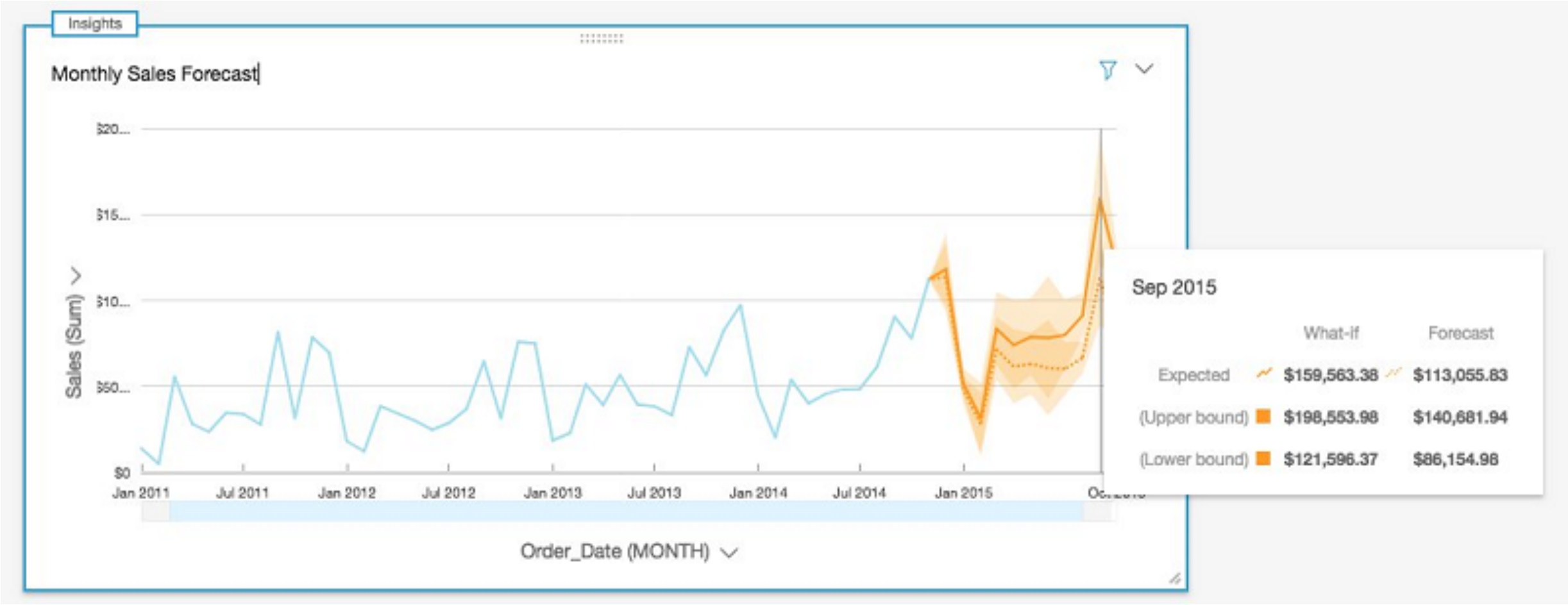
Visualize and build predictive dashboards with Amazon SageMaker models

aws

Anomaly Detection



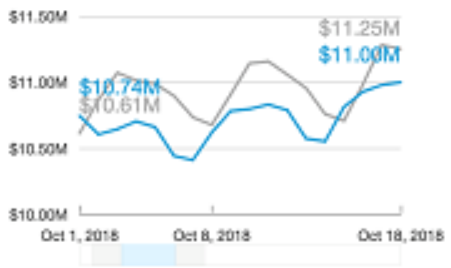
Forecasting



Natural Language Narratives

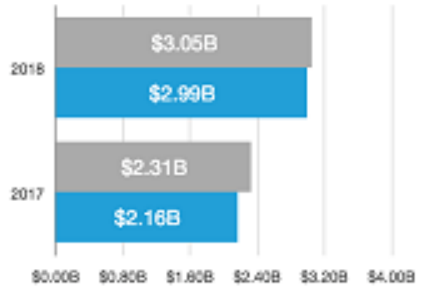
Daily Revenue

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.

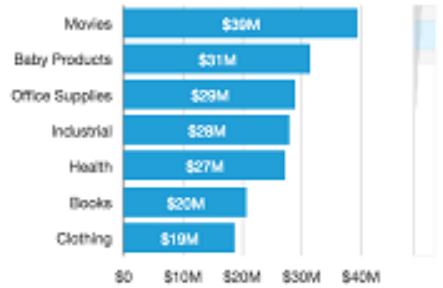


Callouts By Product and Country

Daily revenue for **Baby Products | Russia** on Nov 22, 2018 was **lower than expected** at $1,170.58.

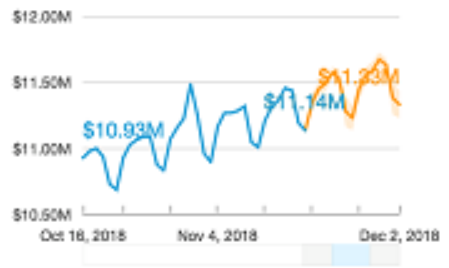
Callouts by Customers

Daily revenue for **MULTIDEL INC.** on Nov 22, 2018 was **higher than expected** at $60,433.95.



Daily Revenue Forecast

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.



Top / Bottom Movers by Product

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

- 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.

Top / Bottom to Plan Variance by Product

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

- 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.**

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

- Digital is **-$236,951.70 below goal.**
- Health is **-$147,834.66 below goal.**

Revenue by Product Category

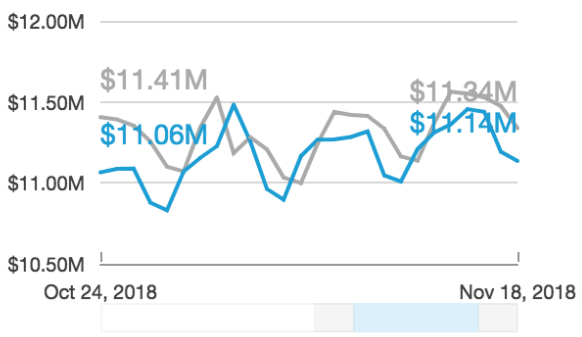
Product Ca...	Nov 18, 2018 Revenue	Nov 17, 2018 Revenue
Arts	$4,988.40	$4,988.18
Automotive	$52,309.00	$52,493.34
Baby Product	$1,354,243.11	$1,368,901.33
Beauty	$8,114.71	$8,116.06
Books	$1,330,700.80	$1,331,300.71
Business	$38,736.02	$41,916.59
Clothing	$770,046.75	$769,561.89
Collectibles	$709.44	$803.63
Computers	$539,996.53	$540,576.75

ML insights

ABCO Daily Sales Report

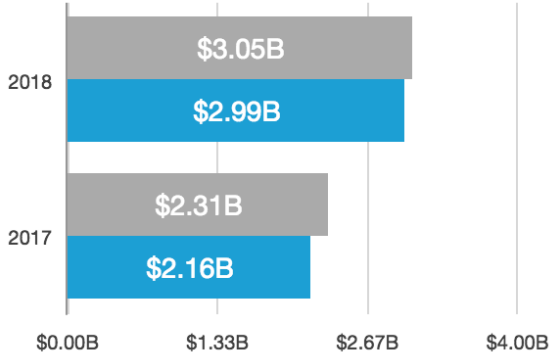
Daily Revenue

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.

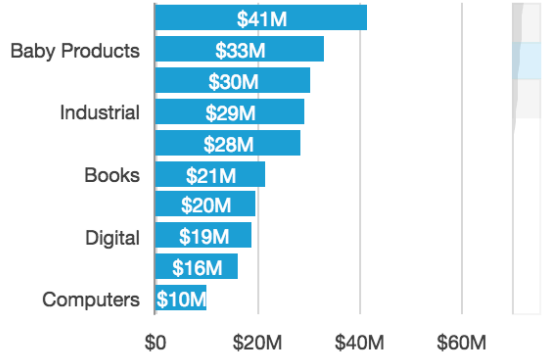


Callouts By Product and Country

Daily revenue for **Baby Products | Russia** on Nov 22, 2018 was **lower than expected** at $1,170.58

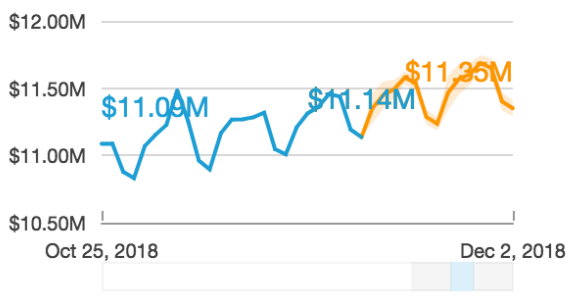
Callouts By Customers

Daily revenue for **MULTIDEL INC.** on Nov 22, 2018 was **higher than expected** at $60,433.95



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

aws

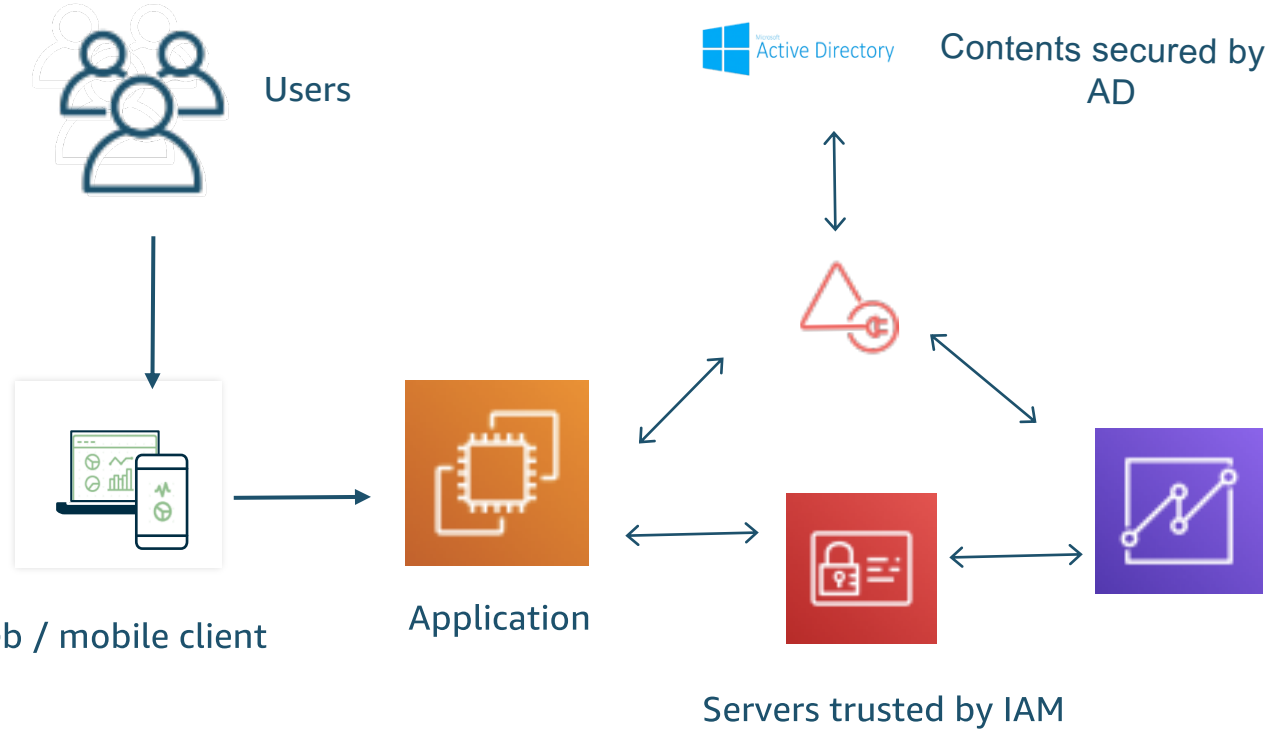
Why Amazon QuickSight for embedded dashboards

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

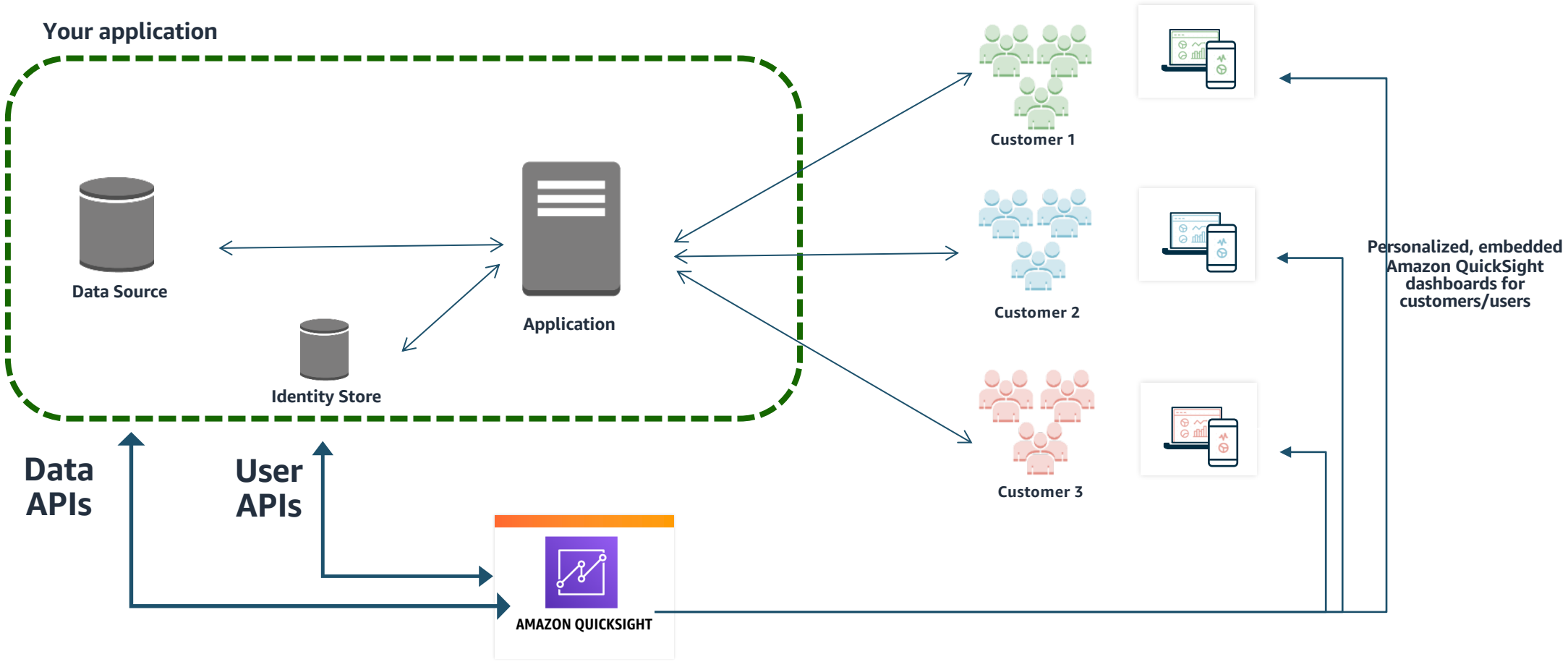
Create templates from analysis
Metadata for dashboard with placeholder for datasets
Templates accessible via API only
Can be copied/referenced across accounts

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

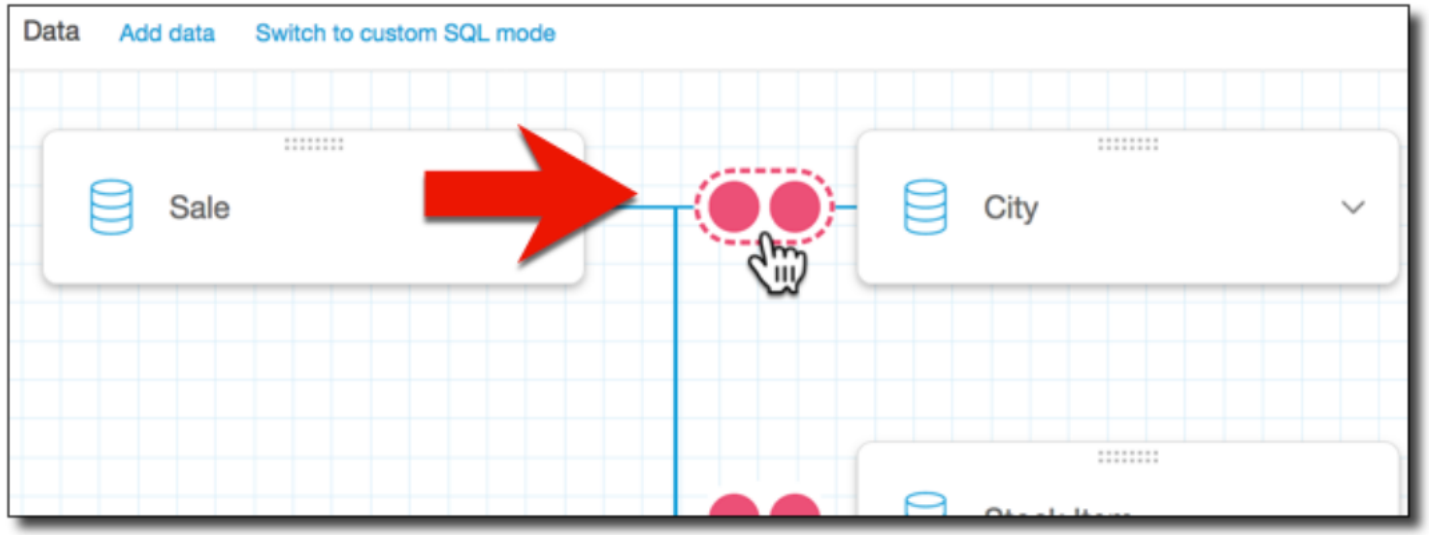
aws

Using APIs for your embedded deployments



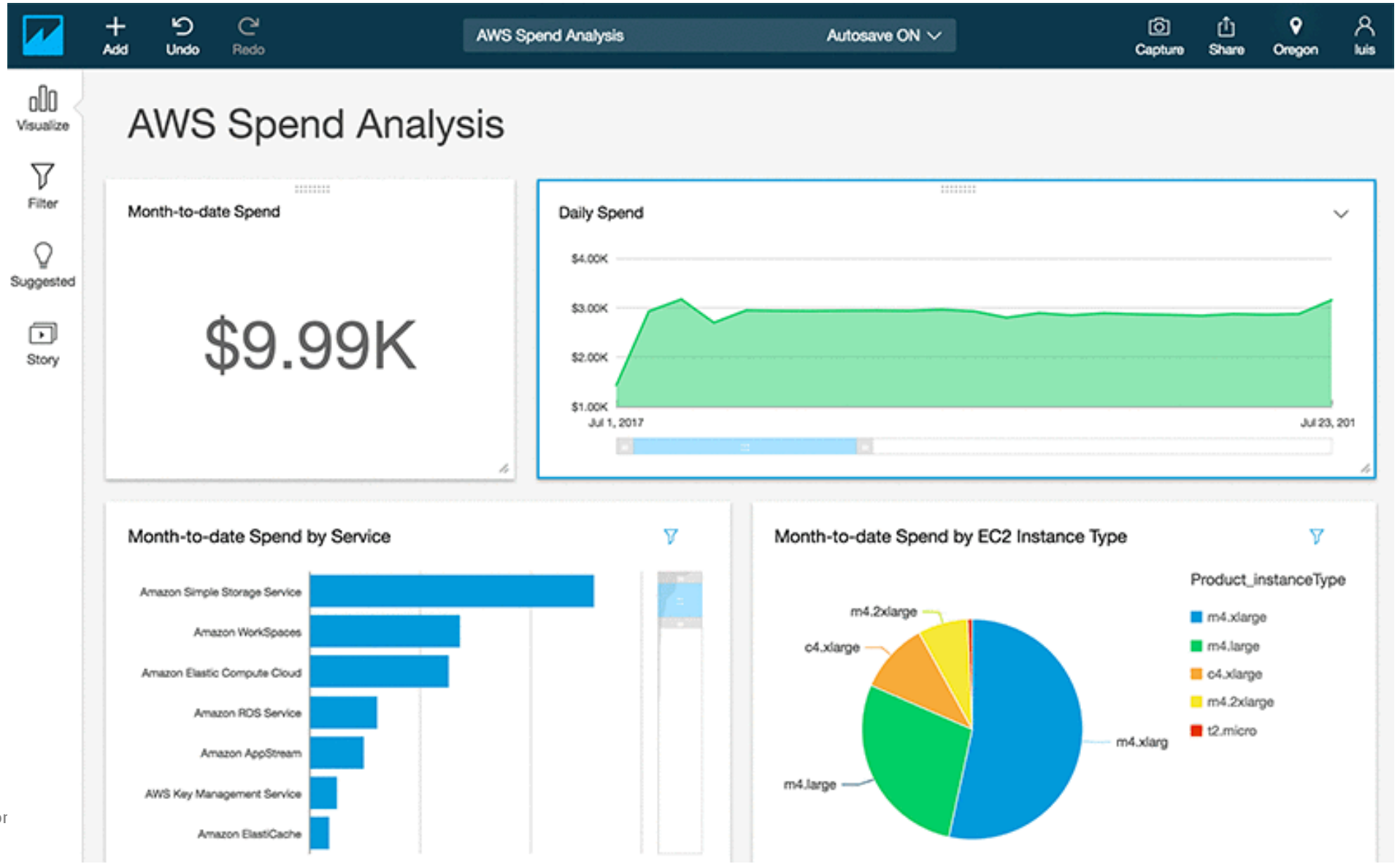
QuickSight Support for Cross Data Source Join

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

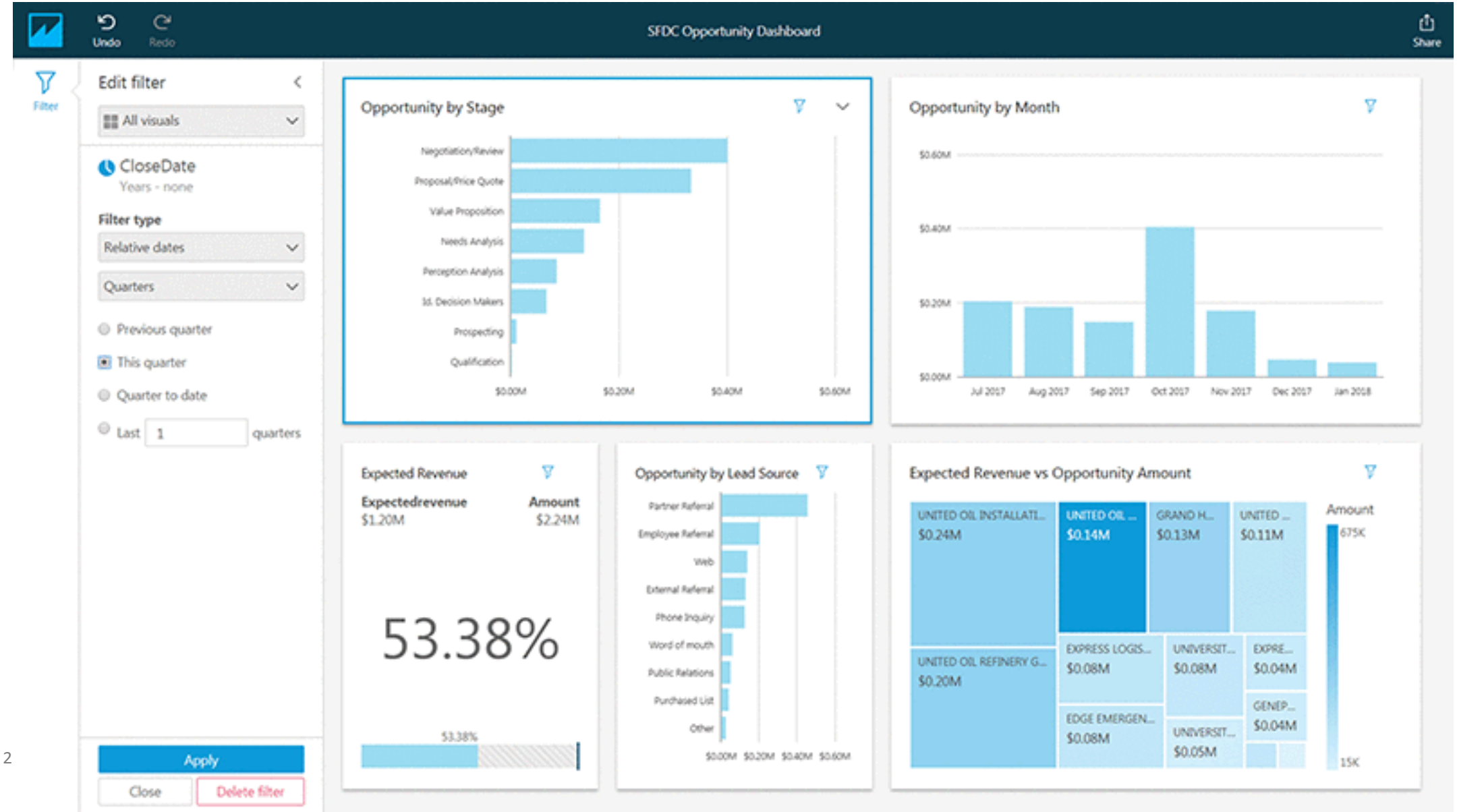


aws

Amazon QuickSight: Examples – AWS Cost & Usage Reporting



Amazon QuickSight: Examples – Salesforce Analytics



Q & A