Getting Started with Amazon Redshift

Greg Khairallah, Director of Analytics Specialists
Harshida Patel, Data Warehouse Specialist SA

aws

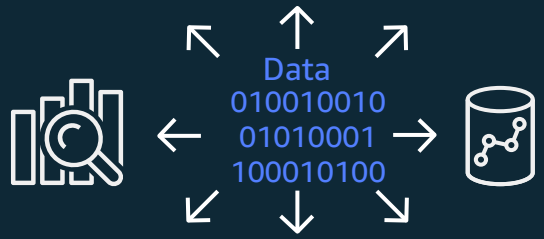
Agenda

- Challenges of data analytics at scale
- Amazon Redshift
    - Introduction and benefits
    - How to get started
    - Scalability
    - Security
- Additional resources
- Q&A

aws

Data warehousing trends

Migrations to the cloud



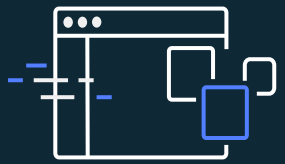
Exponential growth of event data

End-to-end insights from analyzing all your data

aws

Challenges of data analytics at scale

Data volume, variety, velocity



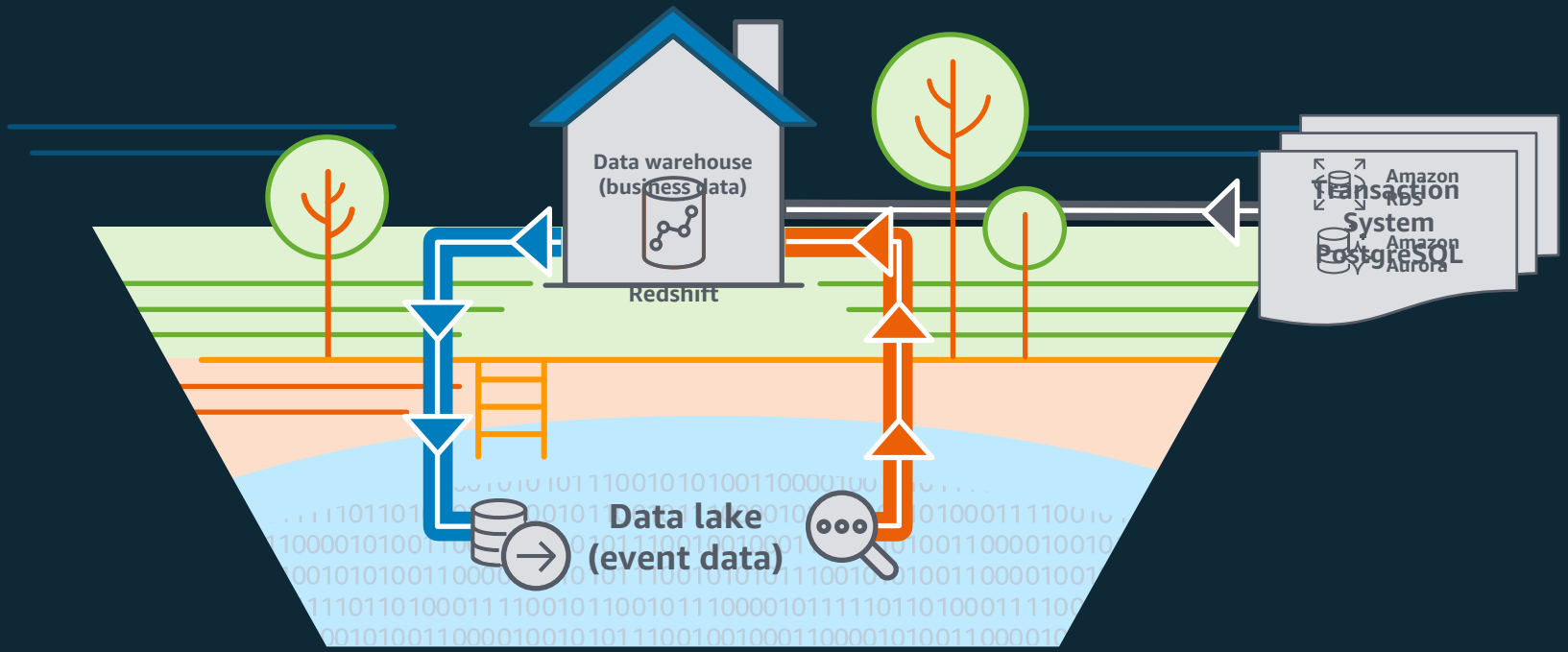
Performance, concurrency

Multiple analytics needs

Security, governance

Increasingly costly, inflexible

aws



Amazon Redshift enables you to have

a real-time lake house approach

aws

Amazon Redshift

is the most popular cloud data warehouse

21st Century Fox
23andMe
amazon
AMGEN
ancestry
AstraZeneca
CIVIS
coinbase
coursera
DOW JONES
edmunds
EA
EQUINOX
FT Financial Times
FOURSQUARE
HEARST
iflix
intuit
Jack in the box
Kimberly-Clark
Klarna
Liberty Mutual
London Stock Exchange
MANTHAN
MERCK
Nasdaq
The Ohio State University
OPTUM
PADDYPOWER
betfair
Pfizer
Prezi
QANTAS
REDFIN
robinhood
SCHOLASTIC
Schumachergroup
SCOPELY
SOUNDCLOUD
Sysco
tinder
yelp

aws

Amazon Redshift benefits

10’s of thousands of customers use Redshift & process EB of data per day

Data lake & AWS integrated

Lake Formation catalog & security, Exabyte scale query (spectrum & federated), AWS integrated (DMS, CloudWatch)

Best performance

Up to 3x faster than other cloud data warehouses

Lowest cost

Up to 75% less than other cloud data warehouses & predictable costs

Most scalable

Virtually unlimited elastic linear scaling

Most secure & compliant

AWS-grade security, (e.g. VPC, encryption with KMS, Cloud Trail), Certifications such as SOC, PCI, DSS, ISO, FedRAMP, HIPAA

Easy to manage

Easy to provision & manage, automated backups, AWS support, 99.9% SLAs

Amazon Redshift has been innovating quickly

Robust result set caching

Large # of tables support ~20000

Copy command support for ORC, Parquet

IAM role chaining

Elastic resize

Groups

Redshift Spectrum: date formats, scalar json and ION file formats support, region expansion, predicate filtering

Auto analyze

Health and performance monitoring w/Amazon Cloud watch

Automatic table distribution style

Cloud watch support for WLM queues

Performance enhancements—hash join, vacuum, window functions, resize ops, aggregations, console, union all, efficient compile code cache

RA3

Auto WLM

~25 Query Monitoring Rules (QMR) support

AQUA

Pause & Resume

Concurrency Scaling

200+

DC1 migration to DC2

Resiliency of ROLLBACK processing

Manage multi-part query in AWS console

Auto analyze for incremental changes on table

new features & enhancement in the past 18 months

Spectrum Request Accelerator

Apply new distribution key

Redshift Spectrum: Row group filtering in Parquet and ORC, Nested data support, Enhanced VPC Routing, Multiple partitions

Performance: Bloom filters in joins, complex queries that create internal table, communication layer

Faster Classic resize with optimized data transfer protocol

Column level access control with AWS lake formation

Spatial Processing

Amazon Lake Formation integration

Auto-Vacuum sort, Auto-Analyze and Auto Table Sort

Auto WLM with query priorities

Snapshot scheduler

Performance: join pushdowns to subquery, mixed workloads temporary tables, rank functions, null handling in join, single row insert

Advisor recommendations for distribution keys

AZ64 compression encoding

Console redesign

Stored procedures

Performance of Inter-Region Snapshot Transfers

Federated Query

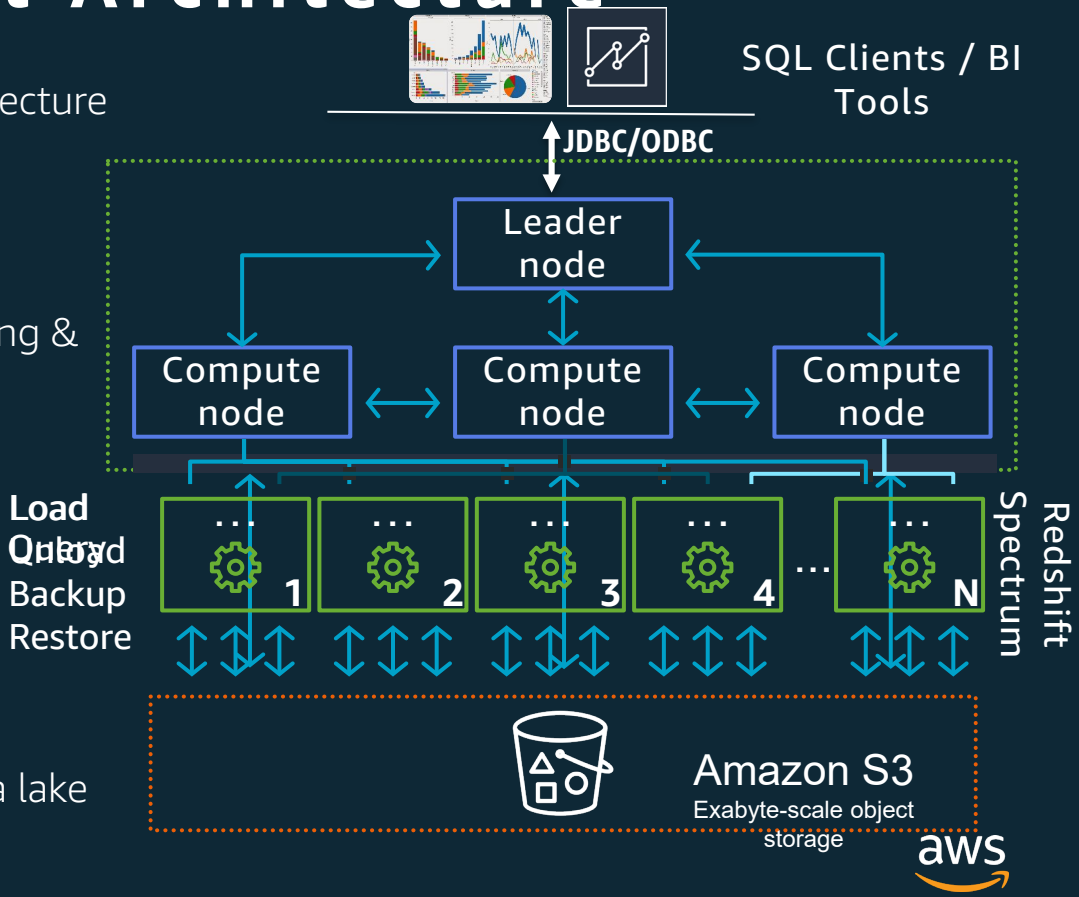
Materialized Views

aws

Amazon Redshift Architecture

Massively parallel, shared nothing architecture

- Leader node
  - SQL endpoint
  - Stores metadata
  - Coordinates parallel SQL processing & ML optimizations
- Compute nodes
  - Local, columnar storage
  - Executes queries in parallel
  - Load, unload, backup, restore from S3
- Amazon Redshift Spectrum nodes
  Execute queries directly against data lake



Amazon Redshift turbo charges query performance with machine learning based automatic optimizations

- Automates table maintenance
- Optimizes for peak performance as data and workloads scale
- Leverages machine learning
- Offers prescriptive recommendations with ability to apply changes dynamically

Automatic Analyze

Automatic Table Distribution Style

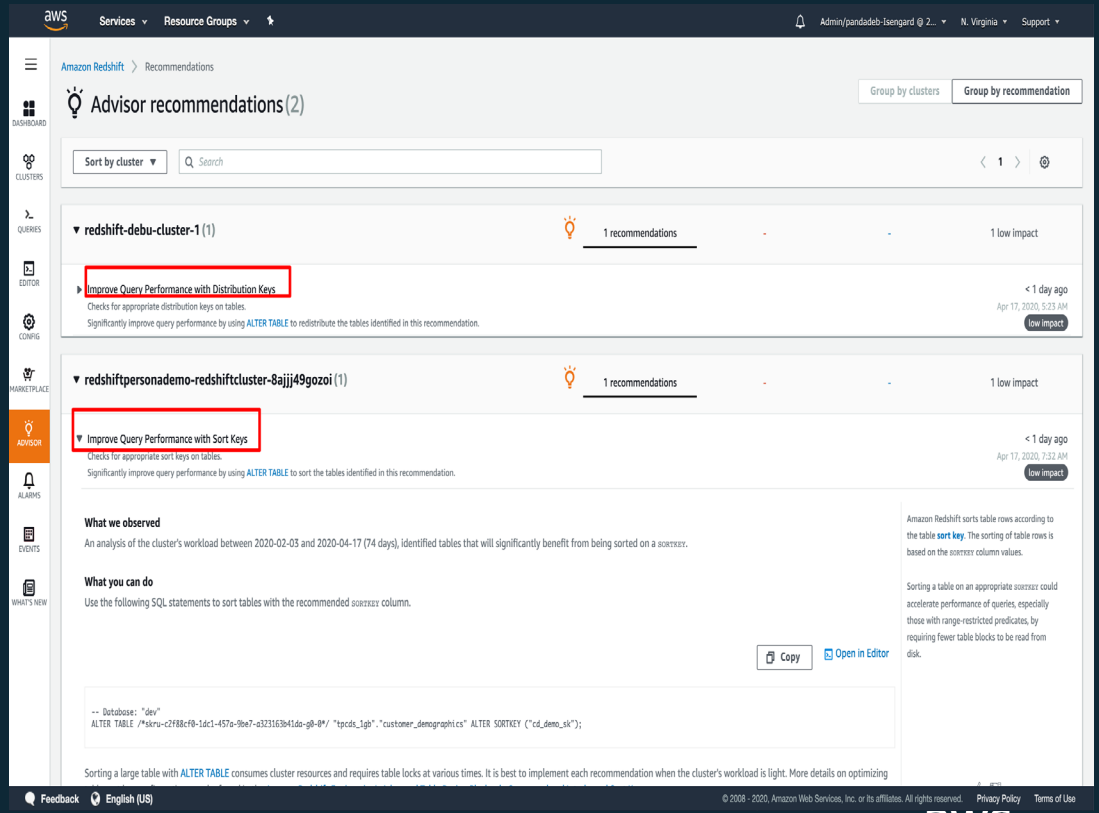
Distribution/Sort Key Advisors

Automatic Vacuum Delete

Automatic Table Sort

Amazon Redshift Advisor

- Offers specific, customized recommendations to improve the performance and decrease the operating costs for your cluster
- Advisor bases its recommendations on machine-learning observations regarding performance statistics or operations data



DEMO – Creation of a new Data Warehouse with Amazon Redshift

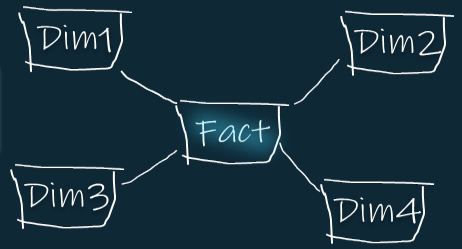
Data Modelling & Data Types

> **Murat Kader**
> **Modacruz, Interim Chief Technology Officer**
>
> Creating an Amazon Redshift instance allows us to segment user types based on what they buy and sell, so we can personalize their experience. This is an important feature that we're able to provide using AWS

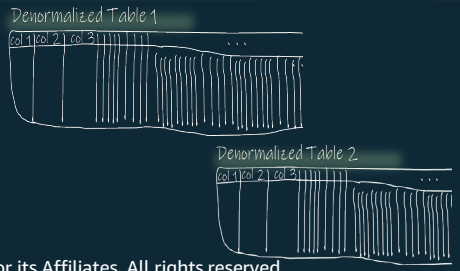
Amazon Redshift: Popular data models

Redshift can be used with a number of data models including…

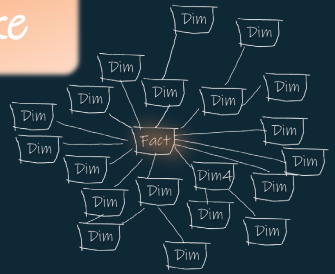
STAR



Highly Denormalized



Snowflake



Amazon Redshift data types

NUMERIC

- SMALLINT - signed two-byte integer
- INTEGER - signed four-byte integer
- BIGINT - signed eight-byte integer
- DECIMAL - exact numeric of selectable precision
- REAL - single precision floating-point number
- DOUBLE PRECISION - double precision floating-point number

TEXT

- CHAR - fixed-length character string
- VARCHAR - variable-length character string with a user-defined limit

DATETIME

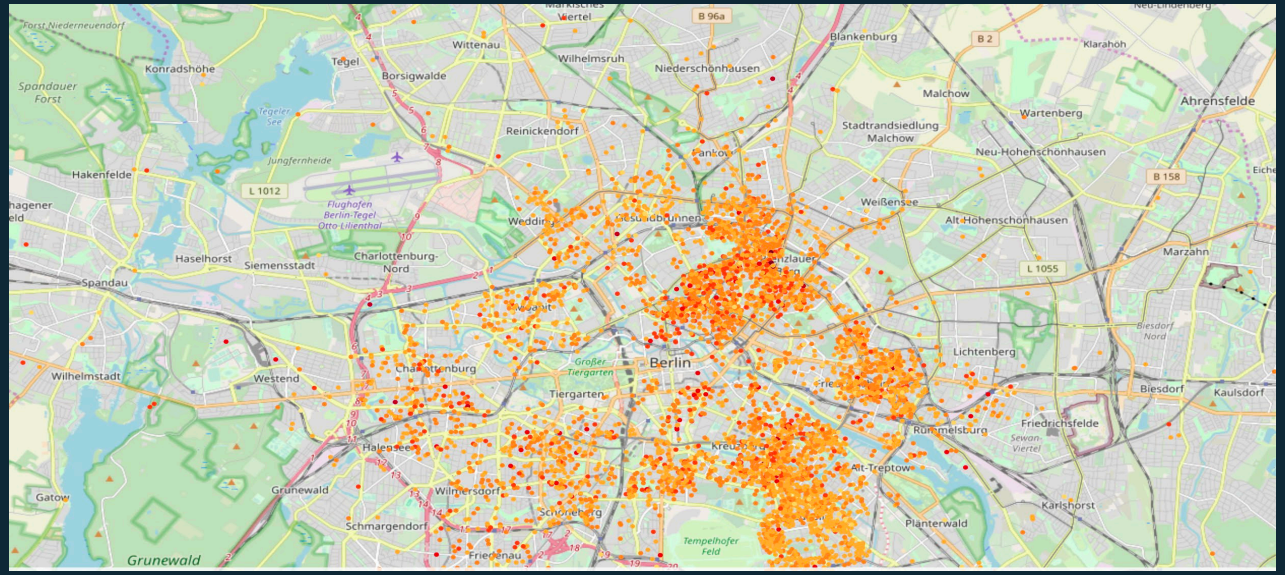
- DATE - calendar date (year, month, day)
- TIMESTAMP - date and time (without time zone)
- TIMESTAMPTZ - date and time (with time zone)

OTHER

- BOOLEAN - logical boolean (true/false)
- GEOMETERY – geospatial data (new)

aws

Spatial Processing - Sample Query




```
SELECT name, ST_X(shape) as lng, ST_Y(shape) as lat, price
FROM accommodations
WHERE ST_Within(shape, ST_GeomFromText( 'POLYGON((13.111839294433596
52.4285942596063, 13.111839294433596 52.60117089057946, 52.4285942596063))', 4326))
LIMIT 5000
```

Data Types

GEOMETRY
Point, Linestring, Polygon,
MultiPoint, MultiLinestring,
MultiPolygon,
GeometryCollection

Spatial Accessors

ST_NumGeometries,
ST_GeometryType,
ST_Dimension, …

Spatial Predicates

ST_Covers, ST_Equals,
ST_Within, ST_DWithin,…

Spatial Functions

ST_Distance,
ST_Azimuth, …

Spatial Formats

WKT/WKB, EWKT/EWKB, GeoJSON
Ingestion: CSV

Loading/Analyzing Data

```
select *
from
  data_warehouses

where
  sophisticated_query_optimization=true and
  scale_out_processing=true and
  super_fast_performance=true and
  support_for_open_formats=true and
  throughput_of_local_disk=true and
  scale_of_S3=true

====
(1) rows returned
```

Redshift

Data Loading: COPY Command

- COPY loads data into Redshift in parallel
- General Syntax

e.g. S3 object name or folder path

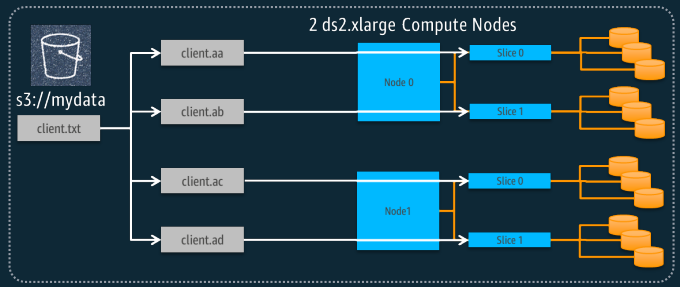
```
COPY <TABLE> from <location>
credentials
"aws_access_key_id=<access_key>;
aws_secret_access_key=<secret_key>"
iam_role "arn"
region;
```

- A number of COPY options can be specified, including
  - CSV, TXT, JSON, ORC, Parquet or Avro Data
  - Compression Options
  - Encryption Options
- Similar syntax for loading from DynamoDB, EMR, SSH

Data Loading

The COPY command typically outperforms single-row inserts, because it engages compute nodes to load data in parallel

```
copy customer from 's3://mydata/'
credentials 'aws_access_key_id=<your-access-key>;
aws_secret_access_key=<your_secret_key>'
delimiter '|';
```

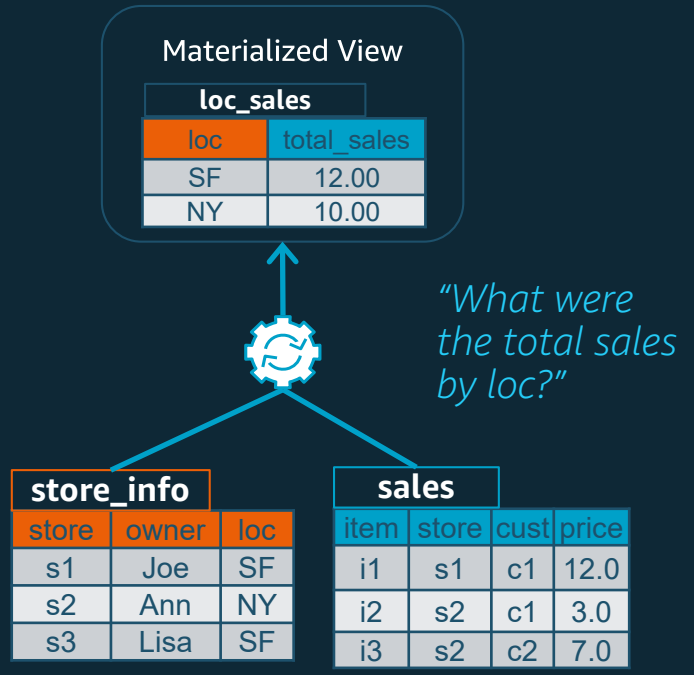


Amazon Redshift Materialized Views
Compute once, query many times

Speed-up queries by orders of magnitude
- Joins, filters, aggregations, and projections

Simplify and accelerate ETL/BI pipelines
- Incremental refresh
- User triggered maintenance

Easier and faster migration to Redshift



aws

Amazon Redshift Stored Procedures

Support to simplify migrations

Bring your existing **stored procedures** to Amazon Redshift

Migrating to Redshift is even easier!

Redshift supports Stored Procedures in PL/pgSQL format

Support for stored procedures provides the ability to run code where the data is to efficiently run ETL, data validation, and custom business logic.

```
CREATE OR REPLACE PROCEDURE test_spl(f1 int, f2 varchar)
AS $$
BEGIN
  RAISE INFO 'f1 = %, f2 = %', f1, f2;
END;
$$ LANGUAGE plpgsql;

call test_spl(5, 'abc');
INFO: f1 = 5, f2 = abc
CALL
```


aws

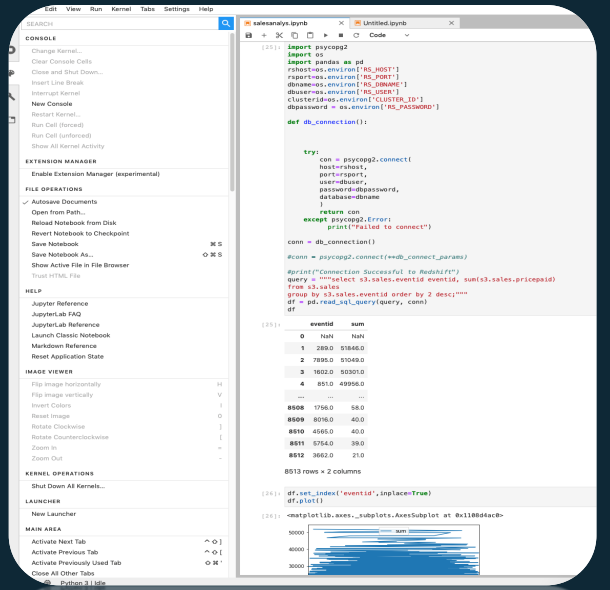
Seamless integration with AWS Services

Migration | ETL | Machine Learning | Data Visualization

- AWS Database Migration Service (DMS)
  Fully-managed migration service
- AWS Glue
  Fully managed extract, transform, and load (ETL) service
- Amazon EMR
  Cloud-native big data platform
- Amazon Sagemaker
  Fully managed service to build, train, and deploy machine learning (ML) models
- Amazon Quicksight
  Fast, cloud-powered business intelligence service

and more…

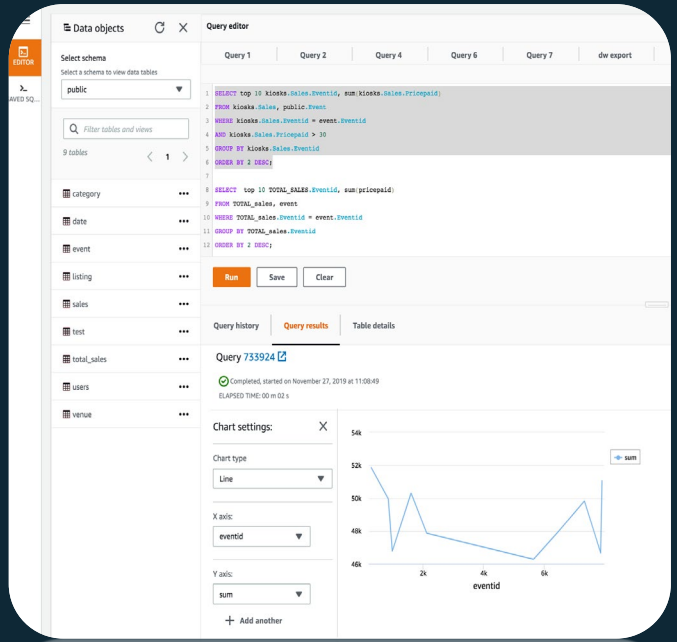
Sagemaker / Jupyter Notebook



DEMO

1. Connect using Amazon Redshift Query Editor
2. Create a table
3. Load data
4. Start Analysis

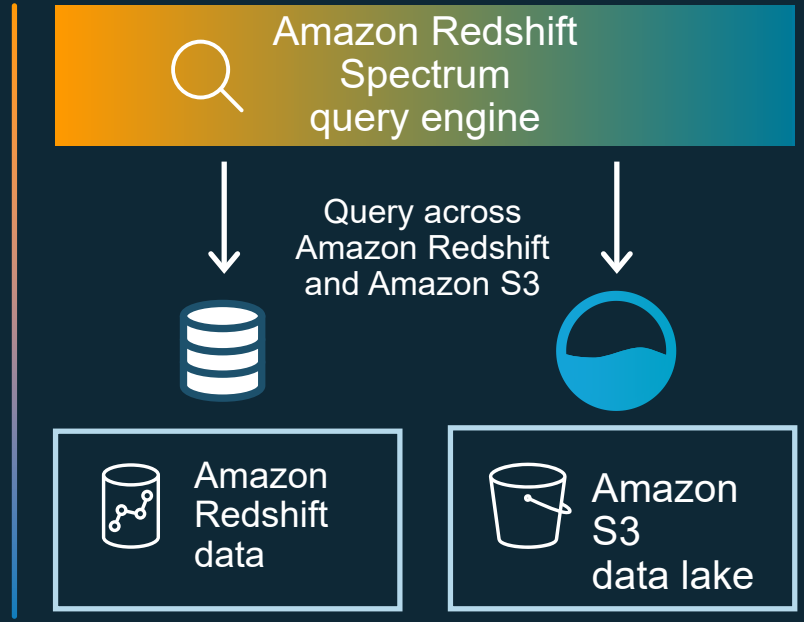
Amazon Redshift Query Editor



Amazon Redshift Spectrum

Extend the data warehouse to exabytes of data in Amazon S3 data lake

- No data loading required
- Scale compute and storage separately
- Directly query data stored in Amazon S3
- Parquet, ORC, Avro, JSON, and CSV data formats
- Spectrum Request Accelerator



aws

Amazon Redshift Federated Query

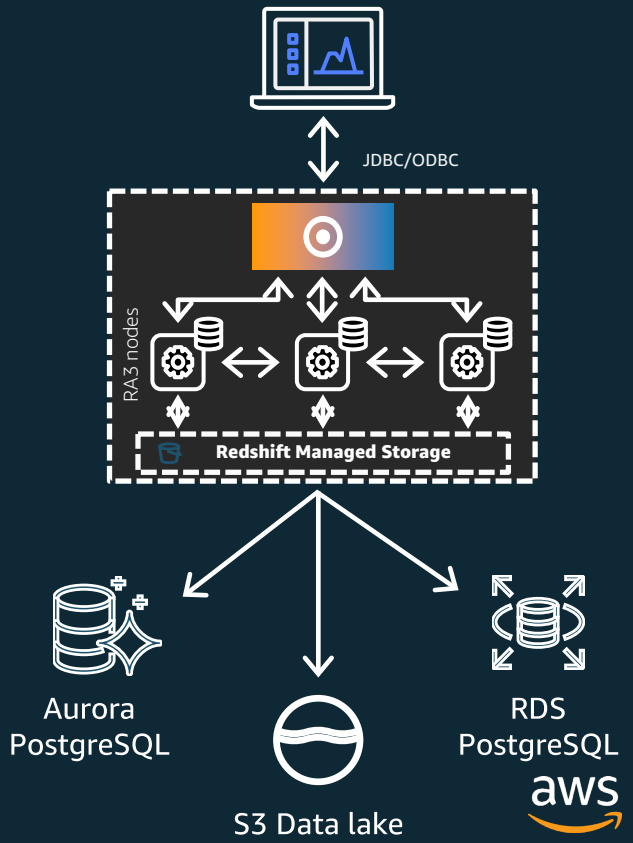
Queries on RDS and Aurora PostgreSQL databases

Analytics on live data without data movement

Unified analytics across data warehouse, data lake and operational databases

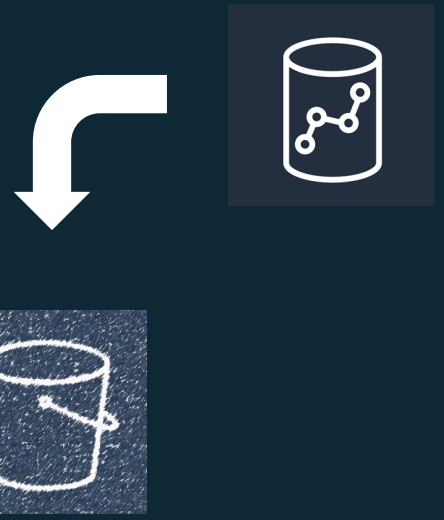
Flexible and easy way to ingest data

Performant and secure access to data



Unloading Data: UNLOAD Command

- `UNLOAD` command is the reverse of `COPY`, in that it outputs data from Amazon Redshift to S3
  - Runs from a `SELECT` statement. `Order By` clause respected by `UNLOAD` if `PARALLEL=OFF`
  - Encryption & compression handled automatically
  - Runs in parallel on all compute nodes
- `UNLOAD` output
  - CSV or Parquet (Data Lake Export) file formats
  - Generates > 1 file per slice for all compute nodes
  - Max file size written on S3 can be controlled (max internal limit 6.2GB)
  - Generates a manifest for all unloaded files (useful for `COPY` into another cluster)
  - Control if files can overwrite existing locations or not

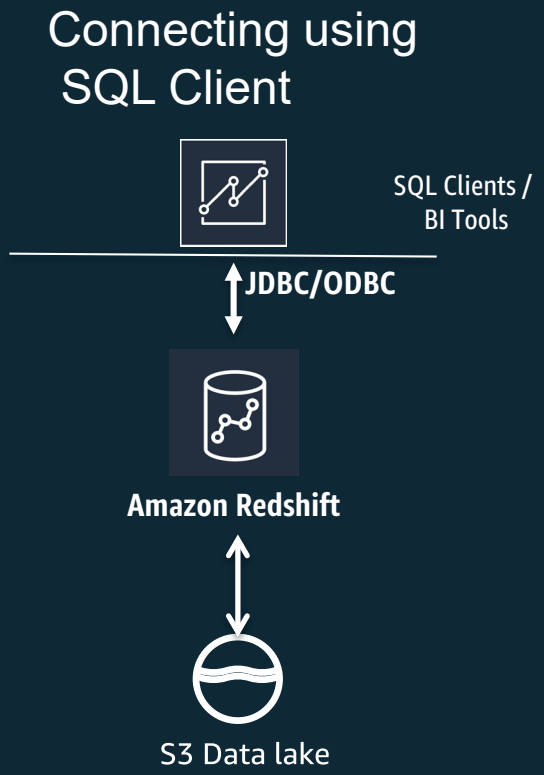


```
UNLOAD ('select-statement')
TO 's3://object-path/name-
prefix'
iam_role “arn” [ option [ ... ]
]
```

DEMO

Lake House architecture with Amazon Redshift Spectrum

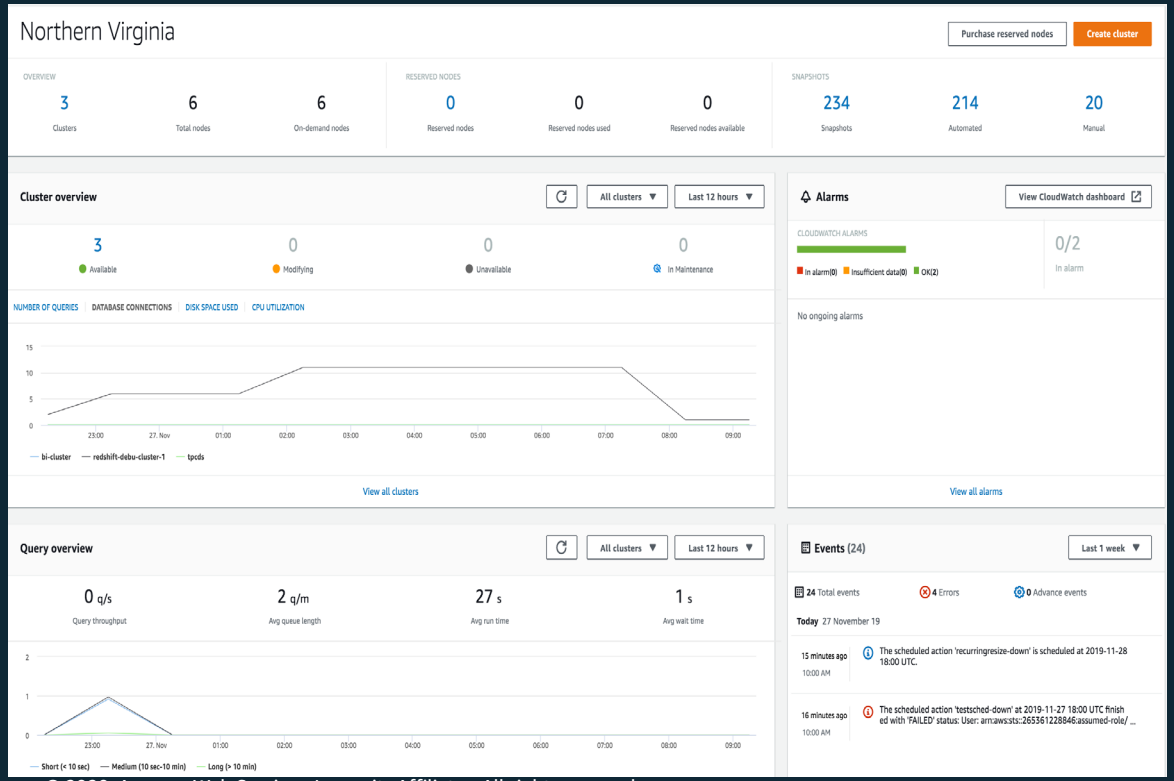
1. Query Product review data in parquet on Amazon S3
2. Analyze purchasing behavior by joining Product review data with Customer tables in Redshift
3. Unload analyzed data to S3 in parquet format



Monitoring

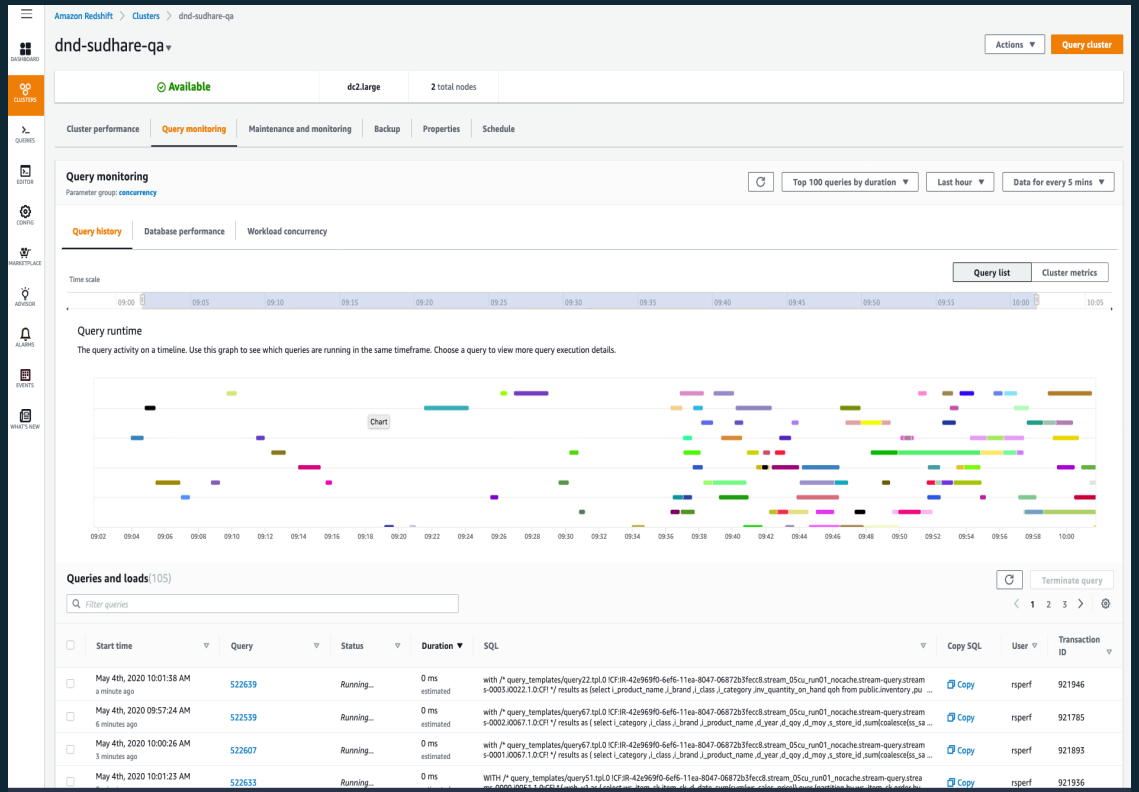
> “
>
> Justin Cunningham
> yelp.com, Business Analytics Tech Lead
>
> We have dozens of Redshift clusters that are owned by different teams. The ability to dynamically create multiple, dedicated clusters eliminates contention issues. Whenever a team wants to run an in-depth analysis on their data, they can do it without needing to consult with any other team. This also decouples development scaling. If one of our teams wants to start a new cluster and bring in a bunch of data, and maybe take it down for a while to facilitate the analysis, they can do that without interfering with other teams.
>
> ”

Monitoring clusters in Amazon Redshift



- Gain visibility to health of all clusters in your account
- Simplify creation and management of clusters
- Reduced time to diagnose user query performance issues
- Share Query Editor with non-admin users

Monitor Performance of User Queries



- Monitor your queries and loads
- Identify slow and failed queries
- Terminate run-away queries

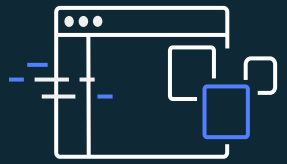
DEMO – Monitoring Your Workload

Scalability

Amazon Redshift serves all your Elastic needs

Compute

Elastic Resize

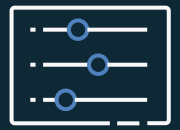


Data

Managed Storage

Users

Concurrency Scaling



Workloads

Queue, Workgroups

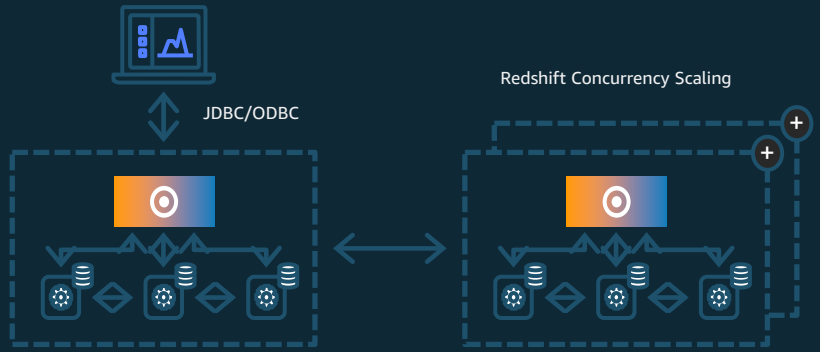
aws

Two forms of compute elasticity

	Horizontal scaling	Vertical scaling
Question	How can I speedup my running jobs?	How do I support spike in users without provisioning for peak demand?
Answer	Add more nodes with Elastic Resize	Enable concurrency scaling

aws

Compute elasticity & scalability with concurrency scaling



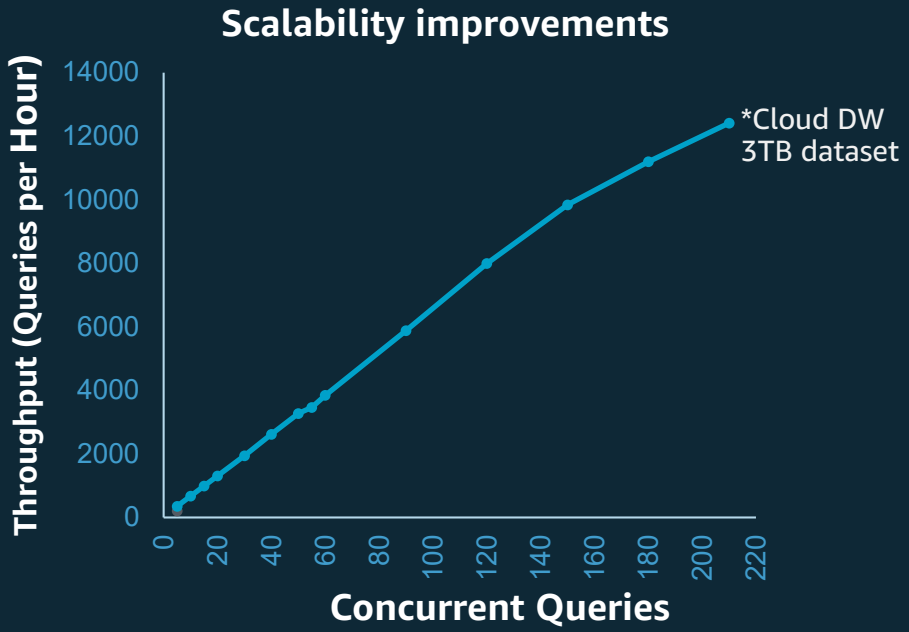
Scale-out to multiple Redshift clusters from a single endpoint in seconds

Support virtually unlimited concurrent users and queries while maintaining SLAs

Per-second billing for additional clusters used

Free 1hr usage per day (free for 97% of clusters)

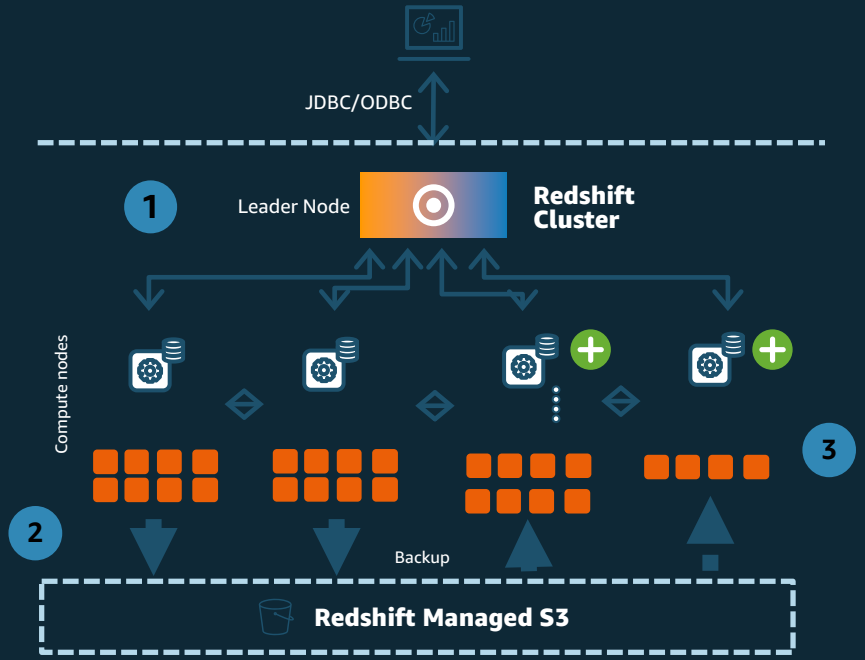
>35x improvement in throughput in 2019



* Cloud DW benchmark is based on TPC-DS (v2.10) with no query modifications are done

Elastic Resize : Scale within minutes

- In-place
  - Add or remove nodes to/from existing cluster
- Scale-Out
  - Performance scales proportionally
- Fast
  - Completes within few minutes
  - Limited disruption to sessions and queries



RA3 nodes with Redshift Managed storage

New!

Scale and pay for compute and storage independently

Managed storage

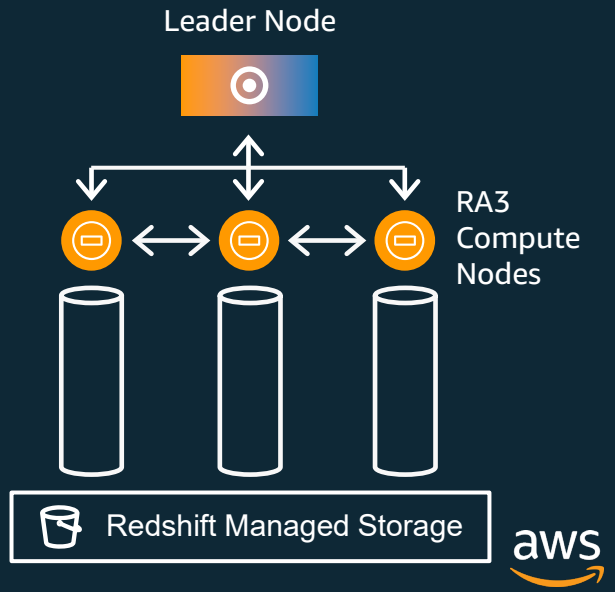
Large high-speed cache

High-bandwidth networking

Size data warehouse only based on steady state compute needs

Scale and pay independently for compute and storage

Automatic, no changes to any workflows, no need to manage storage



Amazon Redshift security is a top priority

AWS is designed to help you build secure, high-performing, resilient, and efficient infrastructure for your applications. Redshift includes the following:


End-to-end data encryption


IAM integration & integration with SAML IdP's for Federation (SSO)


Amazon VPC for network isolation


Database security model (users, groups, privileges)


Audit logging and notifications


Certifications that include SOC 1/2/3, PCI-DSS, FedRAMP, & HIPAA

aws

Next steps

Want more basic information about Redshift and data warehousing? Visit https://aws.amazon.com/redshift/

Ready for more advanced training on features and use cases?

- https://aws.amazon.com/blogs/big-data/amazon-redshift-at-reinvent-2019/
- https://aws.amazon.com/blogs/big-data/speed-up-your-elt-and-bi-queries-with-amazon-redshift-materialized-views/
- https://github.com/awslabs/amazon-redshift-utils/tree/master/src/CloudDataWarehouseBenchmark/Cloud-DWB-Derived-from-TPCDS

aws