Amazon Redshift

COSC 516 – Cloud Databases

Amazon Redshift

Amazon Redshift is a cloud based data warehouse system hosted on Amazon Web Services(AWS).

Redshift differs from other Amazon offerings such as ***RDS*** as Redshift is based solely on PostgresSQL.

Unlike other cloud based databases Redshifts main use case is data analytics as a service.

Different from a normal Database

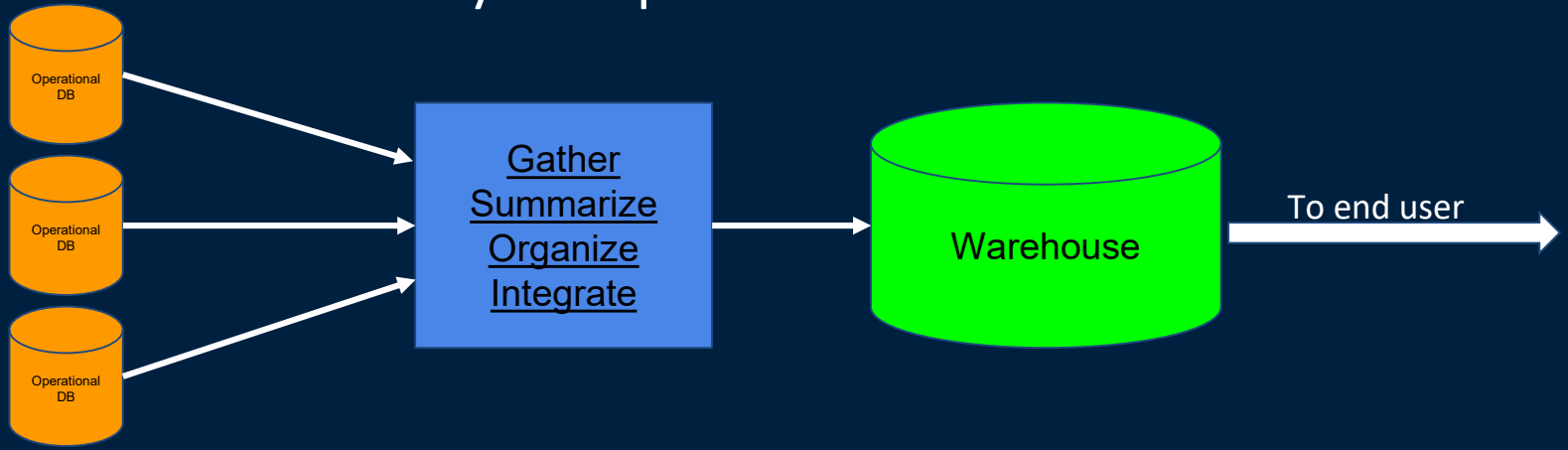
Normal databases are designed to handle large volumes of small transactions (Inserts, Updates, Deletes) and are referred to as *online transaction processing (OLTP)* systems.

Data warehouses are *online analytical processing (OLAP)* systems and handle small volumes of highly complex queries on large amounts of data and are used to support applications that analyze said data.

What is a Data Warehouse?

A data warehouses are databases that are used for data reporting and data analysis.

Do so by summarizing, organizing , and integrating data from one or more operational databases into a central repository in a format that is more efficient for analytical queries.



Queries on Redshift

Redshift is a relational data warehouse so querying is done using SQL.

Can query data on Redshift by using the Amazon query editor.

Connect to the Redshift cluster through a SQL tool such as SQL workbench.

Supports connection through SQL client tools such as JDBC or ODBC, there are used on client side much like RDS.

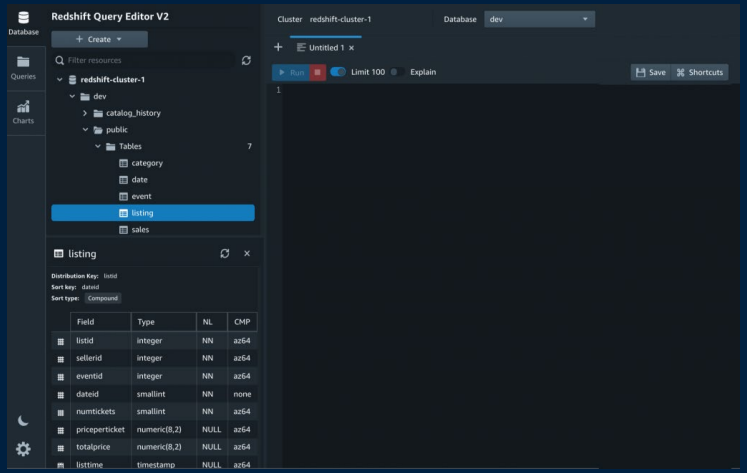
Queries on Redshift will be automatically optimized for best performance.

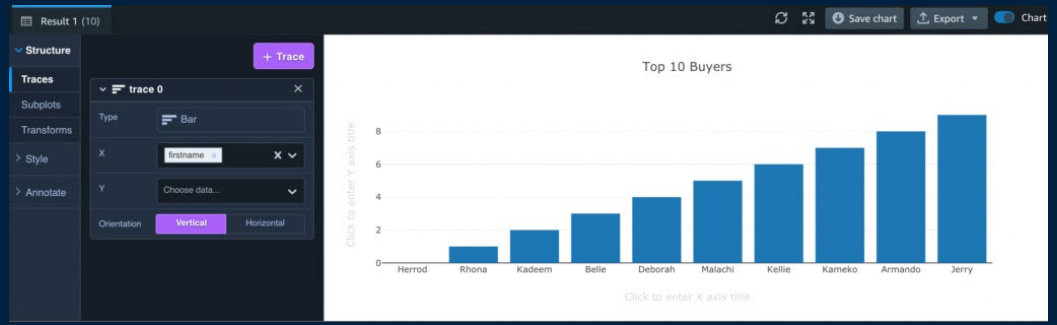
Query Editor

Web-based workbench to explore, query, analyze data using SQL

Enable the visualization of results with charts and graphs.

More in depth overview of query editor and it's setup here.





Scalability

Node Scaling: Called Elastic resize, can add or remove nodes to cluster. It is also possible to change the node type during this operation. Read-only queries still function.

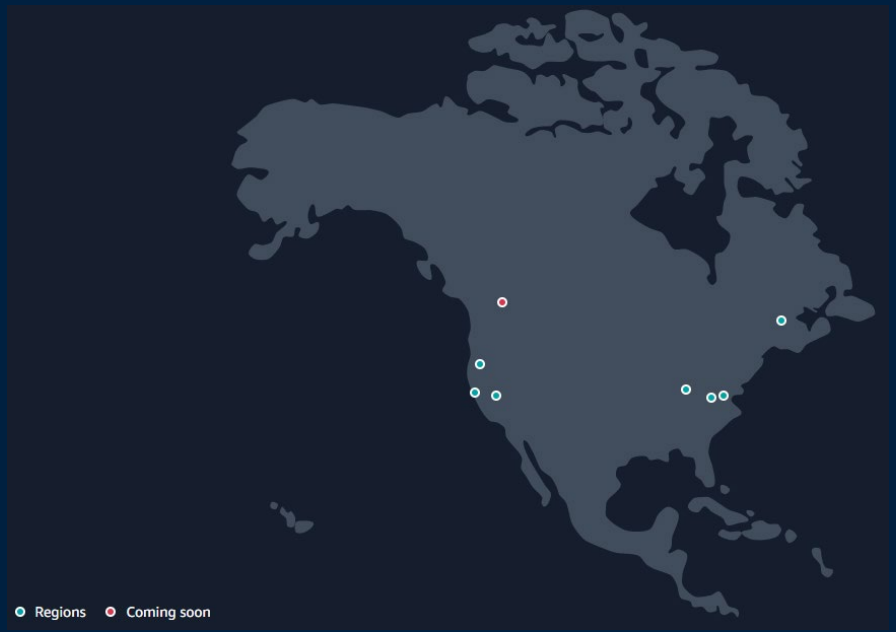
Cluster Addition: Can create new clusters of nodes to query on the same data by restoring the data from a Redshift snapshot.

Redshift Spectrum: Can switch to directly querying data stored on Amazon S3 storage.

Availability Regions and Zone

Amazon Redshift uses the same AWS infrastructure as RDS.

- Same regions and zone available to Redshift instances.



Region	Availability Zone
US West (Oregon)	4
US West (North Cal)	3
US East (North Virginia)	6
US East (Ohio)	3
Canada (Central)	3
Canada(West-Calgary)	coming soon

Security

Similar security to Amazon RDS with some unique to Redshift.

Cluster security groups: Grant users inbound access to an Amazon Redshift cluster on a per cluster bias.

Load data encryption: encrypt table load data files when uploaded to Amazon S3, you can use either server-side encryption or client-side encryption.

Data level access control: Can limit access to Column or Rows without having to implement view-based access control.

More Redshift security feature can be found here.

Instance types

On-demand: Provisioned capacity enabled for on demand computing by node type when needed.

Node types:

- Dense Compute DC2 (2-32 vCPU, 15-244 GB memory, 0.16-2.56 TB ssd).
- RA3 with Redshift Managed Storage (4-48 vCPU, 32-384 GB memory, 32-128TB Redshift Managed Storage (RMS).

Reserved: Functions similarly to On-demand except user reserves nodes for their continuous use. Uses same nodes and storage as On-demand, best for steady-state production workloads.

Redshift Spectrum: Directly run queries on data in long term storage (Amazon S3).

Redshift Serverless: Serverless instance option, no control over node type.

Serverless

Automatic scaling of serverless instance. Scales in units of Redshift processing units.

Allows users to access and analyze data without needing to set up a provision cluster.

No down time for maintenance.

Retention of serverless data through data snapshots with indefinite retention period as the default.

Allows for query exhaustion time management, automatically stops any query that goes for max allotted time.

Use Cases: Self-service Analytics, Auto Scaling.

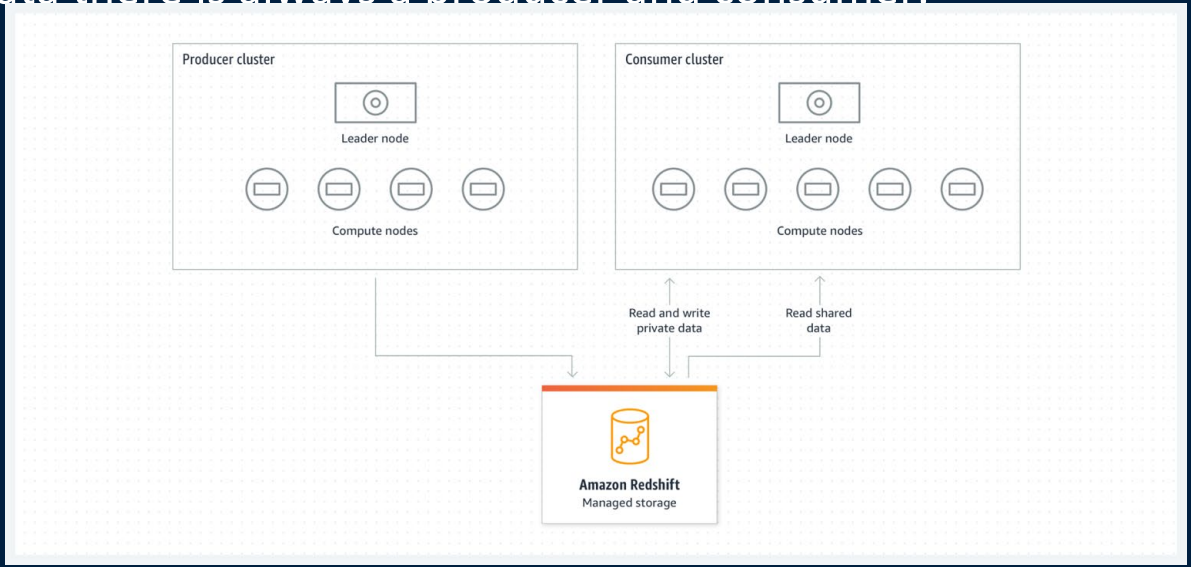
Data sharing

Redshift allows for data sharing between clusters without copying data.

Data shared across clusters is live up-date data and is consistent between the clusters.

Also allows data to be shared across AWS account.

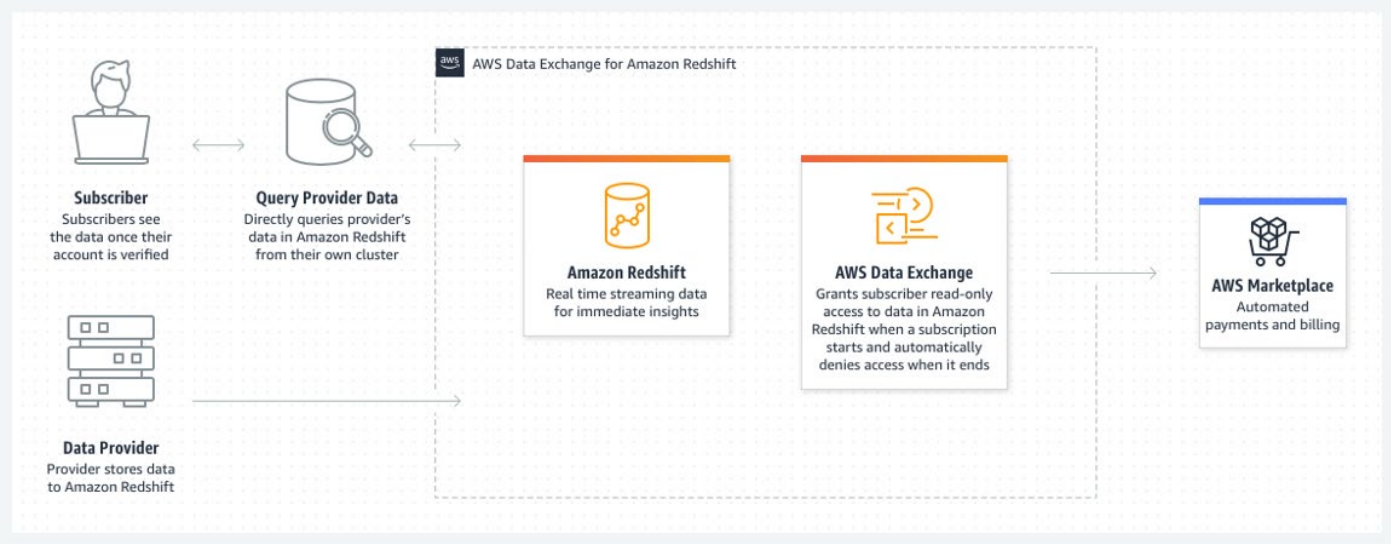
When sharing data there is always a producer and consumer.



Data Sharing: Exchange

Redshift allows for third-party data to be access through the AWS Data exchange.

Redshift data can be licenced to be accessible to other through the AWS Data exchange.



Machine learning Overview

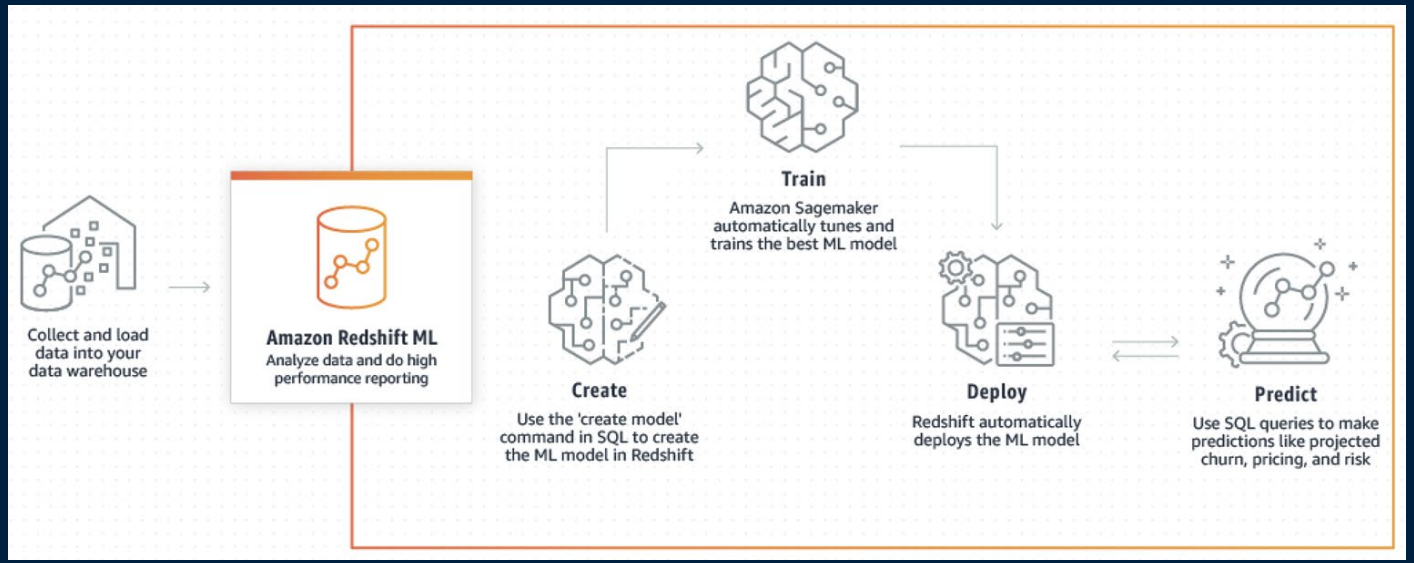
Redshift can be integrated with Amazons Machine learning (ML) service SageMaker.

Allows for fully managed ML models to be directly applied to the data warehouse with no prior ML knowledge.

Models can be created and trained on Redshift using SageMaker or models that the user has created can be trained on SageMaker then applied to the Redshift warehouse.

ML how it works

- Connect SageMaker to the Redshift instance.
- Use SQL statements to create a model and specify training data.
- Models then applied to SQL queries and analytics in RedShift.



On-Demand Cost

- Pay for provisioned capacity by the hour for each node type.
- Partial hours are billed in one-second increments.
- Pause anytime to suspend billing.
- Pause and resume can be manual or scheduled.

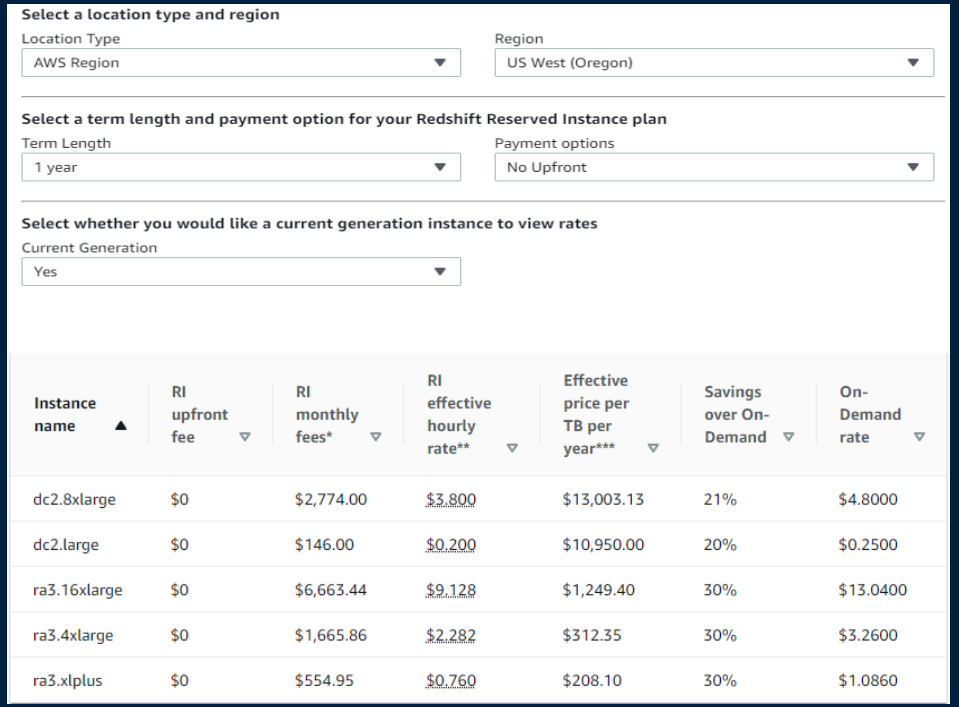
	vCPU	Memory	Addressable storage capacity	I/O	Price
Dense Compute DC2					
dc2.large	2	15 GiB	0.16TB SSD	0.60 GB/s	$0.25 per Hour
dc2.8xlarge	32	244 GiB	2.56TB SSD	7.50 GB/s	$4.80 per Hour
RA3 with Redshift Managed Storage*					
ra3.xlplus	4	32 GiB	32TB RMS	0.65 GB/s	$1.086 per Hour
ra3.4xlarge	12	96 GiB	128TB RMS	2.00 GB/s	$3.26 per Hour
ra3.16xlarge	48	384 GiB	128TB RMS	8.00 GB/s	$13.04 per Hour

Reserved Cost

Discounts over On-demand pricing but locked into 1 or 3 year terms for the node type.

3 pricing options:

- No upfront payment
- Partial upfront payment
- All upfront payment



Select a location type and region

Location Type: AWS Region

Region: US West (Oregon)

Select a term length and payment option for your Redshift Reserved Instance plan

Term Length: 1 year

Payment options: No Upfront

Select whether you would like a current generation instance to view rates

Current Generation: Yes

Instance name	RI upfront fee	RI monthly fees*	RI effective hourly rate**	Effective price per TB per year***	Savings over On-Demand	On-Demand rate
dc2.8xlarge	$0	$2,774.00	$3.800	$13,003.13	21%	$4.8000
dc2.large	$0	$146.00	$0.200	$10,950.00	20%	$0.2500
ra3.16xlarge	$0	$6,663.44	$9.128	$1,249.40	30%	$13.0400
ra3.4xlarge	$0	$1,665.86	$2.282	$312.35	30%	$3.2600
ra3.xlplus	$0	$554.95	$0.760	$208.10	30%	$1.0860

Spectrum and Serverless cost

Spectrum:

- Cost calculated on number of bytes scanned, Rounded up to the next megabyte (Min 10 MB).
- Queries of external data in Amazon S3 are not billed for separately.
- For US West Region works out to: $5 per TB.

Serverless:

- Cost based on compute capacity measured in Redshift Processing Units (RPUs).
- Cost calculated in Redshift RPU-hours on a per-second basis (minimum 60-seconds).
- For US West Region works out to: $0.36 per RPU hour.

Conclusion

Amazon Redshift is a cloud based data warehouse service specialized for data analytics on large quantities of data.

Available in all AWS regions it handles the hardware , software, and storage needed for a data warehouse to function.

With features such as integrated machine learning and 3rd party data Redshift can accelerate the analysis of your data with no up-front cost or a pay as you go setup.

Objectives

- Short introduction to data warehousing and how it differs from a normal database.
- Overview of querying on Redshift.
- Understand the availability of Redshift.
- Look at the security implementation available on Redshift.
- Discuss instance types that are available on Redshift.
- Understand some of the features available to use on Redshift such as:
  - Serverless instance type and its benefits.
  - Data sharing and the availability of 3rd party data.
  - Machine learning integration into data analysis.
- Understand the cost of using Redshifts various instance types.


THE UNIVERSITY OF BRITISH COLUMBIA