Amazon Redshift

March 3, 2020

Reference

- https://docs.aws.amazon.com/redshift/latest/gsg/getting-started.html
- https://docs.aws.amazon.com/redshift/latest/dg/welcome.html

What is Amazon Redshift

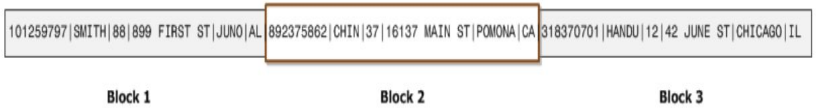
- ✓ It is a petabyte-scale, fully managed, fast and scalable data warehouse service in the cloud.
- ✓ Redshift makes it simple and cost-effective to analyze all your data across your data warehouse and data lake.
- ✓ Redshift delivers ten times faster performance than other data warehouses by using machine learning, massively parallel query execution, and columnar storage on high-performance disk.
- ✓ Deploying a new data warehouse takes few minutes.
- ✓ Redshift enables you to run queries across petabytes of data in your Redshift data warehouse, and exabytes of data in your data lake built on Amazon S3.

	database	***data warehouse***	***data lake***
type of data	structured	structured	unstructured
is data normalized?	yes	not generally as it is used for reporting and analytics	no
type of operation	online transaction processing(OLTP)	online analytical processing(OLAP), aggregate queries, analytics	raw data for analysis
cost of storing data	high	high	low
agility(can i change the data structure easily?)	no	no	yes
data security	matured and robust	matured and robust	still evolving as most of the technologies are open source
audience - type of user	anyone	large companies with petabytes of data	data scientists
use case	ecommerce website - a website creating an order for a product it has sold	conducting market research by analyzing large volumes of data in depth	ingestion of semi-structured and unstructured data sources (aka big data) such as equipment readings, telemetry data, logs, streaming data

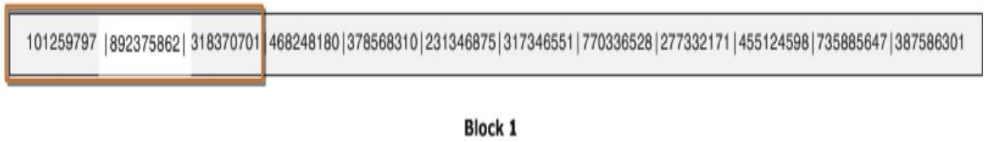
Columnar Storage

This first illustration shows how records from database tables are typically stored into disk blocks by row.

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



Benefits of Amazon Redshift

- Faster performance: Amazon redshift uses machine learning, a massively parallel architecture, computed-optimized hardware, and result set caching to deliver high throughput and sub-second response times, even with thousands of concurrent queries.
- Easy to set up, deploy, and manage:
  - ✓ It is simple to use, enabling you to deploy a new data warehouse in minutes.
  - ✓ It automates most of the common administrative tasks to manage, monitor, and scale your data warehouse. This helps you break free from the complexities of managing on-premises data warehouses.
- Cost-effective: There are no upfront costs with Redshift, and you only pay for what you use.

Benefits of Amazon Redshift

- Scale quickly to meet your needs:
  - It enables you to scale from querying gigabytes to exabytes of data across your Redshift data warehouse and Amazon S3 data lake. Quickly analyze any size of data in S3 with no loading or ETL required. (since a lot of data warehouses has scalability problems)
  - Easily resize Redshift cluster with just a few clicks on the console. You can scale up or down as your needs change.
- Query Your data lake:
  - You can directly query open data formats stored in Amazon S3 with Redshift Spectrum, a feature of Redshift, without the need for unnecessary data movement.
  - Amazon Redshift enables you to analyze data across your data warehouse and data lake, together, with a single service.
- Secure:
  - It can be encrypted using AWS KMS(Key Management Service) or HSM(Hardware Security Module). (Sensitive data can be encrypted, such as SSN)
  - Redshift is compliant with SOC1, SOC2, SOC3, PCI DSS Level 1 requirements, and FedRAMP, and HIPAA eligible with BAA available from AWS. (most of healthcare and financial institutions will need data warehouses and they also need such types of requirements)

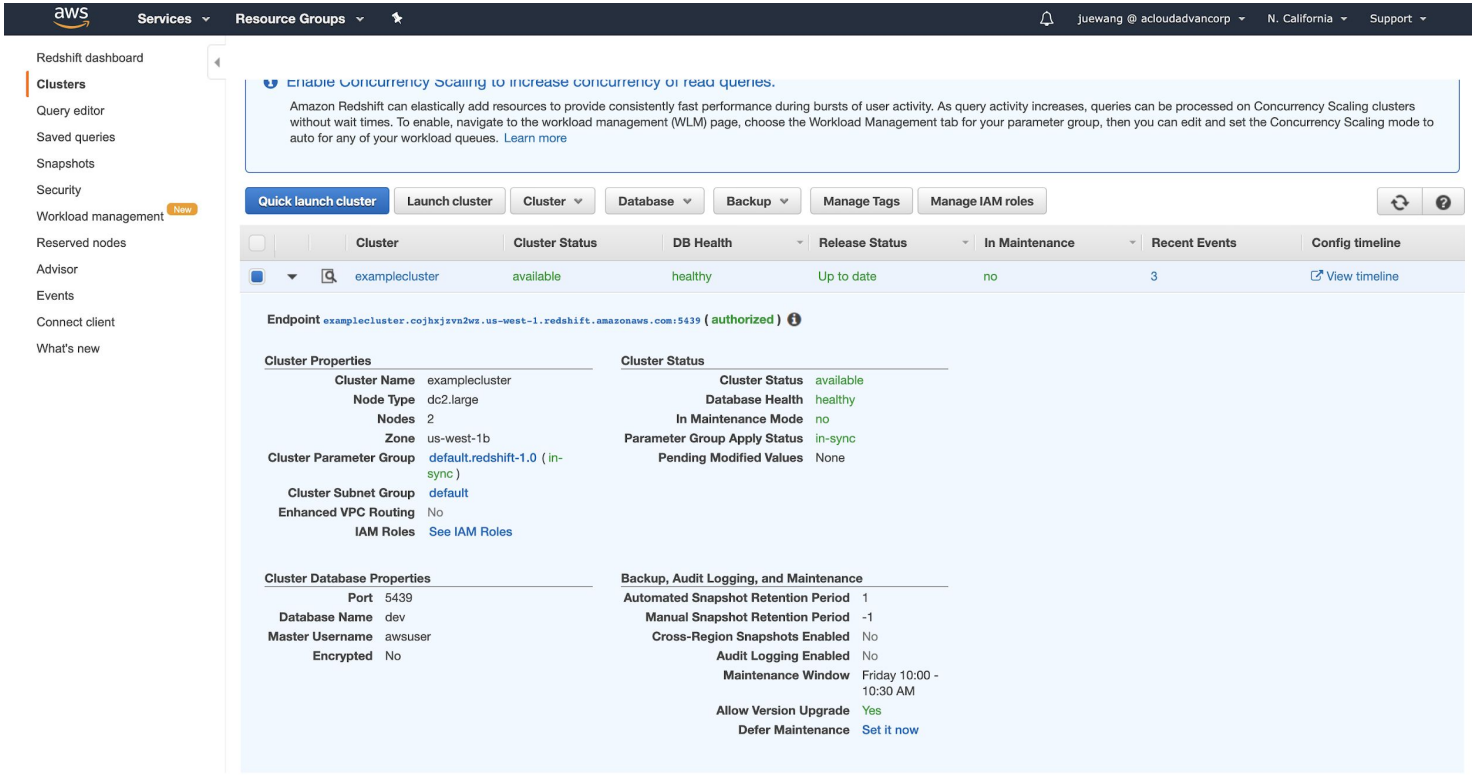
Use cases for Amazon Redshift

- Accelerate all your analytics workloads:
  - you can use Amazon Redshift to get sub-second results for reports, dashboards, and interactive analysis, and to get fast results for complex queries on any scale of data.
- Unified data warehouse and data lake:
  - you can use Amazon Redshift to run queries across your data warehouse and data lake to unlock insights that you would not be able to obtain by querying independent data silos.
- Modernize your on-premises data warehouse:
  - You can modernize your on-premises data warehouse to a fast, scalable, easy to manage, and cost-effective cloud data warehouse running on Amazon Redshift.

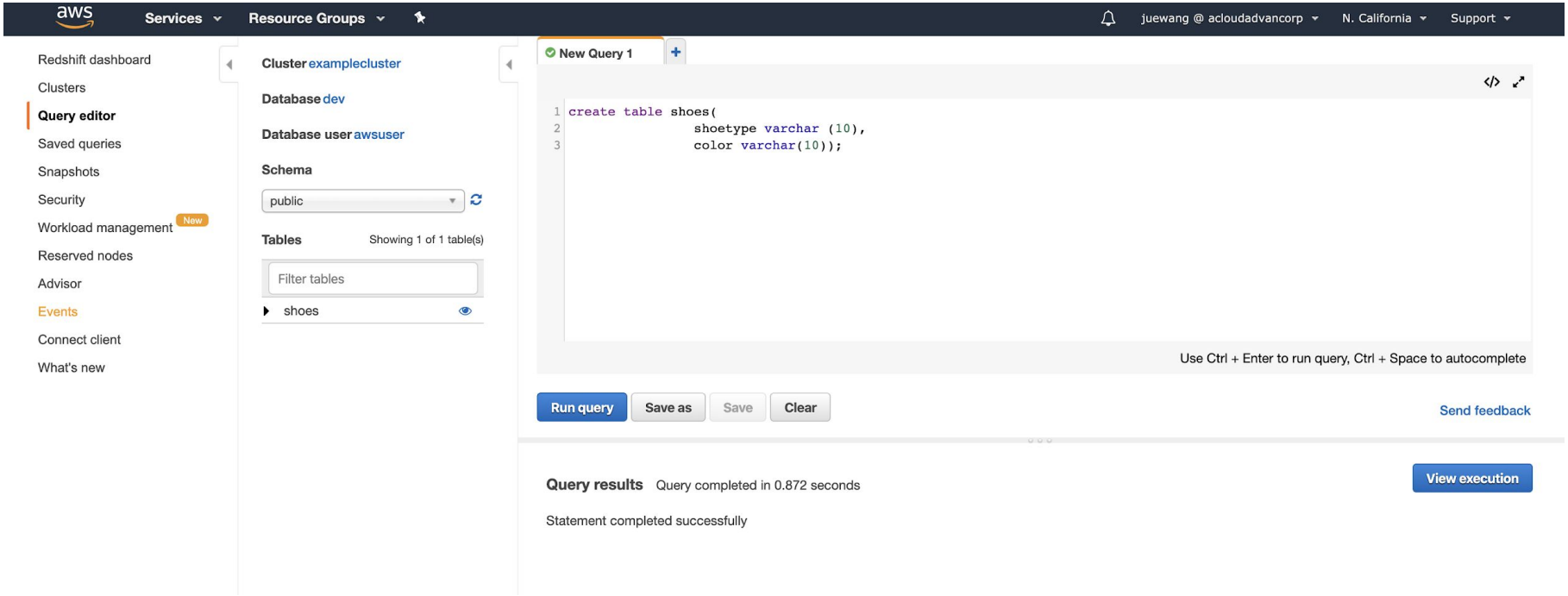
Next...

- Launch a sample Amazon Redshift Cluster
  - Step1: Amazon Redshift uses port 5439 by default
- Create data/table in Amazon Redshift
- Run Queries
- https://docs.aws.amazon.com/redshift/latest/gsg/getting-started.html

Create a Cluster



Creating Table in Redshift Demo



Query in Redshift Demo