
AWS
re:Invent

EMB012

Introduction to AQUA for Amazon Redshift

Andrew Caldwell

Senior Principal Engineer, Amazon Redshift

AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What to expect

1

Hear

more details on the AQUA announcement

2

Understand

our technology evolution and how it can benefit your organization

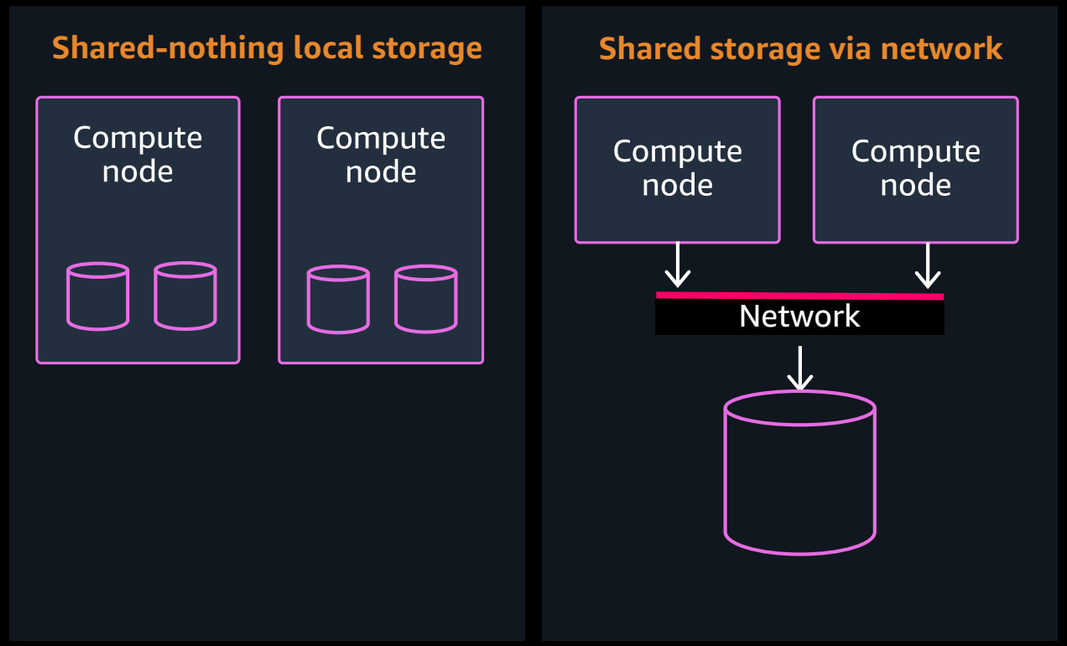
3

Plan

how your data warehouse can evolve in the future

Evolution

SHARED NOTHING TO DISAGGREGATED COMPUTE



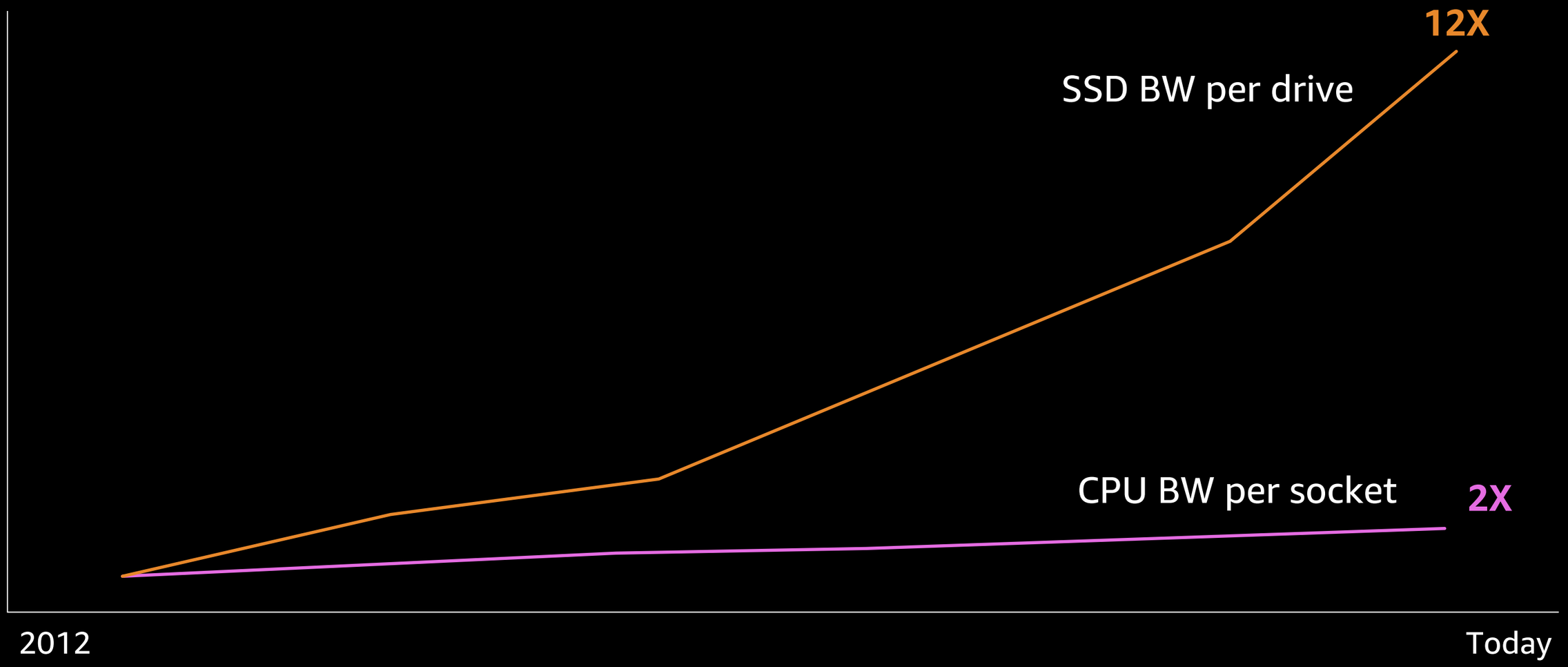
There has to be a better way?

Local storage has enabled the fastest cloud-based data warehouses

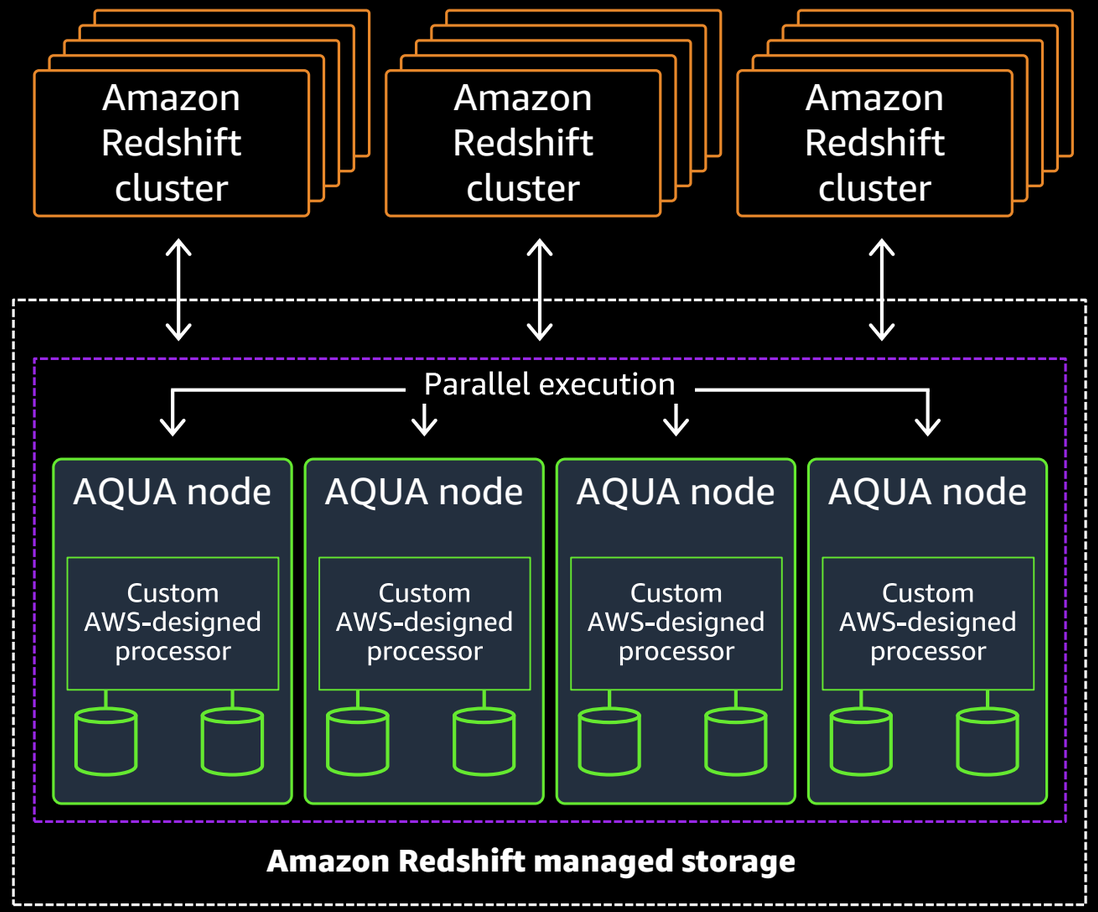
Shared storage enables flexibility at the cost of performance

What if we could get the benefits of both without a network performance penalty?

Diverging speeds in hardware



AQUA (Advanced Query Accelerator) for Amazon Redshift



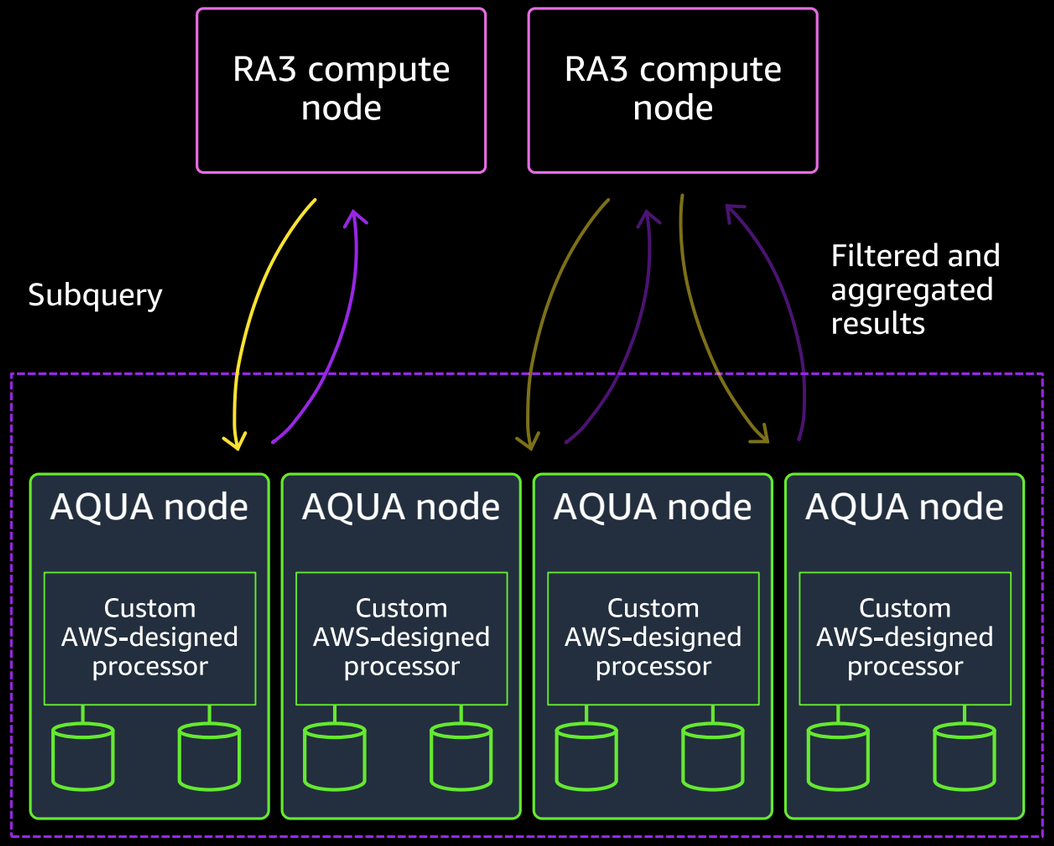
New distributed and hardware-accelerated cache

With AQUA, Amazon Redshift is up to 10x faster than any other cloud data warehouse

Available in preview now with RA3 16XL and 4XL
No code changes required and no extra cost

AQUA

A PUSH-DOWN AND SCALE-OUT COMPUTATION LAYER

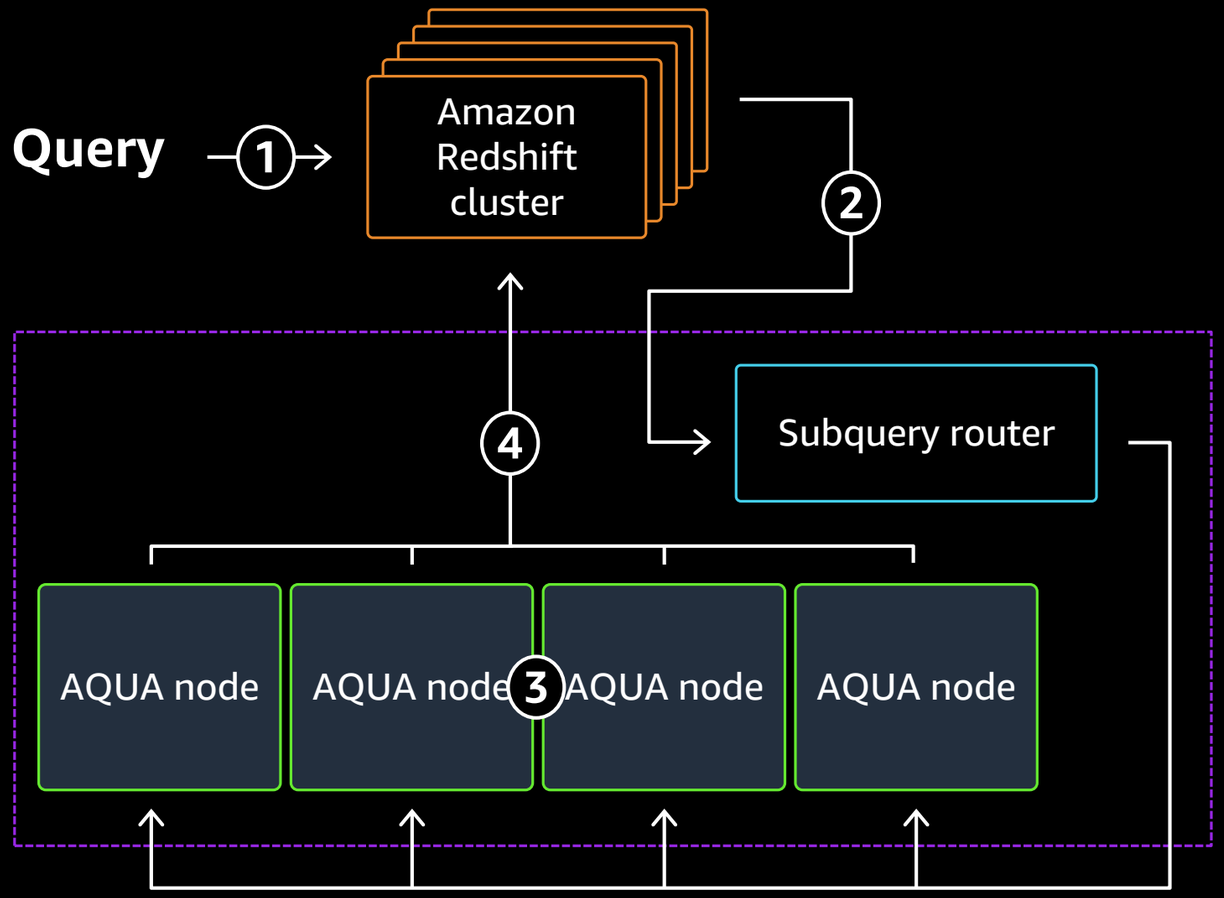


Minimize data movement over the network by pushing down operations to AQUA nodes

Scale out SCAN and AGG operations across AQUA

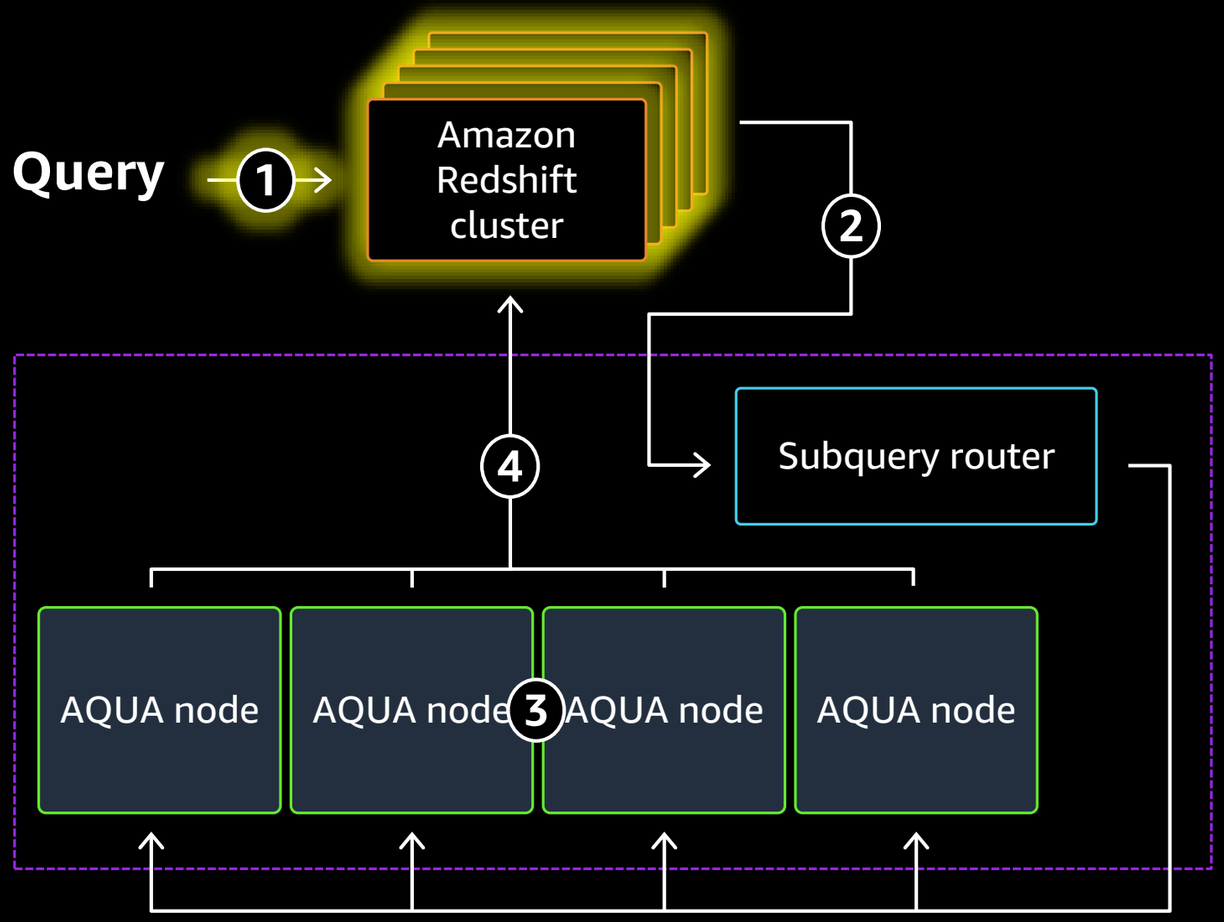
Query lifecycle in AQUA

1. Amazon Redshift receives a query; builds and optimizes a query tree (execution plan)
2. Ships subqueries with SCAN / AGG to AQUA
3. The AQUA subquery router receives the query and distributes it across multiple AQUA nodes to process in parallel
4. AQUA nodes process the query, including hydrating data from Amazon S3, and send the results back to Amazon Redshift



Query lifecycle in AQUA

1. Amazon Redshift receives a query; builds and optimizes a query tree (execution plan)
2. Ships subqueries with SCAN / AGG to AQUA
3. The AQUA subquery router receives the query and distributes it across multiple AQUA nodes to process in parallel
4. AQUA nodes process the query, including hydrating data from Amazon S3, and send the results back to Amazon Redshift

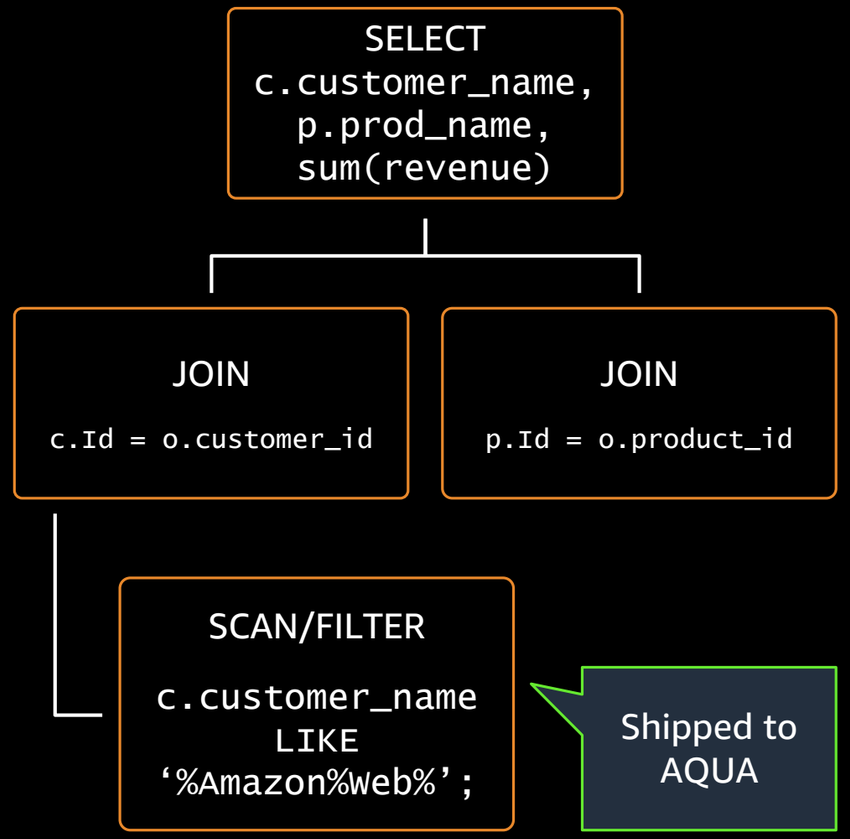


What happens in Amazon Redshift?

Amazon Redshift

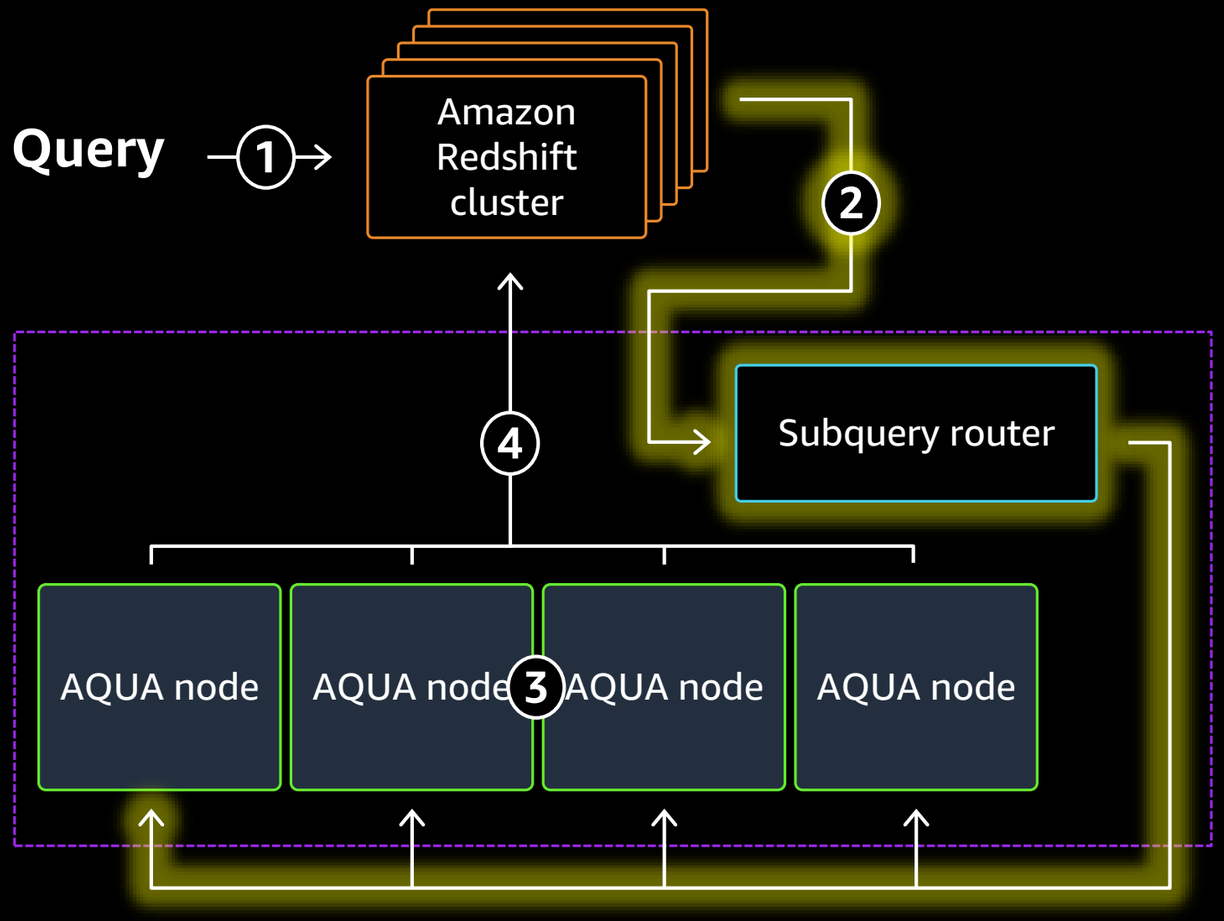
1. Receives a query
2. Builds and optimizes a query tree (execution plan)
3. Groups SCAN / AGG operations on tables
4. Determines which would be faster if sent to AQUA
5. Sends subquery to AQUA, waits for results …

```
SELECT  c.customer_name, p.prod_name, sum(revenue)
FROM orders o
      JOIN customers c ON c.id = o.customer_id
      JOIN products p ON p.id = o.product_id
WHERE c.customer_name LIKE '%Amazon%web%';
```



Query lifecycle in AQUA

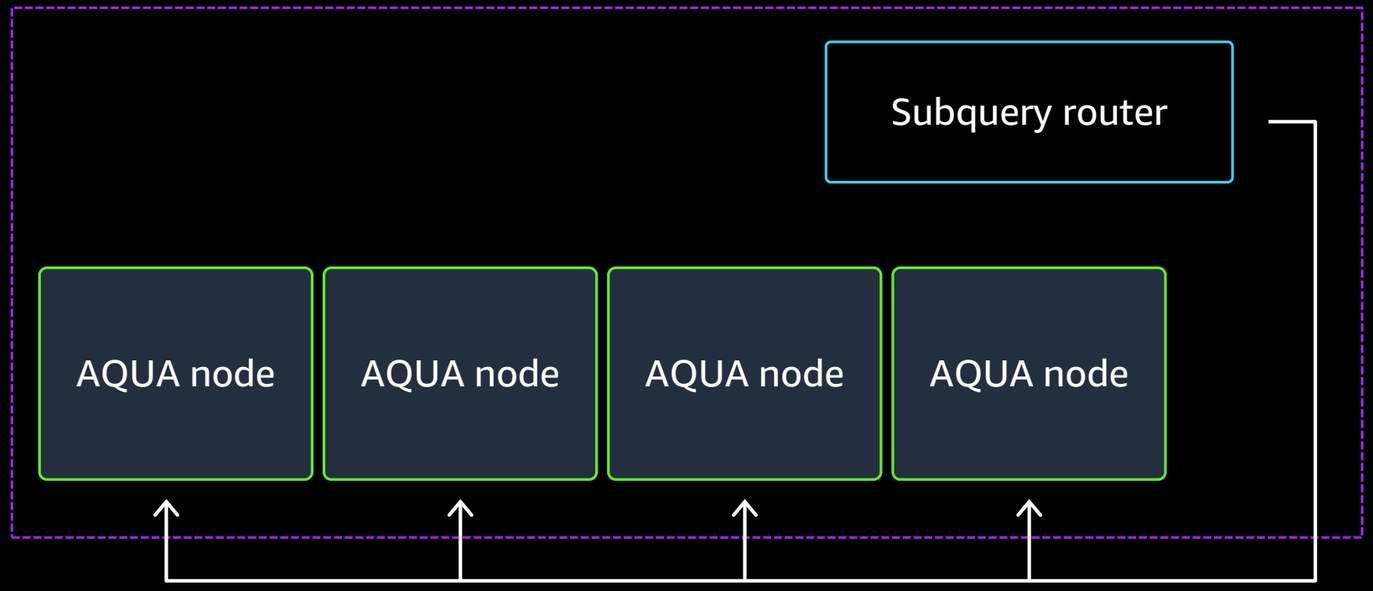
1. Amazon Redshift receives a query; builds and optimizes a query tree (execution plan)
2. Ships subqueries with SCAN / AGG to AQUA
3. The AQUA subquery router receives the query and distributes it across multiple AQUA nodes to process in parallel
4. AQUA nodes process the query, including hydrating data from Amazon S3, and send the results back to Amazon Redshift



AQUA receives subquery …

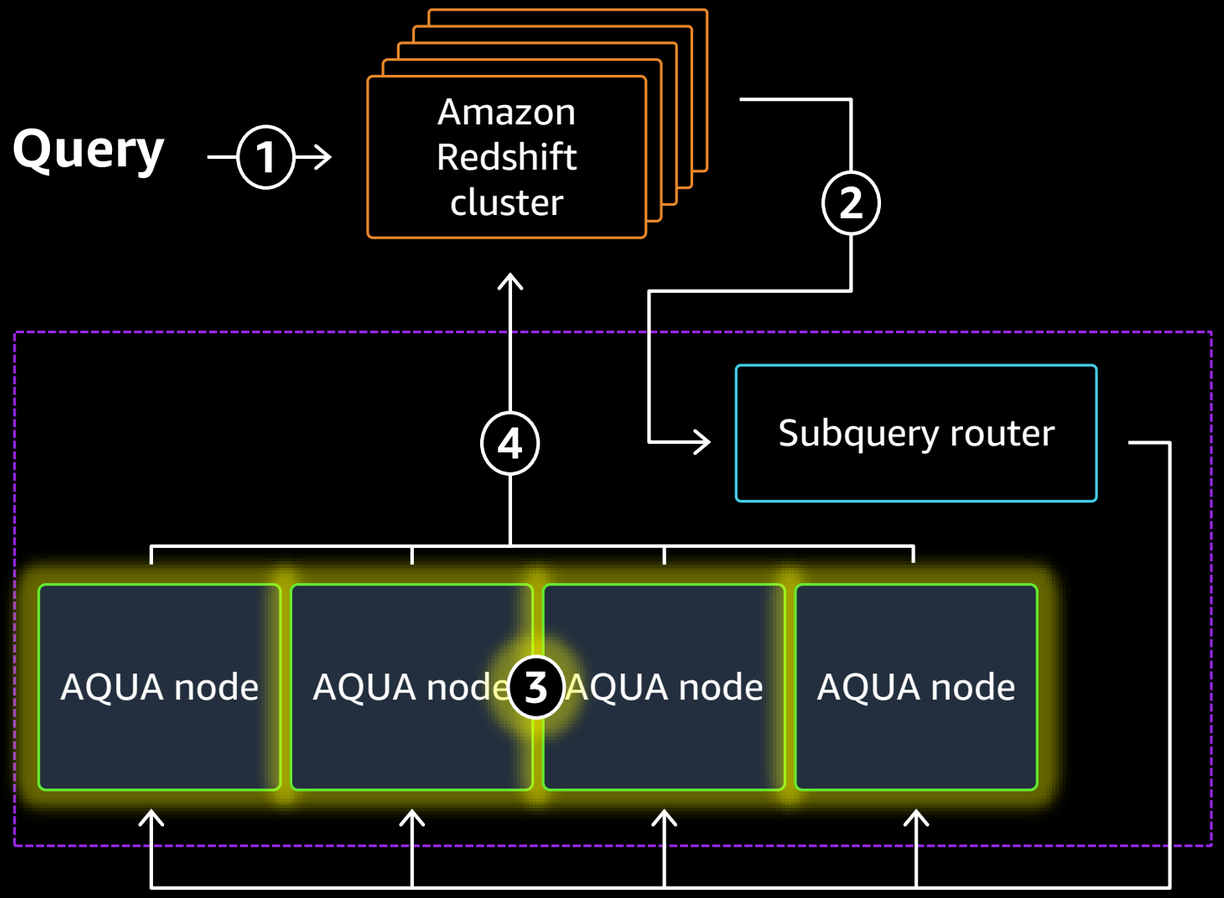
AQUA “load balancer” fans parts of subquery across fleet of AQUA nodes

- Faster execution from more parallelism
- Minimize hydration by sending operations on data to the nodes that recently operated on it



Query lifecycle in AQUA

1. Amazon Redshift receives a query; builds and optimizes a query tree (execution plan)
2. Ships subqueries with SCAN / AGG to AQUA
3. The AQUA subquery router receives the query and distributes it across multiple AQUA nodes to process in parallel
4. AQUA nodes process the query, including hydrating data from Amazon S3, and send the results back to Amazon Redshift



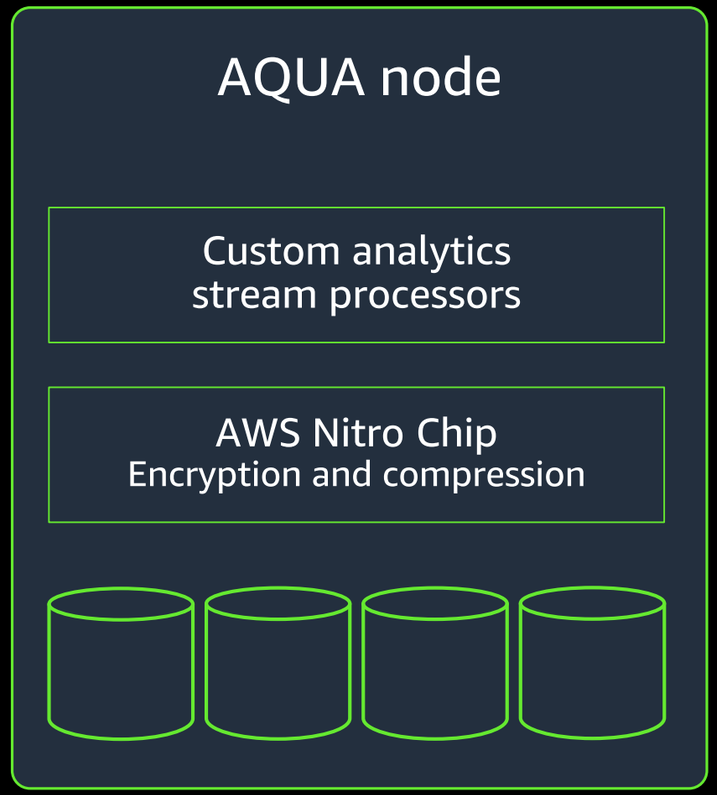
AQUA nodes execute SCAN / AGG subquery

1. AQUA compiles each request across multiple HW modules – CPU, custom ASICs, and query accelerators implemented in FPGAs

2. New / cold data hydrated from S3

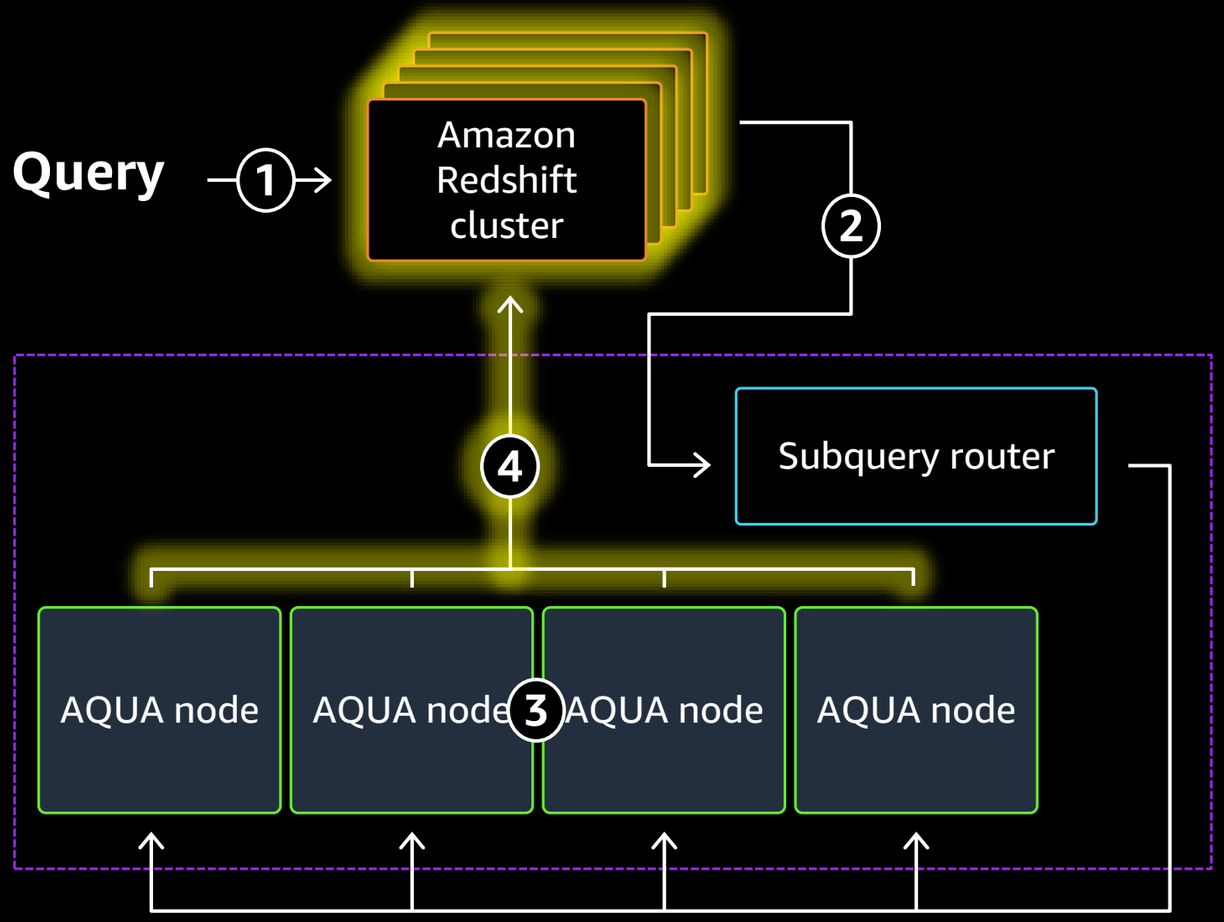
3. Hot data already cached in local SSDs

4. Hardware-accelerated encryption, compression, filtering, hashing, and aggregation to maximize throughput



Query lifecycle in AQUA

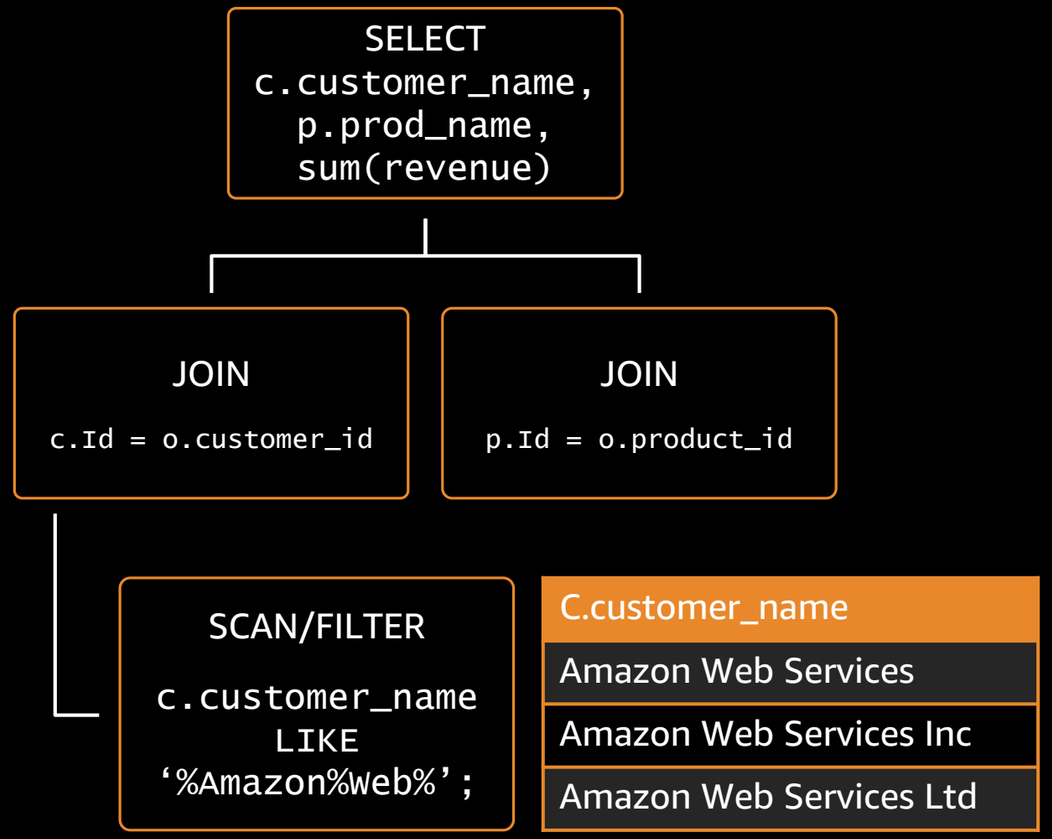
1. Amazon Redshift receives a query; builds and optimizes a query tree (execution plan)
2. Ships subqueries with SCAN / AGG to AQUA
3. The AQUA subquery router receives the query and distributes it across multiple AQUA nodes to process in parallel
4. AQUA nodes process the query, including hydrating data from Amazon S3, and send the results back to Amazon Redshift



RS receives results and completes query

1. AQUA nodes send results back to RS – typically <5% of data scanned is returned

2. Amazon Redshift resumes execution, starting from AQUA results, and completes the query



Query lifecycle in AQUA

HOW AQUA DELIVERS UP TO 10X QUERY PERFORMANCE

Caches recently used data, so subsequent queries run faster

Uses fanout and custom hardware to rapidly scan large volumes of data

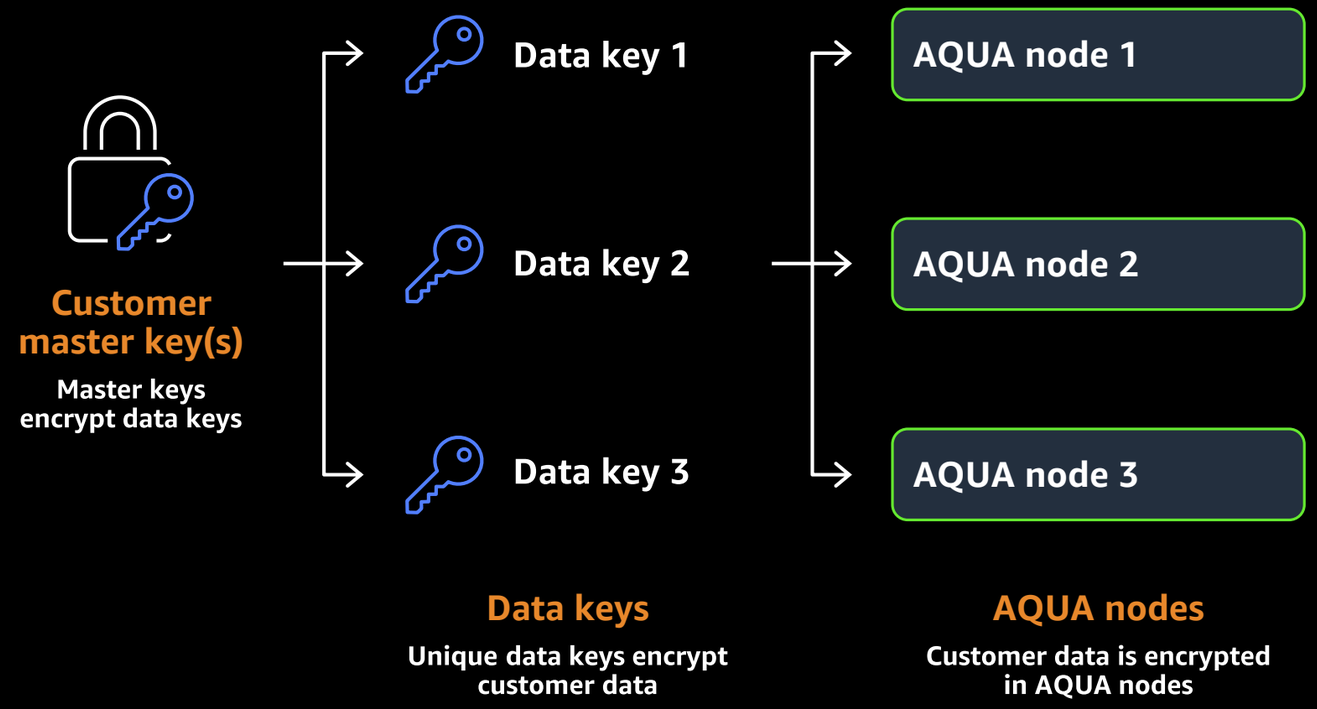
Efficient processing of LIKE, SIMILAR TO; large speedup for those predicates can make formerly inefficient workloads very practical

No code changes; works on Amazon Redshift RA3 16XL / 4XL nodes

Data security in AQUA

Encryption

- TLS encrypted channel over network
- Data encrypted at rest and in memory using customer managed keys



AQUA

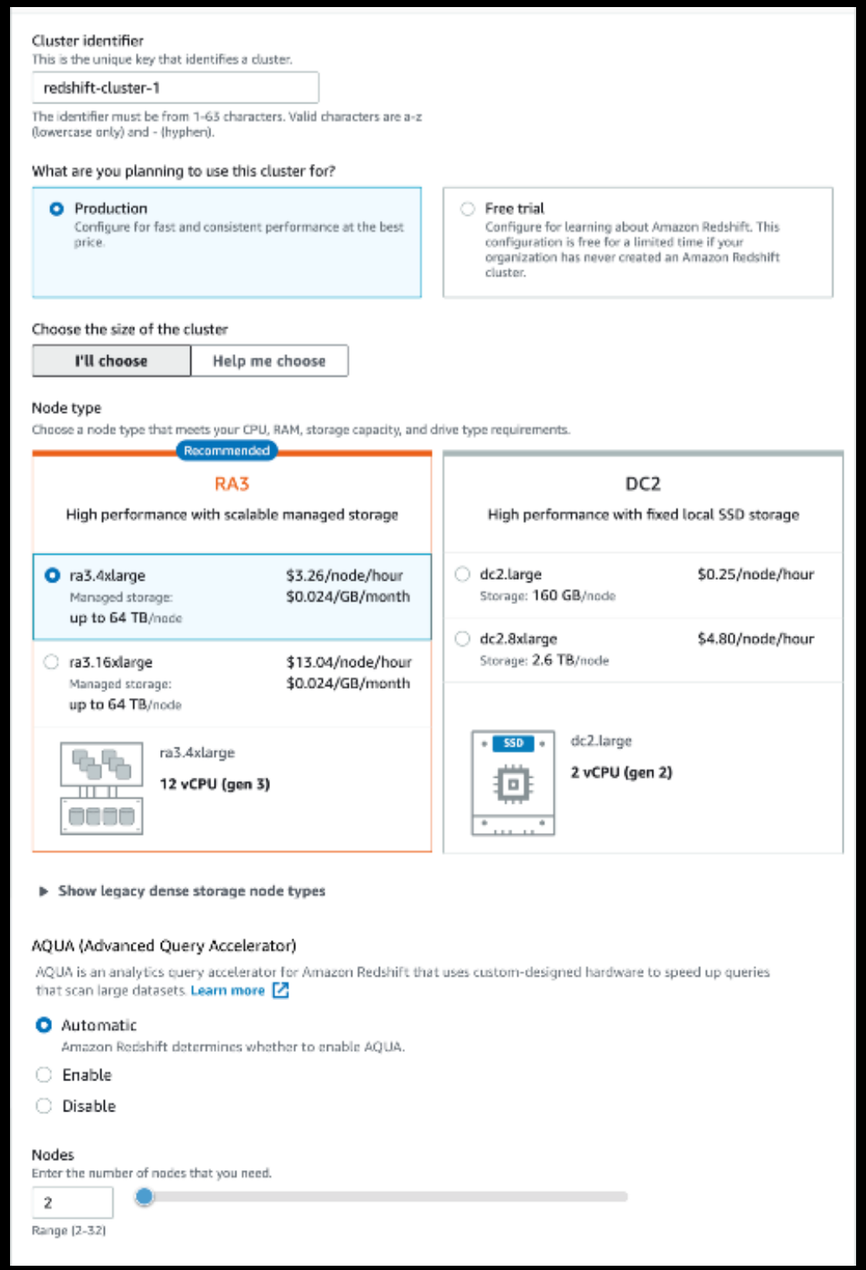
HOW TO GET STARTED?

Upgrade to Amazon Redshift RA3 nodes

Enable AQUA in the Amazon Redshift console or using the CLI / API

Run queries that perform large scan and aggregate operations with LIKE and SIMILAR_TO predicates

Check the query execution plan or query the system tables to see AQUA's impact



AQUA

USE CASES

BI and analytics acceleration – queries, reports, and dashboards up to 10X faster, providing up-to-date information for timely decision-making

---

eDiscovery, search, recommendations, text and sentiment analysis, clickstream analytics

- Compute-intensive use cases in: education, gaming, hospitality, retail, telecomm, and travel. For example:

- Show me all the emails with subject line LIKE ‘%Confidential: Redshift%’

- Show me all the products with name LIKE ‘%Vacuums’

- Show me all the hotels with name LIKE ‘%Hyatt%Regency%’

AQUA

WHAT TO EXPECT BEYOND GENERAL AVAILABILITY

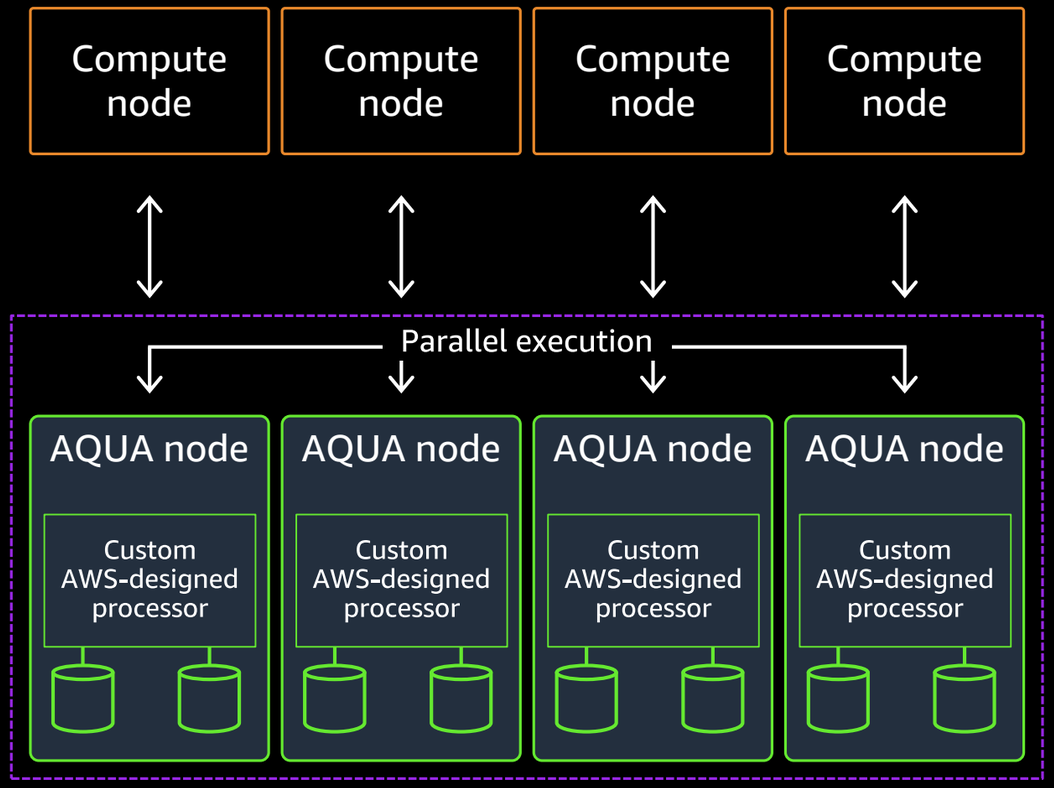
Continued push-down and acceleration of long-running queries

Support for Amazon Redshift features like concurrency scaling, pause and resume

Support in more Regions

AQUA (Advanced Query Accelerator)

A NEW DISTRIBUTED AND HARDWARE-ACCELERATED CACHE THAT WILL MAKE AMAZON REDSHIFT UP TO 10X FASTER THAN ANY OTHER CLOUD DATA WAREHOUSE WITHOUT INCREASING COST



Minimize data movement over the network by pushing down operations to AQUA nodes

AQUA nodes with custom AWS-designed analytics processors to make operations (compression, encryption, filtering, and aggregations) faster than traditional CPUs

Available only with RA3
No cost and no code changes required

Available in preview now

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.