



Optimizing Data Analysis *with* Amazon Redshift

A Step-by-Step Guide to Setup, Management,
and Performance Tuning

what is Amazon Redshift ?

a cloud-based data warehouse solution that offers -

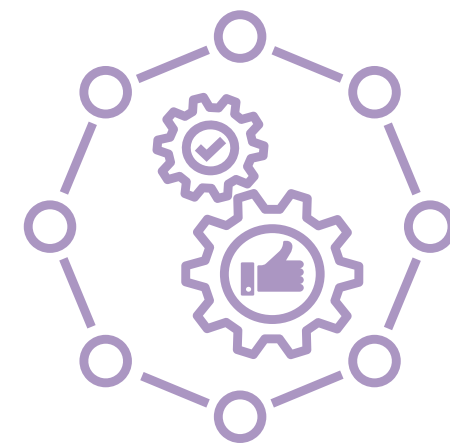
scalable storage



powerful query capabilities



integration with AWS



For more granular control

Create, configure, and manage your cluster to control computing resources.

Create cluster

setting up the *redshift cluster*

Node type [Info](#)

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

dc2.large

Number of nodes

Enter the number of nodes that you need.

2

Range (1-32)

Publicly accessible

For more information, see [Learn more about Redshift clusters security groups](#) 

☒ Turn on Publicly accessible

Allow public connections to Amazon Redshift.

Data Preparation & *Uploading to S3*

Summarize data preparation: downloading an e-commerce dataset and uploading it to an S3 bucket.

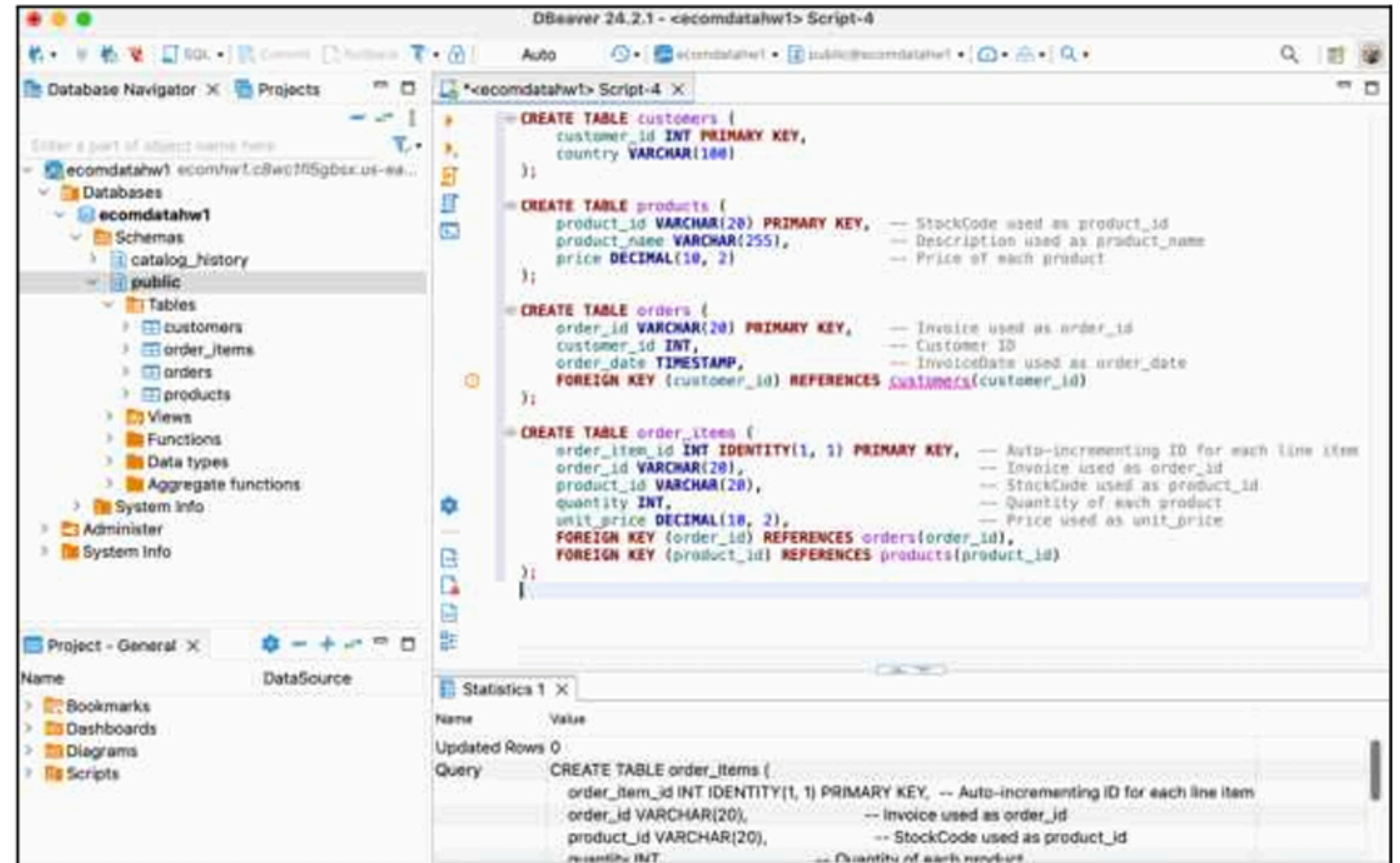


- Note IAM roles for Redshift-S3 integration, specifically AmazonS3ReadOnlyAccess.

Data Loading & *Schema design*

Star schema creation with tables for

- customers
- products
- orders
- order items



Data Analysis & *Querying*

Run queries to explore the data trends such as

total sales per product



customer purchase history



sales by country



total sales per product

The screenshot shows a SQL IDE with a query editor and a results grid. The query in the editor is:

```
SELECT p.product_name, SUM(oi.quantity * oi.unit_price) AS total_sales
FROM order_items oi
JOIN products p ON oi.product_id = p.product_id
GROUP BY p.product_name
ORDER BY total_sales DESC;
```

The results grid displays the following data:

	A-Z product_name	123 total_sales
1	REGENCY CAKESTAND 3 TIER	261,110.95
2	CREAM HANGING HEART T-LIGHT HOLDER	237,833.96
3	JUMBO BAG RED RETROSPOT	165,751.82
4	ASSORTED COLOUR BIRD ORNAMENT	123,631.87
5	POSTAGE	110,338.51
6	PARTY BUNTING	102,089.38
7	PAPER CHAIN KIT 50'S CHRISTMAS	75,388.48
8	CHILLI LIGHTS	68,453.5
9	JUMBO BAG PINK POLKADOT	66,904.73
10	DOORMAT UNION FLAG	63,910.54

sales by country

The screenshot shows the DBeaver SQL editor with a query to calculate total sales by country. The query is:

```
SELECT c.country, SUM(oi.quantity * oi.unit_price) AS total_sales
FROM customers c
JOIN orders o ON c.customer_id = o.customer_id
JOIN order_items oi ON o.order_id = oi.order_id
GROUP BY c.country
ORDER BY total_sales DESC;
```

The results pane below shows a table with 10 rows of data, sorted by total sales in descending order:

	A-Z country	123 total_sales
1	United Kingdom	13,482,121.18
2	EIRE	573,308.2
3	Netherlands	548,330.7
4	Germany	411,862.11
5	France	320,384.66
6	Australia	167,427.2
7	Spain	101,934.43
8	Switzerland	98,253.1
9	Sweden	87,421.52
10	Denmark	58,057.24

customer purchase history

The screenshot shows a SQL IDE with a query editor and a results pane. The query editor contains the following SQL code:

```
SELECT c.customer_id, COUNT(o.order_id) AS total_orders
FROM customers c
JOIN orders o ON c.customer_id = o.customer_id
GROUP BY c.customer_id
ORDER BY total_orders DESC;
```

The results pane shows a table with 10 rows and 2 columns: **customer_id** and **total_orders**. The data is sorted by **total_orders** in descending order.

	customer_id	total_orders
1	14,911	512
2	12,748	366
3	17,841	289
4	15,311	270
5	14,606	261
6	13,089	247
7	14,156	202
8	14,527	191
9	14,646	164
10	13,694	164

DISTKEY

(customer_id)

This helps ensure that joins between the orders and customers tables are efficient, reducing the need for inter-node communication.

SORTKEY

(customer_id)

Queries that filter or aggregate data by date will benefit from faster data retrieval, especially for operations like date-based sales trends.

Optimizing Redshift *Queries*

Impact of Optimizations

SORTKEY

(product_id)

Queries involving product-related analytics (like total sales per product) will be optimized.

DISTKEY

(order_date)

This makes joins between orders and order_items efficient, as the order_id is distributed across nodes.

Advanced *Features*

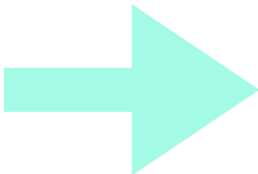
Create Materialized View

*<ecomdatahw1> Script- *<ecomdatahw1> Consol *<eco

```
CREATE MATERIALIZED VIEW mv_sales_summary AS
SELECT p.product_name,
       COUNT(o.order_id) AS total_orders,
       SUM(oi.quantity * oi.unit_price) AS total_sales
FROM order_items oi
JOIN orders o ON oi.order_id = o.order_id
JOIN products p ON oi.product_id = p.product_id
GROUP BY p.product_name;
```

Statistics 1 X

Name	Value
Updated Rows	0
Query	CREATE MATERIALIZED VIEW mv_sales_summary AS SELECT p.product_name, COUNT(o.order_id) AS total_orders, SUM(oi.quantity * oi.unit_price) AS total_sales FROM order_items oi JOIN orders o ON oi.order_id = o.order_id JOIN products p ON oi.product_id = p.product_id GROUP BY p.product_name
Start time	Sun Oct 06 15:01:06 PDT 2024
Finish time	Sun Oct 06 15:01:17 PDT 2024



Perform Analytics Using the Materialized View

Query 1: Top 5 Products by Total Sales

*<ecomdatahw1> Script- *<ecomdatahw1> Consol *<ecomd

```
SELECT product_name, total_sales
FROM mv_sales_summary
ORDER BY total_sales DESC
LIMIT 5;
```

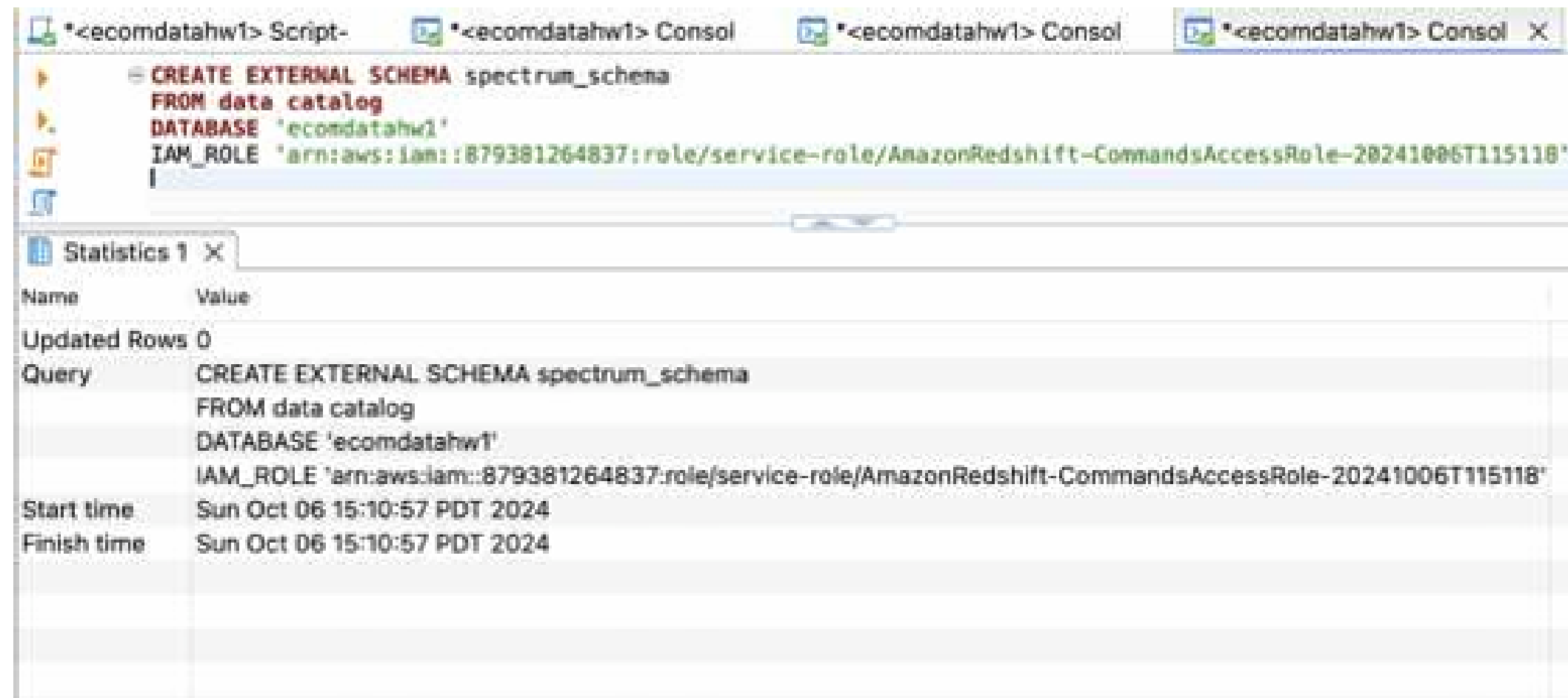
mv_tbl_mv_sales_summary__0 1 X

SELECT product_name, total_s | Enter a SQL expression to filter results (t

	A-Z product_name	total_sales
1	POSTAGE	13,240,621.2
2	WHITE HANGING HEART T-LIGHT HOLDER	2,378,339.6
3	REGENCY CAKESTAND 3 TIER	1,827,776.65
4	JUMBO BAG RED RETROSPOT	1,491,766.38
5	PARTY BUNTING	918,804.42

Use Redshift *Spectrum*

Design and run 2 queries to perform meaningful analytics involving the external schema and draw valuable insights that can support decision making



The screenshot displays the Amazon Redshift console interface. At the top, there are four tabs: '*<ecomdatahw1> Script-', '*<ecomdatahw1> Consol', '*<ecomdatahw1> Consol', and '*<ecomdatahw1> Consol X'. The active tab shows a SQL script for creating an external schema. Below the script, there is a 'Statistics 1 X' tab. The statistics table shows that the query was executed successfully, with 0 rows updated. The query text is: 'CREATE EXTERNAL SCHEMA spectrum_schema FROM data catalog DATABASE 'ecomdatahw1' IAM_ROLE 'arn:aws:iam::879381264837:role/service-role/AmazonRedshift-CommandsAccessRole-20241006T115118''. The start and finish times are both 'Sun Oct 06 15:10:57 PDT 2024'.

Name	Value
Updated Rows	0
Query	CREATE EXTERNAL SCHEMA spectrum_schema FROM data catalog DATABASE 'ecomdatahw1' IAM_ROLE 'arn:aws:iam::879381264837:role/service-role/AmazonRedshift-CommandsAccessRole-20241006T115118'
Start time	Sun Oct 06 15:10:57 PDT 2024
Finish time	Sun Oct 06 15:10:57 PDT 2024

Scalable Data Management

Amazon Redshift efficiently handles large-scale data, offering flexible and scalable storage solutions.



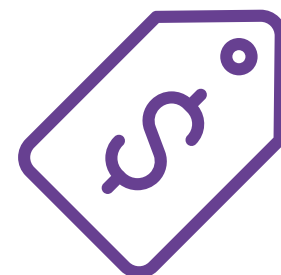
Robust Data Analysis Capabilities

Perform complex queries quickly, enabling deep insights into datasets for better decision-making.



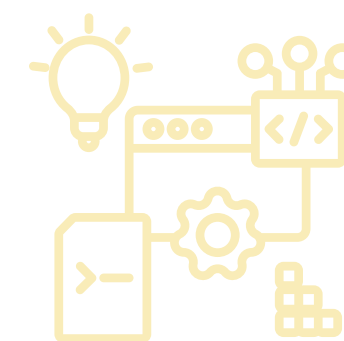
Cost-Effective Cloud Solution

Pay-as-you-go pricing ensures cost efficiency for businesses of all sizes.



Integration with AWS Ecosystem

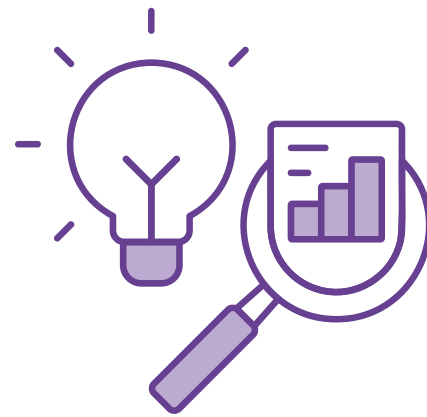
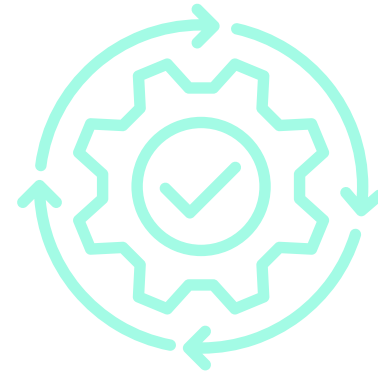
Seamless integration with AWS services like S3, Glue, and Redshift Spectrum extends its functionality and simplifies workflows



Summary & Key Takeaways

Optimization Drives Performance

Use features like DISTKEY, SORTKEY, materialized views, and query tuning for faster performance and resource efficiency



Actionable Insights for Business Growth

Queries on sales trends, customer segmentation, and product performance drive strategic marketing and operational decisions.

Future-Ready Data Warehouse

Advanced features like Redshift Spectrum enable handling structured and semi-structured data for evolving analytics needs.

