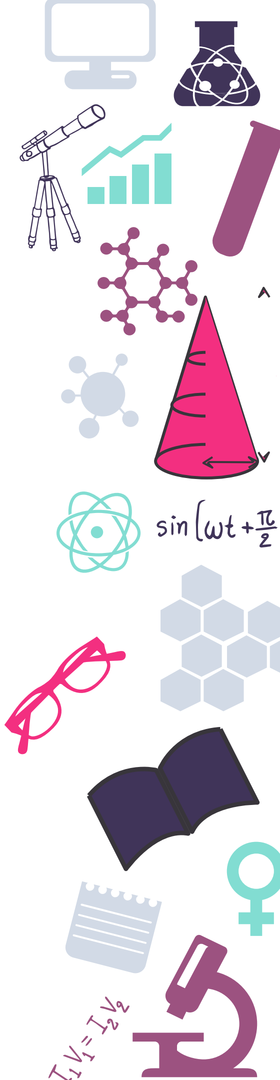
Data Science At AWS

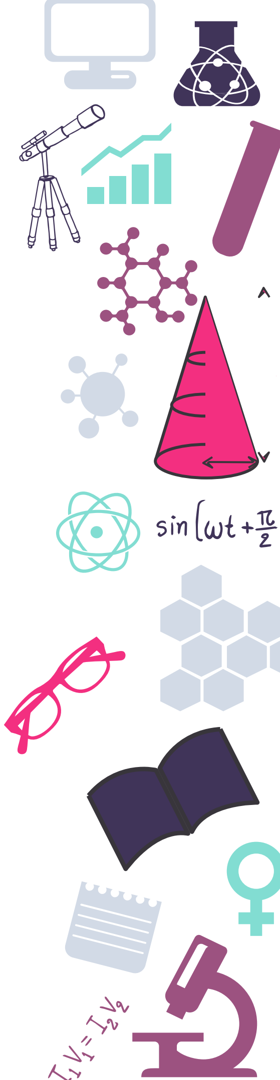
Rui LIU, Bruno BALDEZ CORREA

1

Data Science Services at AWS



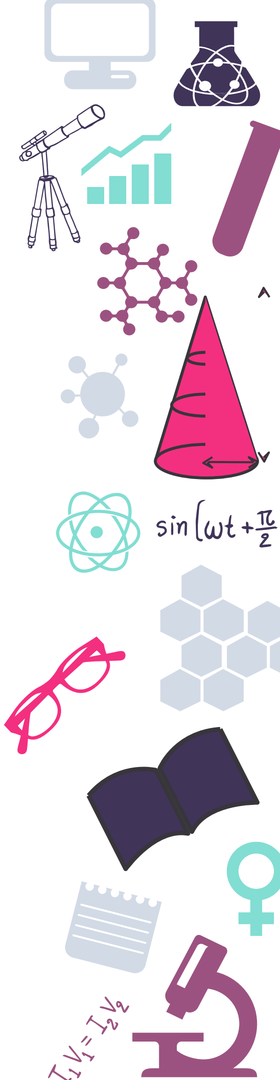
2 Amazon Redshift



Amazon Redshift is a fast, fully managed, petabyte-scale cloud-base data warehouse solution offered by Amazon Web Services that provides simple and cost-effective functionalities to analyze all your data using standard SQL and BI techniques.

Characteristics

- ✓ Fast
- ✓ Inexpensive
- ✓ Extensible
- ✓ Simple
- ✓ Scalable
- ✓ Secure
- ✓ Compatible



Architecture

Clusters

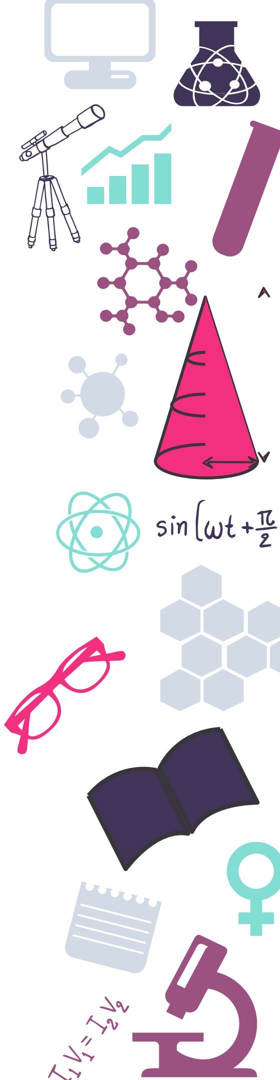
- ✓ 1 or more compute nodes.

Leader node

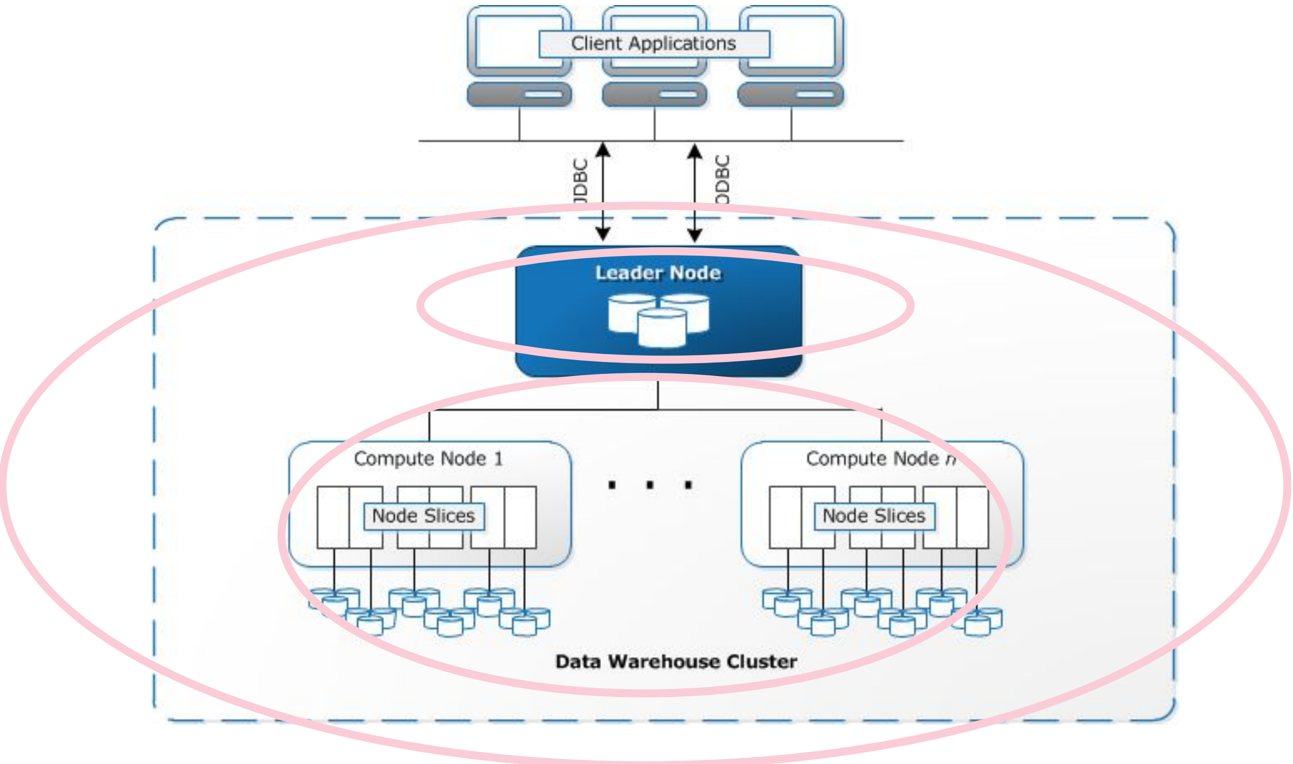
- ✓ compiles code
- ✓ distributes it

Compute nodes

- ✓ compiles code
- ✓ execute the compiled code
- ✓ send intermediate results
- ✓ **dense storage nodes** or **dense compute nodes**



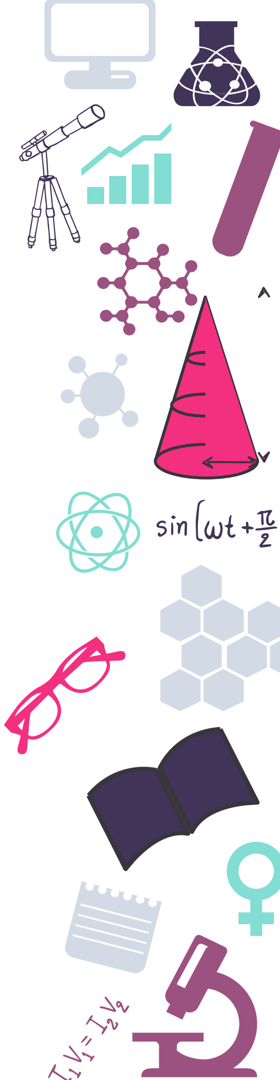
Architecture



Data Distribution

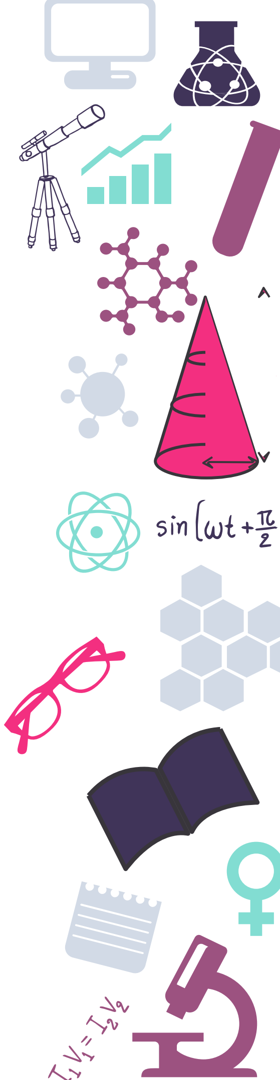
- ALL
- EVEN
- KEY

- SortKey
- Primary and Foreign keys

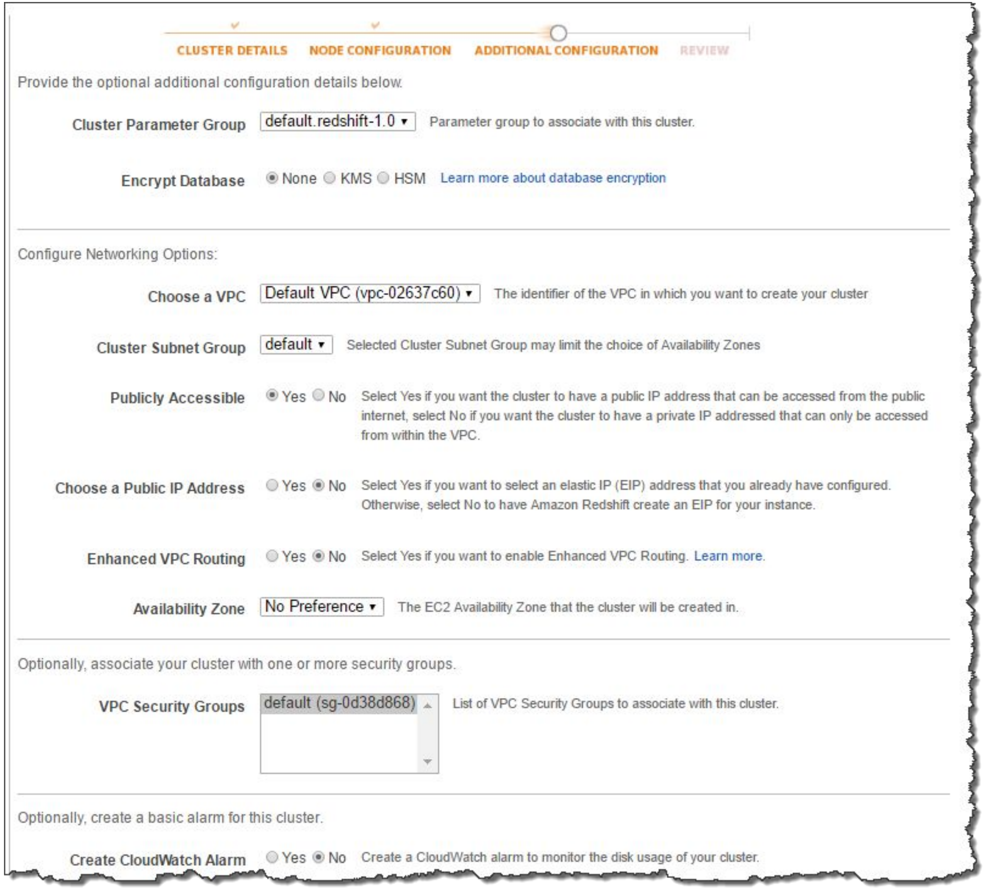


Getting Started with Amazon Redshift

Step 1: Set Up Prerequisites
Step 2: Create an IAM Role
Step 3: Launch a Sample Amazon Redshift Cluster
Step 4: Authorize Access to the Cluster
Step 5: Connect to the Sample Cluster
Step 6: Load Sample Data from Amazon S3
Step 7: Find Additional Resources and Reset Your Environment



Redshift Use Scenario



Provide the optional additional configuration details below.

Cluster Parameter Group: default.redshift-1.0 ▾ Parameter group to associate with this cluster.

Encrypt Database: ● None ○ KMS ○ HSM Learn more about database encryption

Configure Networking Options:

Choose a VPC: Default VPC (vpc-02637c60) ▾ The identifier of the VPC in which you want to create your cluster

Cluster Subnet Group: default ▾ Selected Cluster Subnet Group may limit the choice of Availability Zones

Publicly Accessible: ● Yes ○ No Select Yes if you want the cluster to have a public IP address that can be accessed from the public internet, select No if you want the cluster to have a private IP addressed that can only be accessed from within the VPC.

Choose a Public IP Address: ○ Yes ● No Select Yes if you want to select an elastic IP (EIP) address that you already have configured. Otherwise, select No to have Amazon Redshift create an EIP for your instance.

Enhanced VPC Routing: ○ Yes ● No Select Yes if you want to enable Enhanced VPC Routing. Learn more.

Availability Zone: No Preference ▾ The EC2 Availability Zone that the cluster will be created in.

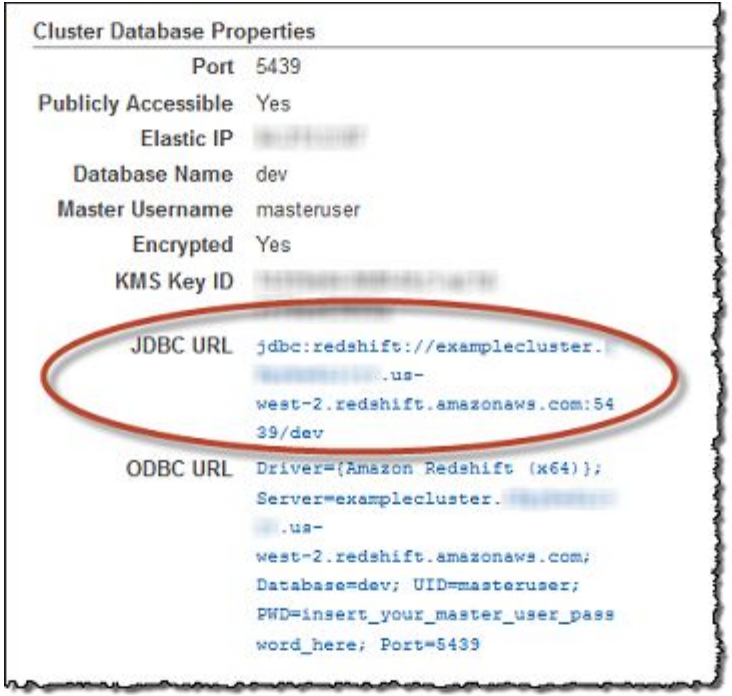
Optionally, associate your cluster with one or more security groups.

VPC Security Groups: default (sg-0d38d868) List of VPC Security Groups to associate with this cluster.

Optionally, create a basic alarm for this cluster.

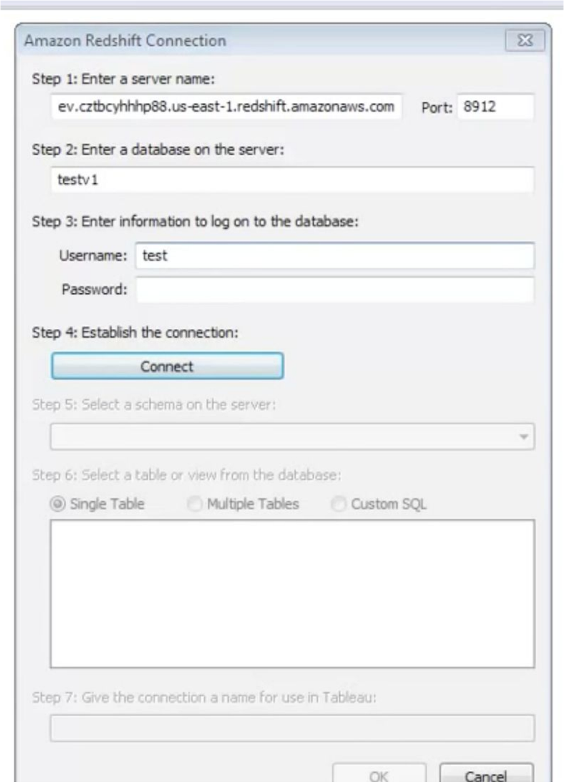
Create CloudWatch Alarm: ○ Yes ● No Create a CloudWatch alarm to monitor the disk usage of your cluster.

Redshift Use Scenario



Cluster Database Properties

Port	5439
Publicly Accessible	Yes
Elastic IP	[redacted]
Database Name	dev
Master Username	masteruser
Encrypted	Yes
KMS Key ID	[redacted]
JDBC URL	jdbc:redshift://examplecluster.[redacted].us-west-2.redshift.amazonaws.com:5439/dev
ODBC URL	Driver={Amazon Redshift (x64)}; Server=examplecluster.[redacted].us-west-2.redshift.amazonaws.com; Database=dev; UID=masteruser; PWD=insert_your_master_user_password_here; Port=5439



Amazon Redshift Connection

Step 1: Enter a server name: ev.cztbcyhhhp88.us-east-1.redshift.amazonaws.com Port: 8912

Step 2: Enter a database on the server: testv1

Step 3: Enter information to log on to the database:

Username: test

Password:

Step 4: Establish the connection:

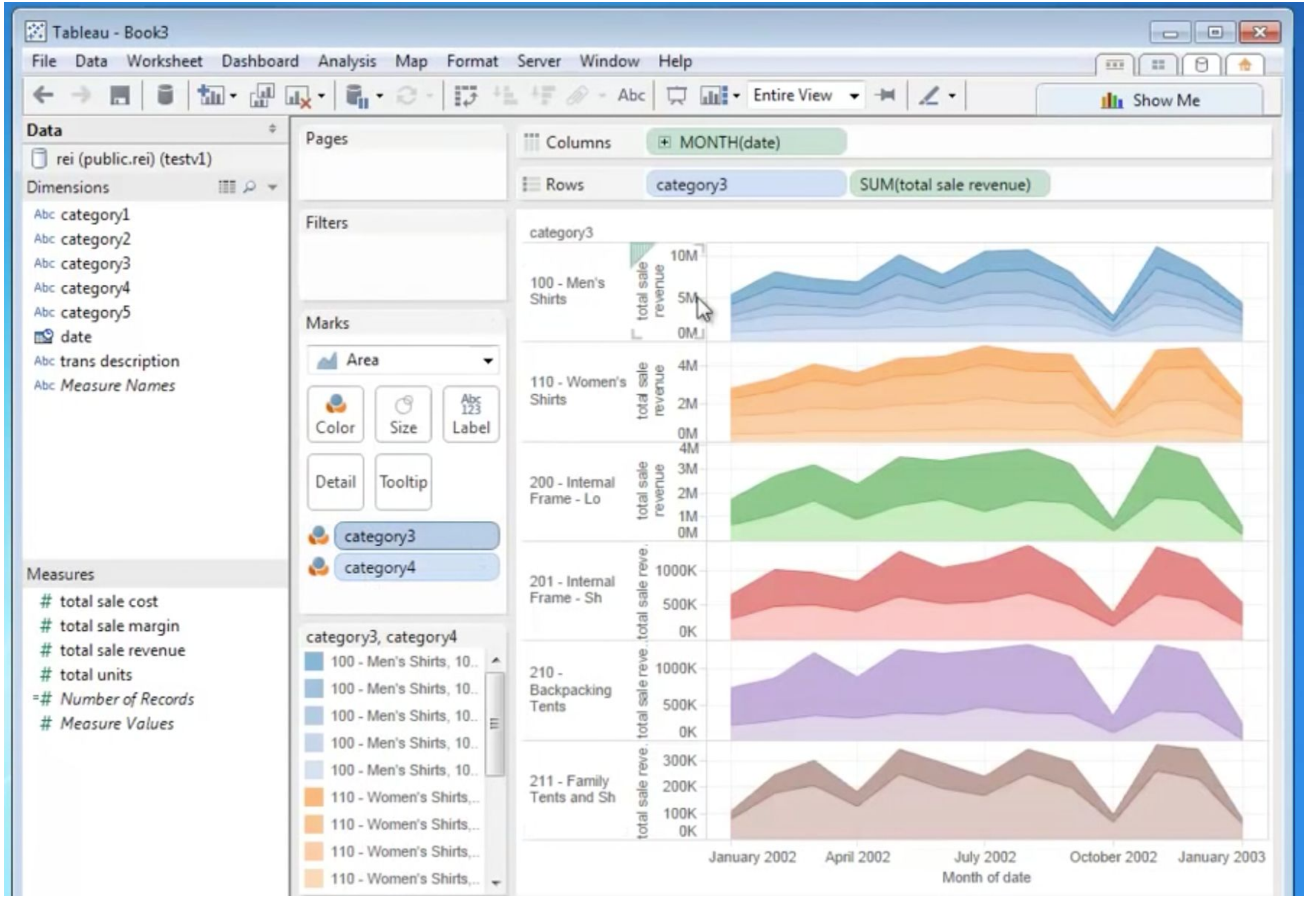
Step 5: Select a schema on the server:

Step 6: Select a table or view from the database:

◉ Single Table ○ Multiple Tables ○ Custom SQL

Step 7: Give the connection a name for use in Tableau:

Redshift Use Scenario



Columns: MONTH(date)

Rows: category3 SUM(total sale revenue)

Marks: Area

3

Amazon Machine Learning

Amazon Machine Learning makes it easy for developers to build machine learning model without learning complex algorithm or hiring experts.

Ideal Usage Pattern

Hard to code rules

- Rules are not explicit
- Number of factors are huge

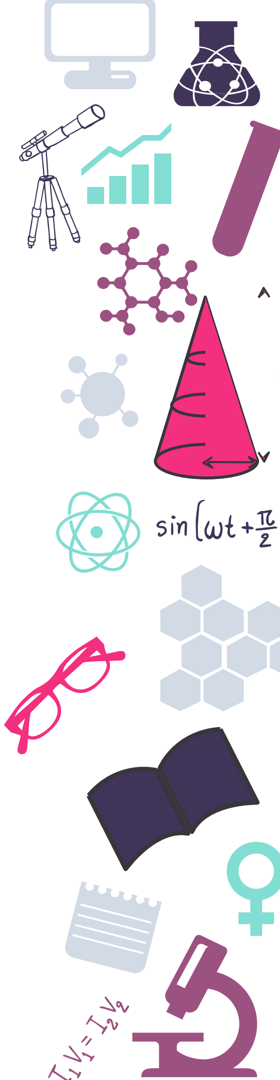
Hard to scale

- Large number of tasks
- Impossible to classify tasks manually

Datasources

Datasource is an object used by Amazon Machine Learning as train data, evaluation data and validation data

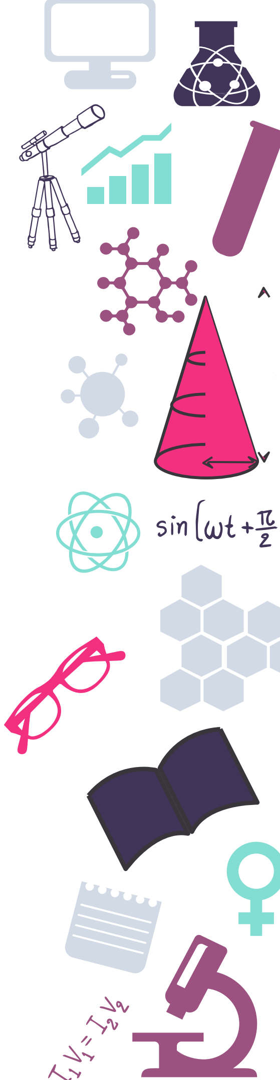
- ✓ Data should be well-formatted
- ✓ Datasources should contain one column as target



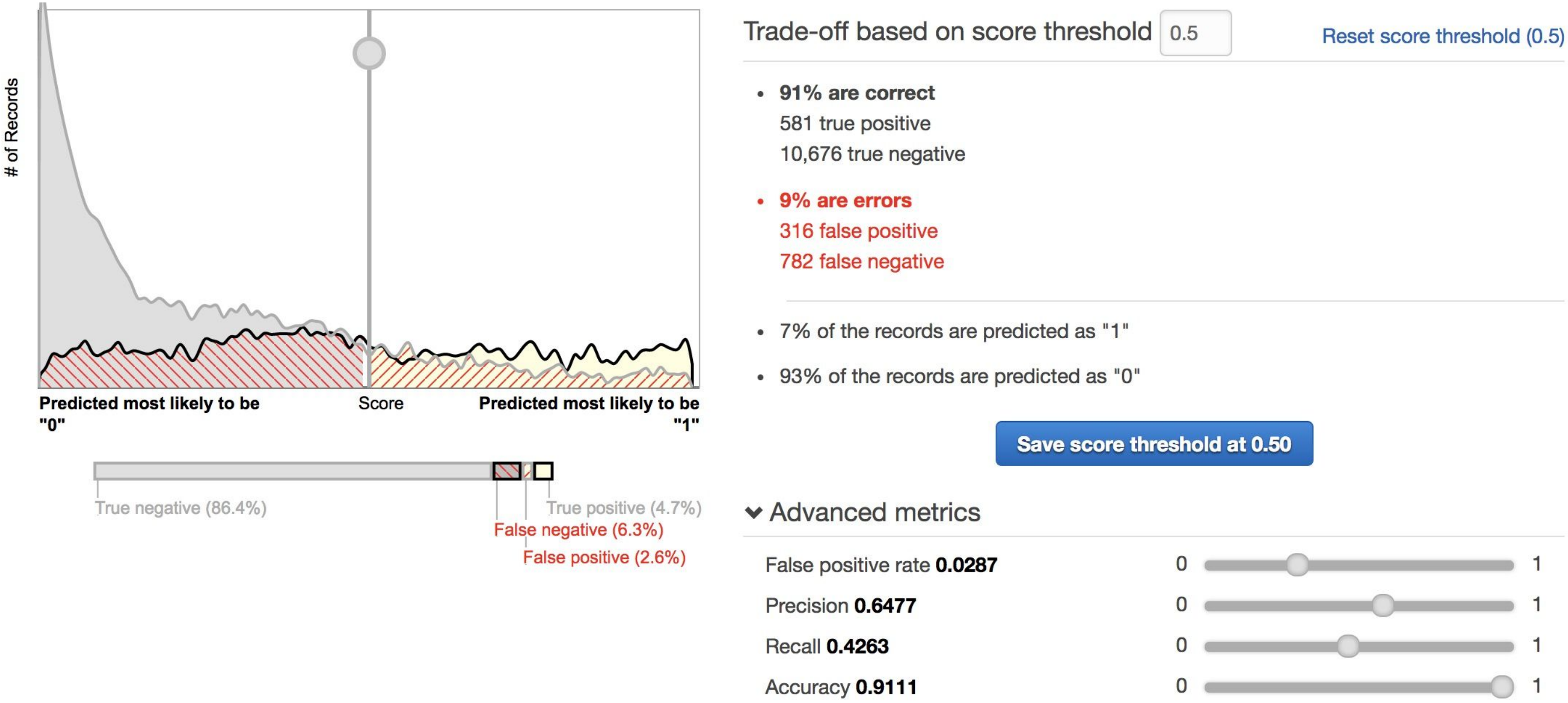
Train ML Models

Amazon ML applies machine learning algorithms automatically

- Binary Classification
- Multiclass Classification
- Regression Model



Evaluate ML Models



Trade-off based on score threshold 0.5

Reset score threshold (0.5)

- **91% are correct**
  581 true positive
  10,676 true negative
- **9% are errors**
  316 false positive
  782 false negative

- 7% of the records are predicted as "1"
- 93% of the records are predicted as "0"

Save score threshold at 0.50

Advanced metrics

False positive rate **0.0287** 0 1

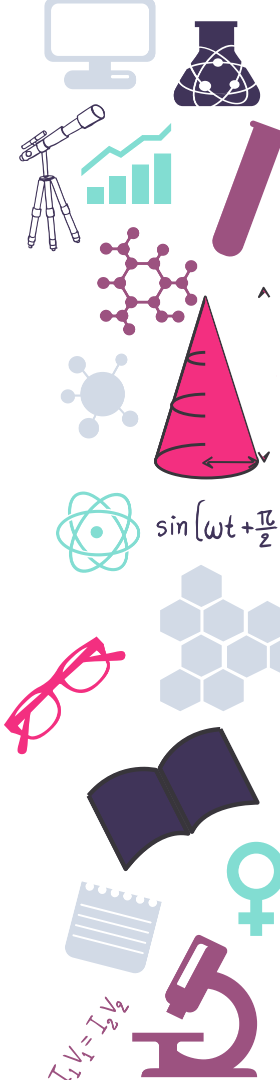
Precision **0.6477** 0 1

Recall **0.4263** 0 1

Accuracy **0.9111** 0 1

Generate and Interpret Prediction

- ✓ Batch Prediction
- ✓ Real-Time Prediction



Cost Model

Size of Model

Number of Predictions

Advantages & Disadvantages

Advantages

- Automatic
- Fast and easy

Disadvantages

- Black box
- Supervised model only

Thanks!

Any questions?