

Creating RDS Database

Select Standard create and Engine Type as MySQL, MySQL Community Edition


Choose a database creation method [Info](#)


☒ **Standard create**
You set all of the configuration options, including ones for availability, security, backups, and maintenance.


☐ **Easy create**
Use recommended best-practice configurations. Some configuration options can be changed after the database is created.


Engine options


Engine type [Info](#)


☐ Amazon Aurora


☒ MySQL


☐ MariaDB



☐ PostgreSQL


☐ Oracle


☐ Microsoft SQL Server


Edition

☒ MySQL Community

 **Known issues/limitations**
Review the [Known issues/limitations](#) to learn about potential compatibility issues with specific database versions.

Select Preferred template

Templates

Choose a sample template to meet your use case.



Production

Use defaults for high availability and fast, consistent performance.



Dev/Test

This instance is intended for development use outside of a production environment.



Free tier

Use RDS Free Tier to develop new applications, test existing applications, or gain hands-on experience with Amazon RDS. [Info](#)

Provide DB Name and setup Master Username and Password

Settings

DB instance identifier [Info](#)

Type a name for your DB instance. The name must be unique across all DB instances owned by your AWS account in the current AWS Region.

The DB instance identifier is case-insensitive, but is stored as all lowercase (as in "mydbinstance"). Constraints: 1 to 60 alphanumeric characters or hyphens. First character must be a letter. Can't contain two consecutive hyphens. Can't end with a hyphen.

▼ Credentials Settings

Master username [Info](#)

Type a login ID for the master user of your DB instance.

1 to 16 alphanumeric characters. First character must be a letter.



Manage master credentials in AWS Secrets Manager

Manage master user credentials in Secrets Manager. RDS can generate a password for you and manage it throughout its lifecycle.



Auto generate a password

Amazon RDS can generate a password for you, or you can specify your own password.

Master password [Info](#)

Constraints: At least 8 printable ASCII characters. Can't contain any of the following: / (slash), ' (single quote), " (double quote) and @ (at sign).

Confirm master password [Info](#)

Select the preferred instance and configure the storage space

Instance configuration

The DB instance configuration options below are limited to those supported by the engine that you selected above.

DB instance class [Info](#)

- ☐ Standard classes (includes m classes)
- ☐ Memory optimized classes (includes r and x classes)
- ☒ Burstable classes (includes t classes)

db.t4g.micro
2 vCPUs 1 GiB RAM Network: 2,085 Mbps

☐ Include previous generation classes

Storage

Storage type [Info](#)

General Purpose SSD (gp2)
Baseline performance determined by volume size

Allocated storage [Info](#)

100

GiB

The minimum value is 20 GiB and the maximum value is 6,144 GiB

Storage autoscaling [Info](#)

Provides dynamic scaling support for your database's storage based on your application's needs.

- ☒ Enable storage autoscaling
Enabling this feature will allow the storage to increase after the specified threshold is exceeded.

Maximum storage threshold [Info](#)


Charges will apply when your database autoscales to the specified threshold

200

GiB

The minimum value is 110 GiB and the maximum value is 6,144 GiB

Configure Connectivity (i.e., VPC, subnet, etc.,)

Connectivity [Info](#) 

Compute resource
Choose whether to set up a connection to a compute resource for this database. Setting up a connection will automatically change connectivity settings so that the compute resource can connect to this database.


☒ **Don't connect to an EC2 compute resource**
Don't set up a connection to a compute resource for this database. You can manually set up a connection to a compute resource later.

☐ **Connect to an EC2 compute resource**
Set up a connection to an EC2 compute resource for this database.

Virtual private cloud (VPC) [Info](#)
Choose the VPC. The VPC defines the virtual networking environment for this DB instance.

Default VPC (vpc-0c7445fe7db0429f6) ▼

Only VPCs with a corresponding DB subnet group are listed.

 After a database is created, you can't change its VPC.

DB Subnet group [Info](#)
Choose the DB subnet group. The DB subnet group defines which subnets and IP ranges the DB instance can use in the VPC that you selected.

default-vpc-0c7445fe7db0429f6 ▼

Public access [Info](#)

☒ **Yes**
RDS assigns a public IP address to the database. Amazon EC2 instances and other resources outside of the VPC can connect to your database. Resources inside the VPC can also connect to the database. Choose one or more VPC security groups that specify which resources can connect to the database.

☐ **No**
RDS doesn't assign a public IP address to the database. Only Amazon EC2 instances and other resources inside the VPC can connect to your database. Choose one or more VPC security groups that specify which resources can connect to the database.

VPC security group (firewall) [Info](#)
Choose one or more VPC security groups to allow access to your database. Make sure that the security group rules allow the appropriate incoming traffic.

☒ **Choose existing**
Choose existing VPC security groups

☐ **Create new**
Create new VPC security group

Existing VPC security groups

Choose one or more options ▼

Configure the database creation under advanced configuration

▼ Additional configuration

Database options, encryption turned on, backup turned off, backtrack turned off, maintenance, CloudWatch Logs, delete protection turned off.

Database options

Initial database name [Info](#)

MAPAssign

If you do not specify a database name, Amazon RDS does not create a database.

DB parameter group [Info](#)

default.mysql8.0 ▼

Option group [Info](#)

default:mysql-8-0 ▼

Backup

☐ Enable automated backups

Creates a point-in-time snapshot of your database

Encryption

☒ Enable encryption

Choose to encrypt the given instance. Master key IDs and aliases appear in the list after they have been created using the AWS Key Management Service console. [Info](#)

AWS KMS key [Info](#)

RHEK ▼

Account

126102239240

KMS key ID

5f838800-a8df-43be-92c6-78e7d57ba838

Click on create database

Maintenance

Auto minor version upgrade [Info](#)

☒ **Enable auto minor version upgrade**
Enabling auto minor version upgrade will automatically upgrade to new minor versions as they are released. The automatic upgrades occur during the maintenance window for the database.

Maintenance window [Info](#)
Select the period you want pending modifications or maintenance applied to the database by Amazon RDS.

☐ Choose a window

☒ No preference

Deletion protection

☐ **Enable deletion protection**
Protects the database from being deleted accidentally. While this option is enabled, you can't delete the database.


Estimated monthly costs

The Amazon RDS Free Tier is available to you for 12 months. Each calendar month, the free tier will allow you to use the Amazon RDS resources listed below for free:

- 750 hrs of Amazon RDS in a Single-AZ db.t2.micro, db.t3.micro or db.t4g.micro Instance.
- 20 GB of General Purpose Storage (SSD).
- 20 GB for automated backup storage and any user-initiated DB Snapshots.

[Learn more about AWS Free Tier.](#)

When your free usage expires or if your application use exceeds the free usage tiers, you simply pay standard, pay-as-you-go service rates as described in the [Amazon RDS Pricing page](#).

 You are responsible for ensuring that you have all of the necessary rights for any third-party products or services that you use with AWS services.

Cancel

Create database

Create a EMR Cluster

Select EMR release, use custom bundle and selected the required tools

Name and applications [Info](#)

Name

Mapreduceassign


Amazon EMR release [Info](#)

A release contains a set of applications which can be installed on your cluster.


emr-5.30.1

Application bundle


Spark



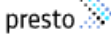
Core
Hadoop




HBase



Presto



Custom



▼ Customize your application bundle

Applications included in bundle

- | | |
|--|--|
| <input type="checkbox"/> Flink 1.10.0 | <input type="checkbox"/> Ganglia 3.7.2 |
| <input checked="" type="checkbox"/> HBase 1.4.13 | <input type="checkbox"/> HCatalog 2.3.6 |
| <input checked="" type="checkbox"/> Hadoop 2.8.5 | <input checked="" type="checkbox"/> Hive 2.3.6 |
| <input checked="" type="checkbox"/> Hue 4.6.0 | <input type="checkbox"/> JupyterHub 1.1.0 |
| <input type="checkbox"/> Livy 0.7.0 | <input type="checkbox"/> MXNet 1.5.1 |
| <input type="checkbox"/> Mahout 0.13.0 | <input type="checkbox"/> Oozie 5.2.0 |
| <input type="checkbox"/> Phoenix 4.14.3 | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input type="checkbox"/> Presto 0.232 | <input type="checkbox"/> Spark 2.4.5 |
| <input checked="" type="checkbox"/> Sqoop 1.4.7 | <input type="checkbox"/> TensorFlow 1.14.0 |
| <input type="checkbox"/> Tez 0.9.2 | <input type="checkbox"/> Zeppelin 0.8.2 |
| <input type="checkbox"/> ZooKeeper 3.4.14 | |

AWS Glue Data Catalog settings

Use the AWS Glue Data Catalog to provide an external metastore for your application.

- ☐ Use for Hive table metadata

Choose HDFS storage settings and select m4.xlarge in cluster configuration

HBase storage settings

Choose the storage layer for your data stored in HBase. The HDFS option uses the HBase default location for the root directory.

- ☒ Hadoop Distributed File System (HDFS)
- ☐ Amazon S3

Custom Amazon Machine Image (AMI) [Info](#)

- ☒ Update all installed packages on reboot

Cluster configuration [Info](#)

Choose a configuration method for the primary, core, and task node groups for your cluster.

- ☒ **Instance groups**
Choose one instance type per node group

- ☐ **Instance fleets**
Choose any combination of instance types within each node group

Instance groups

Primary

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: - Lowest Spot price: -

Actions ▼

- ☐ **Use multiple primary nodes**
To improve cluster availability, use 3 primary nodes with the same configuration and bootstrap actions. You can not use multiple primary nodes with instance fleets.

Set the cluster size and number of instances

Core

Remove instance group

Choose EC2 instance type

m4.xlarge
4 vCore 16 GiB memory EBS only storage
On-Demand price: - Lowest Spot price: -

Actions ▼

▶ Node configuration - optional

Add task instance group

You can add up to 48 more task instance groups.

▶ EBS root volume - optional

Cluster scaling and provisioning option [Info](#)

Amazon EMR console only supports EMR-managed scaling. To create a cluster with auto-scaling, use CLI or SDK.

☒ Set cluster size manually
Use this option if you know your workload patterns in advance.

☐ Use EMR-managed scaling
Monitor key workload metrics so that EMR can optimize the cluster size and resource utilization.

Name	Instance type	Size	Use Spot purchasing option
Core	m4.xlarge	<input type="text" value="3"/>	instance(s) <input type="checkbox"/>

Setup a security configuration

Security configuration and permissions [Info](#)

Security configuration - *optional*
Select your cluster encryption, authentication, and instance metadata service settings.

Choose a security configuration ▼

Browse

Create security configuration

Amazon EC2 key pair for SSH to the cluster - *optional* [Info](#)

Browse

Create key pair

Service role for Amazon EMR
Use a service role to call other AWS services when you provision resources and perform service-level actions.

EMR_DefaultRole ▼

Create IAM role

IAM role for instance profile
Use an instance profile so that application processes can call other AWS services when they run in the Hadoop framework on cluster instances.

EMR_EC2_DefaultRole ▼

Create IAM role

Click on create EMR cluster

Wait for EMR cluster to initialize

Amazon EMR > EMR on EC2: Clusters > Mapreduceassign

Actions ▼

Mapreduceassign

Summary			
Cluster ID j-MK866Z4KV32	Status Waiting	Creation time January 13, 2023, 15:34 (UTC+05:30)	Elapsed time 19 minutes
Capacity 1 Primary 3 Core 0 Task	Amazon EMR version emr-5.30.1	Applications HBase 1.4.13, Hadoop 2.8.5, Hive 2.3.6, Hue 4.6.0, Pig 0.17.0, Sqoop 1.4.7	After last step completes Cluster waits
Cluster configuration Instance groups	Primary node public DNS ec2-44-211-32-6.compute-1.amazonaws.com		

Add the initial 2 files to using EMR instance

[illegible]

Access the database using MySQL and create the table

```
hadoop@ip-172-31-72-58 mysql-connector-java-8.0.25]$ mysql -h mapreduceamitarjun.coucn8jmbfls.us-east-1.rds.amazonaws.com -u adminl -p
Enter password:
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 44
Server version: 8.0.28 Source distribution
```

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

```
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

```
MySQL [(none)]> Show Databases;
```

```

-----+
Database |
-----+
MAPAssign |
information_schema |
mysql |
performance_schema |
sys |
-----+
6 rows in set (0.01 sec)

```

```
mysql [(none)]> use MAPAssign;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
```

Database changed

```
mysql [MAPAssign]> create table lTaxi (
-> VendorID INT Not null,
-> tpep_pickup_datetime TIMESTAMP NOT NULL DEFAULT '0000-00-00 00:00:00',
-> tpep_dropoff_datetime TIMESTAMP NOT NULL DEFAULT '0000-00-00 00:00:00',
-> passenger_count INT,
-> trip_distance Double,
-> RatecodeID VARCHAR(255),
-> store_and_fwd_flag VARCHAR(255),
-> PULocationID VARCHAR(255),
-> DOLocationID VARCHAR(255),
-> payment_type VARCHAR(255),
-> fare_amount Double,
-> extra Double,
-> mta_tax Double,
-> tip_amount Double,
-> tolls_amount Double,
-> improvement_surcharge Double,
-> total_amount Double,
-> airport_fee Double
-> )
```

Query OK, 0 rows affected (0.03 sec)

Upload the dataset on to EMR before uploading the dataset on to MySQL

```
hadoop@ip-172-31-72-58:~/mapr_assignment/input_dataset
[hadoop@ip-172-31-72-58 ~]$ cd mapr_assignment/
[hadoop@ip-172-31-72-58 mapr_assignment]$ mkdir input_dataset
[hadoop@ip-172-31-72-58 mapr_assignment]$ cd input_dataset/
[hadoop@ip-172-31-72-58 input_dataset]$ wget "https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-07.csv"
--2023-01-13 11:44:58-- https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-07.csv
Resolving nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com]... 52.217.74.84, 52.217.235.241, 52.216.33.49, ...
Connecting to nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com] (52.217.74.84):443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 809052076 (772M) [text/csv]
Saving to: 'yellow_tripdata_2017-07.csv'

100%[-----] 809,052,076 66.6MB/s in 12s

2023-01-13 11:45:07 (66.4 MB/s) - 'yellow_tripdata_2017-07.csv' saved [809052076/809052076]

[hadoop@ip-172-31-72-58 input_dataset]$ wget https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
--2023-01-13 11:45:43-- https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-01.csv
Resolving nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com]... 52.216.112.59, 52.216.209.33, 52.217.17.48, ...
Connecting to nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com] (52.216.112.59):443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 914029540 (872M) [text/csv]
Saving to: 'yellow_tripdata_2017-01.csv'

100%[-----] 914,029,540 41.1MB/s in 23s

2023-01-13 11:46:06 (37.7 MB/s) - 'yellow_tripdata_2017-01.csv' saved [914029540/914029540]

[hadoop@ip-172-31-72-58 input_dataset]$ wget https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
--2023-01-13 11:46:14-- https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-02.csv
Resolving nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com]... 54.231.200.249, 3.8.7.18, 3.8.10.16, ...
Connecting to nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com] (54.231.200.249):443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 863487050 (823M) [text/csv]
Saving to: 'yellow_tripdata_2017-02.csv'

100%[-----] 863,487,050 37.0MB/s in 22s

2023-01-13 11:46:36 (37.5 MB/s) - 'yellow_tripdata_2017-02.csv' saved [863487050/863487050]

[hadoop@ip-172-31-72-58 input_dataset]$ wget https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
--2023-01-13 11:46:44-- https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-03.csv
Resolving nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com]... 3.5.20.102, 3.8.29.153, 52.217.164.49, ...
Connecting to nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com] (3.5.20.102):443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 969809025 (928M) [text/csv]
Saving to: 'yellow_tripdata_2017-03.csv'

100%[-----] 969,809,025 36.7MB/s in 24s

2023-01-13 11:47:08 (38.5 MB/s) - 'yellow_tripdata_2017-03.csv' saved [969809025/969809025]

[hadoop@ip-172-31-72-58 input_dataset]$ wget https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv
--2023-01-13 11:47:19-- https://nyct-tlc-upgrad.s3.amazonaws.com/yellow_tripdata_2017-04.csv
Resolving nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com]... 52.216.41.149, 52.216.145.75, 52.216.153.204, ...
Connecting to nyct-tlc-upgrad.s3.amazonaws.com [nyct-tlc-upgrad.s3.amazonaws.com] (52.216.41.149):443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 946349461 (903M) [text/csv]
Saving to: 'yellow_tripdata_2017-04.csv'

43%[-----] 415,446,286 21.0MB/s in 14s

Cannot write to 'yellow_tripdata_2017-04.csv' (Success).
[hadoop@ip-172-31-72-58 input_dataset]$
```

Upload the data from EMR to RDS

```
MySQL [MAPAssign]> Load data local infile '/home/hadoop/mapr_assignment/input_dataset/yellow_tripdata_2017-01.csv'
-> Into Table lTaxi
-> fields terminated by ','
-> lines terminated by '\n'
-> ignore 1 Lines;
```

Verify the data loaded

```
MySQL [MAPAssign]> Select Count(*) from lTaxi;
+-----+
| Count(*) |
+-----+
| 9710820 |
+-----+
1 row in set (25.94 sec)

MySQL [MAPAssign]> Select * from lTaxi Limit 10;
+-----+
| VendorID | tpep_pickup_datetime | tpep_dropoff_datetime | passenger_count | trip_distance | RatecodeID | store_and_fwd_flag | FULocationID | DOLocationID | payment_type | fare_amount | extra | mta_tax | tip_amount |
+-----+
| 1 | 2017-01-01 00:32:05 | 2017-01-01 00:37:48 | 1 | 1.2 | 1 | N | 140 | 236 | 2 | 6.5 | 0.5 | 0.5 | 0 |
| 0.3 | 7.8 | 0 |
| 1 | 2017-01-01 00:43:25 | 2017-01-01 00:47:42 | 2 | 0.7 | 1 | N | 237 | 140 | 2 | 5 | 0.5 | 0.5 | 0 |
| 0.3 | 6.3 | 0 |
| 1 | 2017-01-01 00:49:10 | 2017-01-01 00:53:53 | 2 | 0.8 | 1 | N | 140 | 237 | 2 | 5.5 | 0.5 | 0.5 | 0 |
| 0.3 | 6.8 | 0 |
| 1 | 2017-01-01 00:36:42 | 2017-01-01 00:41:09 | 1 | 1.1 | 1 | N | 41 | 42 | 2 | 6 | 0.5 | 0.5 | 0 |
| 0.3 | 7.3 | 0 |
| 1 | 2017-01-01 00:07:41 | 2017-01-01 00:18:16 | 1 | 3 | 1 | N | 48 | 263 | 2 | 11 | 0.5 | 0.5 | 0 |
| 0.3 | 12.3 | 0 |
| 1 | 2017-01-01 00:20:52 | 2017-01-01 00:24:59 | 2 | 0.7 | 1 | N | 236 | 262 | 2 | 5 | 0.5 | 0.5 | 0 |
| 0.3 | 6.3 | 0 |
| 1 | 2017-01-01 00:33:49 | 2017-01-01 00:42:38 | 2 | 1.6 | 1 | N | 236 | 238 | 1 | 8 | 0.5 | 0.5 | 1.85 |
| 0.3 | 11.15 | 0 |
| 1 | 2017-01-01 00:48:22 | 2017-01-01 00:52:15 | 2 | 0.6 | 1 | N | 238 | 239 | 1 | 5 | 0.5 | 0.5 | 1.25 |
| 0.3 | 7.55 | 0 |
| 1 | 2017-01-01 00:57:12 | 2017-01-01 01:06:18 | 2 | 1 | 1 | N | 239 | 48 | 1 | 7.5 | 0.5 | 0.5 | 1.75 |
| 0.3 | 10.55 | 0 |
| 1 | 2017-01-01 00:10:25 | 2017-01-01 00:29:06 | 1 | 1 | 1 | N | 246 | 48 | 2 | 12 | 0.5 | 0.5 | 0 |
| 0.3 | 13.3 | 0 |
+-----+
10 rows in set (0.01 sec)
```