# **Amazon Redshift FAQs**

<https://aws.amazon.com/redshift/faqs/>

General

**Q: What is Amazon Redshift?**

Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in [the cloud](https://aws.amazon.com/what-is-cloud-computing/). Customers can start small for just $0.25 per hour with no commitments or upfront costs and scale to a petabyte or more for $1,000 per terabyte per year, less than a tenth of most other data warehousing solutions.

Traditional data warehouses require significant time and resource to administer, especially for large datasets. In addition, the financial cost associated with building, maintaining, and growing self-managed, on-premise data warehouses is very high. Amazon Redshift not only significantly lowers the cost of a data warehouse, but also makes it easy to analyze large amounts of data very quickly.

Amazon Redshift gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections. Queries are distributed and parallelized across multiple physical resources. You can easily scale an Amazon Redshift data warehouse up or down with a few clicks in the AWS Management Console or with a single API call. Amazon Redshift automatically patches and backs up your data warehouse, storing the backups for a user-defined retention period. Amazon Redshift uses replication and continuous backups to enhance availability and improve data durability and can automatically recover from component and node failures. In addition, Amazon Redshift supports Amazon Virtual Private Cloud (Amazon VPC), SSL, AES-256 encryption and Hardware Security Modules (HSMs) to protect your data in transit and at rest.

As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use. Amazon Redshift lets you pay as you go. You can even try [Amazon Redshift for free.](https://aws.amazon.com/redshift/partners/)

Q: What does Amazon Redshift manage on my behalf?

Amazon Redshift manages the work needed to set up, operate, and scale a data warehouse, from provisioning the infrastructure capacity to automating ongoing administrative tasks such as backups, and patching. Amazon Redshift automatically monitors your nodes and drives to help you recover from failures.

Q: How does the performance of Amazon Redshift compare to most traditional databases for data warehousing and analytics?

Amazon Redshift uses a variety of innovations to achieve up to ten times higher performance than traditional databases for data warehousing and analytics workloads:

* *Columnar Data Storage*: Instead of storing data as a series of rows, Amazon Redshift organizes the data by column. Unlike row-based systems, which are ideal for transaction processing, column-based systems are ideal for data warehousing and analytics, where queries often involve aggregates performed over large data sets. Since only the columns involved in the queries are processed and columnar data is stored sequentially on the storage media, column-based systems require far fewer I/Os, greatly improving query performance.
* *Advanced Compression:* Columnar data stores can be compressed much more than row-based data stores because similar data is stored sequentially on disk. Amazon Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores. In addition, Amazon Redshift doesn't require indexes or materialized views and so uses less space than traditional relational database systems. When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.
* *Massively Parallel Processing (MPP):* Amazon Redshift automatically distributes data and query load across all nodes. Amazon Redshift makes it easy to add nodes to your data warehouse and enables you to maintain fast query performance as your data warehouse grows.

Q: How do I get started with Amazon Redshift?  
  
You can sign up and get started within minutes from the Amazon Redshift detail page or via the AWS Management Console. If you don't already have an AWS account, you'll be prompted to create one. Visit our [Getting Started](https://aws.amazon.com/redshift/getting-started/)Page to see how to try Amazon Redshift for free.

Q: How do I create an Amazon Redshift data warehouse cluster?  
  
You can easily create an Amazon Redshift data warehouse cluster by using the [AWS Management Console](https://aws.amazon.com/console/) or the [Amazon Redshift APIs.](http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html) You can start with a single node, 160GB data warehouse and scale all the way to a petabyte or more with a few clicks in the [AWS Console](https://aws.amazon.com/console/) or a single API call.

The single node configuration enables you to get started with Amazon Redshift quickly and cost-effectively and scale up to a multi-node configuration as your needs grow. The multi-node configuration requires a leader node that manages client connections and receives queries, and two compute nodes that store data and perform queries and computations. The leader node is provisioned for you automatically and you are not charged for it.

Simply specify your preferred Availability Zone (optional), the number of nodes, node types, a master name and password, security groups, your preferences for backup retention, and other system settings. Once you've chosen your desired configuration, Amazon Redshift will provision the required resources and set up your data warehouse cluster.

Q: What does a leader node do? What does a compute node do?   
  
A leader node receives queries from client applications, parses the queries and develops execution plans, which are an ordered set of steps to process these queries. The leader node then coordinates the parallel execution of these plans with the compute nodes, aggregates the intermediate results from these nodes and finally returns the results back to the client applications.

Compute nodes execute the steps specified in the execution plans and transmit data among themselves to serve these queries. The intermediate results are sent back to the leader node for aggregation before being sent back to the client applications.

Q: What is the maximum storage capacity per compute node? What is the recommended amount of data per compute node for optimal performance?   
  
You can create a cluster using either Dense Storage (DS) nodes or Dense Compute nodes (DC). Dense Storage nodes allow you to create very large data warehouses using hard disk drives (HDDs) for a very low price point. Dense Compute nodes allow you to create very high performance data warehouses using fast CPUs, large amounts of RAM and solid-state disks (SSDs).

Dense Storage (DS) nodes are available in two sizes, Extra Large and Eight Extra Large. The Extra Large (XL) has 3 HDDs with a total of 2TB of magnetic storage, whereas Eight Extra Large (8XL) has 24 HDDs with a total of 16TB of magnetic storage. DS2.8XL has 36 Intel Xeon E5-2676 v3 (Haswell) virtual cores and 244GiB of RAM, and DS2.XL has 4 Intel Xeon E5-2676 v3 (Haswell) virtual cores and 31GiB of RAM. Please see our pricing page for more detail. You can get started with a single Extra Large node, 2TB data warehouse for $0.85 per hour and scale up to a petabyte or more. You can pay by the hour or use reserved instance pricing to lower your price to under $1,000 per TB per year.

Dense Compute (DC) nodes are also available in two sizes. The Large has 160GB of SSD storage, 2 Intel Xeon E5-2670v2 (Ivy Bridge) virtual cores and 15GiB of RAM. The Eight Extra Large is sixteen times bigger with 2.56TB of SSD storage, 32 Intel Xeon E5-2670v2 virtual cores and 244GiB of RAM. You can get started with a single Large node for $0.25 per hour and and scale all the way up to 128 8XL nodes with 326TB of SSD storage, 3,200 virtual cores and 24TiB of RAM.

Amazon Redshift's MPP architecture means you can increase your performance by increasing the number of nodes in your data warehouse cluster. The optimal amount of data per compute node depends on your application characteristics and your query performance needs.

Q: How many nodes can I specify per Amazon Redshift data warehouse cluster?  
  
An Amazon Redshift data warehouse cluster can contain from 1-128 compute nodes, depending on the node type. For details please see [our documentation](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html).

Q: How do I access my running data warehouse cluster?  
  
Once your data warehouse cluster is available, you can retrieve its endpoint and JDBC and ODBC connection string from the AWS Management Console or by using the Redshift APIs. You can then use this connection string with your favorite database tool, programming language, or Business Intelligence (BI) tool. You will need to authorize network requests to your running data warehouse cluster. For a detailed explanation please refer to our [Getting Started Guide](http://docs.aws.amazon.com/redshift/latest/gsg/welcome.html).

Q: When would I use Amazon Redshift vs. Amazon RDS?  
  
Both Amazon Redshift and [Amazon RDS](https://aws.amazon.com/rds/) enable you to run traditional relational databases in the cloud while offloading database administration. Customers use Amazon RDS databases both for online-transaction processing (OLTP) and for reporting and analysis. Amazon Redshift harnesses the scale and resources of multiple nodes and uses a variety of optimizations to provide order of magnitude improvements over traditional databases for analytic and reporting workloads against very large data sets. Amazon Redshift provides an excellent scale-out option as your data and query complexity grows or if you want to prevent your reporting and analytic processing from interfering with the performance of your OLTP workload.

Q: When would I use Amazon Redshift vs. Amazon Elastic MapReduce (Amazon EMR)?  
  
Amazon Redshift is ideal for large volumes of structured data that you want to persist and query using standard SQL and your existing BI tools. [Amazon EMR](https://aws.amazon.com/emr/) is ideal for processing and transforming unstructured or semi-structured data to bring in to Amazon Redshift and is also a much better option for data sets that are relatively transitory, not stored for long-term use.

Q: Why should I use Amazon Redshift instead of running my own MPP data warehouse cluster on Amazon EC2?  
  
Amazon Redshift automatically handles many of the time-consuming tasks associated with managing your own data warehouse including:

* *Setup:* With Amazon Redshift, you simply create a data warehouse cluster, define your schema, and begin loading and querying your data. Provisioning, configuration and patching are all managed for you.
* *Data Durability:*Amazon Redshift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3, which is designed for eleven nines of durability. Amazon Redshift mirrors each drive's data to other nodes within your cluster. If a drive fails, your queries will continue with a slight latency increase while Redshift rebuilds your drive from replicas. In case of node failure(s), Amazon Redshift automatically provisions new node(s) and begins restoring data from other drives within the cluster or from Amazon S3. It prioritizes restoring your most frequently queried data so your most frequently executed queries will become performant quickly.
* *Scaling:* You can add or remove nodes from your Amazon Redshift data warehouse cluster with a single API call or via a few clicks in the AWS Management Console as your capacity and performance needs change.
* *Automatic Updates and Patching:* Amazon Redshift automatically applies upgrades and patches your data warehouse so you can focus on your application and not on its administration.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Billing

Q: How will I be charged and billed for my use of Amazon Redshift?  
  
You pay only for what you use, and there are no minimum or setup fees. You are billed based on:

* *Compute node hours* – Compute node hours are the total number of hours you run across all your compute nodes for the billing period. You are billed for 1 unit per node per hour, so a 3-node data warehouse cluster running persistently for an entire month would incur 2,160 instance hours. You will not be charged for leader node hours; only compute nodes will incur charges.
* *Backup Storage* – Backup storage is the storage associated with your automated and manual snapshots for your data warehouse. Increasing your backup retention period or taking additional snapshots increases the backup storage consumed by your data warehouse. There is no additional charge for backup storage up to 100% of your provisioned storage for an active data warehouse cluster. For example, if you have an active Single Node XL data warehouse cluster with 2TB of local instance storage, we will provide up to 2TB-Month of backup storage at no additional charge. Backup storage beyond the provisioned storage size and backups stored after your cluster is terminated are billed at [standard Amazon S3 rates](https://aws.amazon.com/s3/pricing/).
* *Data transfer* – There is no Data Transfer charge for data transferred to or from Amazon Redshift outside of Amazon VPC. Data Transfer to or from Redshift in [Amazon VPC](https://aws.amazon.com/vpc/) accrues standard AWS data transfer charges.

For Amazon Redshift pricing information, please visit the [Amazon Redshift pricing page.](https://aws.amazon.com/redshift/pricing/)

Q: When does billing of my Amazon Redshift data warehouse clusters begin and end?  
  
Billing commences for a data warehouse cluster as soon as the data warehouse cluster is available. Billing continues until the data warehouse cluster terminates, which would occur upon deletion or in the event of instance failure.

Q: What defines billable Amazon Redshift instance hours?  
  
Node usage hours are billed for each hour your data warehouse cluster is running in an available state. If you no longer wish to be charged for your data warehouse cluster, you must terminate it to avoid being billed for additional node hours. Partial node hours consumed are billed as full hours.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. [Learn more](https://aws.amazon.com/c-tax-faqs/).

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Data Integration and Loading

Q: How do I load data into my Amazon Redshift data warehouse?  
  
You can load data into Amazon Redshift from a range of data sources including [Amazon S3](https://aws.amazon.com/s3/), [Amazon DynamoDB](https://aws.amazon.com/dynamodb/), [Amazon EMR](https://aws.amazon.com/emr/), [AWS Data Pipeline](https://aws.amazon.com/datapipeline/) and or any SSH-enabled host on Amazon EC2 or on-premises. Amazon Redshift attempts to load your data in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster. For more details on loading data into Amazon Redshift please view our [Getting Started Guide](http://docs.aws.amazon.com/redshift/latest/gsg/welcome.html).

Q: Can I load data using SQL ‘INSERT' statements?  
  
Yes, clients can connect to Amazon Redshift using ODBC or JDBC and issue 'insert' SQL commands to insert the data. Please note this is slower than using S3 or DynamoDB since those methods load data in parallel to each compute node while SQL insert statements load via the single leader node.

Q: How do I load data from my existing Amazon RDS, Amazon EMR, Amazon DynamoDB, and Amazon EC2 data sources to Amazon Redshift?  
  
You can use our [COPY command](http://docs.aws.amazon.com/redshift/latest/dg/r_COPY.html) to load data in parallel directly to Amazon Redshift from Amazon EMR, Amazon DynamoDB, or any SSH-enabled host. Moreover, many [ETL companies](https://aws.amazon.com/redshift/partners/) have certified Amazon Redshift for use with their tools, and a number are offering [free trials](https://aws.amazon.com/redshift/partners/) to help you get started loading your data. Finally, [AWS Data Pipeline](https://aws.amazon.com/datapipeline/) provides a high performance, reliable, fault tolerant solution to load data from a variety of AWS data sources. You can use AWS Data Pipeline to specify the data source, desired data transformations, and then execute a pre-written import script to load your data into Amazon Redshift.

Q: I have a lot of data for initial loading into Amazon Redshift. Transferring via the Internet would take a long time. How do I load this data?  
  
You can use [AWS Import/Export](https://aws.amazon.com/snowball/) to transfer the data to Amazon S3 using portable storage devices. In addition, you can use [AWS Direct Connect](https://aws.amazon.com/directconnect/) to establish a private network connection between your network or datacenter and AWS. You can choose 1Gbit/sec or 10Gbit/sec connection ports to transfer your data.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Security

Q: How does Amazon Redshift keep my data secure?  
  
Amazon Redshift encrypts and keeps your data secure in transit and at rest using industry-standard encryption techniques. To keep data secure in transit, Amazon Redshift supports SSL-enabled connections between your client application and your Redshift data warehouse cluster. To keep your data secure at rest, Amazon Redshift encrypts each block using hardware-accelerated AES-256 as it is written to disk. This takes place at a low level in the I/O subsystem, which encrypts everything written to disk, including intermediate query results. The blocks are backed up as is, which means that backups are encrypted as well. By default, Amazon Redshift takes care of key management but you can choose [to manage your keys using your own hardware security modules (HSMs)](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-HSM.html) or manage your keys through [AWS Key Management Service](https://aws.amazon.com/kms/).

Q: Can I use Amazon Redshift in Amazon Virtual Private Cloud (Amazon VPC)?  
  
Yes, you can use Amazon Redshift as part of your VPC configuration. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon Redshift data warehouse cluster.

Q: Can I access my Amazon Redshift compute nodes directly?  
  
No. Your Amazon Redshift compute nodes are in a private network space and can only be accessed from your data warehouse cluster's leader node. This provides an additional layer of security for your data.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Availability and Durability

Q: What happens to my data warehouse cluster availability and data durability if a drive on one of my nodes fails?  
  
Your Amazon Redshift data warehouse cluster will remain available in the event of a drive failure however you may see a slight decline in performance for certain queries. In the event of a drive failure, Amazon Redshift will transparently use a replica of the data on that drive which is stored on other drives within that node. In addition, Amazon Redshift will attempt to move your data to a healthy drive or will replace your node if it is unable to do so. Single node clusters do not support data replication. In the event of a drive failure you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability in the event of individual node failure?  
  
Amazon Redshift will automatically detect and replace a failed node in your data warehouse cluster. The data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB. Amazon Redshift makes your replacement node available immediately and loads your most frequently accessed data from S3 first to allow you to resume querying your data as quickly as possible. Single node clusters do not support data replication. In the event of a drive failure you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability in the event if my data warehouse cluster's Availability Zone (AZ) has an outage?  
  
If your Amazon Redshift data warehouse cluster's Availability Zone becomes unavailable, you will not be able to use your cluster until power and network access to the AZ are restored. Your data warehouse cluster's data is preserved so you can start using your Amazon Redshift data warehouse as soon as the AZ becomes available again. In addition, you can also choose to restore any existing snapshots to a new AZ in the same Region. Amazon Redshift will restore your most frequently accessed data first so you can resume queries as quickly as possible.

Q: Does Amazon Redshift support Multi-AZ Deployments?  
  
Currently, Amazon Redshift only supports Single-AZ deployments. You can run data warehouse clusters in multiple AZ's by loading data into two Amazon Redshift data warehouse clusters in separate AZs from the same set of Amazon S3 input files. In addition, you can also restore a data warehouse cluster to a different AZ from your data warehouse cluster snapshots.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Backup and Restore

Q: How does Amazon Redshift back up my data?  
  
Amazon Redshift replicates all your data within your data warehouse cluster when it is loaded and also continuously backs up your data to S3. Amazon Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3). Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery.

Q: How long does Amazon Redshift retain backups? Is it configurable?  
  
By default, Amazon Redshift retains backups for 1 day. You can configure this to be as long as 35 days.

Q: How do I restore my Amazon Redshift data warehouse cluster from a backup?  
  
You have access to all the automated backups within your backup retention window. Once you choose a backup from which to restore, we will provision a new data warehouse cluster and restore your data to it.

Q: Do I need to enable backups for my data warehouse cluster or is it done automatically?  
  
By default, Amazon Redshift enables automated backups of your data warehouse cluster with a 1-day retention period. Free backup storage is limited to the total size of storage on the nodes in the data warehouse cluster and only applies to active data warehouse clusters. For example, if you have total data warehouse storage of 8TB, we will provide at most 8TB of backup storage at no additional charge. If you would like to extend your backup retention period beyond one day, you can do so using the [AWS Management Console](https://aws.amazon.com/console/) or the [Amazon Redshift APIS](http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html). For more information on automated snapshots, please refer to the Amazon Redshift [Management Guide](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-snapshots.html). Amazon Redshift only backs up data that has changed so most snapshots only use up a small amount of your free backup storage.

Q: How do I manage the retention of my automated backups and snapshots?  
  
You can use the [AWS Management Console](https://aws.amazon.com/console/) or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended).

Q: What happens to my backups if I delete my data warehouse cluster?  
  
When you delete a data warehouse cluster, you have the ability to specify whether a final snapshot is created upon deletion, which enables a restore of the deleted data warehouse cluster at a later date. All previously created manual snapshots of your data warehouse cluster will be retained and billed at [standard Amazon S3 rates](https://aws.amazon.com/s3/pricing/), unless you choose to delete them.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Scalability

Q: How do I scale the size and performance of my Amazon Redshift data warehouse cluster?  
  
If you would like to increase query performance or respond to CPU, memory or I/O over-utilization, you can increase the number of nodes within your data warehouse cluster via the [AWS Management Console](https://aws.amazon.com/console/) or the ModifyCluster API. When you modify your data warehouse cluster, your requested changes will be applied immediately. Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the [AWS Management Console](https://aws.amazon.com/console/) or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via [Amazon Cloudwatch's](https://aws.amazon.com/cloudwatch/) custom metric functionality.

Q: Will my data warehouse cluster remain available during scaling?  
  
The existing data warehouse cluster remains available for read operations while a new data warehouse cluster gets created during scaling operations. When the new data warehouse cluster is ready, your existing data warehouse cluster will be temporarily unavailable while the canonical name record of the existing data warehouse cluster is flipped to point to the new data warehouse cluster. This period of unavailability typically lasts only a few minutes, and will occur during the maintenance window for your data warehouse cluster, unless you specify that the modification should be applied immediately. Amazon Redshift moves data in parallel from the compute nodes in your existing data warehouse cluster to the compute nodes in your new cluster. This enables your operation to complete as quickly as possible.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

Querying and Analysis

Q: Is Amazon Redshift compatible with my preferred business intelligence software package and ETL tools?  
  
Amazon Redshift uses industry-standard SQL and is accessed using standard JDBC and ODBC drivers. You can download Amazon Redshift custom JDBC and ODBC drivers from the Connect Client tab of our [Console](https://console.aws.amazon.com/redshift/). We have validated integrations with [popular BI and ETL vendors](https://aws.amazon.com/redshift/partners/), a number of which are offering [free trials](https://aws.amazon.com/redshift/partners/) to help you get started loading and analyzing your data. You can also go to the [AWS Marketplace](https://aws.amazon.com/redshift/getting-started/) to deploy and configure solutions designed to work with Amazon Redshift in minutes.

Monitoring

Q: How do I monitor the performance of my Amazon Redshift data warehouse cluster?

Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the [AWS Management Console](https://aws.amazon.com/console/) or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via [Amazon Cloudwatch’s](https://aws.amazon.com/cloudwatch/) custom metric functionality. In addition to CloudWatch metrics, Amazon Redshift also provides information on query and cluster performance via the AWS Management Console. This information enables you to see which users and queries are consuming the most system resources and diagnose performance issues. In addition, you can see the resource utilization on each of your compute nodes to ensure that you have data and queries that are well balanced across all nodes.

Maintenance

Q: What is a maintenance window? Will my data warehouse cluster be available during software maintenance?

Amazon Redshift periodically performs maintenance to apply fixes, enhancements and new features to your cluster. You can change the scheduled maintenance windows by modifying the cluster, either programmatically or by using the Amazon Redshift Console. During these maintenance windows, your Amazon Redshift cluster is not available for normal operations. For more information about maintenance windows and schedules by region, see [Maintenance Windows](https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-maintenance-windows) in the *Amazon Redshift Management Guide*.