



Masterclass



Amazon Redshift

Ian Massingham – Technical Evangelist

 @IanMmmm

Masterclass

- A technical deep dive beyond the basics
- Help educate you on how to get the best from AWS technologies
- Show you how things work and how to get things done
- Broaden your knowledge in ~45 mins

Amazon Redshift

A fast, fully managed, petabyte-scale data warehouse

Makes it simple & cost-effective to analyse all your data using existing tools

Start small & scale to a petabyte or more for less than \$1,000/terabyte/year

Optimised for
Data Warehousing

Scalable

No Up-Front Costs

Amazon Redshift

Compatible

Secure

Simple



Agenda

Why Run Your Data Warehouse on AWS? ◀

Getting Started ◀

Table Design ◀

Data Loading ◀

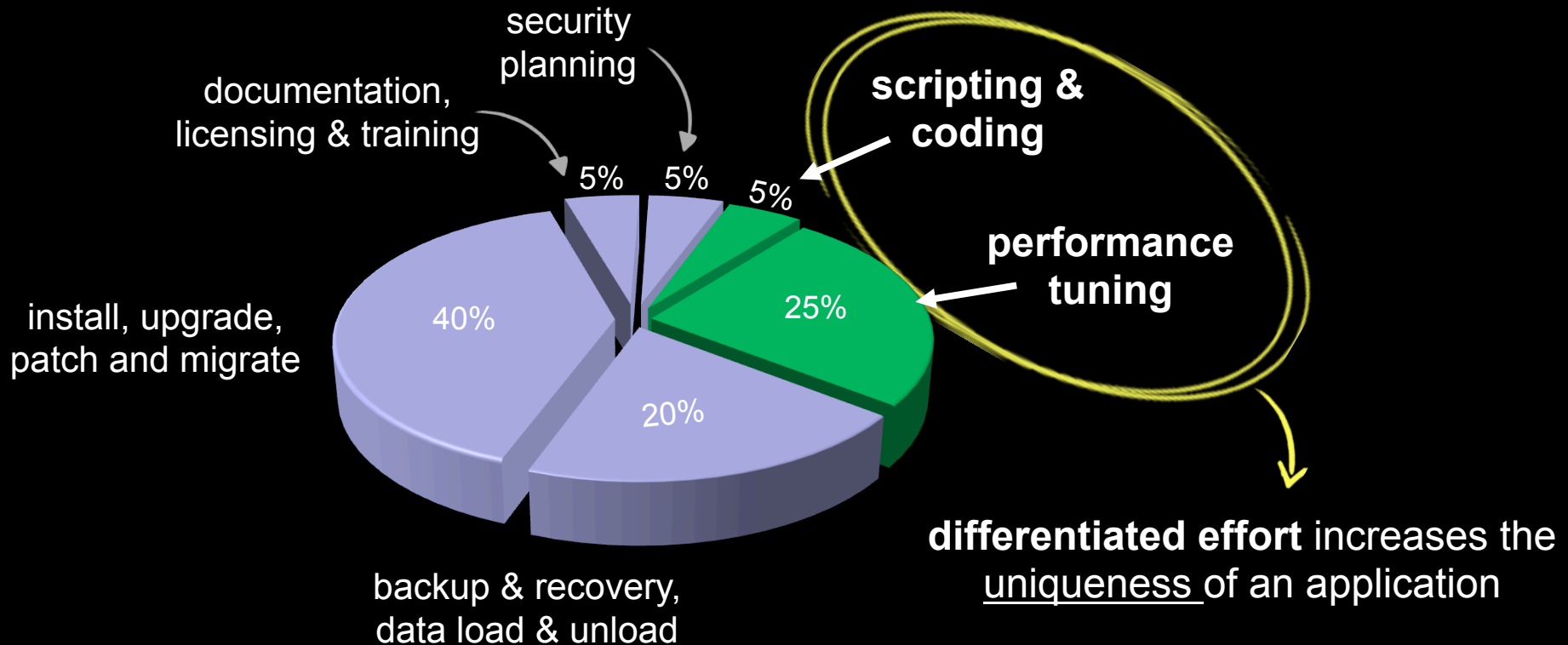
Working with Data ◀

Backup and Restoration ◀

Upgrading & Scaling ◀

Why Run Your Data
Warehouse on AWS?

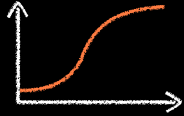
Why Managed Databases?



Our customers asked us for data warehousing done the AWS way



Easy to provision and scale up massively



No upfront costs, pay as you go



Really fast performance at a really low price



Open and flexible with support for popular BI tools

Amazon Redshift parallelizes and distributes everything

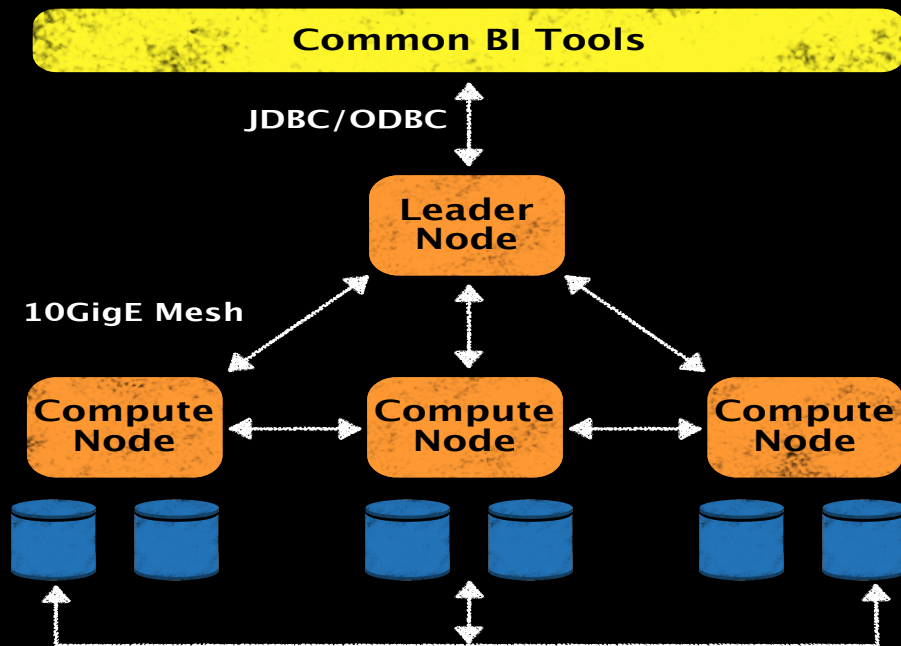
Query

Load

Backup

Restore

Resize



Redshift lets you start small and grow big

Extra Large Node (dw1.xl & dw2.xl)

3 spindles, 15GiB RAM
2 virtual cores, 10GigE

Single Node (160GB SSD or 2TB Magnetic)

XL

Cluster 2-32 Nodes (320GB SSD – 64TB Magnetic)

XL	XL	XL	XL	XL	XL	XL	XL
XL	XL	XL	XL	XL	XL	XL	XL
XL	XL	XL	XL	XL	XL	XL	XL
XL	XL	XL	XL	XL	XL	XL	XL

8 Extra Large Node (dw1.8xl & dw2.8xl)

24 spindles, 120GiB RAM, 1.2TB SSD or 16TB
Magnetic, 16 virtual cores, 10GigE

Cluster 2-100 Nodes (2.4TB SSD – 1.6PB Magnetic)

[illegible]

Redshift simplifies operations

- Built-in security in transit, at rest, when backed up
- Backup to S3 is continuous, incremental, and automatic
- Streaming restores let you resume querying faster
- Disk failures are transparent; nodes recover automatically



Amazon Redshift dramatically reduces IO

Column storage

Data compression

Zone maps

Direct-attached storage

Large data block sizes

Id	Age	State
123	20	CA
345	25	WA
678	40	FL



Row storage

Column storage

amazon.com

Initial Pilot Results

Current production environment

32 nodes, 128 CPUs, 4.2TB RAM, 1.6 PB disk

Tested 2B row data set, 6 representative queries on

2-node Amazon Redshift cluster



queries ran > 10x faster

amazon.com

TEST:

2 BILLION RECORDS

6 REPRESENTATIVE REQUESTS

amazon.com

AMAZON REDSHIFT 2x HS1.8XL

Vs.

32 NODES, 4.2TB RAM, 1.6PB

amazon.com

12x TO 150x FASTER

<u>Query Impact</u>	<u>Before (s)</u>	<u>After (s)</u>	<u>Speedup</u>
29m 58s -> 12s	1,798	12	149.8x
22m 2s -> 16s	1,322	16	82.6x
7m 48s -> 3s	468	3	156.x
5m 30s -> 5s	330	5	66.x
49m 8s -> 35s	2,948	35	84.2x
2m 9s -> 10s	129	10	12.9x
Total Speedup	6,995	81	86.4x

- Needed a way to increase speed, performance and flexibility of data analysis at a low cost
- Using AWS enabled FT to run queries 98% faster than previously—helping FT make business decisions quickly
- Easier to track and analyze trends
- Reduced infrastructure costs by 80% over traditional data center model



When our analysts first started to do queries on Amazon Redshift, they thought it was broken because it was working so fast.

John O'Donovan
CTO, Financial Times

FINANCIAL TIMES

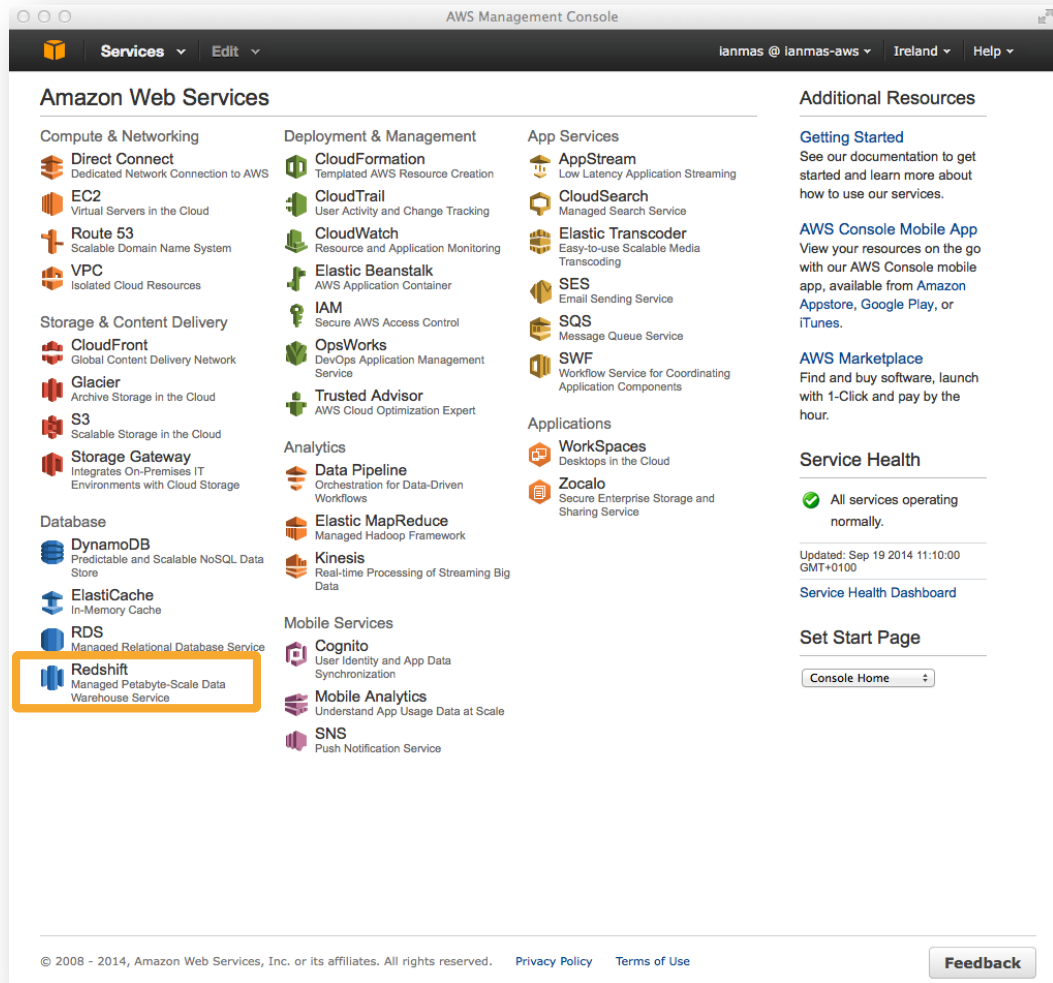


Financial Times is one of the world's leading business news organizations, with a combined paid print and digital circulation of 665,000.

Getting
Started

Create a Data Warehouse in Minutes with Amazon Redshift

Creating a Redshift Data Warehouse via the AWS Console



Creating a Redshift Data Warehouse via the AWS Console

The screenshot shows the Amazon Redshift console in the AWS Management Console. The top navigation bar includes the AWS logo, 'Services', 'Edit', and user information 'ianmas @ ianmas-aws' with a dropdown menu showing 'Ireland' and 'Help'. The left sidebar lists navigation options: 'Amazon Redshift', 'Clusters', 'Snapshots', 'Security', 'Parameter Groups', 'Reserved Nodes', and 'Events'. The main content area is titled 'Welcome to Amazon Redshift' and includes a message: 'You do not appear to have any clusters in the EU West (Ireland) region.' Below this is a paragraph explaining Amazon Redshift as a fast and powerful, fully managed, petabyte-scale data warehouse service. A 'Try Amazon Redshift for free!' section details a two-month free trial of the dw2.large node, 750 hours per month, and 160GB of compressed SSD storage. A 'Launch Cluster' button is highlighted with an orange border. Below the button, a section titled 'Get up and running immediately' contains three cards: 'Create Cluster' (with a cube icon), 'Manage & Configure' (with a cube and gear icon), and 'Load & Query Data' (with a cube and data flow icon). Each card has a 'Learn More' link. At the bottom, an 'Additional Information' section links to 'Overview of Features', 'New Features', 'Subscribe to Announcements', and 'Documentation'. The footer contains copyright information '© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.', links for 'Privacy Policy' and 'Terms of Use', and a 'Feedback' button.

Redshift - AWS Console

Services Edit ianmas @ ianmas-aws Ireland Help

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

Welcome to Amazon Redshift

You do not appear to have any clusters in the EU West (Ireland) region.

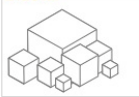
Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Amazon Redshift offers you fast query performance when analyzing virtually any size data set using the same SQL-based tools and business intelligence applications you use today. With a few clicks in the AWS Management Console, you can launch a Redshift cluster, starting with a few hundred gigabytes of data and scaling to a petabyte or more, for under \$1,000 per terabyte per year.

Try Amazon Redshift for free! If you've never created an Amazon Redshift cluster, you're eligible for a two month free trial of our dw2.large node. You get 750 hours per month for free, enough hours to continuously run one dw2.large node with 160GB of compressed SSD storage. You can also build clusters with multiple nodes to test larger data sets, which will consume your free hours more quickly. Once your two month free trial expires or your usage exceeds 750 hours per month, you can shut down your cluster, avoiding any charges, or keep it running at our standard [On-Demand Rate](#). For more information, please see the [free trial FAQ page](#).

Launch Cluster

Get up and running immediately

Create Cluster



[Learn More](#)

Manage & Configure



[Learn More](#)

Load & Query Data



[Learn More](#)

Additional Information

- [Overview of Features](#)
- [New Features](#)
- [Subscribe to Announcements](#)
- [Documentation](#)

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Feedback](#)

Creating a Redshift Data Warehouse via the AWS Console

Redshift · AWS Console

Services Edit

ianmas @ ianmas-aws Ireland Help

Amazon Redshift

Clusters

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

CLUSTER DETAILS

NODE CONFIGURATION

ADDITIONAL CONFIGURATION

REVIEW

Provide the details of your cluster. Fields marked with * are required.

Cluster Identifier*

This is the unique key that identifies a cluster. This parameter is stored as a lowercase string. (e.g. my-dw-instance)

Database Name

Optional. A default database named dev is created for the cluster. Optionally, specify a custom database name (e.g. mydb) to create an additional database.

Database Port*

5439

Port number on which the database accepts connections.

Master User Name*

Name of master user for your cluster. (e.g. awsuser)

Master User Password*

Password must contain 8 to 64 printable ASCII characters excluding: /, *, ', \, and @. It must contain 1 uppercase letter, 1 lowercase letter, and 1 number.

Confirm Password*

Confirm Master User Password.

Cancel

Continue

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Feedback

Creating a Redshift Data Warehouse via the AWS Console

Redshift · AWS Console

Services Edit

ianmas @ ianmas-aws Ireland Help

Amazon Redshift

Clusters

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

CLUSTER DETAILS

NODE CONFIGURATION

ADDITIONAL CONFIGURATION

REVIEW

Provide the details of your cluster. Fields marked with * are required.

Cluster Identifier*

mycluster

This is the unique key that identifies a cluster. This parameter is stored as a lowercase string. (e.g. my-dw-instance)

Database Name

Optional. A default database named dev is created for the cluster. Optionally, specify a custom database name (e.g. mydb) to create an additional database.

Database Port*

5439

Port number on which the database accepts connections.

Master User Name*

master

Name of master user for your cluster. (e.g. awsuser)

Master User Password*

.....

Password must contain 8 to 64 printable ASCII characters excluding: /, *, ', \, and @. It must contain 1 uppercase letter, 1 lowercase letter, and 1 number.

Confirm Password*

.....

Confirm Master User Password.

Cancel

Continue

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Feedback

Creating a Redshift Data Warehouse via the AWS Console

Redshift - AWS Console

Services Edit

ianmas @ ianmas-aws Ireland Help

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

CLUSTER DETAILS NODE CONFIGURATION ADDITIONAL CONFIGURATION REVIEW

Choose a number of nodes and Node Type below. Number of Compute Nodes is required for multi-node clusters.

Node Type Specifies the compute, memory, storage, and I/O capacity of the cluster's nodes.

CPU 7 EC2 Compute Units (2 virtual cores) per node

Memory 15 GiB per node

Storage 160GB SSD storage per node

I/O Performance Moderate

Cluster Type

Number of Compute Nodes* Single Node clusters consist of a single node which performs both leader and compute functions.

Maximum 1

Minimum 1

Cancel Previous Continue

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use Feedback

Creating a Redshift Data Warehouse via the AWS Console

Redshift - AWS Console

Services Edit

ianmas @ ianmas-aws Ireland Help

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

CLUSTER DETAILS **NODE CONFIGURATION** ADDITIONAL CONFIGURATION REVIEW

Choose a number of nodes and Node Type below. Number of Compute Nodes is required for multi-node clusters.

Node Type Specifies the compute, memory, storage, and I/O capacity of the cluster's nodes.

CPU 7 EC2 Compute Units (2 virtual cores) per node

Memory 15 GiB per node

Storage 160GB SSD storage per node

I/O Performance Moderate

Cluster Type

Number of Compute Nodes* Single Node clusters consist of a single node which performs both leader and compute functions.

Maximum 1

Minimum 1

Cancel Previous **Continue**

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Feedback](#)

Creating a Redshift Data Warehouse via the AWS Console

Redshift · AWS Console

Services Edit

ianmas @ ianmas-aws Ireland Help

Amazon Redshift

Clusters

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

CLUSTER DETAILS

NODE CONFIGURATION

ADDITIONAL CONFIGURATION

REVIEW

Provide the optional additional configuration details below.

Cluster Parameter Group

A default parameter group will be associated with this cluster.

Encrypt Database

No

Select Yes to encrypt all data within the cluster and in backups at a small cost to performance.

Use HSM

No

You have not created any HSM Connections. You must create an HSM Connection to use HSM. You must also create at least one HSM Client Certificate.

Configure Networking Options:

Choose a VPC

Default VPC (vpc-b7b8b3d5)

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

Select Yes if you want the cluster to be accessible from the public internet. Select No if you want it to be accessible only from within your private VPC network

Choose a Public IP Address

No

Select Yes if you want to select your own public IP address from a list of elastic IP (EIP) addresses that are already configured for your cluster's VPC. Select No if you want Amazon Redshift to provide an EIP for you instead.

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

launch-wizard-1 (sg-21f13...)

ElasticMapReduce-slave (s...

default (sg-8da1b6ef)

aws-101-sg (sg-0ba94d6e)

List of VPC Security Groups to associate with this cluster.

Optionally, create a basic alarm for this cluster.

Create CloudWatch Alarm

Yes

Create a CloudWatch alarm to monitor the disk usage of your cluster.

Disk Usage Threshold

80%

Threshold at which the alarm will trigger when disk usage across all nodes reaches this percentage.

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Privacy Policy

Terms of Use

Feedback

Creating a Redshift Data Warehouse via the AWS Console

Amazon Redshift

Clusters

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

Services

Edit

ianmas @ ianmas-aws

Ireland

Help

Configure Networking Options:

Choose a VPC

Default VPC (vpc-b7b8b3d5)

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

Select Yes if you want the cluster to be accessible from the public internet. Select No if you want it to be accessible only from within your private VPC network

Choose a Public IP Address

No

Select Yes if you want to select your own public IP address from a list of elastic IP (EIP) addresses that are already configured for your cluster's VPC. Select No if you want Amazon Redshift to provide an EIP for you instead.

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

aws-101-sg (sg-0ba94d6e)
ssh-tools (sg-bb0e12d9)
default_elb_5f1d280e-34fd...
ElasticMapReduce-master (...)

List of VPC Security Groups to associate with this cluster.

Optionally, create a basic alarm for this cluster.

Create CloudWatch Alarm

Yes

Create a CloudWatch alarm to monitor the disk usage of your cluster.

Disk Usage Threshold

80%

Threshold at which the alarm will trigger when disk usage across all nodes reaches this percentage.

Use Existing Topic

No

Use an existing SNS topic or create a new one. SNS is a Simple Notification Service which will send email notifications to the recipients of the SNS topic when the alarm triggers.

Topic

mycluster-default-alarms

Name of the SNS topic that will be created.

Recipients

Recipients of this SNS topic. If you have multiple recipients, separate the recipients with a comma.

Cancel

Previous

Continue

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Privacy Policy

Terms of Use

Feedback

Creating a Redshift Data Warehouse via the AWS Console

Redshift · AWS Console

Services Edit

ianmas @ ianmas-aws Ireland Help

Amazon Redshift

Clusters

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

CLUSTER DETAILS

NODE CONFIGURATION

ADDITIONAL CONFIGURATION

REVIEW

You are about to launch a cluster with the following specifications:

Cluster Properties

Database Configuration

These attributes specify the name of your cluster, what type of virtual hardware it will run on, how many nodes it will contain, and the availability zone in which it will be located.

These properties specify the database name, port, and username you will use to connect to the database. The parameter group contains configuration values used by the database.

Cluster Identifier: mycluster

Database Name: A default database will be created (dev)

Node Type: dw2.large

Database Port: 5439

Number of Compute Nodes: 1 (leader and compute run on a single node)

Master User Name: master

Availability Zone: No Preference

Cluster Parameter Group: A default parameter group will be created when the cluster is launched.

Security, Access, and Encryption

CloudWatch Alarms

These settings control whether your cluster will be created in an existing VPC to allow for simpler integration with other AWS Services, and the security groups which define access rules to your cluster.

CloudWatch alarms are used to notify if metrics for your cluster are within a certain threshold. All recipients under the SNS topic specified for your alarm will receive notifications once an alarm is triggered.

Virtual Private Cloud: vpc-b7b8b3d5

Basic alarms will not be created for this cluster.

Cluster Subnet Group:

Publicly Accessible: Yes

Elastic IP: Not used

VPC Security Groups: sg-21f13e44

Encrypt Database: No

Use HSM: No

Unless you are eligible for the free trial, you will start accruing charges as soon as your cluster is active.

Applicable charges:

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Feedback

Creating a Redshift Data Warehouse via the AWS Console

Amazon Redshift

Clusters

Snapshots

Security

Parameter Groups

Reserved Nodes

Events

Services

Edit

ianmas @ ianmas-aws

Ireland

Help

Node Type: dw2.large

Number of Compute Nodes: 1 (leader and compute run on a single node)

Availability Zone: No Preference

Database Port: 5439

Master User Name: master

A default parameter group will be created when the cluster is launched.

Security, Access, and Encryption

CloudWatch Alarms

These settings control whether your cluster will be created in an existing VPC to allow for simpler integration with other AWS Services, and the security groups which define access rules to your cluster.

CloudWatch alarms are used to notify if metrics for your cluster are within a certain threshold. All recipients under the SNS topic specified for your alarm will receive notifications once an alarm is triggered.

Virtual Private Cloud: vpc-b7b8b3d5

Cluster Subnet Group:

Publicly Accessible: Yes

Elastic IP: Not used

VPC Security Groups: sg-21f13e44

Encrypt Database: No

Use HSM: No

Basic alarms will not be created for this cluster.

⚠

Unless you are eligible for the free trial, you will start accruing charges as soon as your cluster is active.

Applicable charges:

The on-demand hourly rate for this cluster will be **\$0.30**, or **\$0.30/node**. If you have purchased reserved nodes in this region for this node type that are active, your costs will be discounted. Additional nodes will be billed at the on-demand rate.

If you are eligible for a free trial, you will receive 750 hours of free usage for each month of the trial, applied across all running dw2.large nodes across all regions. Regardless of when you start your trial, you will receive two full months of free usage. Once your trial expires or your usage exceeds 750 hours/month, you can shut down your cluster, avoiding any charges, or keep it running at our standard [On-Demand Rate](#).

For more information, see [Amazon Redshift Free Trial FAQ](#), [Amazon Redshift Pricing](#), and [Reserved Nodes Documentation](#).

Cancel

Previous

Launch Cluster

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved.

[Privacy Policy](#)

[Terms of Use](#)

Feedback

Creating a Redshift Data Warehouse via the AWS Console

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

Clusters

[Launch Cluster](#)

Cluster	Cluster Status	DB Health	In Maintenance	Recent Events
mycluster	creating	unknown	unknown	1

Additional Information

- [Overview of Features](#)
- [New Features](#)
- [Subscribe to Announcements](#)
- [Documentation](#)
- [Getting Started](#)
- [Management Guide](#)
- [DB Developer Guide](#)
- [Free Trial](#)
- [Free Trials from ETL and BI Partners](#)

Best Practices

- [Connecting to Clusters](#)
- [Loading Data](#)
- [Tutorial](#)
- [Tuning Table Design](#)
- [Tutorial](#)
- [Tuning Query Performance](#)

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Feedback](#)

Creating a Redshift Data Warehouse via the AWS Console

The screenshot displays the AWS Redshift console interface. At the top, the header bar includes the AWS logo, 'Services', 'Edit', and user information 'ianmas @ ianmas-aws' with a dropdown menu, 'Ireland', and a 'Help' link. The left sidebar lists navigation options: 'Amazon Redshift', 'Clusters' (highlighted with an orange bar), 'Snapshots', 'Security', 'Parameter Groups', 'Reserved Nodes', and 'Events'. The main content area is titled 'Clusters' and features a 'Launch Cluster' button. Below this is a table with columns: 'Cluster', 'Cluster Status', 'DB Health', 'In Maintenance', and 'Recent Events'. A single cluster named 'mycluster' is listed with a status of 'available', health of 'healthy', and no maintenance. Below the table, there are sections for 'Additional Information' and 'Best Practices', each containing several links to documentation and resources. The footer contains copyright information, links to 'Privacy Policy' and 'Terms of Use', and a 'Feedback' button.

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

Clusters

[Launch Cluster](#)

Cluster	Cluster Status	DB Health	In Maintenance	Recent Events
mycluster	available	healthy	no	1

Additional Information

- [Overview of Features](#)
- [New Features](#)
- [Subscribe to Announcements](#)
- [Documentation](#)
- [Getting Started](#)
- [Management Guide](#)
- [DB Developer Guide](#)
- [Free Trial](#)
- [Free Trials from ETL and BI Partners](#)

Best Practices

- [Connecting to Clusters](#)
- [Loading Data](#)
- [Tutorial](#)
- [Tuning Table Design](#)
- [Tutorial](#)
- [Tuning Query Performance](#)

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Feedback](#)

Creating a Redshift Data Warehouse via the AWS Console

The screenshot displays the AWS Redshift console interface. On the left, a navigation menu lists 'Amazon Redshift', 'Clusters', 'Snapshots', 'Security', 'Parameter Groups', 'Reserved Nodes', and 'Events'. The 'Clusters' section is active, showing a list of clusters with 'mycluster' selected. The main panel shows the 'Configuration' tab for 'mycluster'. It includes sub-tabs for 'Cluster', 'Database', and 'Backup'. The 'Cluster Properties' section lists details such as Cluster Name (mycluster), Cluster Type (Single Node), Node Type (dw2.large), Nodes (1), Zone (eu-west-1a), Created Time (September 19, 2014 12:22:19 PM UTC+1), Cluster Version (1.0.827), VPC ID (vpc-b7b8b3d5), Cluster Subnet Group (default), VPC Security Groups (launch-wizard-1), and Cluster Parameter Group (default.redshift-1.0). The 'Cluster Database Properties' section provides the Endpoint, Port (5439), Publicly Accessible status (Yes), Database Name (dev), Master Username (master), Encrypted status (No), JDBC URL, and ODBC URL. The 'Cluster Status' section shows Cluster Status (available), Database Health (healthy), In Maintenance Mode (no), Parameter Group Apply Status (in-sync), and Pending Modified Values (None). The 'Backup, Audit Logging, and Maintenance' section shows Automated Snapshot Retention Period (1), Cross-Region Snapshots Enabled (No), Audit Logging Enabled (No), Maintenance Window (fri:23:30-sat:00:00), and Allow Version Upgrade (Yes). The footer contains copyright information, privacy policy, terms of use, and a feedback button.

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

Cluster: mycluster

Cluster Type: Single Node

Node Type: dw2.large

Nodes: 1

Zone: eu-west-1a

Created Time: September 19, 2014 12:22:19 PM UTC+1

Cluster Version: 1.0.827

VPC ID: vpc-b7b8b3d5 (View VPCs)

Cluster Subnet Group: default

VPC Security Groups: launch-wizard-1 (sg-21f13e44) (active)

Cluster Parameter Group: default.redshift-1.0 (in-sync)

Cluster Properties

Cluster Name: mycluster

Cluster Type: Single Node

Node Type: dw2.large

Nodes: 1

Zone: eu-west-1a

Created Time: September 19, 2014 12:22:19 PM UTC+1

Cluster Version: 1.0.827

VPC ID: vpc-b7b8b3d5 (View VPCs)

Cluster Subnet Group: default

VPC Security Groups: launch-wizard-1 (sg-21f13e44) (active)

Cluster Parameter Group: default.redshift-1.0 (in-sync)

Cluster Database Properties

Endpoint: mycluster.c1ihbrvjsaew.eu-west-1.redshift.amazonaws.com

Port: 5439

Publicly Accessible: Yes

Database Name: dev

Master Username: master

Encrypted: No

JDBC URL: jdbc:postgresql://mycluster.c1ihbrvjsaew.eu-west-1.redshift.amazonaws.com:5439/dev?tcpKeepAlive=true

ODBC URL: Driver={PostgreSQL}; Server=mycluster.c1ihbrvjsaew.eu-west-1.redshift.amazonaws.com;

Cluster Status

Cluster Status: available

Database Health: healthy

In Maintenance Mode: no

Parameter Group Apply Status: in-sync

Pending Modified Values: None

Backup, Audit Logging, and Maintenance

Automated Snapshot Retention Period: 1

Cross-Region Snapshots Enabled: No

Audit Logging Enabled: No

Maintenance Window: fri:23:30-sat:00:00

Allow Version Upgrade: Yes

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Feedback

Creating a Redshift Data Warehouse via the AWS Console

The screenshot displays the AWS Redshift console interface. On the left, a navigation menu lists 'Clusters', 'Snapshots', 'Security', 'Parameter Groups', 'Reserved Nodes', and 'Events'. The main content area is titled 'Amazon Redshift' and shows the configuration for a cluster named 'mycluster'. The 'Configuration' tab is selected, displaying the 'Cluster Database Properties' section, which is highlighted with an orange border. This section includes the following details:

- Endpoint: `mycluster.clihbrvjsaew.eu-west-1.redshift.amazonaws.com`
- Port: `5439`
- Publicly Accessible: `Yes`
- Database Name: `dev`
- Master Username: `master`
- Encrypted: `No`
- JDBC URL: `jdbc:postgresql://mycluster.clihbrvjsaew.eu-west-1.redshift.amazonaws.com:5439/dev?tcpKeepAlive=true`
- ODBC URL: `Driver={PostgreSQL}; Server=mycluster.clihbrvjsaew.eu-west-1.redshift.amazonaws.com; Database=dev; UID=master; PWD=insert_your_master_user_password_here; Port=5439`

Other sections visible include:

- Backup, Audit Logging, and Maintenance:** Automated Snapshot Retention Period: `1`; Cross-Region Snapshots Enabled: `No`; Audit Logging Enabled: `No`; Maintenance Window: `fri:23:30-sat:00:00`; Allow Version Upgrade: `Yes`.
- Capacity Details:** Current Node Type: `dw2.large`; CPU: `7 EC2 Compute Units (2 virtual cores) per node`; Memory: `15 GiB per node`; Storage: `160GB SSD storage per node`; I/O Performance: `Moderate`; Platform: `64-bit`.
- SSH Ingestion Settings:** Cluster Public Key: `ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAQCyF33ZEvYVR4+dhzcnp0uS/cosZV1FBizlUA+MoGeQhZqAF0xqMyQidnx3A8Rv+lg36OGwpEKuCB9DwtHapij/udCN4msWWGfUe9/wtyZLX5aunDtgkPHHdoh/daQZ9+T8iIT0d/r7SLBdaeQOed4IRgzcVTLUEx/CtjTdWuoRHg9QKcrlFNScZqzGAatXEuq5OPfgVsP+Y2almW35Z1uP5fvaE8zzR+dCHSdzdW++smXpda3XVv39AQnKyrj39qrGyTL6wy7uolwV7RKzaMntTcl0nY5dyBgn ebN5ibVpgNKu8Fyt+ggy1kykfbwucpxTKumxnWWf93wu0JF Amazon-Redshift`
- Node IP Addresses:** A table showing the public and private IP addresses for the cluster nodes.

Node	Public IP	Private IP
Shared	54.194.54.168	172.31.35.36

At the bottom of the console, there is a footer with copyright information and links to the Privacy Policy and Terms of Use. A 'Feedback' button is also present in the bottom right corner.

More Getting Started Resources

2 Month Free Trial

Comprehensive Getting Started Guide

Table Design & Data Loading Tutorials

Table Design

Table Design

- ▶ How to choose optimal table design ?
 - ▶ Choosing the right data type
 - ▶ Distributing and sorting data

Data Types

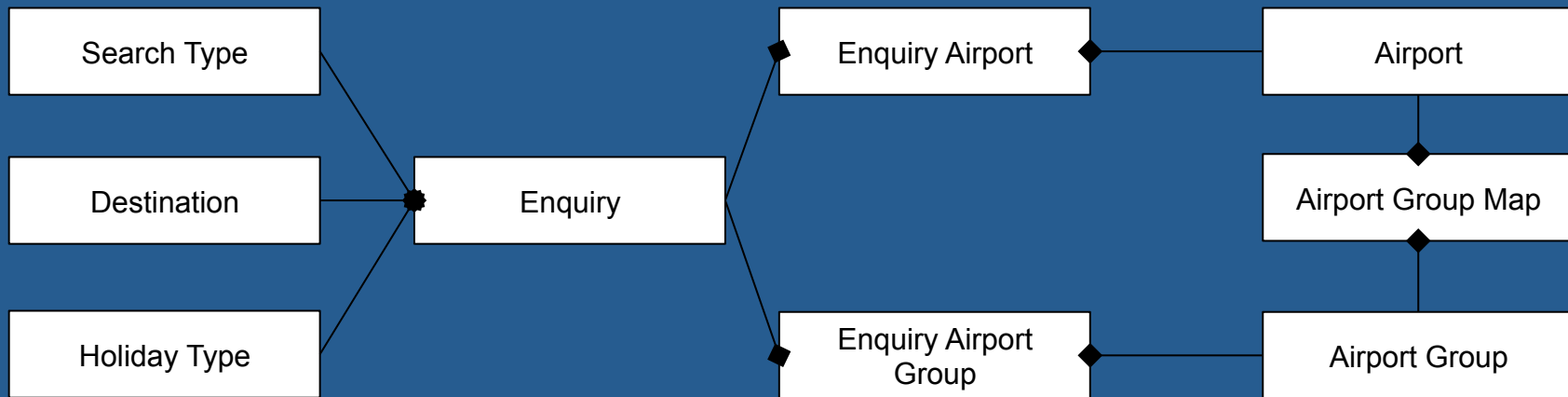
Data type	Aliases	Description
SMALLINT	INT2	Signed two-byte integer
INTEGER	INT, INT4	Signed four-byte integer
BIGINT	INT8	Signed eight-byte integer
DECIMAL	NUMERIC	Exact numeric of selectable precision
REAL	FLOAT4	Single precision floating-point number
DOUBLE PRECISION	FLOAT8	Double precision floating-point number
BOOLEAN	BOOL	Logical Boolean (true/false)
CHAR	CHARACTER	Fixed-length character string
VARCHAR	CHARACTER VARYING	Variable-length character string with a user-defined limit
DATE		Calendar date (year, month, day)
TIMESTAMP		Date and time (without time zone)

The VARCHAR data type supports multi-byte characters up to a maximum of three bytes.
Four-byte or longer characters are not supported.

For example, VARCHAR(12) can support four Chinese characters (three bytes / character):
 $4 \text{ characters} \times 3 \text{ bytes per character} = 12 \text{ bytes}$

Your Schema

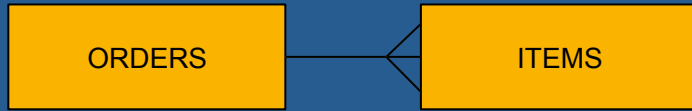
► 3NF



Distributing Data

- ▶ Redshift is a distributed system:
 - ▶ A cluster contains a leader node & compute nodes
 - ▶ A compute node contains slices (one per core)
 - ▶ A slice contains data
- ▶ Slices are chosen based on two types of distribution:
 - ▶ Round Robin (automated)
 - ▶ Based on a distribution key (hash of a defined column)
- ▶ Queries run on all slices in parallel: optimal query throughput can be achieved when data is evenly spread across slices

Unoptimised Distribution



Order 1: Dave Smith, Total £195

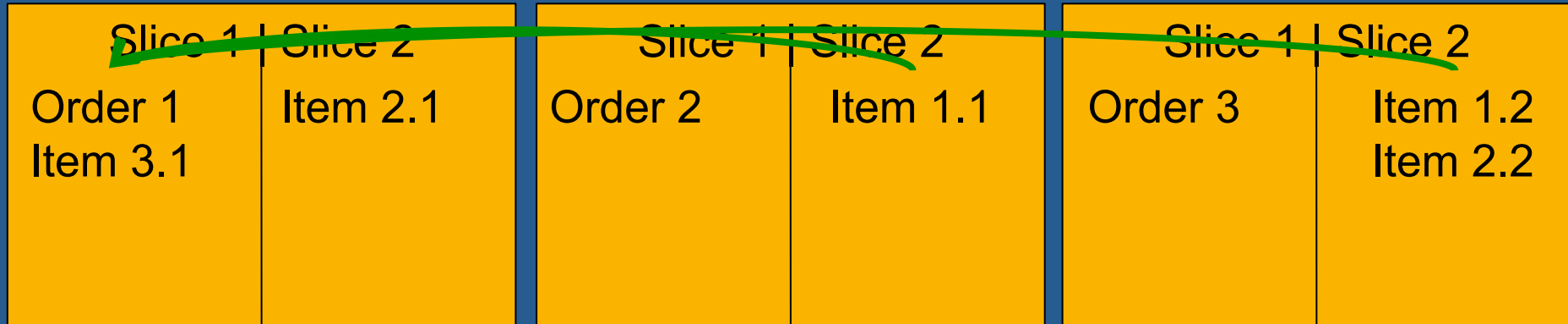
Item 1.1: Order 1, Kindle Fire HD 7", £159

Item 1.2: Order 1, Kindle Fire Case, £36

Node 1

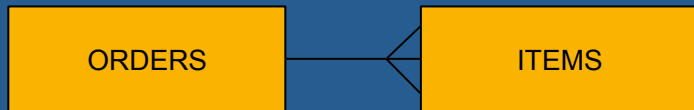
Node 2

Node 3



Default (No Distribution Key, Round Robin Order)

Optimised Distribution

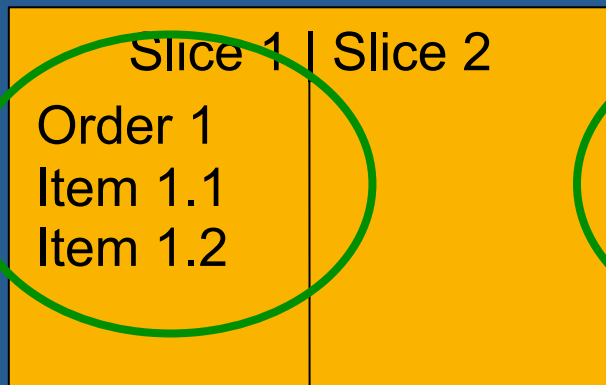


Order 1: Dave Smith, Total £195

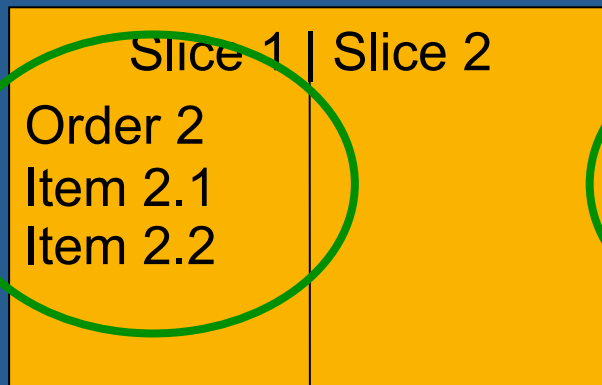
Item 1.1: Order 1, Kindle Fire HD 7", £159

Item 1.2: Order 1, Kindle Fire Case, £36

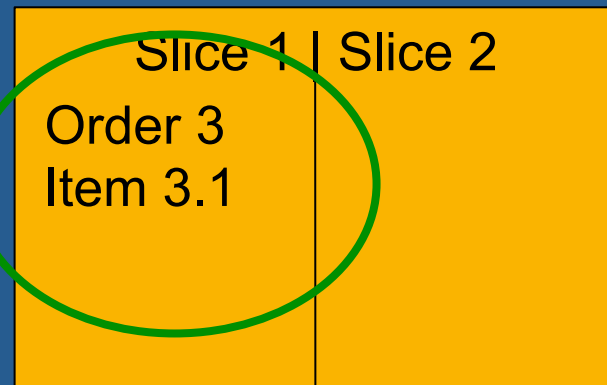
Node 1



Node 2



Node 3



Customised (ORDERS.ORDER_ID DISTKEY, ITEMS.ORDER_ID DISTKEY)

Choosing a Distribution Key

- ▶ Frequently Joined
 - ▶ By most commonly run queries
 - ▶ By queries which consume the most CPU
- ▶ High Cardinality
 - ▶ Large number of discrete values
- ▶ Low Skew
 - ▶ Uniform Distribution
 - ▶ No Hotspots
 - ▶ Query STV_BLOCKLIST for Skew Factor

Sorting Data

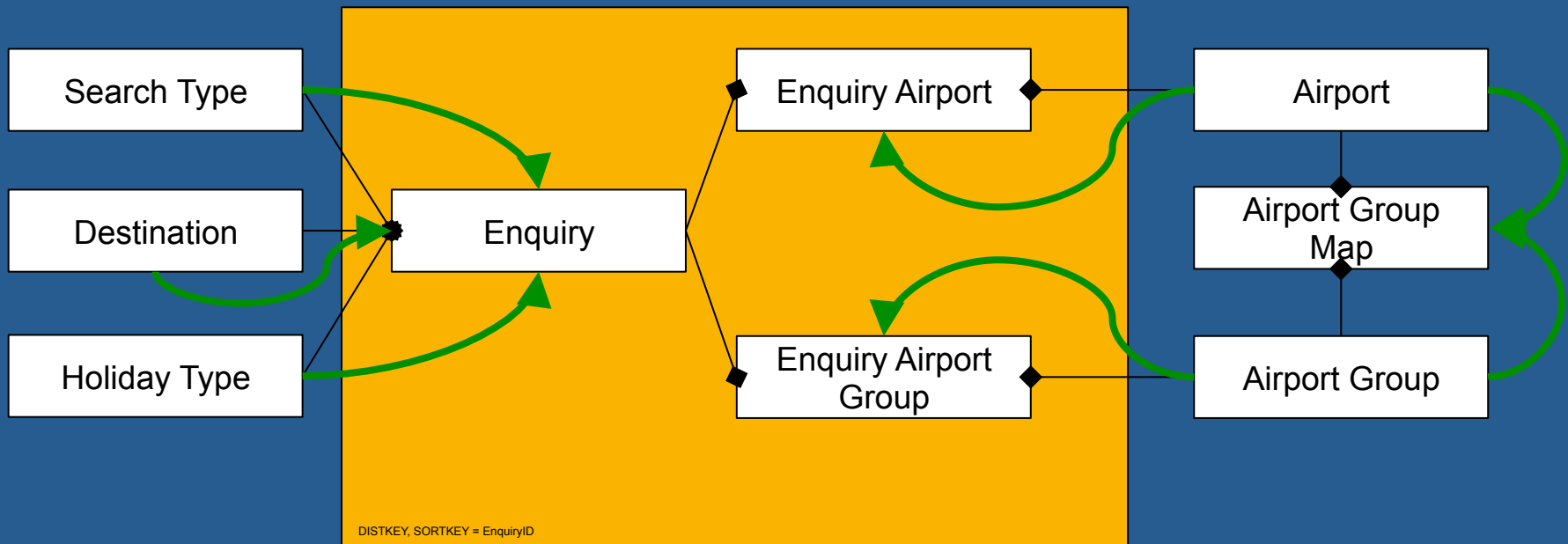
- ▶ In the slices (on disk), the data is sorted by a sort key (if none uses the insertion order)
- ▶ Choose a sort key that is frequently used in your queries
 - ▶ As a query predicate (date, identifier, ...)
 - ▶ As a join parameter (it can also be the hash key)
- ▶ The sort key allows Redshift to avoid reading entire blocks based on predicates
- ▶ E.g.: a table containing a timestamp sort key, and where only recent data is accessed, will skip blocks containing “old” data

Schema Design

- ▶ Optimizing a database for querying
 - ▶ Analyse using Automatic Compression to ensure optimal IO
 - ▶ Co-locate frequently joined tables: use distribution key wisely (avoids data transfers between nodes)
 - ▶ For joined tables, use sort keys on the joined columns, allowing fast merge joins
 - ▶ Compression allows you to de-normalize without penalizing storage, simplifying queries and limiting joins

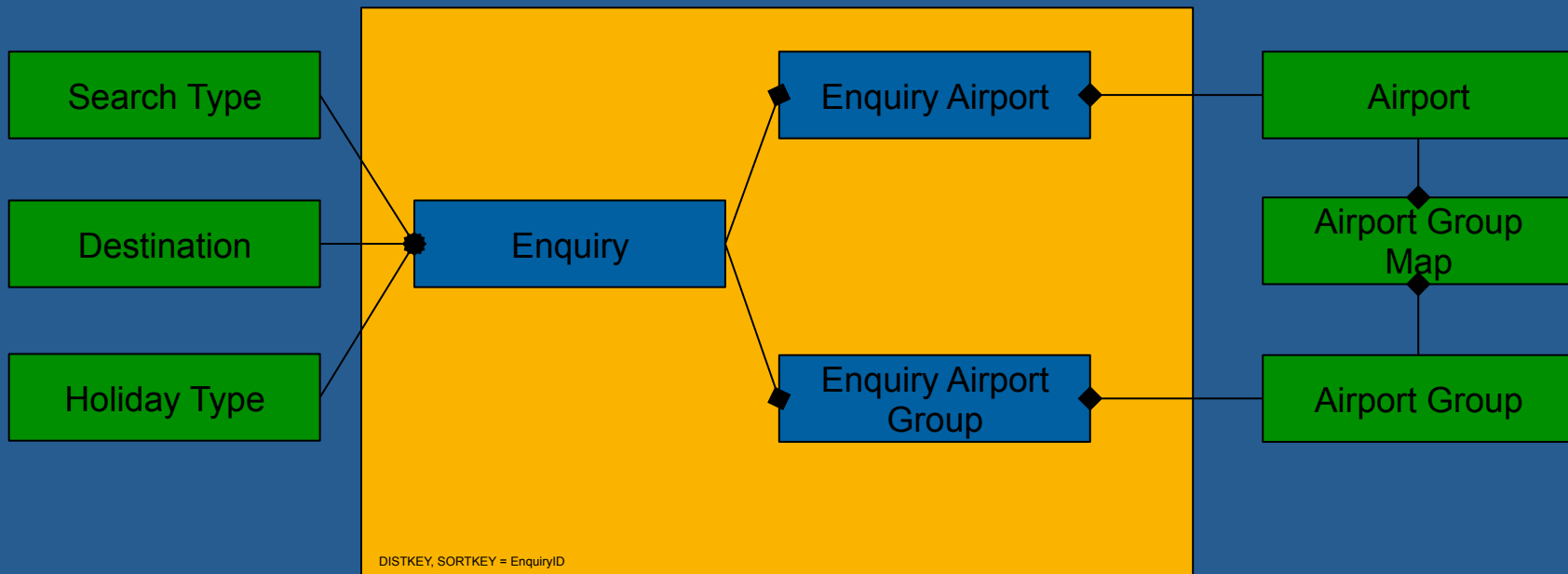
Possible Schema Optimisations

- ▶ Denormalise commonly used join attributes onto large tables



Possible Schema Optimisations

► Cache Small Tables with DISTYLE ALL



More on Table Design

Table Design Documentation

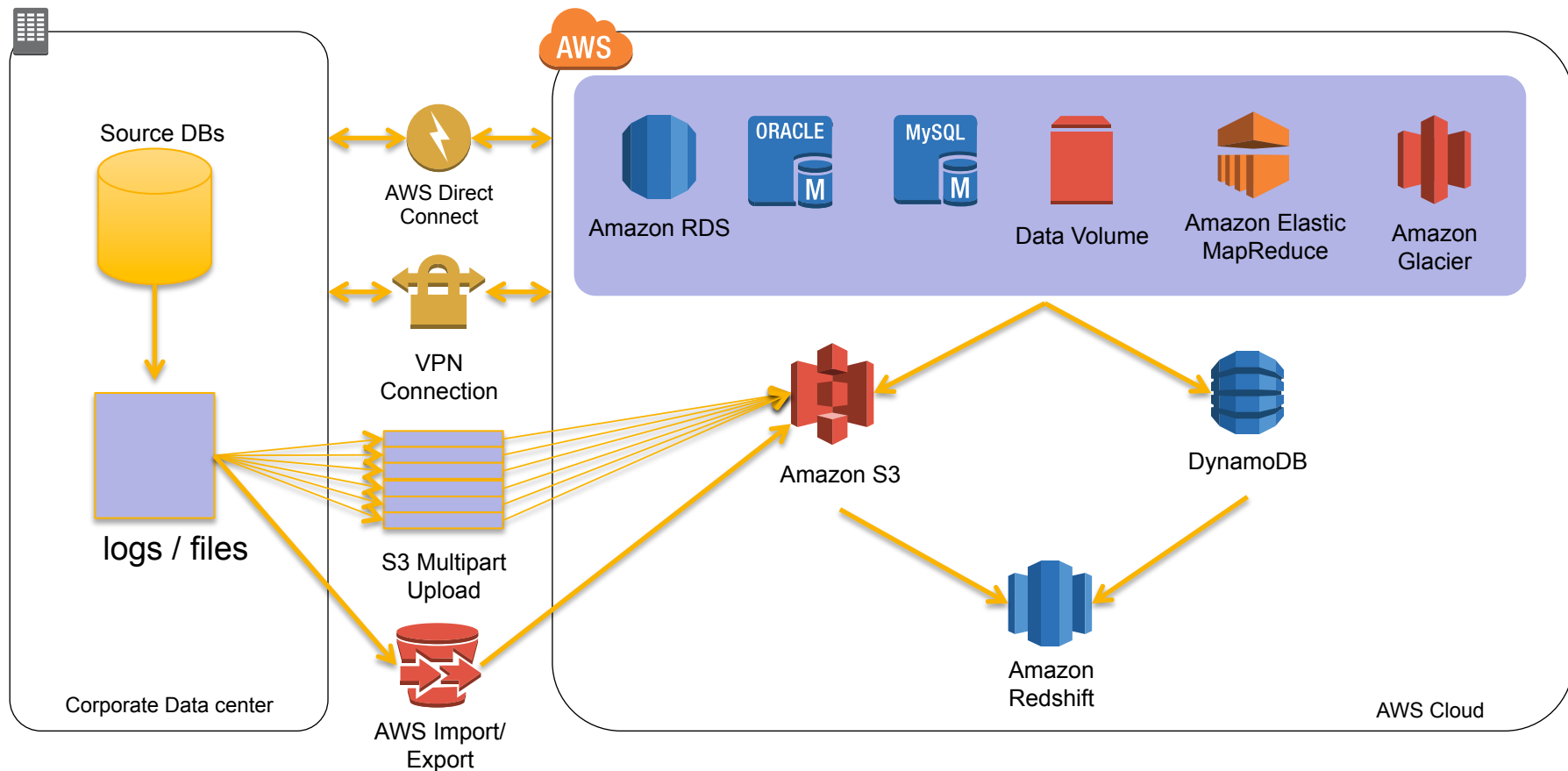
http://docs.aws.amazon.com/redshift/latest/dg/t_Creating_tables.html

Table Design Tutorial

<http://docs.aws.amazon.com/redshift/latest/dg/tutorial-tuning-tables.html>

Data Loading

Amazon Redshift Loading Data Overview

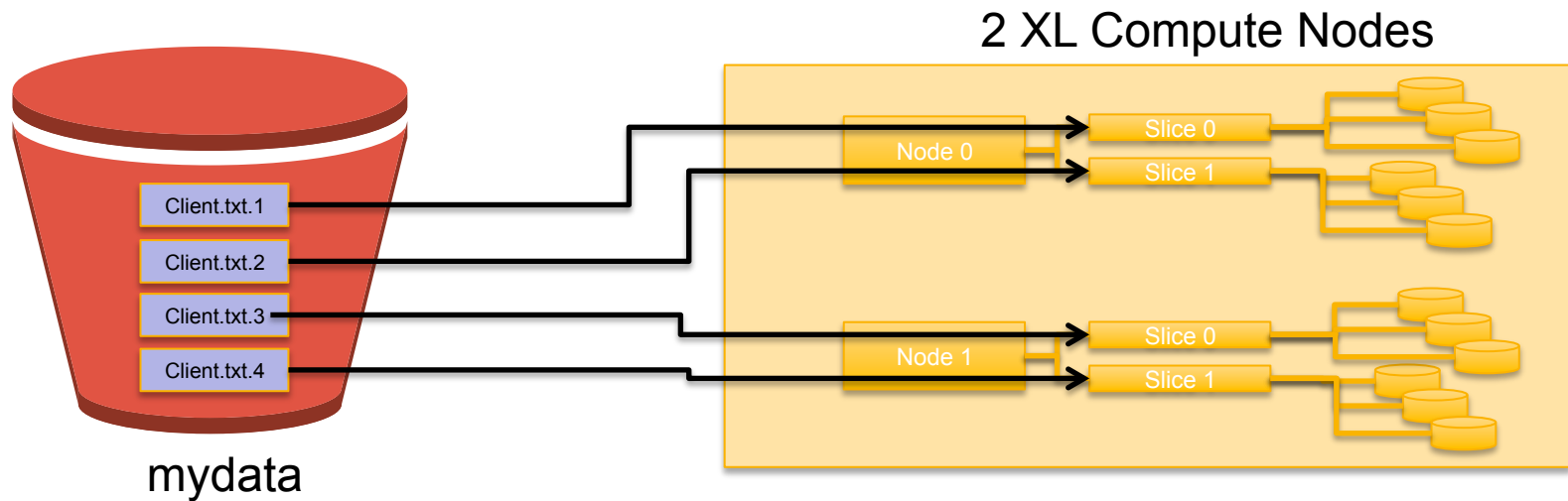


Sharing Your Data

- ▶ Data Dumped from DB
- ▶ CSV Files Encrypted
 - ▶ Review Encryption Code
- ▶ Uploaded to S3
 - ▶ Bucket <bucket>
- ▶ Data shared with AWS

```
{
  "Id": "Policy1234567890",
  "Statement": [
    {
      "Action": [
        "s3:Get*",
        "s3:List*"
      ],
      "Effect": "Allow",
      "Principal": {
        "AWS": ["arn:aws:iam::887210671223:root"]
      },
      "Resource": [
        "arn:aws:s3:::<bucket-name>",
        "arn:aws:s3:::<bucket-name>/*"
      ],
      "Sid": "Stmt0987654321"
    }
  ]
}
```

Splitting Data Files

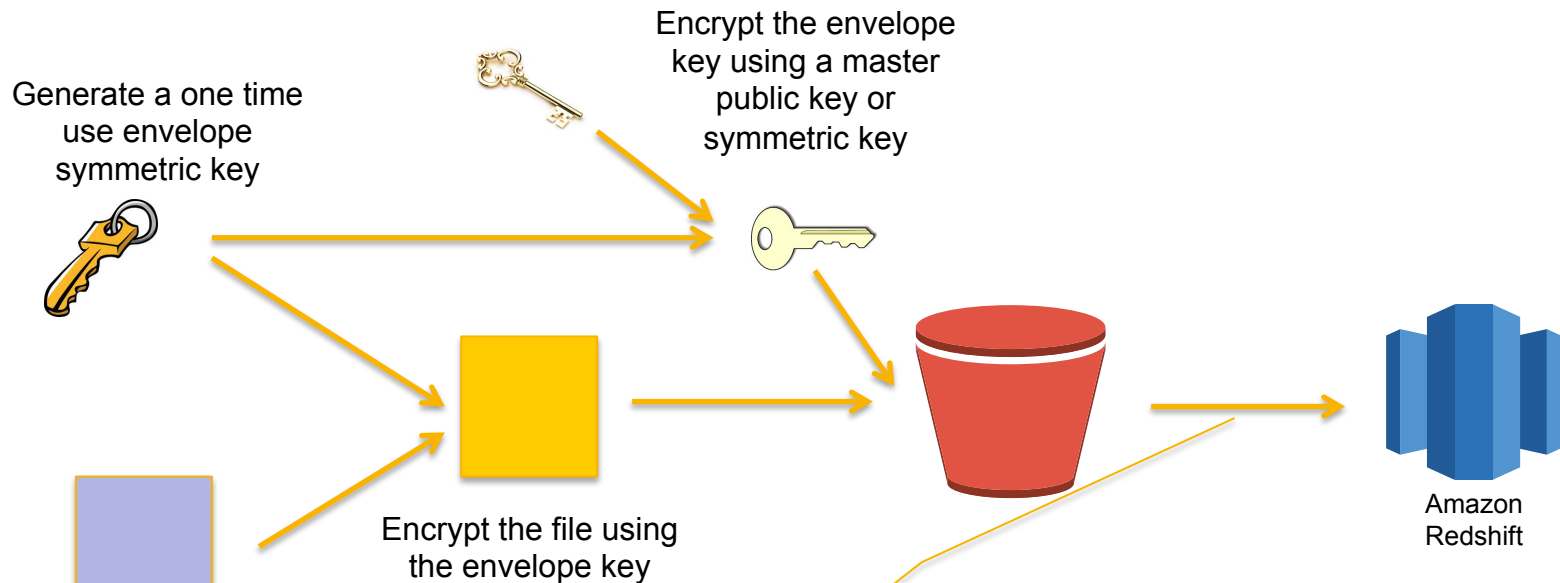


Copy customer from 's3://mydata/client.txt'

Credentials 'aws_access_key_id=<your-access-key>; aws_secret_access_key=<your_secret_key>'

Delimiter '|';

Loading Encrypted Data Files



*Copy customer from 's3://mydata/client.txt'
Credentials 'aws_access_key_id=<your-access-key>;
aws_secret_access_key=<your_secret_key>;
Master_symmetric_key=<master_key>' Delimiter '|' encrypted;*

Typical Load Issues

- ▶ Mismatch between data types in table and values in input data fields
- ▶ Mismatch between number of columns in table and number of fields in input data
- ▶ Mismatched quotes
 - ▶ Redshift supports both single and double quotes; however, these quotes must be balanced appropriately
- ▶ Incorrect format for date/time data in input files
 - ▶ Use DATEFORMAT and TIMEFORMAT to control
- ▶ Out-of-range values in input files (for numeric columns)
- ▶ Number of distinct values for a column exceeds the limitation for its compression encoding

Direct SQL

- ▶ Redshift supports standard DML commands
 - ▶ INSERT, UPDATE, DELETE
- ▶ Redshift does not support single-command merge (upsert) statement
 - ▶ Load data into a staging table
 - ▶ Joining the staging table with the target table
 - ▶ UPDATE data where row exists
 - ▶ INSERT where no row exists
- ▶ All Direct SQL Commands go via Leader Node

Data Loading Best Practices

- ▶ Use a COPY Command to load data
- ▶ Use a single COPY command
- ▶ Split your data into multiple files
- ▶ Compress your data files with GZIP
- ▶ Use multi-row inserts if COPY is not possible
- ▶ Bulk insert operations (INSERT INTO...SELECT and CREATE TABLE AS) provide high performance data insertion

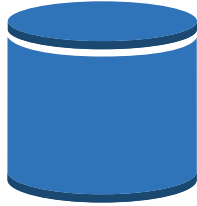
Data Loading Best Practices

- ▶ Load your data in sort key order to avoid needing to vacuum
- ▶ Organize your data as a sequence of time-series tables, where each table is identical but contains data for different time ranges
- ▶ Use staging tables to perform an upsert
- ▶ Run the VACUUM command whenever you add, delete, or modify a large number of rows, unless you load your data in sort key order
- ▶ Increase the memory available to a COPY or VACUUM by increasing `wlm_query_slot_count`
- ▶ Run the ANALYZE command whenever you've made a non-trivial number of changes to your data to ensure your table statistics are current

Working with data in Amazon Redshift



OLTP
ERP



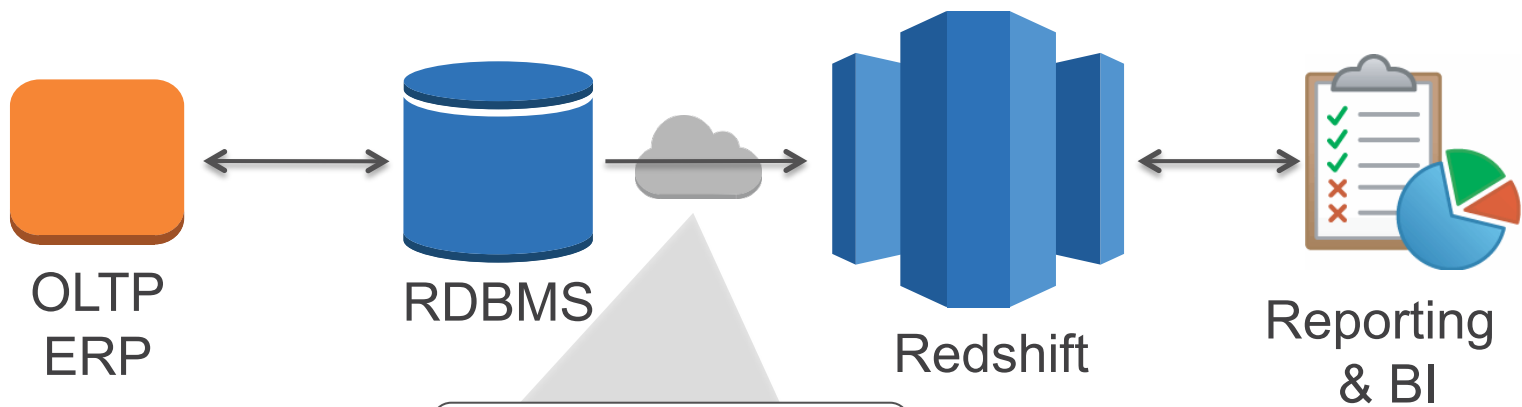
RDBMS



Redshift



Reporting
& BI



 **ATTUNITY**
CloudBeam

INFORMATICA®

talend*
integration at any scale



Amazon Redshift

JDBC/ODBC



[Shop All Categories](#) ▾

Search AWS Marketplace

GO

▶ You

Categories

[All Categories](#)[Business Software](#)[Business Intelligence](#)[Amazon Redshift](#)

Filters

Operating System

☒ All Linux/Unix

Software Pricing Plans

- ☐ Hourly (2)
- ☐ Annual (2)
- ☐ Bring Your Own License (1)

Delivery Method

- ☐ Amazon Machine Image (3)
- ☐ SaaS (3)

Average Rating

★★★★★ & up (2)

★★★★★ & up (2)

★★★★★ & up (2)

Amazon Redshift

Amazon Redshift is a fast and powerful, fully managed, petabyte-scale data warehouse service in the cloud. Amazon Redshift offers you fast query performance when analyzing virtually any size data set using the same SQL-based tools and business intelligence applications you use today. With a few clicks in the AWS Management Console, you can launch an Amazon Redshift cluster, starting with a few hundred gigabytes of data and scaling to a petabyte or more. [Read more](#)

Featured Products



Jaspersoft Reporting and Analytics fo...

Jaspersoft
Starting from **\$0.48/hr** or
from **\$3,750/yr** for
software



Free MicroStrategy Suite

MicroStrategy
\$0.00/hr for software +
Charges for EC2 with
Windows



Attunity CloudBeam

Attunity
From \$49.95 per month

Resources

[Amazon Redshift](#)

Fast and Powerful, Petabyte-Scale Warehouse Service

[AWS Partners](#)

AWS Partners can help you with cloud warehousing, BI and more

[Amazon Web Services Home](#)[Sign in](#) or [Create a new account](#)[Your Account](#) | [Help](#) | [Sell on AWS Marketplace](#)[Shop All Categories](#) ▾

Search AWS Marketplace

GO

[▶ Your Software](#)

Jaspersoft Reporting and Analytics for AWS (Hourly)

Sold by: [Jaspersoft](#) | [See product video](#) 

Jaspersoft for AWS is a commercial open source reporting and analytics server built for AWS that can run standalone or be embedded in your application. It is priced very aggressively with a low hourly rate that has no data or user limits and no additional fees. A multi-tenant version is available as a separate Marketplace listing. Our business intelligence suite allows you to easily create beautiful, interactive reports, dashboards and data visualizations. Designed to quickly connect to your Amazon RDS, Redshift and EMR data sources, you can be analyzing your data and building reports in under ... [Read more](#)

Customer Rating ★★★★★ (21 Customer Reviews)**Latest Version** 5.6 (Other available versions)**Base Operating System** Linux/Unix, Amazon Linux 2014.03.01**Delivery Method** 64-bit Amazon Machine Image (AMI) ([Learn more](#))**Support** [See details below](#)**AWS Services Required** Amazon EC2, Amazon EBS**Highlights**

- Full BI Server for Cents/Hour: no user or data limits and no additional fees. Suite includes ad hoc query and reporting.

[Continue](#)

You will have an opportunity to review your order before launching or being charged.

Pricing Details

For region

US East (Virginia)

Hourly Fees

Total hourly fees will vary by instance type and EC2 region.

Software Pricing: ☒ Hourly ☐ Annual

Customer Rating ★★★★★  (21 Customer Reviews)

Latest Version 5.6 (Other available versions)

Base Operating System Linux/Unix, Amazon Linux 2014.03.01

Delivery Method 64-bit Amazon Machine Image (AMI) ([Learn more](#))

Support [See details below](#)

AWS Services Required Amazon EC2, Amazon EBS

- Highlights**
- Full BI Server for Cents/Hour: no user or data limits and no additional fees. Suite includes ad hoc query and reporting, dashboards, data analysis, data visualization and data virtualization.
 - 10 Minutes to Your AWS Data: purpose-built for AWS, our reporting and analytics server allows you to quickly and easily connect to Amazon RDS, Redshift and EMR. In under 10 minutes you can be reporting on and analyzing your data.
 - BI for Your Business or App: built to modern web standards with a HTML5 UI and JavaScript and REST APIs, our flexible BI suite can be used to analyze your business or deliver stunning interactive reports and dashboards inside your app.

Recommended Products

Continue

You will have an opportunity to review your order before launching or being charged.

Pricing Details

For region

US East (Virginia)

Hourly Fees

Total hourly fees will vary by instance type and EC2 region.

Software Pricing: **Hourly** ☐ **Annual** ☒

Software annual pricing savings over hourly: 10% - 11% 

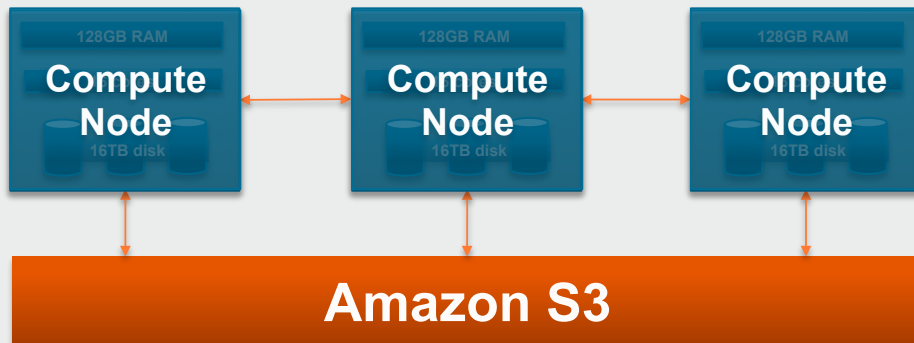
EC2 Instance Type	EC2 Usage	Software	Total
m1.medium	\$0.087/hr	\$0.48/hr	\$0.567/hr
m1.large	\$0.175/hr	\$1.08/hr	\$1.255/hr
m1.xlarge	\$0.35/hr	\$2.16/hr	\$2.51/hr
m2.xlarge	\$0.245/hr	\$1.78/hr	\$2.025/hr
m2.2xlarge	\$0.49/hr	\$3.69/hr	\$4.18/hr
m2.4xlarge	\$0.98/hr	\$7.38/hr	\$8.36/hr
c1.xlarge	\$0.52/hr	\$2.67/hr	\$3.19/hr
m3.medium	\$0.07/hr	\$0.48/hr	\$0.55/hr
m3.large	\$0.14/hr	\$0.99/hr	\$1.13/hr
m3.xlarge	\$0.28/hr	\$2.03/hr	\$2.31/hr
m3.2xlarge	\$0.56/hr	\$4.06/hr	\$4.62/hr

EBS Magnetic volumes

\$0.05 per GB-month of provisioned storage

Backup and Restoration

BACKUP AND RESTORATION



Backups to Amazon S3 are automatic, continuous and incremental

Configurable system snapshot retention period

User snapshots on-demand

Streaming restores enable you to resume querying faster

CLUSTER DATA REPLICATION

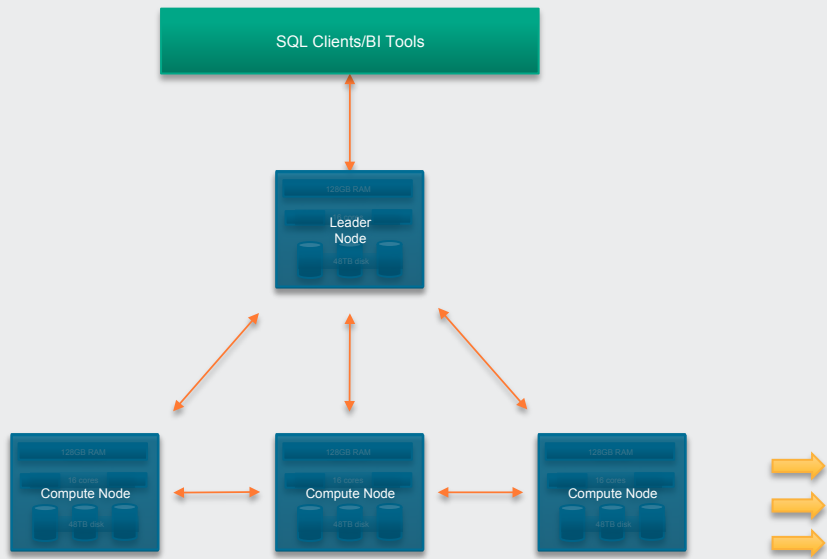
+

**AUTOMATED BACKUPS ON
AMAZON S3**

+

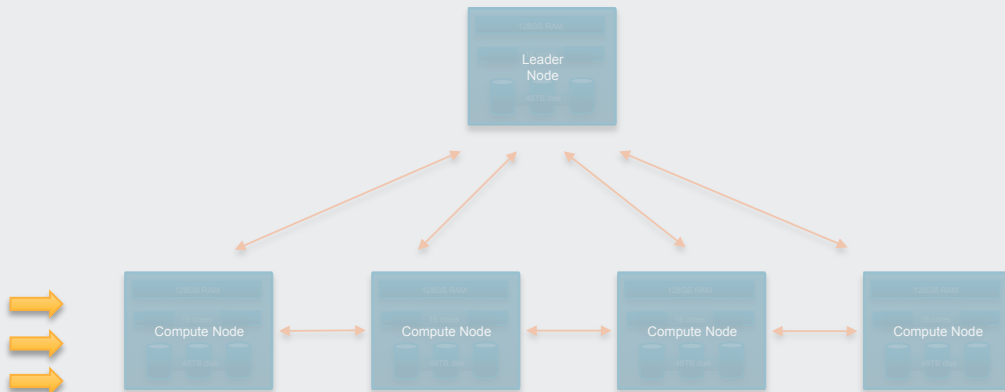
NODE MONITORING

Upgrading and Scaling



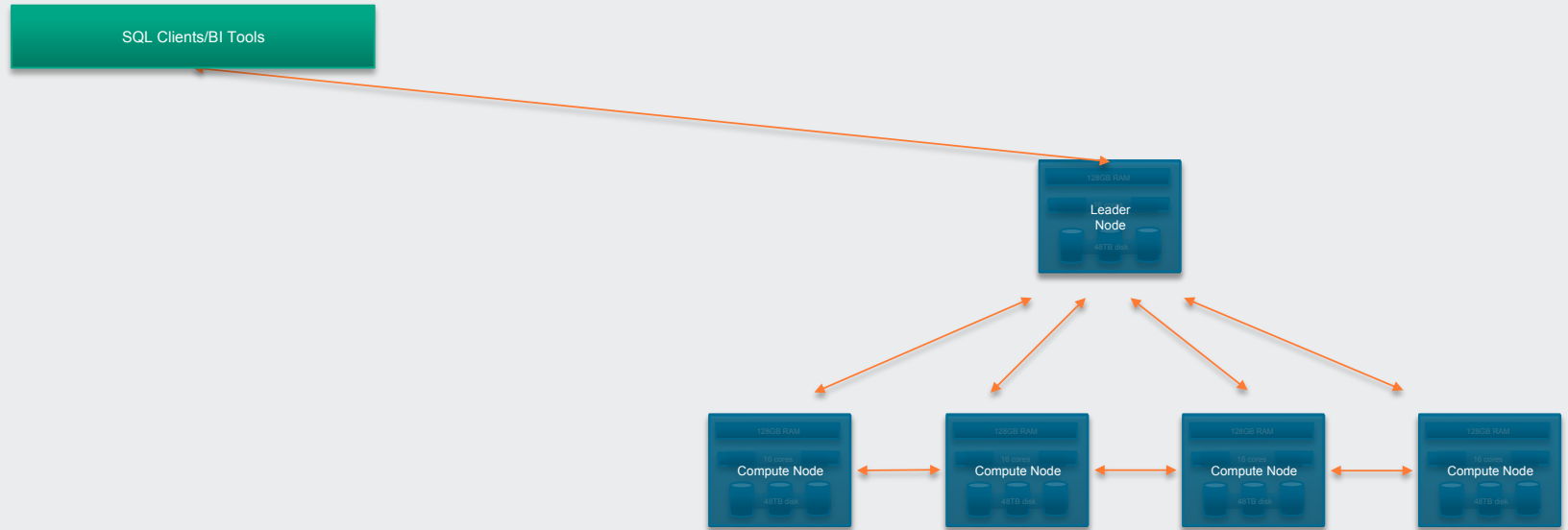
Resize while remaining online

Provision a new cluster in the background



Copy data in parallel from node to node

Only charged for source cluster



Automatic SQL endpoint switchover via DNS

Decommission the source cluster

Resizing a Redshift Data Warehouse via the AWS Console

The screenshot displays the Amazon Redshift console interface. The top navigation bar includes the AWS logo, 'Services', 'Edit', and user information 'ianmas @ ianmas-aws' with a dropdown menu showing 'Ireland' and 'Help'. The left sidebar lists navigation options: Amazon Redshift, Clusters (highlighted with an orange bar), Snapshots, Security, Parameter Groups, Reserved Nodes, and Events. The main content area is titled 'Clusters' and features a 'Launch Cluster' button and a refresh icon. Below this is a table with columns: Cluster, Cluster Status, DB Health, In Maintenance, and Recent Events. A single cluster named 'mycluster' is listed with a status of 'available', health of 'healthy', and no maintenance. Underneath the table, there are sections for 'Additional Information' (including Overview of Features, New Features, Documentation, and Free Trial) and 'Best Practices' (including Connecting to Clusters, Loading Data, Tuning Table Design, and Tuning Query Performance). The footer contains copyright information, links to Privacy Policy and Terms of Use, and a Feedback button.

Amazon Redshift

- Clusters
- Snapshots
- Security
- Parameter Groups
- Reserved Nodes
- Events

Clusters

[Launch Cluster](#)

Cluster	Cluster Status	DB Health	In Maintenance	Recent Events
mycluster	available	healthy	no	1

Additional Information

- [Overview of Features](#)
- [New Features](#)
- [Subscribe to Announcements](#)
- [Documentation](#)
 - [Getting Started](#)
 - [Management Guide](#)
 - [DB Developer Guide](#)
- [Free Trial](#)
 - [Free Trials from ETL and BI Partners](#)

Best Practices

- [Connecting to Clusters](#)
- [Loading Data](#)
 - [Tutorial](#)
- [Tuning Table Design](#)
 - [Tutorial](#)
- [Tuning Query Performance](#)

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Feedback](#)

Resizing a Redshift Data Warehouse via the AWS Console

The screenshot displays the AWS Redshift console interface. On the left, the navigation menu includes 'Clusters', 'Snapshots', 'Security', 'Parameter Groups', 'Reserved Nodes', and 'Events'. The main content area is titled 'Cluster: mycluster' and shows the 'Configuration' tab. A dropdown menu is open over the 'Cluster' button, showing options: 'Modify', 'Resize' (highlighted in orange), 'Shut Down', and 'Reboot'.

Cluster: mycluster

Cluster Properties

- Cluster Name: mycluster
- Cluster Type: Multi Node
- Node Type: dw2.large
- Nodes: 4
- Zone: eu-west-1b
- Created Time: September 19, 2014 3:08:10 PM UTC+1
- Cluster Version: 1.0.827
- VPC ID: vpc-b7b8b3d5 (View VPCs)
- Cluster Subnet Group: default
- VPC Security Groups: launch-wizard-1 (sg-21f13e44) (active)
- Cluster Parameter Group: default.redshift-1.0 (in-sync)

Cluster Status

- Cluster Status: available
- Database Health: healthy
- In Maintenance Mode: no
- Parameter Group Apply Status: in-sync
- Pending Modified Values: None

Cluster Database Properties

- Endpoint: mycluster.c1hbrvjsaew.eu-west-1.redshift.amazonaws.com
- Port: 5439
- Publicly Accessible: Yes
- Database Name: dev
- Master Username: myuser
- Encrypted: No
- JDBC URL: jdbc:postgresql://mycluster.c1hbrvjsaew.eu-west-1.redshift.amazonaws.com:5439/dev?tcpKeepAlive=true
- ODBC URL: Driver={PostgreSQL}; Server=mycluster.c1hbrvjsaew.eu-west-1.redshift.amazonaws.com; Database=dev; UID=myuser; PWD=insert_your_master_user_password_here; Port=5439

Backup, Audit Logging, and Maintenance

- Automated Snapshot Retention Period: 1
- Cross-Region Snapshots Enabled: No
- Audit Logging Enabled: No
- Maintenance Window: tue:00:00-tue:00:30
- Allow Version Upgrade: Yes

Capacity Details

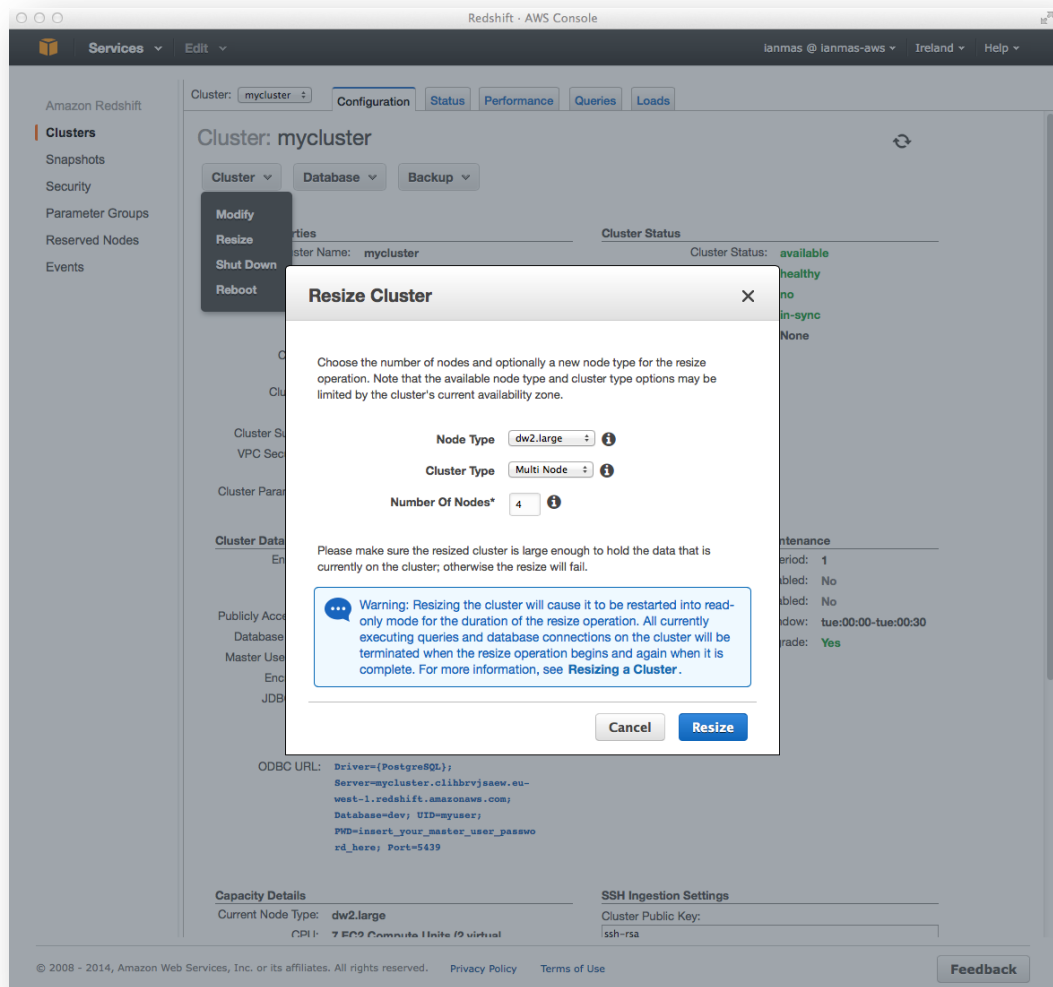
- Current Node Type: dw2.large
- CPU: 7 EC2 Compute Units (9 virtual)

SSH Ingestion Settings

- Cluster Public Key: ssh-rsa

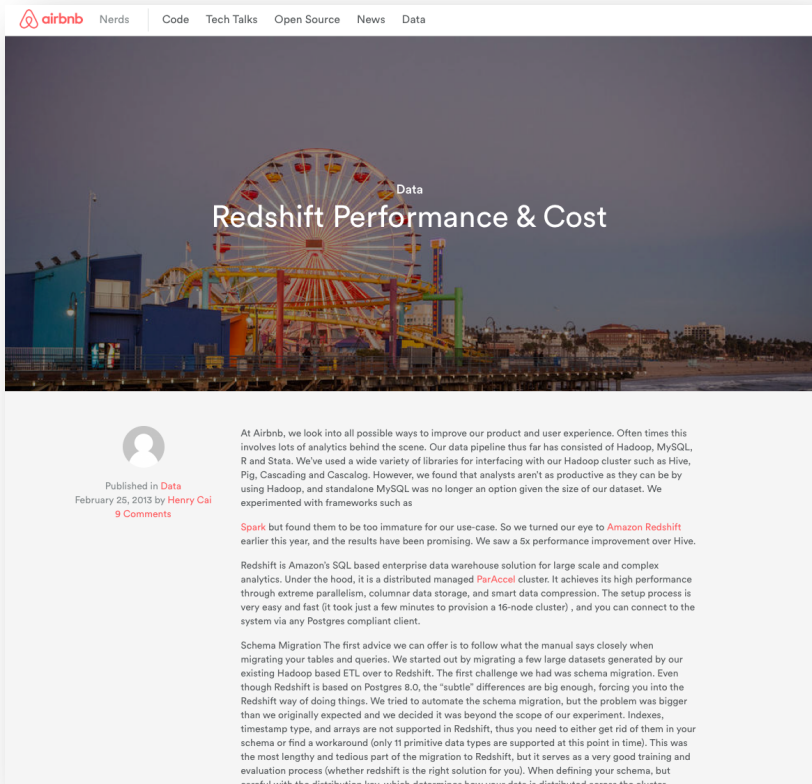
© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#) [Feedback](#)

Resizing a Redshift Data Warehouse via the AWS Console



Some other things...

Airbnb



The screenshot shows the Airbnb Data blog page. The header includes the Airbnb logo and navigation links: Nerds, Code, Tech Talks, Open Source, News, and Data. The main image is a Ferris wheel at night with the word 'Data' overlaid. The title 'Redshift Performance & Cost' is prominently displayed. Below the image, there is a byline for Henry Cai, the publication date (February 25, 2013), and 9 comments. The article text discusses the challenges of scaling analytics at Airbnb and the benefits of using Amazon Redshift.

airbnb Nerds Code Tech Talks Open Source News Data

Data

Redshift Performance & Cost

Published in Data
February 25, 2013 by Henry Cai
9 Comments

At Airbnb, we look into all possible ways to improve our product and user experience. Often times this involves lots of analytics behind the scene. Our data pipeline thus far has consisted of Hadoop, MySQL, R and Stata. We've used a wide variety of libraries for interfacing with our Hadoop cluster such as Hive, Pig, Cascading and Cascalog. However, we found that analysts aren't as productive as they can be by using Hadoop, and standalone MySQL was no longer an option given the size of our dataset. We experimented with frameworks such as Spark but found them to be too immature for our use-case. So we turned our eye to Amazon Redshift earlier this year, and the results have been promising. We saw a 5x performance improvement over Hive.

Redshift is Amazon's SQL based enterprise data warehouse solution for large scale and complex analytics. Under the hood, it is a distributed managed ParAccel cluster. It achieves its high performance through extreme parallelism, columnar data storage, and smart data compression. The setup process is very easy and fast (it took just a few minutes to provision a 16-node cluster), and you can connect to the system via any Postgres compliant client.

Schema Migration The first advice we can offer is to follow what the manual says closely when migrating your tables and queries. We started out by migrating a few large datasets generated by our existing Hadoop based ETL over to Redshift. The first challenge we had was schema migration. Even though Redshift is based on Postgres 8.0, the "subtle" differences are big enough, forcing you into the Redshift way of doing things. We tried to automate the schema migration, but the problem was bigger than we originally expected and we decided it was beyond the scope of our experiment. Indexes, timestamp type, and arrays are not supported in Redshift, thus you need to either get rid of them in your schema or find a workaround (only 11 primitive data types are supported at this point in time). This was the most lengthy and tedious part of the migration to Redshift, but it serves as a very good training and evaluation process (whether Redshift is the right solution for you). When defining your schema, but careful with the distribution key, which determines how your data is distributed across the cluster.

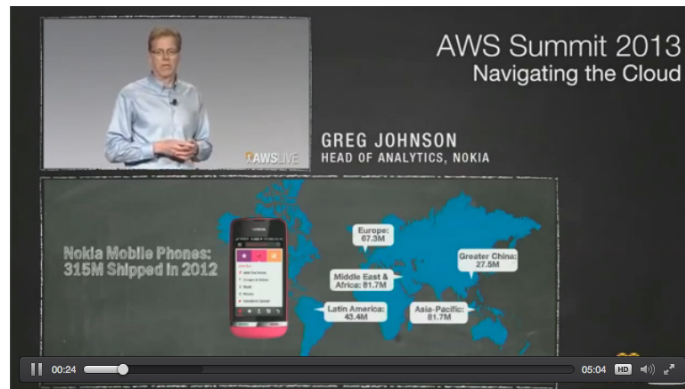
Nokia

Nokia on AWS

NOKIA

About Nokia

Telecommunications giant, [Nokia Corporation](#), uses its Xpress Internet Services platform to provide mobile Internet services for emerging markets in India, Asia Pacific, Africa, and South America. The platform runs on 2200 servers and collects 800 GB of log data daily. The volume of data became too large for the traditional relational database and Nokia could no longer scale the database and generate reports. By moving to AWS and using Amazon Redshift as a fast, fully managed data warehouse, Nokia is able to run queries twice as fast as its previous solution and can use business intelligence tools to mine and analyze big data at a 50% costs savings.



The screenshot shows a presentation slide from the AWS Summit 2013. The title is 'AWS Summit 2013 Navigating the Cloud'. The speaker is Greg Johnson, Head of Analytics at Nokia. The slide features a world map with data points for Nokia mobile phones shipped in 2012: Europe (87.5M), Middle East & Africa (81.7M), Latin America (45.4M), and Asia Pacific (81.7M). A smartphone is shown with the text 'Nokia Mobile Phones: 315M Shipped In 2012'. The video player interface at the bottom shows a progress bar at 00:24 and a total duration of 05:04.

AWS Summit 2013
Navigating the Cloud

GREG JOHNSON
HEAD OF ANALYTICS, NOKIA

Nokia Mobile Phones:
315M Shipped In 2012

Europe: 87.5M
Middle East & Africa: 81.7M
Latin America: 45.4M
Asia Pacific: 81.7M

00:24 05:04

Next Step

To learn more about how AWS can help your enterprise needs, visit our Enterprise on AWS details page: <https://aws.amazon.com/enterprise-tv/>.

Find out more about these and other AWS case studies here: aws.amazon.com/solutions/case-studies/

Resources to learn more

Financial Times Case Study

aws.amazon.com/solutions/case-studies/financial-times

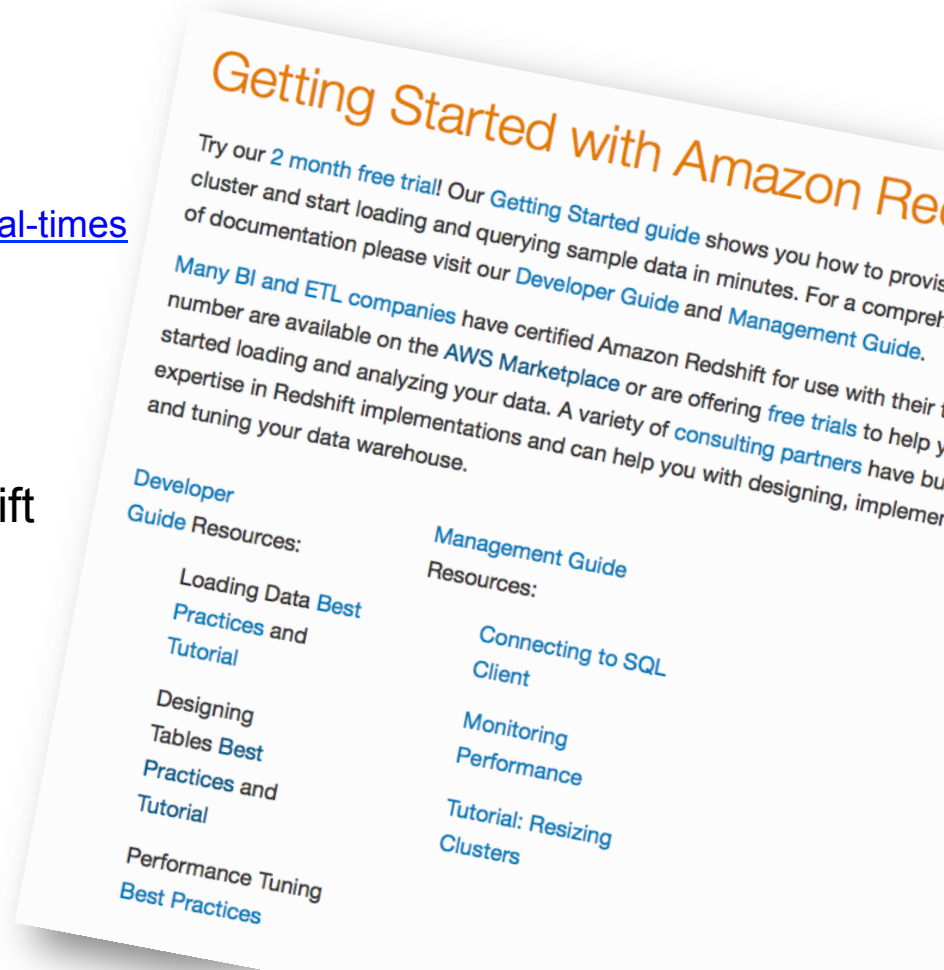
Getting Started with Amazon Redshift

aws.amazon.com/redshift/getting-started/

Uses & Best Practices for Amazon Redshift AWS Summit 2014 San Francisco (video)

<http://youtu.be/reQtXquDpzo>

Your local AWS Solutions Architect ☺



Summary

Optimised for
Data Warehousing

Scalable

No Up-Front Costs

Amazon Redshift

Compatible

Secure

Simple



Find out more:

aws.amazon.com/redshift

AWS Training & Certification

Self-Paced Labs



Try products, gain new skills, and get hands-on practice working with AWS technologies

aws.amazon.com/training/self-paced-labs

Training



Skill up and gain confidence to design, develop, deploy and manage your applications on AWS

aws.amazon.com/training

Certification



Demonstrate your skills, knowledge, and expertise with the AWS platform

aws.amazon.com/certification

Follow us for more
events & webinars



Ian Massingham – Technical Evangelist

 @IanMmmm

 @AWS_UKI for local AWS events & news

 @AWScloud for Global AWS News and Announcements

We typically see customers start by trying our services

All Products

Compute & Networking

Storage

Database

Application Services

Development & Management

AWS Marketplace Software

FAQ »

Find answers to common questions about the AWS Free Tier.



Amazon EC2 »

Web service that provides resizable compute capacity in the cloud.



Amazon S3 »

Highly-scalable, reliable, and low-latency data storage.



Amazon RDS »

Managed MySQL, Oracle and SQL Server databases.



Amazon CloudWatch »

Monitoring for AWS cloud resources and applications.



AWS Data Pipeline »

Orchestration for data-driven workflows.



Amazon DynamoDB »

Fully managed NoSQL database service with seamless scalability.



Amazon EBS »

Highly available, highly reliable, predictable storage volumes.



Amazon ELB »

Web service that provides scalability and high availability.



Amazon ElastiCache »

Managed scale-out caching.



Amazon SNS »

Web service to set up, operate, and send notifications from the cloud.



Amazon Elastic Transcoder »

Convert your media files easily, at low cost and at scale.



Amazon SQS »

Scalable queue for storing messages as they travel between computers.



Amazon SWF »

Workflow service for building scalable, resilient applications.



AWS Marketplace »

Partner software pre-configured to run on AWS.

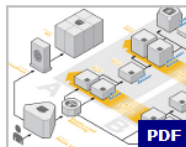
Get started now at : aws.amazon.com/getting-started



Design your application for the AWS Cloud

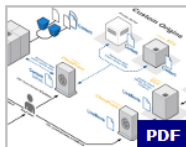
AWS Reference Architectures

The flexibility of AWS allows you to design your application architectures the way you like. AWS Reference Architecture Datasheets provide you with the architectural guidance you need in order to build an application that takes full advantage of the AWS cloud. Each datasheet includes a visual representation of the architecture and basic description of how each service is used.



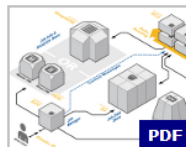
Web Application Hosting

Build highly-scalable and reliable web or mobile-web applications ([PDF](#))



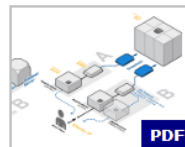
Content and Media Serving

Build highly reliable systems that serve massive amounts of content and media ([PDF](#))



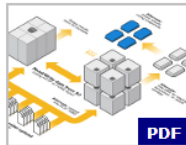
Batch Processing

Build auto-scalable batch processing systems like video processing pipelines ([PDF](#))



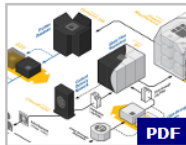
Fault tolerance and High Availability

Build systems that quickly failover to new instances in an event of failure ([PDF](#))



Large Scale Processing and Huge Data sets

Build high-performance computing systems that involve Big Data ([PDF](#))



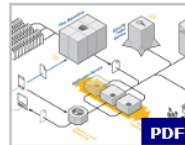
Ad Serving

Build highly-scalable online ad serving solutions ([PDF](#))



Disaster Recovery for Local Applications

Build cost-effective Disaster Recovery solutions for on-premises applications ([PDF](#))



File Synchronization

Build simple file synchronization service ([PDF](#))