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

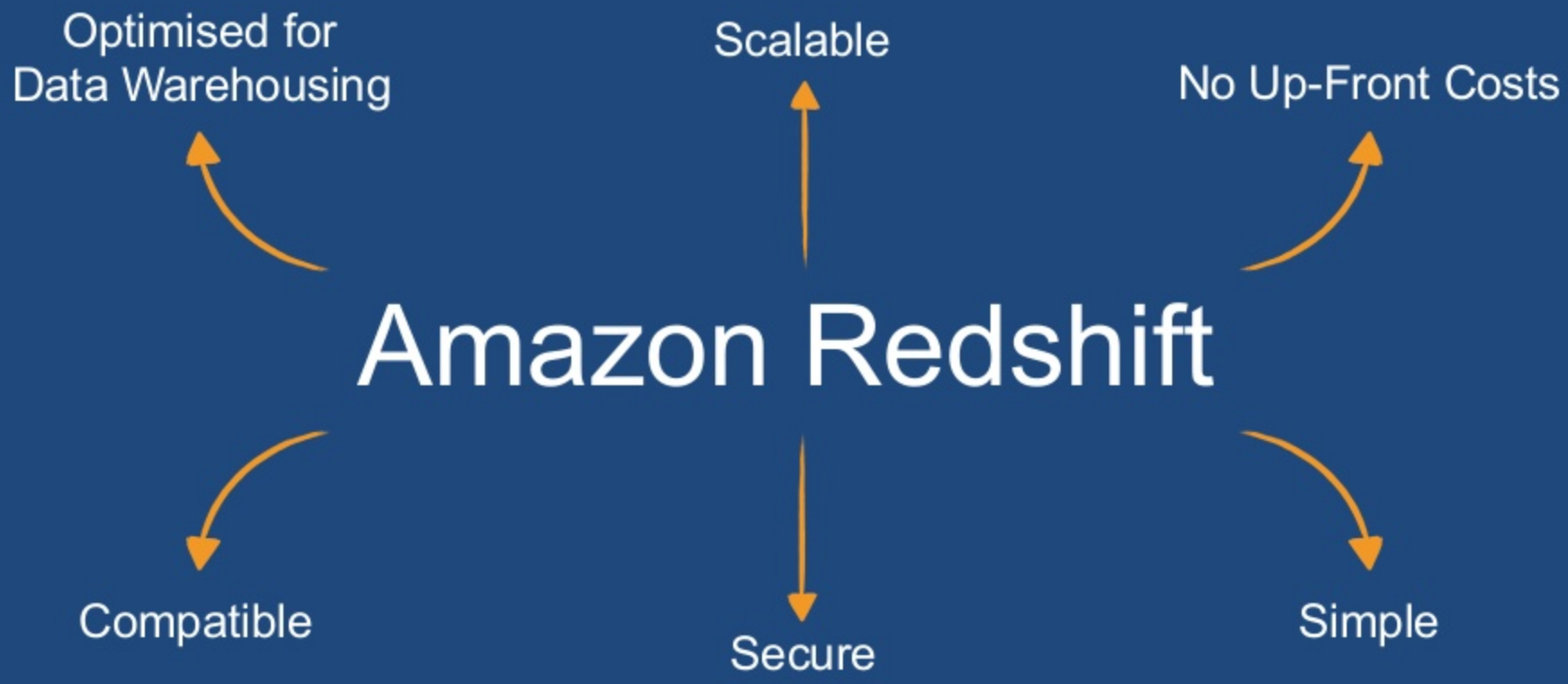
Broaden your knowledge in ~45 mins

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

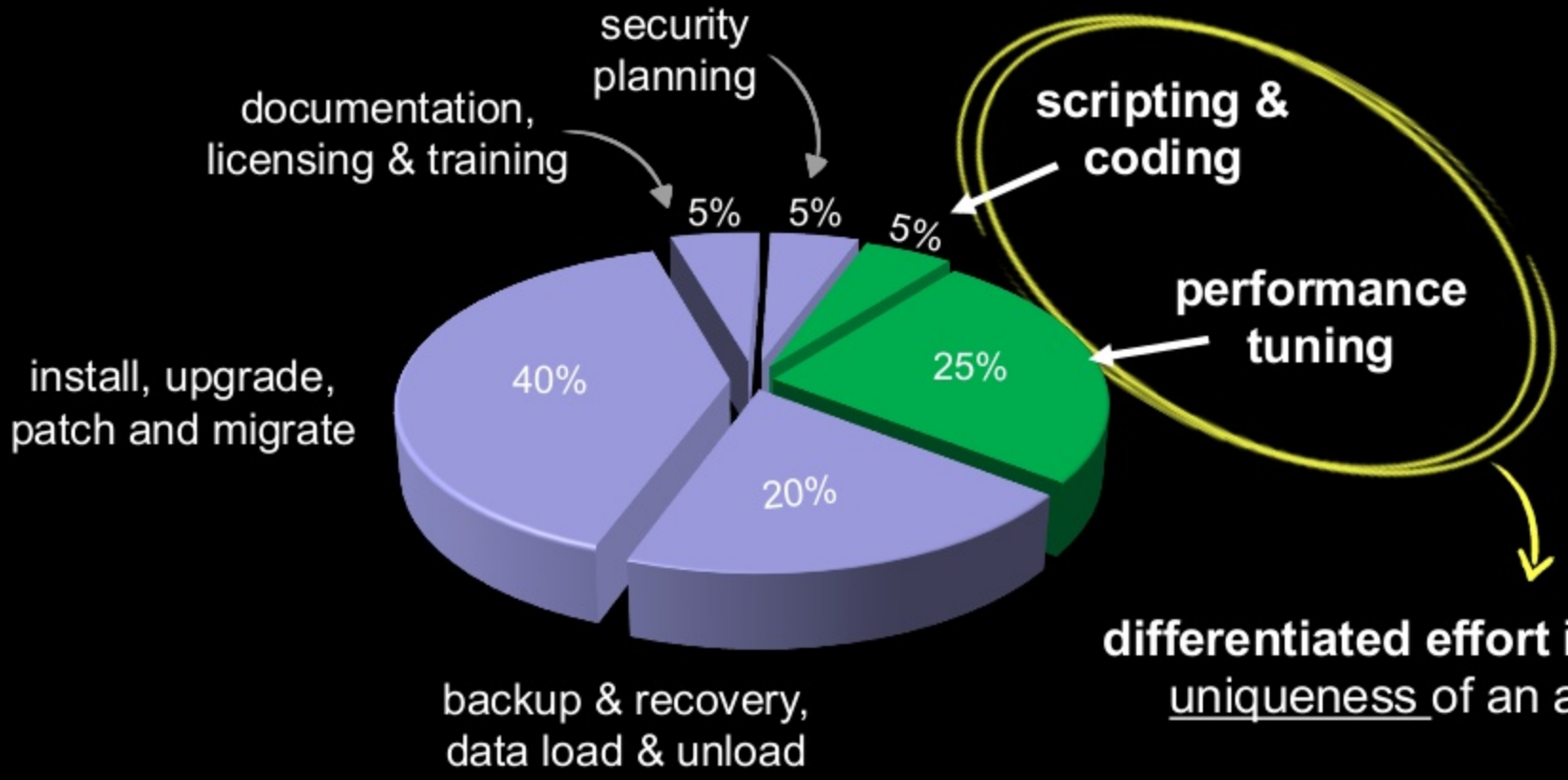

Optimised for
Data Warehousing
Scalable
No Up-Front Costs
Amazon Redshift
Compatible
Secure
Simple

Agenda

- Why Run Your Data Warehouse on AWS?
- Getting Started
- Table Design
- Data Loading
- Working with Data
- Backup and Restoration
- Upgrading & Scaling

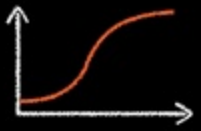
Why Run Your Data Warehouse on AWS?

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

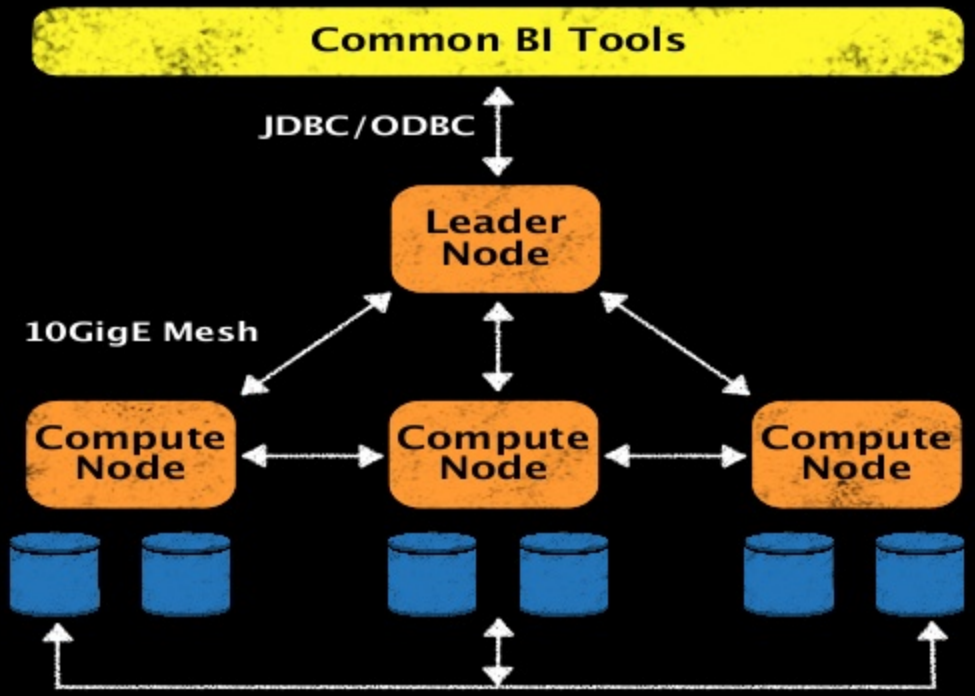
Query

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

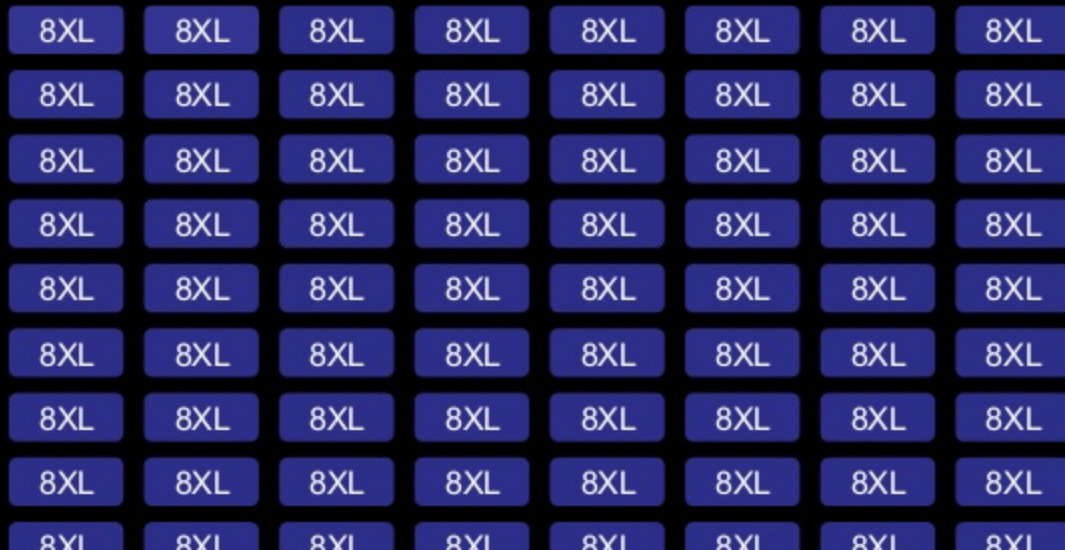
Single Node (160GB SSD or 2TB Magnetic)

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

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)



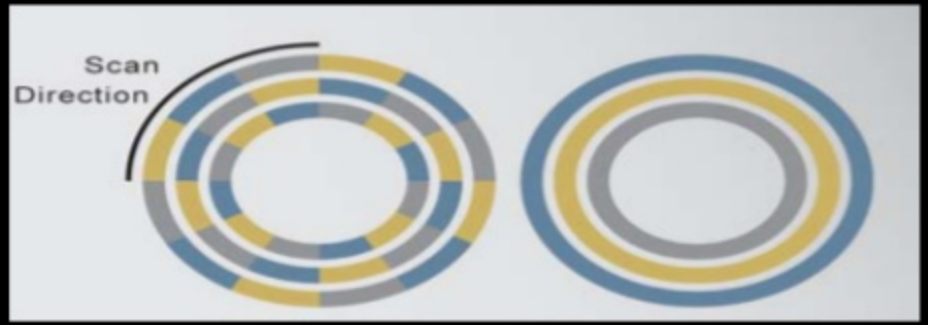
Redshift simplifies operations

- 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

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

TEST:

2 BILLION RECORDS

6 REPRESENTATIVE REQUESTS

AMAZON REDSHIFT 2x HS1.8XL

Vs.

32 NODES, 4.2TB RAM, 1.6PB

amazon.com

12x TO 150x FASTER

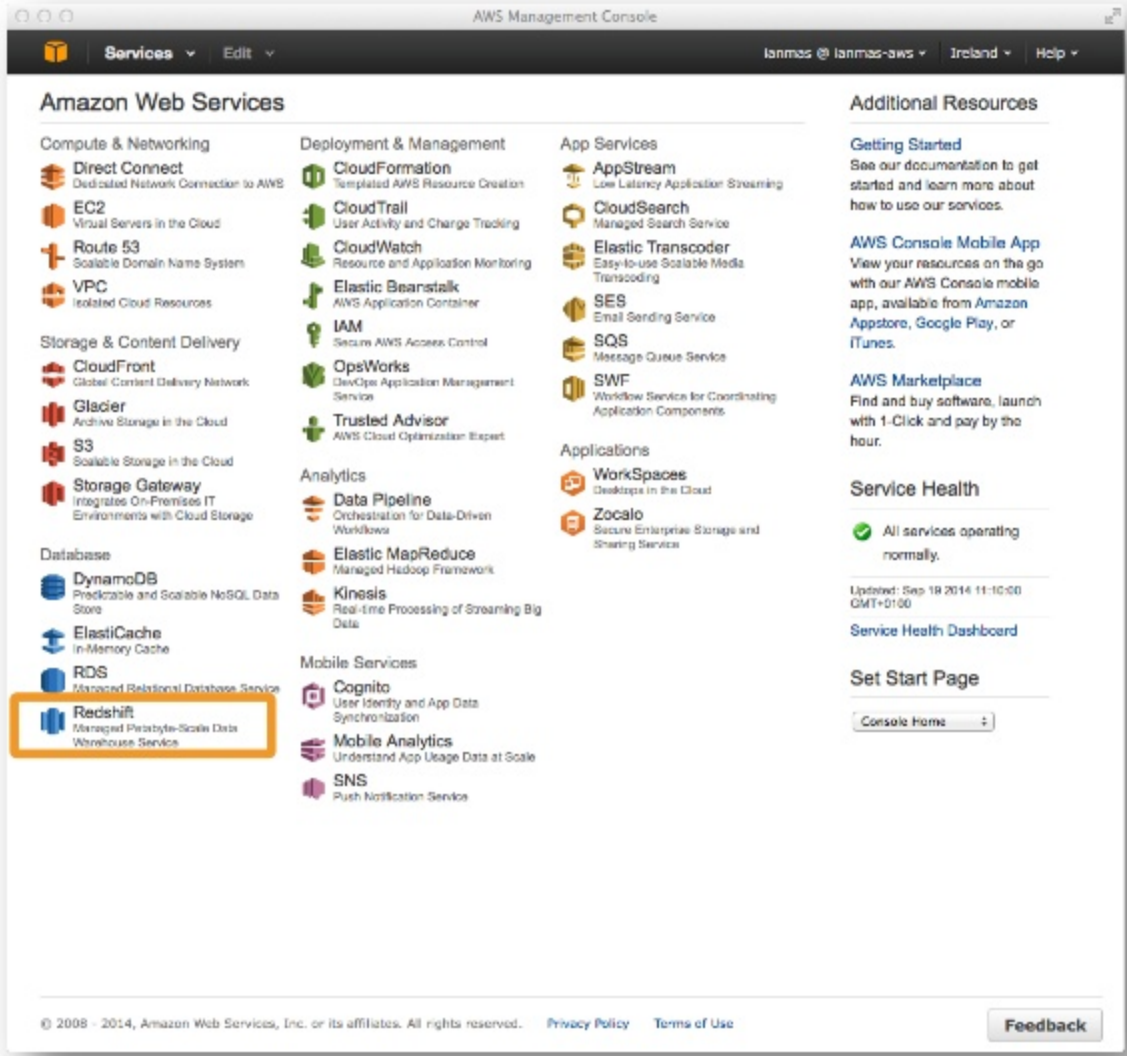
Query Impact	Before (s)	After (s)	Speedup
29m 58s -> 12s	1,798	12	149.8x
22m 2s -> 16s	1,322	16	82.6x
7m 48s -> 3s	468	3	156.x
5m 30s -> 5s	330	5	66.x
49m 8s -> 35s	2,948	35	84.2x
2m 9s -> 10s	129	10	12.9x
Total Speedup	6,995	81	86.4x

- 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

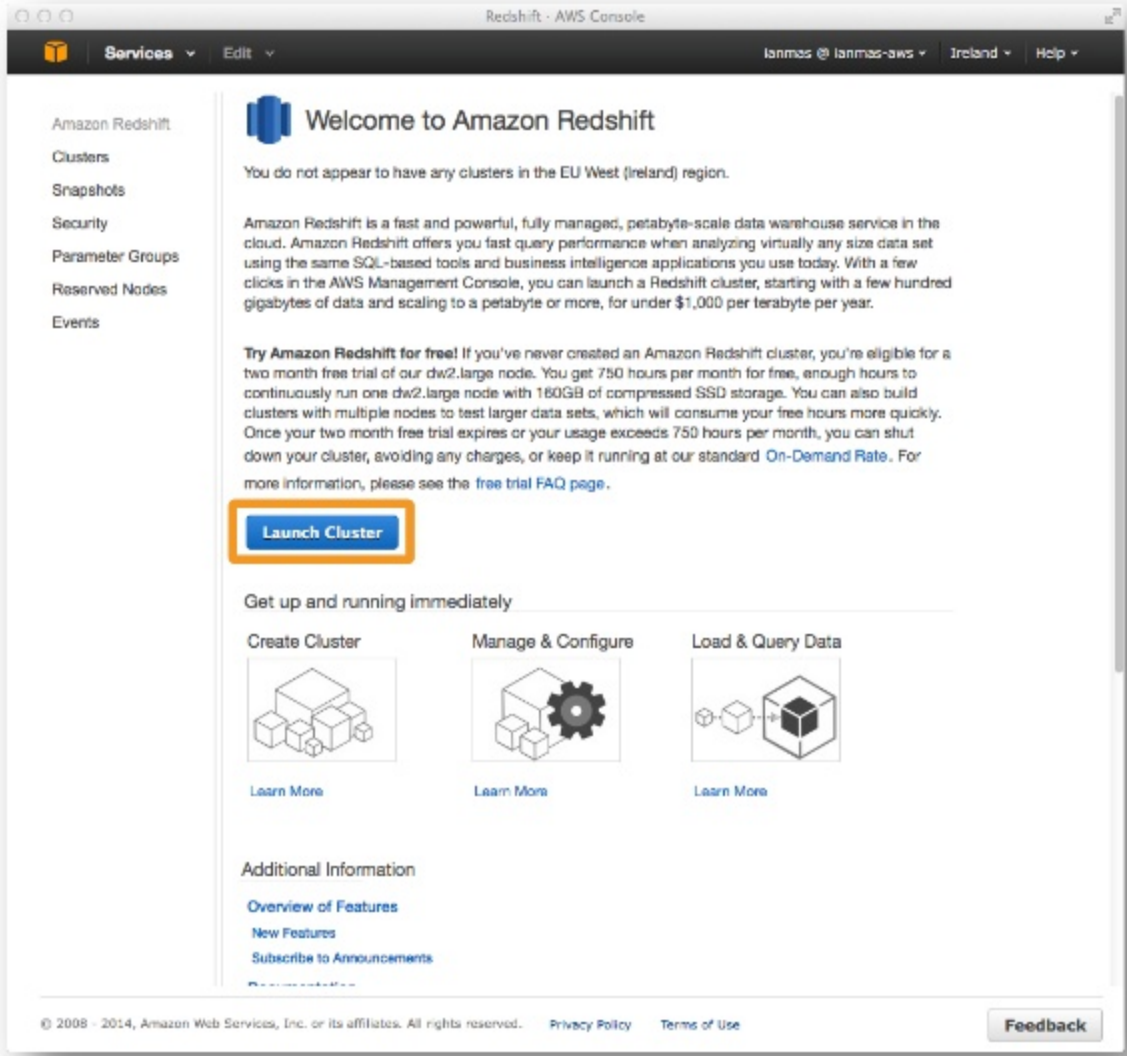
Getting Started

Create a Data Warehouse in Minutes with Amazon Redshift

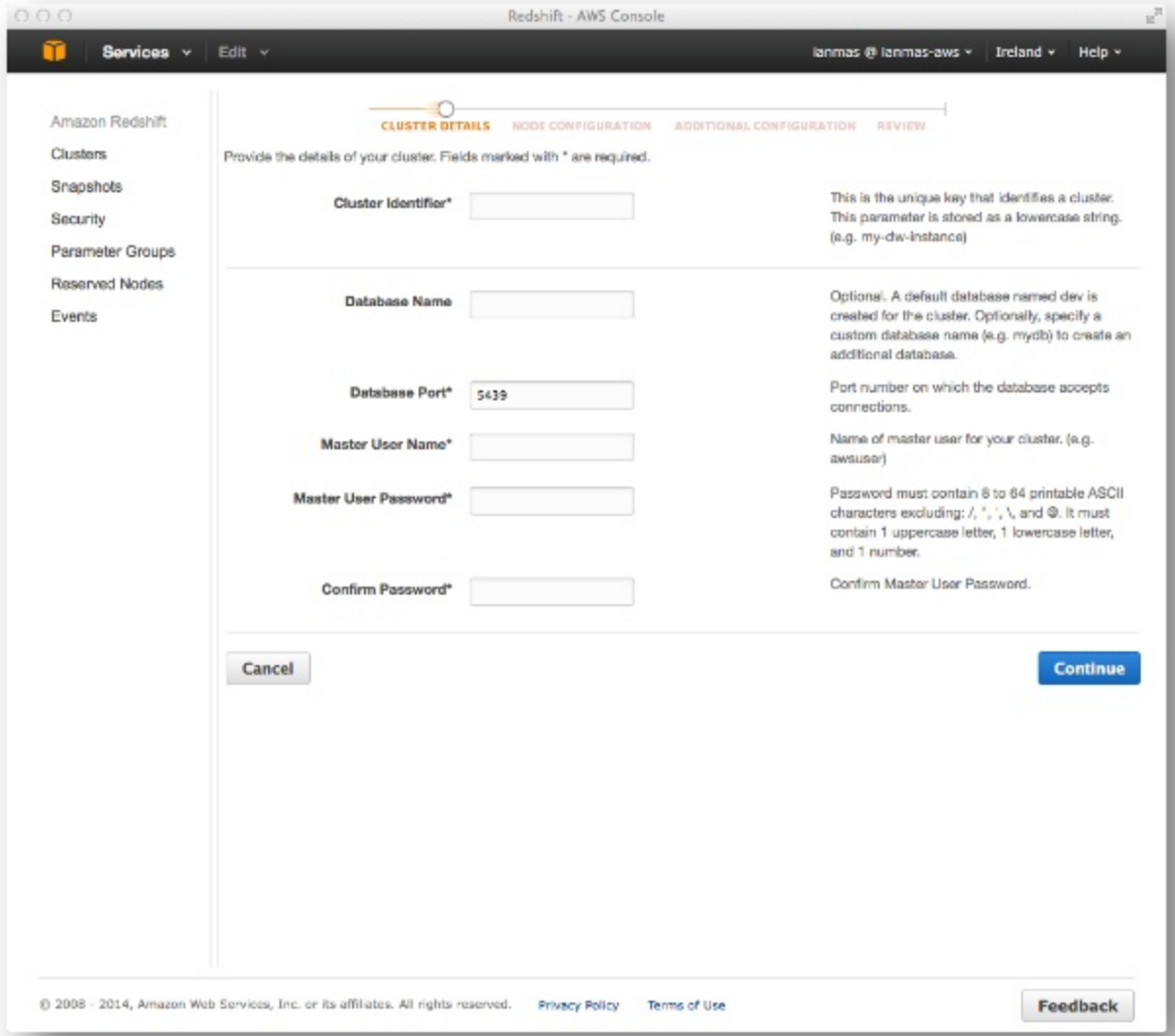
Creating a Redshift Data Warehouse via the AWS Console



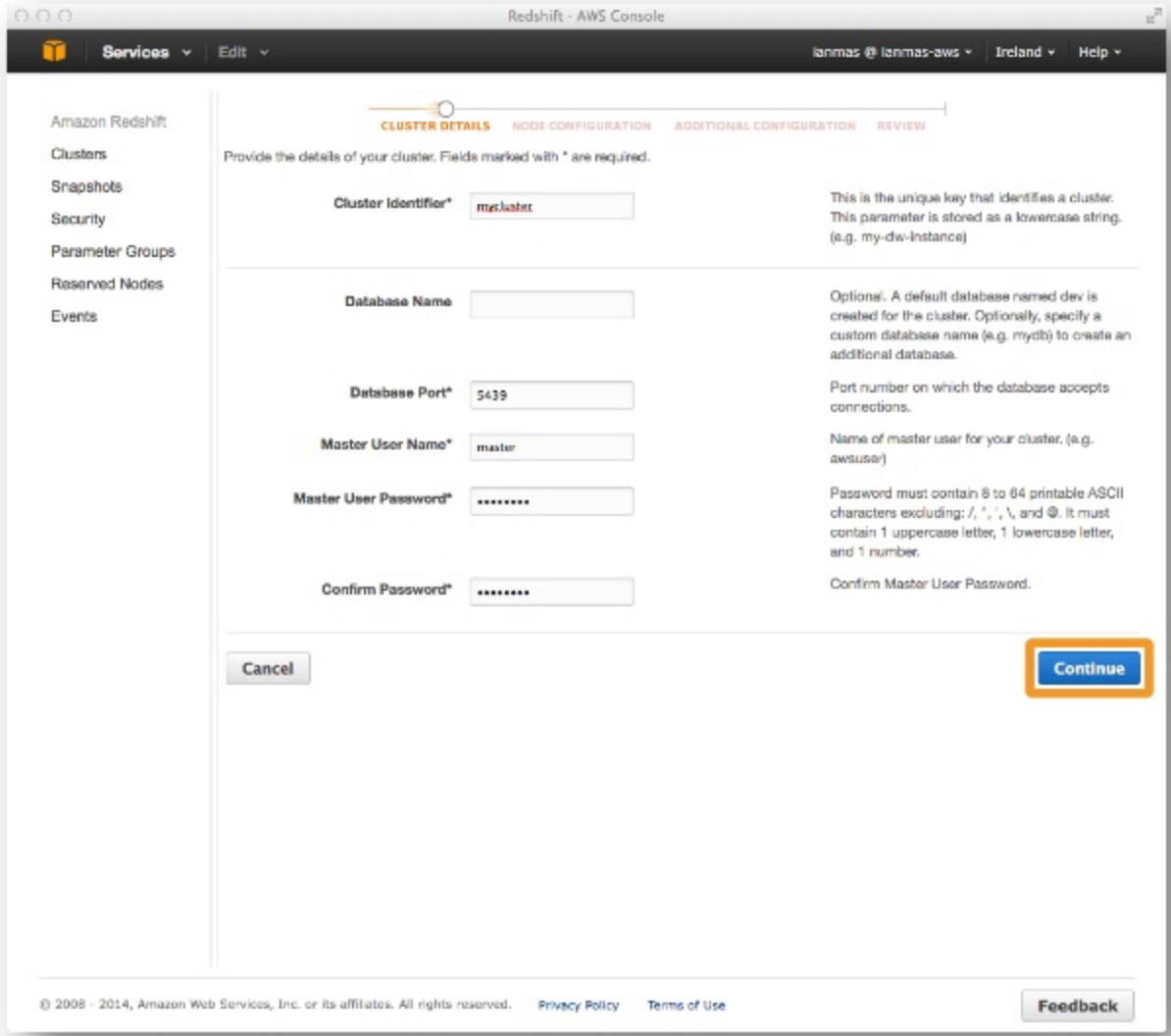
Creating a Redshift Data Warehouse via the AWS Console



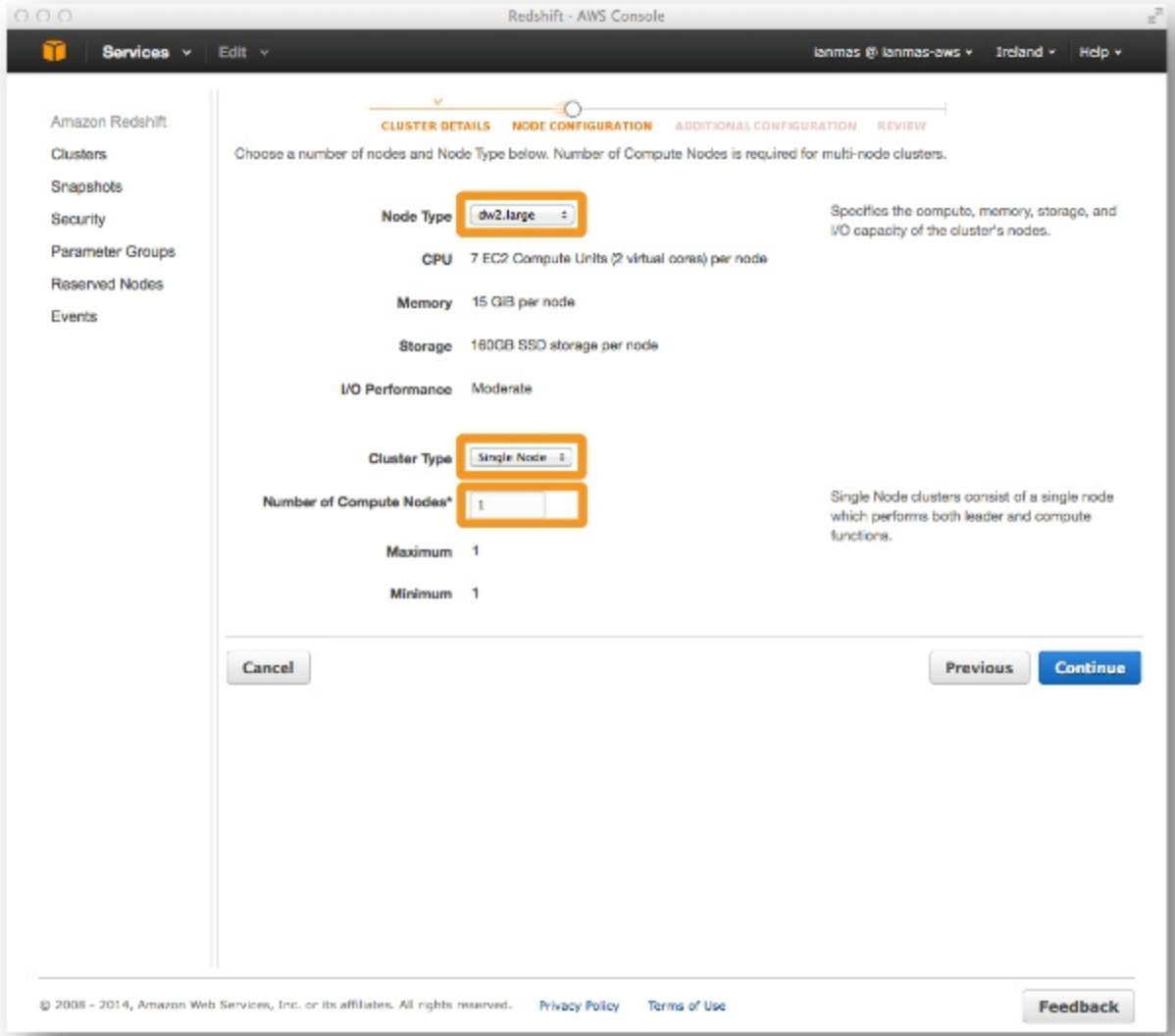
Creating a Redshift Data Warehouse via the AWS Console



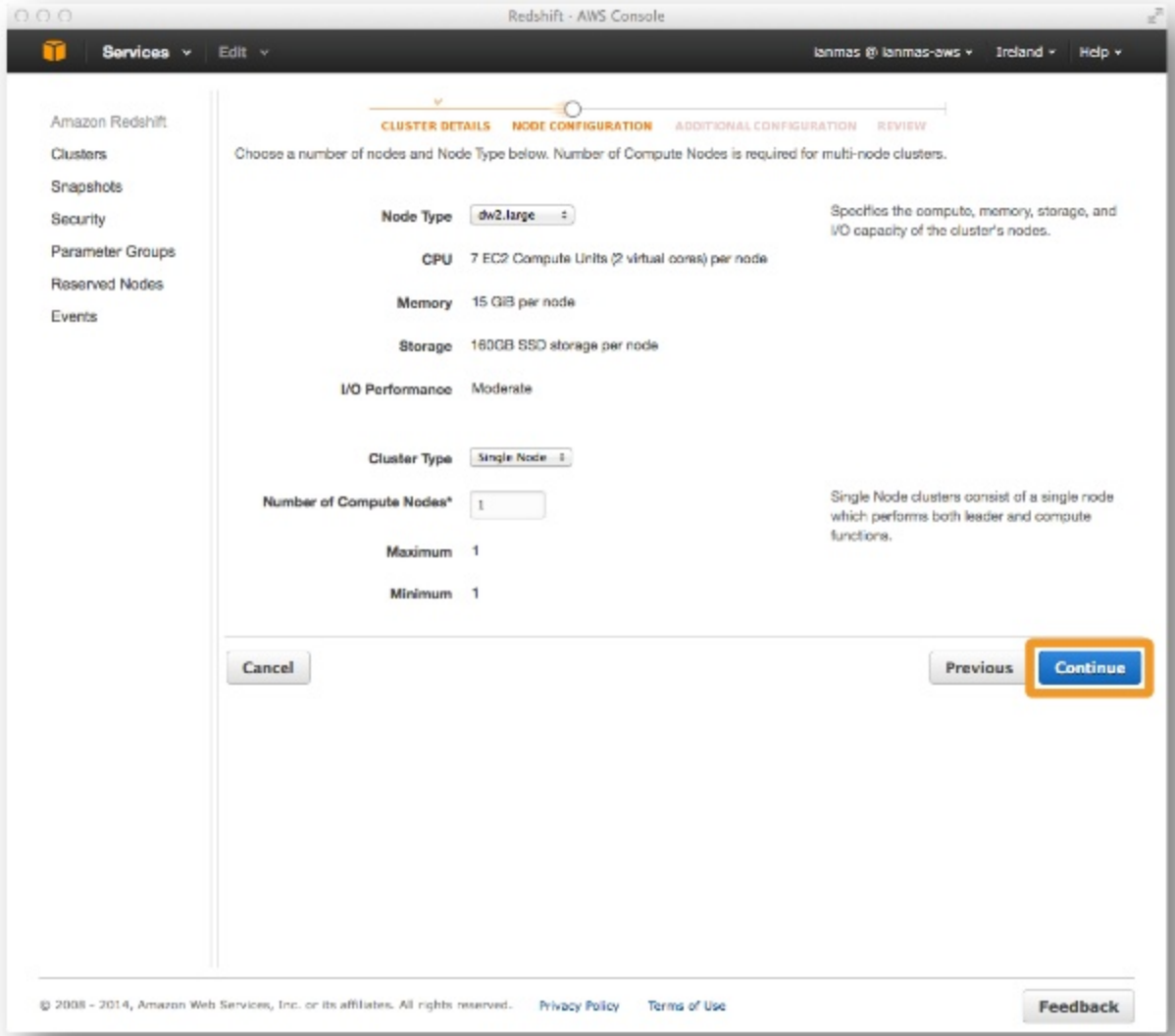
Creating a Redshift Data Warehouse via the AWS Console



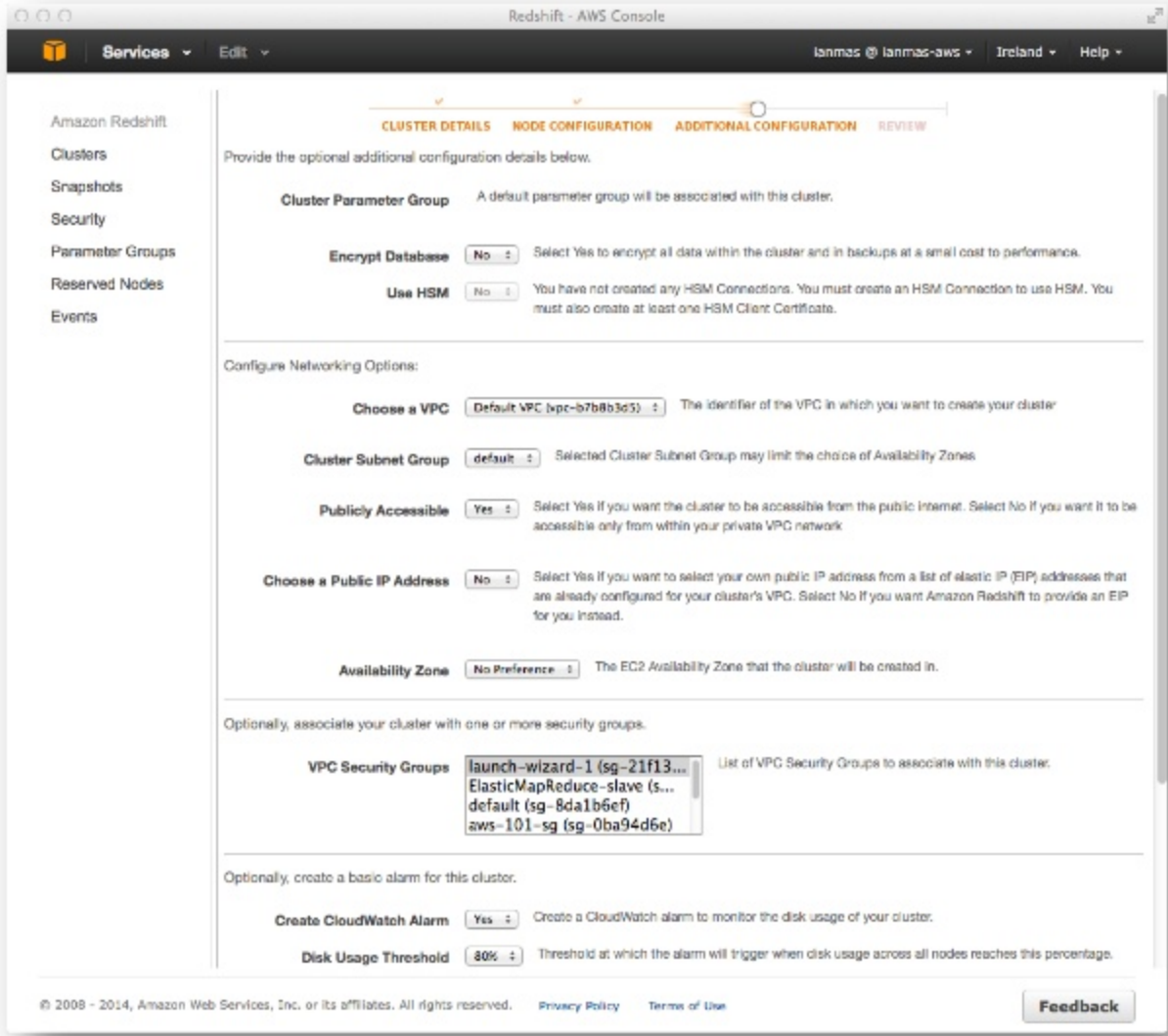
Creating a Redshift Data Warehouse via the AWS Console



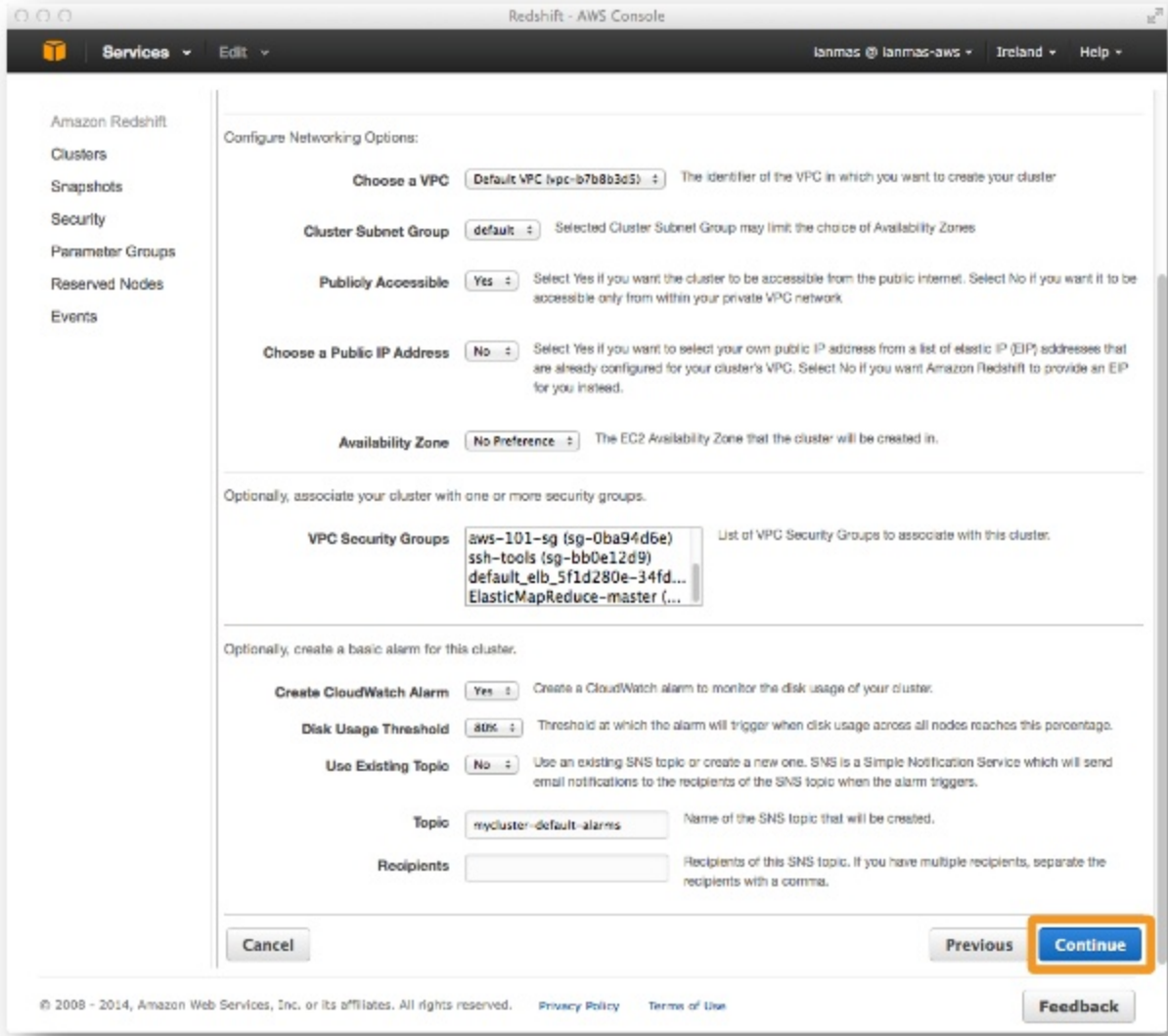
Creating a Redshift Data Warehouse via the AWS Console



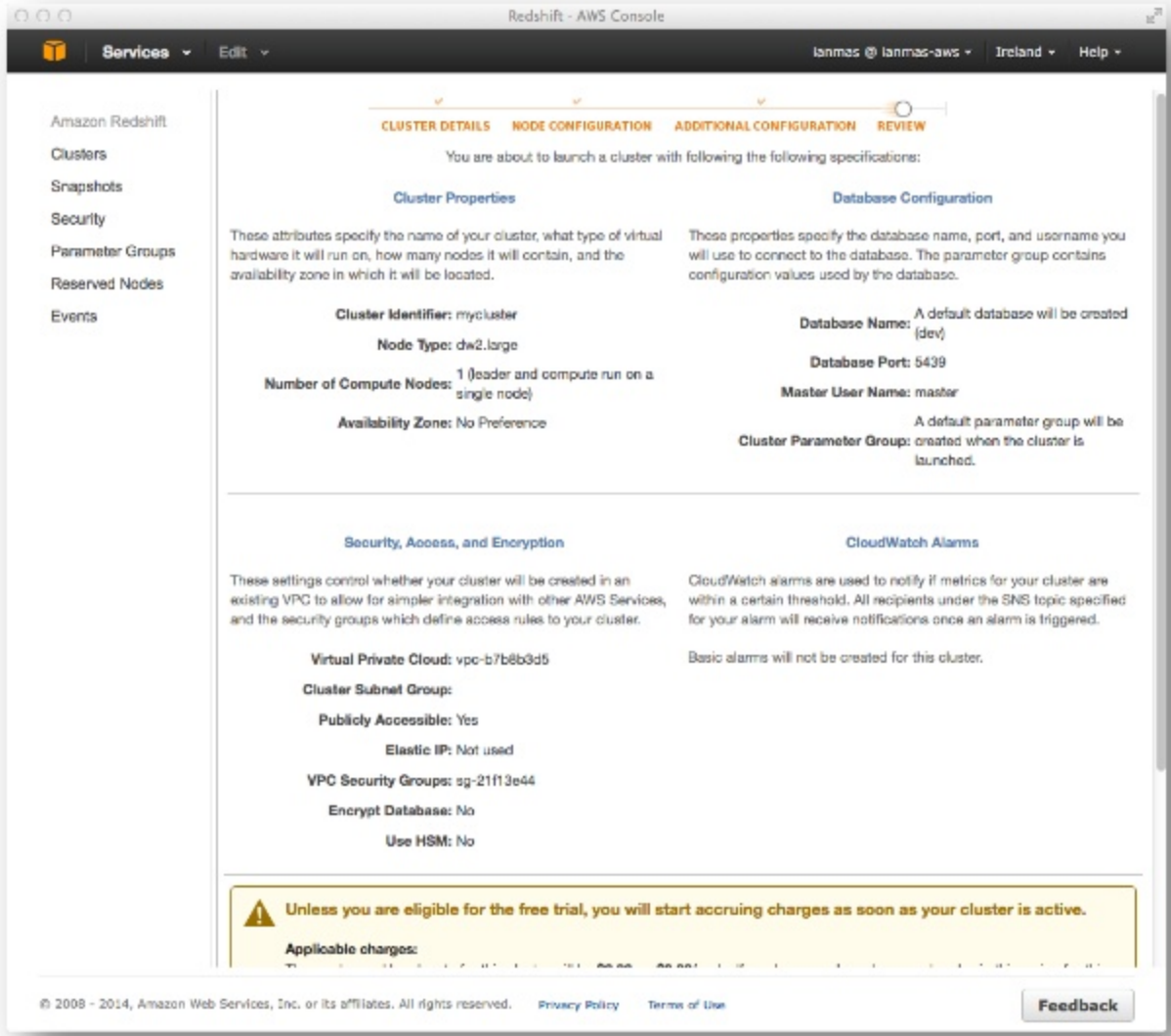
Creating a Redshift Data Warehouse via the AWS Console



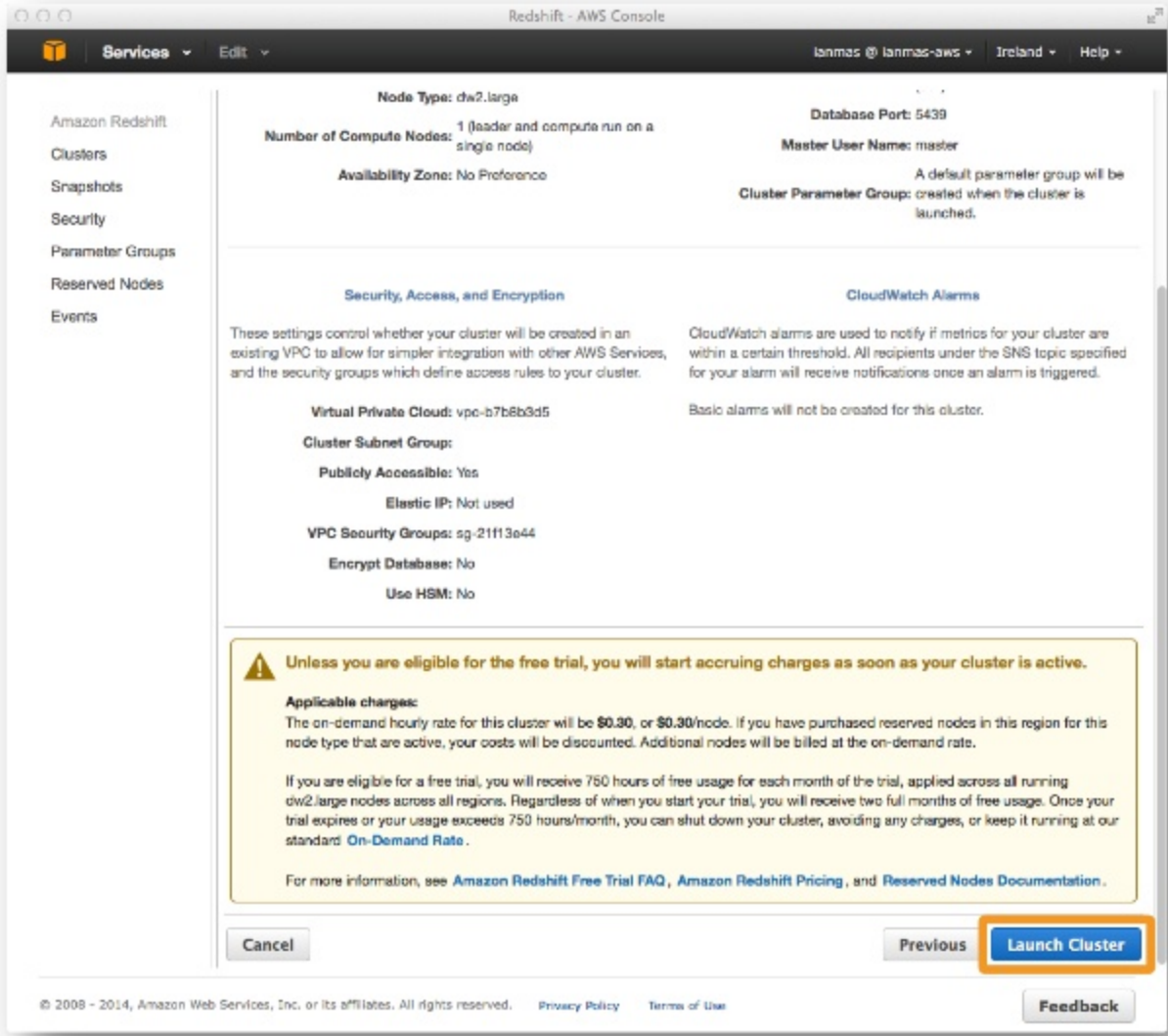
Creating a Redshift Data Warehouse via the AWS Console



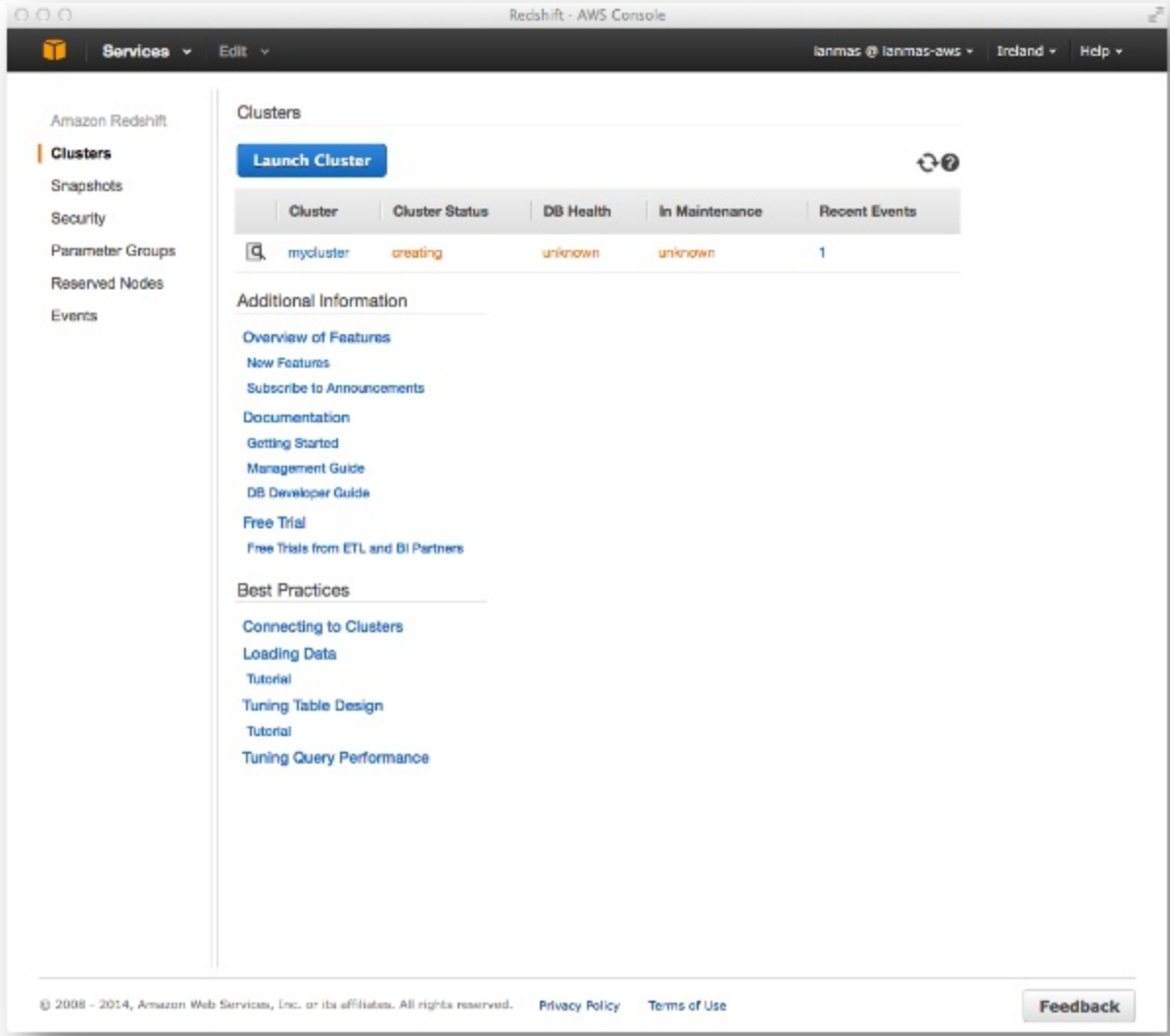
Creating a Redshift Data Warehouse via the AWS Console



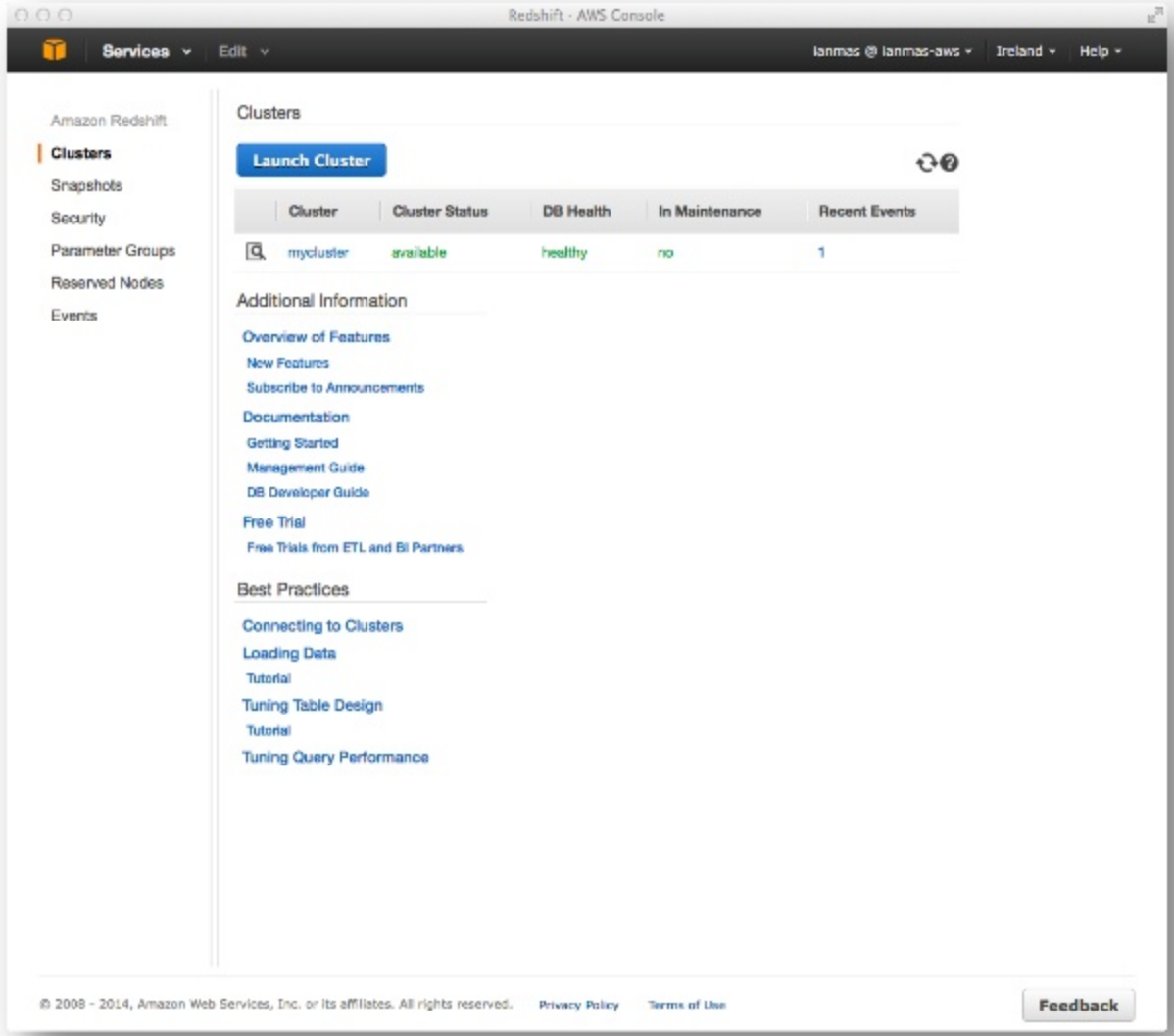
Creating a Redshift Data Warehouse via the AWS Console



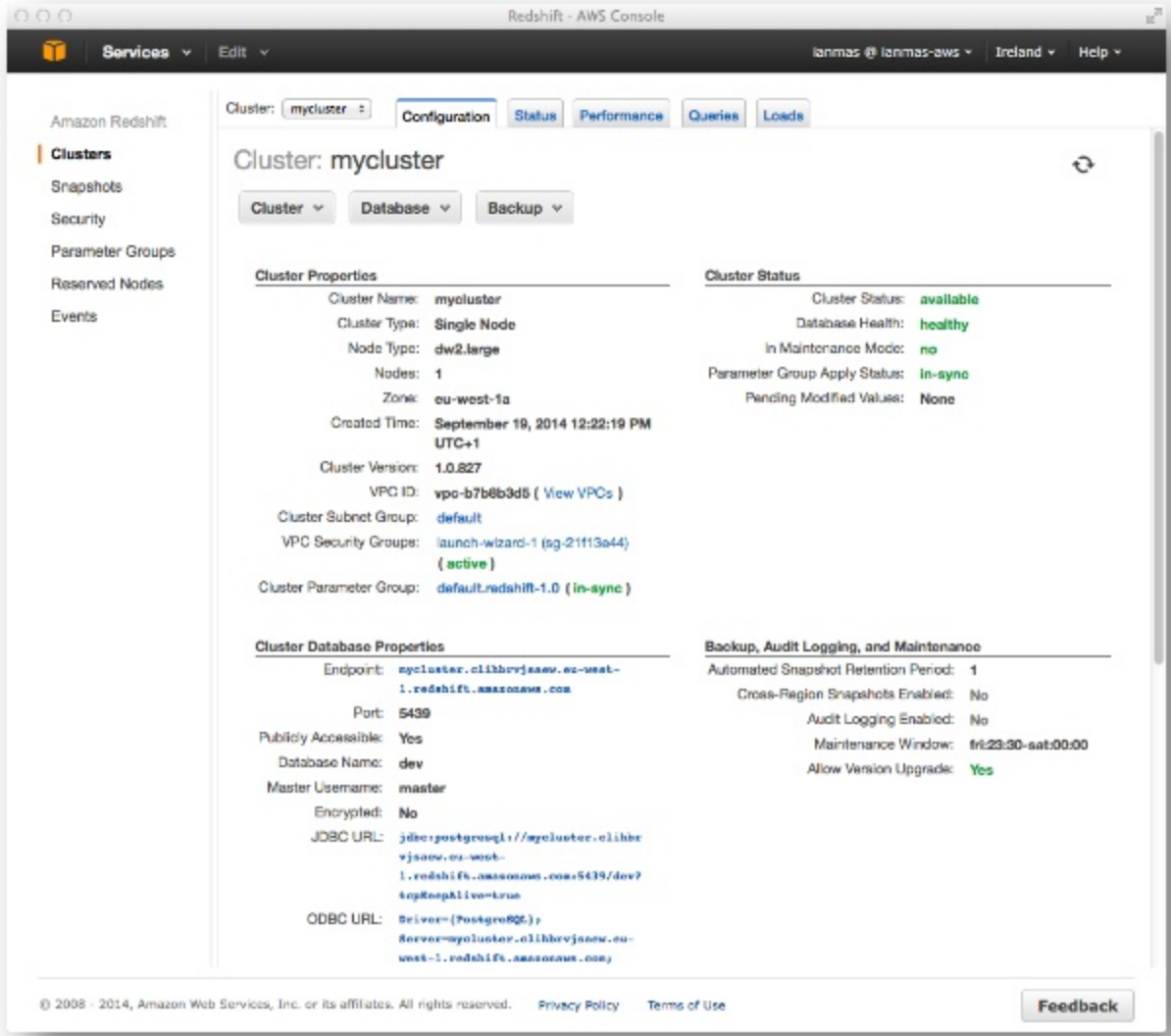
Creating a Redshift Data Warehouse via the AWS Console



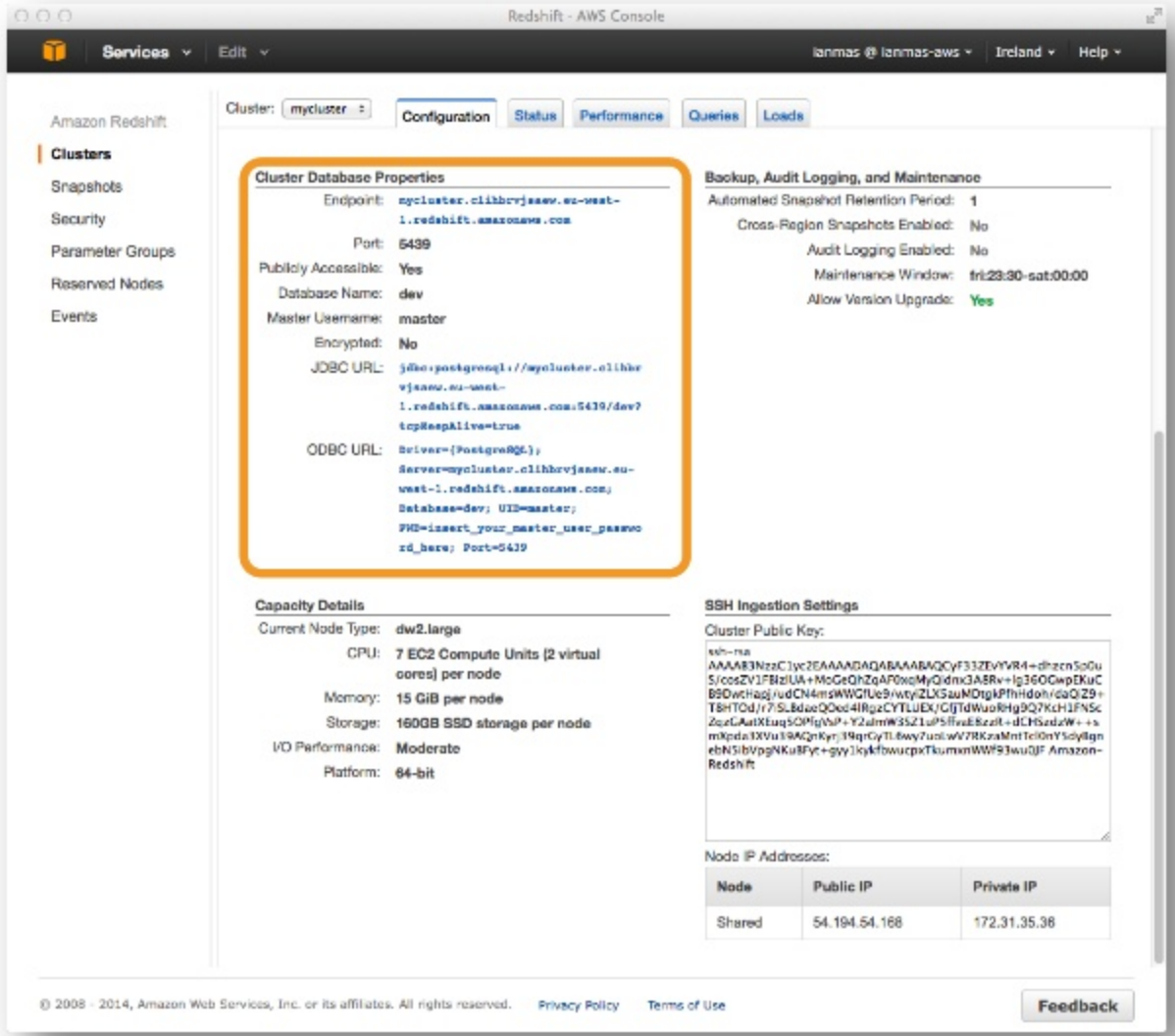
Creating a Redshift Data Warehouse via the AWS Console



Creating a Redshift Data Warehouse via the AWS Console



Creating a Redshift Data Warehouse via the AWS Console



More Getting Started Resources

2 Month Free Trial

Comprehensive Getting Started Guide

Table Design & Data Loading Tutorials

Find out more about getting started with Amazon Redshift: aws.amazon.com/redshift/getting-started/

Table Design

Table Design

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

Data Types

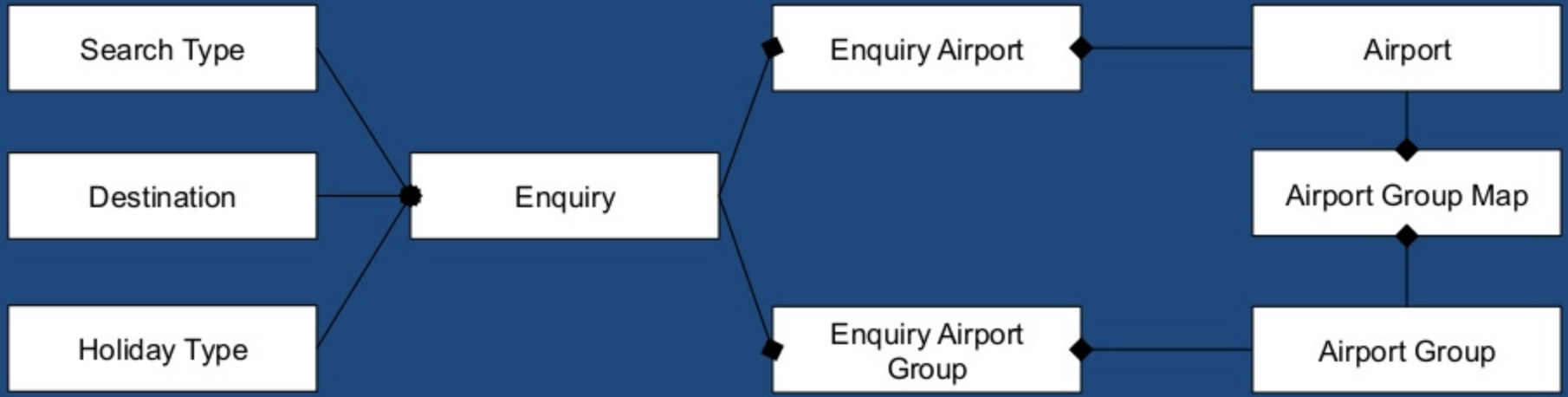
Data type	Aliases	Description
SMALLINT	INT2	Signed two-byte integer
INTEGER	INT, INT4	Signed four-byte integer
BIGINT	INT8	Signed eight-byte integer
DECIMAL	NUMERIC	Exact numeric of selectable precision
REAL	FLOAT4	Single precision floating-point number
DOUBLE PRECISION	FLOAT8	Double precision floating-point 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 characters x 3 bytes per character = 12 bytes

Your Schema

- 3NF



Distributing Data

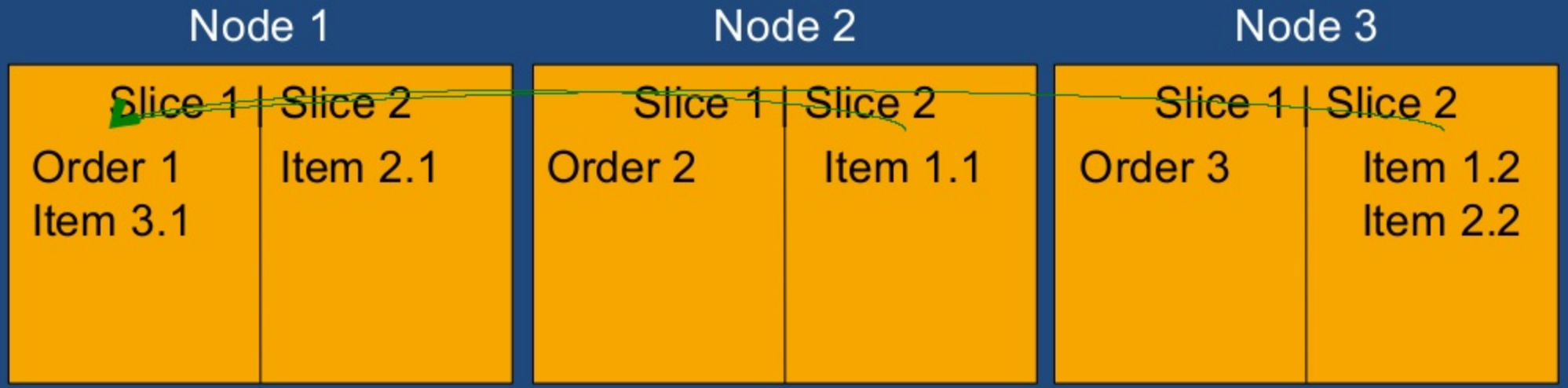
- Redshift is a distributed system:
  - A cluster contains a leader node & compute nodes
  - A compute node contains slices (one per core)
  - A slice contains data
- Slices are chosen based on two types of distribution:
  - Round Robin (automated)
  - Based on a distribution key (hash of a defined column)
- 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



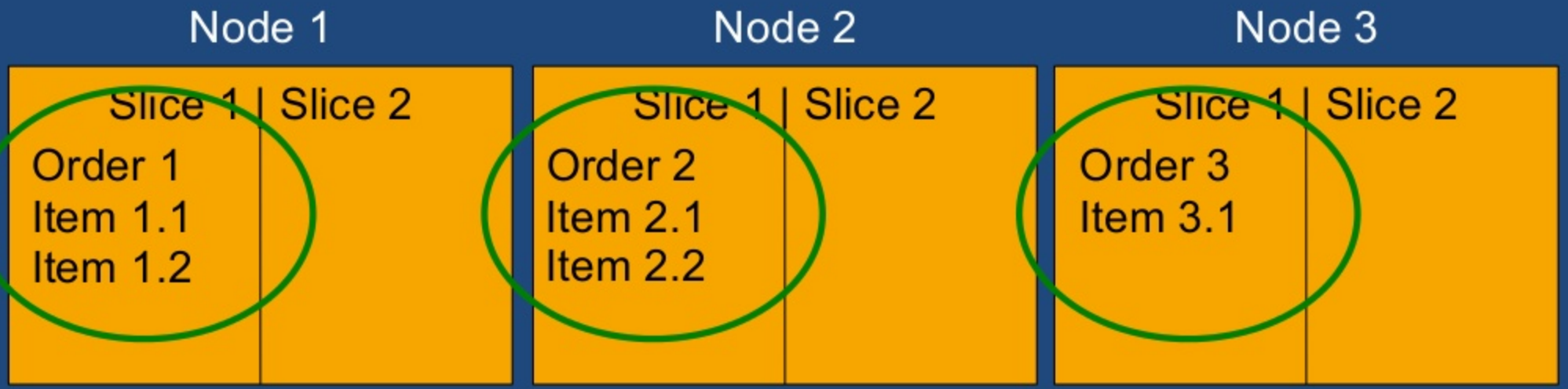
Default (No Distribution Key, Round Robin Order)

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



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

http://docs.aws.amazon.com/redshift/latest/dg/c_analyzing-table-design.html

Sorting Data

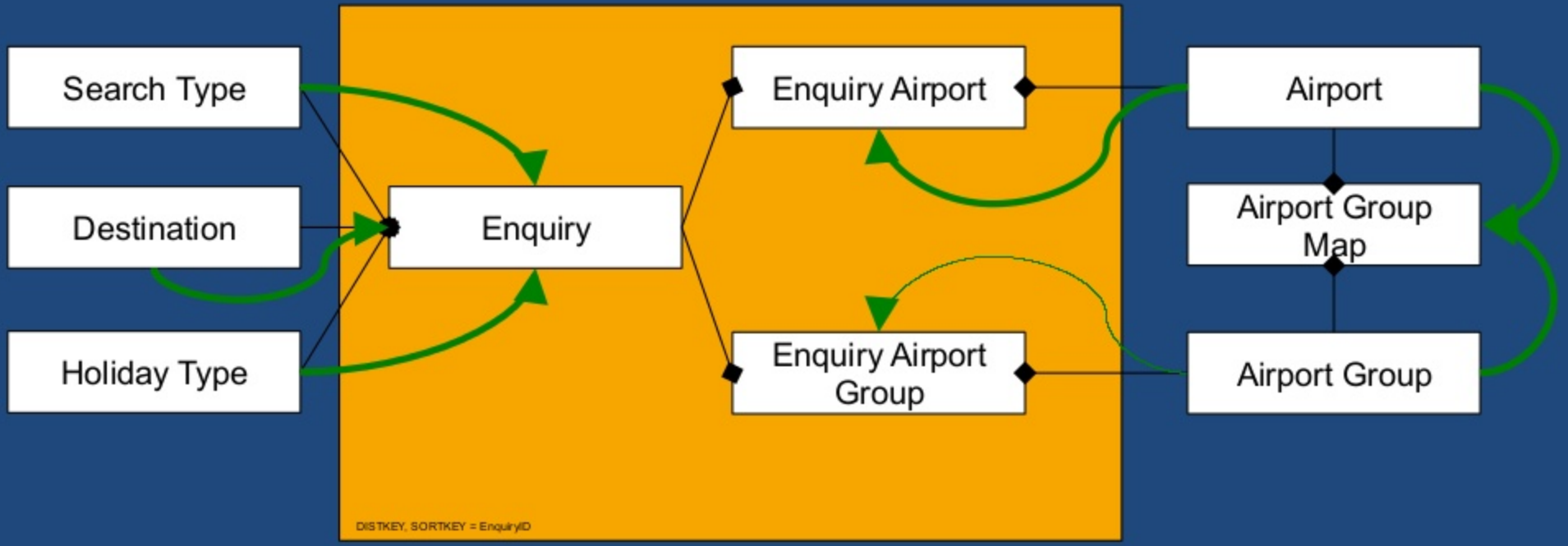
- In the slices (on disk), the data is sorted by a sort key (if none uses the insertion order)
- Choose a sort key that is frequently used in your queries
  - As a query predicate (date, identifier, …)
  - As a join parameter (it can also be the hash key)
- The sort key allows Redshift to avoid reading 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

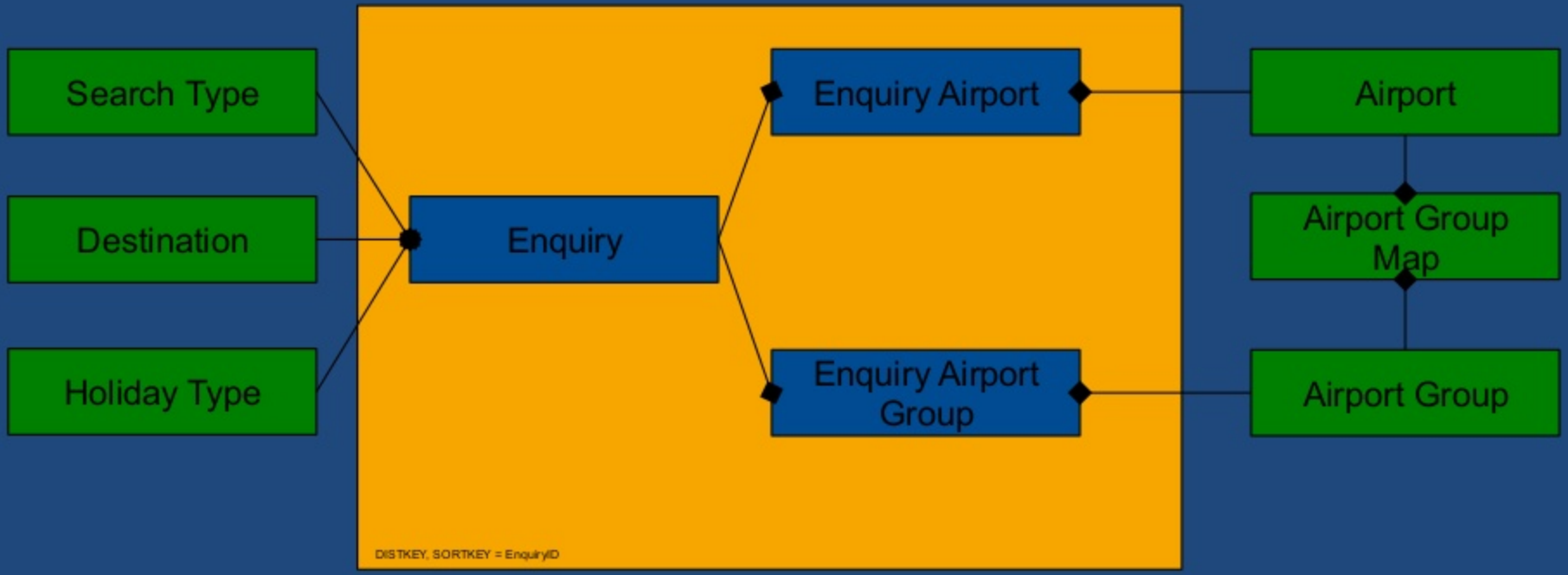
Possible Schema Optimisations

- Denormalise commonly used join attributes onto large tables



Possible Schema Optimisations

- Cache Small Tables with DISTYLE ALL



More on Table Design

Table Design Documentation

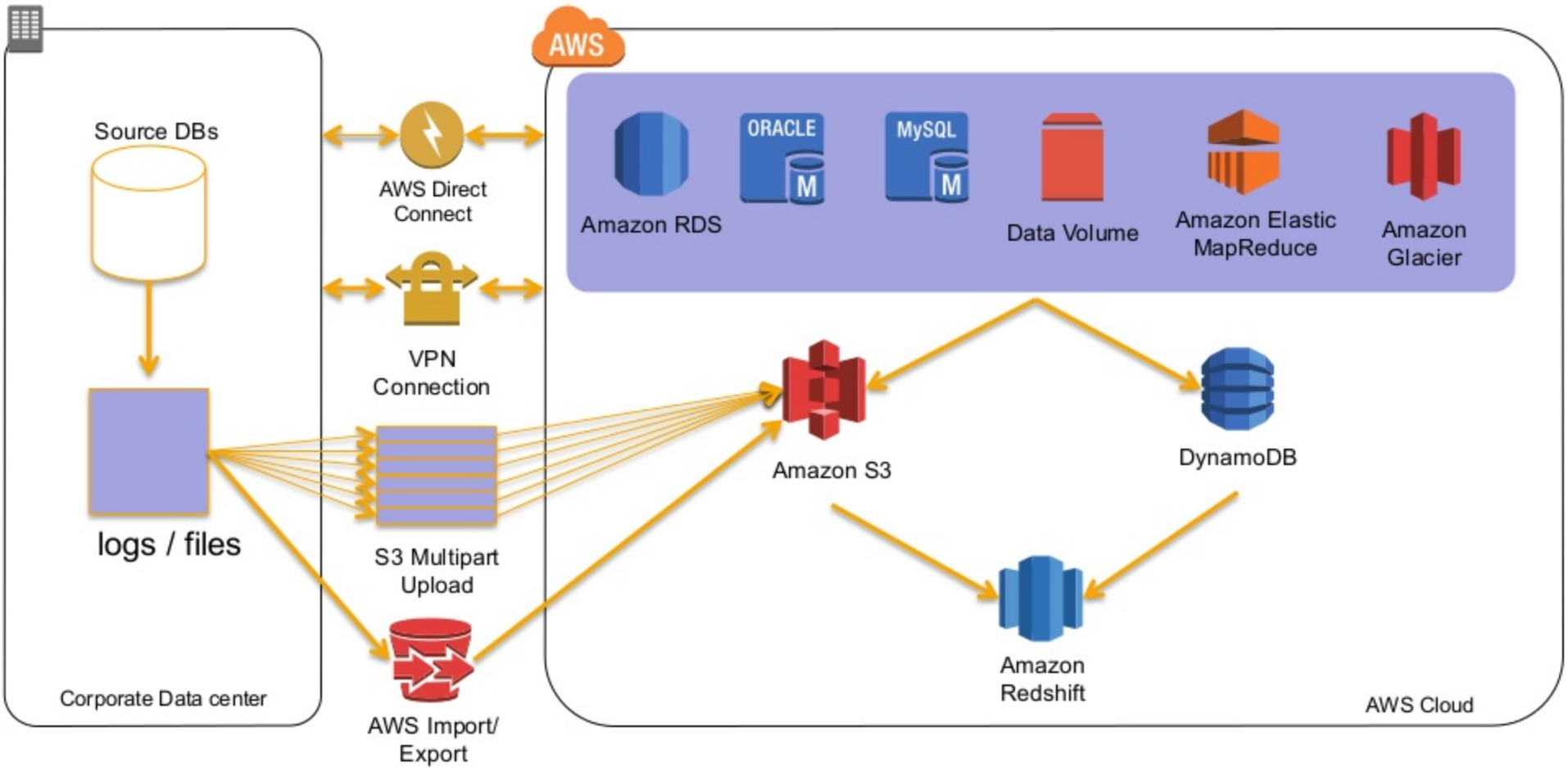
http://docs.aws.amazon.com/redshift/latest/dg/t_Creating_tables.html

Table Design Tutorial

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

Data Loading

Amazon Redshift Loading Data Overview

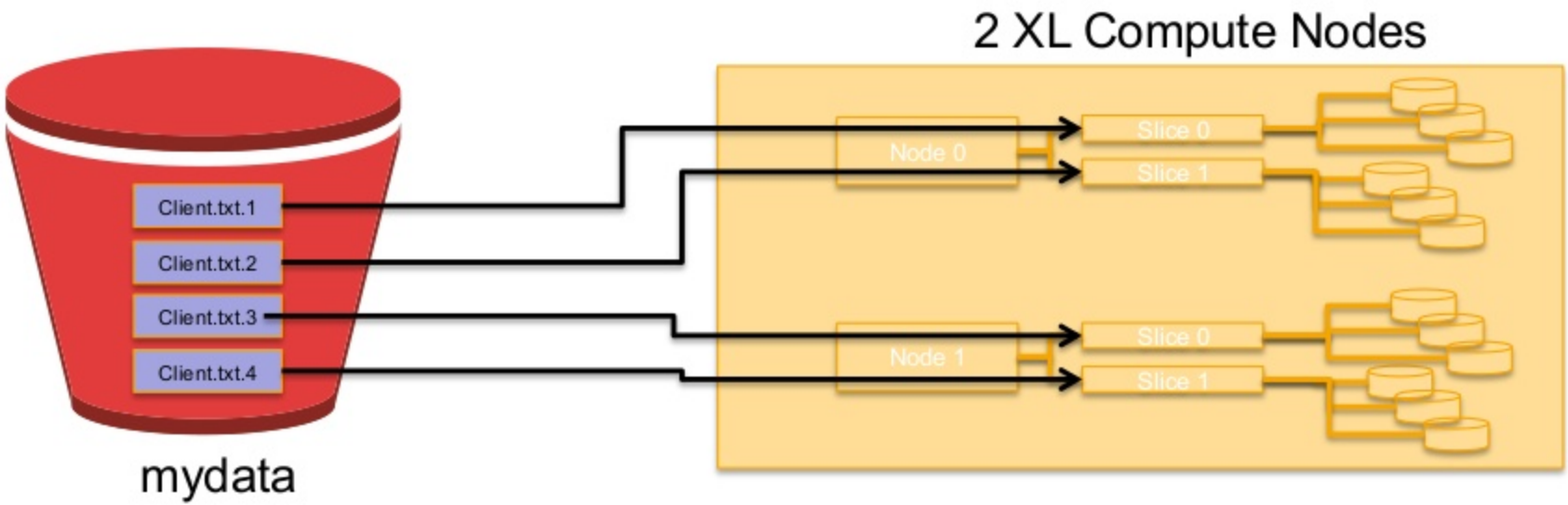


Sharing Your Data

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

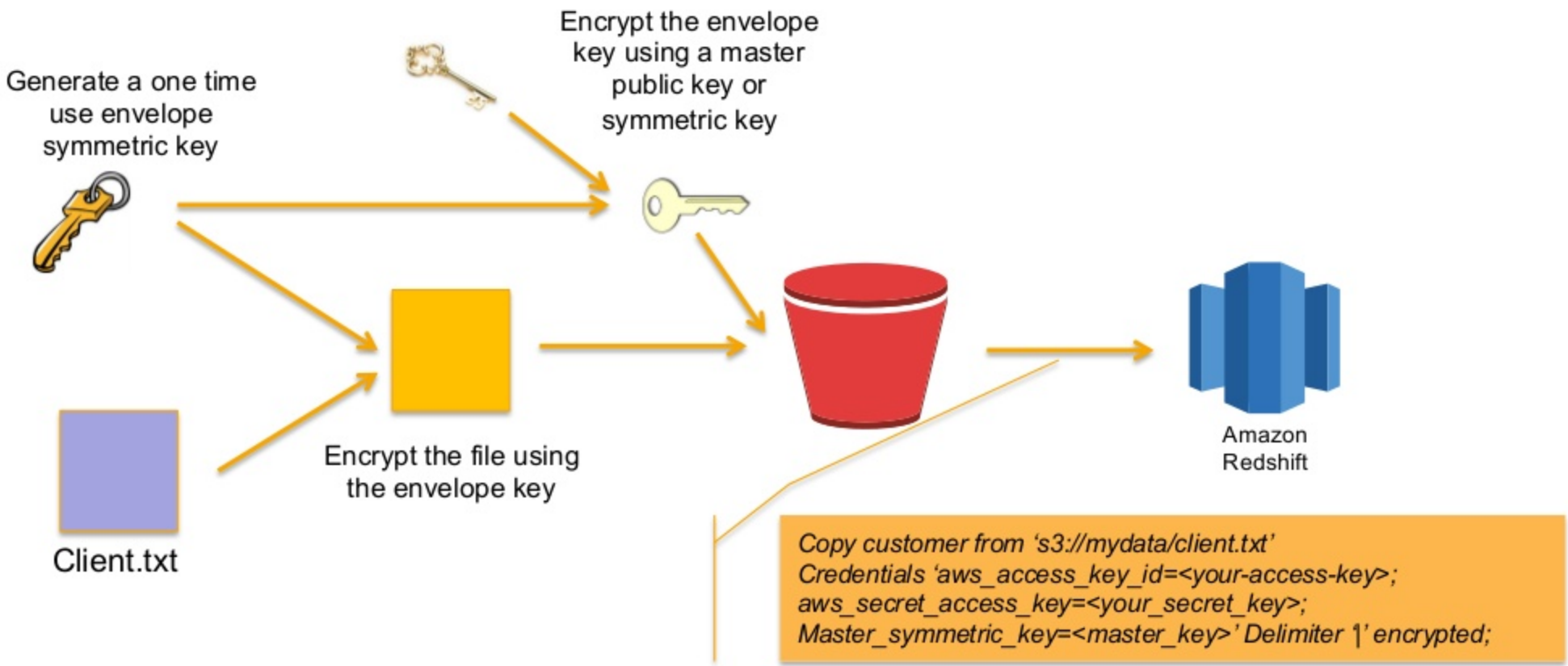
```
{
  "Id": "Policy1234567890",
  "Statement": [
    {
      "Action": [
        "s3:Get*",
        "s3:List*"
      ],
      "Effect": "Allow",
      "Principal": {
        "AWS": ["arn:aws:iam::887210671223:root"]
      },
      "Resource": [
        "arn:aws:s3:::<bucket-name>",
        "arn:aws:s3:::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



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

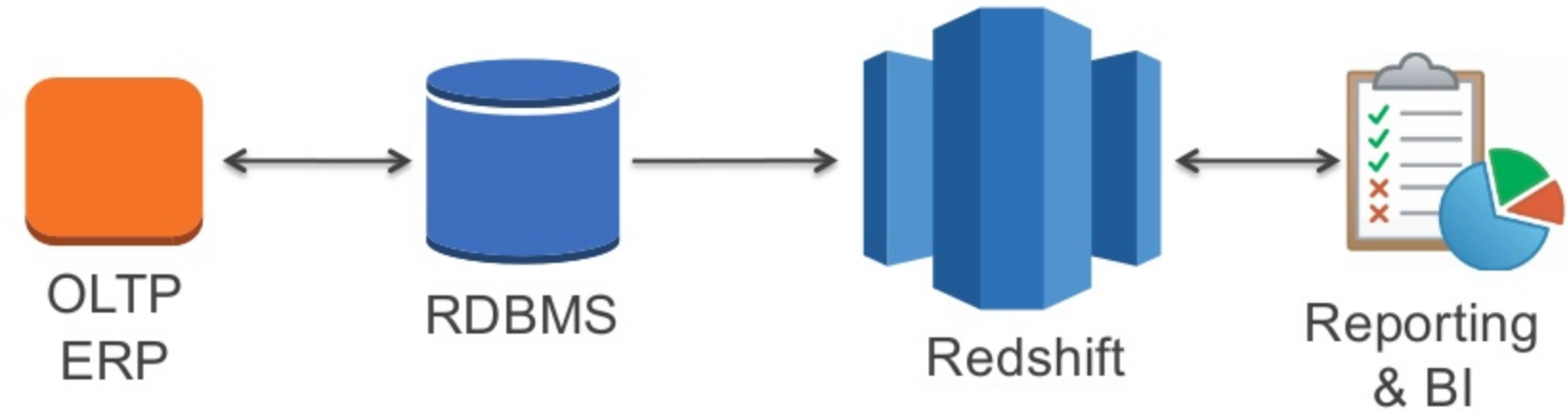
Data Loading Best Practices

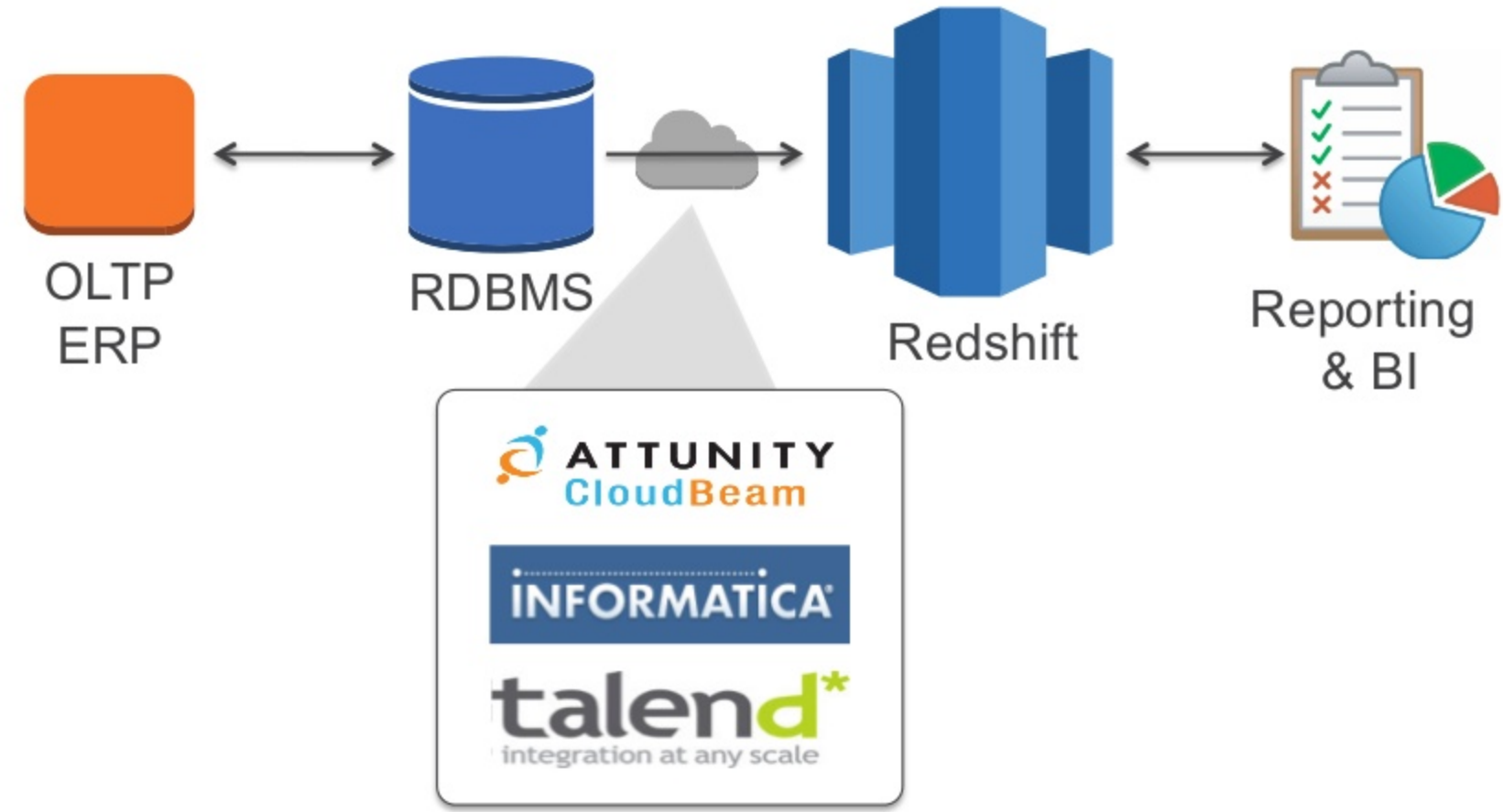
- Use a COPY Command to load data
- Use a single COPY command
- Split your data into multiple files
- Compress your data files with GZIP
- Use multi-row inserts if COPY is not possible
- Bulk insert operations (INSERT INTO…SELECT and CREATE TABLE AS) provide high performance data insertion

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


OLTP
ERP
RDBMS
Redshift
Reporting
& BI
ATTUNITY
CloudBeam
INFORMATICA
talend*
integration at any scale

Amazon Redshift

JDBC/ODBC

ACTUATE
The BIRT Company

birst

JASPERSOFT
the intelligence inside

JReport
JINFONET SOFTWARE

Logi
ANALYTICS

MicroStrategy

PERVASIVE

pentaho

SiSense

tableau
SOFTWARE

awsmarketplace

Amazon Web Ser

Sign in or Create a new account

Your Account | Help | Sell on AWS

Shop All Categories ▾ | Search AWS Marketplace | GO | Yc

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

TIBCO Jaspersoft

Jaspersoft Reporting and Analytics fo...
Jaspersoft
Starting from **$0.48/hr** or from **$3,750/yr** for software

MicroStrategy

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

ATTUNITY CloudBeam

Attunity CloudBeam
Attunity
From $49.95 per month

Business Software > Business Intelligence >

Resources

Amazon Redshift
Fast and Powerful, Petabyte-Scal Warehouse Service

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

Jaspersoft Reporting and Analytics for AWS (Hourly)

Sold by: Jaspersoft | See product video

TIBCO Jaspersoft

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

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%

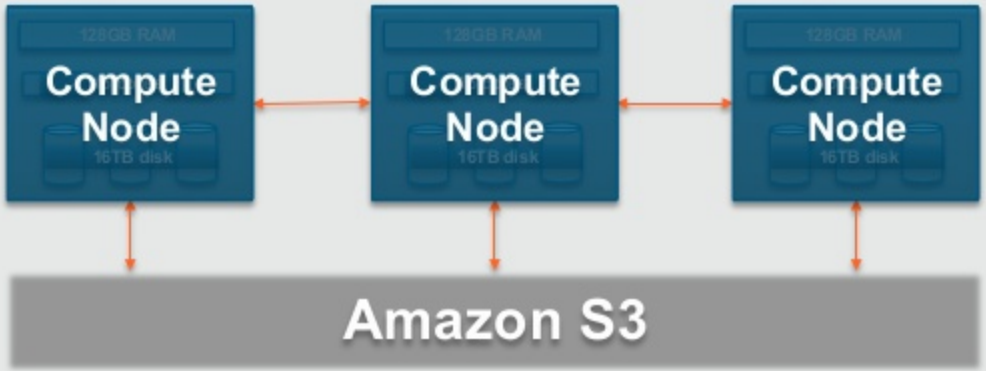
EC2 Instance Type	EC2 Usage	Software	Total
m1.medium	$0.087/hr	$0.48/hr	$0.567/hr
m1.large	$0.175/hr	$1.08/hr	$1.255/hr
m1.xlarge	$0.35/hr	$2.16/hr	$2.51/hr
m2.xlarge	$0.245/hr	$1.78/hr	$2.025/hr
m2.2xlarge	$0.49/hr	$3.69/hr	$4.18/hr
m2.4xlarge	$0.98/hr	$7.38/hr	$8.36/hr
c1.xlarge	$0.52/hr	$2.67/hr	$3.19/hr
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

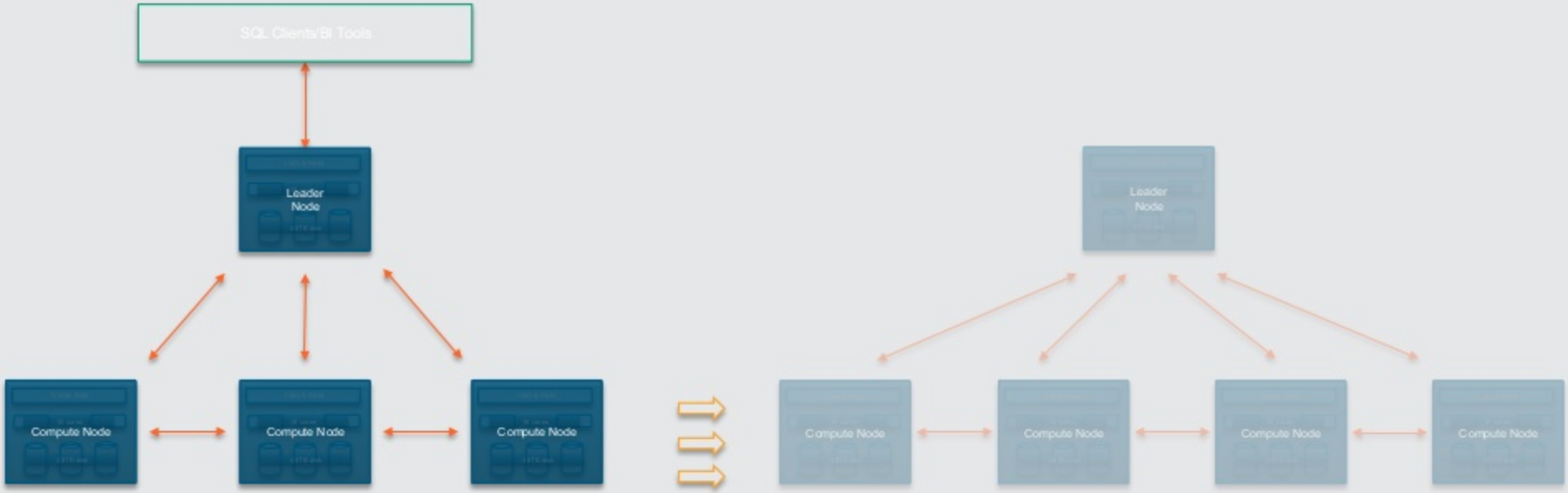
+

AUTOMATED BACKUPS ON AMAZON S3

+

NODE MONITORING

Upgrading and Scaling

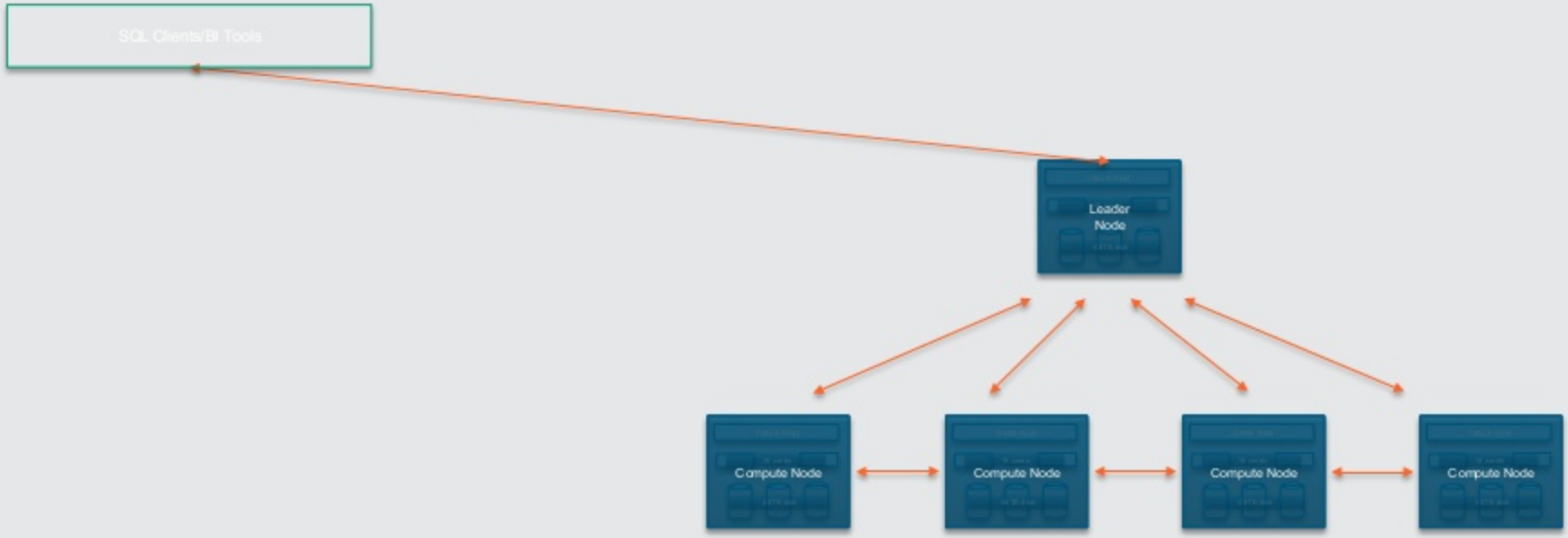


Resize while remaining online

Provision a new cluster in the background

Copy data in parallel from node to node

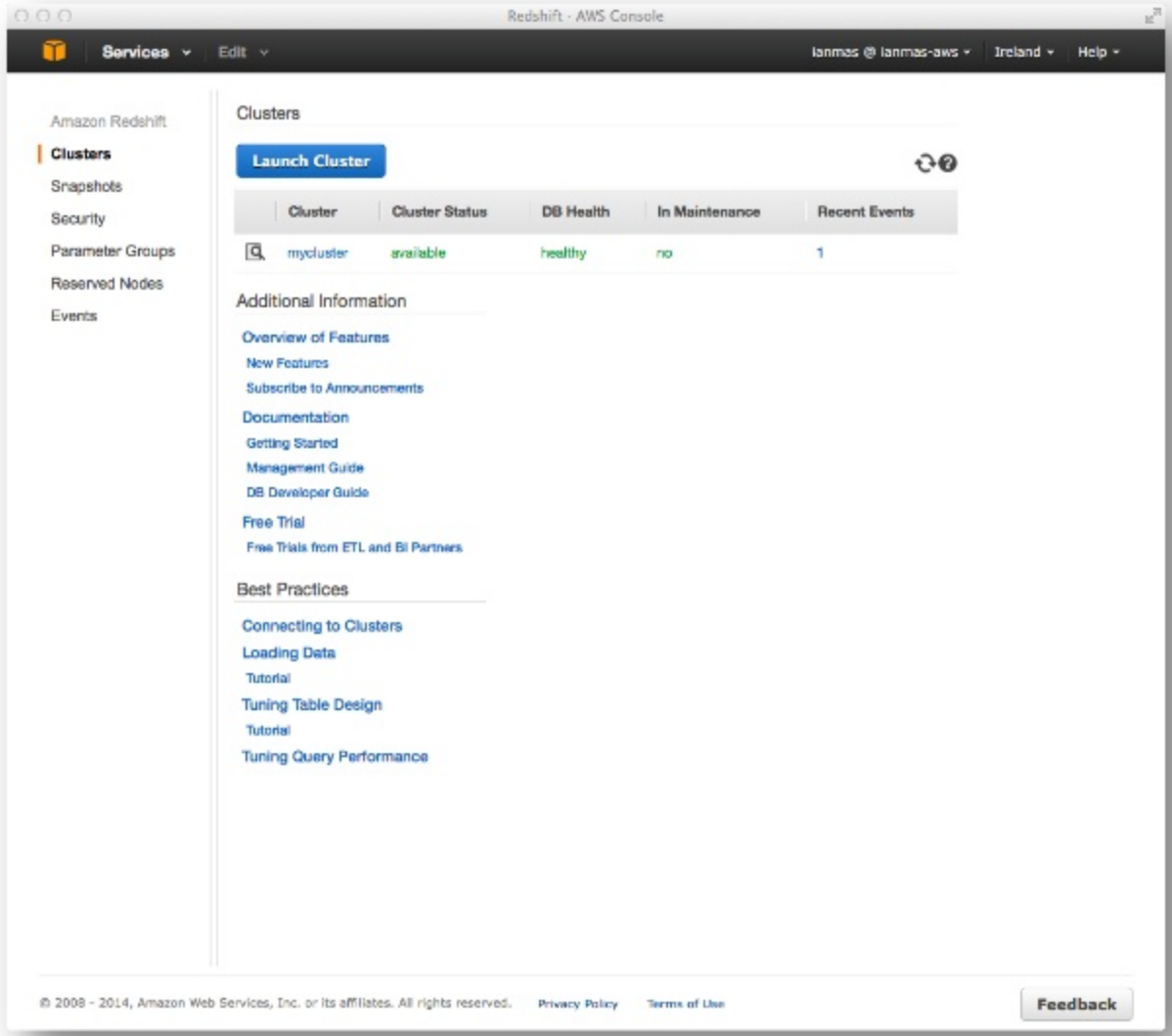
Only charged for source cluster



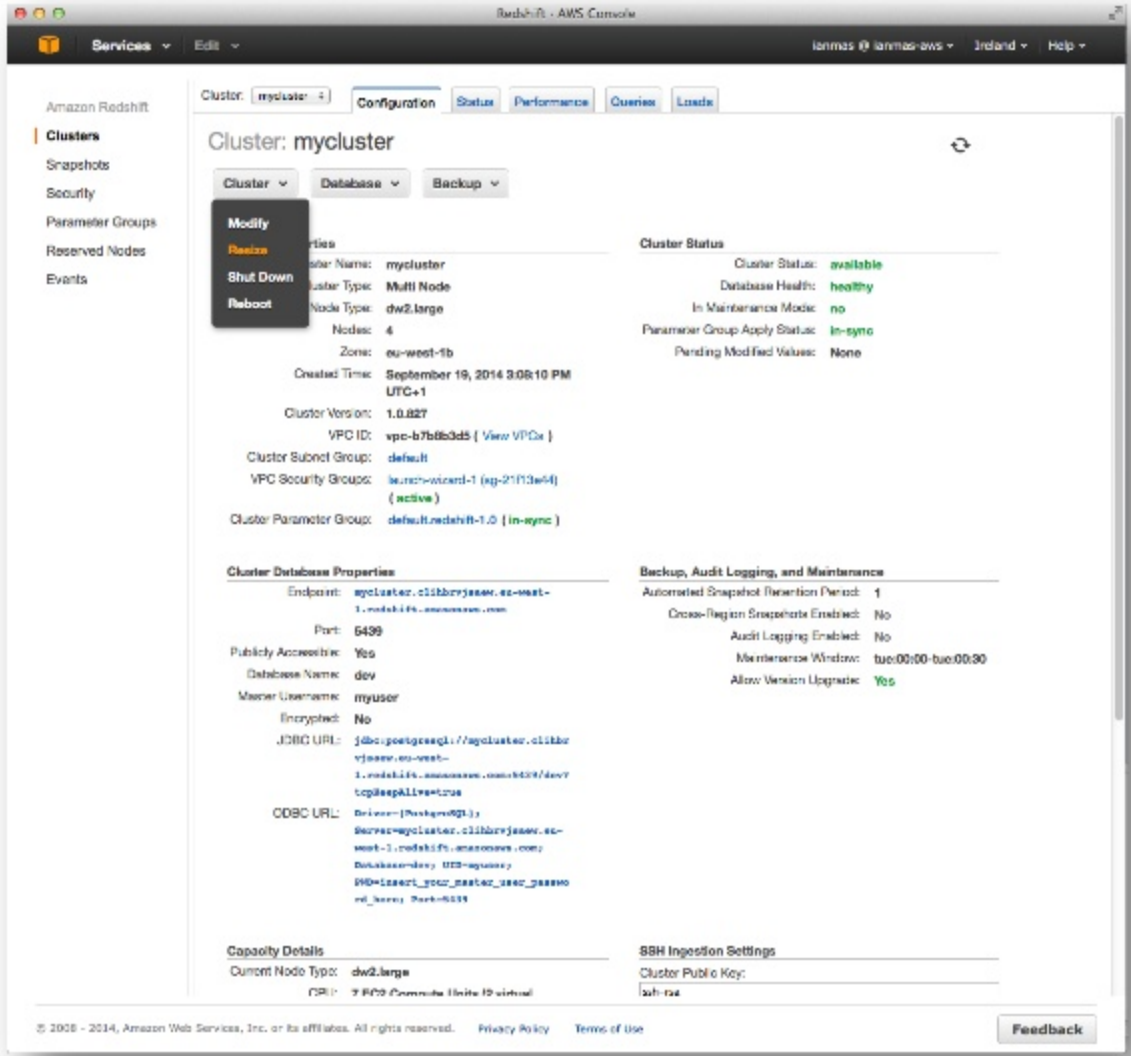
Automatic SQL endpoint switchover via DNS

Decommission the source cluster

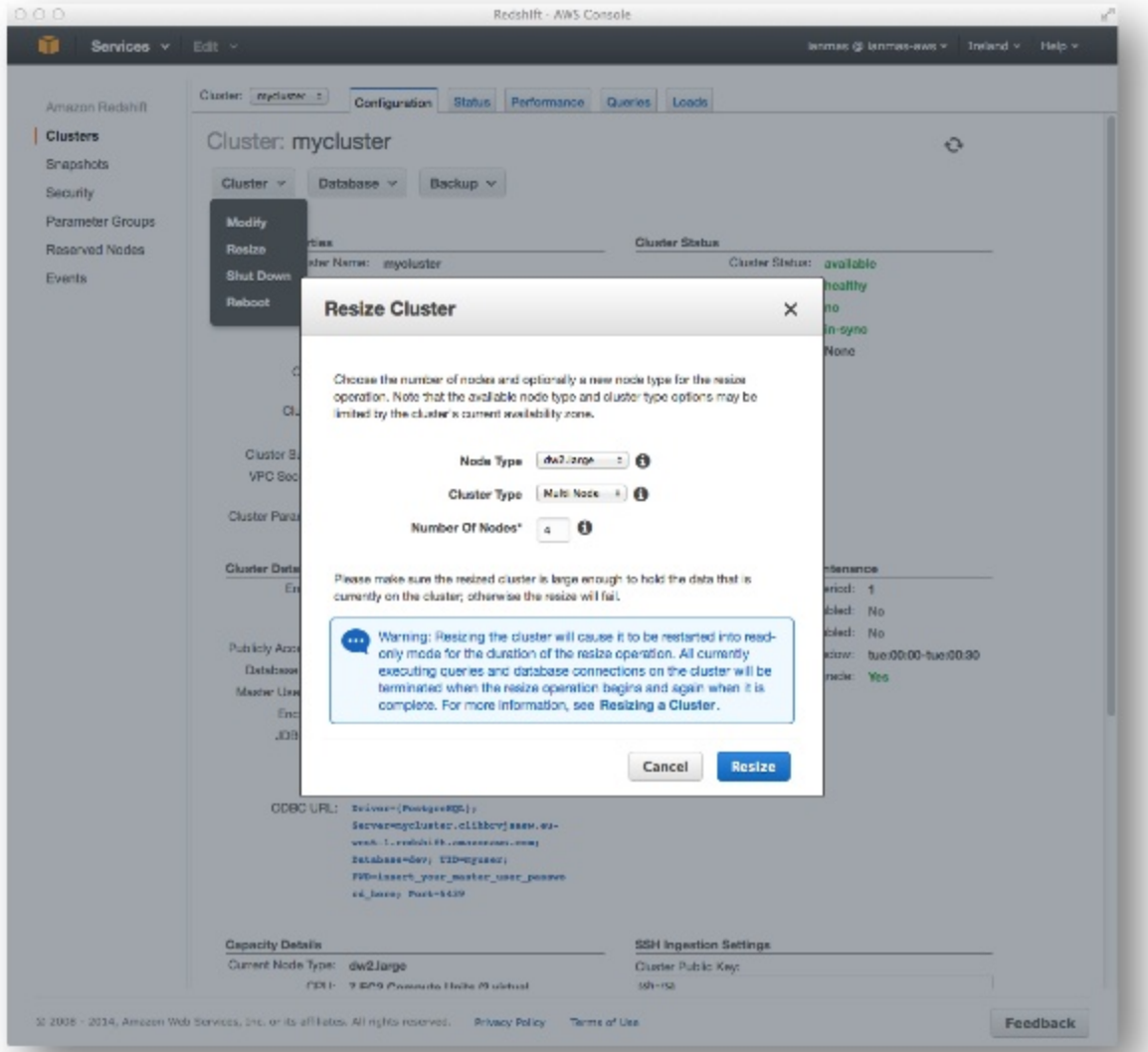
Resizing a Redshift Data Warehouse via the AWS Console



Resizing a Redshift Data Warehouse via the AWS Console

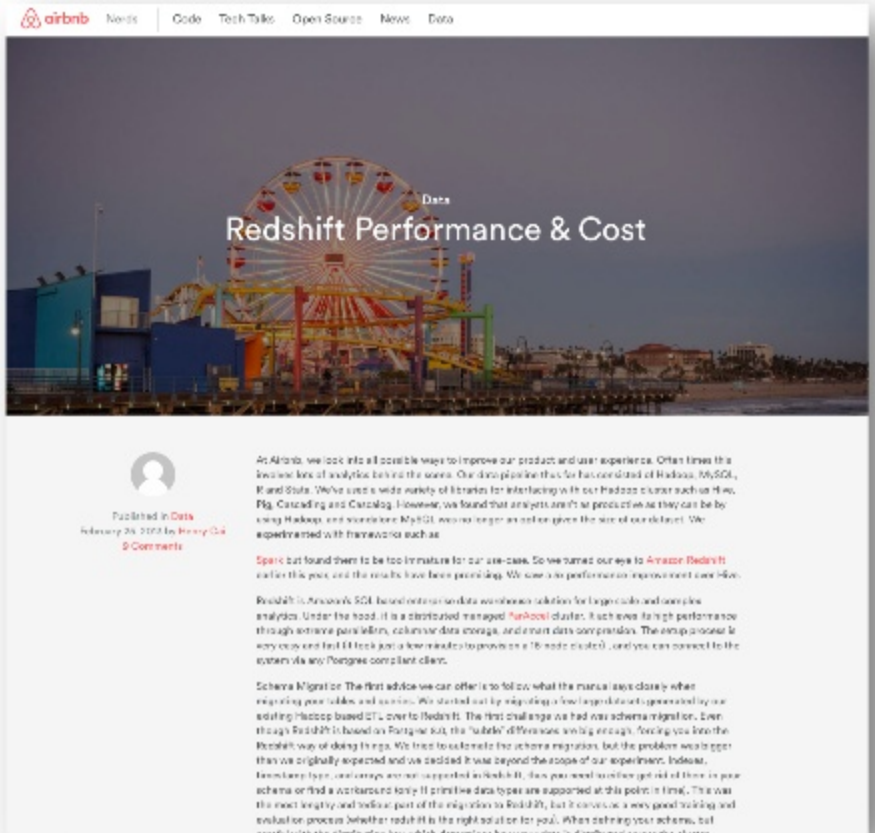


Resizing a Redshift Data Warehouse via the AWS Console

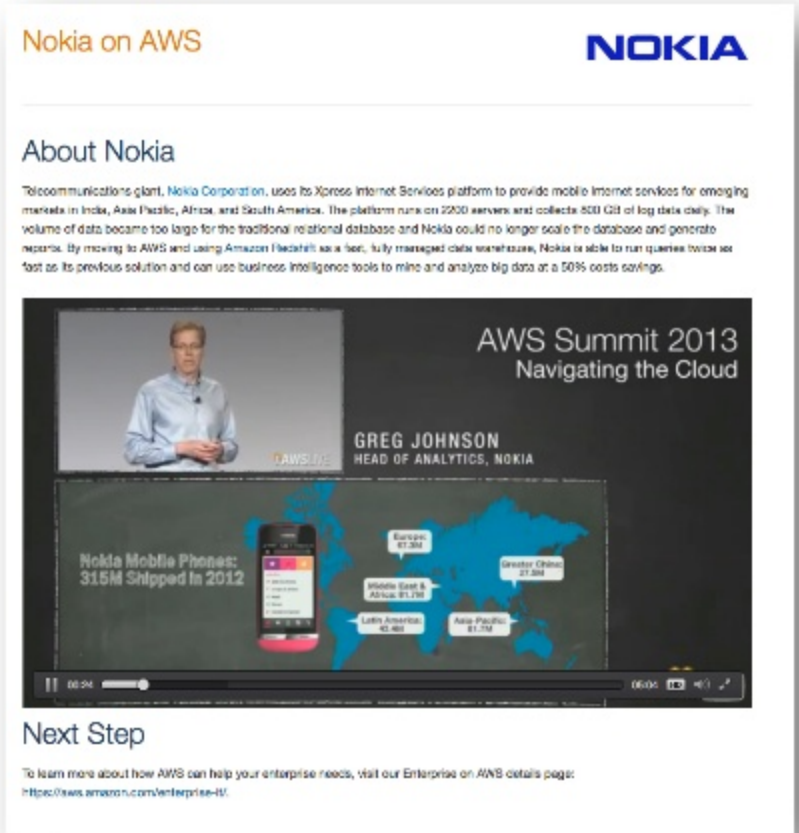


Some other things…

Airbnb



Nokia



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

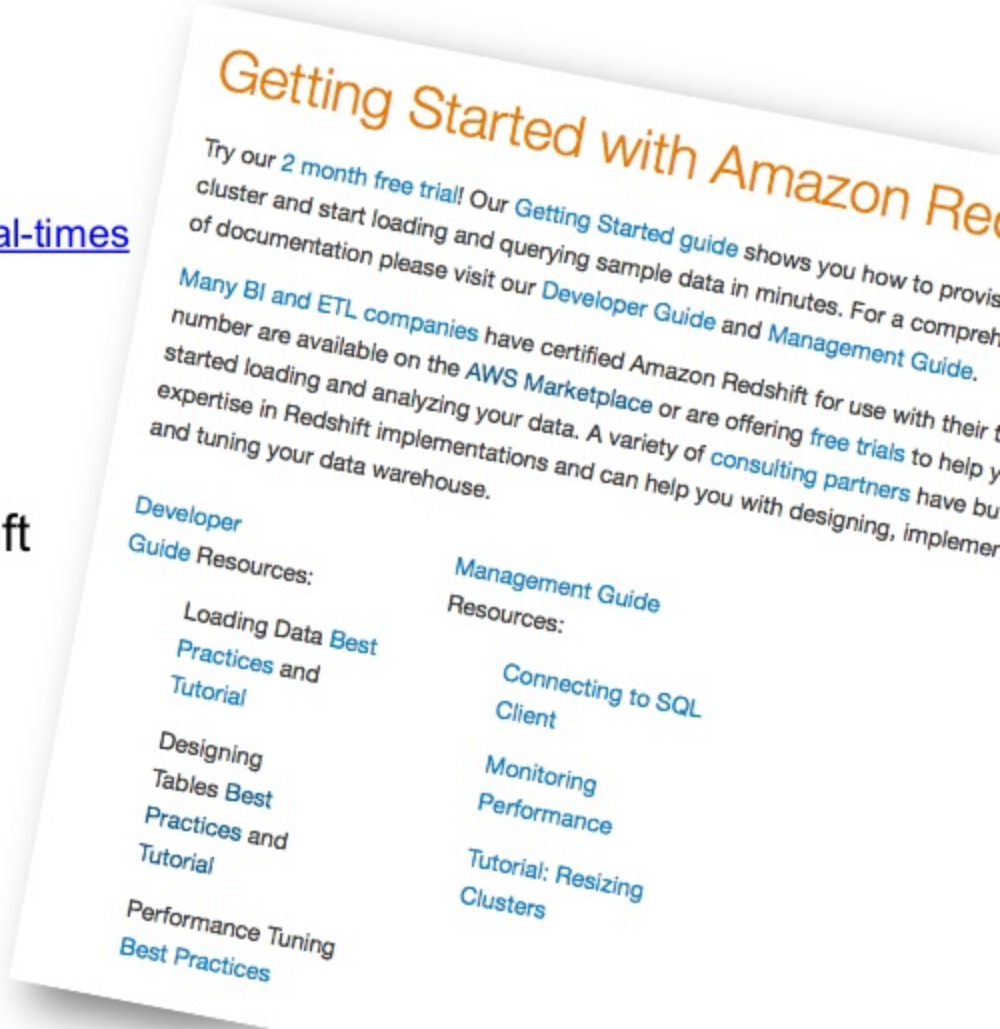
Resources to learn more

Financial Times Case Study
aws.amazon.com/solutions/case-studies/financial-times

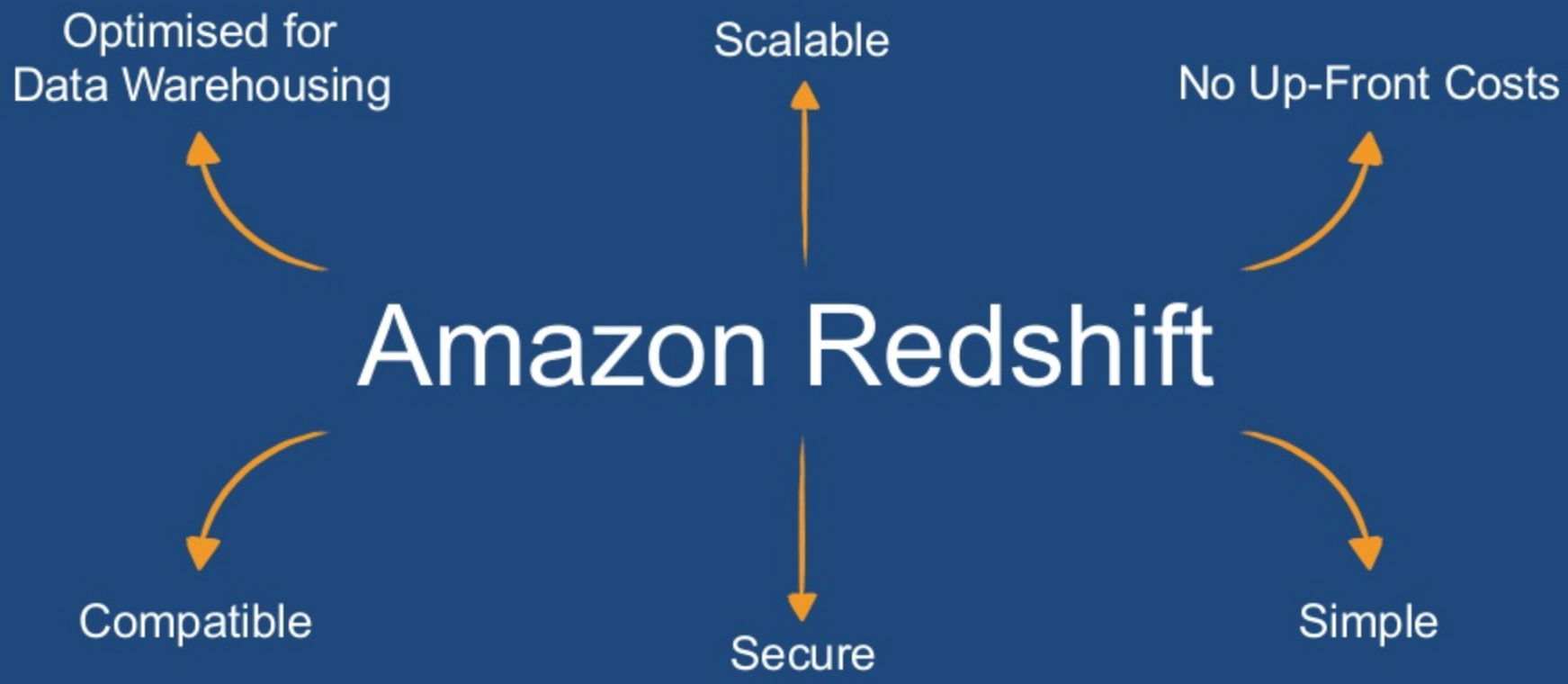
Getting Started with Amazon Redshift
aws.amazon.com/redshift/getting-started/

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

amazon web services

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

amazon web services

Design your application for the AWS Cloud

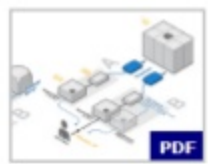
AWS Reference Architectures

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)

More details on the AWS Architecture Center at : aws.amazon.com/architecture

amazon web services