

Cloud Computing & Visualization

Workflows

- Distributed Computation with Spark
- Data Warehousing with Redshift
- Visualization with Tableau

#FIUSCIS

School of Computing & Information Sciences, Florida International University, Miami. 2018

Introduction

- Distributed Computation

- Elastic Map Reduce
- Spark
- Ganglia

- Data Warehousing

- Redshift
- RDS

- Visualization

- Tableau Desktop
- Tableau Prep



Distributed Computation

Spark is a computing platform designed to be **fast** and **general-purpose**.



Data Warehousing

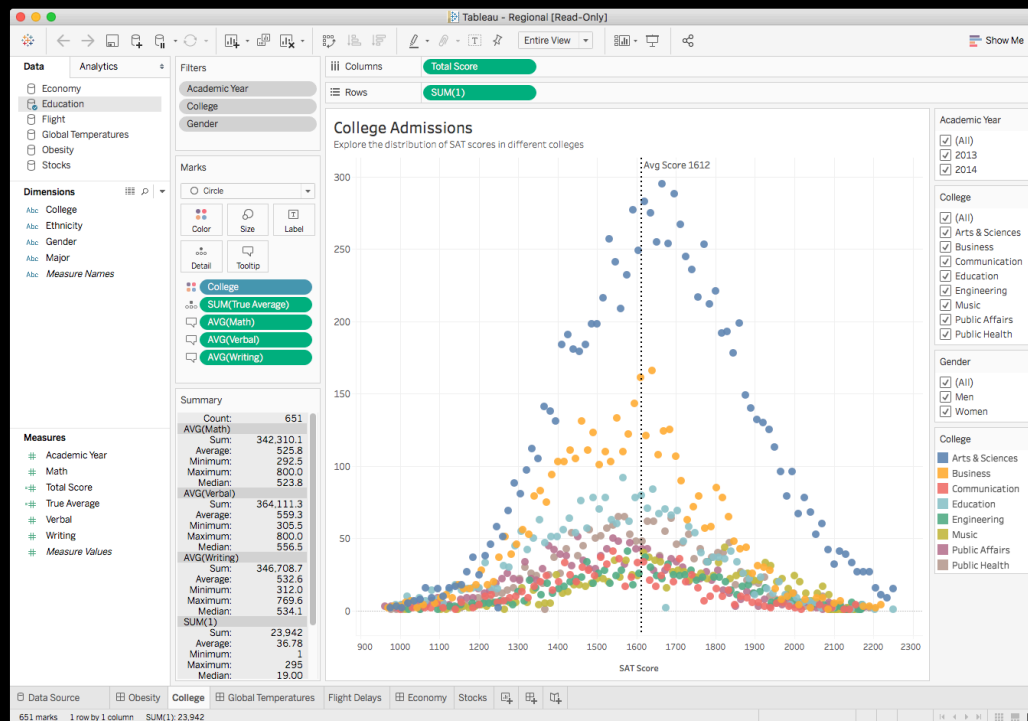
Amazon Redshift is a **fully managed**, petabyte-scale **data warehouse** service in the cloud.



Visualization



Tableau is a Business Intelligence tool for visually analyzing data.



Cloud Computing

- Cloud computing is **shared pools** of configurable computer system resources and higher-level services that can be rapidly **provisioned with minimal management effort**, often over the Internet.
- Third-party cloud providers enable organizations to **focus on core tasks** instead of expending resources on computer infrastructure and maintenance.



aws.amazon.com





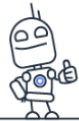
aws.amazon.com



Building Serverless Applications

Build and run your applications and services without thinking about servers

[Learn more](#)



Lightsail
Everything you need to get started on AWS—for a low, predictable price



AWS Fargate
Run containers without managing servers

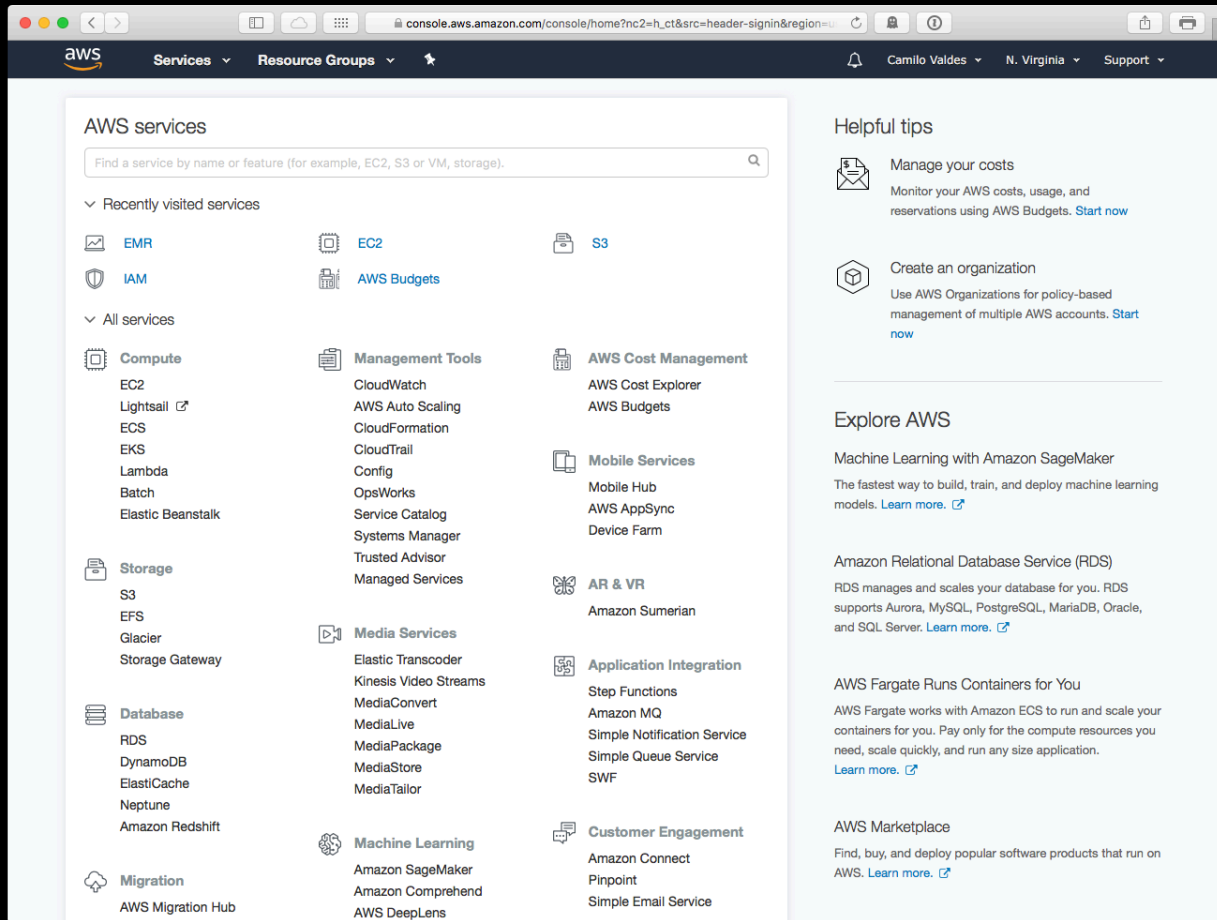


Batch Processing at Any Scale
Run hundreds of thousands of jobs on EC2, fully managed by AWS Batch



90,000+ Databases Migrated to AWS
Save time & cost—migrate to fully managed databases

<https://aws.amazon.com>



- Analytics
 - EMR
- Database
 - Redshift
 - RDS

The image shows a screenshot of the Tableau website homepage. At the top, the Tableau logo is on the left, and navigation links for Products, Solutions, Learning, Community, Support, and About are in the center. On the right, there are links for PRICING, SIGN IN, and a TRY NOW button. A green banner below the navigation bar reads "Upgrade to Tableau 2018.3 for heatmaps, set actions, and more!". The main content area features a large background image of a person working at a laptop. Overlaid on this is a smaller image of a laptop displaying a Tableau dashboard. The dashboard shows a bar chart titled "Revenue by Department" with a color-coded area chart underneath. The chart has a y-axis labeled "Revenue" ranging from 0 to 1,000,000 and an x-axis labeled "Year" ranging from 2010 to 2015. The bars are blue, and the area below them is filled with a gradient of colors (orange, yellow, green). The Tableau interface includes a sidebar with various data fields and a top navigation bar.

www.tableau.com

Products Solutions Learning Community Support About

PRICING SIGN IN TRY NOW

Upgrade to Tableau 2018.3 for heatmaps, set actions, and more!

Changing the way you think about data

THE TABLEAU PLATFORM SEE IT IN ACTION

Department Analysis

Revenue by Department

Revenue

Year

Department

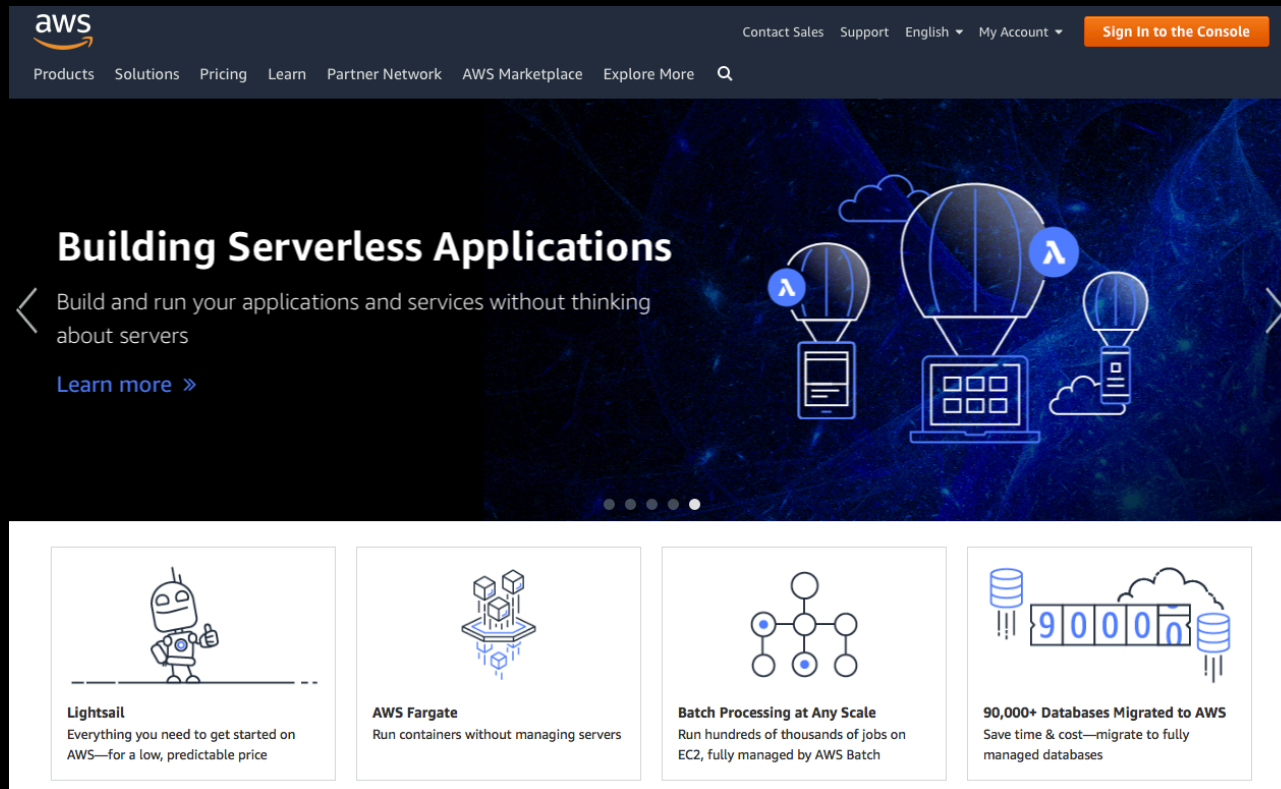
Marketing Sales Operations

MacBook Pro

Harness the power of your data. Unleash the potential of your people.

<https://www.tableau.com>

Distributed Computation with Spark



The screenshot shows the AWS website's navigation bar with the AWS logo, links for Products, Solutions, Pricing, Learn, Partner Network, AWS Marketplace, and Explore More, along with a search icon. On the right, there are links for Contact Sales, Support, English, My Account, and a Sign In to the Console button. The main content area features a dark blue background with a glowing network pattern. A central heading reads "Building Serverless Applications" with a subtext "Build and run your applications and services without thinking about servers" and a "Learn more" link. To the right is an illustration of three hot air balloons, each with a Lambda icon, floating above a laptop and a smartphone. Below this is a grid of four service cards: Lightsail (robot icon), AWS Fargate (server rack icon), Batch Processing at Any Scale (network icon), and 90,000+ Databases Migrated to AWS (database icon).

aws Contact Sales Support English My Account [Sign In to the Console](#)

Products Solutions Pricing Learn Partner Network AWS Marketplace Explore More

Building Serverless Applications

Build and run your applications and services without thinking about servers


[Learn more](#) »

Lightsail
Everything you need to get started on AWS—for a low, predictable price

AWS Fargate
Run containers without managing servers

Batch Processing at Any Scale
Run hundreds of thousands of jobs on EC2, fully managed by AWS Batch

90,000+ Databases Migrated to AWS
Save time & cost—migrate to fully managed databases



Apache Spark



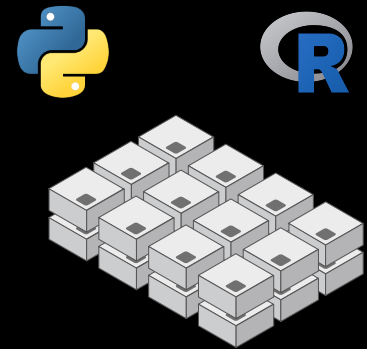
- Spark is a **Big Data Processing Engine** — a Fast, General-Purpose, Cluster-computing Platform.
- Handles the **Scheduling, Distribution, and Monitoring** of applications spanning many worker machines.
- Has a **Rich API** to **distribute data across the cluster**, and process it in parallel.
- Supports a variety of workloads such as **Machine Learning** (MLlib), **Streaming**, interactive queries, graph programming and SQL.
- Execution Frameworks have language support for **Python, R, Java, and Scala**.



Spark — Unified Stack



- The Spark project contains multiple high-level specialized components (MLlib, Streaming, etc.).
- Spark's main programming abstraction are **Resilient Distributed Datasets (RDDs)**, a data structure distributed across nodes that can be worked on in parallel.
- Spark's multiple components operate on RDDs, which allows for close interoperability and tight integration.
- Applications that use **multiple processing models** can be written without high maintenance and development costs.



Spark — Main Benefits

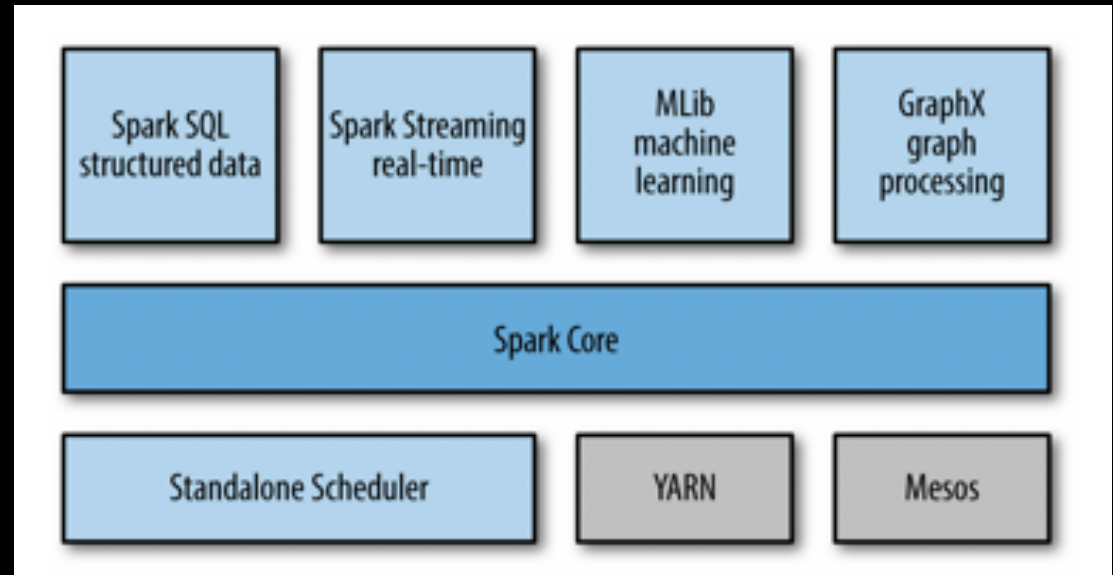


Solve problems faster, and on a much larger scale

- **Ease of Use** — Rich, high level APIs
- **Speed** — Fast parallel execution
- **General Engine** — Combine processing models
- **Open Source** — Freely Available
- Makes developing General Purpose Distributed programs easier, less painful.
- Reduces the **management burden** of maintaining separate tools.
- Allows the **close Interoperability** of high-level components

Spark Core

- Spark Core contains the basic functionality of Spark, including components for **task scheduling, memory management, fault recovery**, interacting with storage systems, and more.



- Spark Core is also home to the API that defines **resilient distributed datasets (RDDs)**, which are Spark's main programming abstraction.
- RDDs represent a collection of **items distributed across many compute nodes** that can be manipulated in parallel.

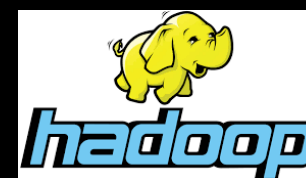
Spark — Data Processing



- Spark provides a simple way to **parallelize applications across clusters**, and hides the complexity of distributed systems programming, network communication, and fault tolerance.
- The system gives control to **monitor, inspect, and tune applications** while allowing implementation of common tasks quickly.
- The **modular nature of the API** (based on passing **distributed collections of objects**) makes it easy to factor work into reusable libraries and test it locally.

Storage Layers for Spark

- Spark can create resilient distributed datasets, RDDs, from any file stored in the Hadoop distributed filesystem (**HDFS**).
- Spark also support other storage systems supported by the Hadoop APIs (including your **local filesystem**, **Amazon S3**, Cassandra, Hive, HBase, etc.).
- It's important to remember that **Spark does not require Hadoop**.
- It simply has support for storage systems implementing the Hadoop APIs.



Spark REPL



- Spark can be used from **Python**, R, Java, or Scala.
- **Spark itself is written in Scala**, and runs on the Java Virtual Machine (JVM).
- To run Spark on either your laptop or a cluster, all you need is an installation of Java 6 or newer.
- If you wish to use the Python API you will also need a Python interpreter (version 2.6 or newer).
- You don't need to have Hadoop.
- Spark comes with **interactive shells** that enable ad hoc data analysis.
- Spark's shells will feel familiar if you have used other shells such as those in R, Python, and Scala,



- Python version of the Spark Shell.

```
Last login: Sat Oct 27 16:23:14 on ttys003
Trajan.>_ pyspark
Python 2.7.14 (default, Mar 10 2018, 00:01:04)
[GCC 4.2.1 Compatible Apple LLVM 9.0.0 (clang-900.0.39.2)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
18/10/30 18:07:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
18/10/30 18:07:48 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to

  ____
 /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\  /--\
/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\/_--\
                                     version 2.2.0

Using Python version 2.7.14 (default, Mar 10 2018 00:01:04)
SparkSession available as 'spark'.
>>> |
```



RDDs



- An RDD is simply a distributed **collection of elements**.
- In Spark all work is expressed as either **creating new RDDs**, **transforming existing RDDs**, or **calling operations** on RDDs to compute a result.
- Spark automatically distributes the data contained in RDDs across your cluster and parallelizes the operations you perform on them.
- An RDD in Spark is simply an **immutable distributed collection** of objects.
- Each RDD is split into **multiple partitions**, which may be computed on different nodes of the cluster.
- RDDs can contain **any type of** Python, Java, or Scala objects, including user-defined classes.
- Once created, RDDs offer two types of operations: **transformations** and **actions**.

RDDs

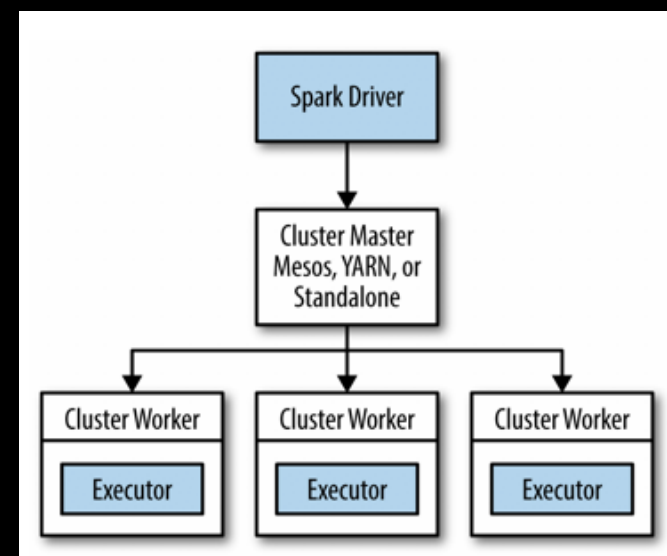


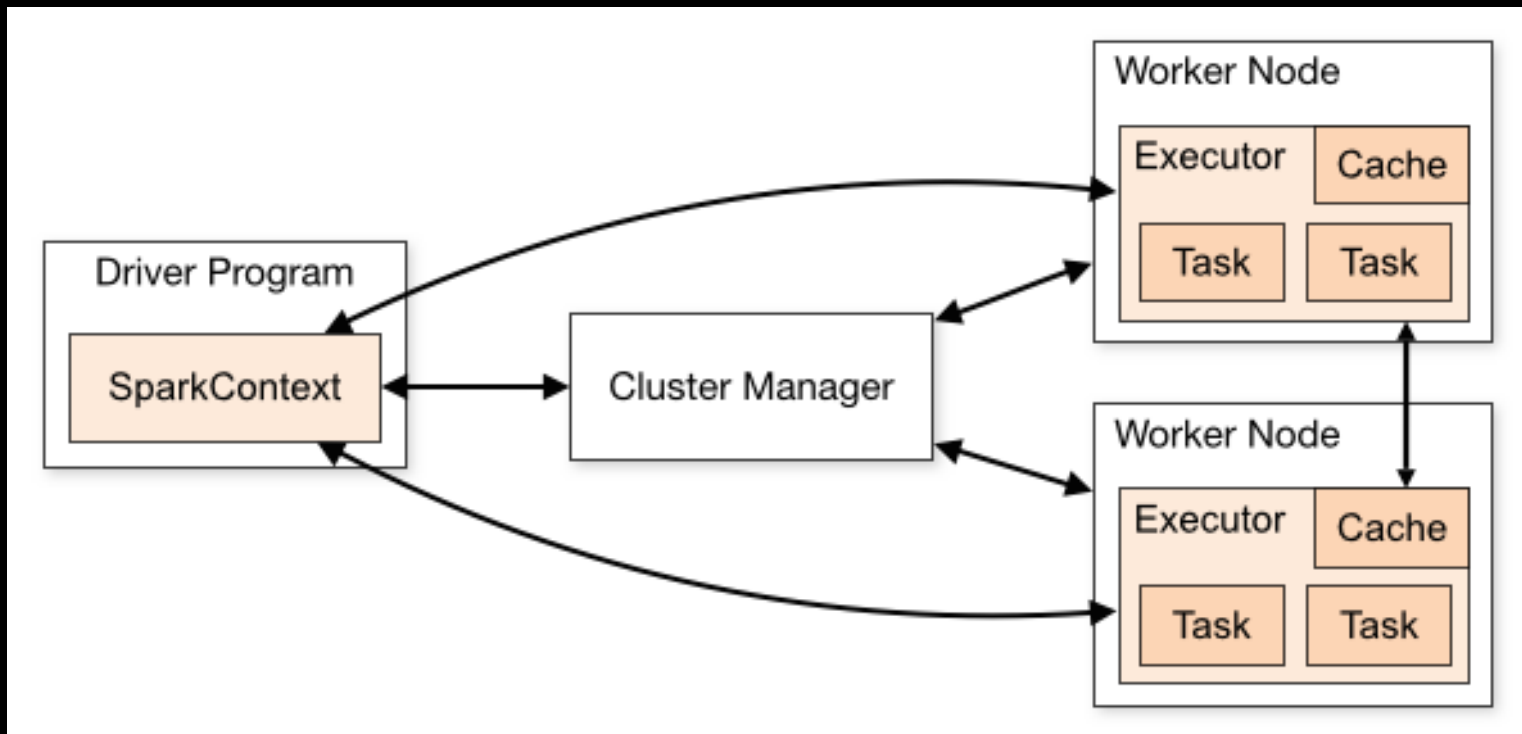
- **Transformations** construct a new RDD from a previous one.
- **Actions** compute a result based on an RDD, and either return it to the driver program or save it to an external storage system.
- Although you can define new RDDs any time, Spark computes them only in a **lazy fashion** — that is, the first time they are used in an action.
- Spark provides two ways to create RDDs
 - **loading an external dataset.**
 - **Parallelizing a collection** in your driver program.

Spark Cluster



- Every Spark application consists of a **driver** program that launches various parallel operations on a cluster.
- The driver program contains your application's **main function** and defines distributed datasets on the cluster, then applies operations to them.
- The driver communicates with a potentially large number of **distributed workers** called executors.
- A driver and its executors are together termed a **Spark application**.







Solve problems faster, and on a Much Larger Scale

Data Warehousing with Redshift

The screenshot shows the AWS website's navigation bar with the AWS logo, links for Products, Solutions, Pricing, Learn, Partner Network, AWS Marketplace, and Explore More, along with a search icon. On the right, there are links for Contact Sales, Support, English, My Account, and a Sign In to the Console button. The main content area features a hero section titled 'Building Serverless Applications' with a sub-headline 'Build and run your applications and services without thinking about servers' and a 'Learn more' link. To the right of the text is an illustration of three hot air balloons, each with an AWS Lambda icon, floating over a laptop and a smartphone. Below the hero section is a grid of four service highlights: Lightsail (robot icon), AWS Fargate (server rack icon), Batch Processing at Any Scale (network icon), and 90,000+ Databases Migrated to AWS (database icon).

aws

Contact Sales Support English My Account Sign In to the Console

Products Solutions Pricing Learn Partner Network AWS Marketplace Explore More

Building Serverless Applications

Build and run your applications and services without thinking about servers

[Learn more »](#)

Lightsail
Everything you need to get started on AWS—for a low, predictable price

AWS Fargate
Run containers without managing servers

Batch Processing at Any Scale
Run hundreds of thousands of jobs on EC2, fully managed by AWS Batch

90,000+ Databases Migrated to AWS
Save time & cost—migrate to fully managed databases



Data Warehouse



- A system used for **reporting** and **data analysis**.
- **Central repositories** of integrated data from one or more disparate sources.
- A data warehouse can store current and historical data in **a single place**.
- *"Subject-oriented, integrated, time-variant and non-volatile collection of data in support of a decision making process"*.
- The data stored in the warehouse is uploaded from different operational systems — systems used to pre-process the data in some way.
- Data sources can also come from clusters such as **Spark** and Hadoop.

Redshift



- Amazon Redshift is a **fully managed**, petabyte-scale **data warehouse** service in the cloud.
- An Amazon Redshift data warehouse is a collection of computing resources called **nodes**, which are organized into a group called a **cluster**.
- Each cluster runs an Amazon Redshift engine and contains one or more **databases**.
- Redshift differs from Amazon's other hosted database offering, **Amazon RDS**, in its **ability to handle analytics workloads on big data** datasets.
- Redshift allows you to analyze data using **Business Intelligence (BI)** tools such as **Spotfire** and **Tableau**.



Redshift



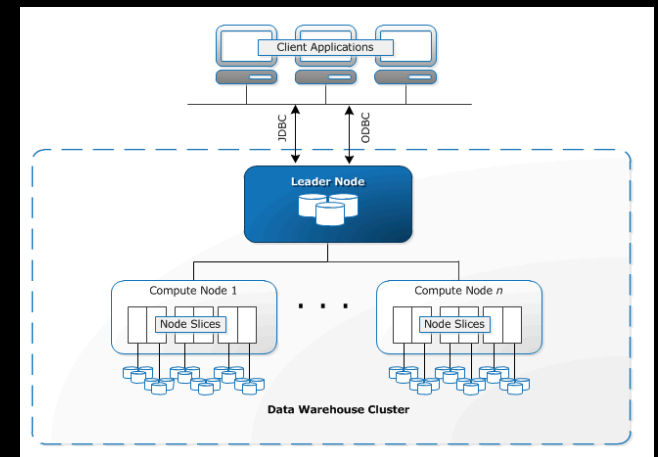
- Redshift is based on **PostgreSQL**.
- An Amazon Redshift data warehouse is an enterprise-class, **relational database query and management system**.
- Amazon Redshift is built around industry-standard SQL, with added functionality to **manage very large datasets** and support **high-performance analysis** and **reporting** of that data.
- Amazon Redshift achieves efficient storage and optimum query performance through a combination of **massively parallel processing**, **columnar data storage**, and very efficient, targeted data compression encoding schemes.

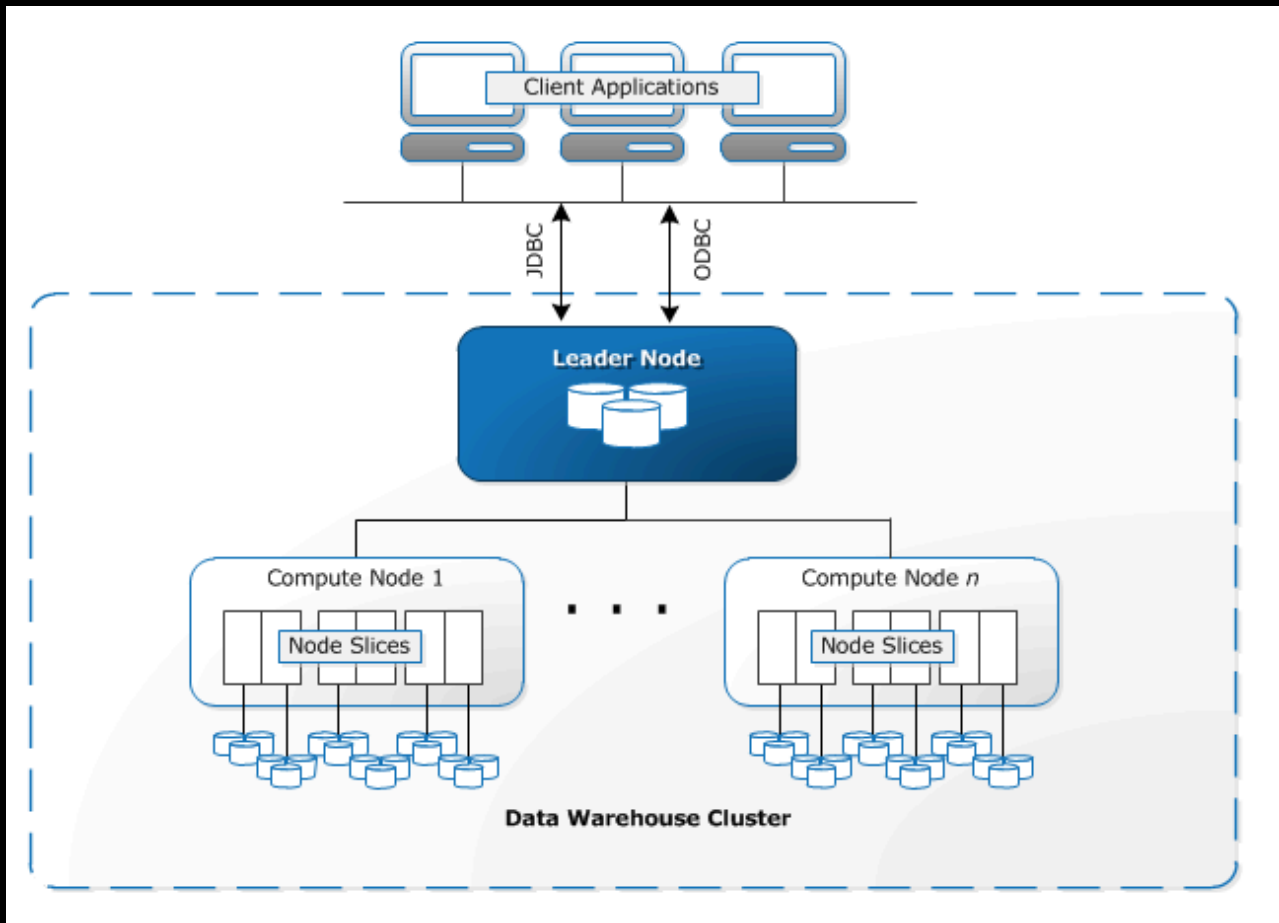


Redshift



- Redshift is based on industry-standard **PostgreSQL**, so most existing SQL client applications will work with only minimal changes.
- A cluster is composed of one or more **compute nodes**.
- If a cluster is provisioned with two or more compute nodes, an additional **leader node** coordinates the compute nodes and handles external communication.
- Your client application interacts directly only with the leader node. Compute nodes are transparent to external applications.



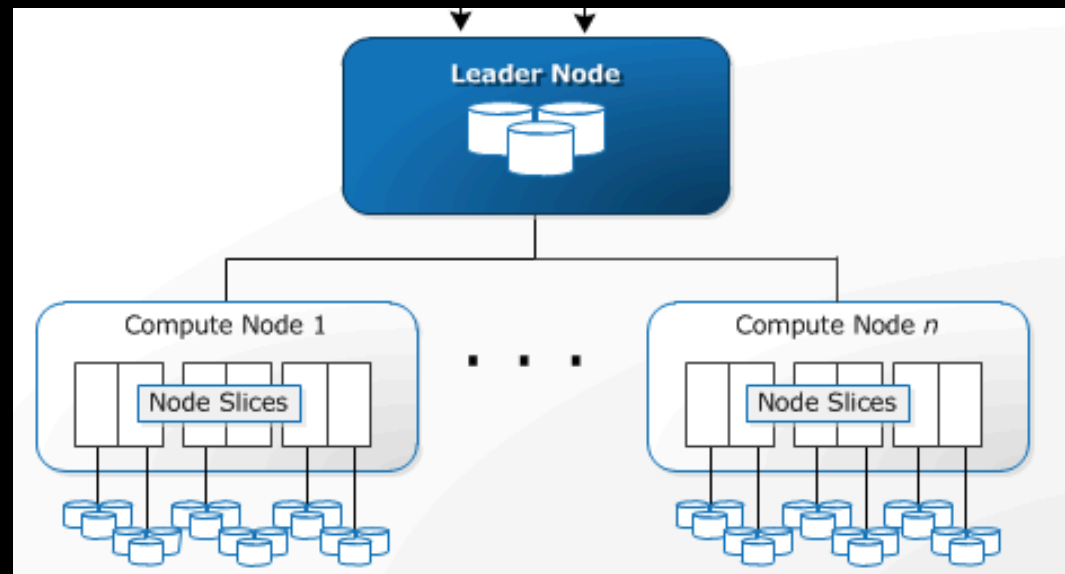


Redshift Cluster



Composed of three (3) main elements

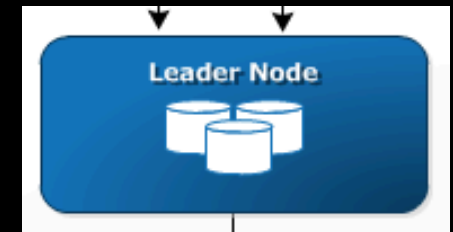
- Leader Node
- Compute Node
- Node Slices



Leader Node



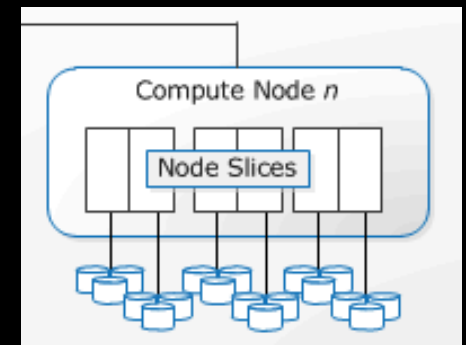
- The leader node **manages communications** with client programs and all communication with compute nodes.
- It **parses and develops execution plans** to carry out database operations, in particular, the **series of steps necessary to obtain results for complex queries**.
- Based on the execution plan, the leader node **compiles code**, distributes the compiled code to the compute nodes, and **assigns a portion of the data to each compute node**.
- The leader node distributes SQL statements to the compute nodes only when **a query references tables** that are stored on the compute nodes.
- All other queries run exclusively on the leader node.



Compute Nodes



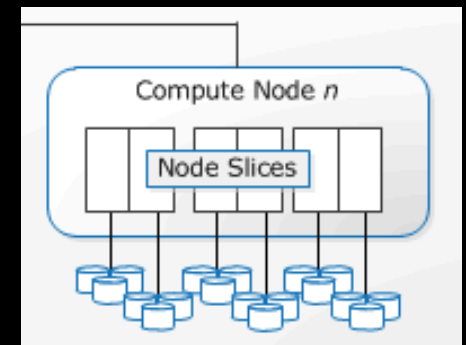
- The leader node compiles code for individual elements of the execution plan and assigns the code to individual compute nodes.
- The compute nodes **execute the compiled code** and **send intermediate results** back to the leader node for final aggregation.
- Each compute node has its own dedicated **CPU, memory,** and **attached disk storage**, which are determined by the node type.
- As your workload grows, you can increase the compute capacity and storage capacity of a cluster by increasing the number of nodes, upgrading the node type, or both.
- You can start with a single **160 GB node** and scale up to **multiple 16 TB nodes** to support a petabyte of data or more



Node Slices



- A compute node is **partitioned into slices**.
- Each slice is **allocated a portion of the node's memory and disk space**, where it **processes a portion of the workload assigned to the node**.
- The leader node manages distributing data to the slices and apportions the workload for any queries or other database operations to the slices.
- The **slices then work in parallel** to complete the operation.
- The number of slices per node is determined by the node size of the cluster.



Redshift Databases



- **User data** is stored on the compute nodes. SQL clients communicate with the leader node, which in turn coordinates query execution with the compute nodes.
- Amazon Redshift is a **relational database management system (RDBMS)**, so it is compatible with other RDBMS applications.
- Although it provides the same functionality as a typical RDBMS, Amazon Redshift is **optimized for high-performance analysis and reporting of very large datasets**.
- Amazon Redshift is based on **PostgreSQL 8.0.2**.
- Redshift and PostgreSQL have a number of very important differences that you need to take into account as you design and develop your data warehouse applications.



console.aws.amazon.com/redshift/home?region=us-east-1#

aws Services Resource Groups Camilo Valdes N. Virginia Support

Redshift dashboard

- Clusters
- Query editor **Now**
- Saved queries
- Snapshots
- Security
- Parameter groups
- Workload management
- Reserved nodes
- Advisor ^{Beta}
- Events
- Connect client
- What's new

Launch cluster

Amazon Redshift is a powerful, fully managed cloud data warehouse service. Redshift Spectrum extends the power of Redshift to query unstructured data in S3 – without loading your data into Redshift. With a few clicks in the AWS Management Console, you can launch a Redshift cluster and get started analyzing your data.

[Quick launch cluster](#) [Launch cluster](#)

Note: Your cluster will launch in the US East (N. Virginia) region

Query Editor **NEW**

Write, run, and save queries directly from the console

You must enable the IAM Policy: [AmazonRedshiftQueryEditor](#) for your account to run queries on eligible clusters. Please attach the IAM policy on the [IAM console](#). See [AWS Managed Policies for Amazon Redshift](#) for more information.

[Launch Query Editor](#)

Note: To use the query editor, ensure the below cluster configuration:

- Allowed node types: **dc1.8xlarge, dc2.large, dc2.8xlarge, or ds2.8xlarge**
- Enhanced VPC routing is not enabled

You need to have at least one supported cluster with an "available" status to query from the console.

Resources

You are using the following Amazon Redshift resources in the US East (N. Virginia) region (used):

- Clusters (0)**
 - Increase cluster limit
- Security**
 - Subnet groups (0)
- Parameter groups (0)**
 - Total Reservations (0)
 - Events (0)
 - Event subscriptions (0)
- Snapshots (0)**
 - Manual (0)
 - Automated (0)

Service health

Current Status	Details
Amazon Redshift (N. Virginia)	Service is operating normally

[View complete service health details](#)

Getting Started

- [Getting Started with Amazon Redshift](#)
- [Overview and features](#)
- [Free Trial](#)
- [Evaluation and POC support](#)
- [Documentation](#)
- [Query your S3 data lake with Redshift Spectrum](#)
- [Pricing and Specs](#)
- [Purchase a Reserved Node](#)

AWS Marketplace

- Matillion ETL for Amazon Redshift**
By Matillion
Rating ★★★★★
Starting from \$1.37/hr or from \$9,950/yr (17% savings) for software + AWS usage fees
[View all Business Intelligence](#)
- Looker Analytics Platform - 10 Users, Multi-node Redshift (plus RDS)**
By Looker (aka Looker Data Sciences)
Rating ★★★★★
\$3,000.00/mo + \$4.17/hr for software + AWS usage fees
[View all Business Intelligence](#)
- Tableau Server (10 users)**
By Tableau
Rating ★★★★★
\$0.63/hr or \$5,500/yr for software + Charges for EC2 with Windows + AWS usage fees
[View all Business Intelligence](#)

[Find more software on AWS Marketplace](#)

Feedback English (US) © 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use



console.aws.amazon.com/redshift/home?region=us-east-1#launch-cluster:

aws Services Resource Groups Camilo Valdes N. Virginia Support

Launch your Amazon Redshift cluster - Advanced settings | [Switch to quick launch](#)

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.

The ds2 and dc2 node types replace the ds1 and dc1 node types, respectively. The newer ds2 and dc2 node types provide higher performance than ds1 and dc1 at no extra cost. [Learn more.](#)

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 15.25 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

Feedback English (US) © 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

console.aws.amazon.com/redshift/home?region=us-east-1#launch-cluster:

aws Services Resource Groups Camilo Valdes N. Virginia Support

Redshift dashboard Clusters Query editor **New** Saved queries Snapshots Security Parameter groups Workload management Reserved nodes Advisor ^{Beta} Events Connect client What's new

Launch your Amazon Redshift cluster - Advanced settings | [Switch to quick launch](#)

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.

Database encryption None KMS HSM [Learn more about database encryption](#)

Configure networking options:

Choose a VPC The identifier of the VPC in which you want to create your cluster

Cluster subnet group Selected Cluster Subnet Group may limit the choice of Availability Zones

Publicly accessible Yes No 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 Yes 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.

Enhanced VPC Routing Yes No Select Yes if you want to enable Enhanced VPC Routing. [Learn more](#)

Availability zone The EC2 Availability Zone that the cluster will be created in.

Associate your cluster with one or more security groups.

VPC security groups List of VPC security groups to associate with this cluster. [Refresh](#)






Optionally, create a basic alarm for this cluster.

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

Optionally, select your maintenance track for this cluster.

Feedback English (US) © 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use



console.aws.amazon.com/redshift/home?region=us-east-1#launch-cluster:  Services  Resource Groups  Camilo Valdes  N. Virginia  Support

Redshift dashboard
 Clusters
 Query editor New
 Saved queries
 Snapshots
 Security
 Parameter groups
 Workload management
 Reserved nodes
 Advisor ^{Beta}
 Events
 Connect client
 What's new

Launch your Amazon Redshift cluster - Advanced settings | [Switch to quick launch](#)

CLUSTER DETAILS NODE CONFIGURATION ADDITIONAL CONFIGURATION REVIEW

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

Cluster properties

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.

Cluster identifier: dw-test-01

Node type: dc2.large

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

Availability zone: No preference

Database configuration

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.

Database name: test

Database port: 5439

Master user name: cvalde03

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

Security, access, and encryption

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.

Virtual private cloud: vpc-aa4963cd

Cluster subnet group:

Publicly accessible: Yes

Elastic IP: Not used

VPC security groups: ElasticMapReduce-master (sg-fed8884)


Enhanced VPC Routing: No

Encrypt database: No



CloudWatch alarms

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.

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 **\$0.25**, or **\$0.25** /node. If you have purchased reserved nodes in this region for this

 Feedback  English (US) © 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#)



console.aws.amazon.com/redshift/home?region=us-east-1#launch-cluster:

Services Resource Groups Camilo Valdes N. Virginia Support

Redshift dashboard Clusters Query editor **New** Saved queries Snapshots Security Parameter groups Workload management Reserved nodes Advisor ^{Beta} Events Connect client What's new

availability zone in which it will be located. Configuration values used by the database.

Cluster identifier: dw-test-01 **Database name:** test

Node type: dc2.large **Database port:** 5439

Number of compute nodes: 1 (leader and compute run on a single node) **Master user name:** cvalde03

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-aa4963cd **CloudWatch alarms:** Basic alarms will not be created for this cluster

Cluster subnet group:

Publicly accessible: Yes

Elastic IP: Not used

VPC security groups: ElasticMapReduce-master (sg-fed8884)

Enhanced VPC Routing: No

Encrypt database: No

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

Applicable charges:
 The on-demand hourly rate for this cluster will be ~~\$0.25~~ **\$0.25 /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 dc2.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**

Feedback English (US) © 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use





✔ Cluster **dw-test-01** is being created. Your cluster may take a few minutes to launch.

You will start accruing charges as soon as your cluster is active.

Applicable charges

The on-demand hourly rate for this cluster will be \$0.25 , or \$0.25 /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.

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



console.aws.amazon.com/redshift/home?region=us-east-1#cluster-list:

aws Services Resource Groups Camilo Valdes N. Virginia Support

Redshift dashboard

- Clusters
- Query editor New
- Saved queries
- Snapshots
- Security
- Parameter groups
- Workload management
- Reserved nodes
- Advisor ^{Beta}
- Events
- Connect client
- What's new

Clusters

[Quick launch cluster](#) [Launch cluster](#) [Cluster](#) [Database](#) [Backup](#) [Manage Tags](#) [Manage IAM roles](#)

	Cluster	Cluster Status	DB Health	Release Status	In Maintenance	Recent Events	Config timeline
<input type="checkbox"/>	dw-test-01	creating	unknown	Not found	unknown	0	View timeline

Open #security;names=;cluster= on this page in a new tab

© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#)

The screenshot displays the AWS Redshift console interface. At the top, the navigation bar includes the AWS logo, 'Services', 'Resource Groups', and user information for 'Camilo Valdes' in 'N. Virginia'. The main content area is titled 'Cluster: dw-test-01' and features several tabs: 'Configuration' (selected), 'Status', 'Cluster Performance', 'Database Performance', 'Queries', 'Loads', and 'Table resto'. A left-hand sidebar lists various Redshift management tools like 'Query editor', 'Saved queries', and 'Snapshots'. The main configuration page is divided into several sections:

- Endpoint:** [Not available: cluster creating]
- Cluster Properties:**
 - Cluster Name: dw-test-01
 - Cluster Type: Single Node
 - Node Type: dc2.large
 - Nodes: 1
 - Zone: us-east-1e
 - Created Time: [Not shown]
 - Maintenance Track: Current
 - Cluster Version: 1.0.4515
 - VPC ID: vpc-aa4963cd (View VPCs)
 - Cluster Subnet Group: default
 - VPC security groups: ElasticMapReduce-master (sg-fedab884) (active)
 - Cluster Parameter Group: default.redshift-1.0 (in-sync)
 - Enhanced VPC Routing: No
 - IAM Roles: See IAM roles
- Cluster Status:**
 - Cluster Status: creating
 - Database Health: unknown
 - In Maintenance Mode: unknown
 - Parameter Group Apply Status: in-sync
 - Pending Modified Values: Master User Password: ****
- Cluster Database Properties:**
 - Port: [Not shown]
 - Publicly Accessible: Yes
 - Database Name: test
 - Master Username: cvalde03
 - Encrypted: No
 - JDBC URL: [Endpoint not available]
 - ODBC URL: [Endpoint not available]
- Backup, Audit Logging, and Maintenance:**
 - Automated Snapshot Retention Period: 1
 - Cross-Region Snapshots Enabled: No
 - Audit Logging Enabled: No
 - Maintenance Window: sat:07:30-sat:08:00
 - Allow Version Upgrade: Yes
- Capacity Details:**
 - Current Node Type: dc2.large
 - CPU: 7 EC2 Compute Units (2 virtual)
- SSH ingestion settings:**
 - Cluster public key: [Input field]

At the bottom of the console, there is a footer with 'Feedback', 'English (US)', and copyright information for Amazon Web Services, Inc. (2008-2018).





Persist large amounts of data.

Visualization with Tableau

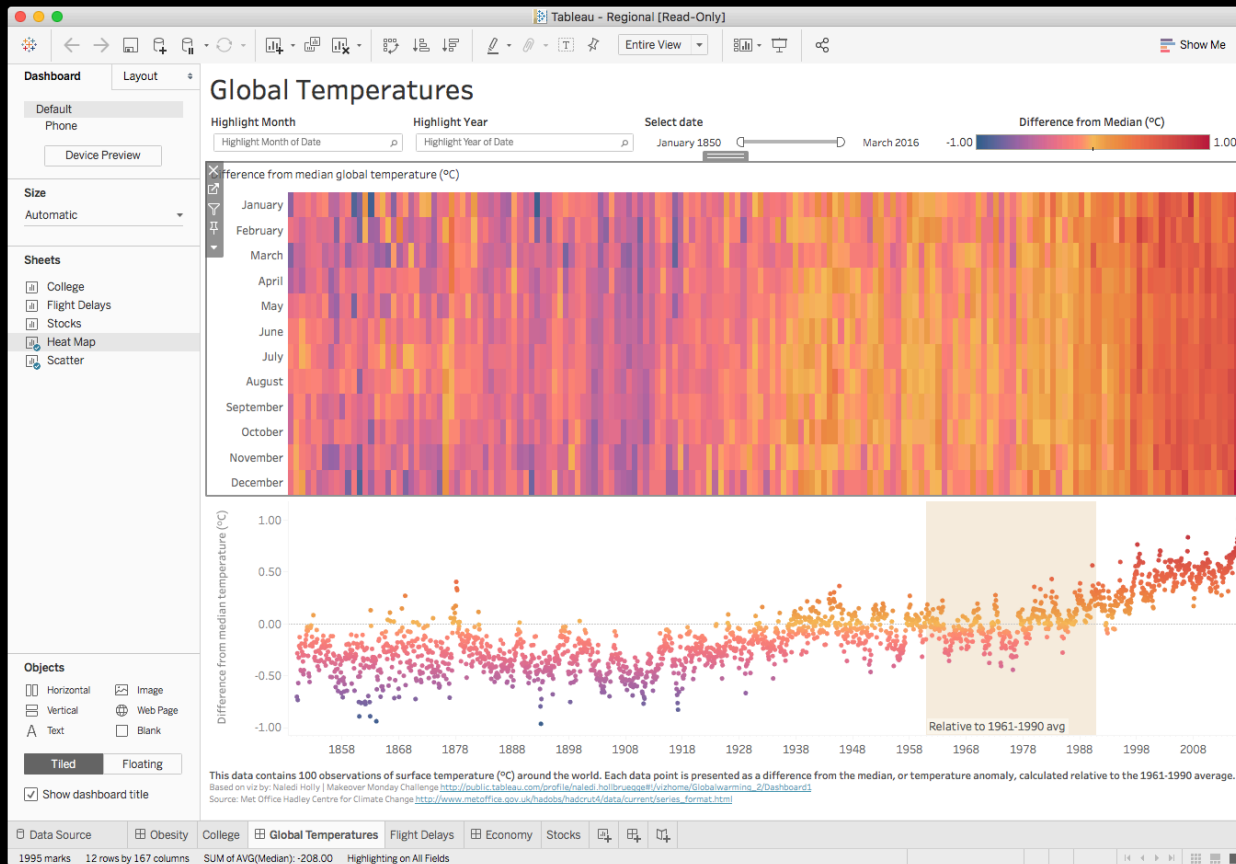


Tableau - Regional [Read-Only]

Connect

To a File

- Microsoft Excel
- Text file
- JSON file
- PDF file
- Spatial file
- Statistical file
- More...

To a Server

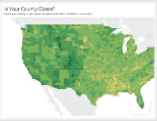


- Tableau Server
- Microsoft SQL Server
- Oracle**
- Amazon Redshift**
- More...

Saved Data Sources

- Sample - Superstore
- World Indicators

Open

[Open a Workbook](#)

- Regional 
- ensemble_taxTree_v40 
- Pathways 

Discover

Training

- Getting Started
- Connecting to Data
- Visual Analytics
- Understanding Tableau
- More training videos...

Sharing



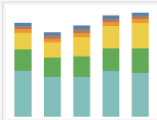
- Learn more about ways to share

Resources

- Get Tableau Prep
- Blog - Heatmaps, set actions, new dashboard formatting, and more—now available in Tableau
- Forums

Sample Workbooks

[More Samples](#)

- Superstore 
- Regional 
- World Indicators 

VIZ OF THE WEEK

Risk of Crises in Middle and Western Africa →

Update to 2018.3 Now

Tableau - Regional [Read-Only]

Connect

To a File

- Microsoft Excel
- Text file
- JSON file
- PDF file
- Spatial file
- Statistical file
- More...

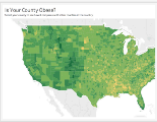
To a Server

- Tableau Server
- Microsoft SQL Server
- MySQL
- Oracle
- Amazon Redshift
- More...

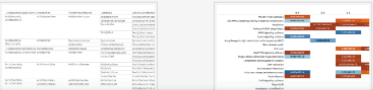
Saved Data Sources

- Sample - Superstore
- World Indicators

Open



Regional



Open a Workbook

Amazon Redshift

Server: Port:

Database:


Enter information to sign in to the database:

Username:

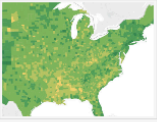
Password:

Require SSL

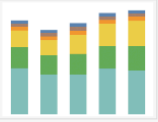
[Initial SQL...](#) Sign In



Superstore



Regional



World Indicators

More Samples

Discover

Training

- Getting Started
- Connecting to Data
- Visual Analytics
- Understanding Tableau
- More training videos...

Sharing

Learn more about ways to share

Resources

- Get Tableau Prep
- Blog - Heatmaps, set actions, new dashboard formatting, and more—now available in Tabl...
- Forums

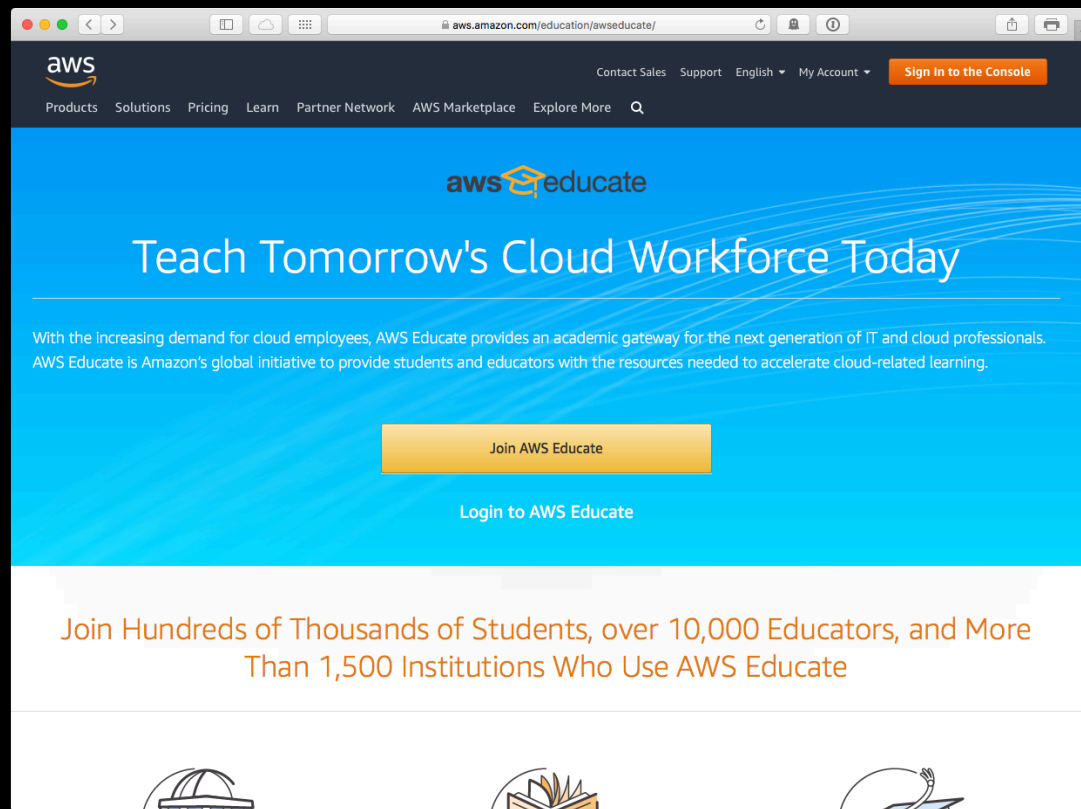
VIZ OF THE WEEK

Risk of Crises in Middle and Western Africa →

Update to 2018.3 Now

AWS Educate

https://www.awseducate.com




www.awseducate.com/student/s/pathways

Explore Cloud Career Pathways

Explore AWS Educate's Cloud Career Pathways to start building the key cloud skills you'll need to be successful in leading technology careers. Earn a completion credential for each pathway and share with prospective employers to show what you've learned.


Check out the roles below to learn more about each pathway and get started!



Cloud Computing 101

Take a crash course on the cloud, its history, solutions, and why companies across the globe are looking for employees with AWS cloud expertise.


[START ▶](#) [LEARN MORE ▶](#)



Application Developer

Curious how App Developers design, test, and improve engaging web and mobile applications in the cloud? Learn more about the skills you'll need.


[START ▶](#) [LEARN MORE ▶](#)



Cloud Support Associate

If you're excited by the future of cloud computing and enjoy working directly with customers, learn more about becoming a Cloud Support Associate.


[START ▶](#) [LEARN MORE ▶](#)



Cloud Support Engineer

Interested in multiple technologies and working with companies to support AWS cloud solutions? Learn more about becoming a Cloud Support Engineer.


[START ▶](#) [LEARN MORE ▶](#)



Cybersecurity Specialist

Cybersecurity Specialists use expertise in networking, programming, and coding to protect customer data every day. Learn more about the skills they use.


[START ▶](#) [LEARN MORE ▶](#)



Data Integration Specialist

Excited about bringing data sources together to tell the story of a product's performance? Discover ways to build and improve products through data.


[START ▶](#) [LEARN MORE ▶](#)



Data Scientist

Curious how discovering patterns in large data sets can translate into new business strategies? Learn more about how Data Scientists do this every day.

[START ▶](#) [LEARN MORE ▶](#)



DevOps Engineer

If you like working behind the scenes to tackle challenges and are curious about skills like scripting and coding, learn more about becoming a DevOps

[START ▶](#) [LEARN MORE ▶](#)