Cloud Computing & Visualization

Workflows

- Distributed Computation with Spark
- Data Warehousing with Redshift
- Visualization with Tableau
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.
- Analytics
  - EMR
- Database
  - Redshift
  - RDS
Changing the way you think about data

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

https://www.tableau.com
Distributed Computation with Spark

Building Serverless Applications
Build and run your applications and services without thinking about servers

Learn more
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,
pyspark

- Python version of the Spark Shell.

```bash
last login: Sat Oct 27 16:23:14 on tty2
Trajan$ pyspark
Python 2.7.14 (default, Mar 19 2018, 00:01:04)
[GCC 4.2.1 Compatible Apple LLVM 9.0.0 (clang-900.0.39.21)] 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).
18/03/19 18:07:42 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in-java classes where applicable
18/03/19 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 19 2018 00:01:04)
SparkSession available as 'spark'.
>>> 
```
In Spark, we express our computation through operations on distributed collections that are automatically parallelized across the cluster.

These collections are called resilient distributed datasets, or RDDs.

RDDs are Spark’s fundamental abstraction for distributed data and computation.
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

Building Serverless Applications

Build and run your applications and services without thinking about servers

Learn more
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.
Launch your Amazon Redshift cluster - Advanced settings | Switch to quick launch

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

- **Node type**: **dc2.large**
  - 7 EC2 Compute Units (2 virtual cores) per node
  - 15.25 GB per node
  - 160GB SSD storage per node
  - I/O performance: Moderate

- **Cluster type**: **Single Node**

- **Number of compute nodes**: 1

---

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.

---

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

---

Single Node clusters consist of a single node which performs both leader and compute functions.
Launch your Amazon Redshift cluster - Advanced settings

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 username: dwuser01
- 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-za9h63cc
- Cluster subnet group:
  - Publicly accessible: Yes
  - Elastic IP: Not used
- VPC security groups: (Default security group used)
- 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. If you have purchased reserved nodes in this region for this
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/node. If you have purchased reserved nodes in this region for this node type that are active, your costs will be discounted.

If you are eligible for a free trial, you will receive 750 hours of service each month of the trial, applied across all running ds2.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.
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
Persist large amounts of data.
Visualization with Tableau

This image shows a Tableau dashboard illustrating global temperature trends. The dashboard is titled "Global Temperatures" and displays data over a timeline from 1851 to 2008. The visualization includes a scatter plot and a heat map, each highlighting differences from the median global temperature. The data is presented as a deviation from the 1961-1990 average, with colors ranging from blue (lower temperatures) to red (higher temperatures). The dashboard allows users to select different data points and highlight specific months or years, providing a dynamic way to analyze long-term climate data.
AWS Educate

Teach Tomorrow’s Cloud Workforce Today

With the increasing demand for cloud employees, AWS Educate provides an academic gateway for the next generation of IT and cloud professionals. AWS Educate is Amazon’s global initiative to provide students and educators with the resources needed to accelerate cloud-related learning.

Join AWS Educate

Login to AWS Educate

Join Hundreds of Thousands of Students, over 10,000 Educators, and More Than 1,500 Institutions Who Use AWS Educate

https://www.awseducate.com
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.

**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.

**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.

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

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

**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.

**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.

**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