BIG DATA on AWS

Data Analysis

Agenda

01 What is Amazon Redshift?

02 Data Warehouse System Architecture

03 Redshift Concepts

04 Designing tables

05 Loading Data to Redshift

06 Using the Redshift Query Editor

07 Tuning Query Performance

08 Best Practices using Redshift

Agenda

09 Amazon SageMaker

10 How SageMaker works?

11 Built-in Algorithms in SageMaker

12 What is Amazon Athena?

13 When should you use Athena?

14 What Is Amazon Elasticsearch Service?

15 How Elasticsearch works?

16 ES Domains

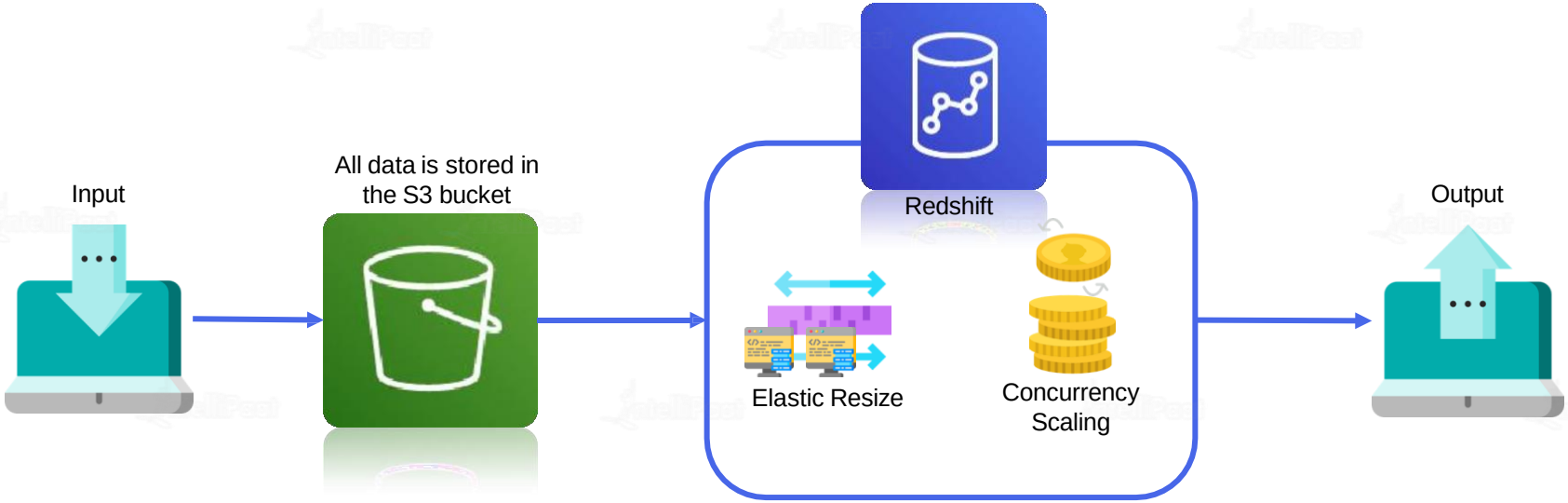
What is Amazon Redshift?

What is Amazon Redshift?

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. You can start with just a few hundred gigabytes of data and scale to a petabyte or more. This enables you to use your data to acquire new insights for your business and customers.

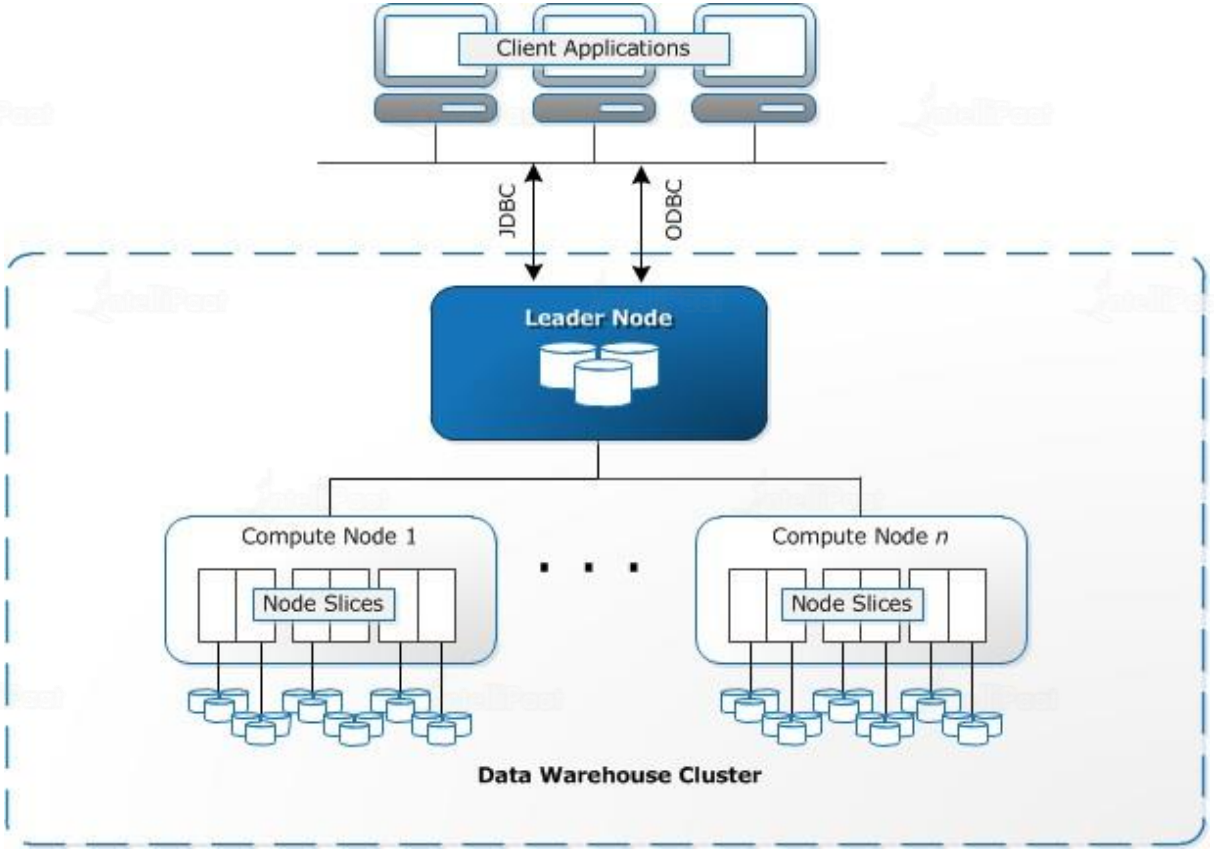
What is Amazon Redshift?

This is a sample use case on how Redshift can be used.



Data Warehouse System Architecture

Data Warehouse System Architecture



Data Warehouse System Architecture

- Client applications
- Clusters
- Leader node
- Compute nodes

Amazon Redshift integrates with various data loading and ETL (extract, transform, and load) tools and business intelligence (BI) reporting, data mining, and analytics tools.

Redshift is based on PostgreSQL and this makes it easy to connect Redshift to most SQL clients. Connections can be made with JDBC and ODBC drivers.

Data Warehouse System Architecture

- Client applications
- Clusters
- Leader node
- Compute nodes

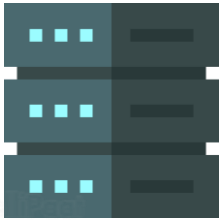
The core infrastructure component of an Amazon Redshift data warehouse is a **cluster**.

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.

Data Warehouse System Architecture

- Client applications
- Clusters
- Leader node
- Compute nodes

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.

Data Warehouse System Architecture

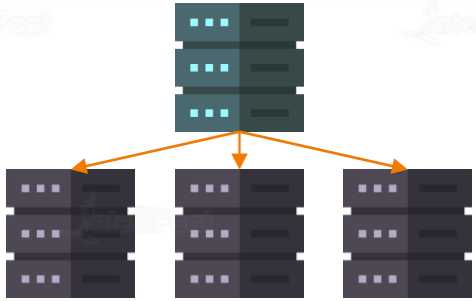
Client applications

Clusters

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.

Data Warehouse System Architecture

Client applications

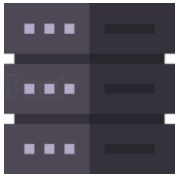
Clusters

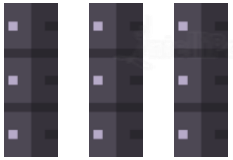
Leader node

Compute nodes

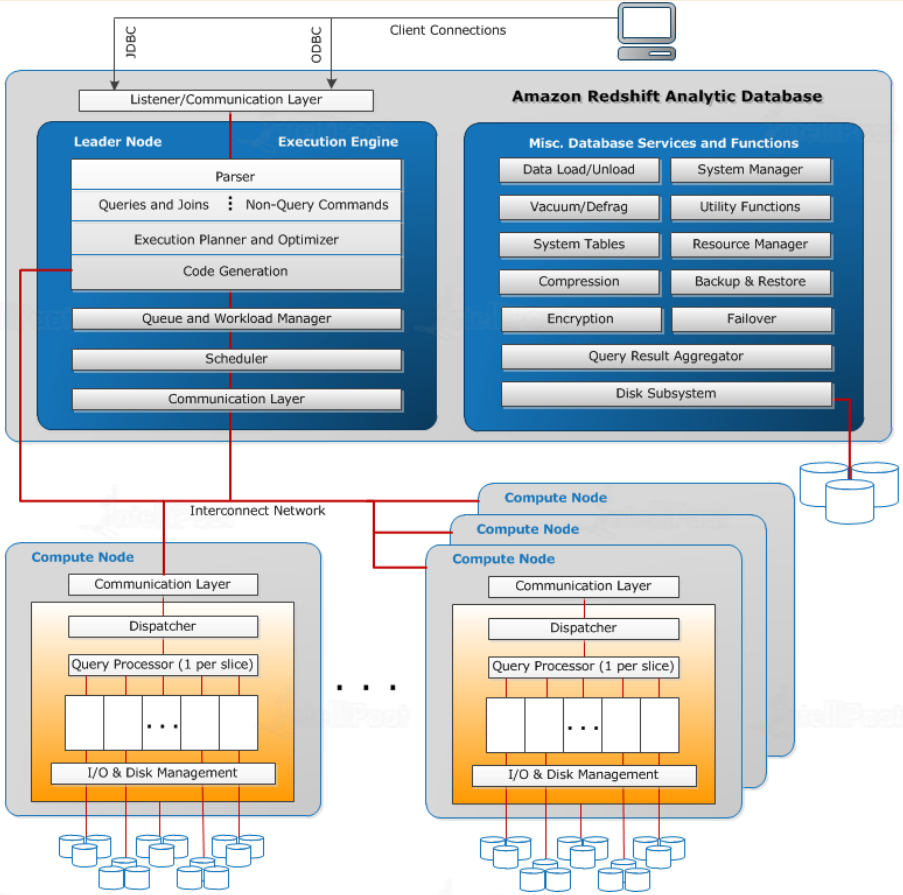
Node Slices

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.





Data Warehouse System Architecture

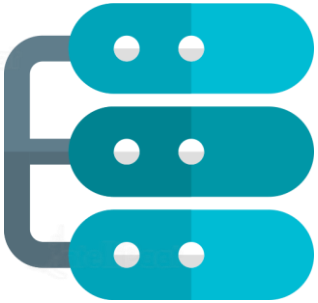


Redshift concepts

Redshift concepts

Massively Parallel Processing

MPP enables fast execution of the most complex queries operating on large amounts of data



Columnar Storage

Workload Management

Multiple compute nodes handle all query processing which is aggregated into a final result, with each core executing the same compiled query segments on portions of the entire data set.

Redshift concepts

Massively Parallel Processing

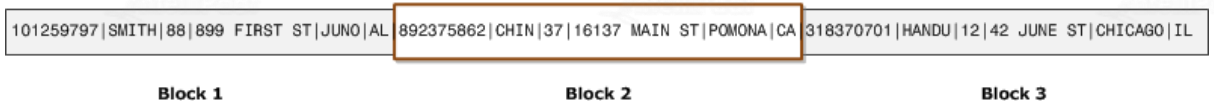
Columnar Storage

Workload Management

Columnar storage helps in optimizing query performance because it reduces the overall disk I/O requirements and reduces the amount of data you need to load from disk.

Typical disk blocks stored in row

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



Redshift concepts

Massively Parallel Processing

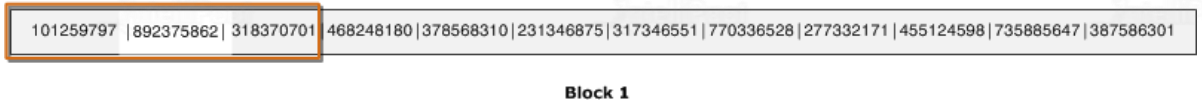
Columnar Storage

Workload Management

Using columnar storage, each data block holds column field values for as many as three times as many records as row-based storage.

Columnar storage

SSN	Name	Age	Addr	City	St
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL



Block 1

Redshift concepts

Massively Parallel Processing

Amazon Redshift WLM enables users to flexibly manage priorities within workloads so that short, fast-running queries won't get stuck in queues behind long-running queries.

Columnar Storage

Workload Management

Automatic WLM manages the resources required to run queries. If you use Manual WLM, you will have to decide the memory and concurrency of queries to run.

Designing tables

Designing tables

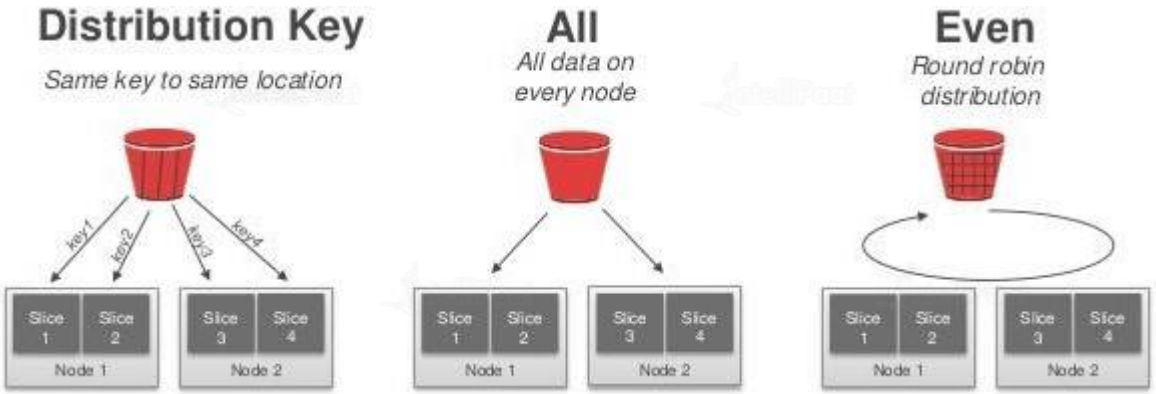
A data warehouse system has very different design goals compared to a typical transaction-oriented relational database system. Redshift enables quick execution of complex analytic queries against large data sets.

The table design plays a major role in making the queries faster

Designing tables

When you load data into a table, Amazon Redshift distributes the rows of the table to each of the node slices according to the table's distribution style.

Distribution Styles



Hands-on: Creating a Redshift Cluster

Loading Data to Redshift

Loading Data to Redshift

Top ways to do load data into redshift

Using COPY command

- The COPY command leverages the Amazon Redshift massively parallel processing (MPP) architecture to read and load data in parallel from files in an Amazon S3 bucket.

Using DML commands

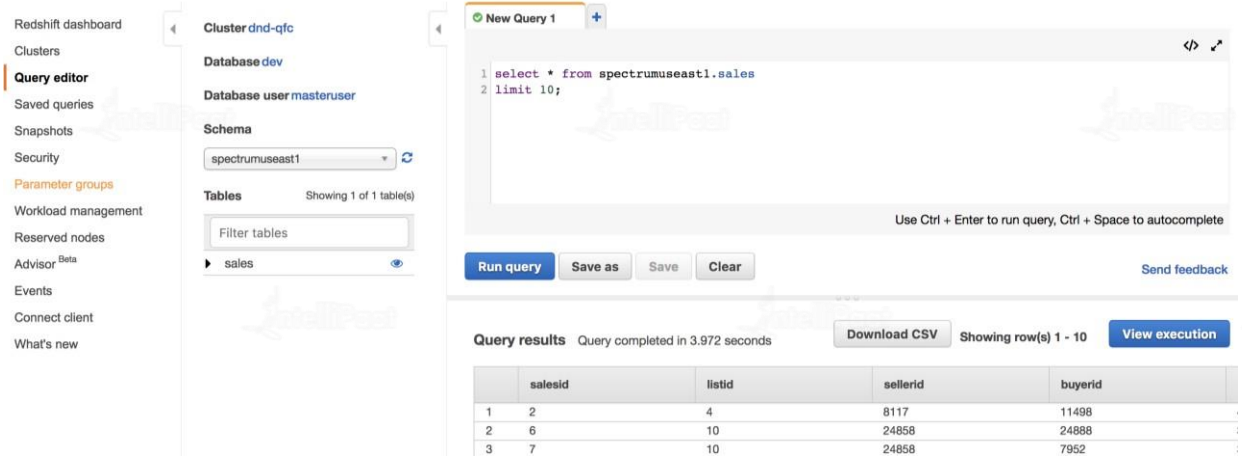
- CREATE, INSERT, UPDATE, and DELETE that you can use to modify rows in tables. If your table already exists in redshift, then it is same as how you write queries in SQL

Hands-on: Loading data into the cluster

Using the Redshift Query Editor

Using the Redshift Query Editor

Connect to your cluster and run queries on the AWS Management Console with the query editor.



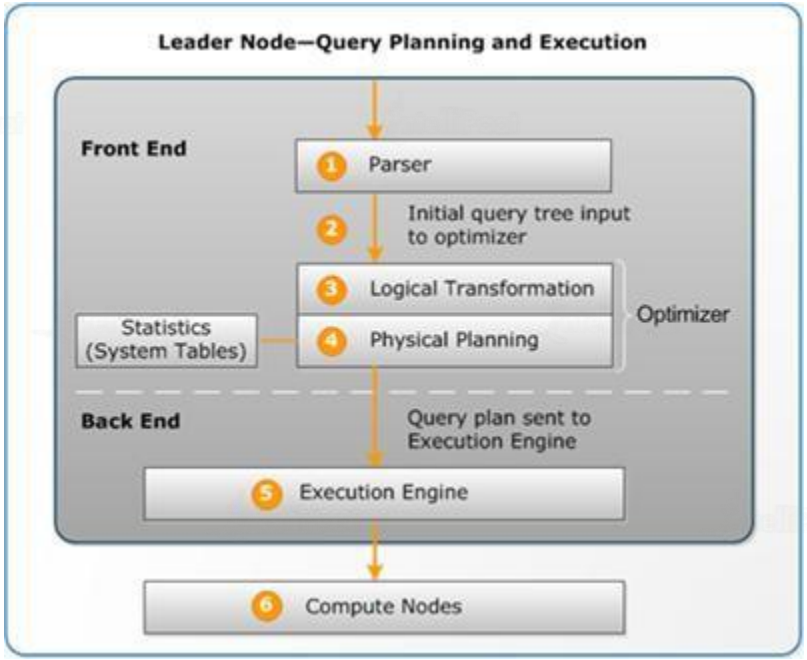
If you use the query editor on the Amazon Redshift console, you don't have to download and set up a SQL client application.

Hands-on: Querying the cluster

Tuning Query Performance

Tuning Query Performance

Amazon Redshift uses queries based on structured query language (SQL) to interact with data and objects in the system.



Tuning Query Performance

Improving query performance

- If ghost rows or uncommitted rows are present, you might see an alert event in STL_ALERT_EVENT_LOG that indicates excessive ghost rows. If data is not loading currently, run VACUUM command to clean those rows.
- If your query returns a very large result set, consider rewriting the query to use UNLOAD to write the results to Amazon S3.
- If unsorted or missorted rows are present, you might see a very selective filter alert event in STL_ALERT_EVENT_LOG. Again run VACUUM to re-sort those rows.

Best Practices using Redshift

Best Practices using Redshift

- If recent data is queried most frequently, specify the timestamp column as the leading column for the sort key
- Change some dimension tables to use ALL distribution
- COPY command will use compression encodings to an empty table automatically as part of the load operation which makes it a faster load
- Consider the largest values you are likely to store in a VARCHAR column
- Amazon Redshift stores DATE and TIMESTAMP data more efficiently than CHAR or VARCHAR

Amazon SageMaker

Amazon SageMaker

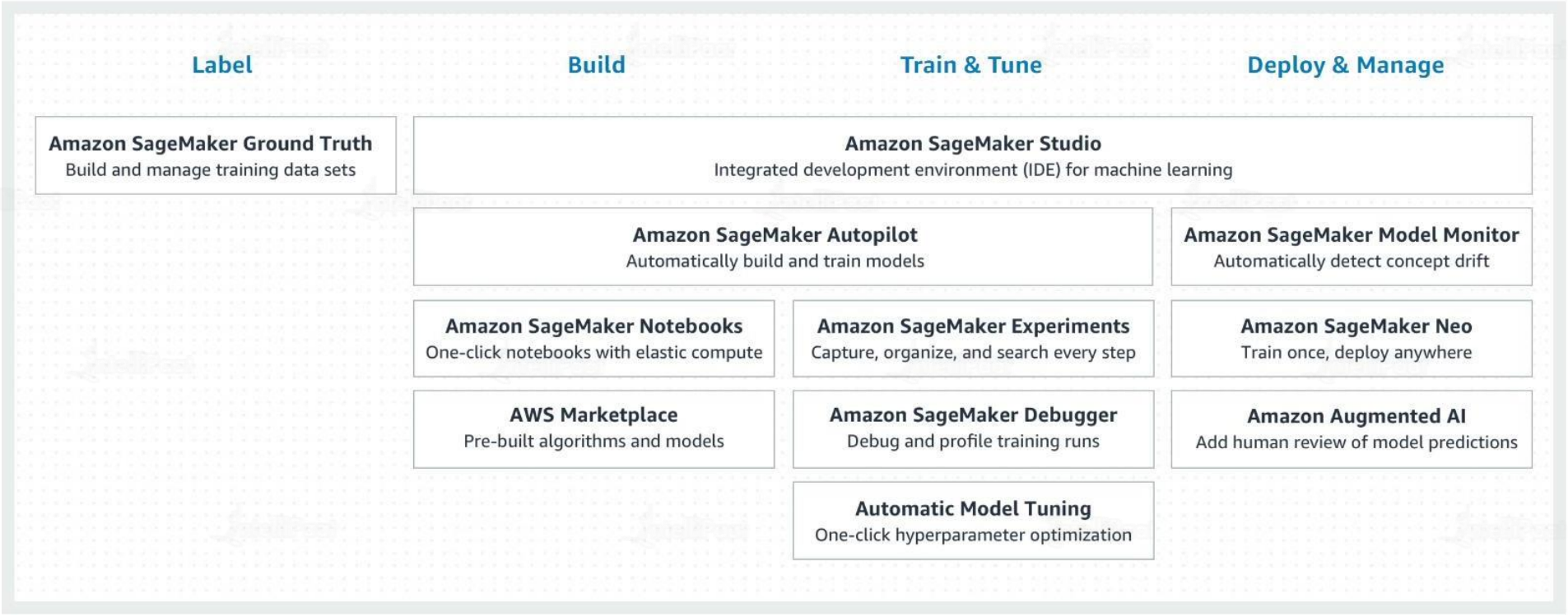
Amazon SageMaker is a fully managed machine learning service. With Amazon SageMaker, data scientists and developers can quickly and easily build and train machine learning models, and then directly deploy them into a production-ready hosted environment.

AWS is the best place to run TensorFlow

90%

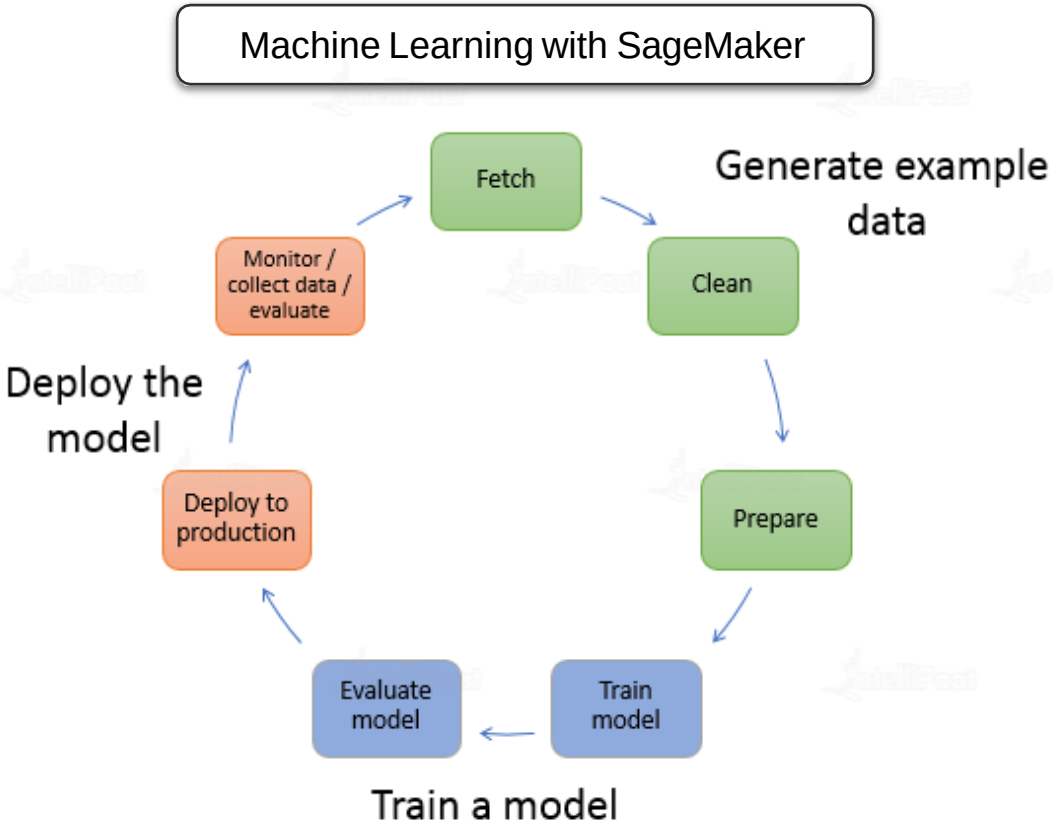
SCALING EFFICIENCY WITH 256 GPUS

Amazon SageMaker



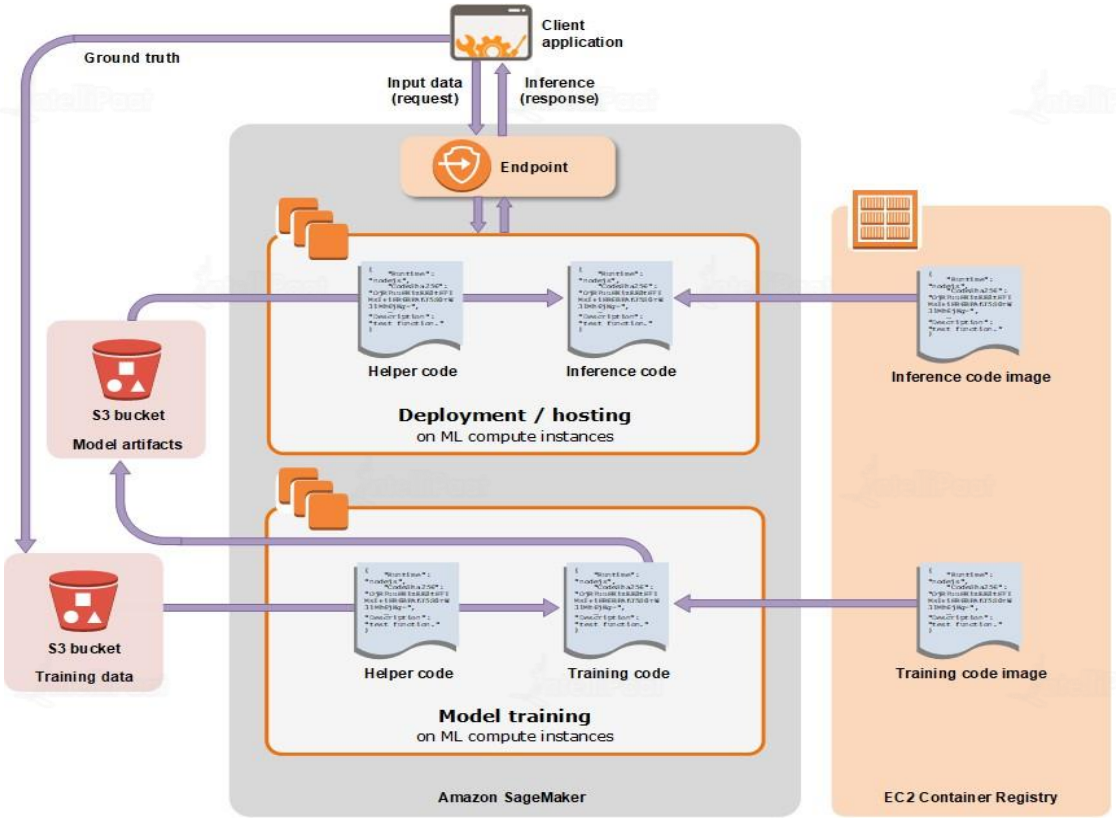
How SageMaker works?

How SageMaker works?



How SageMaker works?

Train and Deploy your model



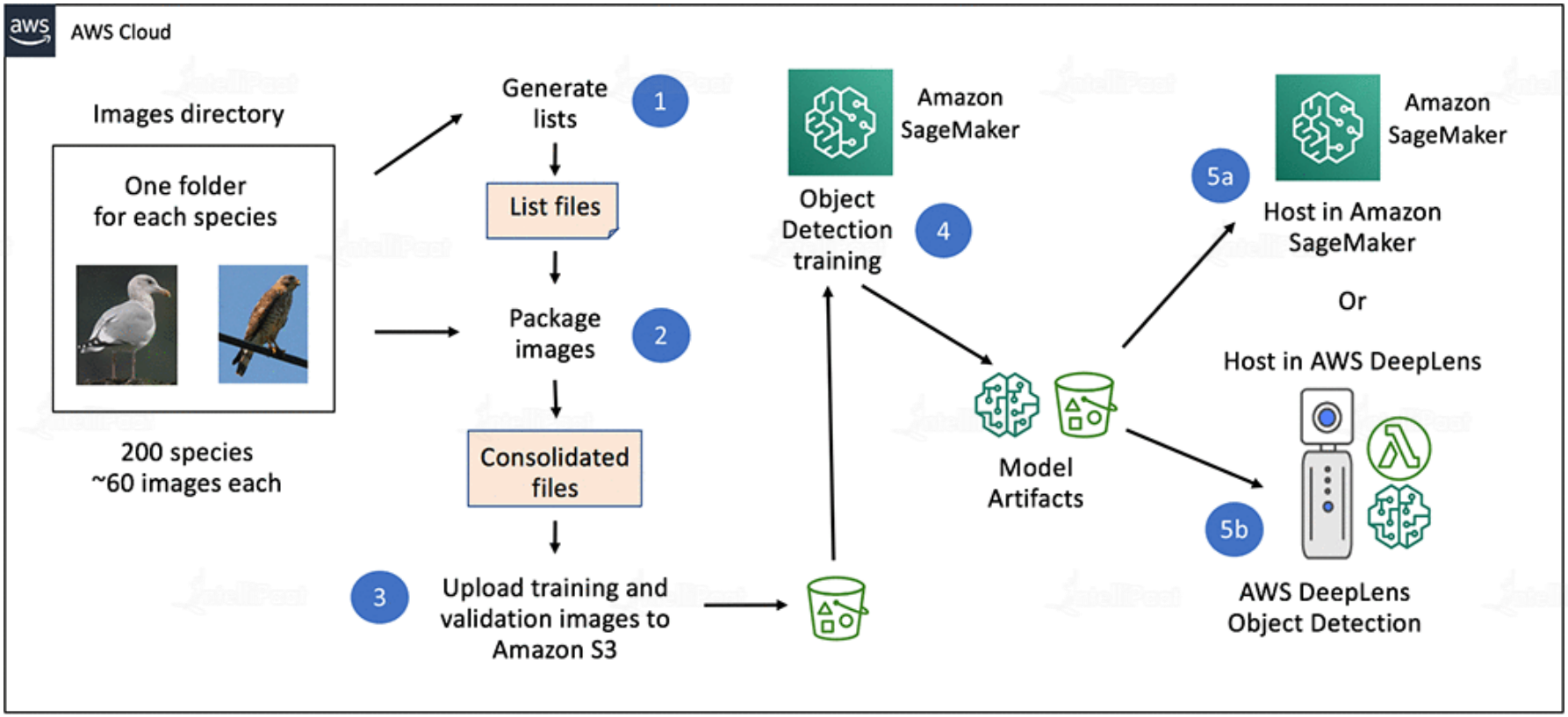
Built-in Algorithms in SageMaker

Built-in Algorithms in SageMaker

A machine learning algorithm uses example data to create a generalized solution (a model) that addresses the business question you are trying to answer.

Amazon SageMaker provides several built-in machine learning algorithms that you can use for a variety of problem types.

Object Detection Algorithm Example

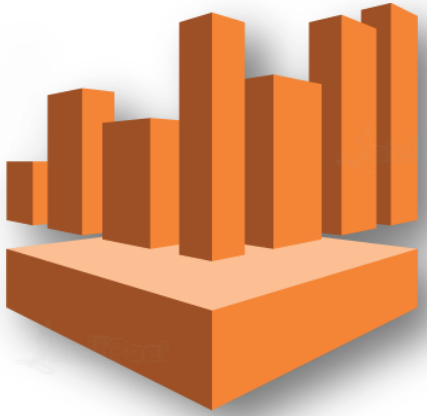


Hands-on: SageMaker

What is Amazon Athena?

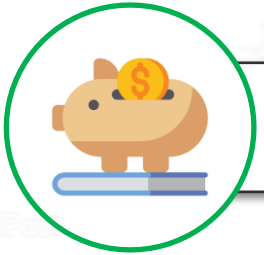
What is Amazon Athena?

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.



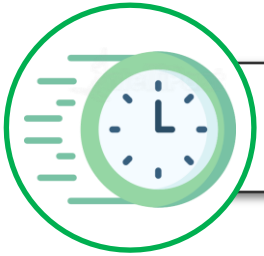
What is Amazon Athena?

Start querying instantly



Pay per query

Open, powerful, standard



Fast, really fast

When should you use Athena?

When should you use Athena?

- Athena helps you analyze unstructured, semi-structured, and structured data stored in Amazon S3.
- Athena integrates with Amazon QuickSight for easy data visualization.
- Easy connection using JDBC or ODBC drivers with SQL or BI clients
- Integration with Glue Data Catalog enables persistent metadata store for data in S3

Hands-on: Creating a database and running queries

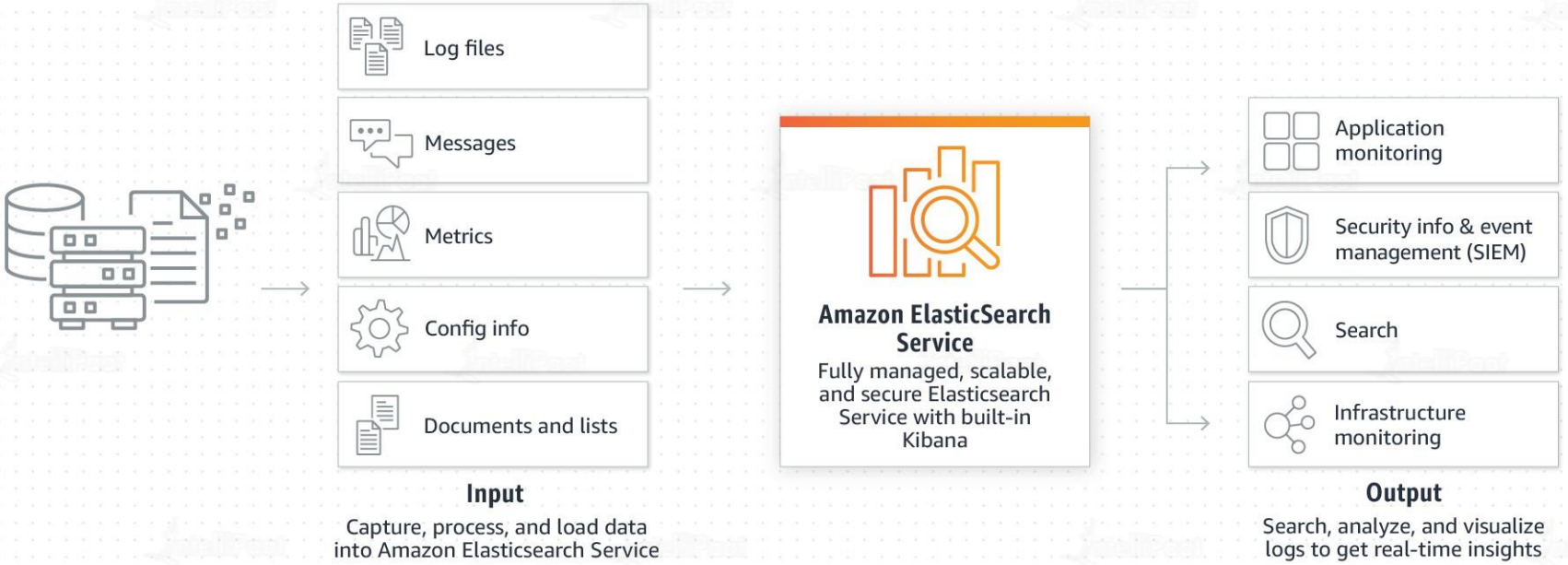
What Is Amazon Elasticsearch Service?

What Is Amazon Elasticsearch Service?

Amazon Elasticsearch Service is a fully managed service that makes it easy for you to deploy, secure, and run Elasticsearch cost effectively at scale.

How Elasticsearch works?

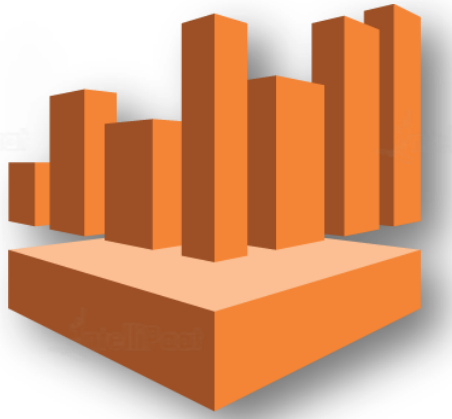
How Elasticsearch works?

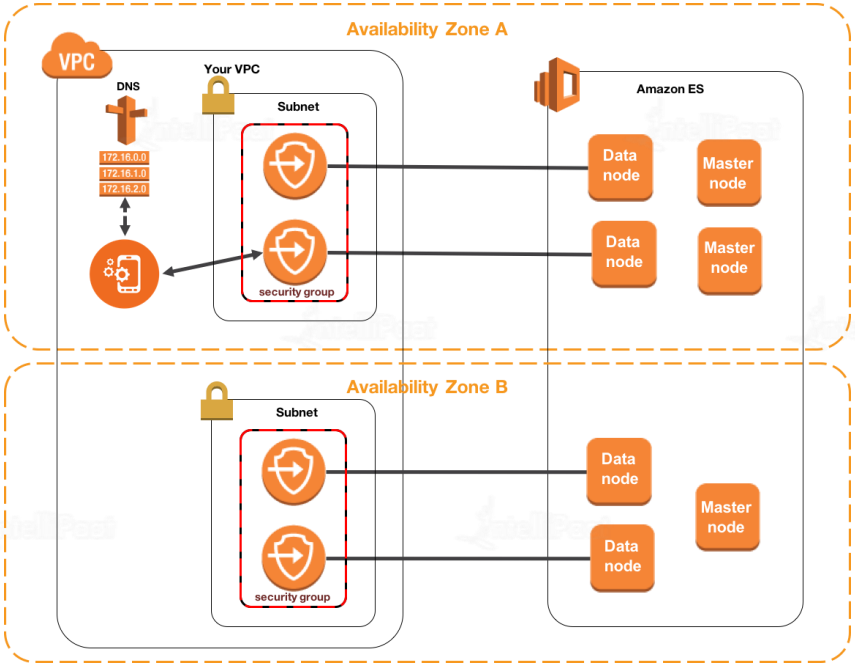


ES Domains

ES Domains

An Amazon ES domain is synonymous with an Elasticsearch cluster. Domains are clusters with the settings, instance types, instance counts, and storage resources that you specify.





+919030485102

rganesh0203@gmail.com

https://topmate.io/rganesh_0203