# Abstract

# Introduction

# Amazon Services Overview

Amazon Redshift is Amazon’s cloud-based data warehousing platform. Like many data warehousing platforms, it consists of clusters which are used to service the data queries of the user. Each cluster is made up of compute nodes. Clusters which have more than one compute node, have a leader node which is defined. The leader node coordinates the compute nodes and acts as the main interface to the client. The data itself is stored in a separate storage tier known as Redshift Managed Storage. This separation allows for scaling of storage space without interrupting the service. Data is stored in as columnar which drastically reduces the over disk I/O and reduces the amount of data you need to load from the disk. Because each column holds the same type of data, compression can be used to further decrease disk I/O. The standard block size used by Redshift is 1 MB.

Amazon Simple Storage Service (Amazon S3) is a cloud object storage service. This provides the base storage location for building a data lake. It integrates into AWS Glue, Redshift and SageMaker and uses common policies for access.

It is important to note that the Redshift databases are structured and optimized to analyze relational data. The use we will be discussing in this paper will be an interface between a data lake stored on Amazon S3 and running a ML operation on Redshift.

# Data Lakes

A data lake is a modern data management strategy and they have become popular due to their ability to store large amounts of data at a low cost. They allow data to be stored in its original raw form without minimal preprocessing or modifications. This means both unstructured and structured data can be stored with no predefine schema. As compared to a data warehouse which requires all its data the be highly and its schema defined on implementation. However, with a data lake, as is with any data collection, it is important to have good data-quality and good data-governance otherwise it will become unmanageable.

Upon data being added to a data lake, it needs to be cataloged and labeled using metadata. This central catalog allows users to discover what datasets are available and speeds any queries run directly on the data.

# Amazon Redshift ML

# Conclusion