An Experimental Study of Redshift

# Abstract

This project will explore the data warehouse and data lake capabilities of Amazon’s Redshift product. It will follow class work on data warehousing and expand to the data lake capabilities of this product. Upon reviewing the data lake capabilities it will explore the its use for machine learning and other data science applications.

# Overview

Redshift is Amazon’s cloud-based data warehousing product which is a part of Amazon’s Web Services platform. Redshift has the ability to handle large analytic workloads stored on a column-oriented DBSM principle. It utilizes parallel-processing and compression to optimize execution time, and this allows Redshift to perform operations on billions of rows at once.

# Timeline

Below is a rough timeline and breakdown of goals for the project.

|  |  |
| --- | --- |
| Deadline target | Goal Description |
| 11/20/22 | Review Product and determine feature set |
| 11/27/22 | Use example data set provided and work through Tutorials |
| 12/3/22 | Create examples for presentation and paper |
| 12/7/22 | Complete paper |

# Definition of Work

For this project I will create a free redshift cluster and work through the tutorials and examples which have been provided to for the product. I will highlight key features and common tasks which are typically done when using a data warehouse or data lake applications. In a data lake, the schema of the database is defined on time of analysis and is in a raw type of format. This means that when the data is captured it can be stored without regard to design or structure. This format is useful in IoT, mobile applications and social media because it does not need to be fit to a schema until the data is read. I will also focus on how Redshift can take these interfaces and use it for machine learning and other data science type applications.