[[1]](#footnote-1)

When does it make sense to use a Data Lake instead of a Data Warehouse?

Cameron Stewart, Nathan Deinlein, Cleveland Johnson

Data Lakes are a new technology in almost all industries. There is not a lot of clarity within Technology organizations on how to best create and leverage Data Lakes. A company’s decision to invest in this new technology is expensive and requires significant thought into initial processes to ensure integrity and quality.

*Index Terms*— data, data lake, data warehouse, mysql, mongoDB

# Data lake vs data warehouse problem

This section describe the problem set that we are working to address.

# methodology

The section details the approach of how we are going to work to understand data lakes, data warehouses and their differences.

We will be researching the different use cases for data lakes versus data warehouses.

# Data Warehouse Results

This section details our “Data Warehouse” including the structure, data within and to the best of our ability details on what insights the data provides.

We will create a data warehouse using MySQL. We will also test ingestion techniques and provide observations on optimization and reporting from the data warehouse.

# Data Lake Results

This section details our “Data Lake” including the “structure”, data within and what insights it is used to find.

We will create a data lake using MongoDB Data Lake. We will test ingestion techniques and provide observations on optimization and reporting from the data lake.

# Data lake vs data warehouse Analysis

This section houses the analysis of data warehouse results compared to data lake results.

This highlights the unique attributes of the data warehouse and also the data warehouse of the data lake. We are working to answer the following questions:

* How is the data different?
* Can we run the same queries against the different data sources?
* Is there a speed difference between the two data sources? If so, what is the difference?
* What makes each data source unique?
* What can we showcase from data lakes that someone would not necessarily know?

# Conclusion

This section will hold the decision tree of when we think one should use a data warehouse and when one should use a data lake. Ultimately confirming or differing from what was originally stated in the methodology section.

REFERENCES

1. Creating a Data Lake in MongoDB: <https://docs.mongodb.com/datalake/>
2. Accessing Twitter data: <https://www.youtube.com/watch?v=rm0mJP1P4eE>

1. [↑](#footnote-ref-1)