Consolidated Lab

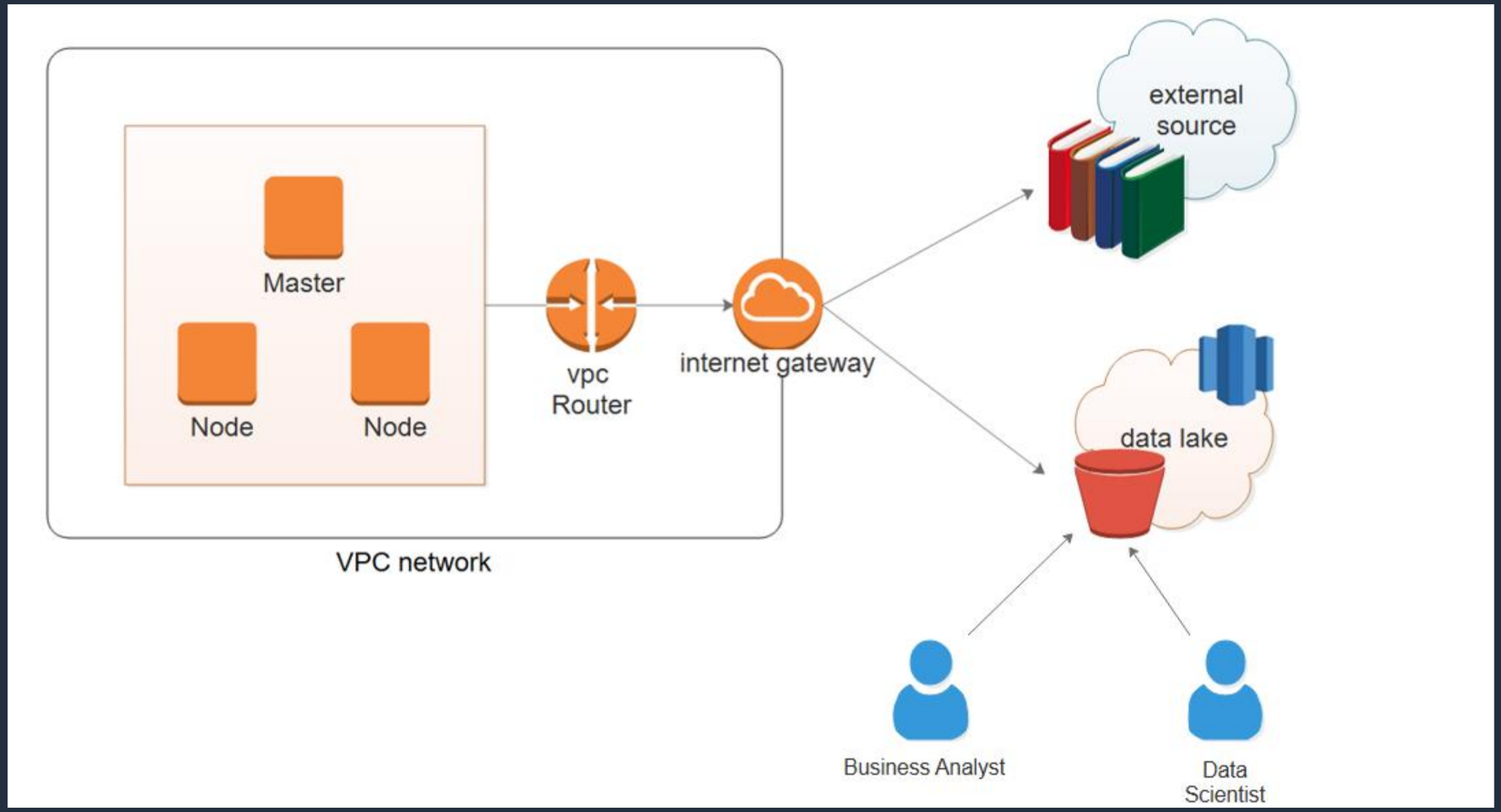
Odwa Yekela
Cloud Support Engineer
AWS

Mission

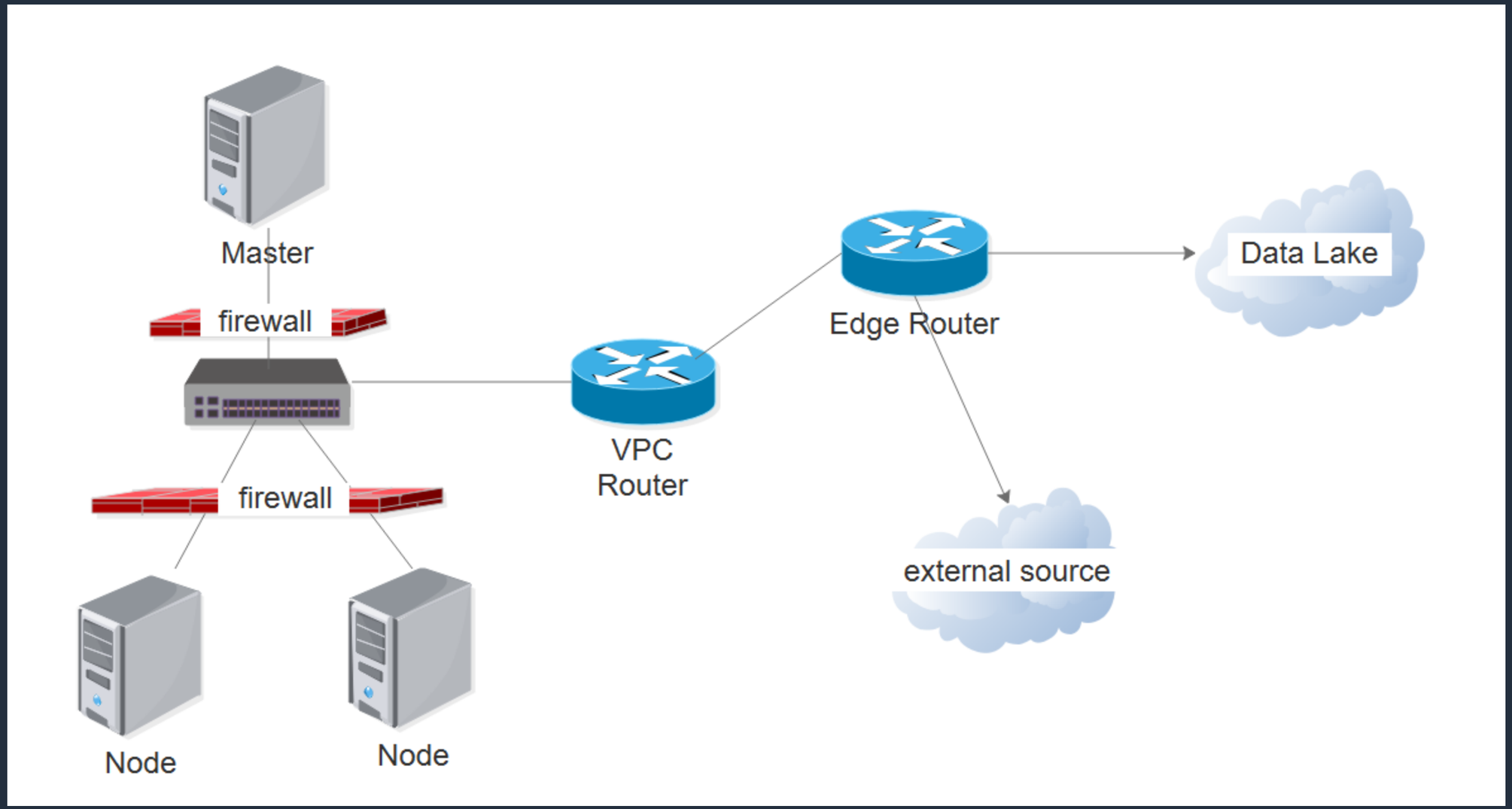
In this lab we will be moving data from an external data lake and move the dataset into HDFS so that we can stream the files from HDFS into spark. This will require some careful thinking regarding how we treat the dataset to optimize our read jobs. At the end of the lab will enjoy a nice visual illustration of our data inside Hue.

1. Download the files for this lab from S3 Data Lake location
2. Package and submit spark stream application to EMR cluster
3. Combining multiple smaller files to few larger files
4. Visualize data in HUE

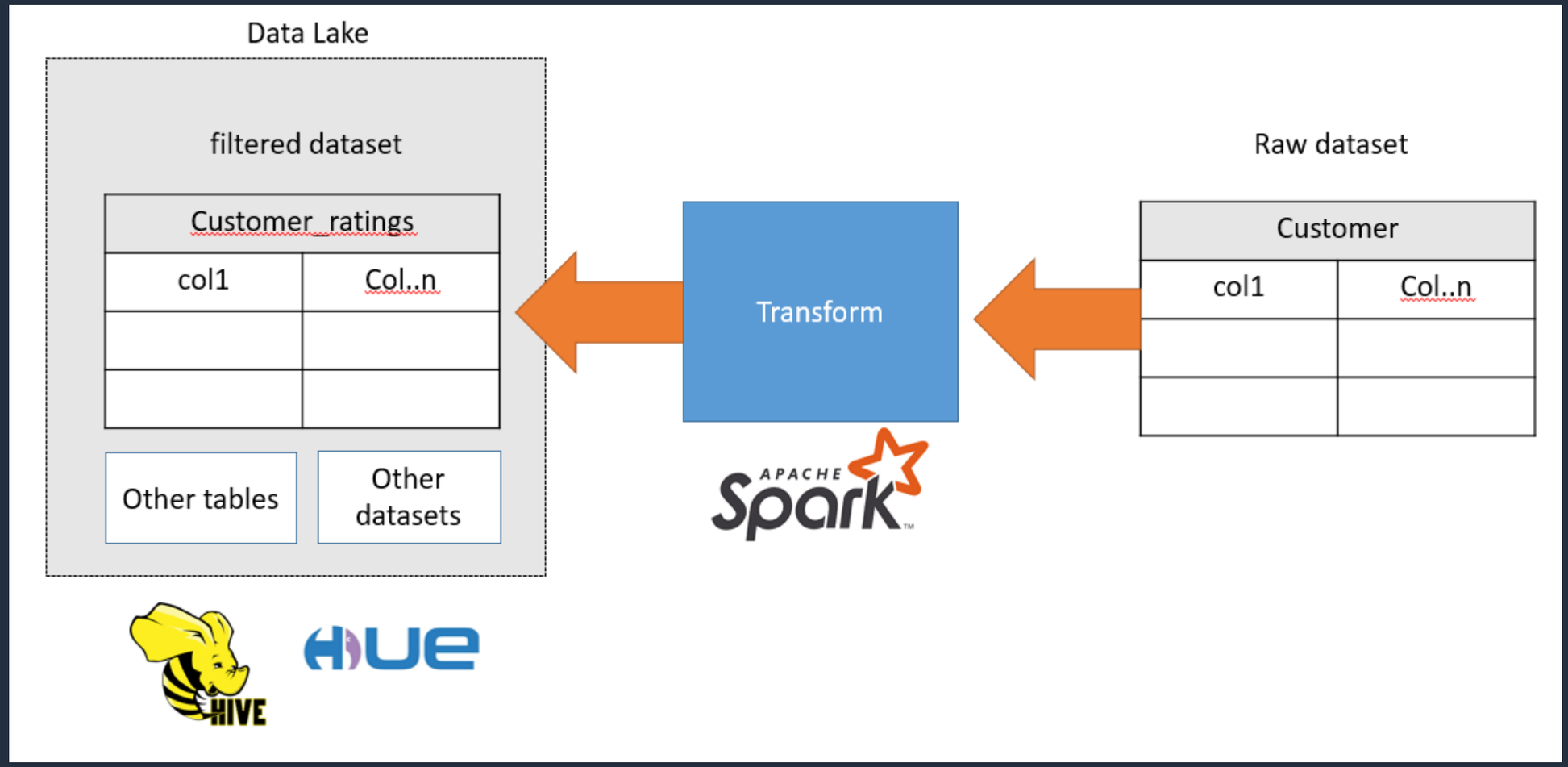
Setup



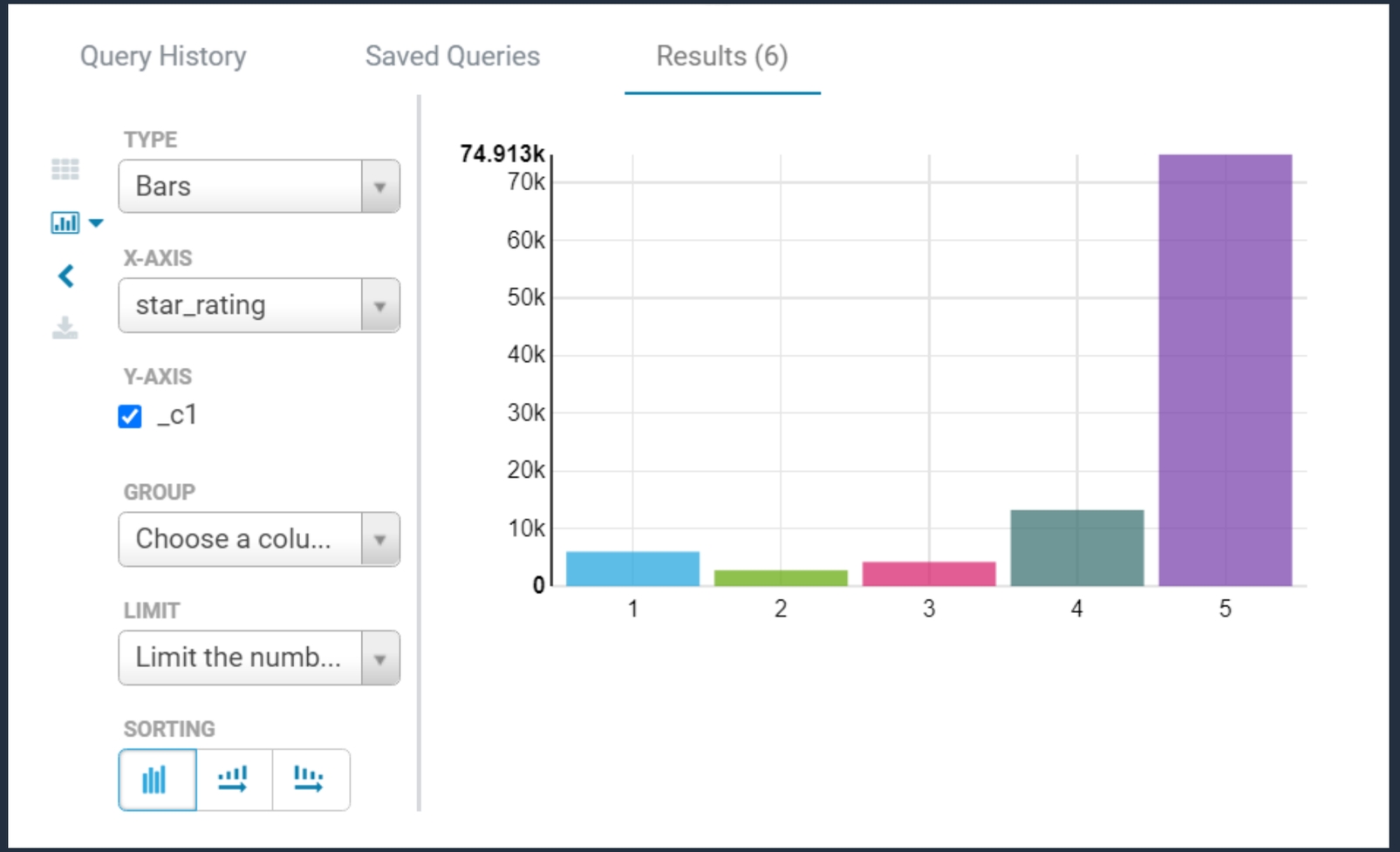
Physical representation (conceptual)



Extract Transform Load



Preview of dataset



Questions?

Thank You