The transformer model has been a recent breakthrough for machine learning models as it uses attention to assist with sequence data predictions. Attention will allow the model to focus on important and relevant parts of the model which will improve robustness and accuracy. While transformers are prominently used for natural language processing, its application in time series application has largely been unexplored. In this project we utilize the transformer model as a tool to address the problem of predicting cloud traffic. Leveraging the Google Cluster Trace dataset for training and testing, we set off to aim to develop a model capable of predicting the usage of key metric such as mean CPU usage. The transformer model will be compared to comparative time series workload prediction models such as LSTM which are widely used for time series prediction problem. A workflow to interact with the transformer from external components has also been implemented to allow for greater modularity within the system. Furthermore, the introduction of a sliding window has been developed, an evaluation of its effectiveness will be carried out. Our work includes comprehensive analysis of the transformer model and all its underlying components. Certain hyperparameters such as data sampling intervals, epochs and sequence length will be optimized to find the best combination for this task. Preliminary results show that optimizing the combination of hyperparameters and input data will largely influence the accuracy of the model. This project aims to highlight the potential of the transformer model in the time series prediction space, with the hope of encouraging peers to undergo further exploration in this field.