Q. What algorithm, machine learning or AI approaches would you take to find anomalies in the duration of a span? And why do you think that approach is a good approach?

Ans - Several machine learning and AI techniques may be taken into consideration in order to identify abnormalities throughout the course of a span. Here are a few popular methods and the reasoning behind them:

1. Statistical Methods:

A. Z-Score or Modified Z-Score: Determine each duration's Z-Score and mark as anomalous any spans with a Z-Score higher than a threshold. Non-Gaussian distributions can be handled with the Modified Z-Score.

B. Quartile-based Methods: To find outliers, use the interquartile range (IQR). Anomalies can be defined as spans that are more than 1.5 times the IQR from the quartiles.

C. Histogram Analysis: See how the durations are distributed and search for any peaks or long tails that could point to abnormalities.

2. Analysis of Time Series:

A. Moving Average: Use moving averages to smooth the time series of durations and mark anomalies in any departures from the smoothed curve.

B. Seasonal Decomposition: Find anomalies in the residual component by breaking down the time series into trend, seasonal, and residual components.

3. Machine Learning Models:

A. Isolation Forest: In high-dimensional datasets, this unsupervised learning technique works well for identifying outliers. By choosing a characteristic at random and then a split value between the maximum and minimum values of that feature, it isolates anomalies.

B. One-Class SVM: This method detects anomalies in each data point that deviates from the learnt distribution after learning the distribution of normal data.

4. Deep Learning Methods:

A. Autoencoders: Use the duration data to train an autoencoder neural network. By comparing each span's reconstruction error with a predetermined threshold, anomalies can be found.

B. LSTM Networks: These networks, which stand for Long Short-Term Memory, are able to recognize patterns that have been learnt and capture temporal relationships in duration data.

The distribution of duration data, the availability of labeled data, the need for interpretability, computational resources, and other considerations all influence the choice of strategy. Although statistical techniques are easily understood and straightforward, they could miss subtle patterns. Complex correlations in the data can be captured by machine learning and deep learning models, but they need more computer power and labeled training data. When working with data of sequential length, time series analysis is appropriate.