Propose a model monitoring pipeline and describe how you would track model drift in 500 words.

In this Automatic Speech Recognition (ASR) project, the ASR system transcribes audio files from the Common Voice dataset, with their transcripts stored in Elasticsearch. A robust model monitoring pipeline should (1) facilitate tracking the system's performance, (2) detect anomalies over time, and (3) react promptly to changes in the data or the environment that may affect transcription accuracy. To this end, it must systematically collect audio input and ground truth labels (when available) for comparison with model outputs.

The first consideration involves establishing key performance indicators (KPIs) for the ASR model. Common metrics include Word Error Rate (WER), which compares predicted transcripts against reference transcripts to quantify transcription accuracy. In addition, model confidence scores can be tracked to identify model drifts. Whisper (from OpenAI) provides confidence scores based on word-level log probabilities while in Wav2Vec (from Meta), confidence scores can be computed from softmax probabilities of the model's logits. Latency and throughput metrics can also be captured to ensure the system meets required real-time or large-scale processing demands. All these metrics can be aggregated and visualized in Kibana dashboards, integrated with Elasticsearch and linked to alerting rules. For example, a line graph tracking WER trends can trigger Slack or email alerts when WER exceeds a 10-15% threshold, suggesting potential model drift due to vocabulary shifts, poor audio quality, or unseen accents.

ASR systems are susceptible to both model and data drifts, influenced by changing speech patterns, linguistic trends, and recording conditions. To track and quantify both types of drift, we can compare historical distributions of key performance metrics (e.g., Word Error Rate (WER) or confidence scores etc.) with new incoming data using statistical methods. The Kolmogorov-Smirnov (KS) test, a non-parametric method, detects changes in continuous distributions, such as confidence scores or WER. This is useful in detecting model drifts, particularly when confidence scores decrease over time, suggesting that the ASR model is struggling with unfamiliar accents, increased background noise, or new vocabulary.

On the other hand, Jensen-Shannon Divergence (JSD) is effective in identifying data drift by comparing the probability distributions of categorical features, such as accent distributions, noise profiles, or speaker demographics. For instance, suppose an ASR system was trained on a dataset with a balanced mix of British, American, and Indian accents, but new data shows a predominance of South African accents, which were not well-represented during training. While this data drift itself does not guarantee model drift, JSD can help measure the extent of the shift and assess whether this affects ASR performance.

Beyond statistical monitoring, maintaining a shadow dataset of ground truth transcripts with diverse acoustic features allows benchmarking against past performance. If WER increases for specific demographics or conditions, active learning strategies can prioritize these cases for annotation and retraining. For instance, confidence scores from the ASR model can be used to identify low-confidence transcriptions and send them for human review and correction. Once human annotators provide corrected transcripts, these transcripts can be added to a gold-standard dataset used to fine tune the ASR model.