Data

Dataset
(BGL or HDFS)

Preprocessing

Tokenization and cleaning → Sentence filtering and shuffling → Normalization (scaled zero to one) → Partition dataset (k-fold)

Generating Set of Anomalies

First autoencoder (Creates Label set) → Decision Forest (generated from label set) → Isolates anomalies (remove normal logs)

Detecting anomalies

Second autoencoder (feature extracts anomalies) → Model testing

Deployment

Apply model → Integrate into Kafka

Possible libraries:
Tensorflow (if not Pytorch)
Sklearn
Pandas
Math Plot Lib