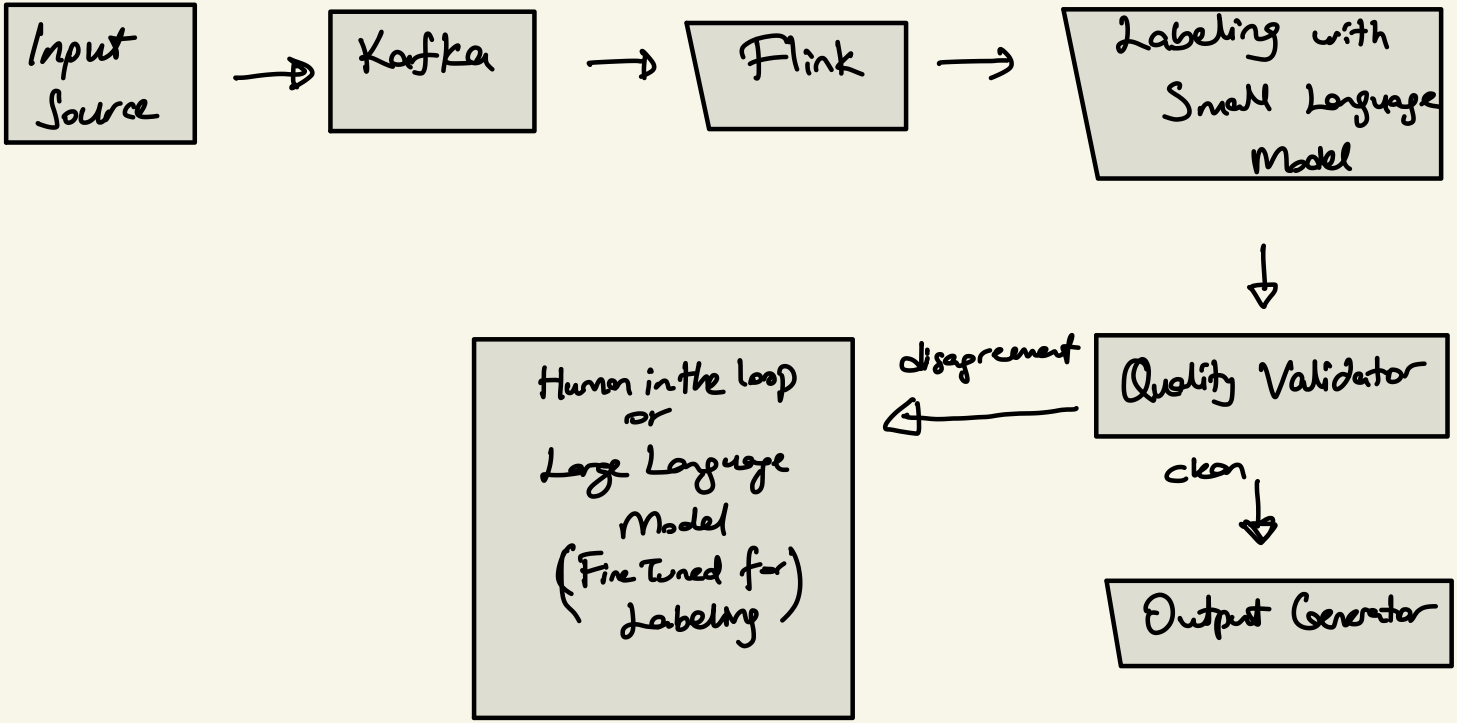
Input Source
Kafka
Flink
Labeling with Small Language Model
Quality Validator
disagreement
Human in the loop
or
Large Language Model
(Fine Tuned for Labeling)
clean
Output Generator

Raw Data Source

↓

Ingestion

Kafka

{ I would use Apache Kafka since it provides continuous streaming]

↓

Streaming Data

Apache Flink

{ Even though Apache Spark can be an option here, it supports micro batches. On the other hand, Flink supports real time streaming with low latency. Schema Registry could provide data integrity.]

↓

Labeling

Fine-Tuned Small Language Model(s)

{ I would use a SLM such as Llama 3.2 3B, to label coming text, with confidence score. model version and prompt must be stored for governance.
I may fine tune the model, if its performance is not satisfied.]

↓

Storing Labeled Data

Object Store / Delta Lake

I would prefer to use S3 Object Store to put output file since it is quite cheap. Using delta lake we can have versions of our files.

↓

Quality Validator & Output Generator

Python

Python has strong library ecosystem for processing data. It supports both file reading, Json format and db connections. Python makes the logic easy to review. Python is the de facto standard for data validation and dataset preparation.