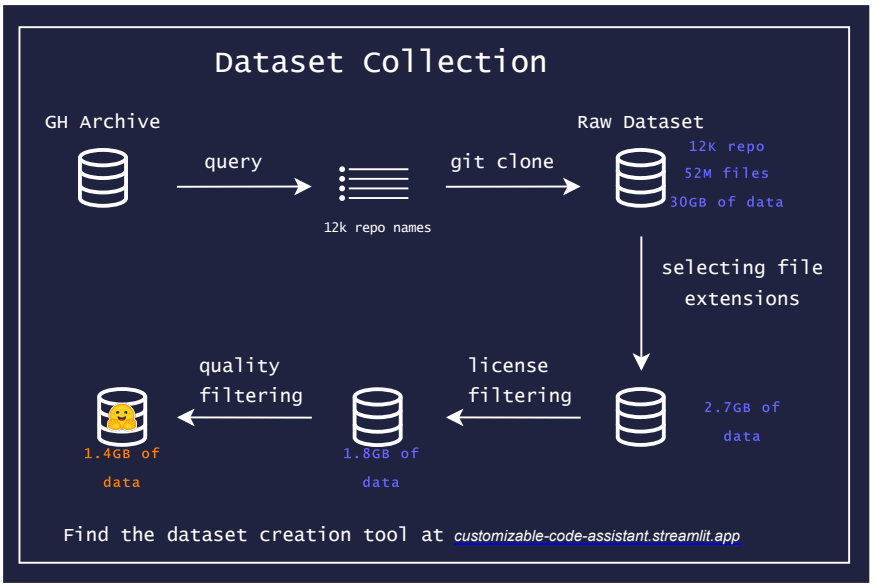
Dataset Collection
GH Archive
query
12k repo names
git clone
Raw Dataset
12K repo
52M files
30GB of data
selecting file extensions
2.7GB of data
license filtering
1.8GB of data
quality filtering
1.4GB of data
Find the dataset creation tool at customizable-code-assistant.streamlit.app