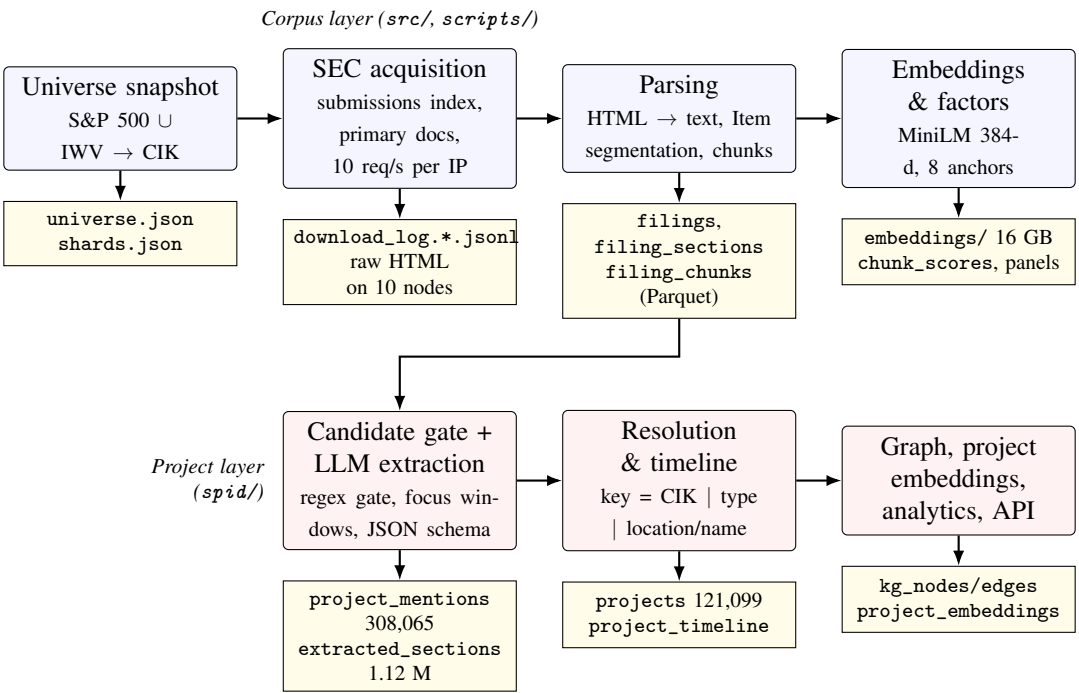

Corpus layer (src/, scripts/)
Universe snapshot
S&P 500 ∪
IWV → CIK
universe.json
shards.json
SEC acquisition
submissions index,
primary docs,
10 req/s per IP
download_log.*.jsonl
raw HTML
on 10 nodes
Parsing
HTML → text, Item
segmentation, chunks
filings,
filing_sections
filing_chunks
(Parquet)
Embeddings
& factors
MiniLM 384-
d, 8 anchors
embeddings/ 16 GB
chunk_scores, panels
Project layer
(spid/)
Candidate gate +
LLM extraction
regex gate, focus win-
dows, JSON schema
project_mentions
308,065
extracted_sections
1.12 M
Resolution
& timeline
key = CIK | type
| location/name
projects 121,099
project_timeline
Graph, project
embeddings,
analytics, API
kg_nodes/edges
project_embeddings