

Combining
Apache Beam
and
Apache Tika
can help to ingest
the content from the files
in most known formats.