TikaWithoutFiles

When does Tika need/create a File rather than an InputStream?

Detection

Tika will be able to do mime-magic detection, for the formats where that's possible, with just a re-windable / mark+reset `InputStream`. Generally Tika uses the first 64kb of the stream for this, but the first 4kb should be enough for the majority of the mime-magic matches

For some file formats, generally Container Based formats, Tika needs the whole of the File (plus the Tika Parsers package!) to be able to correctly detect the file type. This is because some formats only have mime-magic which identify the Container itself (eg Zip), and you need to look inside the Container to find the specific subtype (eg XLSX = `application/vnd.openxmlformats-officedocument.spreadsheetml.sheet`). For this either the File object is needed, or Tika would need to spool the stream out to a Temporary File.

It is best to make use of `TikaInputStream.get(File)` or `TikaInputStream.get(InputStream)` to have your stream/file wrapped for possible temp file creation / file access, if possible

Note - Tika should be given the filename if possible, to help guide mime-magic and container detection

Parsing

Some formats require a File to be able to be parsed. This is because of restrictions in the underlying Java libraries being used

Some formats work better with a File when parsing. This is because it permits lower memory use, as back-and-forth searching can be handled without buffering the whole stream

As of Apache Tika 1.17

As of Apache Tika 1.17 (note - other versions may differ because of changes in the underlying Java libraries), the following formats require Files or create temporary Files if not:

jpeg, zip (for detection) and derived (docx, xlsx, pptx), ole2 (for detection) and derived (doc, xls, ppt), mdb, pst, rar, 7zip, sqlite...