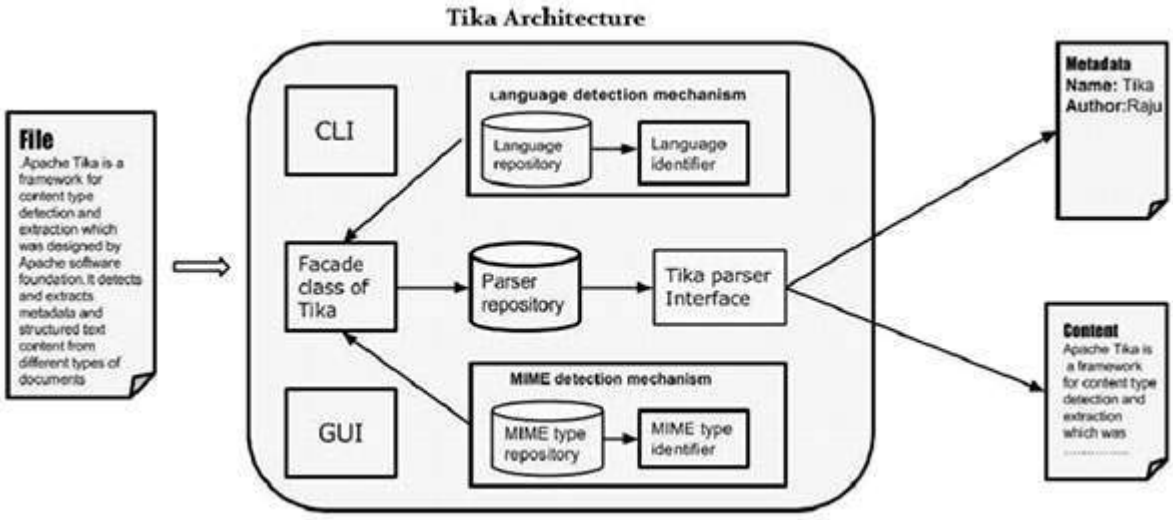
TIKA - ARCHITECTURE

Application-Level Architecture of Tika

Application programmers can easily integrate Tika in their applications. Tika provides a Command Line Interface and a GUI to make it user friendly.

In this chapter, we will discuss the four important modules that constitute the Tika architecture. The following illustration shows the architecture of Tika along with its four modules:

- Language detection mechanism.
- MIME detection mechanism.
- Parser interface.
- Tika Facade class.



Language Detection Mechanism

Whenever a text document is passed to Tika, it will detect the language in which it was written. It accepts documents without language annotation and adds that information in the metadata of the document by detecting the language.

To support language identification, Tika has a class called **Language Identifier**in the package **org.apache.tika.language**, and a language identification repository inside which contains algorithms for language detection from a given text. Tika internally uses N-gram algorithm for language detection.

MIME Detection Mechanism

Tika can detect the document type according to the MIME standards. Default MIME type detection in Tika is done using org.apache.tika.mime.mimeTypes. It uses the org.apache.tika.detect.Detector interface for most of the content type detection.

Internally Tika uses several techniques like file globs, content-type hints, magic bytes, character encodings, and several other techniques.

Parser Interface

The parser interface of org.apache.tika.parser is the key interface for parsing documents in Tika. This Interface extracts the text and the metadata from a document and summarizes it for external users who are willing to write parser plugins.

Using different concrete parser classes, specific for individual document types, Tika supports a lot of document formats. These format specific classes provide support for different document formats, either by directly implementing the parser logic or by using external parser libraries.

Tika Facade Class

Using Tika facade class is the simplest and direct way of calling Tika from Java, and it follows the facade design pattern. You can find the Tika facade class in the org.apache.tika package of Tika API.

By implementing basic use cases, Tika acts as a broker of landscape. It abstracts the underlying complexity of the Tika library such as MIME detection mechanism, parser interface, and language detection mechanism, and provides the users a simple interface to use.

Features of Tika

- **Unified parser Interface** : Tika encapsulates all the third party parser libraries within a single parser interface. Due to this feature, the user escapes from the burden of selecting the suitable parser library and use it according to the file type encountered.
- **Low memory usage** : Tika consumes less memory resources therefore it is easily embeddable with Java applications. We can also use Tika within the application which run on platforms with less resources like mobile PDA.
- **Fast processing** : Quick content detection and extraction from applications can be expected.
- **Flexible metadata** : Tika understands all the metadata models which are used to describe files.
- **Parser integration** : Tika can use various parser libraries available for each document type in a single application.
- **MIME type detection** : Tika can detect and extract content from all the media types included in the MIME standards.
- **Language detection** : Tika includes language identification feature, therefore can be used in documents based on language type in a multi lingual websites.

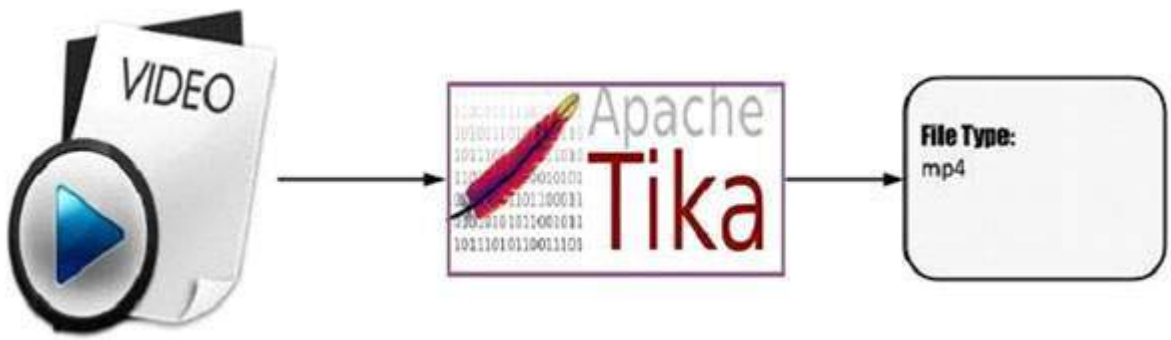
Functionalities of Tika

Tika supports various functionalities:

- Document type detection
- Content extraction
- Metadata extraction
- Language detection

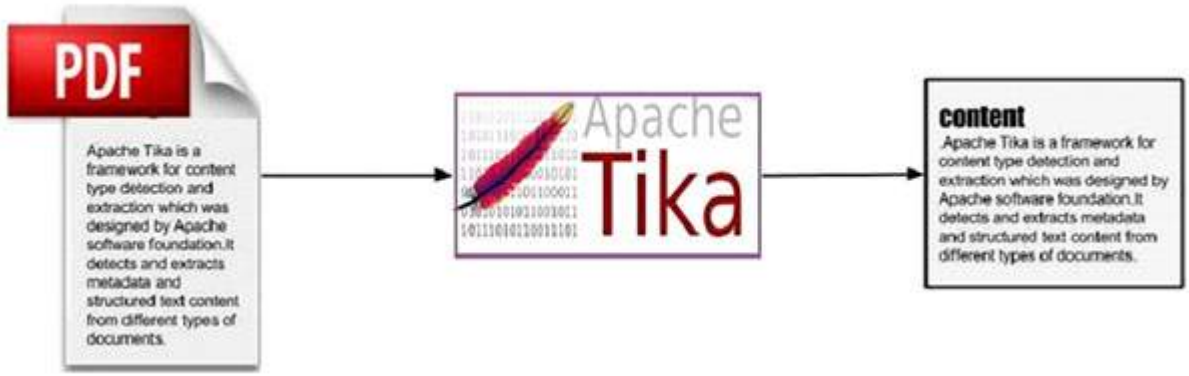
Document Type Detection

Tika uses various detection techniques and detects the type of the document given to it.



Content Extraction

Tika has a parser library that can parse the content of various document formats and extract them. After detecting the type of the document, it selects the appropriate parser from the parser repository and passes the document. Different classes of Tika have methods to parse different document formats.



Metadata Extraction

Along with the content, Tika extracts the metadata of the document with the same procedure as in content extraction. For some document types, Tika have classes to extract metadata.

Language Detection

Internally, Tika follows algorithms like n-gram to detect the language of the content in a given document. Tika depends on classes like Language identifier and Profiler for language identification.