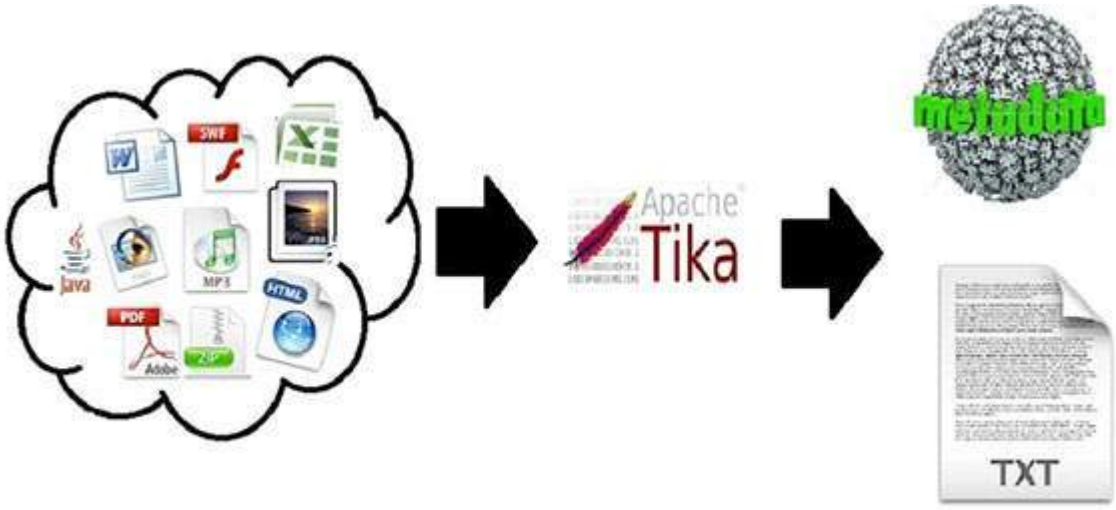
TIKA - OVERVIEW

What is Apache Tika?

- Apache Tika is a library that is used for document type detection and content extraction from various file formats.
- Internally, Tika uses various existing document parsers and document type detection techniques to detect and extract data.
- Using Tika, one can develop a universal type detector and content extractor to extract both structured text as well as metadata from different types of documents such as spreadsheets, text documents, images, PDFs and even multimedia input formats to a certain extent.
- Tika provides a single generic API for parsing different file formats. It uses 83 existing specialized parser ibraries for each document type.
- All these parser libraries are encapsulated under a single interface called the **Parser interface**.



Why Tika?

According to filext.com, there are about 15k to 51k content types, and this number is growing day by day. Data is being stored in various formats such as text documents, excel spreadsheet, PDFs, images, and multimedia files, to name a few. Therefore, applications such as search engines and content management systems need additional support for easy extraction of data from these document types. Apache Tika serves this purpose by providing a generic API to detect and extract data from multiple file formats.

Apache Tika Applications

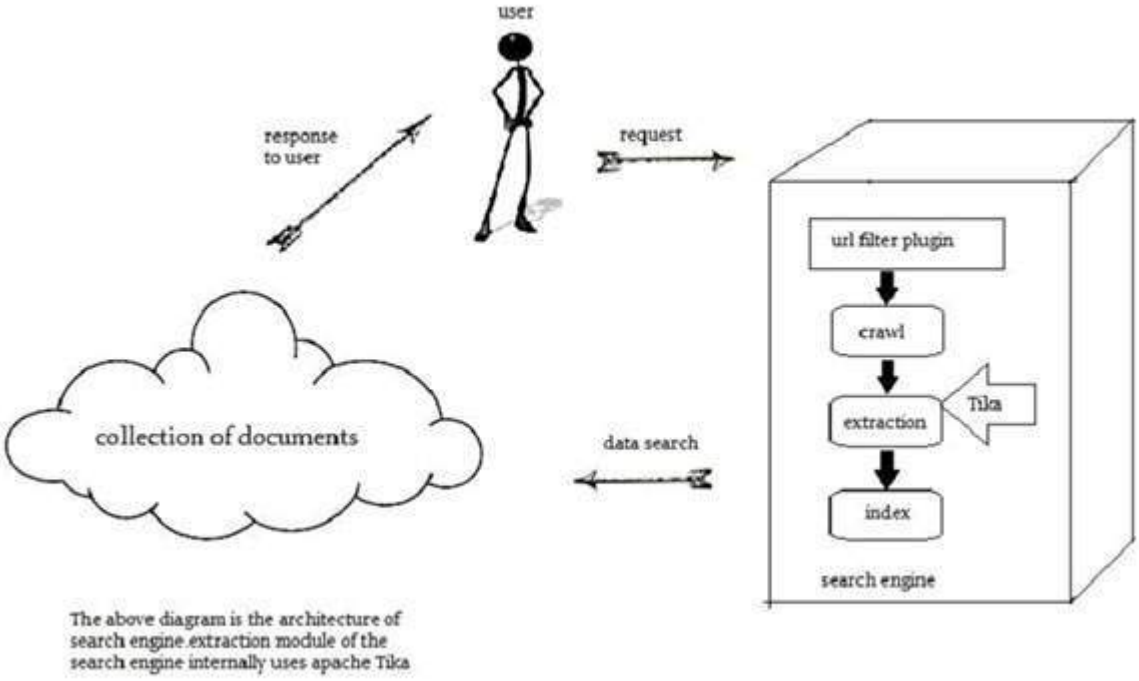
There are various applications that make use of Apache Tika. Here we will discuss a few prominent applications that depend heavily on Apache Tika.

Search Engines

Tika is widely used while developing search engines to index the text contents of digital documents.

- Search engines are information processing systems designed to search information and indexed documents from the Web.
- Crawler is an important component of a search engine that crawls through the Web to fetch the documents that are to be indexed using some indexing technique. Thereafter, the crawler transfers these indexed documents to an extraction component.

- The duty of extraction component is to extract the text and metadata from the document. Such extracted content and metadata are very useful for a search engine. This extraction component contains Tika.
- The extracted content is then passed to the indexer of the search engine that uses it to build a search index. Apart from this, the search engine uses the extracted content in many other ways as well.



The above diagram is the architecture of search engine extraction module of the search engine internally uses apache Tika

Document Analysis

- In the field of artificial intelligence, there are certain tools to analyze documents automatically at semantic level and extract all kinds of data from them.
- In such applications, the documents are classified based on the prominent terms in the extracted content of the document.
- These tools make use of Tika for content extraction to analyze documents varying from plain text to digital documents.

Digital Asset Management

- Some organizations manage their digital assets such as photographs, e-books, drawings, music and video using a special application known as digital asset management DAM.
- Such applications take the help of document type detectors and metadata extractor to classify the various documents.

Content Analysis

- Websites like Amazon recommend newly released contents of their website to individual users according to their interests. To do so, these websites follow **machine learning techniques**, or take the help of social media websites like Facebook to extract required information such as likes and interests of the users. This gathered information will be in the form of html tags or other formats that require further content type detection and extraction.
- For content analysis of a document, we have technologies that implement machine learning techniques such as **UIMA** and **Mahout**. These technologies are useful in clustering and analyzing the data in the documents.

- **Apache Mahout** is a framework which provides ML algorithms on Apache Hadoop – a cloud computing platform. Mahout provides an architecture by following certain clustering and filtering techniques. By following this architecture, programmers can write their own ML algorithms to produce recommendations by taking various text and metadata combinations. To provide inputs to these algorithms, recent versions of Mahout use Tika to extract text and metadata from binary content.
- **Apache UIMA** analyzes and processes various programming languages and produces UIMA annotations. Internally it uses Tika Annotator to extract document text and metadata.

History

Year	Development
2006	The idea of Tika was projected before the Lucene Project Management Committee.
2006	The concept of Tika and its usefulness in the Jackrabbit project was discussed.
2007	Tika entered into Apache incubator.
2008	Versions 0.1 and 0.2 were released and Tika graduated from the incubator to the Lucene sub-project.
2009	Versions 0.3, 0.4, and 0.5 were released.
2010	Version 0.6 and 0.7 were released and Tika graduated into the top-level Apache project.
2011	Tika 1.0 was released and the book on Tika "Tika in Action” was also released in the same year.