**A SIMPLE CRAWLER IN JAVA TO CRAWL AND DOWNLOAD ALL MAILS FOR A PARTICULER YEAR**

**Logic behind this simple crawler system in brief:**

At the very first stage, Crawler parse the given url and recessively parse the urls residing in the pages. While parsing the urls recursively it also check for the url if it is already parsed or not. If the url parsed already it did not parse it again and continue with the other urls. While parsing the urls it checks the url format if it is the actual mail url containing the particular year as a substring and not having extra parameter in the url. If url pass matched with the above criteria it store the url in some application cache. Once all the urls will be parsed, Crawler start dumping the cache data to DB as well as in a file.

**System required resources:**

Oracle driver jar : oracle14.jar

Jsoup jar : jsoup-1.9\*.jar

Oracle data base 11g

Table must be created in DB schema scott or whatever we provide in the crawler.conf file

CREATE TABLE T\_CRAWLER\_DATA

(

"F\_COUNTER" NUMBER PRIMARY KEY,

"F\_MAIL\_URL" VARCHAR2(3000)

)

**crawler.conf**

Above file has the DB, URL and Year configurable values which we read in our java program.

**Jsoup library**

Jsoup is *a java html parser*. It is a java library that is used to parse HTML document. Jsoup provides api to extract and manipulate data from URL or HTML file. It uses DOM, CSS and Jquery-like methods for extracting and manipulating file.

High level design diagram for the system

File

http://\*\*2015

[http://\*\*2015](http://**2015)

………

DB

Starting url