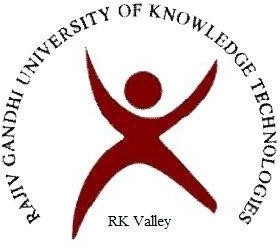
#### A Project Report on

**Web Scraping**

##### Submitted by

Nalla Parvathi R161459

##### Submitted to

****

Under the supervision of

**Ms.Suneetha**

**Project Guide**

**Department of Computer Science and Engineering**

**As a part of**

fulfilment of the degree of Bachelor of Technology in Computer Science and Engineering.

**Date: 2-05-2022**

## CERTIFICATE

This is to certify that the report submitted by **Nalla Parvathi** in fulfilment of the requirements for the award of Technology in Computer Science and Engineering is a bonafide work carried out by her under my supervision and guidance.

##### Ms.Suneetha, Ms. C.Ratna Kumari,

Project Coordinator, Head of the Department,

Assistant Professor, Assistant Professor,

Computer Science and Engineering, Computer Science and Engineering,

RGUKT, R.K Valley. RGUKT, R.K Valley.

**Declaration**

I Nalla Parvathi(R161459) hereby declare that this report submitted by me under the guidance and supervision of Ms.Suneetha madam is a bonafide work.

I also declare that it has not been submitted previously in part or in full to this university or other university or institution for the award of any degree or diploma.

Date:21-04-2022, Nalla Parvathi(R161459).

Place: RK Valley.

# Acknowledgments

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and we would not forget to thank them.

We would like to express our special thanks of gratitude to **Mr.Ravindran** (Supporting Manager)for their able guidance and support in completing our Project.

We are grateful to **Ms.Suneetha** for encouragement and more over timely support and gudiance.

At the outset, I would like to thank Rajiv Gandhi University of Knowledge and Technologies, R.K Valley for providing all the necessary resources for the successful completion of my course work.

With Sincere Regards, Nalla Parvathi.

# Abstract

Main objective of Web Scraping is to extract information from one or many websites and process it into simple structures such as spreadsheets, database or CSV file.

The project uses Ruby programming language and YAML data serialization language.

Our project uses different proxy sources,proxy countries and request Type for fetching the pages.

Xpath and Regular expression plays crucial role for extracting the data from the webpage.

We get the csv file that contains all the information like product name,product price, product Description Etc., for all the fetched pages in the website [www.bmstore.co.uk](http://www.bmstores.co.uk/)

# Table of Content

# 1 [Introduction………………………………………………………………………………](#_TOC_250017).7

* 1. Web Scraping[. ……………..7](#_TOC_250016)
     1. [Introduction. …..7](#_TOC_250015)

1. Training…………………………………………………………………………………….7

2.1 Basic Concepts……………………………………………………………………...7

2.1.1 Vim(Editor)…………………………………………………………………….7

2.1.2 Git……………………………………………………………………………………..7

2.1.3 Github…………………………………………………………………………………8

2.1.4 Jenkins………………………………………………………………………….8

* 1. Ruby………………………………………………………………………………..8

2.2.1 Regular Expression…………………………………………………………….9

2.2.2 DOM10

2.2.3 Nokogiri [11](#_TOC_250009)

2.2.4 [Xpath 11](#_TOC_250008)

1. Web crawler 11
   1. Crawling [1](#_TOC_250006)2

3.1.1 Recrawl[. 1](#_TOC_250005)2

3.1.2 FeedCrawl [1](#_TOC_250004)2

* 1. Extraction[. 1](#_TOC_250003)3
  2. Dedup [1](#_TOC_250002)3
  3. Normalization[. 1](#_TOC_250001)3
  4. Upload…………………………………………………………………………………..13

1. S[ource Code ……………………..1](#_TOC_250000)4

4.1 DRL code………………………………………………………………………………..14

4.2 RSS Code………………………………………………………………………………..15

4.3 Diskfetcher Code………………………………………………………………………...17

4.4 Extraction Code………………………………………………………………………….19

1. **Result…………………………………………………………………………………….20-51**
2. **Future Work 52**
3. **Conclusion…………………………………………………………………………………..52**

# 1. Introduction

# 1.1 Web Scraping

#### 1.1.1 Introduction

Web scraping is a technique for extracting information from websites. Web scraping allows you to acquire non-tabular or poorly structured data from websites and convert it into a usable, structured format, such as a .csv file or spreadsheet. Scraping is about more than just acquiring data: it can also help you archive data and track changes to data online.

**2. Training**

**2.1 Basic Topics**

**2.1.1 Vim(Editor)**

It is a text editor similar to notepad, where we can write any text document or any  ruby files. It has its own set of commands to write, edit, or manipulate files in any way.

vim works in three mode

1. Escape mode

2. Insertion mode

3. Visual mode - shift v , use escape to exit this or any mode to go to normal  mode

to create a new file,write command

vim abc.txt -> a file with name abc.txt will be created and will be opened,till now file is in escape mode(i.e. we cant write anything in it),so to bring it in insertion mode so that we can edit the file press i (now your file is ready to edit).

**2.1.2 Git**

Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

**Git Command:**

1)git init

2) git clone

3) git add

4)git status

5) git checkout

6) git new

7)git commit

8)git push

**2.1.3 Github**

We use Github for code review purpose. Github is a free, web-based team software code review tool.github lets you and others work together on projects from anywhere.

**2.****1.4 Jenkins**

 jenkins is the first reviewer of all our codes and scripts. Jenkins makes a code walk-through.

**2.2 Ruby**

It is a general-purpose object oriented programming language.To run any program with ruby , use ruby program\_filename.rb. Everything in Ruby is an object except the blocks but there are replacements too for it i.e procs and lambda. The objective of Ruby’s development was to make it act as a sensible buffer between human programmers and the underlying computing machinery.

It is an interpreted scripting language which means most of its implementations execute instructions directly and freely, without previously compiling a program into machine-language instructions.

#### 2.2.1 Regular Expression

It is a special sequence of symbols that is used to match the string and in technical terms you can say it is a special sequence of characters that helps us match or find another string or a set of string using a specialized syntex held in pattern.

A *regular expression* is a special sequence of characters that helps you match or find other strings or sets of strings using a specialized syntax held in a pattern.

A *regular expression literal* is a pattern between slashes or between arbitrary delimiters followed by %r as follows

line1 = "Cats are smarter than dogs";

line2 = "Dogs also like meat";

if ( line1 =~ /Cats(.\*)/ )

   puts "Line1 contains Cats"

end

if ( line2 =~ /Cats(.\*)/ )

   puts "Line2 contains  Dogs"

End

**output**:Line1 contains Cats

Some of the most important String methods that use regular expressions are sub and gsub, and their in-place variants sub! and gsub!.

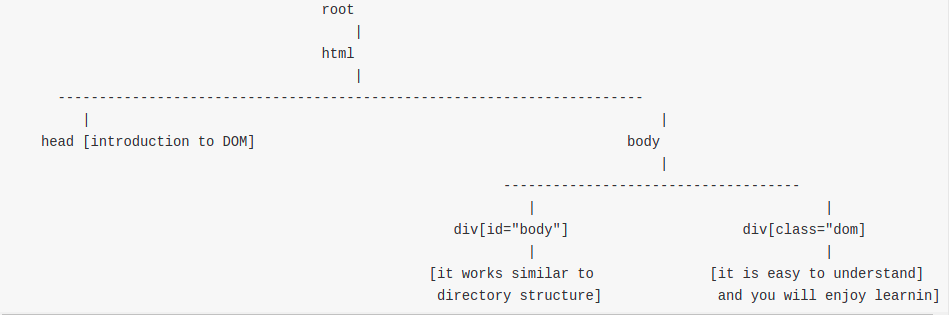
**2.2.2 DOM(Document Object Model)**

The Document Object Model (DOM) is a programming interface for web documents. It represents the page so that programs can change the document structure, style, and content. The DOM represents the document as nodes and objects; that way, programming languages can interact with the page.

dom is a tree-like structure for a given document.

to make more understand about it

* <'html>
* <'head> introduction to DOM <'/head>
* <'body>
* <'div id="body">it works similar to directory structure <'/div>
* <'div class="dom"> it is easy to understand and you will enjoy learning it <'/div>
* <'/body>
* <'/html>
* for the above html code DOM will look like this save this code in the path as below

  
once a tree for a document is formed,it is easy to traverse to a node[i.e head,body,div ..] in a tree.

/html/body/div[@id="body"]

//div[@id="body"]

Both the above commands will produce the same result.

**2.2.3Nokogiri**

Nokogiri makes it easy and painless to work with XML and HTML from Ruby. It provides a sensible, easy-to-understand API for reading, writing, modifying, and querying documents.

Dom is tree like structure for a given document. to make that tree like structure i.e to form. DOM you need a tool called nokogiri that will produce a tree like structure for an HTML or an XML page.

Before you start building a DOM you are required to install 'NOKOGIRI' on your system. once you have installed you can start building DOM and then can traverse through your tree to fetch the required data.

here is a sample code to fetch the data from the html page that lies in /home/demo.html dom= Nokogiri::HTML(File.read("/home/demo.html")) -> this will produce a tree like structure(same as above) for demo.html

str\_node = dom.xpath('/html/body/div[@id="body"]').first ->will return div node whose attribute "id" has value "body"

str = str\_node.content -> return the content that lies between <'div> <'/div> i.e it works similar to directory st

**2.2.4 XPATH**

it is a mechanism to traverse in a DOM to fetch the required data.it is similar to directory structure in our file system.!!

XPath is a technology that uses path expressions to select nodes or node- sets in an XML document (or in our case an HTML document). Even if XPath is not a programming language in itself, it allows you to write expressions that can access directly to a specific HTML element without having to go through the entire HTML tree.

xpath is a perfect tool for web scraping.

**3.Web crawler**

A Web crawler starts with a list of [URLs](https://en.wikipedia.org/wiki/Uniform_Resource_Locator) to visit. Those first URLs are called the *seeds*. As the crawler visits these URLs, by communicating with [web servers](https://en.wikipedia.org/wiki/Web_server) that respond to those URLs, it identifies all the [hyperlinks](https://en.wikipedia.org/wiki/Hyperlink) in the retrieved web pages and adds them to the list of URLs to visit, called the [*crawl frontier*](https://en.wikipedia.org/wiki/Crawl_frontier). URLs from the frontier are [recursively](https://en.wikipedia.org/wiki/Recursion) visited according to a set of policies. If the crawler is performing archiving of [websites](https://en.wikipedia.org/wiki/Website) (or [web archiving](https://en.wikipedia.org/wiki/Web_archiving)), it copies and saves the information as it goes. The archives are usually stored in such a way they can be viewed, read and navigated as if they were on the live web, but are preserved as 'snapshots'.

The archive is known as the *repository* and is designed to store and manage the collection of [web pages](https://en.wikipedia.org/wiki/Web_page). The [repository](https://en.wikipedia.org/wiki/Repository_(version_control)) only stores [HTML](https://en.wikipedia.org/wiki/HTML) pages and these pages are stored as distinct files. A repository is similar to any other system that stores data, like a modern-day database.The number of possible URLs crawled being generated by server-side software has also made it difficult for web crawlers to avoid retrieving [duplicate content](https://en.wikipedia.org/wiki/Duplicate_content). Endless combinations of [HTTP GET](https://en.wikipedia.org/wiki/HTTP) (URL-based) parameters exist, of which only a small selection will actually return unique content.

**3.1 Crawl**

Crawl is the process of fetching a webpage in html format. Here are two stages in pipeline recrawl and Feedcrawl. Recrawl is discovering product URLs from the webpage. Feedcrawl is fetching that product page from the internet in html format.

**3.1.1 Recrawl**

Its purpose is to get almost all urls to be crawled.we will start with seedurl (i.e., https://bmstores.co.uk/ ) for getting all the category urls from the seedurl we need to write xpath, proxy source and request type.we will define a method called **get\_category\_urls** for fetching the corresponding urls.Then returned category urls will get into the next depth.

List of category URLS:

https://www.bmstores.co.uk/products/home-and-furniture/furniture/storage-and-shelving <https://www.bmstores.co.uk/products/toys-and-games/browse-by-age/toys-for-3-4-years> <https://www.bmstores.co.uk/products/food-and-drink/food/breakfast>

Fetching page URLS:

It will take category urls as an input url.For getting all the pagination urls from the category url, we define function **get\_pagination\_urls**.

List of pagination\_urls :

https://www.bmstores.co.uk/products/home-and-furniture/furniture/storage-and-shelving?page=1

https://www.bmstores.co.uk/products/home-and-furniture/furniture/storage-and-shelving?page=2

<https://www.bmstores.co.uk/products/home-and-furniture/furniture/storage-and-shelving?page=3>

Fetching product URLS:

It will take pagination urls as an input url.For getting all the product urls from the pagination url, we define function **get\_product\_urls**.

List of pagination\_urls :

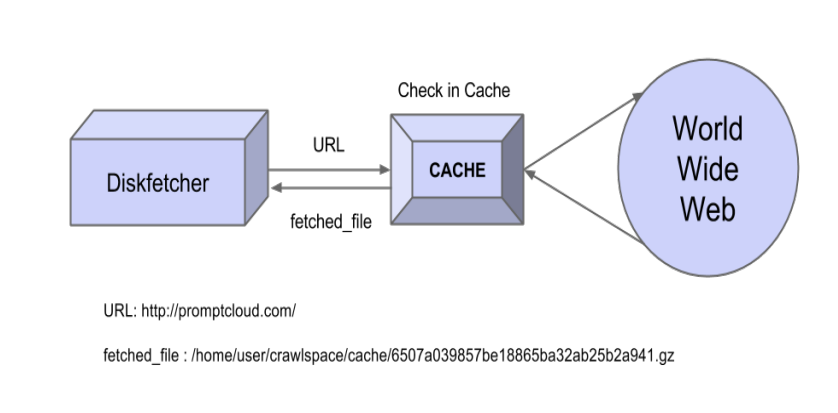
https://www.bmstores.co.uk/products/skinny-crunch-cereal-choco-hoops-and-marshmallows-375g-388554

https://www.bmstores.co.uk/products/back-to-the-future-multigrain-chocolate-balls-cereal-375g-382740

<https://www.bmstores.co.uk/products/skinny-crunch-cereal-375g-388555>

**3.1.2 Feedcrawl**

Diskfetcher is a special type of fetcher which have the following functionality:

* Fetch page and store gzipped page content in the cache directory
* Before fetching a URL from the web it checks whether it is available in the cache or not?
* There is a max\_age(default 24 hours) for the above check, if the fetched file is older than this fetched date then we fetch the page again from the web. The cached file gets overwritten with the fresh content from the web.

**3.2 Extraction**

Extraction is the process of extracting essential information from the crawled HTML page, and form a XML structure of this information.

Fetching product urls

● The final product page that we had fetched through feedcrawl will send it to extraction for extracting the essential information like product\_name,product\_price,product\_description,product\_page\_url,product\_rating and product\_brand etc.

● For every product , we will generate a uniq\_id and crawl\_time\_stamp for it.these are the require fields for every product.For getting each details of the page, we will take the help of xpath.

**3.3 Dedup**

Dedup aka De-duplication is the process of removing duplicate entries(data) from the extracted XML page.

**3.4 Normalization**

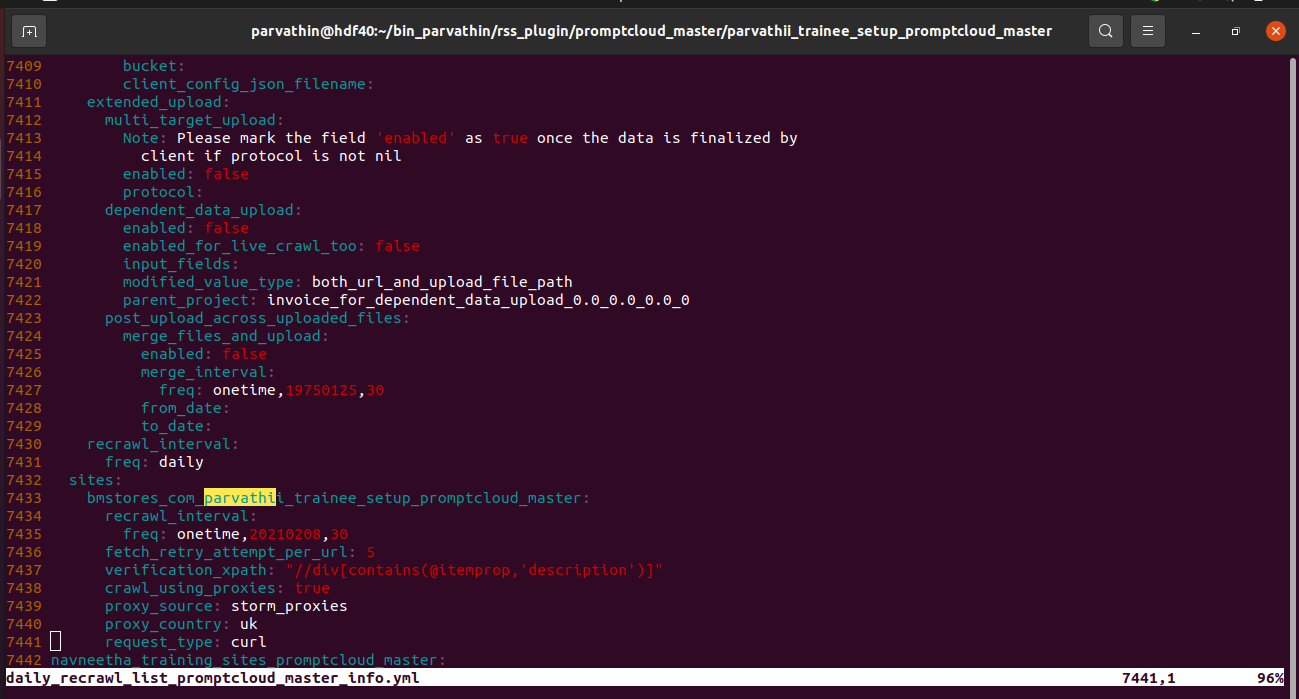
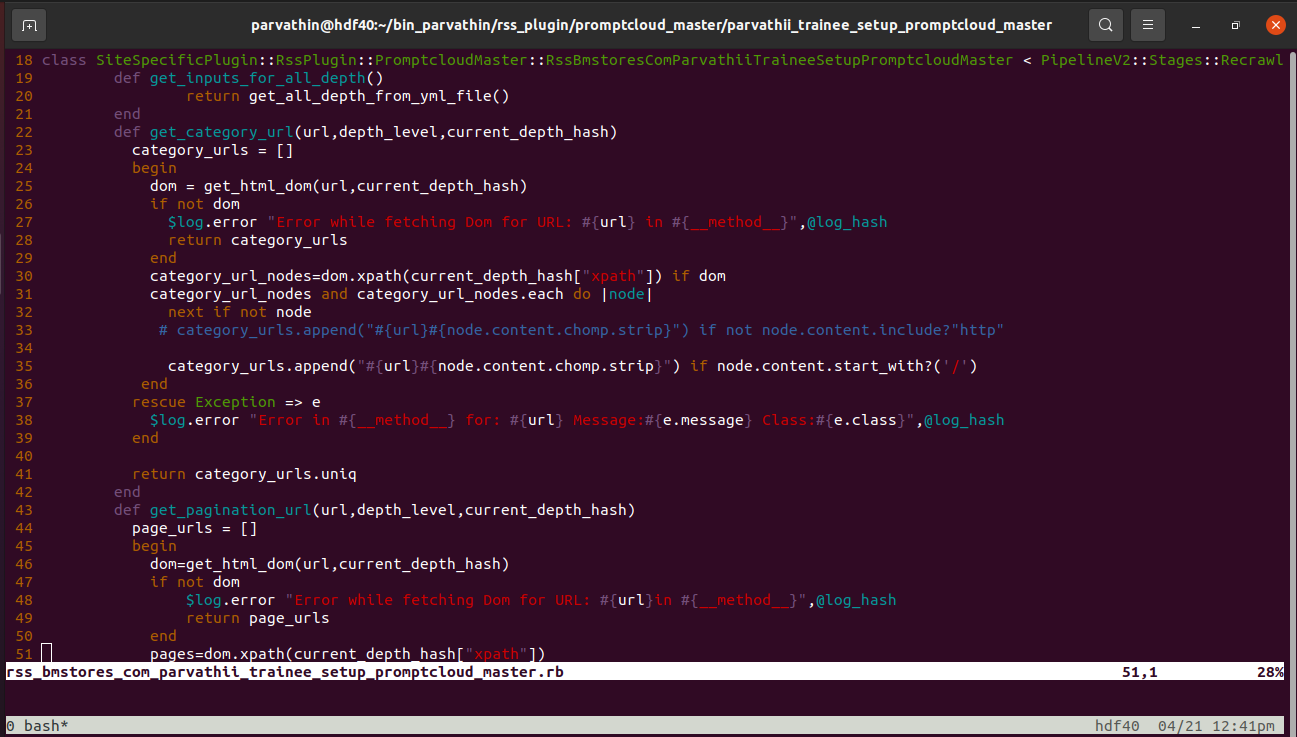
Normalization is the process of transforming the deduped XML according the client requirements, this may include transformation of XML into CSV, etc.

**3.5 Upload**

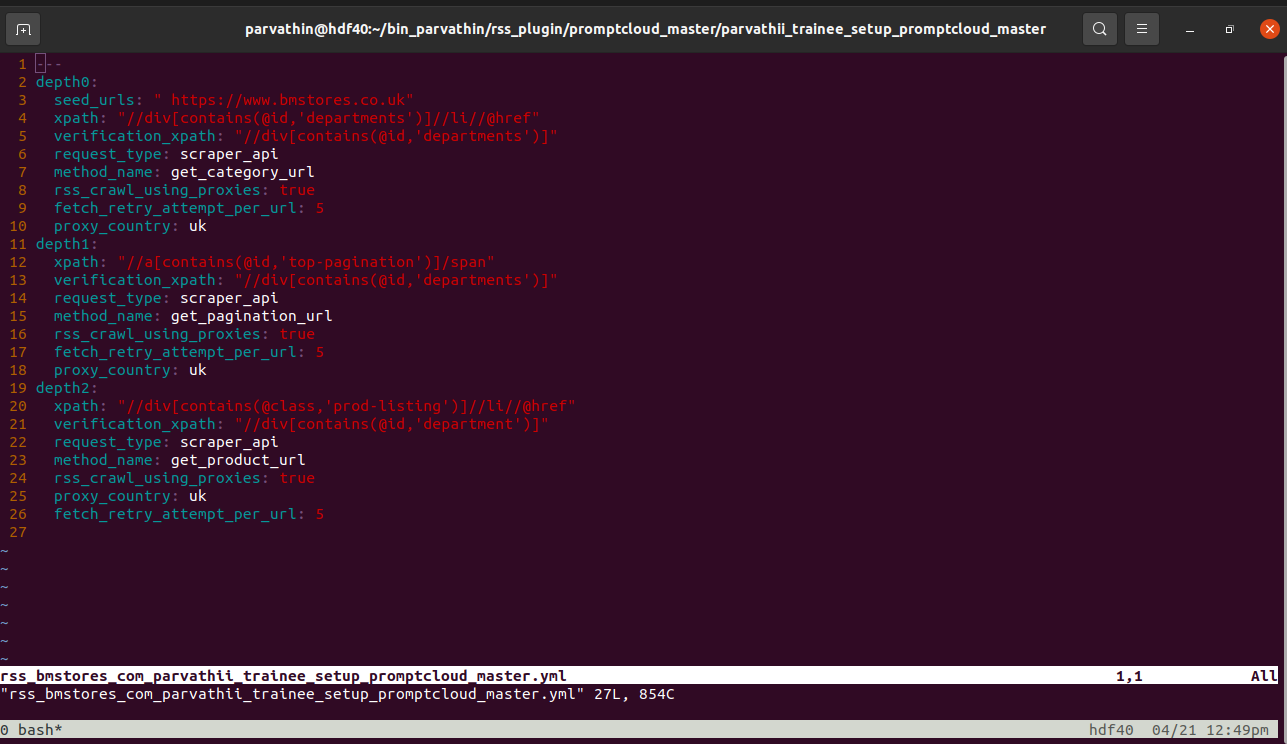
Upload is the process of uploading the normalized XML, CSV, etc data file to the PromptCloud's client API, and after that we are uploading to target location as well based on requirement so that client can access the final result.

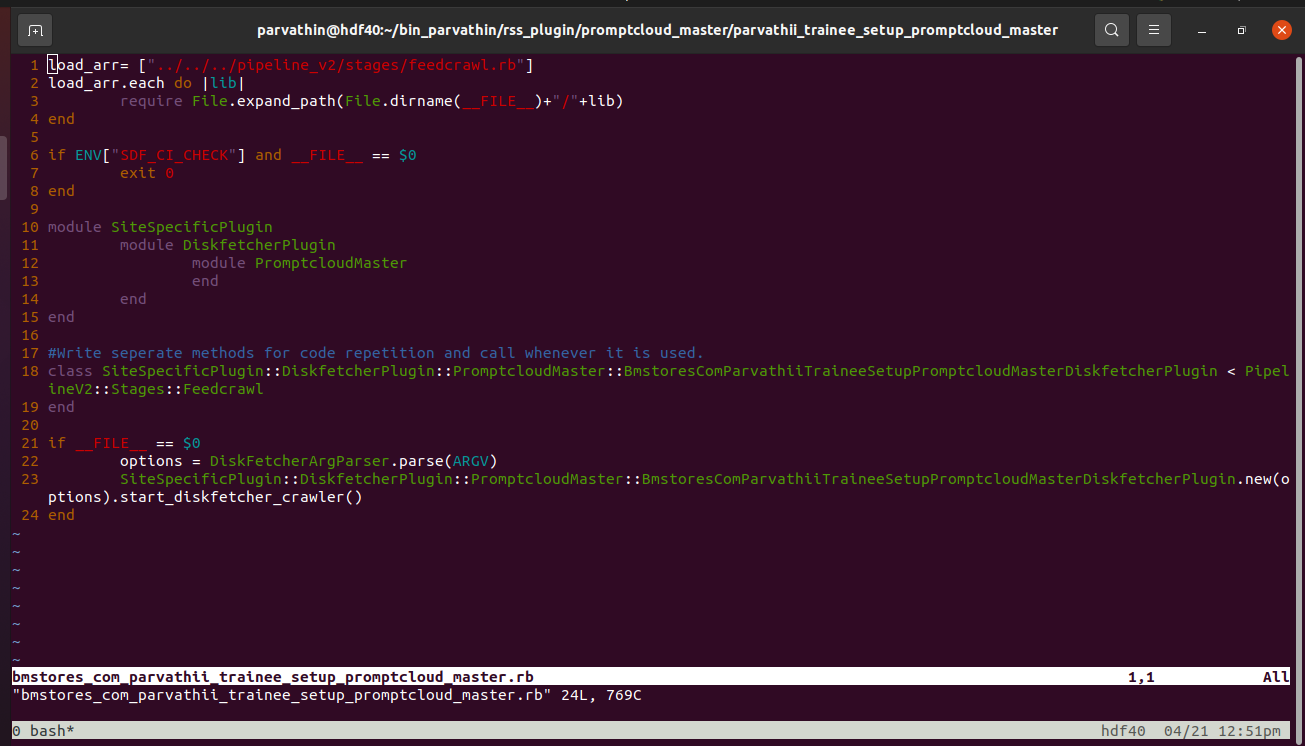
**4.0 Source code**

**DRL Code**

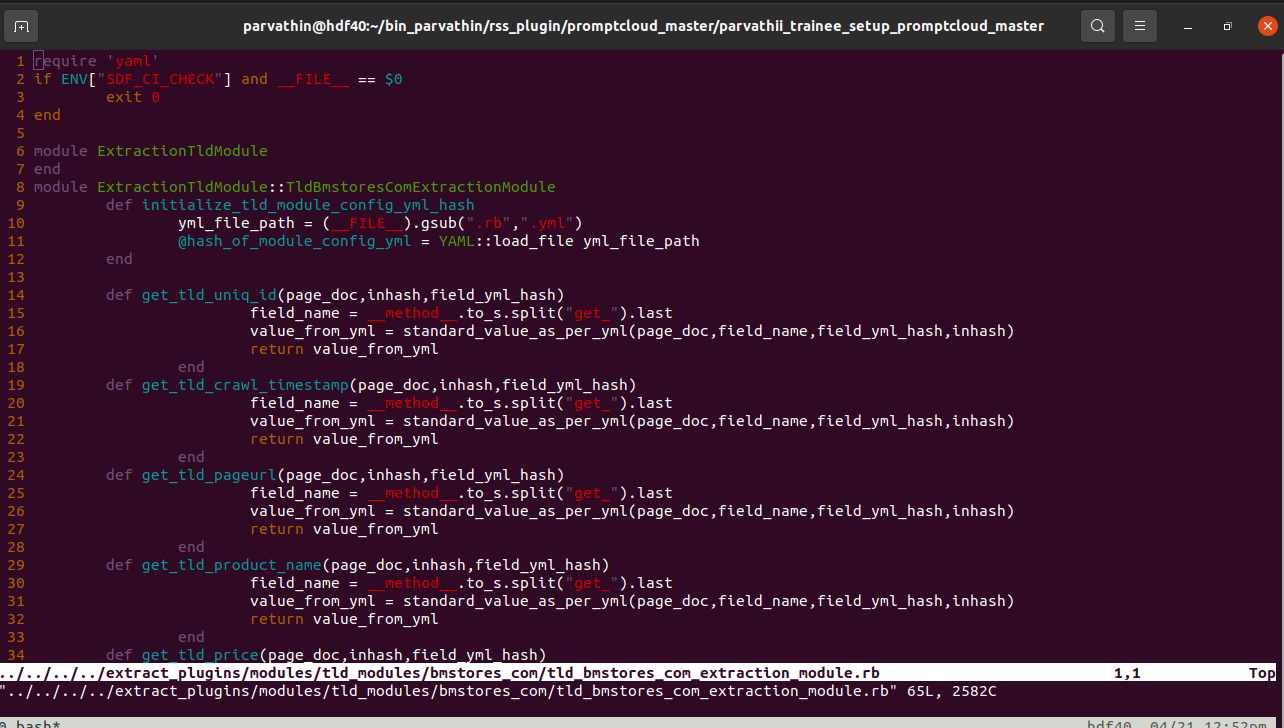
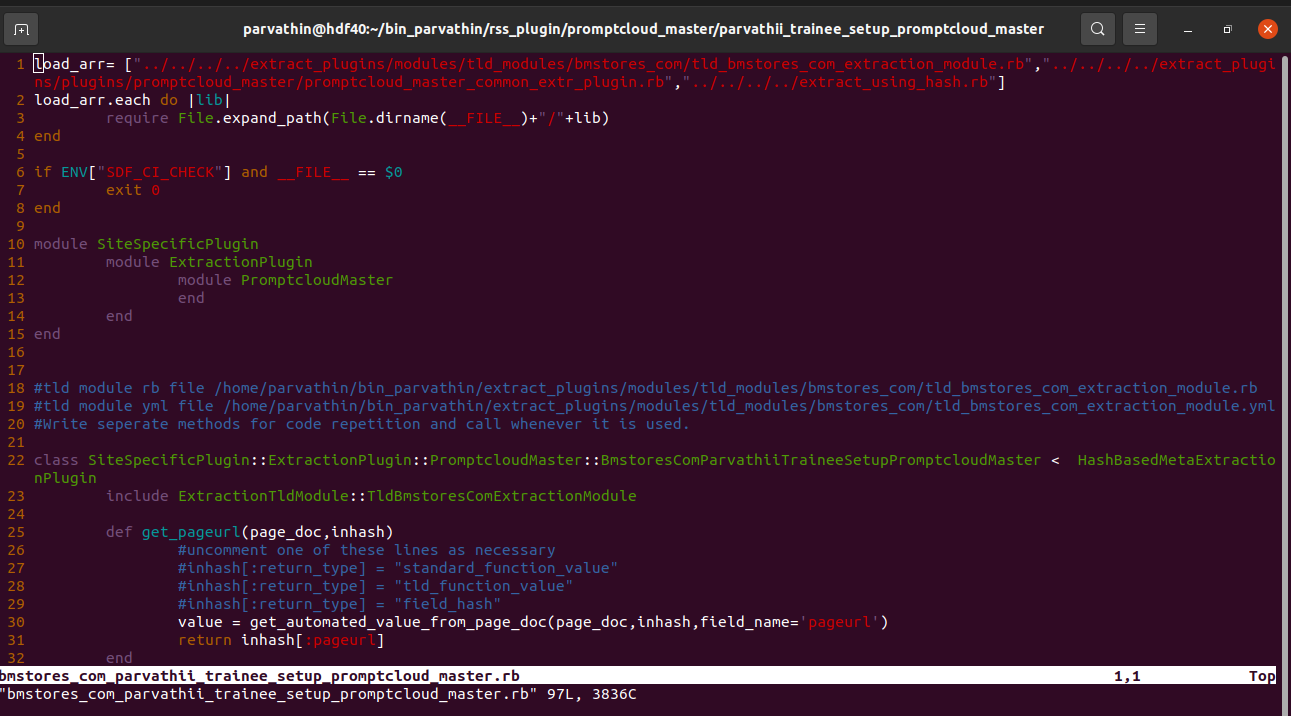
**RSS Code**

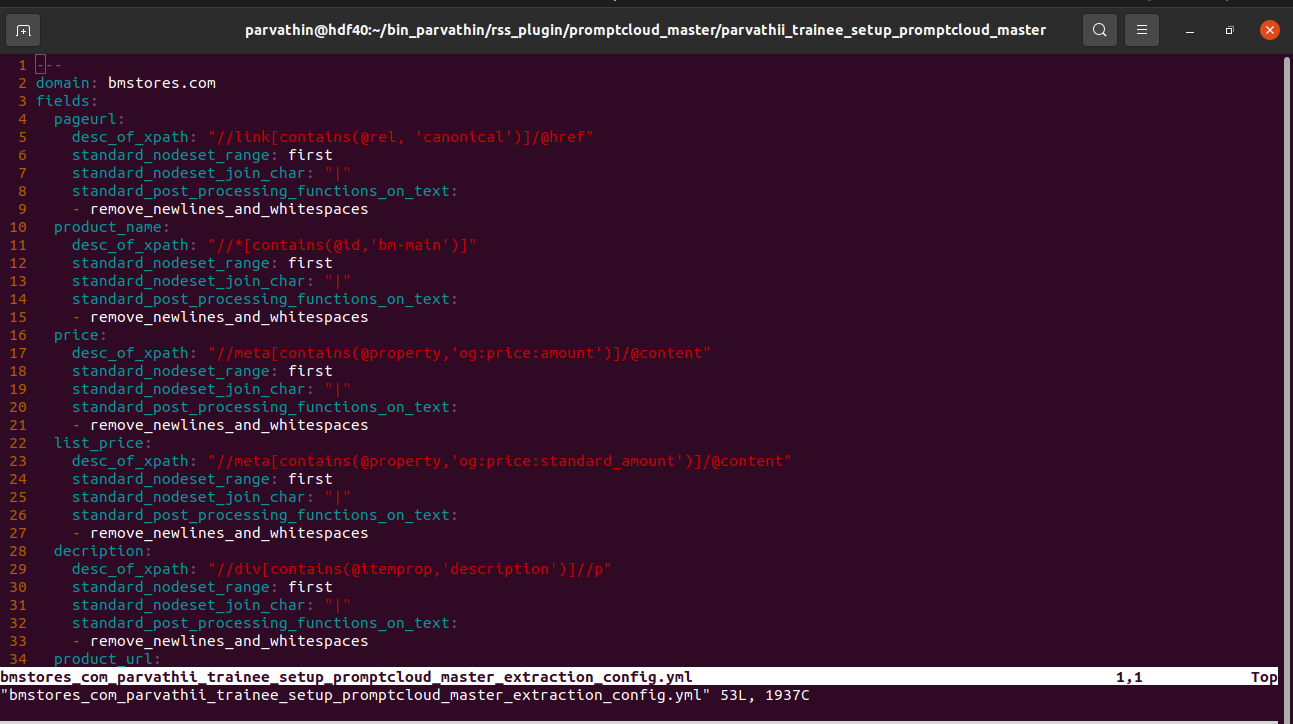
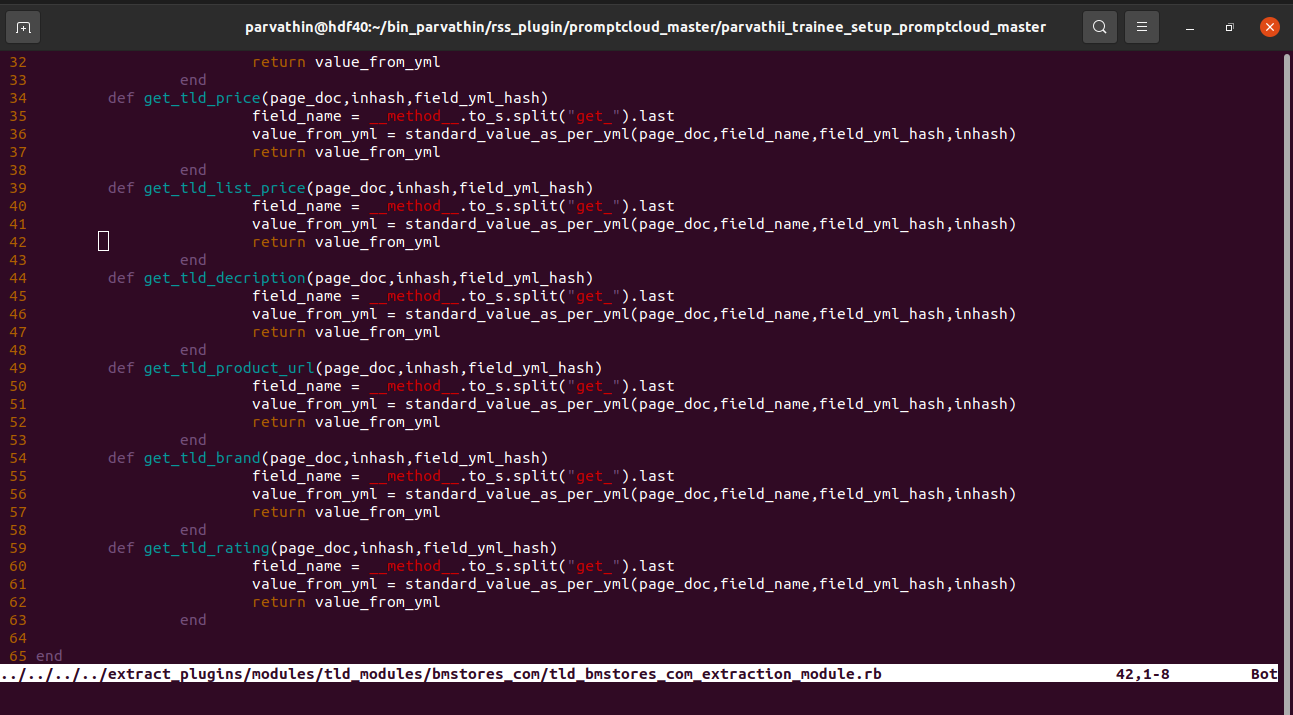


**Dsk code**

****

**Extraction code**





**5.0 RESULT (Output)**

**This page information is extracted output**

**https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625**

{

:msg => " urls are [\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\"]",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "965",

:function\_name => "block(2levels)parse",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "running pipeline in local mode which might need to interact with db or rss\_queue, so it is advised to execute pf.sh in different tab",

:level => "WARN",

:hostname => "hdf40",

:line\_number => "127",

:function\_name => "process",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "parameters received {:site=>\"bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master\", :run\_extraction=>true, :urls=>[\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\"]}",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "168",

:function\_name => "process",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

could not find the config entry for sdf\_hosts->is\_data\_api\_secondary\_machine\_on\_any\_data\_api\_cluster in /home/nbolisetty/bin\_nbolisetty/get\_sdf\_config.rb at 190

{

:msg => ">>> Starting process for site bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master and project parvathin\_training\_sites\_promptcloud\_master <<<\n\n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "174",

:function\_name => "process",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => ">>>> taking urls from command line",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "476",

:function\_name => "run\_extraction",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => ">>> starting diskfetcher <<<\n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "293",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "success till here\n>>> size of uniq urls\_chunk is 1",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "327",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => 10,

:level => "INFO",

:hostname => "hdf40",

:line\_number => "333",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

/home/sdf/.rvm/gems/ruby-2.6.6@20200422\_sdfprod/gems/awesome\_print-1.2.0/lib/awesome\_print/formatter.rb:378: warning: constant ::Fixnum is deprecated

{

:msg => "class\_name of df plugin is SiteSpecificPlugin::DiskfetcherPlugin::PromptcloudMaster::BmstoresComNavneethaTrainingSitesPromptcloudMasterDiskfetcherPlugin\n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "363",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Updating the arguments parameter based on drl entry for bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "143",

:function\_name => "check\_and\_update\_args\_hash\_with\_site\_drl\_info",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "request type for bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master is curl as per drl site hash",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "269",

:function\_name => "get\_valid\_request\_type\_from\_drl",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "max\_pop\_urls\_count: 2000",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "331",

:function\_name => "get\_max\_rss\_queue\_pop\_count",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Argument hash is successfully updated with site\_info from drl",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "84",

:function\_name => "initialize",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Creating a new segment.",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "314",

:function\_name => "get\_outdir\_for\_live\_crawl",

:file\_name => "feedcrawl.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "feedcrawl is valid pipeline\_stage\_name",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "152",

:function\_name => "check\_for\_valid\_pipeline\_stage\_name",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master is valid site",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "159",

:function\_name => "check\_for\_valid\_site\_name",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0 is generated",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "188",

:function\_name => "generate\_uniq\_segment\_id",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "current\_pipeline\_stage\_status\_directory: /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/current\_pipeline\_stage\_status/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "206",

:function\_name => "get\_current\_pipeline\_stage\_status\_directory\_per\_segment\_id",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "creating the directory /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/current\_pipeline\_stage\_status/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "296",

:function\_name => "create\_directory\_if\_not\_present",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "touching file /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/current\_pipeline\_stage\_status//feedcrawl\_started",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "301",

:function\_name => "touch\_file",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "segment\_full\_path: /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "137",

:function\_name => "create\_a\_uniq\_segment\_for\_site\_per\_pipeline\_stage",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "creating the directory /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "296",

:function\_name => "create\_directory\_if\_not\_present",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "creating the directory /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "296",

:function\_name => "create\_directory\_if\_not\_present",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "saving time 2022-04-04 07:26:16 +0000 in file /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/start\_time",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "397",

:function\_name => "save\_current\_time\_in\_file",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "creating the directory /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/segments/20220404072616/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "296",

:function\_name => "create\_directory\_if\_not\_present",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "No class type specified for cloud-type . Returning nil.",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "52",

:function\_name => "get\_klass\_for\_cloud\_type",

:file\_name => "sdf\_temp\_cloud\_machine\_manager\_runner.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "NOTE: As per DRL Request type is curl\n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "398",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "out\_dir is ==> /home/nbolisetty/crawlspace//test\_run\_dir/parvathin\_training\_sites\_promptcloud\_master/bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master\n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "405",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "Running DFv3",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "415",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "WARN:: In diskfetcher: Not a record\_url\_json. Hence failed to parse with MultiJson. Error message: MultiJson::ParseError: 767: unexpected token at 'https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625'",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1314",

:function\_name => "rescueget\_modified\_url\_hash\_per\_record",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "TIMEOUT set per page fetch :: 180",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1132",

:function\_name => "get\_fetch\_result\_hash\_for\_ordinary\_single\_fetch",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => ">>>> max\_age is 86400 <<<<",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1456",

:function\_name => "get\_page\_content\_hash\_from\_cache\_dir",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"mail\_group" => "diskfetcher",

"pipeline\_stage" => "feedcrawl",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "File may not be in cache, fetching from web for url https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625, will put it in the local path /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/segments/20220404072616/bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-78ce3a957c44b72fd13863c129a478ed.gz",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1408",

:function\_name => "get\_page\_content\_hash\_internal\_should\_not\_be\_overridden\_in\_child\_classes",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:16+00:00",

"mail\_group" => "diskfetcher",

"pipeline\_stage" => "feedcrawl",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "Different Proxy Sources are => [\"premium\_proxy\", \"tor\_proxy\", \"docker\_proxy\", \"spys\_proxy\", \"proxynova\_proxy\", \"cool\_proxy\", \"us\_proxy\_org\", \"free\_proxy\", \"didsoft\_premium\_proxy\", \"proxy\_rotator\_proxy\", \"micro\_leaves\_non\_amazon\_proxies\", \"micro\_leaves\_amazon\_proxies\", \"storm\_proxies\", \"rotating\_proxies\", \"rack\_proxies\", \"flipnode\_proxies\", \"scraper\_api\_proxies\"]",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "71",

:function\_name => "get\_all\_proxy\_sources",

:file\_name => "proxy\_utility\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Loading all\_other\_proxies\_hash",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_other\_proxies\_hash",

:file\_name => "proxy\_utility\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Fetching storm\_proxies: uk",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "727",

:function\_name => "get\_countrywise\_storm\_proxy\_list",

:file\_name => "proxy\_utility\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Proxy\_source: storm\_proxies",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "284",

:function\_name => "get\_customized\_proxy\_list",

:file\_name => "proxy\_utility\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "customized\_proxy\_list: []",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "285",

:function\_name => "get\_customized\_proxy\_list",

:file\_name => "proxy\_utility\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Fetching storm\_proxies: uk",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "727",

:function\_name => "get\_countrywise\_storm\_proxy\_list",

:file\_name => "proxy\_utility\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Max\_retry\_attempt per url => 5",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1452",

:function\_name => "get\_max\_retry\_count",

:file\_name => "sdf\_fetch\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "\*\*\*\*\*\*\*\*Fetching Page using CURL request URL-> https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\*\*\*\*\*\*\*\*\*",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "938",

:function\_name => "get\_page\_content\_hash\_by\_curl\_request\_internal\_should\_not\_be\_overridden\_in\_child\_classes",

:file\_name => "sdf\_fetch\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

{

:msg => "Using Proxy server => ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "953",

:function\_name => "get\_page\_content\_hash\_by\_curl\_request\_internal\_should\_not\_be\_overridden\_in\_child\_classes",

:file\_name => "sdf\_fetch\_module.rb",

:log\_ts => "2022-04-04T07:26:16+00:00"

}

% Total % Received % Xferd Average Speed Time Time Time Current

Dload Upload Total Spent Left Speed

100 957k 100 957k 0 0 2130k 0 --:--:-- --:--:-- --:--:-- 2127k

{

:msg => "Checking verification\_xpath => //div[contains(@itemprop,'description')]",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "877",

:function\_name => "is\_page\_fetch\_valid?",

:file\_name => "sdf\_fetch\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "Checking verification\_xpath => //div[contains(@itemprop,'description')]",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "877",

:function\_name => "is\_page\_fetch\_valid?",

:file\_name => "sdf\_fetch\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "Segment\_directory ==>> /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/segments/20220404072616/",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1535",

:function\_name => "save\_to\_disk",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"mail\_group" => "diskfetcher",

"pipeline\_stage" => "feedcrawl",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran",

"module" => "SdfPageFetchModule",

"hostname" => "hdf40",

"url" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625"

}

{

:msg => "File\_path is /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/segments/20220404072616/bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-78ce3a957c44b72fd13863c129a478ed.gz",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1542",

:function\_name => "save\_to\_disk",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"mail\_group" => "diskfetcher",

"pipeline\_stage" => "feedcrawl",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran",

"module" => "SdfPageFetchModule",

"hostname" => "hdf40",

"url" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625"

}

{

:msg => "Adaptive crawl obj is nil. Nothing to do in update\_adaptive\_crawl\_db()",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1748",

:function\_name => "update\_adaptive\_crawl\_db",

:file\_name => "diskfetcher\_base\_v3.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"mail\_group" => "diskfetcher",

"pipeline\_stage" => "feedcrawl",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran",

"module" => "SdfPageFetchModule",

"hostname" => "hdf40",

"url" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625"

}

{

:msg => "Fetch store: {\"bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-78ce3a957c44b72fd13863c129a478ed.gz\"=>\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\"}",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "437",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "Meta store: {\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\_pipeline\_entry\"=>{}}",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "438",

:function\_name => "run\_diskfetcher",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "NOTE :=> Records in meta\_store are {\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\_pipeline\_entry\"=>{}}",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "501",

:function\_name => "run\_extraction",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => ">>>>starting extracion <<<<<< \n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "502",

:function\_name => "run\_extraction",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "fetched\_file\_path =>/home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/feedcrawl/segments/20220404072616/bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-78ce3a957c44b72fd13863c129a478ed.gz \n",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "517",

:function\_name => "blockrun\_extraction",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

/home/sdf/.rvm/gems/ruby-2.6.6@20200422\_sdfprod/gems/awesome\_print-1.2.0/lib/awesome\_print/formatter.rb:378: warning: constant ::Fixnum is deprecated

{

:msg => {

:mode => "Test",

:path => "/home/nbolisetty/bin\_nbolisetty/extract\_plugins/plugins/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master.rb"

},

:level => "INFO",

:hostname => "hdf40",

:line\_number => "1417",

:function\_name => "get\_extr\_plugin\_from\_file",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

/home/sdf/.rvm/gems/ruby-2.6.6@20200422\_sdfprod/gems/awesome\_print-1.2.0/lib/awesome\_print/formatter.rb:378: warning: constant ::Fixnum is deprecated

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "Cound not find valid\_field\_qa\_functions for any of the fields for bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master. Returning empty hash",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4060",

:function\_name => "get\_valid\_field\_qa\_function\_map\_for\_site",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "included\_modules [ExtractionTldModule::TldBmstoresComExtractionModule, SDF, RabbitMqClusterMapping, DataApiClusterMapping, SdfEnv, ExtractionHelperMethodsModule, DataQaMethodsModule, ActiveSupport::ToJsonWithActiveSupportEncoder, Nori::CoreExt::Object, ActiveSupport::Tryable, ProxyUtilityModule, AlertsUtils, JSON::Ext::Generator::GeneratorMethods::Object, PP::ObjectMixin, Kernel]",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "119",

:function\_name => "initialize",

:file\_name => "plugin\_and\_wtplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "going to initialize included module ExtractionTldModule::TldBmstoresComExtractionModule yml hash",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "122",

:function\_name => "blockinitialize",

:file\_name => "plugin\_and\_wtplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master -> extended\_upload -> dependent\_data\_upload -> enabled: false",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "134",

:function\_name => "initialize",

:file\_name => "dependent\_data\_upload\_runner.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master -> extended\_upload -> dependent\_data\_upload -> enabled\_for\_live\_crawl\_too: false",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "142",

:function\_name => "initialize",

:file\_name => "dependent\_data\_upload\_runner.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "returning /home/nbolisetty/bin\_nbolisetty/daily\_recrawl\_lists/projects//promptcloud\_master/daily\_recrawl\_list\_promptcloud\_master\_info.yml",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "2836",

:function\_name => "get\_drl\_file\_for\_client",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "site\_dir\_hierarchy is promptcloud\_master , parvathin\_training\_sites\_promptcloud\_master , bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "2962",

:function\_name => "convert\_site\_dir\_hierarchy\_to\_pipeline\_paths\_and\_vice\_versa",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "returning /home/nbolisetty/bin\_nbolisetty/daily\_recrawl\_lists/projects//promptcloud\_master/daily\_recrawl\_list\_promptcloud\_master\_info.yml",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "2836",

:function\_name => "get\_drl\_file\_for\_client",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "schema is like",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "98",

:function\_name => "initialize",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => {

"required\_fields" => [

[0] "uniq\_id",

[1] "crawl\_timestamp",

[2] "pageurl"

],

"schema" => {

"uniq\_id" => "",

"crawl\_timestamp" => "",

"pageurl" => "",

"product\_name" => "",

"price" => "",

"list\_price" => "",

"decription" => "",

"product\_url" => "",

"rating" => "",

"brand" => ""

}

},

:level => "INFO",

:hostname => "hdf40",

:line\_number => "99",

:function\_name => "initialize",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "getting fetched data charset bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "2763",

:function\_name => "get\_site\_fetched\_data\_charset",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "site\_name paseed to sdf\_utils to get log\_hash is bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "4008",

:function\_name => "get\_log\_hash\_for\_es",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "creating the directory /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/extraction",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "296",

:function\_name => "create\_directory\_if\_not\_present",

:file\_name => "pipeline\_v2\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "complete\_segment\_path: /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "139",

:function\_name => "get\_extraction\_log\_file\_for\_segment\_path",

:file\_name => "plugin\_and\_wtplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pipeline\_version: ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "3835",

:function\_name => "get\_pipeline\_version\_for\_site",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pipeline\_version is not found in DRL, so returning default 1",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "3838",

:function\_name => "get\_pipeline\_version\_for\_site",

:file\_name => "sdf\_utils.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "Making uniq\_id from :pageurl. Override get\_uniq\_id if you have any other logic.",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "414",

:function\_name => "get\_uniq\_id",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "Making uniq\_id from :pageurl. Override get\_uniq\_id if you have any other logic.",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "414",

:function\_name => "get\_uniq\_id",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

{

:msg => "pageurl processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl value will be searched with xpath //link[@rel='canonical']/@href ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl ,first node being used for //link[@rel='canonical']/@href ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "124",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl gave https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625 after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl gave https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625 after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "pageurl gave https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625 after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name value will be searched with xpath //p[contains(@class,'text-secondary')] ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name ,first node being used for //p[contains(@class,'text-secondary')] ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "124",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name gave Luxury Seat Pads - Navy 2pk after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name gave Luxury Seat Pads - Navy 2pk after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_name gave Luxury Seat Pads - Navy 2pk after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price value will be searched with xpath //span[contains(@itemprop,'price')][2]/@content ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price ,first node being used for //span[contains(@itemprop,'price')][2]/@content ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "124",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price gave 8.00 after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price gave 8.00 after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "price gave 8.00 after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price value will be searched with xpath //meta[contains(@property, 'og:price:amount')]/@content ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price ,first node being used for //meta[contains(@property, 'og:price:amount')]/@content ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "124",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price gave 8.00 after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price gave 8.00 after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "list\_price gave 8.00 after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription value will be searched with xpath //div[contains(@class,'tab-content')]//p[position()<10] ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription ,all nodes being used for //div[contains(@class,'tab-content')]//p[position()<10] ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "130",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription gave Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&M.|TOTAL SATISFACTION GUARANTEED\n\nIf you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\*\n\nCustomers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging.\n\nConditions apply to some products:\n\nRefunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title.\n\nUnwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned.\n\nFor reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned.\n\nMattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition. after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription gave Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&M.|TOTAL SATISFACTION GUARANTEED\n\nIf you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\*\n\nCustomers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging.\n\nConditions apply to some products:\n\nRefunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title.\n\nUnwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned.\n\nFor reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned.\n\nMattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition. after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "decription gave Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&M.|TOTAL SATISFACTION GUARANTEED If you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\* Customers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging. Conditions apply to some products: Refunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title. Unwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned. For reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned. Mattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition. after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url value will be searched with xpath //link[@rel='canonical']/@href ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url ,first node being used for //link[@rel='canonical']/@href ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "124",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url gave https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625 after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url gave https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625 after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "product\_url gave https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625 after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "please define method get\_rating,returning nil till then",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "188",

:function\_name => "process\_string",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand processing started with return\_type standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "42",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand will be looked up in plugin hierarchy",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "65",

:function\_name => "get\_automated\_value\_from\_page\_doc",

:file\_name => "generic\_meta\_baseplugin.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand value will be obtained via standard\_function\_value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "79",

:function\_name => "get\_value\_for\_field\_name\_depending\_on\_return\_type",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand value will be searched with xpath //meta[contains(@property,'og:brand')]/@content ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "111",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand ,first node being used for //meta[contains(@property,'og:brand')]/@content ",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "124",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand gave Alfresco Living after xpath application ,will apply post\_processing\_functions to this value",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "166",

:function\_name => "standard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand gave Alfresco Living after going through function get\_encoding\_fixed\_content\_base",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "brand gave Alfresco Living after going through function remove\_newlines\_and\_whitespaces",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "179",

:function\_name => "blockstandard\_value\_as\_per\_yml",

:file\_name => "extraction\_helper\_methods\_module.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "output record for https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "451",

:function\_name => "extract",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => {

"uniq\_id" => "e423a4de44a15abb00a8229c9795d6d6",

"crawl\_timestamp" => "2022-04-04 07:26:17 UTC",

"pageurl" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625",

"product\_name" => "Luxury Seat Pads - Navy 2pk",

"price" => "8.00",

"list\_price" => "8.00",

"decription" => "Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&M.|TOTAL SATISFACTION GUARANTEED If you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\* Customers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging. Conditions apply to some products: Refunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title. Unwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned. For reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned. Mattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition.",

"product\_url" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625",

"rating" => nil,

"brand" => "Alfresco Living"

},

:level => "INFO",

:hostname => "hdf40",

:line\_number => "452",

:function\_name => "extract",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "skipping key rating,since its value is nil or empty",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "585",

:function\_name => "blockpost\_process\_hash",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "output record after post processing for https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "457",

:function\_name => "extract",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => {

"uniq\_id" => "e423a4de44a15abb00a8229c9795d6d6",

"crawl\_timestamp" => "2022-04-04 07:26:17 UTC",

"pageurl" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625",

"product\_name" => "Luxury Seat Pads - Navy 2pk",

"price" => "8.00",

"list\_price" => "8.00",

"decription" => "Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&M.|TOTAL SATISFACTION GUARANTEED If you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\* Customers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging. Conditions apply to some products: Refunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title. Unwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned. For reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned. Mattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition.",

"product\_url" => "https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625",

"brand" => "Alfresco Living"

},

:level => "INFO",

:hostname => "hdf40",

:line\_number => "458",

:function\_name => "extract",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "Please override this method if any post processing need to be done after extraction.",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "339",

:function\_name => "meta\_page\_safe\_extract\_end\_hook",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "Output Record Array to be returned from Hash based extraction : [{\"uniq\_id\"=>\"e423a4de44a15abb00a8229c9795d6d6\", \"crawl\_timestamp\"=>\"2022-04-04 07:26:17 UTC\", \"pageurl\"=>\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\", \"product\_name\"=>\"Luxury Seat Pads - Navy 2pk\", \"price\"=>\"8.00\", \"list\_price\"=>\"8.00\", \"decription\"=>\"Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&M.|TOTAL SATISFACTION GUARANTEED If you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\* Customers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging. Conditions apply to some products: Refunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title. Unwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned. For reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned. Mattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition.\", \"product\_url\"=>\"https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625\", \"brand\"=>\"Alfresco Living\"}]",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "306",

:function\_name => "safe\_extract",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

{

:msg => "dependent\_data\_upload is not enabled for the site bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "327",

:function\_name => "safe\_extract",

:file\_name => "extract\_using\_hash.rb",

:log\_ts => "2022-04-04T07:26:17+00:00"

}

<?xml version="1.0" encoding="UTF-8"?>

<root>

<page>

<pageurl>https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625</pageurl>

<entity>

<record>

<uniq\_id>e423a4de44a15abb00a8229c9795d6d6</uniq\_id>

<crawl\_timestamp>2022-04-04 07:26:17 UTC</crawl\_timestamp>

<pageurl>https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625</pageurl>

<product\_name>Luxury Seat Pads - Navy 2pk</product\_name>

<price>8.00</price>

<list\_price>8.00</list\_price>

<decription>Luxury Seat Pads.|Bring maximum comfort to your garden with these wonderful Luxury Seat Pads.|This shower proof seat pad is perfect for picnics, and suitable for indoor and outdoor use.|2 pack, spot clean only.|Available Colours: Black, Navy, Grey|Materials:|Weight: 0.42kg (Approx.)|Dimensions: 39 x 42 x 3cm (Approx.)|Browse even more Gaden Furniture online, and buy in-store at B&amp;M.|TOTAL SATISFACTION GUARANTEED If you are not happy with a purchase you make at our stores, simply return it in its original condition, with proof of purchase, within 30 days and we’ll give you a full refund or replacement.\* Customers are advised that refunds and credit notes will only be given where proof of purchase (receipt, credit card statement, etc) is provided. To qualify for a refund or replacement, products need to be complete with all components and original packaging. Conditions apply to some products: Refunds and exchanges cannot be given against Audio/DVD and Computer Games. Should one of these products prove to be faulty it can be returned with your receipt within 30 days and we will exchange it for one of the same title. Unwanted Food or Drink Products - Once supply conditions are broken, there are a number of factors outside of our control that can affect the quality of a product. Therefore perishable goods such as food and drink cannot be returned. For reasons of hygiene and safety, personal grooming products, cosmetics or items of intimate clothing cannot be returned. Mattresses, mattress toppers, duvets and pillows can only be returned if they are unused, in the original packaging and in a resaleable condition.</decription>

<product\_url>https://www.bmstores.co.uk/products/luxury-seat-pads-navy-2pk-379625</product\_url>

<brand>Alfresco Living</brand>

</record>

</entity>

</page>

</root>

{

:msg => "Ext output file path is >>\n >> /home/nbolisetty/crawlspace/pipeline\_v2/crawl\_and\_postcrawl\_data/20220404/promptcloud\_master/parvathin\_training\_sites\_promptcloud\_master/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0/extraction/live\_crawl\_count\_bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master-20220404\_072616664777394\_\_0.xml <<",

:level => "INFO",

:hostname => "hdf40",

:line\_number => "593",

:function\_name => "run\_extraction",

:file\_name => "run\_test\_pipeline.rb",

:log\_ts => "2022-04-04T07:26:17+00:00",

"site\_name" => "bmstores\_com\_parvathin\_training\_sites\_promptcloud\_master",

"pipeline\_stage" => "run\_test\_pipeline",

"log\_yyyymmdd" => "20220404",

"client\_name" => "promptcloud\_master",

"person\_responsible" => "ravindran"

}

**6.0 Future Work**

With tools like Selendroid or Appium, the scenario is completely different. You can easily write the scenario that you need to test and automatically run that test script again and again. We have decided to use Appium with Android Emulator.

Android emulators are known as impossible tools to deal with for mobile development, though with the release of x86 emulators, things have started to work easily and it feels like working the applications inside laptops run quicker than the physical devices themselves.

Later, we created a Docker container using Ubuntu 16.04, Appium and Android x86 emulator to start the test about how many of those we could run at the same time.

**7.0 CONCLUSION**

The web crawling technology provides for an opportunity to improve statistical data quality and to reduce the overall workload for data collection. Using automatic price collection methods enables statisticians to react better to the increasing amount of data sources on the internet. Any implementation of the method requires thorough planning in various fields. Legal and data security aspects need to be dealt with in the beginning. Necessary IT resources and IT training required to maintain the automatic data collection system have to be estimated in the course of a pilot project and should not be underestimated.