Json structure

Input parameters

Json using url:


```
{
   "id": 1,
   "crawlerType": 4,
   "items": [
      {
         "siteId": "0",
         "urlContentResponse": null,
         "siteObj": {
            "fetchType": 1,
            "id": "0",
            "uDate": null,
            "tcDate": null,
            "cDate": null,
            "resources": null,
            "iterations": null,
            "description": null,
            "urls": [],
            "filters": [],
            "properties": {},
            "state": null,
            "priority": null,
            "maxURLs": null,
            "maxResources": null,
            "maxErrors": null,
            "maxResourceSize": null,
            "requestDelay": null,
            "httpTimeout": null,
            "errorMask": null,
            "errors": null,
            "urlType": null,
            "contents": null,
            "processingDelay": null,
            "size": null,
            "avgSpeed": null,
            "avgSpeedCounter": null,
            "userId": null,
            "recrawlPeriod": null,
            "recrawlDate": null,
            "maxURLsFromPage": null,
            "collectedURLs": null
         },
         "urlObj": {
            "status": 2,
            "linksI": 0,
            "linksE": 0,
            "contentMask": 0,
            "processingTime": 0,
            "CDate": null,
            "mRateCounter": 0,
            "httpTimeout": 60000,
            "size": 0,
            "urlPut": null,
            "batchId": 0,
            "lastModified": null,
            "tagsCount": 0,
            "mRate": 0,
            "charset": "",
            "state": 0,
            "httpCode": 0,
            "priority": 0,
            "maxURLsFromPage": 0,
            "processingDelay": 1000,
            "crawlingTime": 0,
            "type": 1,
            "processed": 0,
            "totalTime": 0,
            "siteSelect": 0,
            "contentType": "",
            "pDate": null,
            "errorMask": 0,
            "httpMethod": "get",
            "eTag": "",
            "siteId": "0",
            "freq": 0,
            "tcDate": null,
            "rawContentMd5": "",
            "crawled": 0,
            "UDate": null,
            "contentURLMd5": "",
            "requestDelay": 1000,
            "depth": 0,
            "parentMd5": "",
            "urlUpdate": null,
            "tagsMask": 0,
            "urlMd5": "cbc10d3e9ce15ab1130f542d77b348ca",
            "url": "https://blog.udemy.com/api-testing/"
         },
         "urlPutObj": {
            "putDict": {},
            "urlMd5": "cbc10d3e9ce15ab1130f542d77b348ca",
            "contentType": 0,
            "siteId": "0",
            "fileStorageSuffix": null,
            "criterions": null
         },
         "properties": {
            "DB_TASK_MODE": "RO",
            "HTTP_REDIRECTS_MAX": 5,
            "HTML_REDIRECTS_MAX": 5,
            "HTML_RECOVER": "0",
            "PROCESSOR_PROPERTIES":
"{\"algorithm\":{\"algorithm_name\":\"user_name_algorithm\"},\"modules\":{\"user_name_algorithm\":[\"ScrapyExtractor\",\"GooseExtractor\",\"NewspaperExtractor\"]}}",
            "template": {
               "templates": [
                  {
                     "output_format": {
                        "name": "json",
                        "header": "[\n",
                        "items_header": "",
                        "item":
"{\n\"pubdate\":\"%pubdate%\",\n\"title\":\"%title%\",\n\"description\":\"%description%\",\n\"media\":\"%media%\",\n\"author\":\"%author%\",\n\"dc_date\":\"%dc_date%\",\n\"link\":\"%link%\",\n\"keywords\":\"%keywords%\",\n\"content_encoded\":\"%content_encoded%\",\n\"html_lang\":\"%html_lang%\",\n\"pubdate_extractor\":\"%pubdate_extractor%\",\n\"title_extractor\":\"%title_extractor%\",\n\"description_extractor\":\"%description_extractor%\",\n\"media_extractor\":\"%media_extractor%\",\n\"author_extractor\":\"%author_extractor%\",\n\"dc_date_extractor\":\"%dc_date_extractor%\",\n\"link_extractor\":\"%link_extractor%\",\n\"keywords_extractor\":\"%keywords_extractor%\",\n\"content_encoded_extractor\":\"%content_encoded_extractor%\",\n\"html_lang_extractor\":\"%html_lang_extractor%\",\n\"crawler_time\":\"%crawler_time%\",\n\"scraper_time\":\"%scraper_time%\",\n\"errors_mask\":\"%errors_mask%\"\n}\n",
                        "items_footer": "",
                        "footer": "]\n"
                     },
                     "tags": {
                        "pubdate": {
                           "default": ""
                        },
                        "title": {
                           "default": ""
                        },
                        "description": {
                           "default": ""
                        },
                        "media": {
                           "default": ""
                        },
```

```
            "author": {
                "default": ""
            },
            "dc_date": {
                "default": ""
            },
            "link": {
                "default": ""
            },
            "keywords": {
                "default": ""
            },
            "content_encoded": {
                "default": ""
            },
            "html_lang": {
                "default": ""
            },
            "errors_mask": {
                "default": ""
            }
        },
        "priority": 100,
        "mandatory": 1,
        "is_filled": 0
      }
    ],
    "select": "first_nonempty"
  }
 },
 "urlId": "cbc10d3e9ce15ab1130f542d77b348ca"
}
]
}
```


Response example:


```
[
  {
    "pubdate": "%pubdate%",
    "title": "Why It Matters, and How To Do It",
    "description": "API testing -- an overall survey. What API testing is, why it's important, how to do it, and the best ways to do it.",
    "media": "<img class=\"alignright size-medium wp-image-80080\" alt=\"api testing\" src=\"https://blog.udemy.com/wp-content/uploads/2014/04/shutterstock_77883016-300x199.jpg\" width=\"300\" height=\"199\">",
    "author": "Michael Churchman",
    "dc_date": "April 18, 2014",
    "link": "https://blog.udemy.com/api-testing/",
    "keywords": "api testing,api,software testing,api call,application programming interface,testing,quality assurance,debugging,bug fix,debug,software development,qa,for students,technology",
    "content_encoded": "API testing — what is it, why is it important, and what do you need to know about it? First and most basic, API stands for Application Programming Interface. An API is a set of procedures, functions, and other points of access which an application, an operating system, a … checking out the broad range of courses which are available online.",
    "html_lang": "en-US",
    "pubdate_extractor": "%pubdate_extractor%",
    "title_extractor": "GooseExtractor",
    "description_extractor": "ScrapyExtractor",
    "media_extractor": "ScrapyExtractor",
    "author_extractor": "ScrapyExtractor",
    "dc_date_extractor": "ScrapyExtractor",
    "link_extractor": "GooseExtractor",
    "keywords_extractor": "GooseExtractor",
    "content_encoded_extractor": "GooseExtractor",
    "html_lang_extractor": "ScrapyExtractor",
    "crawler_time": "1.2",
    "scraper_time": "0.33 sec.",
    "errors_mask": "0"
  }
]
```


Output parameters:


```
{
    "errorCode": 0,
    "errorMessage": "",
    "itemsList": [
        {
            "errorCode": 0,
            "errorMessage": "",
            "host": "localhost",
            "id": 3663140995,
            "itemObject": [
                {
                    "contentURLMd5": "",
                    "cookies": [],
                    "dbFields": {
                        "BatchId": 0,
                        "Charset": "utf-8",
                        "ContentType": "text/html",
                        "Crawled": 1,
                        "ErrorMask": 0,
                        "HttpCode": 200,
                        "Processed": 0,
                        "Size": 12240,
                        "TagsCount": 9,
```

```
                        "TagsMask": 66751
                    },
                    "headers": [],
                    "meta": [],
                    "processedContents": [
                        {
                            "buffer":
"Wwp7CiJwdWJkYXRlIjoiV2VkLCAyMiBBcHIgMjAxNSAxMDo0MTowMCAtMDAwMCIsCiJ0aXRsZSI6IuODgeODpeODvOODquODg+O
Dl+OBp+aPj+OBj+OCtOODg+ODm+OBruiCluWDj+OAgeayoeW+jDEyNeW5tOOBpyDjgqrjg6njg7Pjg4DjgIDlhpnnnJ8y5p6a44C
A5Zu96Zqb44OL44Ol44O844K577yaQUZQQkIgTmV3cyIsCiJkZXNjcmlwdGlvbiI6IuOAkDTmnIgyMuaXpSBBRlDjgJHjgqrjg6n
jg7Pjg4Djg7vjg6rjg4PjgrvvvIhMaXNzZe+8ieOBq+OBguOCi+OCreODpeODvOOCseODs+ODm+ODle+8iEtldWtlbmhvZu+8ieW
FrOWckuOBruODleODqeODr+ODvOOCrOODvOODh+ODs+OBp+OAgeWQjOWbveOBrueUu+WutuODk+ODs+OCu+ODs+ODiOODu+ODleO
CoeODs+ODu+OCtOODg+ODm++8iFZpbmNlbnQgVmFuIEdvZ2jvvInjga7ogpblg4/jgYzjg4Hjg6Xjg7zjg6rjg4Pjg5fjgaflho3
nj77jgZXjgozjgZ/jgIIiLAoibWVkaWEiOiI8aW1nIGlkPVwibWFpbi1pbWdcIiBzcmM9XCJodHRwOi8vYWZwYmIuaXNtY2RuLmp
wL213aW1ncy9lL2QvNTAweDQwMC9pbWdfZWQ5Yzg5ZDRkMDlmNjdkZTMwY2I5YWE3NjZhYWZkNDIyNzAzMTQuanBnXCIgYWx0PVw
i44OB44Ol44O844Oq44OD44OX44Gn5o+P44GP44K044OD44Ob44Gu6IKW5YOP44CB5rKh5b6MMTI15bm044GnIOOCquODqeODs+O
DgFwiIHRpdGxlPVwi44OB44Ol44O844Oq44OD44OX44Gn5o+P44GP44K044OD44Ob44Gu6IKW5YOP44CB5rKh5b6MMTI15bm044G
nIOOCquODqeODs+ODgFwiPiIsCiJhdXRob3IiOiIlYXV0aG9yJSIsCiJkY19kYXRlIjoiV2VkLCAyMiBBcHIgMjAxNSAxMDo0MTo
0NyArMDkwMCIsCiJsaW5rIjoiaHR0cDovL3d3dy5hZnBiYi5jb20vYXJ0aWNsZXMvLS8zMDQ2MDg3IiwKImtleXdvcmRzIjoi5Zu
96Zqb44OL44Ol44O844K5LOOCueODneODvOODhOODi+ODpeODvOOCuSxBRlAs44OL44Ol44O844K5LE5ld3MsQUZQQkIiLAoiY29
udGVudF9lbmNvZGVkIjoi44OB44Ol44O844Oq44OD44OX44Gn5o+P44GP44K044OD44Ob44Gu6IKW5YOP44CB5rKh5b6MMTI15bm
044GnIOOCquODqeODs+ODgDIwMTXlubQwNOaciDIy5pelIDEwOjQx44CA55m65L+h5Zyw77ya44Oq44OD44K7L+OCquODqeODs+O
DgOODqeOCpOODlSDjgZPjgbzjgozoqbHjgZPjga7jg4vjg6Xjg7zjgrnjgpLjgrfjgqfjgqJUd2VldOODoeODvOODq+OBp+mAgeO
CiyAgMS8y5YmN44G45qyh44G45YaZ55yf5ouh5aSn4paywqDjgq3jg6Pjg5fjgrfjg6fjg7Pooajnp LrDl+OCquODqeODs+ODgOO
Du+ODquODg+OCu+OBruOCreODpeODvOOCseODs+ODm+ODleWFrOWckuOBq+ODgeODpeODvOODquODg+ODl+OBp+aPj+OBi+OCjOO
Bn+ODk+ODs+OCu+ODs+ODiOODu+ODleOCoeODs+ODu+OCtOODg+ODm+OBruiCluWDj++8iDIwMTXlubQ05pyIMjHml6Xmkq7lvbH
vvInjgIIoYylBRlAvQU5QL1JFTUtPIERFIFdBQUzjgJDjg6Hjg4fjgqPjgqLjg7vloLHpgZPplqLkv4Ljg7vms5Xkurrjga7mlrn
jgJHlhpnnnJ/os7zlhaXjga7jgYrllY/lkIjjgZvjga/jgZPjgaHjgonplqLpgKPlhpnnnJ/jgJA05pyIMjLml6UgQUZQ44CR44K
q44Op44Oz44OA44O744Oq44OD44K777yITGlzc2XvvInjgavjgYLjgovjgq3jg6Xjg7zjgrHjg7Pjg5vjg5XvvIhLZXVrZW5ob2b
vvInlhazlnJLjga7jg5Xjg6njg6/jg7zjgqzjg7zjg4fjg7PjgafjgIHlkIzlm73jga7nlLvlrrbjg5Pjg7Pjgrvjg7Pjg4jjg7v
jg5XjgqHjg7Pjg7vjgrTjg4Pjg5vvvIhWaW5jZW50IFZhbiBHb2do77yJ44Gu6IKW5YOP44GM44OB44Ol44O844Oq44OD44OX44G
n5YaN54++44GV44KM44Gf44CC5LuK5bm044Gv44K044OD44Ob5rKh5b6MMTI15bm044Gr44GC44Gf44KL44GT44Go44GL44KJ44C
B44OV44Op44Ov44O844Ks44O844OH44Oz44Gu44OG44O844Oe44KC44CM44OV44Kh44Oz44O744K044OD44Ob44CN44Go44Gq44G
j44Gm44GE44KL44CCKGMpQUZQ44Op44Kk44OVICDkuIDopqfjgbhQUlBS44GK44GZ44GZ44KB5YaZ55yf54m56ZuGNDjjgJDnibn
pm4bjgJFGMTbmiKbpl5jmqZ/jgIzjg5XjgqHjgqTjg4bjgqPjg7PjgrDjg7vjg5XjgqHjg6vjgrPjg7PjgI01MeOAkOeJuembhuO
AkUFGUOOBruODmeOCueODiOOCt+ODp+ODg+ODiO+8mjfmnIgyNe+9njMx5pelMTnpgb/mmpHjgavjgYbjgaPjgabjgaTjgZHvvJ/
ku4/jg6Ljg7Pjg5bjg6njg7Pjga7msLfnqbTjgajjgqLjg6vjg5fjgrnjga7ls7DjgIUyNuOAkOWGmeecn+eJuembhuOAkeWGpee
Oi+aYn+OAgeaOouafu+apn+OBjOacgOaOpei/kSDmmI7jgonjgYvjgavjgarjgovjgZ3jga7lp78yNeOAjE1DTeOCs+ODn+ODg+O
Cr+ODu+OCs+ODs+OAjemWi+WCrOOAgeaXpeacrOS9nOWTgeOBruOCs+OCueODl+ODrOOCguWkmuaVsCDoi7ExN+ODluODqeOCuOO
Dq+OBp+OAjOOCteODs+ODkeOCpuODreODu+OCv+ODiOOCpeODvOODu+OCpuOCo+ODvOOCrzIwMTXjgI3plovlgqzjgYrjgZnjgZn
jgoHli5XnlLvnibnpm4bli5XnlLvvvJo0MzPlj7DjgYzkuIDmlonpo5vooYzjgIHnqbrjgpLlvanjgovprq7jgoTjgYvnhrHmsJf
nkIMg5LuP5YuV55S777ya57Gz44Gn44OQ44O844OT44O85Lq65b2i44Gu44Kz44Oz44OZ44Oz44K344On44Oz44CB44OV44Kh44O
z5aSn6ZuG5ZCI5YuV55S777ya44Or44O844OW44Or44O744Op44Oz44K5576O6KGT6aSo44Gu44CM5pmC5Luj6aCG6Zmz5YiX44C
NQ09TQ+OBruOCr+ODreODjuODoeODvOOCv+ODvOiqjeWumuOBqOOBr+KAleOAjOOCr+ODreODjuODoeODiOODquODvOOAjeesrDX
oqbHli5XnlLvvvJrjgq/jg6njgrfjg4Pjgq/jgqvjg7zjgYzpgYTjgZbjgarjgYbjgIHjgq3jg6Xjg7zjg5Djga7
poqjmma/li5XnlLvvvJrnpL7kvJrnmoTmipHlnKfjgavjgYLjgonjgYzjgYTnpL7kvJrjgavpgLLlh7rjgZnjgovjgqvjgr/jg7z
jg6vlpbPmgKciLAoiaHRtbF9sYW5nIjoiamEiLAoicHViZGF0ZV9leHRyYWN0b3IiOiJTY3JhcHlFeHRyYWN0b3IiLAoidGl0bGV
fZXh0cmFjdG9yIjoiR29vc2VFeHRyYWN0b3IiLAoiZGVzY3JpcHRpb25fZXh0cmFjdG9yIjoiU2NyYXB5RXh0cmFjdG9yIiwKIm1
lZGlhX2V4dHJhY3RvciI6IlNjcmFweUV4dHJhY3RvciIsCiJhdXRob3JfZXh0cmFjdG9yIjoiJWF1dGhvcl9leHRyYWN0b3IlIiw
KImRjX2RhdGVfZXh0cmFjdG9yIjoiU2NyYXB5RXh0cmFjdG9yIiwKImxpbmtfZXh0cmFjdG9yIjoiR29vc2VFeHRyYWN0b3IiLAo
ia2V5d29yZHNfZXh0cmFjdG9yIjoiR29vc2VFeHRyYWN0b3IiLAoiY29udGVudF9lbmNvZGVkX2V4dHJhY3RvciI6IlNjcmFweUV
4dHJhY3RvciIsCiJodG1sX2xhbmdfZXh0cmFjdG9yIjoiU2NyYXB5RXh0cmFjdG9yIiwKImNyYXdsZXJfdGltZSI6IjQuNzY1Iiw
KInNjcmFwZXJfdGltZSI6IjAuNzkiLAoiZXJyb3JzX21hc2siOiIwIgp9Cl0K",
                            "cDate": 0,
                            "typeId": 10
                        }
                    ],
                    "rawContentMd5": "",
                    "rawContents": [],
                    "requests": [],
                    "siteId": 0,
                    "status": 7,
                    "url": "http://www.afpbb.com/articles/-/3046087",
                    "urlMd5": null
                }
            ],
            "node": "r014_data",
            "port": "5733",
            "time": 13543
        }
    ]
}
```


Input parameters

Json using html content:


```
{
    "id": 1928771667,
    "crawlerType": 4,
```

```
    "items": [
        {
            "siteId": "0",
            "urlContentResponse": null,
            "siteObj": {
                "fetchType": 1,
                "id": "0",
                "uDate": null,
                "tcDate": null,
                "cDate": null,
                "resources": null,
                "iterations": null,
                "description": null,
                "urls": [],
                "filters": [],
                "properties": {},
                "state": null,
                "priority": null,
                "maxURLs": null,
                "maxResources": null,
                "maxErrors": null,
                "maxResourceSize": null,
                "requestDelay": null,
                "httpTimeout": null,
                "errorMask": null,
                "errors": null,
                "urlType": null,
                "contents": null,
                "processingDelay": null,
                "size": null,
                "avgSpeed": null,
                "avgSpeedCounter": null,
                "userId": null,
                "recrawlPeriod": null,
                "recrawlDate": null,
                "maxURLsFromPage": null,
                "collectedURLs": null
            },
            "urlObj": {
                "status": 2,
                "linksI": 0,
                "linksE": 0,
                "contentMask": 1,
                "processingTime": 0,
                "CDate": null,
                "mRateCounter": 0,
                "httpTimeout": 60000,
                "size": 0,
                "urlPut": {
                    "urlMd5": "317c812ea5d933ce3b035a1b56ca0c3f",
                    "contentType": 0,
                    "siteId": "0",
                    "fileStorageSuffix": null,
                    "criterions": null,
                    "putDict": {
                        "data":
"PCFET0NUWVBFIGh0bWw+PGh0bWw+PGhlYWQ+PG1ldGEgY2hhcnNldD11dGYtOD48dGl0bGU+VGl0bGU8L3RpdGxlPjwvaGVhZD4
8Ym9keT48cD5QYWdlPC9wPjxcL2JvZHk+PC9odG1sPg=="
                    }
                },
                "batchId": 1928771667,
                "lastModified": null,
                "tagsCount": 0,
                "mRate": 0,
                "charset": "utf-8",
                "state": 0,
                "httpCode": 200,
                "priority": 0,
                "maxURLsFromPage": 0,
                "processingDelay": 0,
                "crawlingTime": 0,
                "type": 1,
                "processed": 0,
                "totalTime": 0,
                "siteSelect": 0,
                "contentType": "text/html",
                "pDate": null,
                "errorMask": 0,
                "httpMethod": "get",
                "eTag": "",
```

```
                "siteId": "0",
                "freq": 0,
                "tcDate": null,
                "rawContentMd5": "",
                "crawled": 0,
                "UDate": null,
                "contentURLMd5": "",
                "requestDelay": 0,
                "depth": 0,
                "parentMd5": "",
                "urlUpdate": null,
                "tagsMask": 0,
                "urlMd5": "317c812ea5d933ce3b035a1b56ca0c3f",
                "url": null
            },
            "urlPutObj": null,
            "properties": {
                "DB_TASK_MODE": "RO",
                "HTTP_REDIRECTS_MAX": 5,
                "HTML_REDIRECTS_MAX": 5,
                "HTML_RECOVER": "0",
                "PROCESSOR_PROPERTIES":
"{\"algorithm\":{\"algorithm_name\":\"user_name_algorithm\"},\"modules\":{\"user_name_algorithm\":[\
"ScrapyExtractor\",\"GooseExtractor\",\"NewspaperExtractor\"]}}",
                "template": {
                    "templates": [
                        {
                            "output_format": {
                                "name": "json",
                                "header": "[\n",
                                "items_header": "",
                                "item":
"{\n\"pubdate\":\"%pubdate%\",\n\"title\":\"%title%\",\n\"description\":\"%description%\",\n\"media\
":\"%media%\",\n\"author\":\"%author%\",\n\"dc_date\":\"%dc_date%\",\n\"link\":\"%link%\",\n\"keywor
ds\":\"%keywords%\",\n\"content_encoded\":\"%content_encoded%\",\n\"html_lang\":\"%html_lang%\",\n\"
pubdate_extractor\":\"%pubdate_extractor%\",\n\"title_extractor\":\"%title_extractor%\",\n\"descript
ion_extractor\":\"%description_extractor%\",\n\"media_extractor\":\"%media_extractor%\",\n\"author_e
xtractor\":\"%author_extractor%\",\n\"dc_date_extractor\":\"%dc_date_extractor%\",\n\"link_extractor
\":\"%link_extractor%\",\n\"keywords_extractor\":\"%keywords_extractor%\",\n\"content_encoded_extrac
tor\":\"%content_encoded_extractor%\",\n\"html_lang_extractor\":\"%html_lang_extractor%\",\n\"crawle
r_time\":\"%crawler_time%\",\n\"scraper_time\":\"%scraper_time%\",\n\"errors_mask\":\"%errors_mask%\
"\n}\n",
                                "items_footer": "",
                                "footer": "]\n"
                            },
                            "tags": {
                                "pubdate": [],
                                "title": [],
                                "description": [],
                                "media": [],
                                "author": [],
                                "dc_date": [],
                                "link": [],
                                "keywords": [],
                                "content_encoded": [],
                                "html_lang": [],
                                "pubdate_extractor": [],
                                "title_extractor": [],
                                "description_extractor": [],
                                "media_extractor": [],
                                "author_extractor": [],
                                "dc_date_extractor": [],
                                "link_extractor": [],
                                "keywords_extractor": [],
                                "content_encoded_extractor": [],
                                "html_lang_extractor": [],
                                "crawler_time": [],
                                "scraper_time": [],
                                "errors_mask": []
                            },
                            "priority": 100,
                            "mandatory": 1,
                            "is_filled": 0
                        }
                    ],
                    "select": "first_nonempty"
                }
            },
            "urlId": "317c812ea5d933ce3b035a1b56ca0c3f"
        }
```

```
    ]
}
```


Request example:

Raw HTML body:

<!DOCTYPE html><html><head><meta charset=utf-8><title>Title</title></head><body> Page <\/body></html>

Response example:


```
{
    "errorCode": 0,
    "errorMessage": "",
    "itemsList": [
        {
            "errorCode": 0,
            "errorMessage": "",
            "host": "localhost",
            "id": 102471721,
            "itemObject": [
                {
                    "contentURLMd5": "",
                    "cookies": [],
                    "dbFields": {
                        "BatchId": 1928771667,
                        "Charset": "utf-8",
                        "ContentType": "text/html",
                        "Crawled": 0,
                        "ErrorMask": 0,
                        "HttpCode": 200,
                        "Processed": 0,
                        "Size": 0,
                        "TagsCount": 3,
                        "TagsMask": 25
                    },
                    "headers": [],
                    "meta": [],
                    "processedContents": [
                        {
                            "buffer":
"Wwp7CiJwdWJkYXRlIjoiJXB1YmRhdGUlIiwKInRpdGxlIjoiVGl0bGUiLAoiZGVzY3JpcHRpb24iOiIlZGVzY3JpcHRpb24lIiw
KIm1lZGlhIjoiOi8vIiwKImF1dGhvciI6IiVhdXRob3IlIiwKImRjX2RhdGUiOiIlZGNfZGF0ZSUiLAoibGluayI6Ik5vbmUiLAo
ia2V5d29yZHMiOiIla2V5d29yZHMlIiwKImNvbnRlbnRfZW5jb2RlZCI6IiVjb250ZW50X2VuY29kZWQlIiwKImh0bWxfbGFuZyI
6IiVodG1sX2xhbmclIiwKInB1YmRhdGVfZXh0cmFjdG9yIjoiJXB1YmRhdGVfZXh0cmFjdG9yJSIsCiJ0aXRsZV9leHRyYWN0b3I
iOiJHb29zZUV4dHJhY3RvciIsCiJkZXNjcmlwdGlvbl9leHRyYWN0b3IiOiIlZGVzY3JpcHRpb25fZXh0cmFjdG9yJSIsCiJtZWR
pYV9leHRyYWN0b3IiOiJOZXdzcGFwZXJFeHRyYWN0b3IiLAoiYXV0aG9yX2V4dHJhY3RvciI6IiVhdXRob3JfZXh0cmFjdG9yJSI
sCiJkY19kYXRlX2V4dHJhY3RvciI6Ik5ld3NwYXBlckV4dHJhY3RvciIsCiJsaW5rX2V4dHJhY3RvciI6Ikdvb3NlRXh0cmFjdG9
yIiwKImtleXdvcmRzX2V4dHJhY3RvciI6IiVrZXl3b3Jkc19leHRyYWN0b3IlIiwKImNvbnRlbnRfZW5jb2RlZF9leHRyYWN0b3I
iOiIlY29udGVudF9lbmNvZGVkX2V4dHJhY3RvciUiLAoiaHRtbF9sYW5nX2V4dHJhY3RvciI6IiVodG1sX2xhbmdfZXh0cmFjdG9
yJSIsCiJjcmF3bGVyX3RpbWUiOiIwLjAiLAoic2NyYXBlcl90aW1lIjoiMC4wNyIsCiJlcnJvcnNfbWFzayI6IjAiCn0KXQo=",
                            "cDate": 0,
                            "typeId": 10
                        }
                    ],
                    "rawContentMd5": "",
                    "rawContents": [],
                    "requests": [],
                    "siteId": 0,
                    "status": 7,
                    "url": "None",
                    "urlMd5": null
                }
            ],
            "node": "r013_data",
            "port": "5732",
            "time": 1078
        }
    ]
}
```


Fields in the json file

Input parameters

Using url	
id	The unique request identifier.
crawlerType	Type crawler
items	the whole structure of the document
siteId	id of the site
urlContentResponse	
siteObj	tags site
fetchType	1 - static (default), 2 - dynamic, 3 - external'
id	
uDate	Update date
tcDate	Touch date, when some action was performed with site or it's URLs
cDate	Creation date, not changed all life time of the Site object inside the DC service.
resources	Number of raw data web-resources received from web servers that are stored in the local file storage. Depends on mode that this parameter received it can reflects total number for all hosts for whole installation or only one host in each response from each host server.
iterations	Crawling iterations counter
description	description
urls	URLs string
filters	
properties	
state	
priority	Priority by default inherited from parent URL or set by another source
maxURLs	Limit of max number of collected URLs in the URLs table for site. Depends on crawling type, re-crawling, auto-remove and resources TTL settings can block Site crawling, define when system will try to remove existing resources and so on. This limit defined and used for each host data node in the installation. Total value returned with grouped results will be calculated as sum of values for each data host. To configure proper way for new site it need to be set as ceil(RequestedMaxNumber/NumberOfDataHosts). Zero value means unlimited resources collection.
maxResources	Limit of max number of processed resources stored in the key-value DB. If this max value reached – processor will ignore resource and skip it from processing. Zero value means unlimited.
maxErrors	Limit of max errors count that happened during the site usage. It can be increased by the crawler, processor or

	another compound or module. It can be reset to zero by state change operations like re-crawling start. If this value reached site processing suspended. Zero value means unlimited.
maxResourceSize	Limit max raw content file size that can be stored after some URL was crawled. In case of raw content size bytes is greater than this limit raw content skipped and file is not stored. Correspondent error is set in ErrorsMask. Zero value means unlimited.
requestDelay	Delay before crawling request, ms. Used to make crawling process more smooth and balanced.
httpTimeout	Timeout of HTTP response, ms.
errorMask	Bit set of errors. Each bit reflects some error state of operation or data of the site. Mostly cumulative representation of errors that are happened during the resources crawling and processing. Per resource errors represented by the same field in the URLs table.
errors	Errors counter, represents total number of error happened during the Site usage from NEW state or state was changed like re-crawl.
urlType	Type of URL by usage in crawling and other processes. Defines behavior of crawler. 0 - Regular, collect URLs and insert only for this site according filters; 1 - Single, do not collect URLs, 3 - collect URLs, create sites and insert for all.
contents	The same of the Resources but scraped contents count in the key-value db. Scraped contents count can differ from Resources because some resources will not be processed by content-type, errors mask or other kind conditions. But, Contents always equal or less than Resources.
processingDelay	Delay before the content processing.
size	Total size of all raw contents crawled for period of the Site usage. Can be reset to zero value after state change operations like re-crawl start.
avgSpeed	Average crawling speed, bytes per second or BPS rate. Calculated for period of the Site usage for all resources.
avgSpeedCounter	Counter of times when average speed rate was calculated.
userId	Unique user Id. Used by client API to identify owner user and permissions restrictions and ACL.
recrawlPeriod	Re-crawl period, min. 0 – means the Site is not re-crawled. This value used in calculations to define the re-crawl date. Re-crawl starts the Site crawling from root URLs and scans all resources the same way as it was done first time after newly created.
recrawlDate	The re-crawl process starts date. It is exact date when (or bit after that because period of state check is not real time) site

	became re-crawled. Re-crawl sets the root URLs in NEW status and pushed next iteration of the Site scan and crawl. The condition is NOW()<=RecrawlDate.
maxURLsFromPage	Limit of max unique URLs that can be collected from one HTML page or RSS feed. Zero value means unlimited.
collectedURLs	Number of collected URLs in the SQL URLs table. The same way as fields above can represent the total number or per host number.
urlObj	
status	0 - Undefined, 1 - New, 2 - selected for crawling, 3 - crawling, 4 - crawled, 5 - selected to process, 6 - processing, 7 - processed, 8 - as 2 for incremental crawling.
linksI	Number of internal links
linksE	Number of external links
contentMask	
processingTime	After crawling processing time, msec
CDate	Creation date
mRateCounter	Counter for AVG mutability rate calculation
httpTimeout	Timeout of HTTP response, ms.
size	Total size of all raw contents crawled for period of the Site usage. Can be reset to zero value after state change operations like re-crawl start.
urlPut	
batchId	
lastModified	Last-Modified tag value
tagsCount	The counter of detected tags by the processing algorithm named the scraping.
mRate	AVG mutability rate, relative value calculated to be used as a measure of frequency of the page changes content.
charset	The charset, can be changed while resource processing cause sequential usage of different detection algorithms.
state	0 – enabled, 1 – disabled, 2 – error, used to control of usage of Site's URLs
httpCode	HTTP response code
priority	It is rate value that used to range sites selected for some operation. Typical usage is a crawling or processing. In

	group operations sites are ranged and then limits for number of sites applied. Higher value signs top of usage list and site will appear in the selected list more often.
maxURLsFromPage	Limit of max unique URLs that can be collected from one HTML page or RSS feed. Zero value means unlimited.
processingDelay	Delay before resource data processing, ms.
crawlingTime	Time of crawling, msec
type	0 – include, 1 – exclude behavior depends on Mode and business logic usage
processed	Number of processed times.
totalTime	Total time – crawling + processing, msec
siteSelect	
contentType	MIME content-type string, can be changed while resource processing cause sequential usage of different detection algorithms.
pDate	The exact date of publication of article detected by the scraping processing.
errorMask	Bit set of errors. Each bit reflects some error state of operation or data of the site. Mostly cumulative representation of errors that are happened during the resources crawling and processing. Per resource errors represented by the same field in the URLs table.
httpMethod	The HTTP method used to fetch resource, GET, POST or another.
eTag	The “Etag” HTTP header’s field value
siteId	Site Id
freq	Frequency of usage of the page URL on other pages of this site that was already crawled.
tcDate	Touch date, updated when some action performed with this URL, for example – the crawling or processing.
rawContentMd5	The md5 of raw content file data
crawled	Number of crawled times.
UDate	Update date
contentURLMd5	The md5 calculated according the correspondent site’s properties settings based on the scraped fields string.
requestDelay	Delay before crawling request, ms. Used to make crawling process more smooth and balanced.

depth	The incremental depth relative with the root URL starting from zero. Can be used to evaluate as far the page was from the root page.
parentMd5	The md5 of the parent page URL, used as unique Id of the parent resource from this resource's URL reference was taken.
urlUpdate	
tagsMask	Tags mask bits set for processing algorithm named the scraper. Results of this kind of processing are tags set. This mask signs detection and successful scraping of some defined tags.
urlMd5	URL Id from URLs table in case of properties applicable only for one URL
url	URL string
urlPutObj	
putDict	
urlMd5	URL Id from URLs table in case of properties applicable only for one URL
contentType	MIME content-type string, can be changed while resource processing cause sequential usage of different detection algorithms.
siteId	
fileStorageSuffix	
criterions	
properties	
DB_TASK_MODE	
HTTP_REDIRECTS_MAX	Max HTTP redirects (hops count)
HTML_REDIRECTS_MAX	Max HTML redirects (hops count)
HTML_RECOVER	Use tidy lib to recover the source HTML content or not. Value "1" – means always recover. Value "2" – means recover if the regular DOM parser failed cause wrong HTML structure. In case of another value – no recovering at all.
PROCESSOR_PROPERTIES	TODO: description need to be updated for common cases of template-based and news type of the scraping. Used to set the processing algorithm and modules, for

	example for the real-time request to return RAW HTML buffer base64-encoded:
template	Site’s template:
templates	Site’s template:
output_format	Output format templates
name	Name of a property, used as unique identifier per site or URL object.
header	Header contents
items_header	Items templates heder
item	Templates item
items_footer	Templates footer
footer	Contents footer
tags	URL string templates
pubdate	Publication date templates
title	Title site templates
description	Description site templates
media	Media in site templates
author	Autor page in site templates
dc_date	
link	Link templates site
keywords	Keywords templates site
content_encoded	
html_lang	
pubdate_extractor	
title_extractor	
description_extractor	

media_extractor	
author_extractor	
dc_date_extractor	
link_extractor	
keywords_extractor	
content_encoded_extractor	
html_lang_extractor	Html lang extractor tags templates
crawler_time	Crawler time tags templates
scraper_time	Scraper time tags templates
errors_mask	Errors mask tags templates
priority	
mandatory	
is_filled	
select	
urlId	

When we use html content added values in **urlPut**

urlPut: {
- urlMd5: "",
- contentType: 0,
- siteId: "0",
- fileStorageSuffix: null,
- criterions: null,
- putDict: {
  - data: ""
- }

}

and **urlPutObj** equivalent to Null.

Using html content

urlPut	
urlMd5	URL Id from URLs table in case of properties applicable only for one URL
contentType	MIME content-type string, can be changed while resource processing cause sequential usage of different detection algorithms.
siteId	
fileStorageSuffix	
criterions	
putDict	
data	

Response example:

Using html content	
errorCode	
errorMessage	
itemsList	
errorCode	
errorMessage	
host	
id	The unique Site object identifier. Normally created from first root URL specified when new site create action performed. Can be defined direct way in SiteNew request or changed after the Site object created. If user lost this identifier site can be found by find operation query. Id returned on SiteNew operation request as the “statuses” field array item value in the GeneralResponse object.
itemObject	
contentURLMd5	The md5 calculated according the correspondent site’s properties settings based on the scraped fields string.
cookies	
dbFields	
BatchId	
Charset	The charset, can be changed while resource processing cause sequential usage of different detection algorithms.
ContentType	MIME content-type string, can be changed while resource processing cause sequential usage of different detection algorithms.

Crawled	Number of crawled times.
ErrorMask	Error mask bits set, detailed description see in the architecture document DC_application_architecture.docx or use the decode utility api/python/bin/dc-urls-mask.py.
HttpCode	HTTP response code
Processed	Number of processed times.
Size	Resource size, byte
TagsCount	The counter of detected tags by the processing algorithm named the scraping.
TagsMask	Tags mask bits set for processing algorithm named the scraper. Results of this kind of processing are tags set. This mask signs detection and successful scraping of some defined tags.
headers	
meta	
processedContents	
buffer	
cDate	Creation date.
typeId	
rawContentMd5	The md5 of raw content file data
rawContents	
requests	
siteId	
status	0 - Undefined, 1 - New, 2 - selected for crawling, 3 - crawling, 4 - crawled, 5 - selected to process, 6 - processing, 7 - processed, 8 - as 2 for incremental crawling.
url	Url site
urlMd5	URL Id from URLs table in case of properties applicable only for one URL
node	
port	
time	