Distributed Crawler Application databases schema

The databases schema specification per key application functional objects

The name of database fields corresponds to name of the Python object field with difference of first character. The db table field name has capitalized first character, but the Python object is not. The Id suffix in the db table field name has subdivided with underscore "_".

The Site object

The sites object data fields represented by the Site object instance or json response on the SITE_STATUS or SITE_FIND operation. This object represented by several tables.

`dc_sites`.`sites`

This object represents main structure object that is managed by the DC service. It defines set of properties and rules for the crawling process and most of its main stages.

Fields list:

Name	Type	Description
Id	String max 32	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.
UDate	DataTime	Update some fields date
TcDate	DataTime	Touch date, when some action was performed with site or it's URLs
CDate	DataTime	Creation date, not changed all life time of the Site object inside the DC service.
Resources	Longint	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.
Contents	Longint	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.
CollectedURLs	Longint	Number of collected URLs in the SQL URLs table. The same way as fields above can represent the total number or per host number.
NewURLs	Longint	Number of new URLs in the SQL URLs table. New URLs are in state 1 and supposed to be crawled.
DeletedURLs	Longint	Number of URLs in the SQL deleted_urls database URLs table. Those URLs are moved from the main DB to deleted DB.
Iterations	Longint	Current iteration number of site periodic crawling process. Typical periodic crawling process is re-crawling. This number reflects counter

		of periodic processes starts but not signs that iterations are finished.
State	Int	Current Site state. Can be 1 - Active, 2 - Disabled, 3 – Suspended. Active state signs that any operation with site is available. Disabled state means that all automated operations like crawling, processing and so on are not applicable for this site and it will not be included in selection. Suspended state means that not all, but some periodic active operations are not available for that site, but regular way it is served as usual. Also, the Disabled state can be related with user owner and per user usage extends permissions for operations available for user.
Priority	Int	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.
MaxURLs	longint	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.
MaxURLsFromPage	Longint	Limit of max unique URLs that can be collected from one HTML page or RSS feed. Zero value means unlimited.
MaxResources	Longint	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	Int	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	LongInt	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	Int	Delay before crawling request, ms. Used to make crawling process more smooth and balanced.
ProcessingDelay	Int	Delay before resource data processing, ms.
HTTPTimeout	Int	Timeout of HTTP response, ms.
ErrorMask	LongInt	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	LongInt	Errors counter, represents total number of error happened during the Site usage from NEW state or state was changed like re-crawl.

Size	Longint	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	Float	Average crawling speed, bytes per second or BPS rate. Calculated for period of the Site usage for all resources.
AVGSpeedCounter	Longint	Counter of times when average speed rate was calculated.
URLType	Itr	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.
User_Id	Int	Unique user Id. Used by client API to identify owner user and permissions restrictions and ACL.
RecrawlPeriod	Int	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	DataTime	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.
FetchType	Int	Type of fetcher used for site. Can be 1 - static (default), 2 - dynamic, 3 – external. The static means that resources are static and represent some formats of documents or web-resources that are not require to be rendered before structure will be parsed. Typically it is the generated HTML pages. The dynamic means that resources requires additional rendering before structure will be parsed. Typically it is the HTML pages with javascript that is used to modify the DOM of the page. For this resources used render machine and document structure processing heavier. The external means that some HTTP URL will be used to fetch rendered content from the external source.

`dc_sites`.`sites_filters`

The filters are list of dependent objects that used as a set of properties in crawling process on different stage of URLs and another detectable data processing to accept or reject some item. Filters created as a set of records during SiteNew operation.

Name	Type	Description
Site_Id	String 32 char	Id of the site
Pattern	String 4K	The pattern string with expression used depends on Type and Mode and business logic usage.
Type	Int	0 – include, 1 – exclude behavior depends on Mode and business logic usage
UDate	DateTime	Update date.
CDate	DateTime	Creation date.
Mode	Int	0 – URLs of site, 1 – URLs of media content.

`dc_sites`.`sites_properties`

The properties are list of dependent objects that used as a universal key-value data on different stage of the Site object life-time. Properties created as a set of records during SiteNew operation.

Name	Type	Description
Site_Id	String 32 char	Id of the site
URLMd5	String 32 char	URL Id from URLs table in case of properties applicable only for one URL
Name	String 64 char	Name of a property, used as unique identifier per site or URL object.
Value	String 8K	Value of a property, used depending on target business logic.
UDate	DateTime	Update date
CDate	DateTime	Creation date

`dc_sites`.`sites_urls`

The URLs is a list of root URLs that used to start the crawling process for the Site object. URLs created as a set of records during SiteNew operation.

Name	Type	Description
Site_Id	String 32 char	Id of the site
URL	String 4K	URL string
State	Int	0 – enabled, 1 – disabled, 2 – error, used to control of usage of Site's URLs
Crawled		Deprecated
Processed		Deprecated
CrDate		Deprecated
PrDate		Deprecated
CDate	DateTime	Creation date
User_Id	Bigint	User Id, used to identify the user correspondence in external system and as a criterion of selection of the Site object's related lists.

The resource object

The resource object represents web-resource that is registered inside the DC service. It can have several statuses: registered URL, crawled resource raw data on this host, crawled resource on another host, processed resource key-value db storage data. Depends on the status it represented as a record in the SQL db only, and as a raw data files in the file storage and record in the key-value storage. The resource objects are stored in the dedicated tables and database `dc_urls`.`urls_<SITE_ID_MD5>`.

Name	Type	Description
Site_Id	String 32 char	Id of the site

URL		The HTTP URL link
Type		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
State		0 - Enabled, 1 - disabled, 2 – error
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.
Crawled		Number of crawled times.
Processed		Number of processed times.
URLMd5		The md5 from the URL field
ContentType		MIME content-type string, can be changed while resource processing cause sequential usage of different detection algorithms.
RequestDelay		Delay before the HTTP request.
ProcessingDelay		Delay before the content processing.
HTTPTimeout		HTTP response timeout, msec
Charset		The charset, can be changed while resource processing cause sequential usage of different detection algorithms.
Batch_Id		The batch task Id from the DTM service, corresponds to the DRCE task id for the hce-node cluster.
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.
CrawlingTime		Time of crawling, msec
ProcessingTime		After crawling processing time, msec
TotalTime		Total time – crawling + processing, msec
HTTPCode		HTTP response code
HTTPMethod		The HTTP method used to fetch resource, GET, POST or another.
Size		Resource size, byte
LinksI		Number of internal links
LinksE		Number of external links
Freq		Frequency of usage of the page URL on other pages of this site that was already crawled.
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.
RawContentMd5		The md5 of raw content file data
ParentMd5		The md5 of the parent page URL, used as unique Id of the parent resource from this resource's URL reference was taken.
LastModified		The "Last-Modified" HTTP header's field value
ETag		The "Etag" HTTP header's field value
MRate		AVG mutability rate, relative value calculated to be used as a measure of frequency of the page changes content.
MRateCounter		Counter for AVG mutability rate calculation
UDate		Update date
CDate		Creation date
TcDate		Touch date, updated when some action performed with this URL, for example – the crawling or processing.

MaxURLsFromPage		Limit max URLs that can be collected from page. 0- means unlimited
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.
TagsCount		The counter of detected tags by the processing algorithm named the scraping.
PDate		The exact date of publication of article detected by the scraping processing.
ContentURLMd5		The md5 calculated according the correspondent site's properties settings based on the scraped fields string.

Depends on crawling strategy and settings of sites database records can be managed different way and at different time. To get understanding the behavior of some field and/or record sees the architecture document specifications on algorithm and source code.