



Automated and AI-based Web Data Extraction and the Practical Use Cases

Lively Impact Technology Limited

Lively Impact

Background

A technology company

- HQ in Science Park
- Established in 2011
- Providing leading data and AI platforms, solutions and services
- HKSTP incubation graduate and acceleration graduate
- ICT awards (on big data solutions)

Award Winning Products and Technologies

International and Regional Awards



Cloudera Data
Impact Award 2013
Top 3



TechCrunch BJ 2015
Top 15 Start-up



IOT GD-HK Most
Creative Award

Local Awards



ITC SERAP Awardee



HKSTP LEAP Graduate



HKSTP Incubatee
Graduate



ICT Award 2015
(Silver)
Big Data Stream



ICT Award 2017
(Certificate of Merit)
Big Data Stream

Lively Impact

Our satisfied customers from different verticals and horizontals



Internet & Computer Companies

Public bodies



Education

Brands and Retails

Banking & Finance

Web Data Extraction – What is that?

An overview about Web Extraction

For businesses of all shape and sizes, whether start-ups or Fortune 100s, scraping the web for data to **fuel your data research efforts** offers the broadest and most insightful perspective of your industry.



Manually acquiring data for market research is a mundane, arduous task - one, fortunately, it can be **automated by intelligently designed web crawlers**.



Market Trend Analysis



Price Monitoring



Research and Development



Competitor Analysis



News/Alerts Monitoring



Profile Analysis

Web Data Extraction – Our Solution

A data automation platform to crawl data from websites continuously



Unstructured data
(the above shows example websites only)



Data Automation Platform
Automated data crawlers
that run on schedule

Data View - FIS_F26_Result_URL API Fetching Back

Simple Query

9256 results in 0.035 seconds

Key	Value	HttpStatus	HttpMessage	WebSiteNo	ReportDate	URL
F26_2021-10-19-14-49_https://inspection.canada.ca/food-recall-warnings-and-allergy-alerts/eng/1351519587174/1351519588221	{'level_1': {'Topic': '', 'Report Date': '', 'URL': '', 'Reporting Authority/Organization': ''}, 'level_2': {'Food Item Involved': '', 'Reason for Recall': '', 'Hazard': '', 'Related Reference': '', 'Product Detail': ''}, 'level_3': {'Notice Title': '', 'Notice Number': ''}, 'Surveillance Report': '', 'report': {'Activity ID': 'F26', 'Source': 'https://inspection.canada.ca/food-recall-warnings-and-allergy-alerts/eng/1351519587174/1351519588221', 'URL': '', 'Commencement Time': '2021-10-19T14:49:26', 'Completion Time': '2021-10-19T14:49:30', 'Status Code': 300, 'Message': ''}, 'files': []}	200	Logged	F26	2021-10-19-14-49	https://inspection.canada/warnings-and-allergy-alerts/eng/135151958717
F26_2021-10-19-17-30_https://inspection.canada.ca/food-recall-warnings-and-allergy-alerts/eng/1351519587174/1351519588221	{'level_1': {'Topic': '', 'Report Date': '', 'URL': '', 'Reporting Authority/Organization': ''}, 'level_2': {'Food Item Involved': '', 'Reason for Recall': '', 'Hazard': '', 'Related Reference': '', 'Product Detail': ''}, 'level_3': {'Notice Title': '', 'Notice Number': ''}, 'Surveillance Report': '', 'report': {'Activity ID': 'F26', 'Source': 'https://inspection.canada.ca/food-recall-warnings-and-allergy-alerts/eng/1351519587174/1351519588221', 'URL': '', 'Commencement Time': '2021-10-19T17:30:27', 'Completion Time': '2021-10-19T17:30:30', 'Status Code': 300, 'Message': ''}}	200	Logged	F26	2021-10-19-17-30	https://inspection.canada/warnings-and-allergy-alerts/eng/135151958717

Structured data ready for consumption
(through files, API, webhooks, etc.)

Web Data Extraction Solution – Data Automation Platform

Data Warehouse – Keeping all data in a repository

Hi, admin SYSTEM_ADMIN 120751

Data Warehouse

An integrated place to store all your data

[Create Data Entity](#)

Data Warehouse Refresh Delete + Create

All Filter by Status All Filter by Tag Search Search in all fields

<input type="checkbox"/>	Name	Tag	Owner	Dataset	No. of Record	Update At	Status	Actions
<input type="checkbox"/>	FIS_F10_Result_URL	Tag	admin		25907	15/08/2022(Wed) - 09:13AM	PRODUCTION READY	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F11_Result_URL	Tag	admin		10208	15/08/2022(Wed) - 09:12AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F14_Result_URL	Tag	admin		8856	15/08/2022(Wed) - 09:12AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F17_Result_URL	Tag	admin		9285	15/08/2022(Wed) - 09:12AM	PRODUCTION READY	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F18_Result_URL	Tag	admin		9119	15/08/2022(Wed) - 09:12AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F13_Result_URL	Tag	admin		8596	15/08/2022(Wed) - 09:12AM	PRODUCTION READY	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F19_Result_URL	Tag	admin		9188	15/08/2022(Wed) - 09:11AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F09_Result_URL	Tag	admin		8966	15/08/2022(Wed) - 09:08AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F05_Result_URL	Tag	admin		9087	15/08/2022(Wed) - 09:08AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete
<input type="checkbox"/>	FIS_F10_Result_URL	Tag	admin		9332	15/08/2022(Wed) - 09:08AM	PRODUCTION READY WITH CHANGES	Search Folder Edit Settings Share Delete

« < 1 2 3 4 ... 7 > »

10 Showing rows 11 to 20 of 67

Web Data Extraction Solution – Data Automation Platform

Data Warehouse – Keeping all data in a repository



Hi, admin SYSTEM_ADMIN 120751

Data View - FIS_F27_Result_URL

Simple Query API Fetching Back

8931 results in 0.002 seconds

Key	Value	Http Status	HttpMessage	Web SiteNo	ReportDate	URL
F27_2021-10-19-14-50_https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1	{'level_1': {'Topic': '', 'Report Date': '', 'URL': '', 'Reporting Authority/Organization': ''}, 'level_2': {'Food Item Involved': '', 'Reason for Recall': '', 'Hazard': '', 'Related Reference': '', 'Product Detail': ''}, 'level_3': {'Notice Title': '', 'Notice Number': '', 'Surveillance Report': ''}, 'report': {'Activity ID': 'F27', 'Source': 'https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1', 'URL': '', 'Commencement Time': '2021-10-19T14:50:27', 'Completion Time': '2021-10-19T14:50:33', 'Status Code': 300, 'Message': '', 'files': []}}	200	Logged	F27	2021-10-19-14-50	https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1
F27_2021-10-19-17-30_https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1	{'level_1': {'Topic': '', 'Report Date': '', 'URL': '', 'Reporting Authority/Organization': ''}, 'level_2': {'Food Item Involved': '', 'Reason for Recall': '', 'Hazard': '', 'Related Reference': '', 'Product Detail': ''}, 'level_3': {'Notice Title': '', 'Notice Number': '', 'Surveillance Report': ''}, 'report': {'Activity ID': 'F27', 'Source': 'https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1', 'URL': '', 'Commencement Time': '2021-10-19T17:30:27', 'Completion Time': '2021-10-19T17:30:33', 'Status Code': 300, 'Message': '', 'files': []}}	200	Logged	F27	2021-10-19-17-30	https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1
F27_2021-10-19-18-30_https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1	{'level_1': {'Topic': '', 'Report Date': '', 'URL': '', 'Reporting Authority/Organization': ''}, 'level_2': {'Food Item Involved': '', 'Reason for Recall': '', 'Hazard': '', 'Related Reference': '', 'Product Detail': ''}, 'level_3': {'Notice Title': '', 'Notice Number': '', 'Surveillance Report': ''}, 'report': {'Activity ID': 'F27', 'Source': 'https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1', 'URL': '', 'Commencement Time': '2021-10-19T18:30:28', 'Completion Time': '2021-10-19T18:30:34', 'Status Code': 300, 'Message': '', 'files': []}}	200	Logged	F27	2021-10-19-18-30	https://healthycanadians.gc.ca/recall-alert-rappel-avis/index-eng.php?cat=1


Web Data Extraction Solution – Data Automation Platform

Data Pipelines – Run all data pipelines systematically with transparency



- Dashboard
- Data Warehouse
- Data Pipeline**
- Smart Query
- Smart Search
- User & Group
- Access Control
- My Profile
- Admin Setting
- Logout

Hi, admin SYSTEM_ADMIN 120751



Name: FIS_F32

Output Status

History: 2572086 (15 Jun 2022 8:51:14)

Key	Value	Total
F32_2022-06-15-08-20_https://www.sfa.gov.sg/food-information/food-alerts-recalls	{level_1: {Topic: ', Report Date: ', URL: ', Reporting Authority/Organization: ', level_2: {Food Item Involved: ', Reason for Recall: ', Hazard: ', Related Reference: ', Product Detail: ', level_3: {Notice Title: ', Notice Number: ', Surveillance Report: ', report: {Activity ID: F32; Source: https://www.sfa.gov.sg/food-information/food-alerts-recalls; URL: ', Commencement Time: '2022-06-15T08:20:04'; Completion Time: '2022-06-15T08:20:05'; Status Code: 300; Message: ', files: []}	1

Data Console

- View the status (success/failed) for the whole data pipeline or individual components
- View the data that pass through each component in a data pipeline

Web Data Extraction Solution – Our Differentiation

Proprietary technologies to enable data crawling at scale even on difficult sites



Anti-ban

Strategies to emulate a human visit session to avoid banning



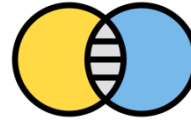
Natural Language Analysis

Extract key phrases, key sentences and perform summarisation if needed



Auto-queuing

Queue up in virtual waiting room just like a human



Data Change Detection

Extract delta change in data to minimize data crawling workload



Login

Login with a valid credential or handle session-related mechanics



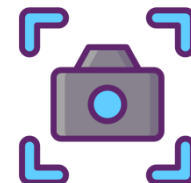
Rotational Proxy

Visit websites from a large pool of IP to decrease latency and improve success rate



Deep Crawling

Crawl into the documents and get the information inside



Screen Capture

Save the screen in PDF file for historical snapshot of the website

Web Data Extraction Solution – Our Differentiation

Proprietary technologies to enable data crawling on difficult websites



The image displays a browser window with search results for '交友' (social networking) on the left and a DevTools console on the right. The search results show a list of posts with filters like 'All', 'Posts', 'Recent posts', etc. The DevTools console shows network logs for 'total posts' and 'Canvas2D: Multiple readback operations using getImageData are faster with the willReadFrequently attribute set to true.' The console also shows a list of posts with their content and metadata.

Note: Showing an anti-blocking mechanism for website such as Facebook

Web Data Extraction Solution – Deep AI integration

Enrich the crawled data with more contextual intelligence – content summarization



Photo illustration

Fanta recalled after full sugar variety labeled as zero sugar

By News Desk on June 27, 2023

Great Lakes Coca-Cola Distribution LLC, of Niles, IL, is recalling over 14,000 12-pack cases of 12 oz. Orange Fanta cans because the packages are labeled as zero sugar, but they contain the full sugar variety.

For some people with certain underlying conditions, too much sugar can lead to serious health problems.

According to the details posted online by the FDA, the recall was initiated on June 1 and is ongoing.

The recalled product was distributed in California.

Recalled product:

- Fanta Orange 12pk (12, 12 oz. cans) Wrap
- UPC 0 4900003073 0
- Date Code: FEB1224DDB1956
- 14074 cases

Anyone who purchased the recalled product intending for it to be a sugar-free product should not consume it.

<https://www.foodsafetynews.com/2023/06/fanta-recalled-after-full-sugar-variety-labeled-as-zero-sugar/#more-229196>

Web Data Extraction Solution – Deep AI integration

Enrich the crawled data with more contextual intelligence – content summarization

▼ **Prompt**

Provide a brief summary for the following article:

Fanta recalled after full sugar variety labeled as zero sugar
By News Desk on June 27, 2023
Great Lakes Coca-Cola Distribution LLC, of Niles, IL, is recalling over 14,000 12-pack cases of 12 oz. Orange Fanta cans because the packages are labeled as zero sugar, but they contain the full sugar variety.

For some people with certain underlying conditions, too much sugar can lead to serious health problems.

According to the details posted online by the FDA, the recall was initiated on June 1 and is ongoing.

The recalled product was distributed in California.

▼ **Response** Markdown

Fanta Orange 12pk (12, 12 oz. cans) Wrap with UPC 0 4900003073 0 and Date Code: FEB1224DDB1956 is being recalled because it is labeled as zero sugar, but it contains the full sugar variety.

i We want your [feedback](#).

Model

Temperature 0.2

Token limit 768

Top-K 40

Top-P 0.8

Safety filter threshold

SUBMIT **RESET PARAMETERS**

Web Data Extraction Solution – Deep AI integration

Enrich the crawled data with more contextual intelligence – content summarization



Photo illustration

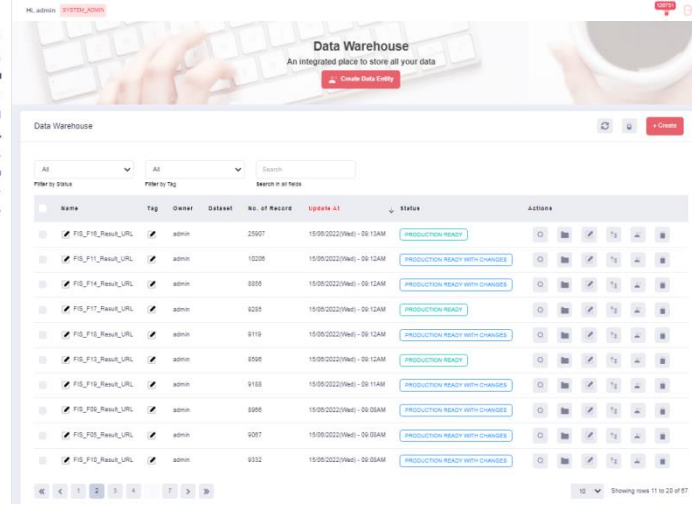
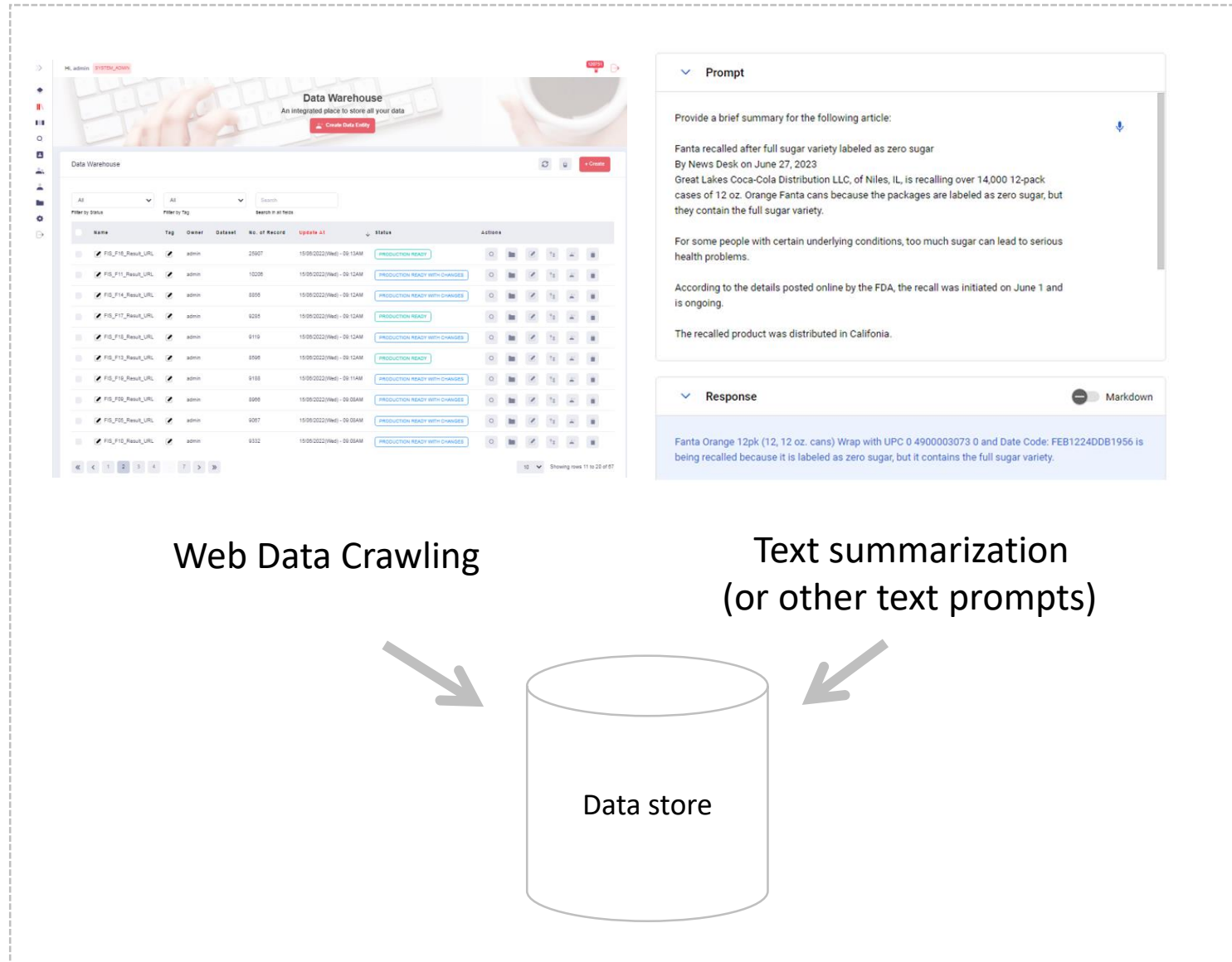
Fanta recalled after full sugar variety labeled as zero sugar

By News Desk on June 27, 2023

Great Lakes Coca-Cola Distribution LLC, of Niles, IL, is recalling over 14,000 12-pack cases of 12 oz. Orange Fanta cans because the packages are labeled as zero sugar, but they contain the full sugar variety.

For some people with certain underlying conditions, too much sugar can lead to serious health problems.

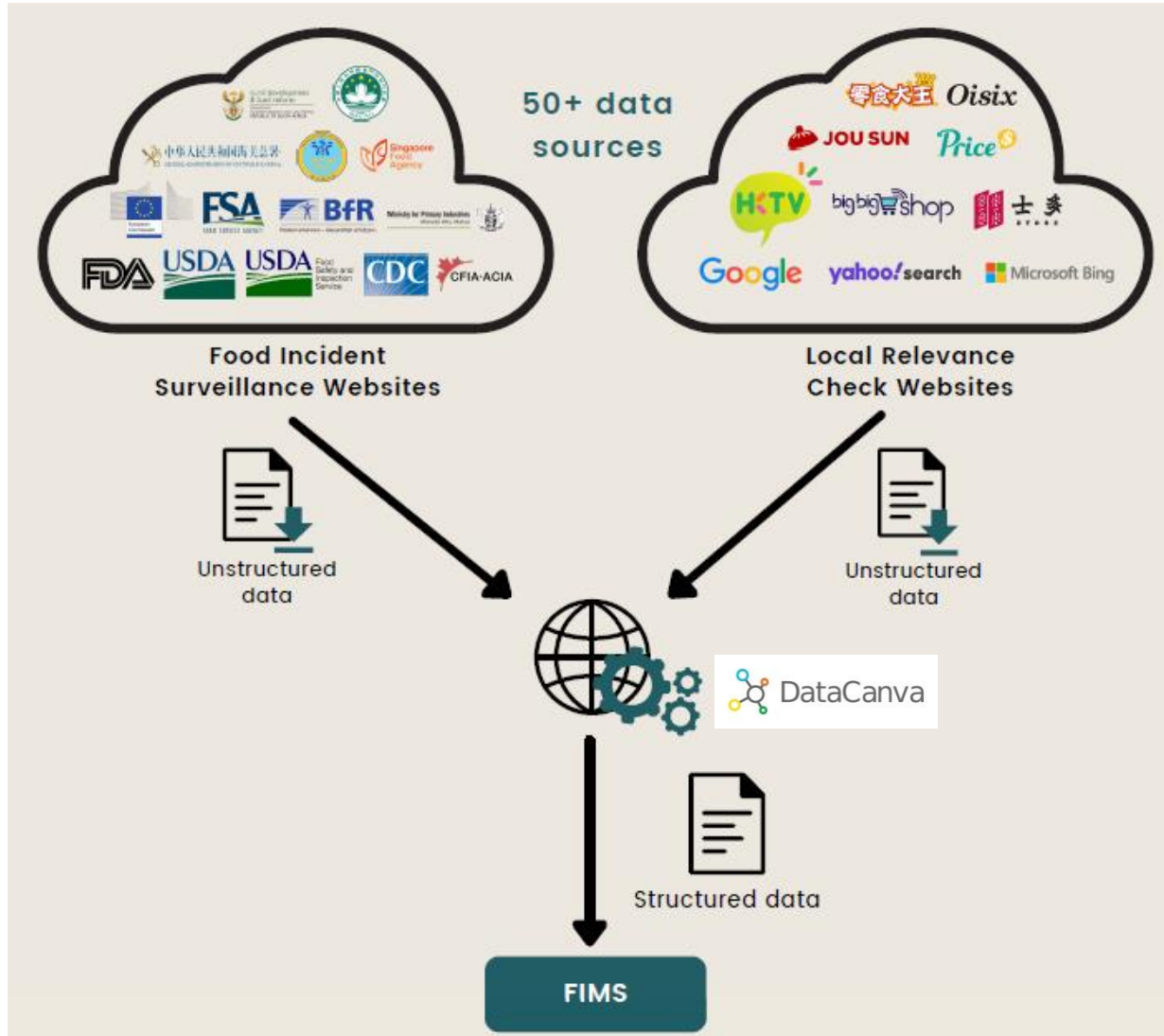
News Article



Name	Tag	Owner	Dataset	No. of Record	Update At	Status	Actions
FIS_F12_Reau_LURL		admin		2507	15/09/2022(09h45) - 09:15AM	PRODUCTION READY	[Icons]
FIS_F11_Reau_LURL		admin		10208	15/09/2022(09h45) - 09:15AM	PRODUCTION READY WITH CHANGES	[Icons]
FIS_F14_Reau_LURL		admin		8810	15/09/2022(09h45) - 09:15AM	PRODUCTION READY WITH CHANGES	[Icons]
FIS_F17_Reau_LURL		admin		8235	15/09/2022(09h45) - 09:15AM	PRODUCTION READY	[Icons]
FIS_F13_Reau_LURL		admin		9119	15/09/2022(09h45) - 09:15AM	PRODUCTION READY WITH CHANGES	[Icons]
FIS_F15_Reau_LURL		admin		8995	15/09/2022(09h45) - 09:15AM	PRODUCTION READY	[Icons]
FIS_F18_Reau_LURL		admin		9188	15/09/2022(09h45) - 09:11AM	PRODUCTION READY WITH CHANGES	[Icons]
FIS_F20_Reau_LURL		admin		8998	15/09/2022(09h45) - 09:05AM	PRODUCTION READY WITH CHANGES	[Icons]
FIS_F25_Reau_LURL		admin		9267	15/09/2022(09h45) - 09:05AM	PRODUCTION READY WITH CHANGES	[Icons]
FIS_F12_Reau_LURL		admin		9312	15/09/2022(09h45) - 09:05AM	PRODUCTION READY WITH CHANGES	[Icons]

Web Data Extraction Solution – Case Study

Data Extraction for Centre for Food Safety (FEHD)



Background

CFS is made aware of food incidents through various sources, such as foreign food safety authorities, World Health Organization (WHO), to identify food incidents that may have potential impact on the public or Hong Kong's food supply.



Before/After Our Solution

Before our solution, the CFS staff manually monitor a predetermined list of websites every day. To meet the needs of web-based surveillance, an automated and system approach is adopted, crawling 40+ websites.



Results

Keeping data freshness in 1 hour interval automatically, improving alertness level by almost 90% (from usually 8 hours to 1 hour), saving days in manpower every month.

Web Data Extraction Solution – Case Study

Data Extraction for Cyber Security Information Sharing (Cybersec InfoHub) (OGCIO)



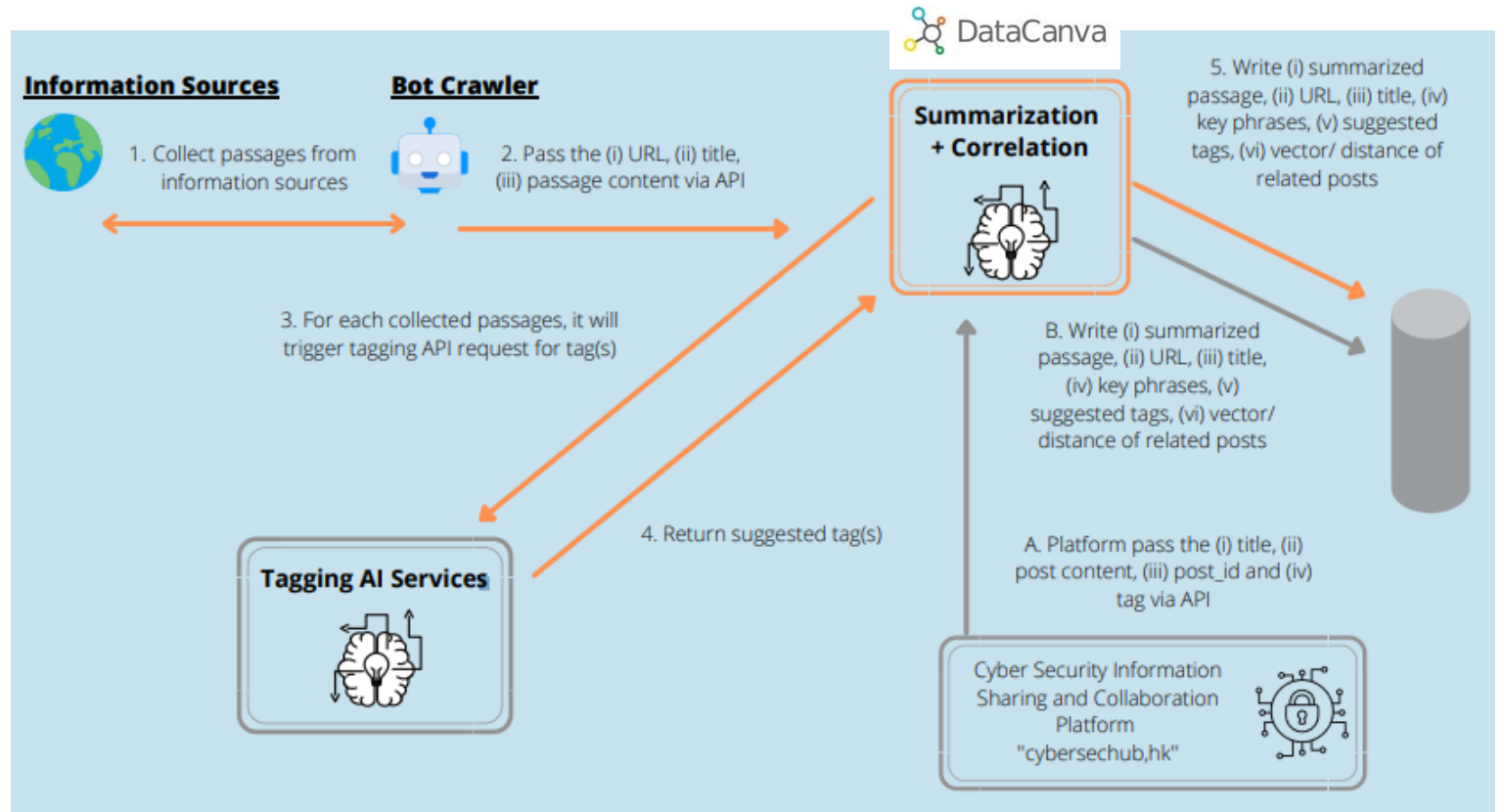
Background

Cyber attacks continue to increase in frequency and sophistication, presenting significant challenges for organisations to protect their digital assets and information systems. It is difficult for individual organisations to scout the Internet continuously by themselves for all happenings in the cyberspace and to find out meaningful learning and formulate informed decisions to guard against cyber threats.



Before/After Our Solution

Before our solution, it relies to practitioners to report cyber security alerts. To meet the needs of web-based surveillance, an automated and system approach is adopted, crawling 20+ websites. In addition, key phrases, sentences and summarisation are automatically done.



Results

Keeping data freshness in 1 hour interval automatically, improving alertness level by 90% and more (from usually 24 hours to 1 hour), saving days in manpower every month.

**Automated and AI-based
Web Data Extraction
and the Practical Use Cases**

Lively Impact Technology Limited

Email: ivan.ng@livelyimpact.com (Founder)

Tel: 92761341 (mobile), 34269508 (office)

