Scraping infrastructure management and solutions

Continuously yield high quality data while growing from 100 to 100M daily requests

Glen Henri J. De Cauwsemaecker
Lead Crawler Engineer @ OTA Insight Ltd.

100,000,000 requests / day
~ 300 data sources
9000 GiB (compressed) raw data

(2022)

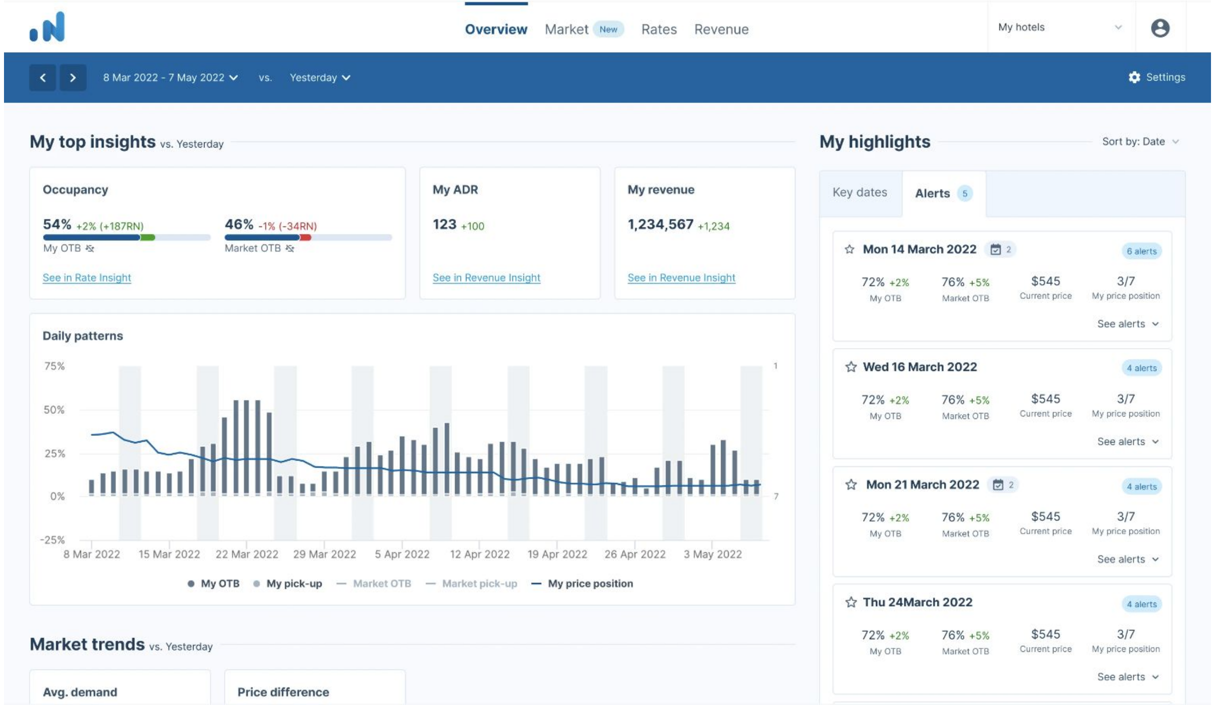
100 requests / day
(2013)

Credits: SpaceX

OTA Insight Ltd.

We deliver smarter revenue, distribution and marketing outcomes.

365+ employees
~ 80 engineers
5 crawler engineers



My top insights vs. Yesterday

Occupancy
54% +2% (+187RN) My OTB
46% -1% (-34RN) Market OTB
See in Rate Insight

My ADR
123 +100
See in Revenue Insight

My revenue
1,234,567 +1,234
See in Revenue Insight

Daily patterns

Market trends vs. Yesterday

Avg. demand

Price difference

My highlights

Sort by: Date

Key dates

Alerts 5

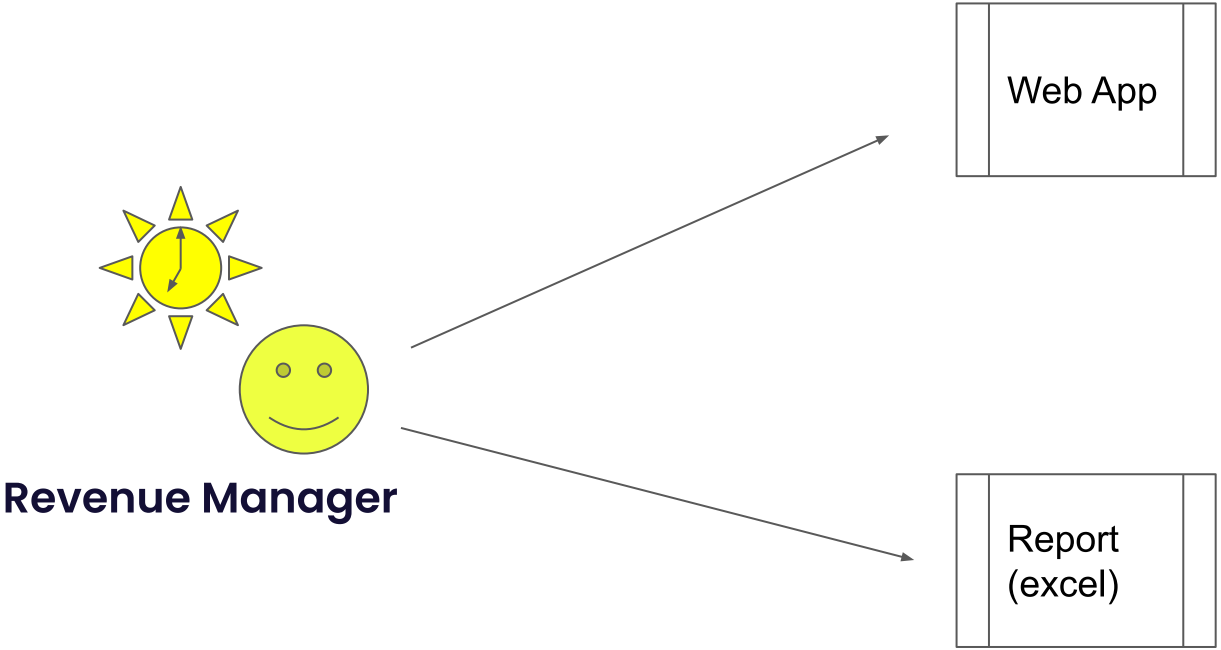
Mon 14 March 2022 2 — 6 alerts
72% +2% My OTB
76% +5% Market OTB
$545 Current price
3/7 My price position
See alerts

Wed 16 March 2022 — 4 alerts
72% +2% My OTB
76% +5% Market OTB
$545 Current price
3/7 My price position
See alerts

Mon 21 March 2022 2 — 4 alerts
72% +2% My OTB
76% +5% Market OTB
$545 Current price
3/7 My price position
See alerts

Thu 24March 2022 — 4 alerts
72% +2% My OTB
76% +5% Market OTB
$545 Current price
3/7 My price position
See alerts

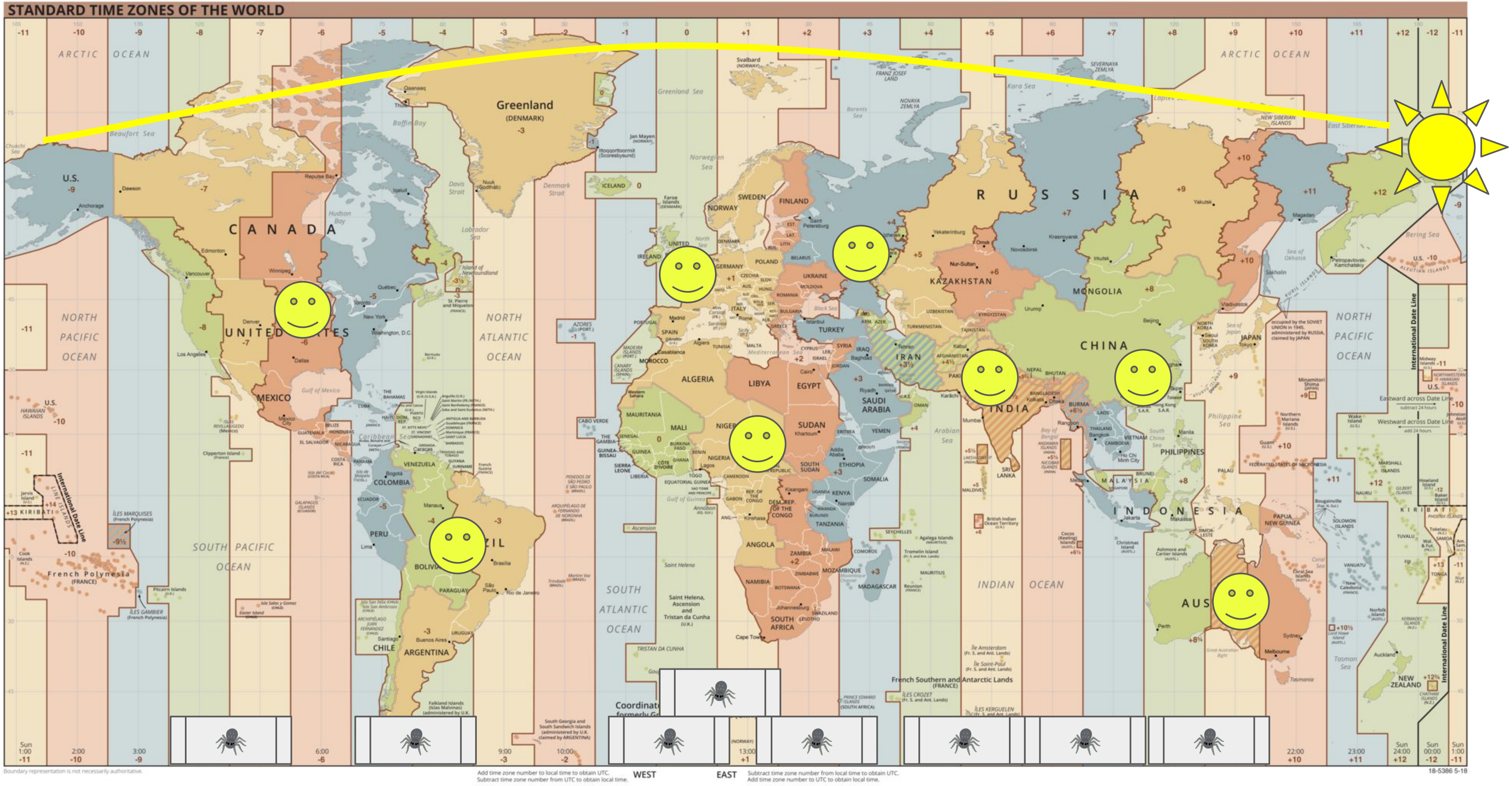
Hotel Rates and Ranking Information



Web App

Revenue Manager

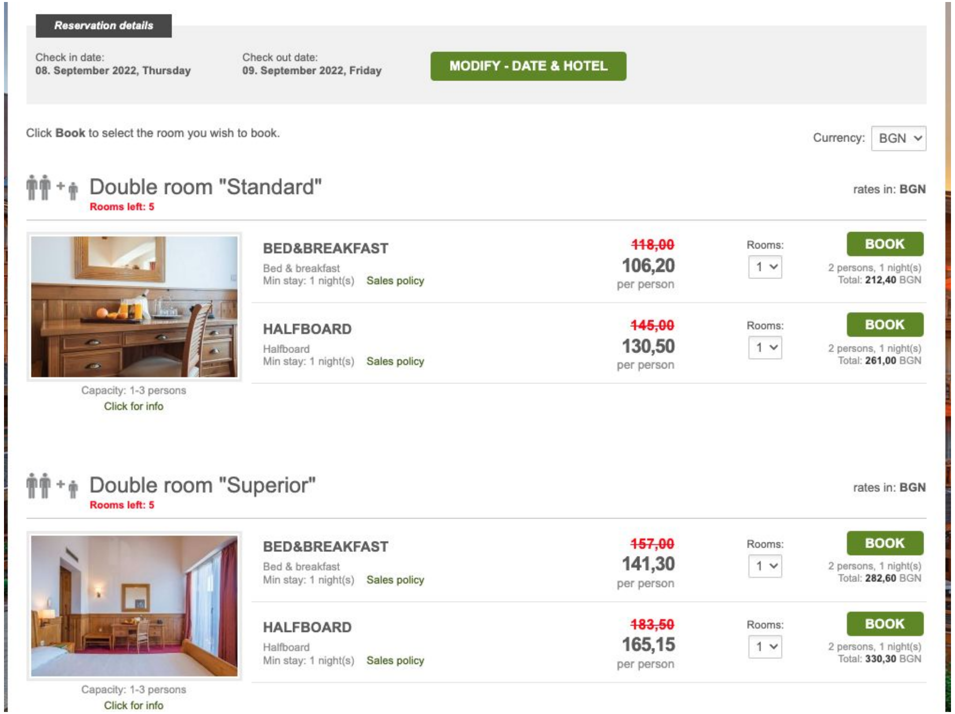
Report (excel)



TimeZonesBoy - US Central Intelligence Agency

Public Domain

Hmmm... Data...



Reservation details

Check in date:
08. September 2022, Thursday

Check out date:
09. September 2022, Friday

MODIFY - DATE & HOTEL

Click **Book** to select the room you wish to book.

Currency: BGN

Double room "Standard"

Rooms left: 5

rates in: **BGN**

Capacity: 1-3 persons
Click for info

BED&BREAKFAST Bed & breakfast Min stay: 1 night(s) Sales policy	~~118,00~~ 106,20 per person	Rooms: 1	BOOK 2 persons, 1 night(s) Total: **212,40** BGN
HALFBOARD Halfboard Min stay: 1 night(s) Sales policy	~~145,00~~ 130,50 per person	Rooms: 1	BOOK 2 persons, 1 night(s) Total: **261,00** BGN

Double room "Superior"

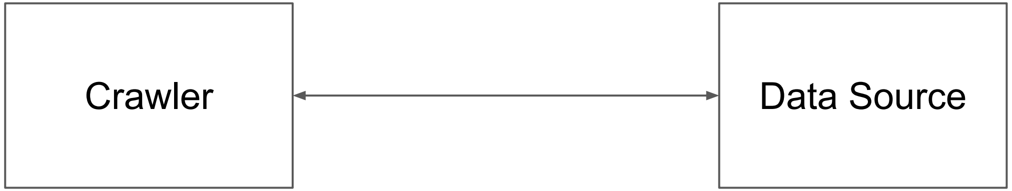
Rooms left: 5

rates in: **BGN**

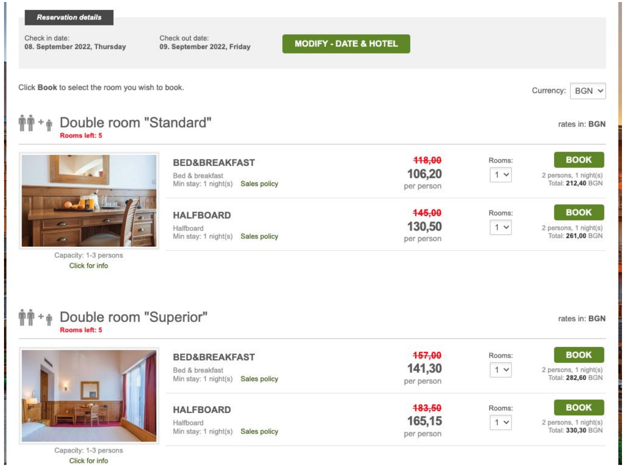
Capacity: 1-3 persons
Click for info

BED&BREAKFAST Bed & breakfast Min stay: 1 night(s) Sales policy	~~157,00~~ 141,30 per person	Rooms: 1	BOOK 2 persons, 1 night(s) Total: **282,60** BGN
HALFBOARD Halfboard Min stay: 1 night(s) Sales policy	~~183,50~~ 165,15 per person	Rooms: 1	BOOK 2 persons, 1 night(s) Total: **330,30** BGN

Data Extraction...



Extract + Transform



Check in date:
08. September 2022, Thursday

Check out date:
09. September 2022, Friday

Click **Book** to select the room you wish to book.

Currency: BGN

Double room "Standard"

Rooms left: 5

rates in: **BGN**

BED&BREAKFAST Bed & breakfast Min stay: 1 night(s) Sales policy	~~118,00~~ **106,20** per person	Rooms: 1	2 persons, 1 night(s) Total: **212,40** BGN
HALFBOARD Halfboard Min stay: 1 night(s) Sales policy	~~145,00~~ **130,50** per person	Rooms: 1	2 persons, 1 night(s) Total: **261,00** BGN

Capacity: 1-3 persons
Click for info

Double room "Superior"

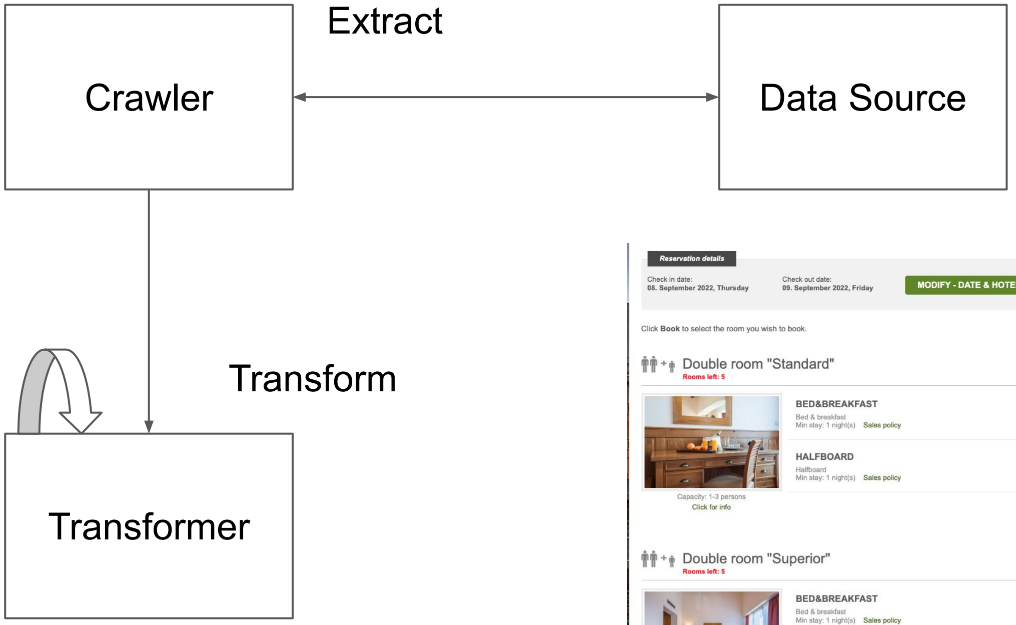
Rooms left: 5

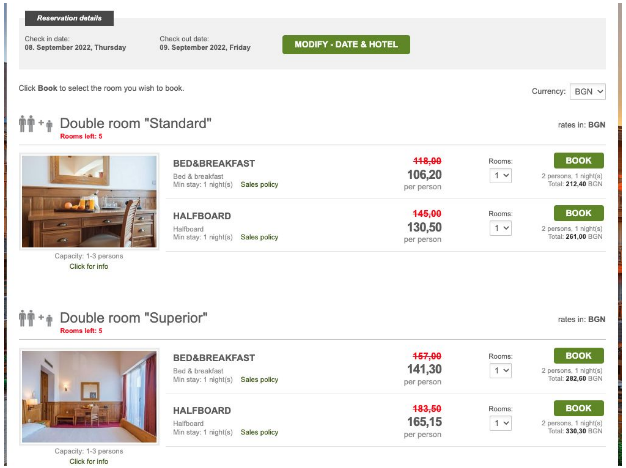
rates in: **BGN**

BED&BREAKFAST Bed & breakfast Min stay: 1 night(s) Sales policy	~~157,00~~ **141,30** per person	Rooms: 1	2 persons, 1 night(s) Total: **282,60** BGN
HALFBOARD Halfboard Min stay: 1 night(s) Sales policy	~~183,50~~ **165,15** per person	Rooms: 1	2 persons, 1 night(s) Total: **330,30** BGN

Capacity: 1-3 persons
Click for info

Data Extraction...





Reservation details

Check in date:
08. September 2022, Thursday

Check out date:
09. September 2022, Friday

Click **Book** to select the room you wish to book.

Currency: BGN

Double room "Standard"

Rooms left: 5

rates in: **BGN**

BED&BREAKFAST
Bed & breakfast
Min stay: 1 night(s) Sales policy
~~118,00~~
106,20
per person
Rooms: 1

2 persons, 1 night(s)
Total: **212,40** BGN

HALFBOARD
Halfboard
Min stay: 1 night(s) Sales policy
~~145,00~~
130,50
per person
Rooms: 1

2 persons, 1 night(s)
Total: **261,00** BGN

Capacity: 1-3 persons
Click for info

Double room "Superior"

Rooms left: 5

rates in: **BGN**

BED&BREAKFAST
Bed & breakfast
Min stay: 1 night(s) Sales policy
~~157,00~~
141,30
per person
Rooms: 1

2 persons, 1 night(s)
Total: **282,60** BGN

HALFBOARD
Halfboard
Min stay: 1 night(s) Sales policy
~~183,50~~
165,15
per person
Rooms: 1

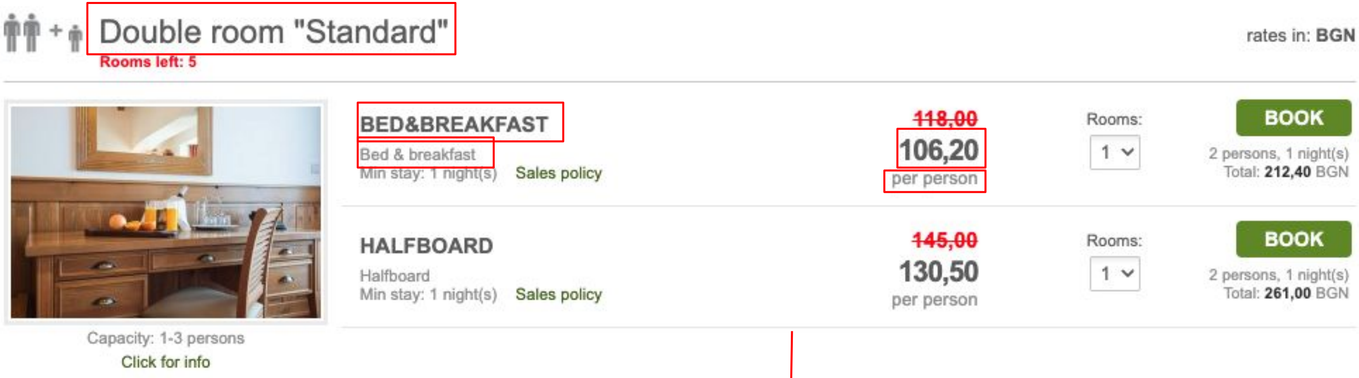
2 persons, 1 night(s)
Total: **330,30** BGN

Capacity: 1-3 persons
Click for info

19222#{137#0#166678 90#19267#1#2####}#1 660788000#0

0 Inputs

1 Extract



Double room "Standard"

Rooms left: 5

rates in: BGN

BED&BREAKFAST

Bed & breakfast

Min stay: 1 night(s) Sales policy

~~118,00~~

106,20

per person

Rooms: 1

BOOK

2 persons, 1 night(s)

Total: **212,40** BGN

HALFBOARD

Halfboard

Min stay: 1 night(s) Sales policy

~~145,00~~

130,50

per person

Rooms: 1

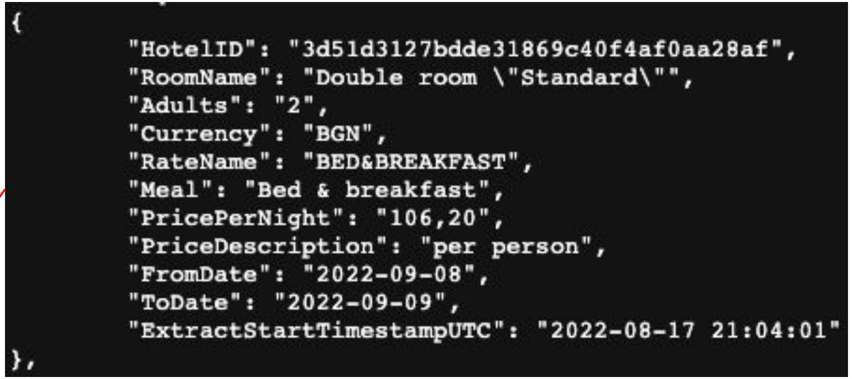
BOOK

2 persons, 1 night(s)

Total: **261,00** BGN

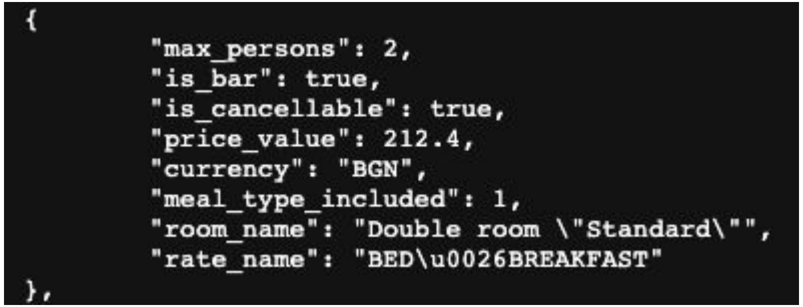
Capacity: 1-3 persons

Click for info



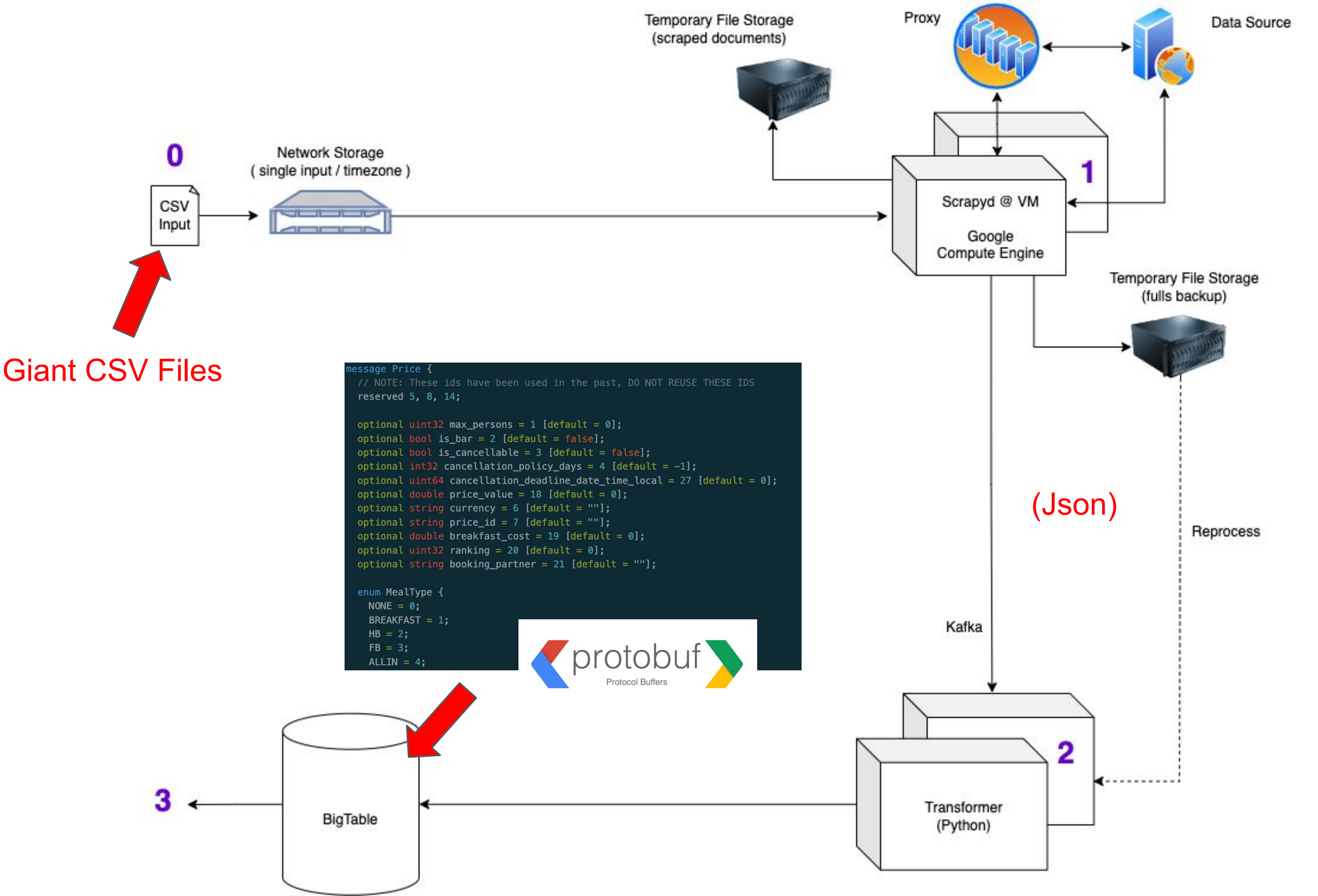
```
{
        "HotelID": "3d51d3127bdde31869c40f4af0aa28af",
        "RoomName": "Double room \"Standard\"",
        "Adults": "2",
        "Currency": "BGN",
        "RateName": "BED&BREAKFAST",
        "Meal": "Bed & breakfast",
        "PricePerNight": "106,20",
        "PriceDescription": "per person",
        "FromDate": "2022-09-08",
        "ToDate": "2022-09-09",
        "ExtractStartTimestampUTC": "2022-08-17 21:04:01"
},
```

Transform 2



```
{
        "max_persons": 2,
        "is_bar": true,
        "is_cancellable": true,
        "price_value": 212.4,
        "currency": "BGN",
        "meal_type_included": 1,
        "room_name": "Double room \"Standard\"",
        "rate_name": "BED\u0026BREAKFAST"
},
```

3



```
message Price {
  // NOTE: These ids have been used in the past, DO NOT REUSE THESE IDS
  reserved 5, 8, 14;

  optional uint32 max_persons = 1 [default = 0];
  optional bool is_bar = 2 [default = false];
  optional bool is_cancellable = 3 [default = false];
  optional int32 cancellation_policy_days = 4 [default = -1];
  optional uint64 cancellation_deadline_date_time_local = 27 [default = 0];
  optional double price_value = 18 [default = 0];
  optional string currency = 6 [default = ""];
  optional string price_id = 7 [default = ""];
  optional double breakfast_cost = 19 [default = 0];
  optional uint32 ranking = 20 [default = 0];
  optional string booking_partner = 21 [default = ""];

  enum MealType {
    NONE = 0;
    BREAKFAST = 1;
    HB = 2;
    FB = 3;
    ALLIN = 4;
```

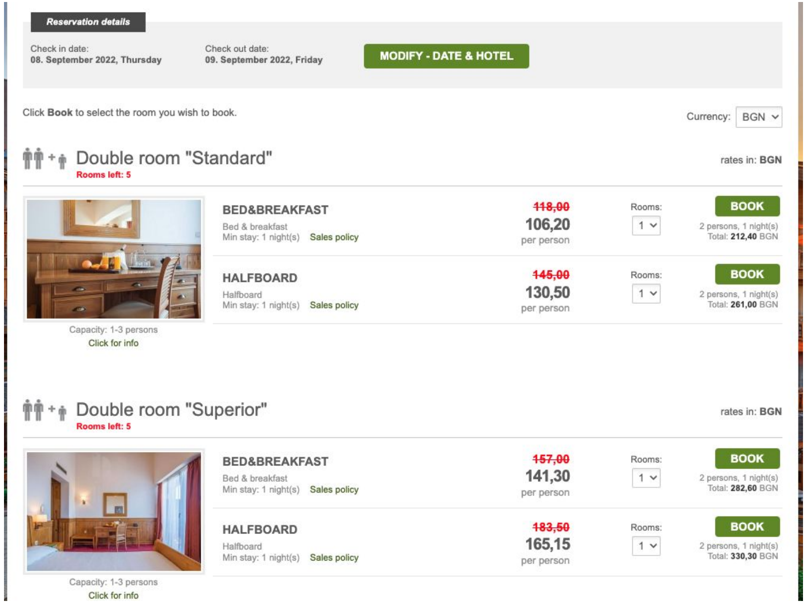
Error Prone

- ⚠️ Manual Intensive
- ⚠️ No Overview
- ⚠️ Costly Retries

Integrities (v1)

`5_1_2_308362_17508_1_rate_17438`

An integrity is one job assignment, and links to one crawl input.



Check in date:
08. September 2022, Thursday

Check out date:
09. September 2022, Friday

Click **Book** to select the room you wish to book.

Currency: BGN

Double room "Standard"
Rooms left: 5

rates in: **BGN**

BED&BREAKFAST
Bed & breakfast
Min stay: 1 night(s) Sales policy
~~118,00~~
106,20
per person
Rooms: 1
2 persons, 1 night(s)
Total: **212,40** BGN

HALFBOARD
Halfboard
Min stay: 1 night(s) Sales policy
~~145,00~~
130,50
per person
Rooms: 1
2 persons, 1 night(s)
Total: **261,00** BGN

Capacity: 1-3 persons
Click for info

Double room "Superior"
Rooms left: 5

rates in: **BGN**

BED&BREAKFAST
Bed & breakfast
Min stay: 1 night(s) Sales policy
~~157,00~~
141,30
per person
Rooms: 1
2 persons, 1 night(s)
Total: **282,60** BGN

HALFBOARD
Halfboard
Min stay: 1 night(s) Sales policy
~~183,50~~
165,15
per person
Rooms: 1
2 persons, 1 night(s)
Total: **330,30** BGN

Capacity: 1-3 persons
Click for info

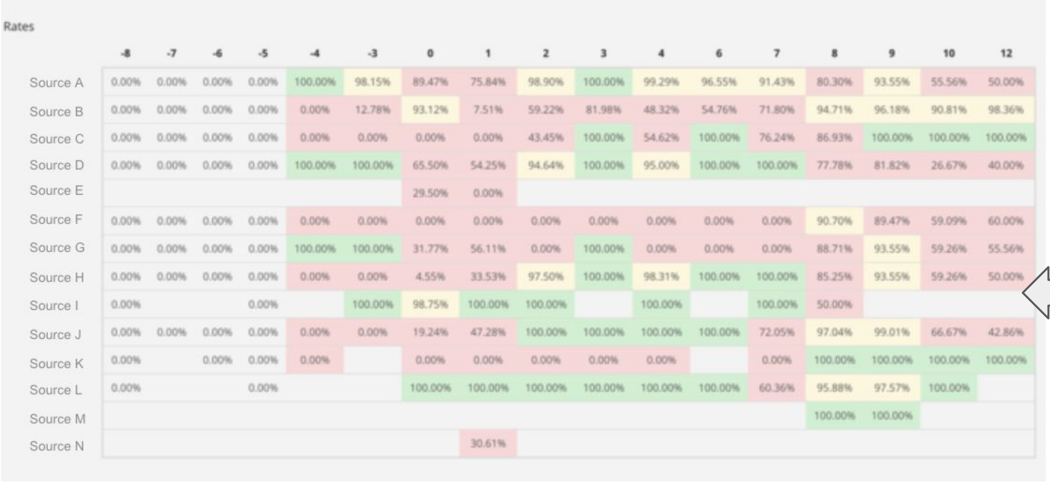
2015

Integrities (v1)

Status tracking per integrity, stored in a MySQL Database:

missing → error or complete

✅ Auto-Retry on Failure

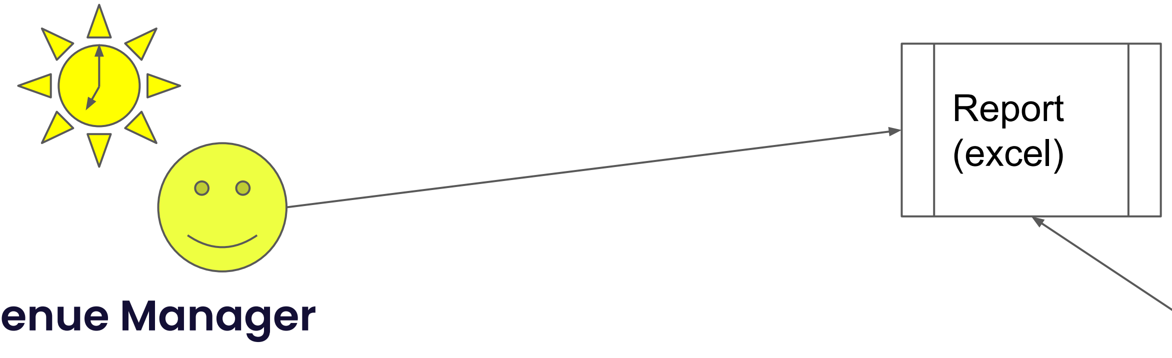


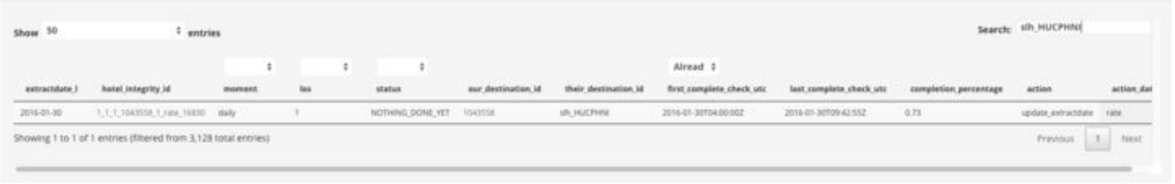
Rates

	-8	-7	-6	-5	-4	-3	0	1	2	3	4	6	7	8	9	10	12
Source A	0.00%	0.00%	0.00%	0.00%	100.00%	98.15%	89.47%	75.84%	98.90%	100.00%	99.29%	96.55%	91.43%	80.30%	93.55%	55.56%	50.00%
Source B	0.00%	0.00%	0.00%	0.00%	0.00%	12.78%	93.12%	7.51%	59.22%	81.98%	48.32%	54.76%	71.80%	94.71%	96.18%	90.81%	98.36%
Source C	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	43.45%	100.00%	54.62%	100.00%	76.24%	86.93%	100.00%	100.00%	100.00%
Source D	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	65.50%	54.25%	94.64%	100.00%	95.00%	100.00%	100.00%	77.78%	81.82%	26.67%	40.00%
Source E							29.50%	0.00%									
Source F	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	90.70%	89.47%	59.09%	60.00%
Source G	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	31.77%	56.11%	0.00%	100.00%	0.00%	0.00%	0.00%	88.71%	93.55%	59.26%	55.56%
Source H	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.55%	33.53%	97.50%	100.00%	98.31%	100.00%	100.00%	85.25%	93.55%	59.26%	50.00%
Source I	0.00%			0.00%		100.00%	98.75%	100.00%	100.00%		100.00%		100.00%	50.00%			
Source J	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	19.24%	47.28%	100.00%	100.00%	100.00%	100.00%	72.05%	97.04%	99.01%	66.67%	42.86%
Source K	0.00%		0.00%	0.00%	0.00%		0.00%	0.00%	0.00%	0.00%	0.00%		0.00%	100.00%	100.00%	100.00%	100.00%
Source L	0.00%			0.00%			100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	60.36%	95.88%	97.57%	100.00%	
Source M														100.00%	100.00%		
Source N								30.61%									

✅ Overview

Grouped Integrities



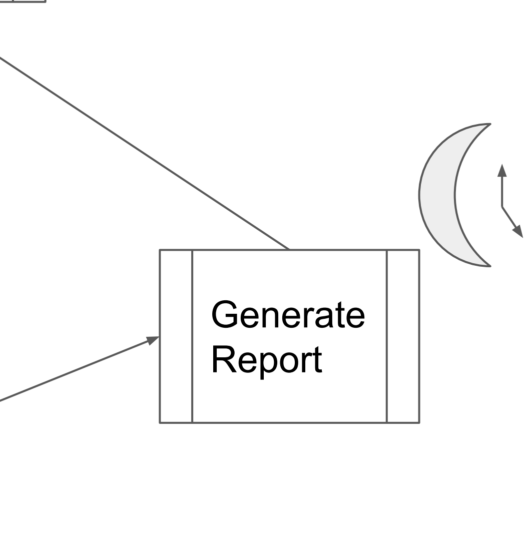

Show 50 entries

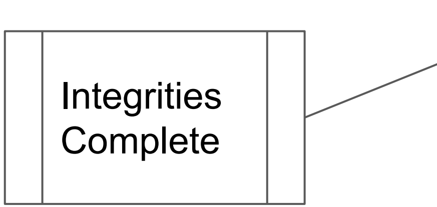
Search: slh_HUCPHM

Alread

extractdate_l	hotel_integrity_id	moment	los	status	our_destination_id	their_destination_id	first_complete_check_utc	last_complete_check_utc	completion_percentage	action	action_dat
2016-01-30	1_1_1_1043558_1_rate_16830	daily	1	NOTHING_DONE_YET	1043558	slh_HUCPHM	2016-01-30T04:00:00Z	2016-01-30T09:42:55Z	0.73	update_extractdate	rate

Showing 1 to 1 of 1 entries (filtered from 3,128 total entries)



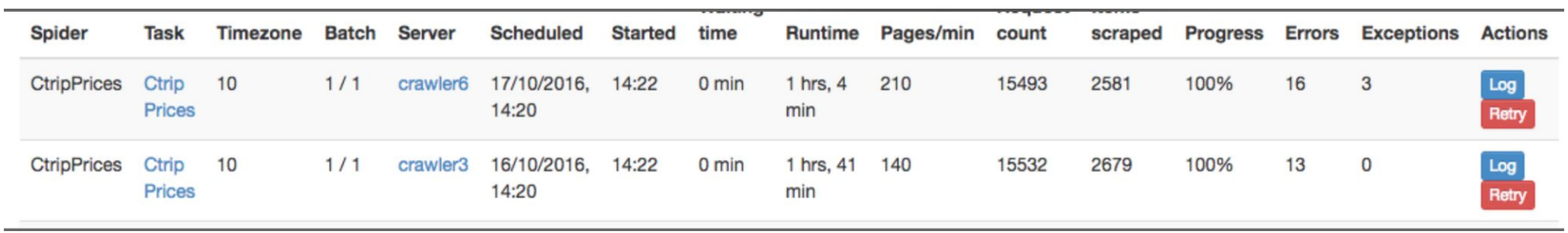


2015

Scheduler (v1)

Introduction of a GUI:

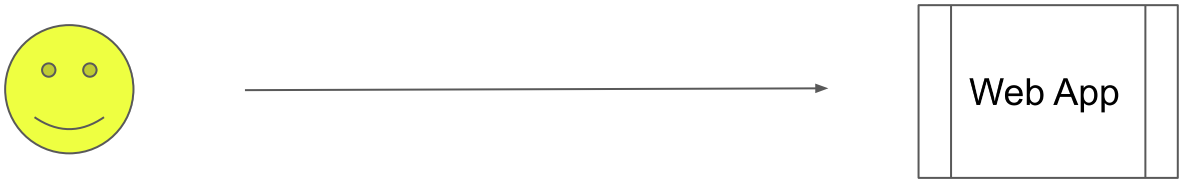
- ✅ High level cross-VM configuration (automate)
- ✅ Ability to orchestrate on a higher level



Spider	Task	Timezone	Batch	Server	Scheduled	Started	Waiting time	Runtime	Pages/min	Request count	Items scraped	Progress	Errors	Exceptions	Actions
CtripPrices	Ctrip Prices	10	1 / 1	crawler6	17/10/2016, 14:20	14:22	0 min	1 hrs, 4 min	210	15493	2581	100%	16	3	Log Retry
CtripPrices	Ctrip Prices	10	1 / 1	crawler3	16/10/2016, 14:20	14:22	0 min	1 hrs, 41 min	140	15532	2679	100%	13	0	Log Retry

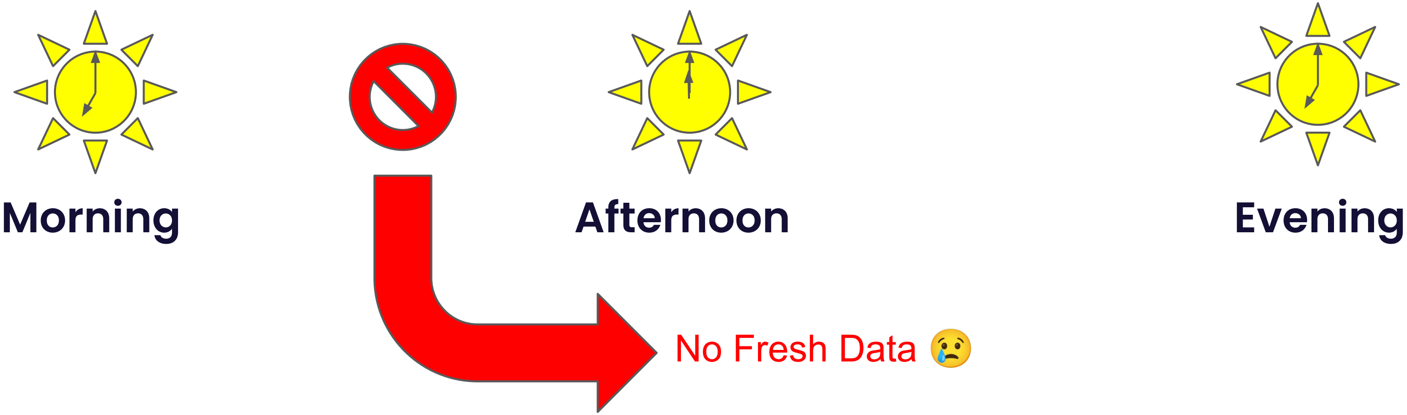
2015

Revenue Manager



Web App

Scheduled Data Pipelines:

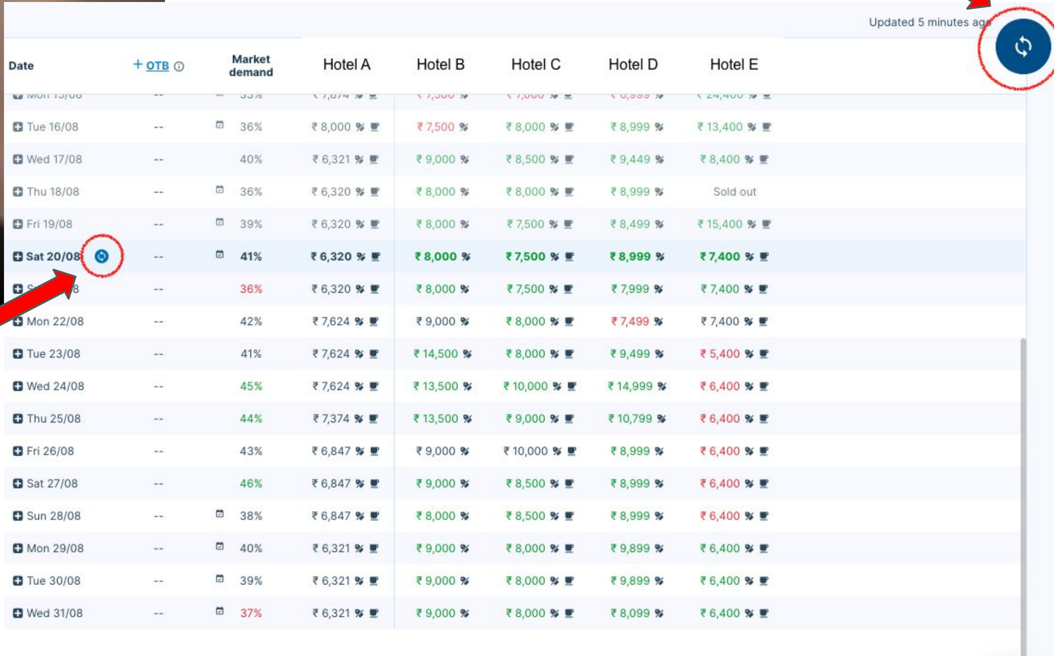


Morning

Afternoon

Evening

No Fresh Data 😢



Updated 5 minutes ago

Date	+ OTB	Market demand	Hotel A	Hotel B	Hotel C	Hotel D	Hotel E
Mon 15/08	--	33%	₹ 7,874	₹ 7,500	₹ 7,000	₹ 6,999	₹ 24,400
Tue 16/08	--	36%	₹ 8,000	₹ 7,500	₹ 8,000	₹ 8,999	₹ 13,400
Wed 17/08	--	40%	₹ 6,321	₹ 9,000	₹ 8,500	₹ 9,449	₹ 8,400
Thu 18/08	--	36%	₹ 6,320	₹ 8,000	₹ 8,000	₹ 8,999	Sold out
Fri 19/08	--	39%	₹ 6,320	₹ 8,000	₹ 7,500	₹ 8,499	₹ 15,400
Sat 20/08	--	**41%**	**₹ 6,320**	**₹ 8,000**	**₹ 7,500**	**₹ 8,999**	**₹ 7,400**
S 8	--	36%	₹ 6,320	₹ 8,000	₹ 7,500	₹ 7,999	₹ 7,400
Mon 22/08	--	42%	₹ 7,624	₹ 9,000	₹ 8,000	₹ 7,499	₹ 7,400
Tue 23/08	--	41%	₹ 7,624	₹ 14,500	₹ 8,000	₹ 9,499	₹ 5,400
Wed 24/08	--	45%	₹ 7,624	₹ 13,500	₹ 10,000	₹ 14,999	₹ 6,400
Thu 25/08	--	44%	₹ 7,374	₹ 13,500	₹ 9,000	₹ 10,799	₹ 6,400
Fri 26/08	--	43%	₹ 6,847	₹ 9,000	₹ 10,000	₹ 8,999	₹ 6,400
Sat 27/08	--	46%	₹ 6,847	₹ 9,000	₹ 8,500	₹ 8,999	₹ 6,400
Sun 28/08	--	38%	₹ 6,847	₹ 8,000	₹ 8,500	₹ 8,999	₹ 6,400
Mon 29/08	--	40%	₹ 6,321	₹ 9,000	₹ 8,000	₹ 9,899	₹ 6,400
Tue 30/08	--	39%	₹ 6,321	₹ 9,000	₹ 8,000	₹ 9,899	₹ 6,400
Wed 31/08	--	37%	₹ 6,321	₹ 9,000	₹ 8,000	₹ 8,099	₹ 6,400

Liveshops…

Live

Always active

Scale

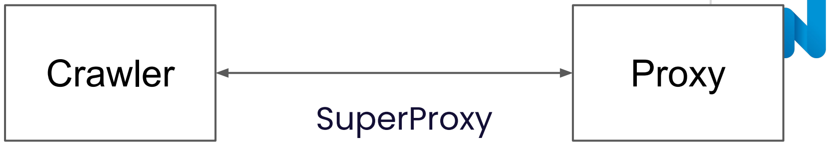
Automatic Resource-based Scaling

Inputs

Stream inputs over Pub/Sub to crawlers on demand.

CSV → Protobuf

2016



SuperProxy

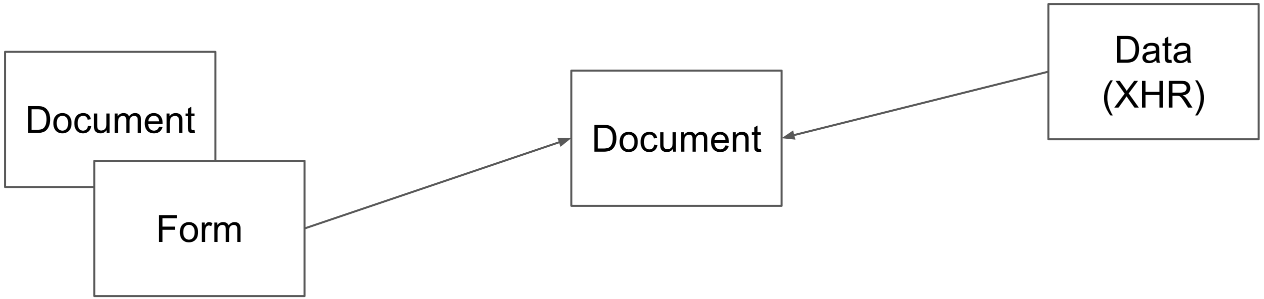
A MITM HTTP(S) Proxy written in Golang.

- Designed to handle high throughputs of requests/second;
- Ability to monitor all incoming and outgoing traffic;
- Delegate to proxy providers;

Only later we also started using it for emulation of HTTP and TLS Fingerprints.

Internet Browsers

I am a browser.

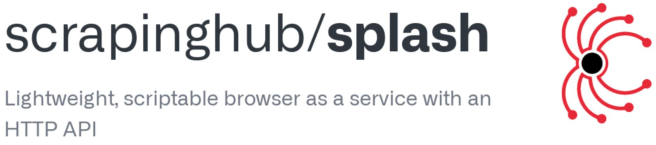


Document

Form

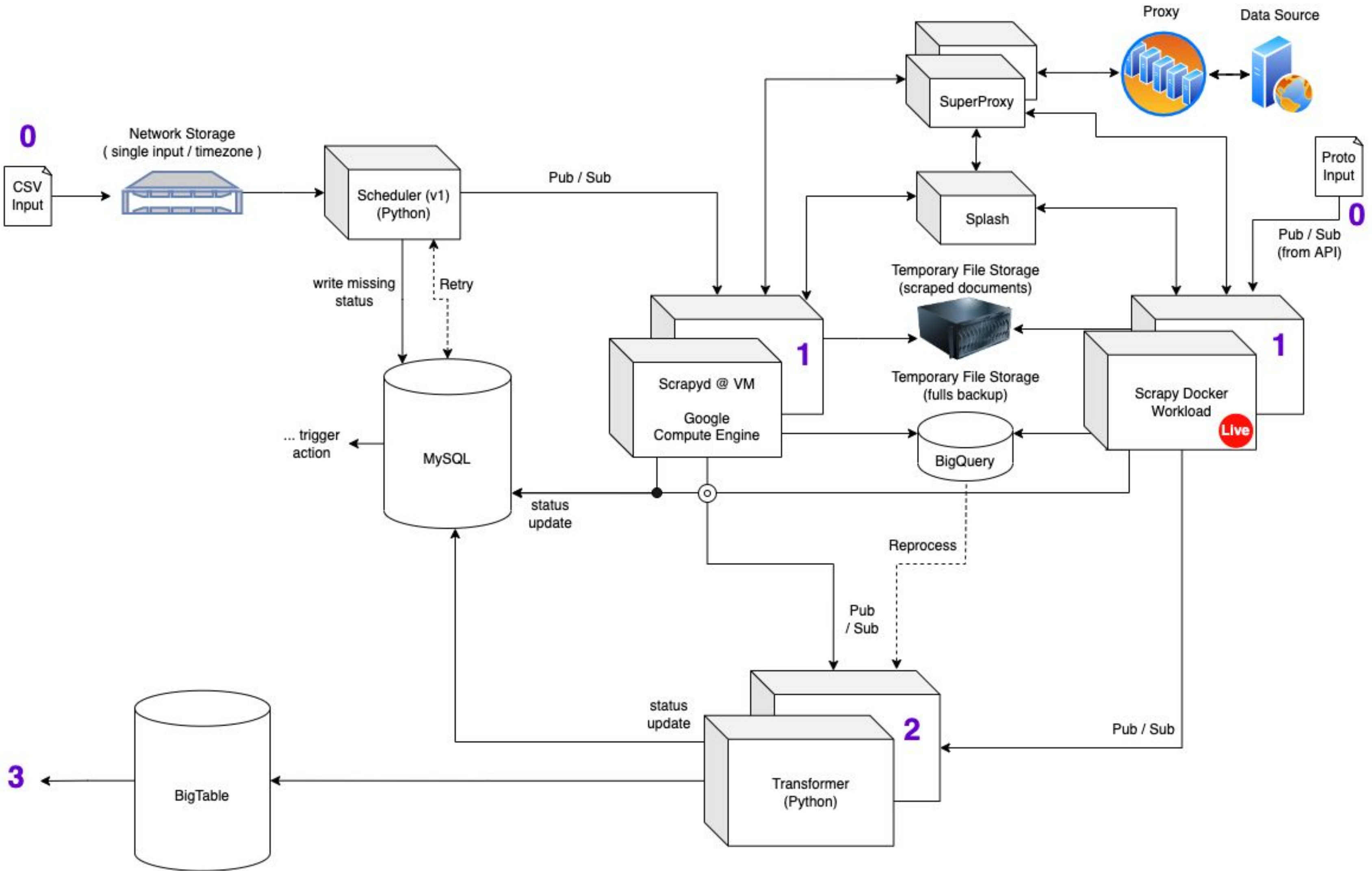
Document

Data (XHR)



scrapinghub/**splash**

Lightweight, scriptable browser as a service with an HTTP API


Proxy
Data Source
SuperProxy
Splash
0
CSV Input
Network Storage (single input / timezone)
Scheduler (v1) (Python)
Pub / Sub
Proto Input
0
Pub / Sub (from API)
write missing status
Retry
Temporary File Storage (scraped documents)
1
Scrapyd @ VM
Google Compute Engine
Temporary File Storage (fulls backup)
1
Scrapy Docker Workload
Live
... trigger action
MySQL
BigQuery
status update
Reprocess
Pub / Sub
status update
2
Pub / Sub
3
BigTable
Transformer (Python)

2013 - 2016

20,000,000

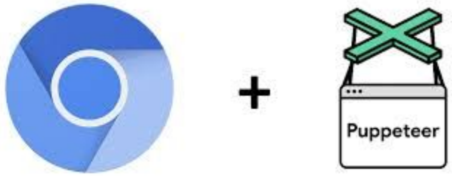
Requests / Day

2017

Internet Browsers

Replace *Splash*

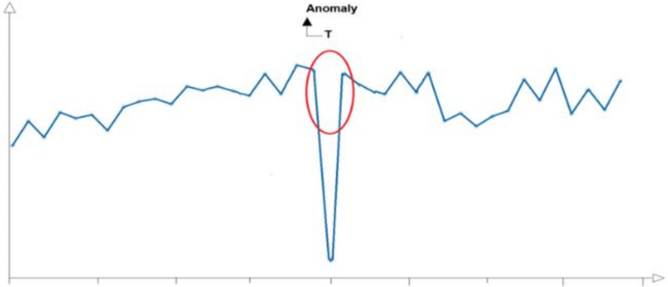
- Our demands outgrew its capabilities
- Javascript > Lua
- Rendering engine easy to detect



Anomaly Detection

Based on historical data

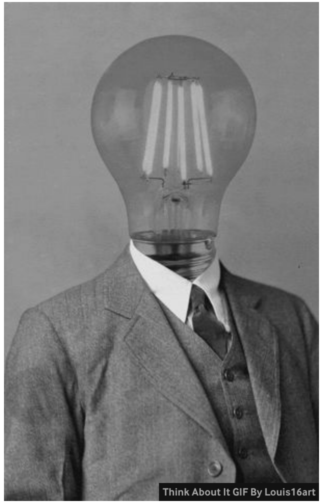
- Pattern-based anomaly detection
- Allows us to keep billions of data points Accurate at any given point
  - One of many solutions to our “scale” challenges



1 Our client and data needs were growing exponentially

2 Infrastructure suffered from Legacy Creep

3 Business needs were harder to integrate into our infrastructure



Think About It GIF By Louis16art

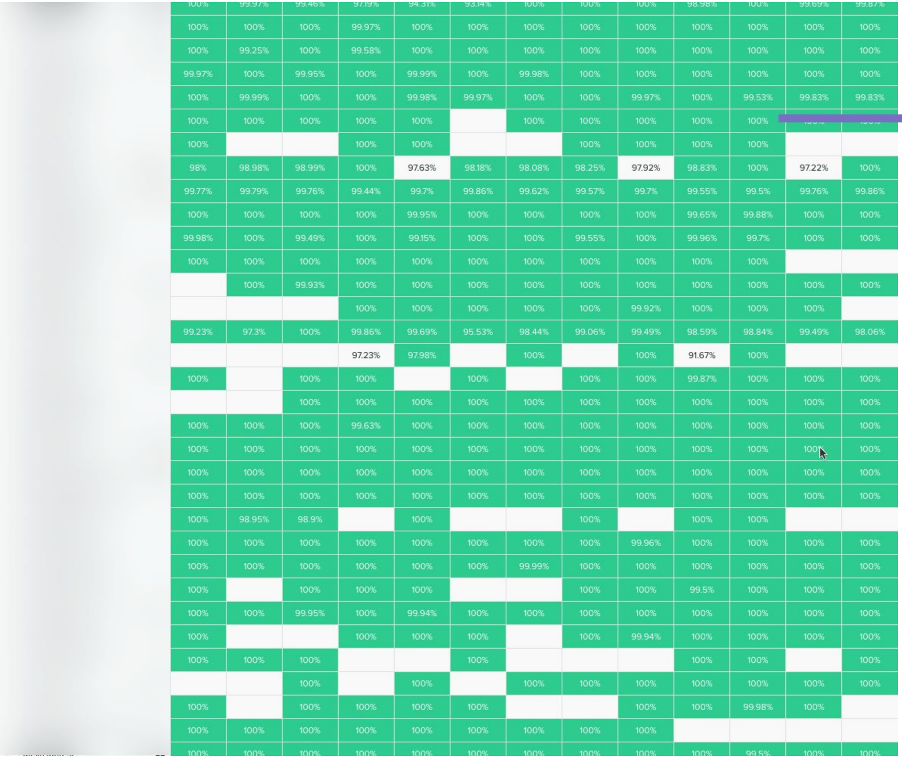
Integrities (v2)

```
19214#{184#0#19246900#19231#1#2####}#1660078800#0
```

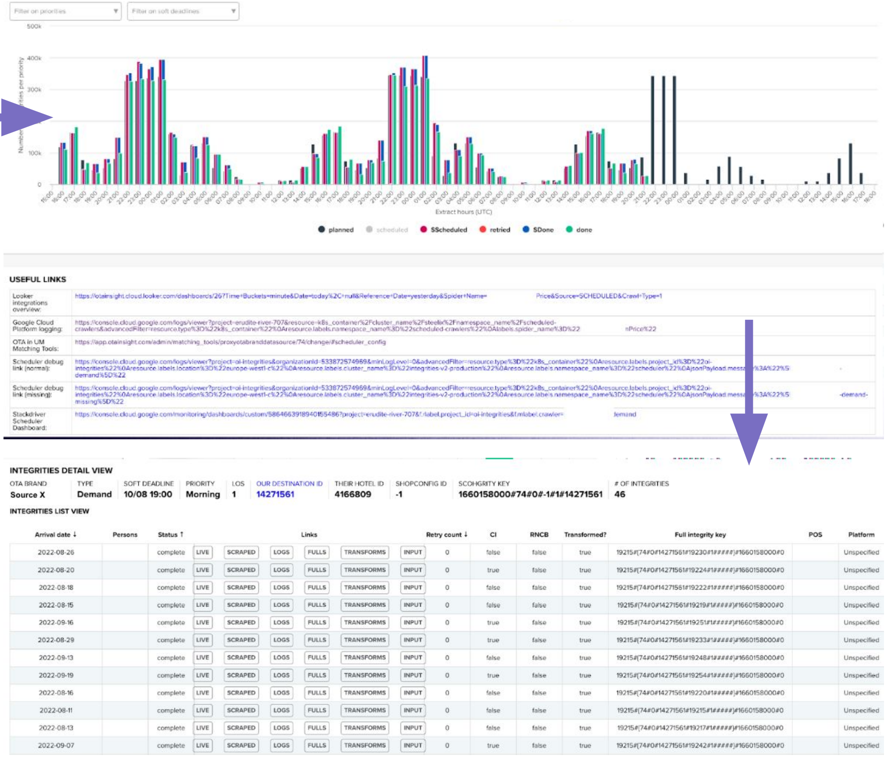
old: `184_1_2_308362_19214_1_rate_19231`

2018

Integrities (v2)



100%	99.97%	99.46%	97.19%	94.31%	93.14%	100%	100%	100%	98.98%	100%	99.69%	99.87%
100%	100%	100%	99.97%	100%	100%	100%	100%	100%	100%	100%	100%	100%
100%	99.25%	100%	99.58%	100%	100%	100%	100%	100%	100%	100%	100%	100%
99.97%	100%	99.95%	100%	99.99%	100%	99.98%	100%	100%	100%	100%	100%	100%
100%	99.99%	100%	100%	99.98%	99.97%	100%	100%	99.97%	100%	99.53%	99.83%	99.83%
100%	100%	100%	100%	100%		100%	100%	100%	100%	100%	100%	100%
100%			100%	100%			100%	100%	100%	100%		
98%	98.98%	98.99%	100%	97.63%	98.18%	98.08%	98.25%	97.92%	98.83%	100%	97.22%	100%
99.77%	99.79%	99.76%	99.44%	99.7%	99.86%	99.62%	99.57%	99.7%	99.55%	99.5%	99.76%	99.86%
100%	100%	100%	100%	99.95%	100%	100%	100%	100%	99.65%	99.88%	100%	100%
99.98%	100%	99.49%	100%	99.15%	100%	100%	99.55%	100%	99.96%	99.7%	100%	100%
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%		
	100%	99.93%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
			100%	100%	100%	100%	100%	99.92%	100%	100%	100%	
99.23%	97.3%	100%	99.86%	99.69%	95.53%	98.44%	99.06%	99.49%	98.59%	98.84%	99.49%	98.06%
			97.23%	97.98%		100%		100%	91.67%	100%		
100%		100%	100%		100%		100%	100%	99.87%	100%	100%	100%
		100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
100%	100%	100%	99.63%	100%	100%	100%	100%	100%	100%	100%	100%	100%
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
100%	98.95%	98.9%		100%			100%		100%	100%		
100%	100%	100%	100%	100%	100%	100%	100%	99.96%	100%	100%	100%	100%
100%	100%	100%	100%	100%	100%	99.99%	100%	100%	100%	100%	100%	100%
100%		100%	100%	100%			100%	100%	99.5%	100%	100%	100%
100%	100%	99.95%	100%	99.94%	100%	100%	100%	100%	100%	100%	100%	100%
100%			100%	100%	100%		100%	99.94%	100%	100%	100%	100%
100%	100%	100%			100%				100%	100%		100%
		100%		100%		100%	100%	100%	100%	100%	100%	100%
100%		100%	100%	100%	100%			100%	100%	99.98%	100%	
100%	100%	100%	100%	100%	100%	100%	100%	100%				
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	99.5%	100%	100%



USEFUL LINKS

Looker integrations overview:	https://otainsight.cloud.looker.com/dashboards/267?Time+Buckets=minute&Date=today%2C+null&Reference+Date=yesterday&Spider+Name= Price&Source=SCHEDULED&Crawl+Type=1
Google Cloud Platform logging:	https://console.cloud.google.com/logs/viewer?project=erudite-river-707&resource=k8s_container%2Fcluster_name%2Fsteelix%2Fnamespace_name%2Fscheduled-crawlers&advancedFilter=resource.type%3D%22k8s_container%22%0Aresource.labels.namespace_name%3D%22scheduled-crawlers%22%0Alabels.spider_name%3D%22 nPrice%22
OTA in UM Matching Tools:	https://app.otainsight.com/admin/matching_tools/proxyotabranddatasource/74/change/#scheduler_config
Scheduler debug link (normal):	https://console.cloud.google.com/logs/viewer?project=oi-integrities&organizationId=533872574969&minLogLevel=0&advancedFilter=resource.type%3D%22k8s_container%22%0Aresource.labels.project_id%3D%22oi-integrities%22%0Aresource.labels.location%3D%22europe-west1-c%22%0Aresource.labels.cluster_name%3D%22integrities-v2-production%22%0Aresource.labels.namespace_name%3D%22scheduler%22%0AjsonPayload.messa %3A%22%5 demand%5D%22
Scheduler debug link (missing):	https://console.cloud.google.com/logs/viewer?project=oi-integrities&organizationId=533872574969&minLogLevel=0&advancedFilter=resource.type%3D%22k8s_container%22%0Aresource.labels.project_id%3D%22oi-integrities%22%0Aresource.labels.location%3D%22europe-west1-c%22%0Aresource.labels.cluster_name%3D%22integrities-v2-production%22%0Aresource.labels.namespace_name%3D%22scheduler%22%0AjsonPayload.messa %3A%22%5 -demand-missing%5D%22
Stackdriver Scheduler Dashboard:	https://console.cloud.google.com/monitoring/dashboards/custom/5864663918940155486?project=erudite-river-707&f.rlabel.project_id=oi-integrities&f.mlabel.crawler= lemand

INTEGRITIES DETAIL VIEW

OTA BRAND	TYPE	SOFT DEADLINE	PRIORITY	LOS	OUR DESTINATION ID	THEIR HOTEL ID	SHOPCONFIG ID	SCOHGRITY KEY	# OF INTEGRITIES
Source X	Demand	10/08 19:00	Morning	1	14271561	4166809	-1	1660158000#74#0#-1#1#14271561	46

INTEGRITIES LIST VIEW

Arrival date ↓	Persons	Status ↑	Links	Retry count ↓	CI	RNCB	Transformed?	Full integrity key	POS	Platform
2022-08-26		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19230#1#####)#1660158000#0		Unspecified
2022-08-20		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	true	false	true	19215#(74#0#14271561#19224#1#####)#1660158000#0		Unspecified
2022-08-18		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19222#1#####)#1660158000#0		Unspecified
2022-08-15		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19219#1#####)#1660158000#0		Unspecified
2022-09-16		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	true	false	true	19215#(74#0#14271561#19251#1#####)#1660158000#0		Unspecified
2022-08-29		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	true	false	true	19215#(74#0#14271561#19233#1#####)#1660158000#0		Unspecified
2022-09-13		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19248#1#####)#1660158000#0		Unspecified
2022-09-19		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	true	false	true	19215#(74#0#14271561#19254#1#####)#1660158000#0		Unspecified
2022-08-16		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19220#1#####)#1660158000#0		Unspecified
2022-08-11		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19215#1#####)#1660158000#0		Unspecified
2022-08-13		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	false	false	true	19215#(74#0#14271561#19217#1#####)#1660158000#0		Unspecified
2022-09-07		complete	LIVE SCRAPED LOGS FULLS TRANSFORMS INPUT	0	true	false	true	19215#(74#0#14271561#19242#1#####)#1660158000#0		Unspecified

Integrities (v2)

Satisfy our needs for atomic writes at scale

FoundationDB

The right tool for the job.

Integrities (v2)

Apply lessons learned from liveshop crawlers to scheduled crawlers

k8s

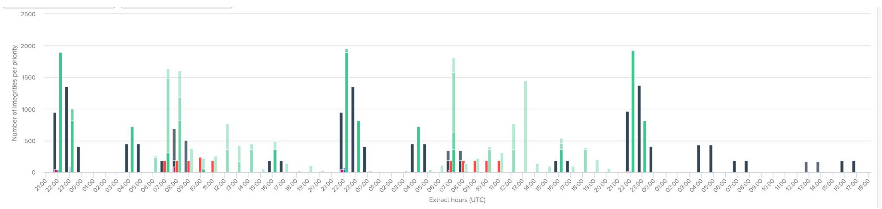
Orchestrated by scheduler (v2)

Protobuf

Drop the CSV Inputs in favour of Protobuf-driven Crawl Inputs

Integrities (v2)

Automate workload definitions



Planner

Forecast and simulate

Scheduler

Orchestrate workloads
and deliver crawl inputs

Integrities (v2)

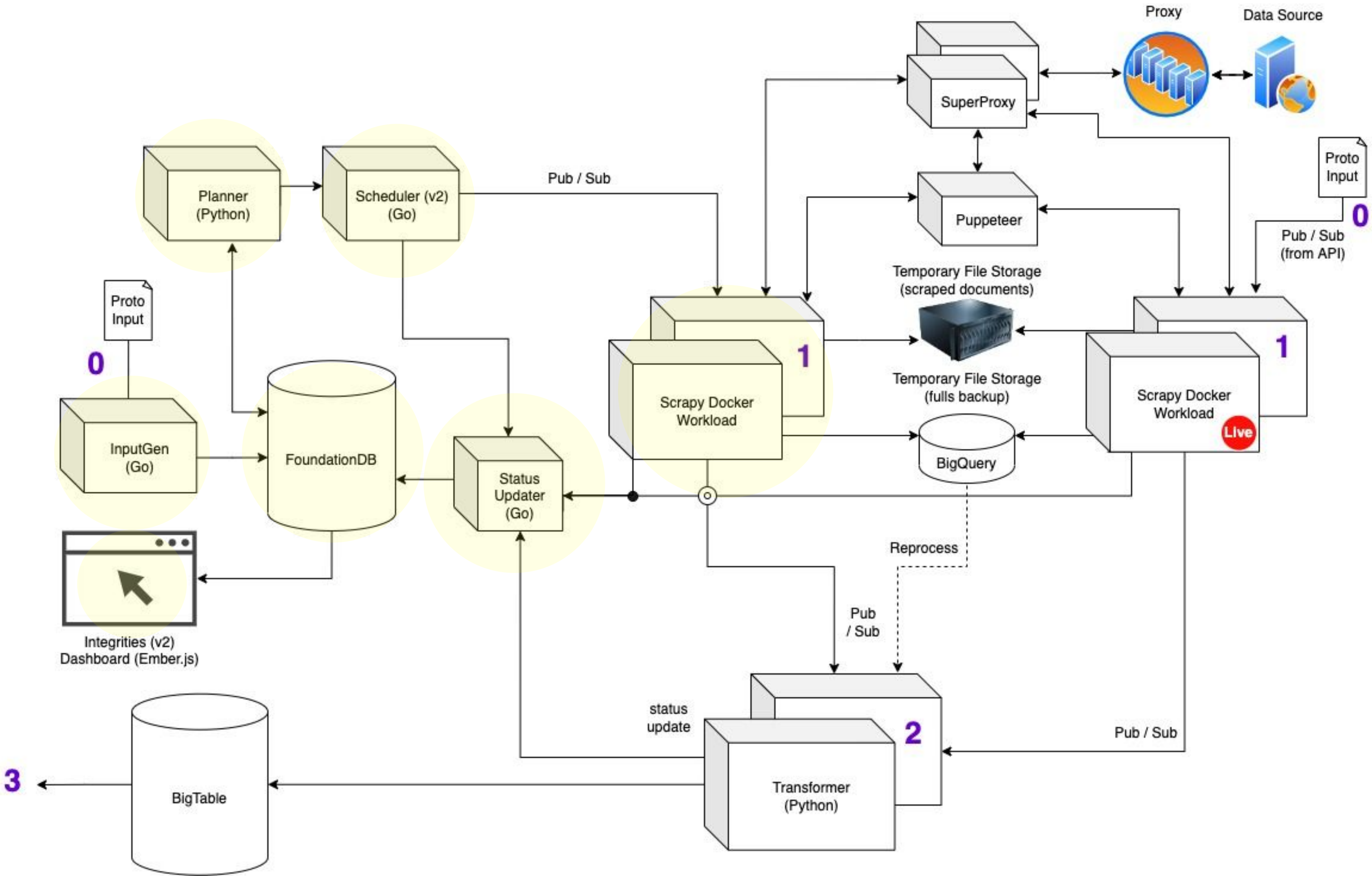
Migration to our new infrastructure

Test

Tests and Traffic Emulators

Parallel

Run both versions of Integrities in parallel for an extended period to compare workloads.

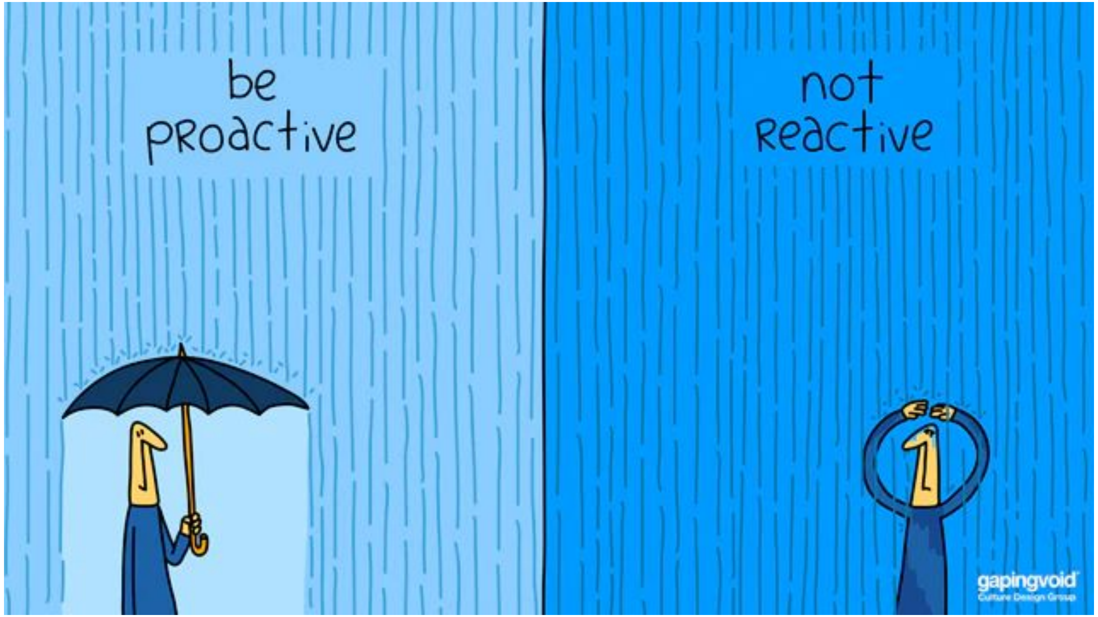

Proxy
Data Source
SuperProxy
Planner (Python)
Scheduler (v2) (Go)
Pub / Sub
Proto Input
0
Pub / Sub (from API)
Puppeteer
Temporary File Storage (scraped documents)
Proto Input
0
1
Scrapy Docker Workload
Temporary File Storage (fulls backup)
1
Scrapy Docker Workload
Live
InputGen (Go)
FoundationDB
Status Updater (Go)
BigQuery
Reprocess
Integrities (v2) Dashboard (Ember.js)
Pub / Sub
status update
2
Transformer (Python)
Pub / Sub
3
BigTable

Learn. Adapt. Evolve.
meme-arsenal.ru

Automate

Priority #1

- Error Monitoring: Sentry
- Alerting: slack + email
- Incident Response: PagerDuty → OpsGenie
- Tooling: Metric exports & Dashboards

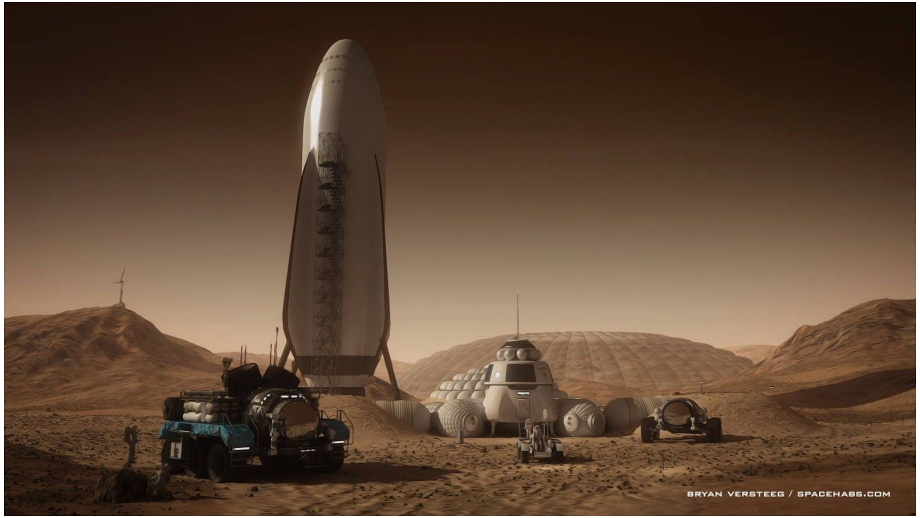

be proactive
not reactive
gapingvoid
Culture Design Group

Be Pragmatic

Cost-Reward Balance

- Proxy?
- Browser?
- Automate?

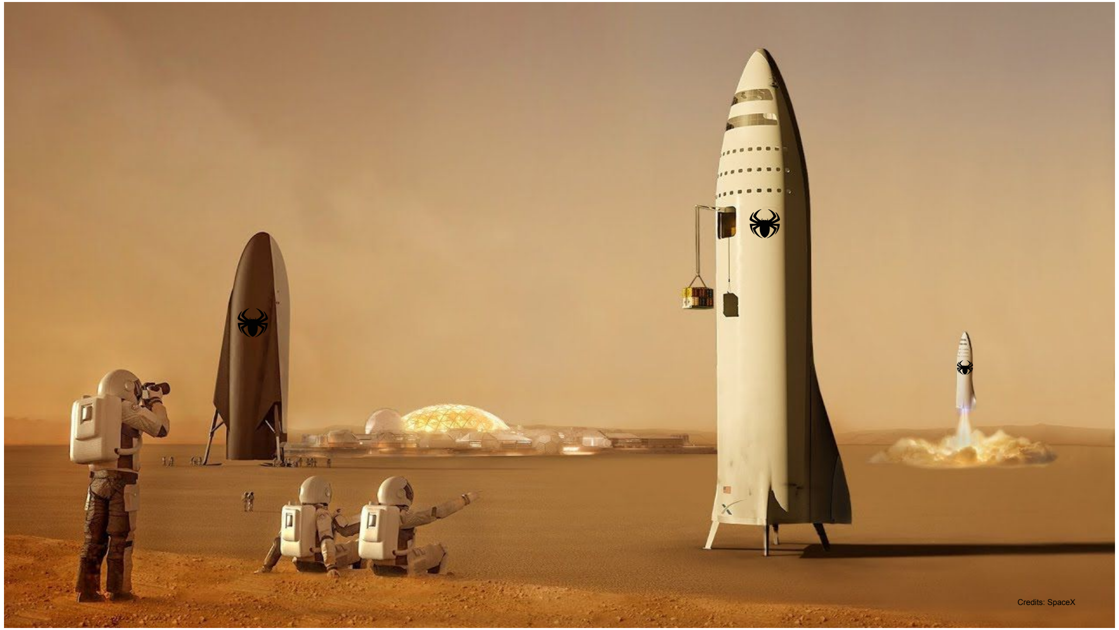
Present to Future Roadmap



- Hire & Educate
- Improve UX
- R&D: Anti-Fingerprint

careers.otainsight.com

Give a crawler engineer scrapy and you can extract data for a day, teach him how to scale and you will go to space.



Credits: SpaceX

Thank You

Web Data Extraction
Summit 2022

Powered by Zyte

extract.summit@glendc.com