# GetYourGuide

Document version

| Version | Updates | Author |
|---|---|---|
| v0.1 | | |
| | | |

## Table of content

---

## Section 1 - to be shared and signed off by customer

Shared Document version

| Version | Updates | Author |
|---|---|---|
| v0.1 | | |
| | | |

## Table of content

## 1.0 Project Overview

The Import.io Technical Services team has been engaged by Customer for the purpose of conducting a production deployment of the Import.io solution. The Import.io team will collaborate with Customer to design and configure the Import.io platform pursuant to the scope and requirements set forth in this document. Data will be retained for 30 days.

**GetYourGuide introduction**

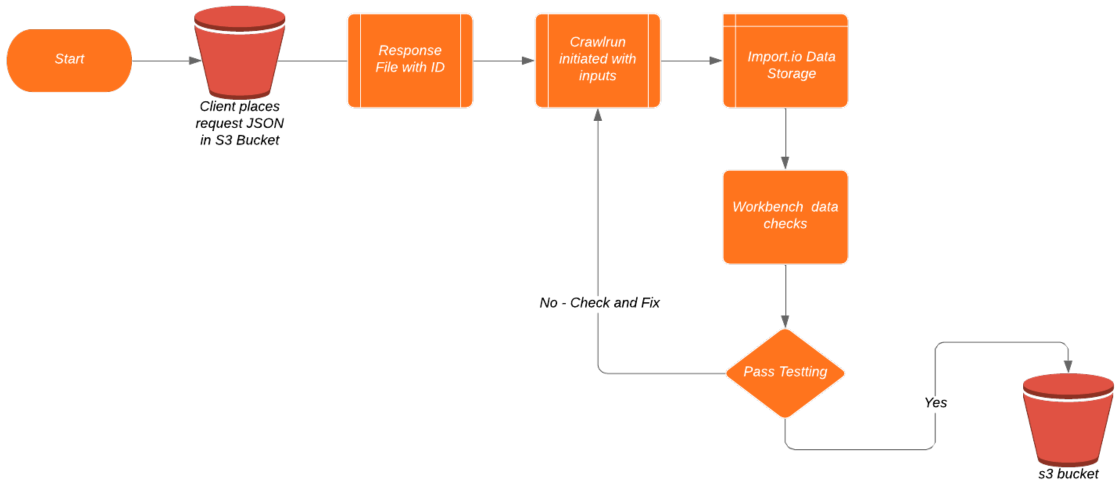
GetYourGuide is a privately held global company headquartered in Berlin, Germany that operates an online marketplace and internet booking engine which is accessible via its website and mobile app. GetYourGuide's global inventory includes tours and excursions, activities like cooking classes, and tickets to numerous tourist attractions. It currently offers more than 40,000 products in destinations around the world. GetYourGuide acts as an online platform for third-party companies to list their products for users to easily find. Businesses offering sightseeing tours, adventure activities, multiple day tours, attractions passes, and other products can upload and manage their products under their own brand. Customers can book these products directly on the website as well as through iOS and Android Apps or through its distribution network.

**Project Requirements**
GetYourGuide would like to track their price competitiveness by examining the prices for the same product as on their website, on the websites of their competitors, starting with Viator, Tiqets and Klook. GetYourGuide already has processes that map their inventory to the URLs of their competitors.

## 2.0 Data Flow Diagram

Data Flow Diagram

andrew rowlands | February 17, 2020



## 3.0 Design of the Data Pipeline

### 3.1 Customer Schema

- Details of Input Schema
  - S3 Input method discussed later in document.
  - 3 Sources
  - One line for each currency requested in input and output.
  - Lookahead period starting with tomorrow (+1, +7,+14,+21) - each input line will results in up to 4 outputs - if all 4 lookahead dates are available.
  - selection should always be one adult

| Data Source | URL |
|---|---|
| Viator | https://www.viator.com/ |
| Tiqets | https://www.tiqets.com/en/ |
| Klook | https://www.klook.com/en-US |

| Name | Description | Type | Notes |
|---|---|---|---|
| URL | The URL of the activity | URL | |

| | | | |
|---|---|---|---|
| Option | The name of the variant | Text | Option must contain this field - can be extraneous data on end but all of this field should be in title unless if option is blank pick first option in list. |
| Currency | The currency | Text | 3-letter code for the currency. Will be used in addition to USD, EUR and GBP if provided. |
| TourOptionId | Pass Thru | Text | Id of Tour Option |

- Details of Output Schema
- Lookahead period starting with tomorrow (+1, +7,+14,+21) - each input line will results in up to 4 outputs - if all 4 lookahead dates are available.

| **Name** | **Description** | **Type** | **Notes** |
|---|---|---|---|
| Activity ID | The ID of this activity | Text | Derived from URL<br>Viator Example: d511-3731COLOSSEUM<br>Tiqets Example: p918256<br>Klook Example: 6396 |
| Competitor | Viator or Tiqets or Klook | Text | Example: Viator |
| Date of collection | Today’s collection date | Date (UTC TZ for consistency) | Example: 2020-02-22 |
| Date of travel | Selected date of travel | Date | Example: 2020-02-24 |
| From price (black) | The *from* price | Price | Example: 35.00 |
| From price (red) | The discounted *from* price | Price | Example: 23.00<br>If available |
| Is Available | Is the tour available on this date of travel | Boolean | Example: TRUE |
| Final price (black) | The *final* price, including all fees | Price | Example: 39.00<br>If available |
| Final price (red) | The discounted *final* price, including all fees | Price | Example: 32.00<br>If available |
| Currency | Three letter currency code | Text | Example: USD |
| TourOptionId | From Input | Text | ABC123 |
| Option | From Input | Text | Option must contain this field - can be extraneous data on end but all of this field should be in title unless if option is blank pick first option in list. |

### 3.2 Input Processing

- Customer places inputs in S3 bucket
- Format of input File.

  -
    ```
    `[{
    "_url": "https://example.com",
    "options": {
    "exact": true,
    "details": ["Show+Champagne",””,"11:00"]
    },
    "currency": "USD",
    "TourOptionId": "BC12"
    },
    "_url": "https://example.com",
    "options": {
    "exact": true,
    "details": ["Show+Dinner",””,"15:00"]
    },
    ```

```
        "currency": "USD",
        "TourOptionId": "BC13"
        }]
```

exact tells us the type of match expected - exact true is a full match - exact false is a contains match as discussed. In both false and true we want the match to NOT be case sensitive.

Options  field now is an array in order to “drill down” into the sites as discussed.

Options will be taken in order they are found in file so for example if first selection is “Show + Champagne” that will be selected - if next option is “First Row Seating” then that will be selected next assuming option is available.

Below is link of example options

Example Links

New Examples

**Input S3 location** (Owned by http://Import.IO) - arn:aws:s3:::importio-getyourguide

**Input file name format** - {unique-id} can be date or any id client uses to identify input files.

Viator-{unique-id}.json

Klook-{unique-id}.json

Tiqets-{unique-id}.json

**Output S3 location** (Owned by GetYourGuide) -bucket name is import-io.gyg.io

**Output file name format** - unique id, source and date will be part of filename to tie back to request.

Viator-{unique-id}-{date}.json

Klook-{unique-id}-{date}.json

Tiqets-{unique-id}-{date}.json

Failed runs will be checked by Import.io and any issues will be remedied and files re-ran - if the issue is in the inputs - that will be reported to client either manually or an error file will be created - future api’s will be developed to report status of runs.

CRAWLRUN refers to import.io job reference for an extraction job - each occurrence has a unique internal guid known as crawlrun-id. Import.IO will monitor the input S3 destination once daily (time to be verified). If file(s) are identified a lambda script on Import.IO side will,

- Notify client of crawlrun id- return file to output S3 location
- Process and post api with json formatted inputs (extractor-api)
- Start crawlrun with those inputs
- Archive input file and store for 30 days - input files will be moved to an archive directory in the input S3 bucket.
- Results will be delivered to output S3 location, as detailed below - unique id - source and date  will be part of filename to tie back to request.
- Passing the customer input into Import
  - API will start extractors with inputs gathered from input S3.
  - S3 bucket for requests will be an import.io specified bucket - credentials will be shared with Client.
  - Required input Validation
    - URL is valid pointer to source developed
    - Variant name is not blank
    - Currency is not blank

### 3.3 Extractors

- type of extractors - TBD
- use of workflow
  - Results will be passed through Workbench QA and workflow
- use of chained extractor
  - None - extractor should be one detail page extraction of URL provided
- error handling
- examples

### 3.4 Data Quality Criteria

- Data Shape Rules

| **Name** | **Description** | **Type** | **Data Rules - EXAMPLES** | **Checks** | **Where** |
|---|---|---|---|---|---|
| Activity ID | The ID of this activity | Text | https://www.tiqets.com/en/checkout/top-of-the-rock-p974124<br>In this example - P974124 is the ID | P974124<br>Check for non blank(WB) | WORKBENCH |
| Competitor | Viator or Tiqets | Text | Viator | Viator, Tiqets or Klook | From input no check needed |
| Date of collection | Today's collection date | Date | Example: 2020-02-22 | Valid Date | WORKBENCH |
| Date of travel | Selected date of travel | Date | Example: 2020-02-24 | Valid Date | WORKBENCH |
| From price (black) | The *from* price | Price | Example: 35.00 | Valid value | Not always available<br>if product is unavailable or deactivated |
| From price (red) | The discounted *from* price | Price | Example: 23.00<br>If available | Valid value | Not always available |
| Is Available | Is the tour available on this date of travel | Boolean | Example: TRUE | TRUE/FALSE | WORKBENCH |
| Final price (black) | The *final* price, including all fees | Price | Example: 39.00<br>If available | NA | Not always available<br>if IS AVAILABLE IS FALSE |
| Final price (red) | The discounted *final* price, including all fees | Price | Example: 32.00<br>If available | Valid value | Not always available |
| Currency | Three letter currency code | Text | Example: USD | Example: USD | From input no check needed |
| TourOptionId | From input | Text | ABC123 | NA | From input no check needed |
| Option | From input | Text | VARIANTNAMEA | NA | From input no check needed- can be blank |

-

### 3.5 Post Processing

- None currently

### 3.6 Data Harmonization Across Sources

- None currently

### 3.7 Screen Shots & storage

### 3.8 Details of Data Delivery Format & Destination

- Delivery of JSON to S3 bucket (Client specified and Import.IO).

## 4.0 Schedule

### 4.1 Crawl run schedule

- Schedule Runs : Once Daily if pickup file(s) are present.

**4.2 Delivery Schedule**

- As completed - automatically via Workbench after QA is passed

Questions

## Section 2 - internal use only

### 6.0 Project Meta Data

**6.1 Document Owner: Andrew Rowlands**

**6.2 Contributor(s):**

**6.3 Project Plan (link) GetYourGuide Project Plan**

**6.4 SOW (link)**

**6.5 Monday Board (link) Monday**

### 7.0 Input/Output Schema ETL from Standard Schema

**7.1 use of JQ transform (in workbench)**

**7.2 scripts required outside workbench**

### 8.0 Open Issue / Concerns?

| Issue/Concern | Responsible Party | Phase/Date impact | Workaround |
|---|---|---|---|
| Time as input for variant | SA/Client | | Have client include in variant name |
| | | | |

### 9.0  Consideration for tech ops (edited)

**9.1 anything that is not automated**

**9.2 Non-standard procedure**