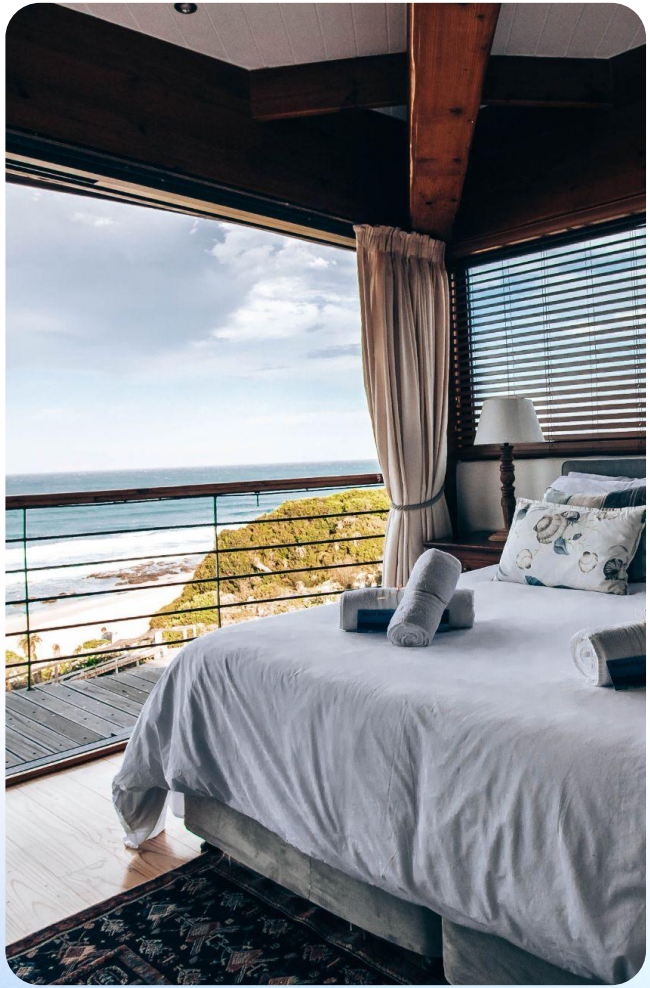
Supply data processing evolution

@ HomeToGo

Darius Kasiulevičius

Vilnius, 2022-03-03

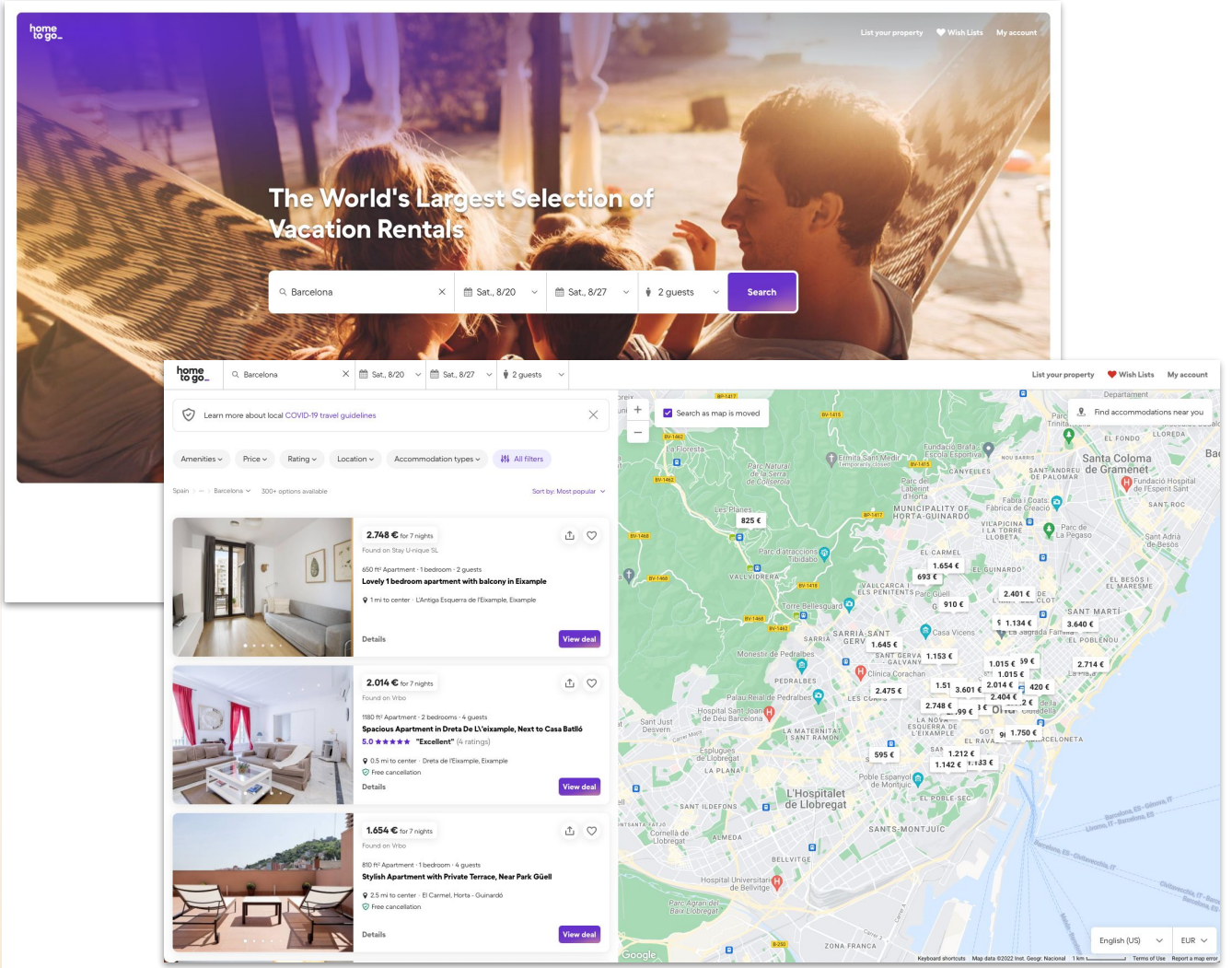


About me

- Darius Kasiulevičius
- +10 years PHP developer
- ~21 years since wrote first code line
- Team lead & Software architect at HomeToGo

home to go

What is HomeToGo?



home to go

Our Challenges

Supply Data is critical for business

- Missing inventory - missed sales opportunity
    - Including search filters, e.g. destination, dates, pool, wifi, pets, etc.
- Data accuracy (price, availabilities) issues leads to poor conversion rates
- Missing data points leads to poor conversion rates. E.g. cancellation policies, room plan, etc.

home to go

First integration

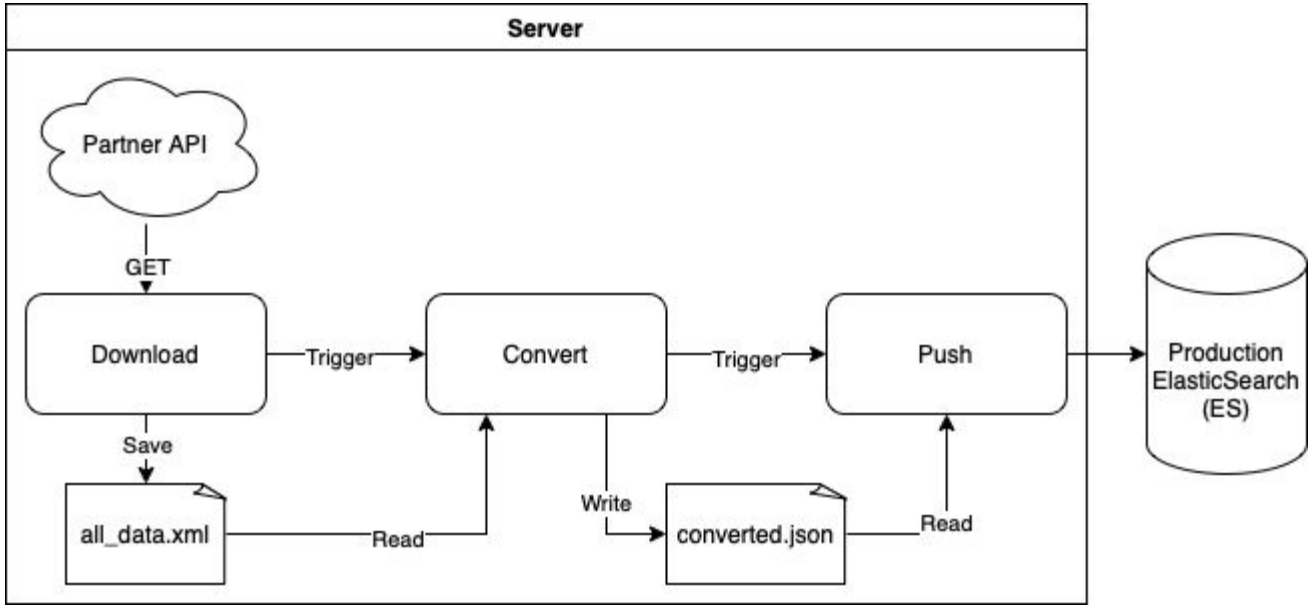
First partner

First partner

- Small one - less than 1000 offers
- One file with full data
- Data file type XML
- Simple download over HTTP

home to go

First pipeline



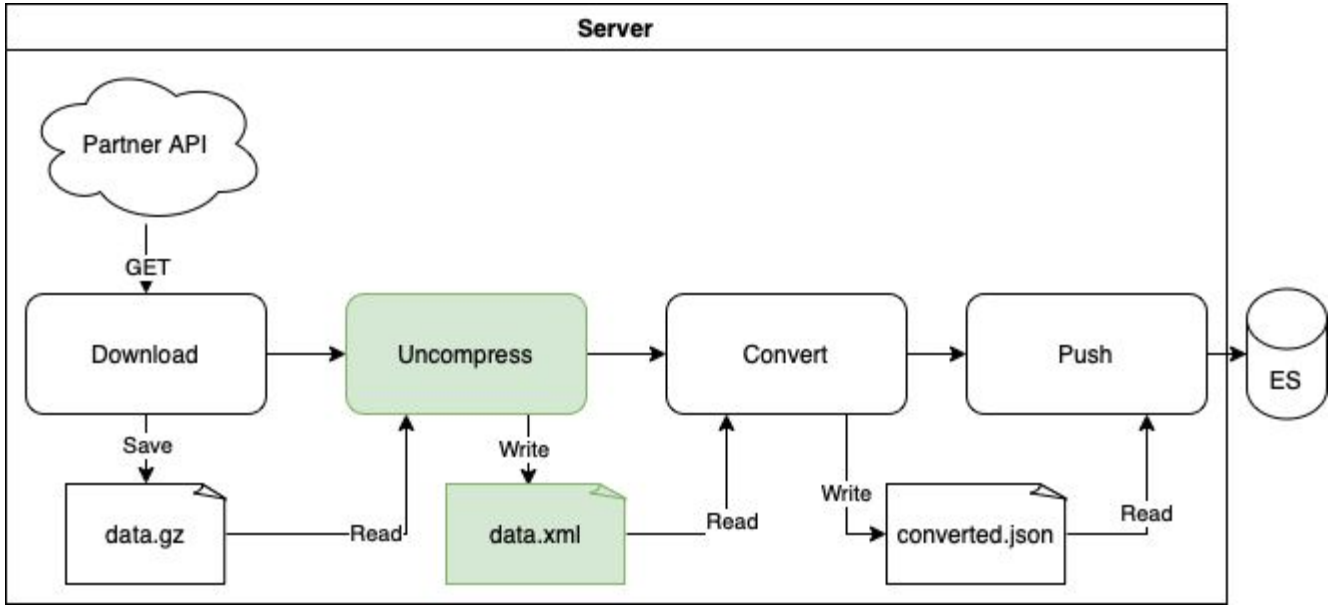
home to go

Next partners

- Compressed file
- Download over FTP, SFTP, SOAP
- File types CSV, JSON
- To get all data need make X GET requests

home to go

Next pipeline



First improvement

Custom pipelines per partner

Download more than one file (units.xml.gz, locations.xml, descriptions.xml.gz)
Uncompress
Validation

Task Manager

- Custom pipeline per partner
- Monitor pipeline status
- Resume pipeline from failed task
- More than one pipeline per partner (locations)

home to go

Task Manager configuration

```
vrt_providers_adria_gate.task_manager.download_manager.config:
    -
        task: 'download_manager'
        depends: { task: ['prepare_db'] }

vrt_providers_adria_gate.task_manager.healer.config:
    -
        task: 'heal'
        params: { source: 'original_source/raw_*.xml' }
        depends: { task: ['download_manager'] }

vrt_providers_adria_gate.task_manager.split_prepare_data.config:
    -
        task: 'prepare_data'
        params:
            source: 'adria_gate'
            aggregationConfig:  '%vrt_providers_adria_gate.aggregation_config%'
        depends: { container: ['vrt_providers_adria_gate.task_manager.healer'] }

vrt_providers_adria_gate.task_manager.after_prepare_data.config:
    -
        task: 'after_prepare'
        depends: {container: ['vrt_providers_adria_gate.task_manager.split_prepare_data']}

vrt_providers_adria_gate.task_manager.template.config:
    -
        template: 'common_from_splitted_convert.yml'
        params:
            inheritableTaskId: 'after_prepare'
            masterTable: SessionUnit
            suffix: 'adria_gate'
```

home to go

Download Manager configuration

```
vrt_providers_adria_gate.download_manager_config:
    _params:
        concurrency: 20
        curl:
            CURLOPT_HTTPAUTH: 1 # CURLAUTH_BASIC
            CURLOPT_USERPWD: XXXXXXXXXXXX
    areas:
        source: 'https://ws.adriagate.com/agws/v03/xml/GetAreas?language=en'
        target: 'areas_en.xml'
    index:
        source: 'https://ws.adriagate.com/agws/v03/xml/GetListingIndex'
        target: 'index.xml'
    objects:
        parent: 'index'
        source: '{$url}'
        target: ''
        generator:
            -
                class: 'Htg\Backend\DownloadManagerBundle\Request\Generators\Ids'
                idPattern: '(?<=(<url>))(.+?)(?=(</url>))'
                pattern: '{$url}'
                ignoreEmpty: true
            -
                class: 'Vrt\Providers\AdriaGateBundle\DownloadManager\DecodeSourceUrl'
            -
                class: 'Vrt\Providers\AdriaGateBundle\DownloadManager\ModifyTarget'
                search: '/(.+id=)(\d+-\d+)(.*)/'
                replace: 'object_$2.xml'
    availabilities:
        parent: 'objects'
        source: 'https://ws.adriagate.com/agws/v03/AdriagateService.svc/xml/GetUnitAvailability?i
        target: 'availabilities_{$id}.xml'
```

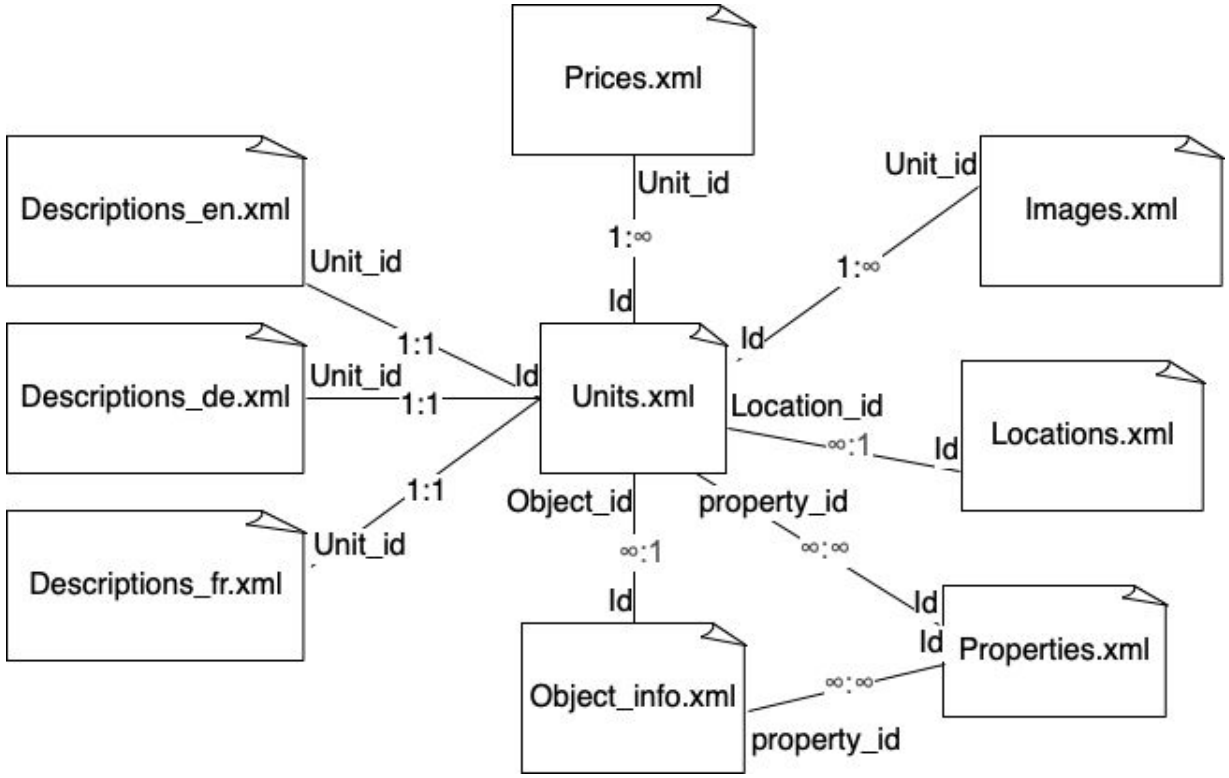
home to go

Second improvement

More than one file to get full data

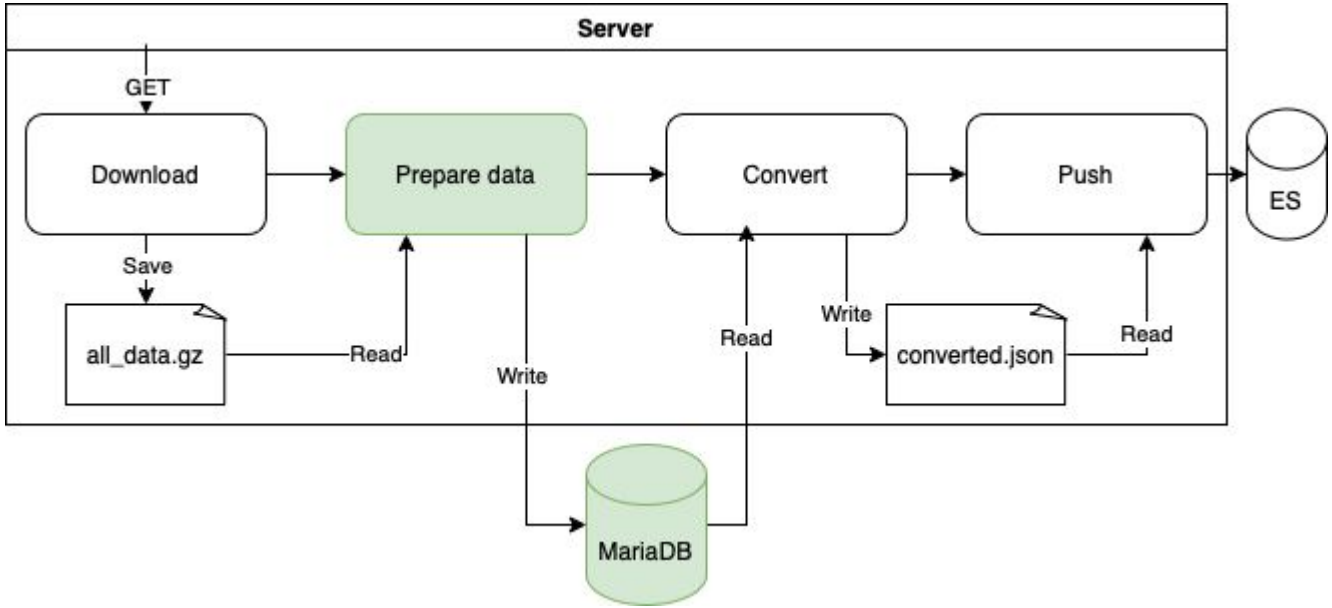
Images.xml
Descriptions_de.xml
Descriptions_en.xml
Descriptions_fr.xml
Prices.xml
Units.xml

Data relationship



Permanent DB

- Separate DB per partner
- Separate table per file type (prices, images, description, etc.)
- Uniq table structure per partner



home to go

Prepare data configuration

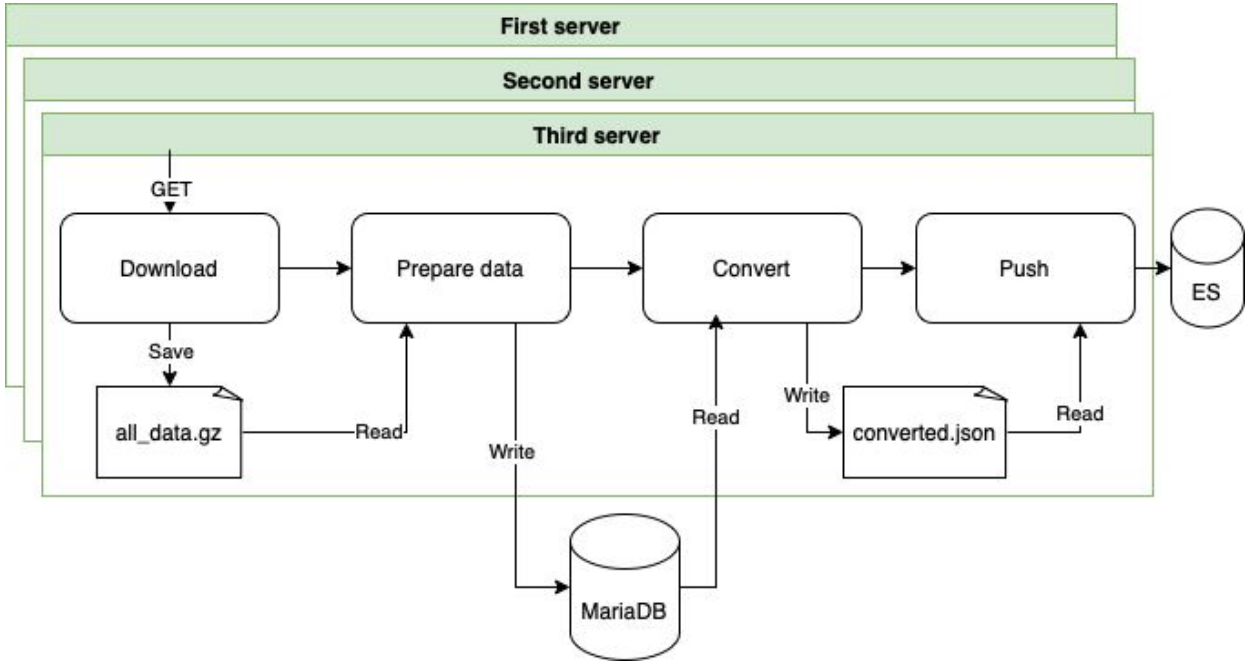
```
vrt_providers_adria_island.aggregation_config:
    objects:
        tag: Property
        xsl: unit.xsl
        entity: Units
        master: true
        types: [default]
        file: 'inventory_*_de.xml'
        type: csv2
        params: <1 key>
        dir: original_source
        source: XmlParserFileSource
    prices:
        tag: Prices
        xsl: price.xsl
        entity: Prices
        types: [default]
        file: 'prices_*.xml'
        type: csv2
        params:
            duplicate: REPLACE
            file_name_parts: 'prices_(\d+)_.*\.xml'
            file_name_parts_mapping: 'year'
        dir: original_source
        source: XmlParserFileSource
    vacancies:
        tag: Availability
        xsl: availability.xsl
        entity: Availabilities
        types: [default]
        file: 'vacancies_*.xml'
        type: csv2
        params:
```

Third improvement

Update data for all partners at same time

Partner count increased

Increasing server count to 3



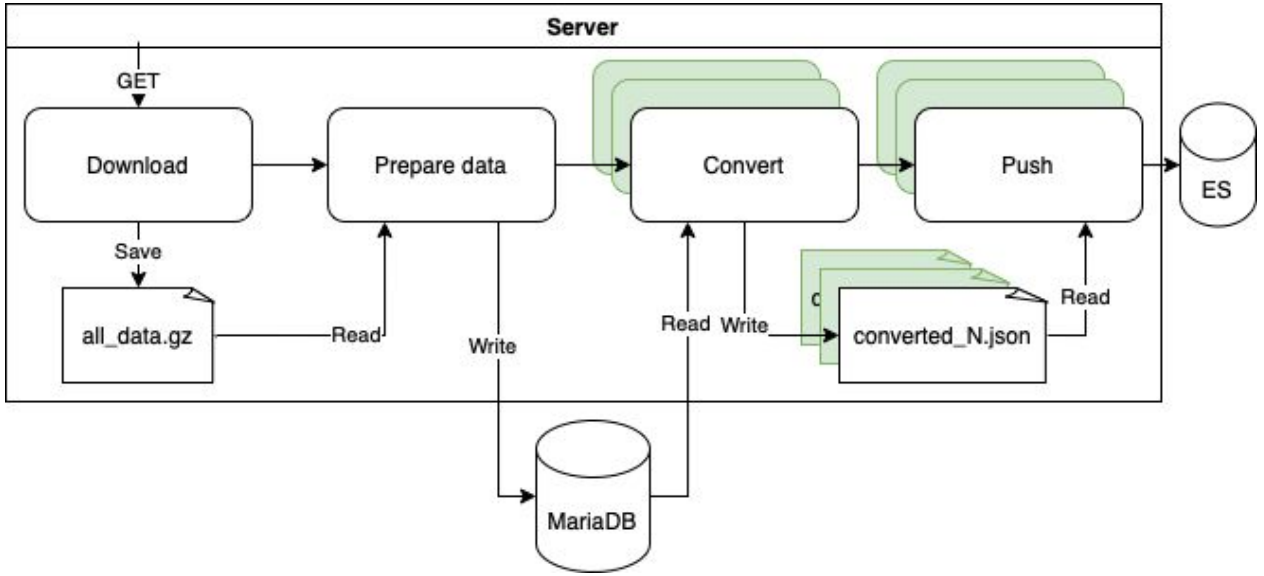
home to go

Fourth improvement

Process run time optimization

Partner with 1 million offers

Split convert



SELECT select_list FROM table_name LIMIT offset, row_count;

First convert run time << second convert run time <<< third convert run time

SELECT select_list FROM table_name WHERE autoincrement_id BETWEEN 5000 AND 10000;

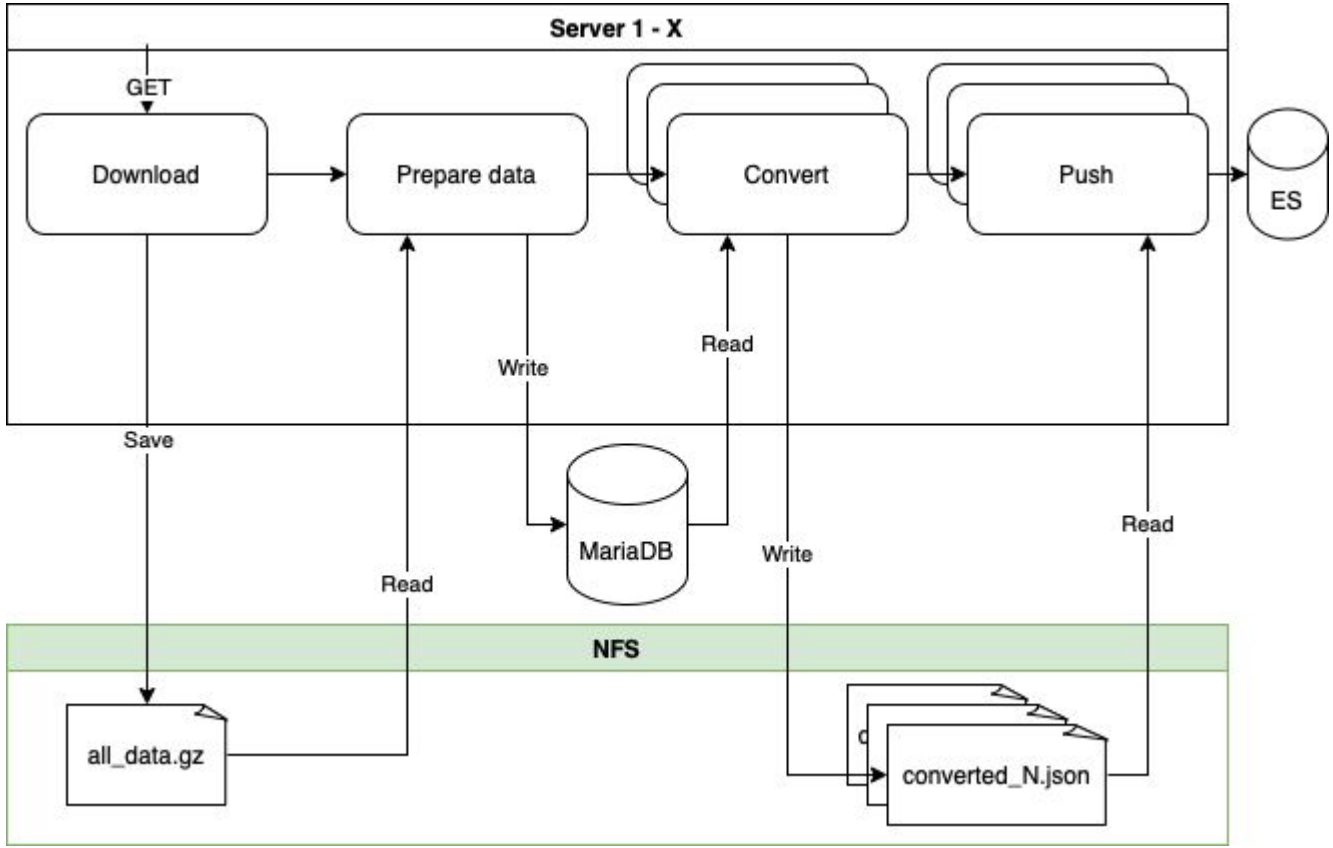
home to go_

Fifth improvement

Equal use of server resources

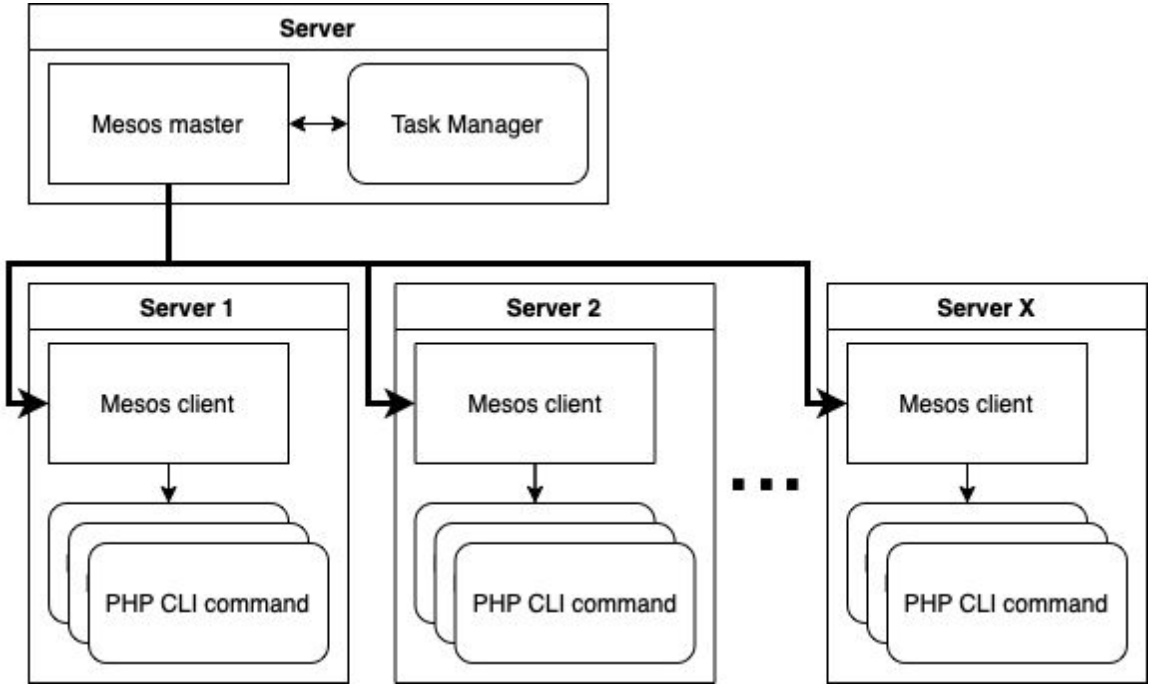
Big partners competes with the small ones

Add NFS



home to go

Start using



- Task Manager changes
  - Communicate with Mesos API
  - Support X servers
  - Partner commands run in any server
  - Support auto scaling

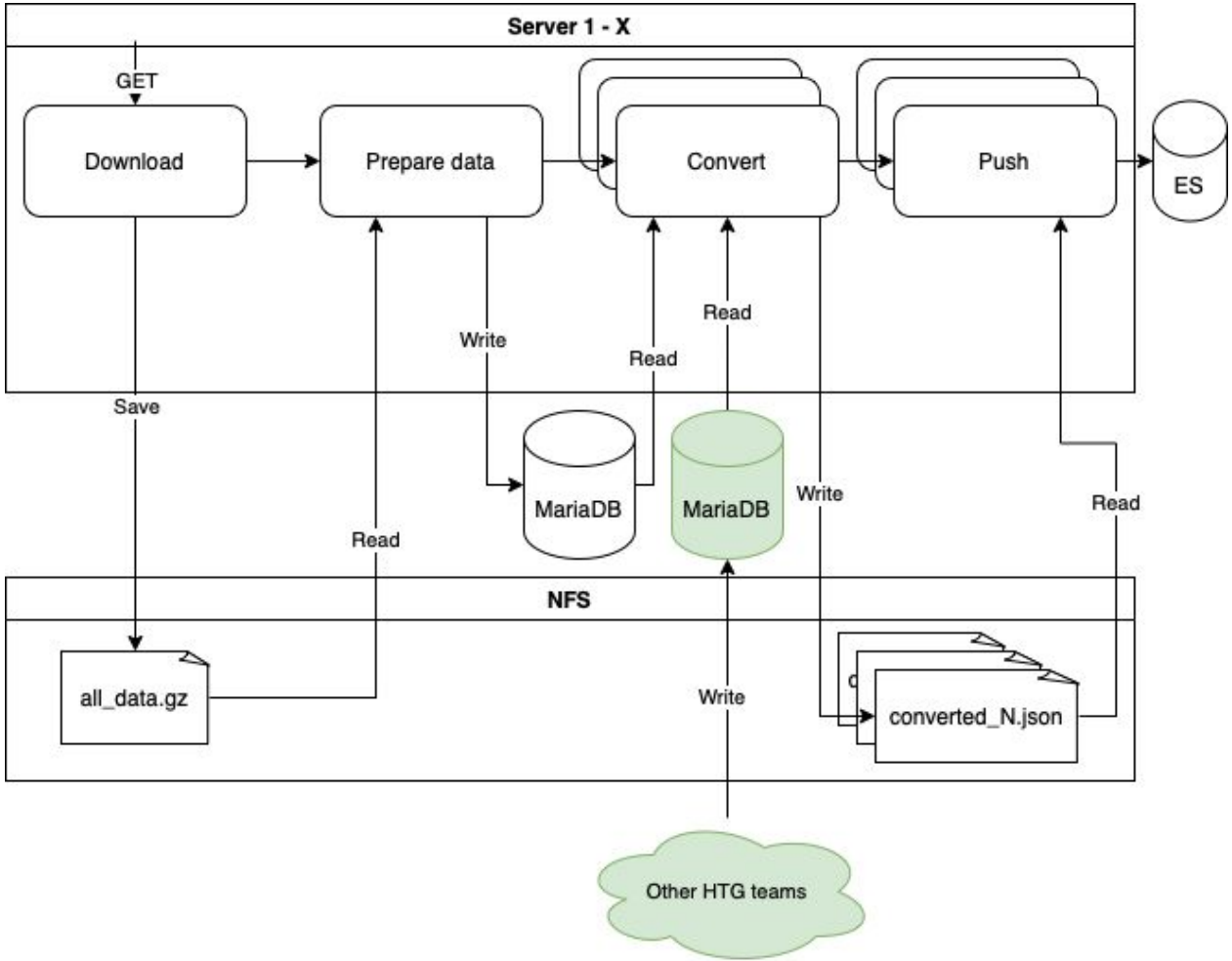
home to go

Sixth improvement

Enrich data from third sources

Distance to water
Sorting data

Enrich data



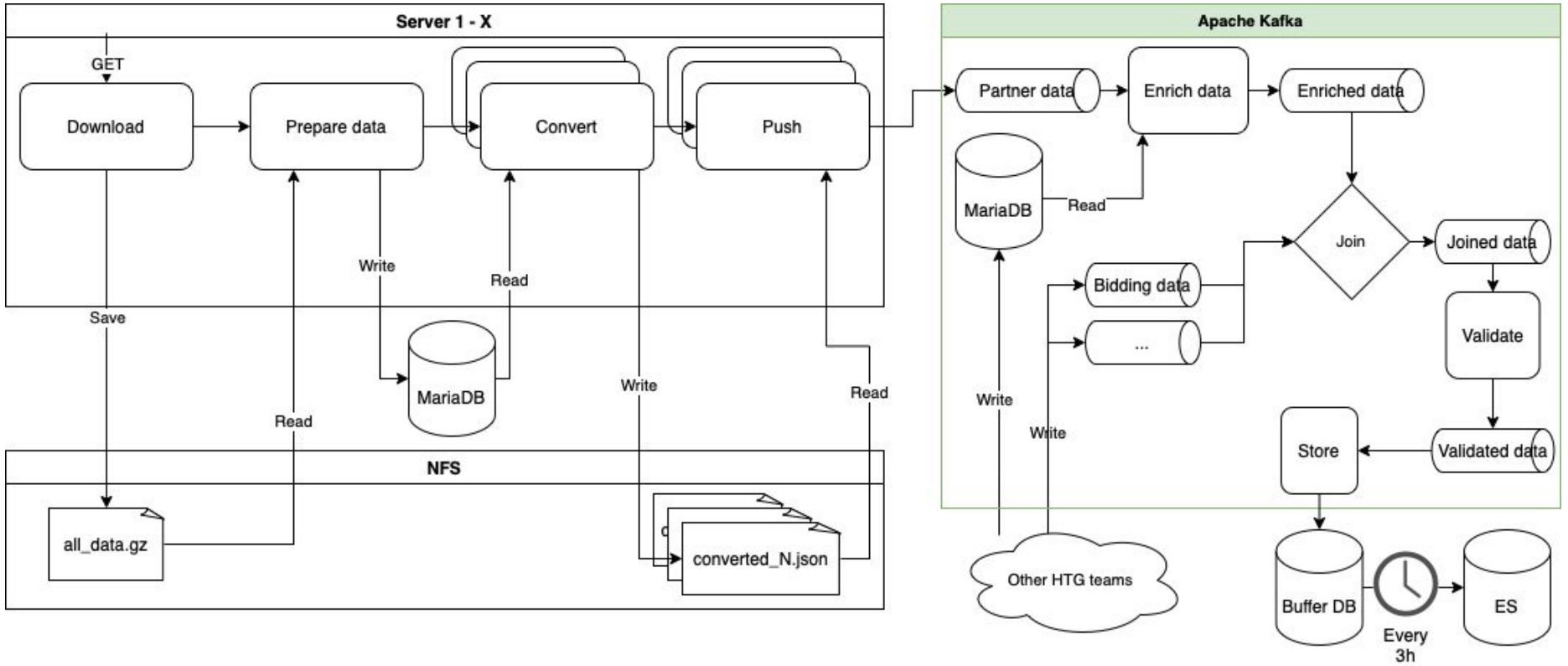
home to go

N improvement

Parallel data update

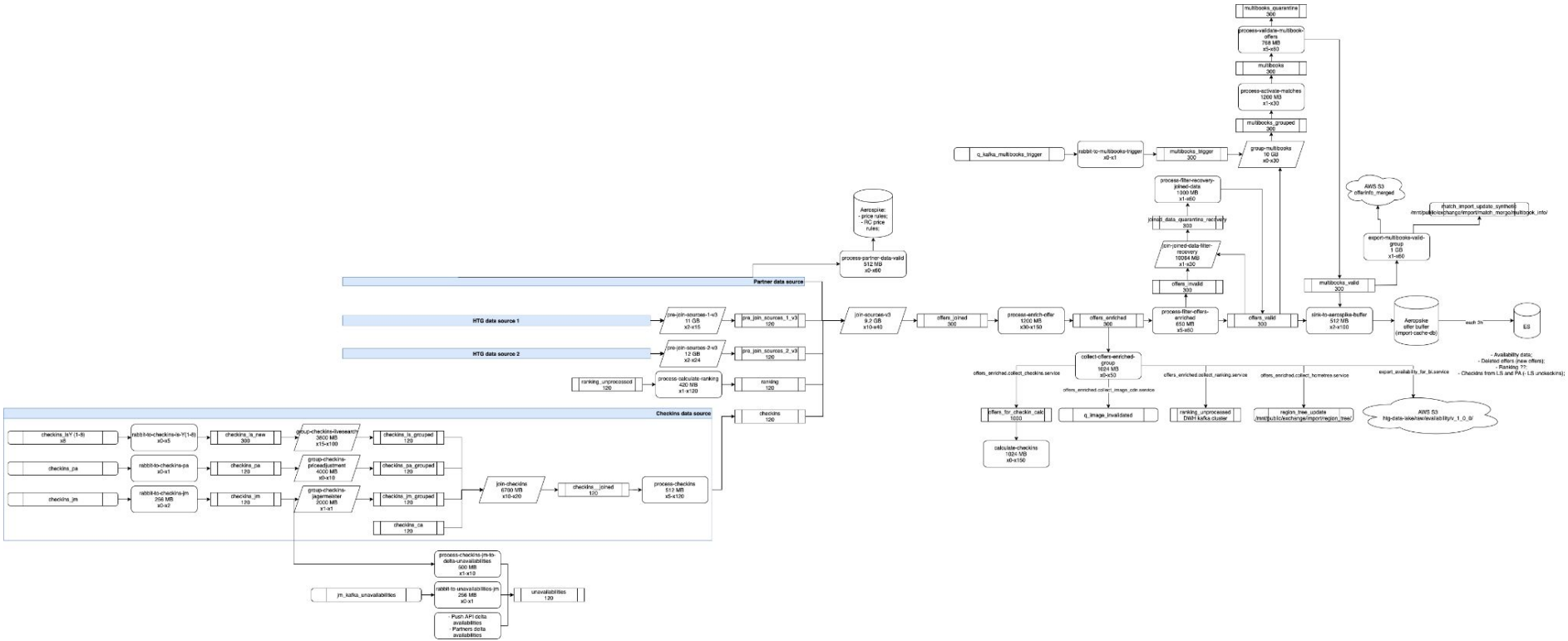
Bidding data
Match and merge offers

Start using APACHE kafka



home to go

Production Kafka pipelines



Production Kafka

- Today we have
  - 131 topics
  - 12 brokers
  - 37 consumer groups (scaling 1 - 300 processes)
  - ~6.5 billion messages per day

Rate	Mean	1 min	5 min	15 min
Messages in /sec	46k	32k	34k	40k
Bytes in /sec	482m	297m	344m	416m
Bytes out /sec	3.1b	2.3b	2.5b	3.1b

home to go

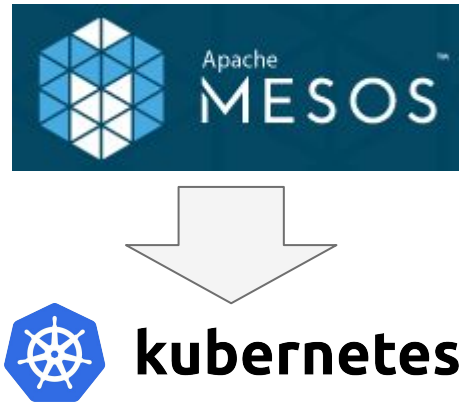
N+1 improvement

Dynamic resource pool

Increased user traffic
Seasons

Changing

- Mesos
  - Last release 24 Nov 2020 version 1.11.0
  - New OS don’t have builded package
- Kubernetes
  - Fast auto scaling
  - Better resource utilization
  - Containerization



home to go

You aren't gonna need it

YAGNI

- From Wiki
  - is a principle that states a programmer should not add functionality until deemed necessary

- From personal experience
  - Business changes a lot
  - We keep discovering new opportunities
  - Do what is needed only today with the opportunity to expand

Let’s connect

Darius Kasiulevičius

https://www.linkedin.com/in/darius-kasiulevicius/