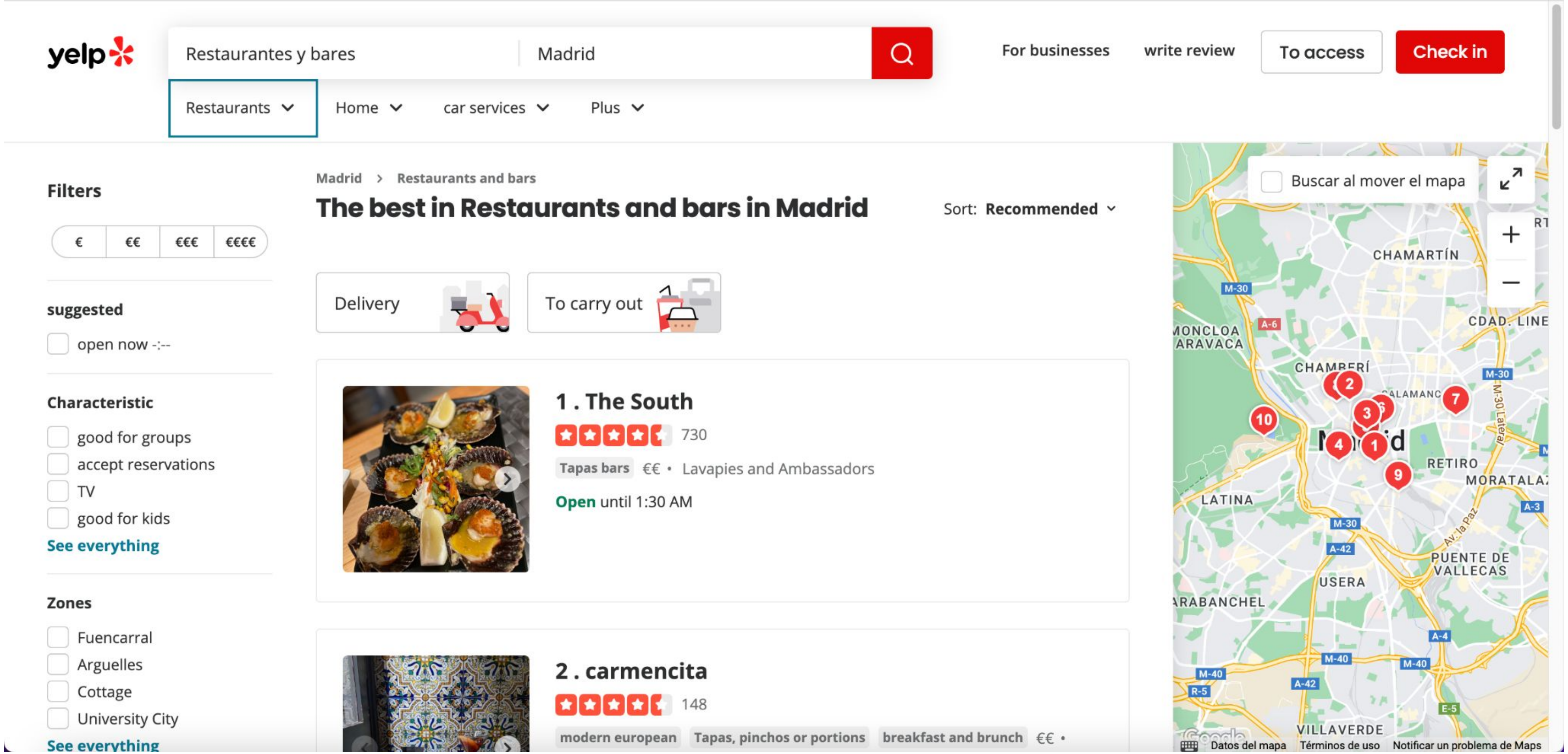
About Yelp



Filters

€ | €€ | €€€ | €€€€

suggested

□ open now -:--

Characteristic

□ good for groups
□ accept reservations
□ TV
□ good for kids

See everything

Zones

□ Fuencarral
□ Arguelles
□ Cottage
□ University City

See everything

Madrid > Restaurants and bars

The best in Restaurants and bars in Madrid

Sort: Recommended

Delivery

To carry out

1 . The South

730

Tapas bars €€ • Lavapies and Ambassadors

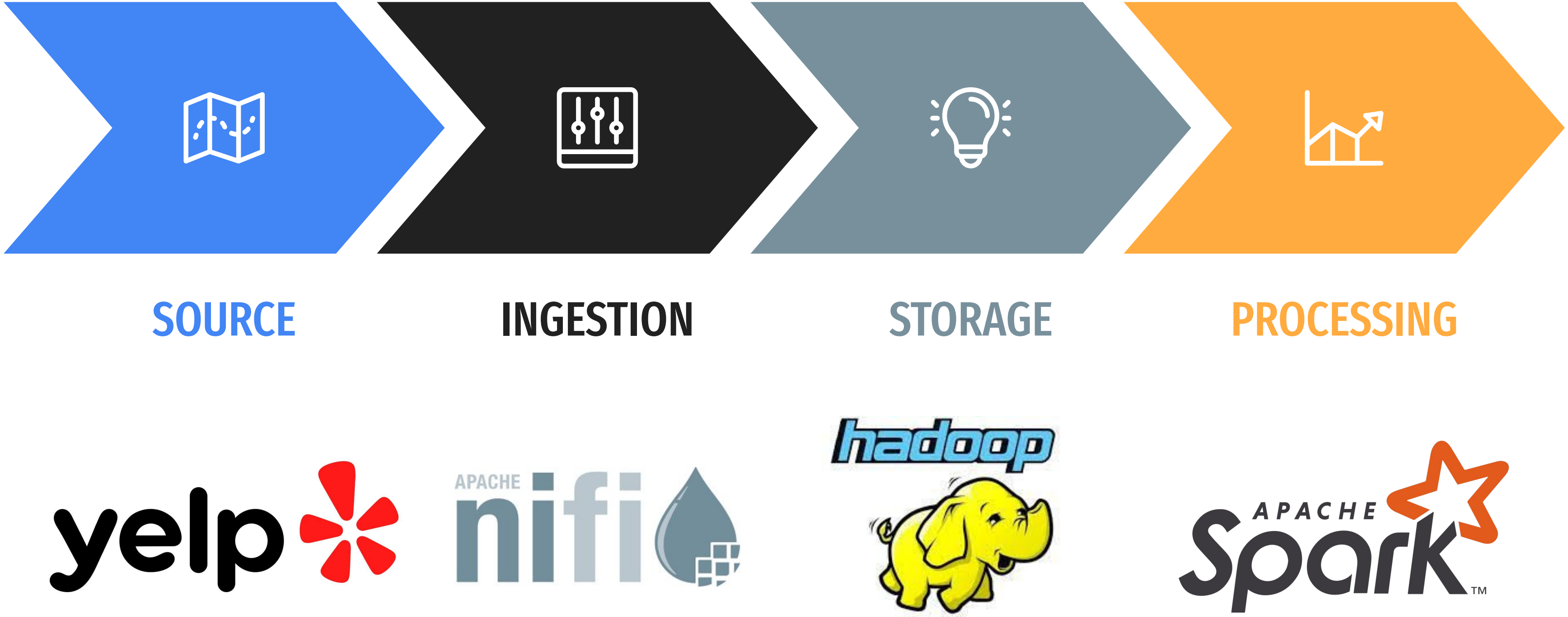
Open until 1:30 AM

2 . carmencita

148

modern european Tapas, pinchos or portions breakfast and brunch €€ •

Data Flow – Big Picture



SOURCE

INGESTION

STORAGE

PROCESSING

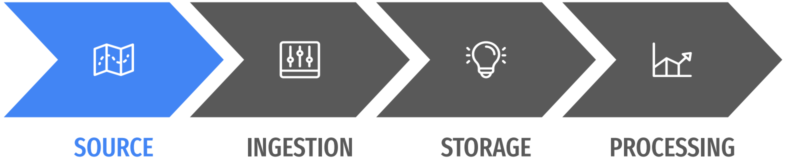
yelp

APACHE nifi

hadoop

APACHE Spark™

About Data Source



Yelp has free REST based API

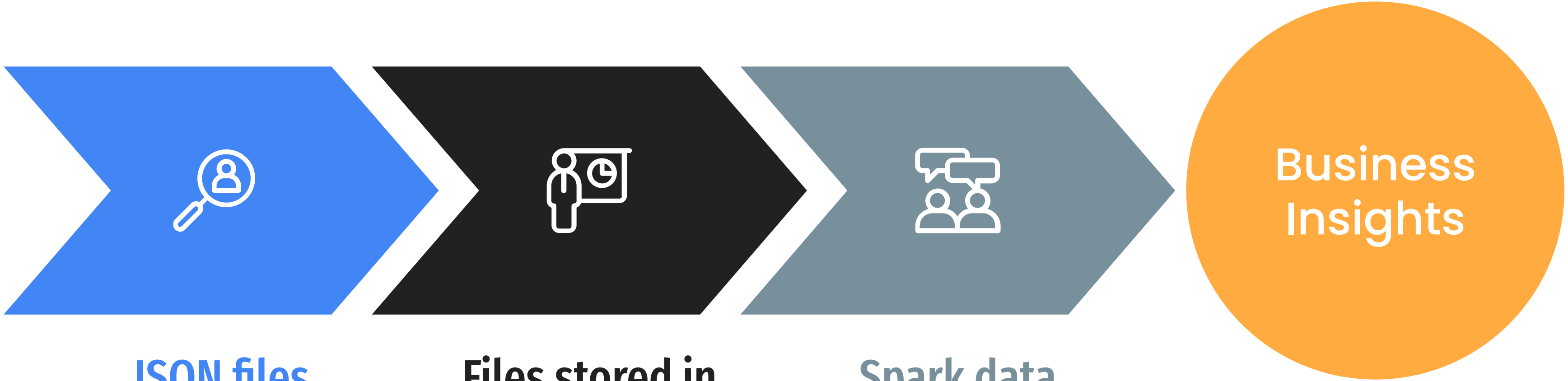
Using the search businesses API call we were able to pull in restaurants in Madrid

The JSON also gave us information on ratings, review counts and how expensive these restaurants are

The API call that we used: https://api.yelp.com/v3/businesses/search?location=Madrid

Sample API response: response.json

Data Flow – What we did?



Business Insights

JSON files through API

JSON files pulled in from yelp API using the invoke HTTP process

Files stored in HDFS

JSON files are stored at given path in HDFS

Spark data processing

Using Spark SQL to process the data

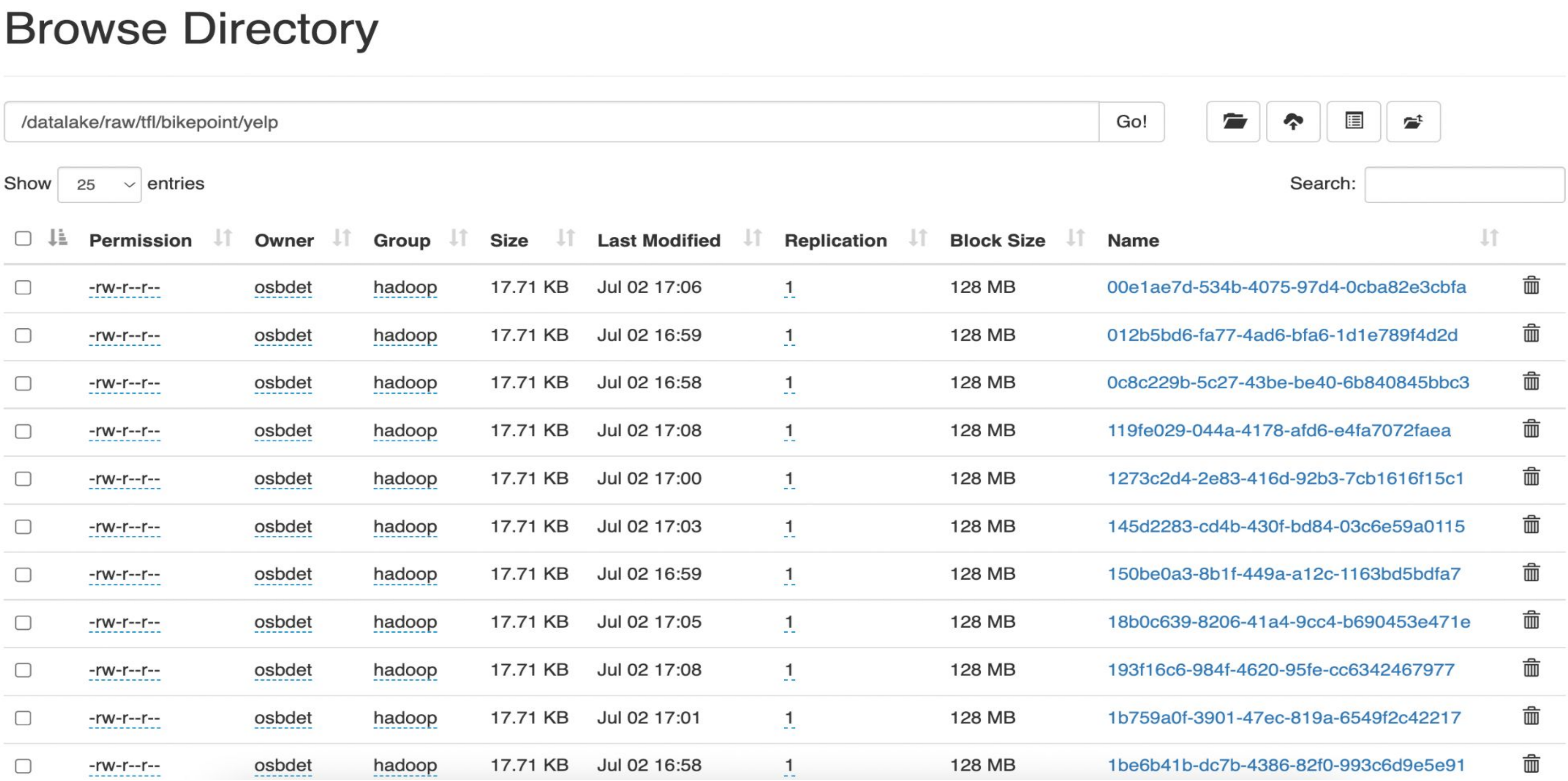
We answer the business questions

NiFi flow

1. Invoke HTTP Process
2. Updates on the flow files
3. Storing the JSON File into HDFS
4. Creating Spark session to answer the business questions

Files stored in HDFS

SOURCE
INGESTION
STORAGE
PROCESSING



Browse Directory

/datalake/raw/tfl/bikepoint/yelp Go!

Show 25 entries

Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:06	1	128 MB	00e1ae7d-534b-4075-97d4-0cba82e3cbfa	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 16:59	1	128 MB	012b5bd6-fa77-4ad6-bfa6-1d1e789f4d2d	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 16:58	1	128 MB	0c8c229b-5c27-43be-be40-6b840845bbc3	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:08	1	128 MB	119fe029-044a-4178-afd6-e4fa7072faea	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:00	1	128 MB	1273c2d4-2e83-416d-92b3-7cb1616f15c1	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:03	1	128 MB	145d2283-cd4b-430f-bd84-03c6e59a0115	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 16:59	1	128 MB	150be0a3-8b1f-449a-a12c-1163bd5bdfa7	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:05	1	128 MB	18b0c639-8206-41a4-9cc4-b690453e471e	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:08	1	128 MB	193f16c6-984f-4620-95fe-cc6342467977	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 17:01	1	128 MB	1b759a0f-3901-47ec-819a-6549f2c42217	🗑
□	-rw-r--r--	osbdet	hadoop	17.71 KB	Jul 02 16:58	1	128 MB	1be6b41b-dc7b-4386-82f0-993c6d9e5e91	🗑

Business Questions

SOURCE

INGESTION

STORAGE

PROCESSING

1 How many restaurants in Madrid have a rating of 4.0, 4.5 and 5.0

2 Distribution of restaurants based on how expensive they are (GroupBy)

3 Correlation between restaurant prices and ratings

4 Correlation between restaurant review counts and ratings

5 Jupyter Notebook

Business Questions

SOURCE

INGESTION

STORAGE

PROCESSING

1 How many restaurants in Madrid have a rating of 4.0, 4.5 and 5.0

4.0	4.5	5.0
267	1424	89

Business Questions

SOURCE

INGESTION

STORAGE

PROCESSING

2 Distribution of restaurants based on how expensive they are

€	€€	€€€	Other
534	623	267	534

Business Questions

SOURCE
INGESTION
STORAGE
PROCESSING

3 Correlation between restaurant prices and ratings

Correlation
0.385

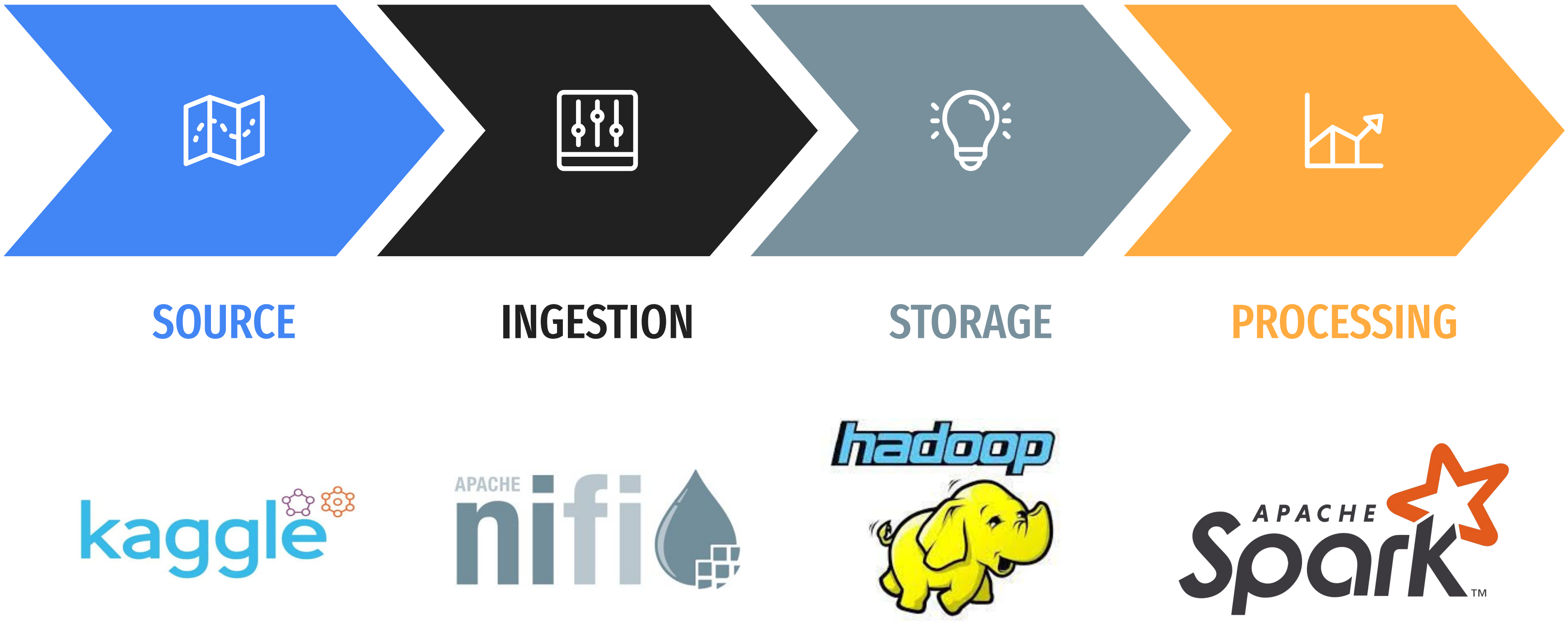
Business Questions

SOURCE INGESTION STORAGE PROCESSING

4 Correlation between restaurant review counts and ratings

Correlation
0.233

Data Flow – Big Picture



SOURCE

INGESTION

STORAGE

PROCESSING

NiFi flow

1. Using the ListFile and Fetch File process on Nifi to pull in the JSON File

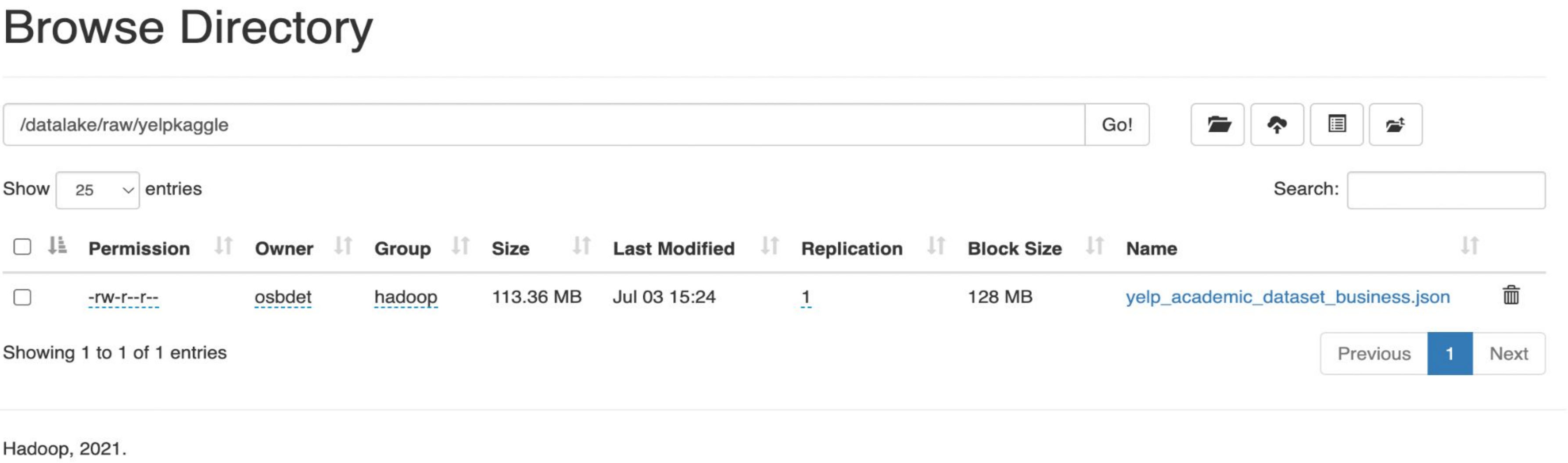
2. Updates on the flow files

3. Storing the JSON File into HDFS at a given path

4. Creating Spark session to answer the business questions

Files stored in HDFS

SOURCE
INGESTION
STORAGE
PROCESSING



Browse Directory

/datalake/raw/yelpkaggle Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	osbdet	hadoop	113.36 MB	Jul 03 15:24	1	128 MB	yelp_academic_dataset_business.json	🗑

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2021.

Business Questions

SOURCE

INGESTION

STORAGE

PROCESSING

1. Top 5 states with the highest average rating but with review_counts more than 1000

2. Top 5 states with the highest number of businesses listed on yelp

3. Which is the highest rated restaurant in New Orleans (review_count > 1000)

4. Jupyter Notebook

Business Questions

SOURCE INGESTION STORAGE PROCESSING

1 Top 5 states with the highest average rating but with review_counts more than 1000

California	Nevada	Idaho	Louisiana	Florida
3.99	3.74	3.71	3.68	3.61

Business Questions

2 Top 5 states with the highest number of businesses listed on yelp

Pennsylvania	Florida	Tennessee	Indiana	Missouri
34,039	26,330	12,056	11,247	10,913

Business Questions

SOURCE INGESTION STORAGE PROCESSING

3 Using Spark SQL, get the highest rated restaurant in New Orleans

District Donuts Sliders Brew
4.5

Appendix

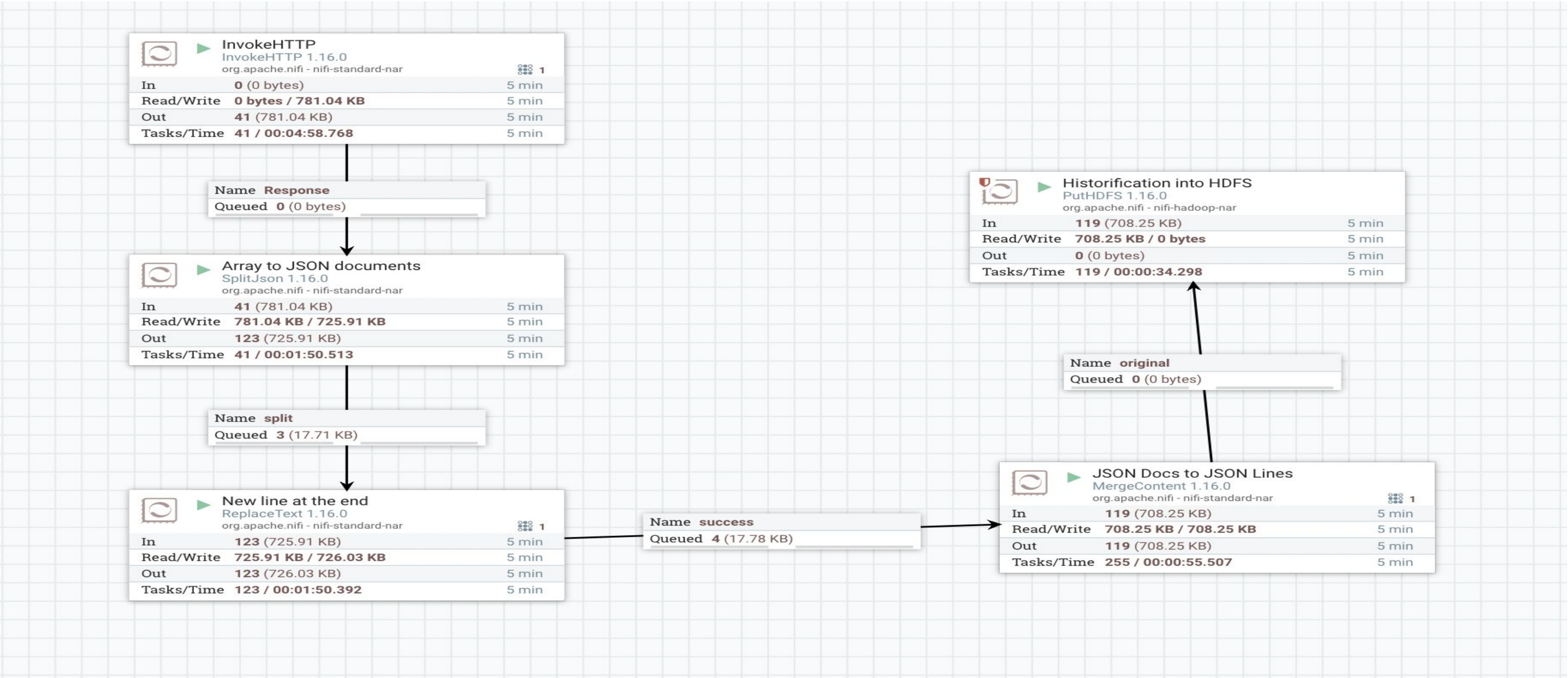
NiFi Flow using API

SOURCE

INGESTION

STORAGE

PROCESSING



InvokeHTTP
InvokeHTTP 1.16.0
org.apache.nifi - nifi-standard-nar 1

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 781.04 KB**	5 min
Out	**41** (781.04 KB)	5 min
Tasks/Time	**41 / 00:04:58.768**	5 min

Name **Response**
Queued **0** (0 bytes)

Array to JSON documents
SplitJson 1.16.0
org.apache.nifi - nifi-standard-nar

In	**41** (781.04 KB)	5 min
Read/Write	**781.04 KB / 725.91 KB**	5 min
Out	**123** (725.91 KB)	5 min
Tasks/Time	**41 / 00:01:50.513**	5 min

Name **split**
Queued **3** (17.71 KB)

New line at the end
ReplaceText 1.16.0
org.apache.nifi - nifi-standard-nar 1

In	**123** (725.91 KB)	5 min
Read/Write	**725.91 KB / 726.03 KB**	5 min
Out	**123** (726.03 KB)	5 min
Tasks/Time	**123 / 00:01:50.392**	5 min

Name **success**
Queued **4** (17.78 KB)

JSON Docs to JSON Lines
MergeContent 1.16.0
org.apache.nifi - nifi-standard-nar 1

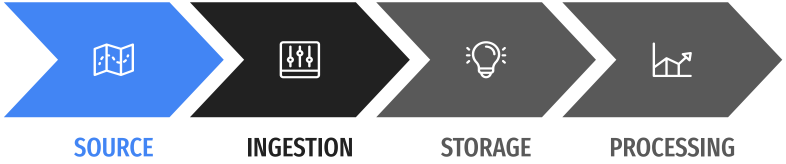
In	**119** (708.25 KB)	5 min
Read/Write	**708.25 KB / 708.25 KB**	5 min
Out	**119** (708.25 KB)	5 min
Tasks/Time	**255 / 00:00:55.507**	5 min

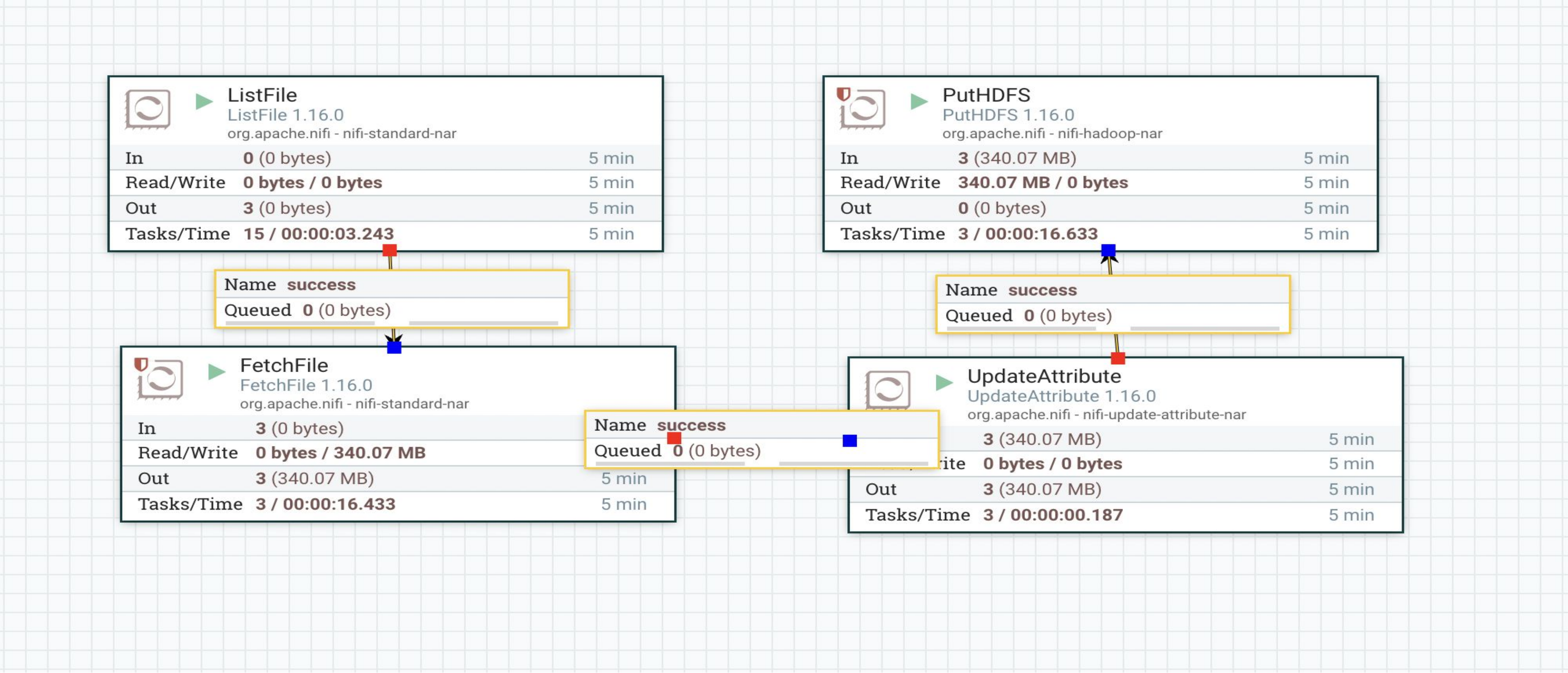
Name **original**
Queued **0** (0 bytes)

Historification into HDFS
PutHDFS 1.16.0
org.apache.nifi - nifi-hadoop-nar

In	**119** (708.25 KB)	5 min
Read/Write	**708.25 KB / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**119 / 00:00:34.298**	5 min

NiFi Flow for Kaggle Dataset





ListFile
ListFile 1.16.0
org.apache.nifi - nifi-standard-nar

In	**0** (0 bytes)	5 min
Read/Write	**0 bytes / 0 bytes**	5 min
Out	**3** (0 bytes)	5 min
Tasks/Time	**15 / 00:00:03.243**	5 min

Name **success**
Queued **0** (0 bytes)

FetchFile
FetchFile 1.16.0
org.apache.nifi - nifi-standard-nar

In	**3** (0 bytes)	
Read/Write	**0 bytes / 340.07 MB**	
Out	**3** (340.07 MB)	5 min
Tasks/Time	**3 / 00:00:16.433**	5 min

Name **success**
Queued **0** (0 bytes)

PutHDFS
PutHDFS 1.16.0
org.apache.nifi - nifi-hadoop-nar

In	**3** (340.07 MB)	5 min
Read/Write	**340.07 MB / 0 bytes**	5 min
Out	**0** (0 bytes)	5 min
Tasks/Time	**3 / 00:00:16.633**	5 min

Name **success**
Queued **0** (0 bytes)

UpdateAttribute
UpdateAttribute 1.16.0
org.apache.nifi - nifi-update-attribute-nar

	3 (340.07 MB)	5 min
ite	**0 bytes / 0 bytes**	5 min
Out	**3** (340.07 MB)	5 min
Tasks/Time	**3 / 00:00:00.187**	5 min

Nifi flows XML and Jupyter Notebooks to download

Project Files

Thank you!