



# Repsol analytics consultancy project

# Data source

- We decided to exploit the Daily Prices for Spanish Gas Stations (2007-2022), retrieved from Kaggle
- Since 2007, Spanish gas stations must send their daily prices to the Spanish regulatory authorities, as described in ITC/2308/2007. This data is published daily in the Spanish Open Data Catalog.
- Our dataset is the aggregated version of this data
- Parquet format (column-based storage format, good for querying and good compression)

# Data source

**Daily Prices for Spanish Gas Stations  
(2007-2022)**



**spanish\_gas\_prices\_000000000000.parquet (79.33 MB)**

# Data source format



# Parquet

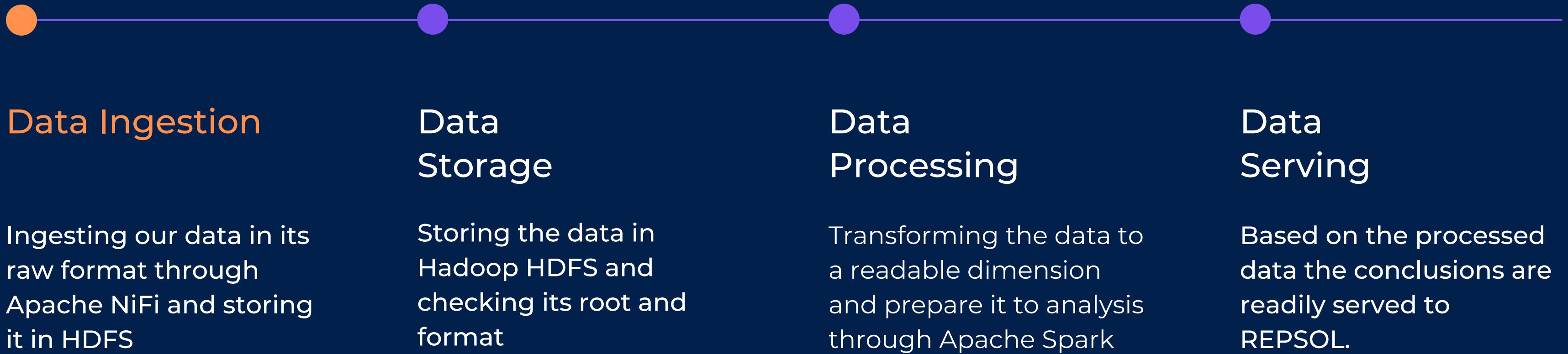
- Structured data tables, useful for storing big data.
- Parquet format (column-based storage format, good for querying and good compression).
- Gives better-summarized data and follows type-specific encoding.



We simulated an analytics consultancy for Repsol. The company asked us to provide them with some useful business insights:

- Price development for the two most common fuel types for the period 2020 – 2022.
- Other top competitors to Repsol in the year 2022.
- Top 10 and bottom 10 gas stations of the Repsol in terms of price ranking.

# Data RoadMap



# Batch Ingestion

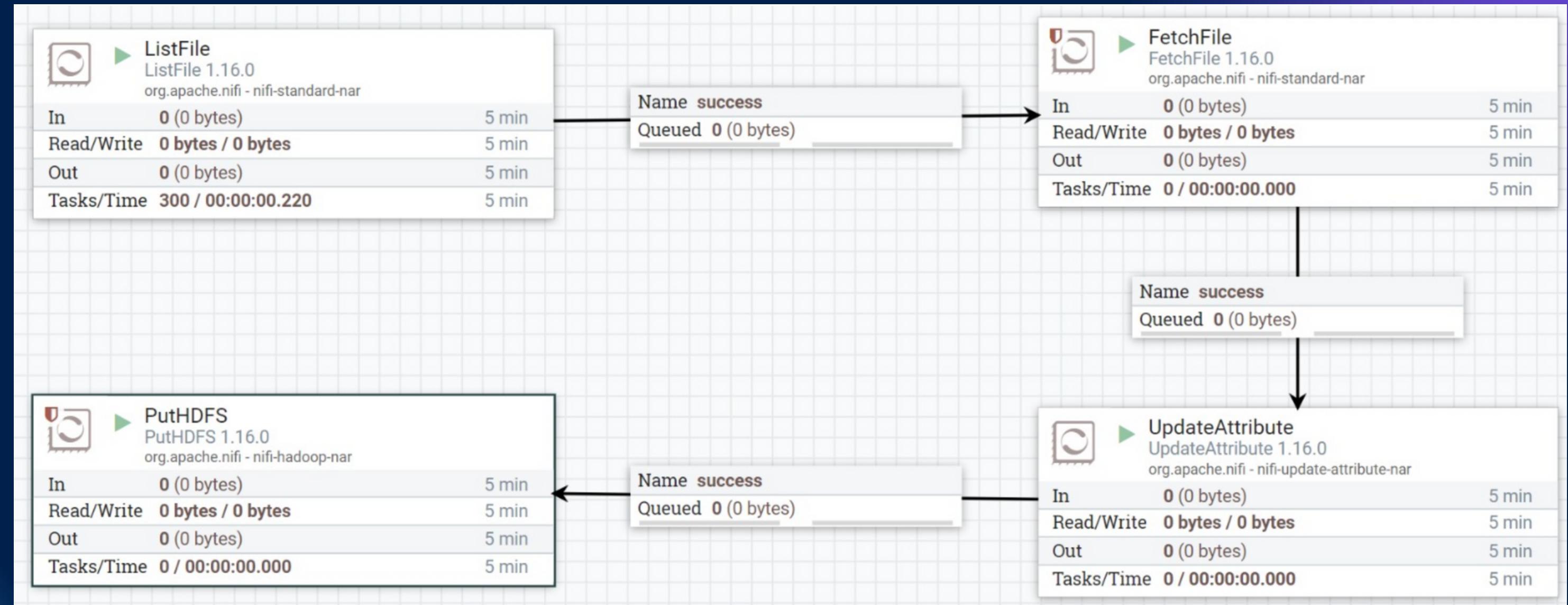
After uploading our main .parquet file in a Folder, we set up the Apache NiFi framework with 4 processors in sequence:

- ListFile
- FetchFile
- UpdateAttribute
- PutHDFS

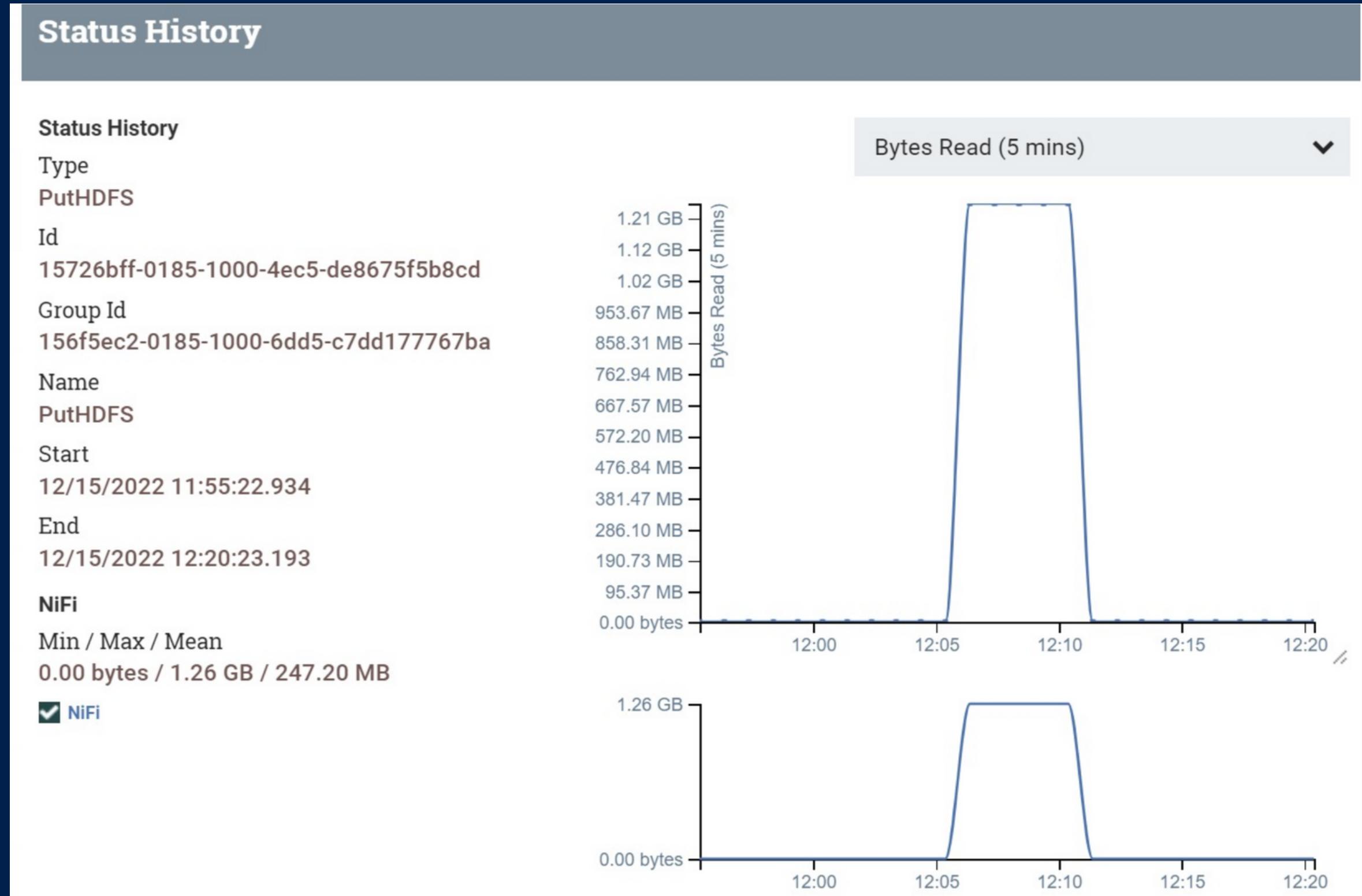


Each of those has been configured to ingest the data, store it in HDFS and delete it from the root source.

# Batch Ingestion



# Data read by NiFi



# NiFi Data Provenance

NiFi Data Provenance

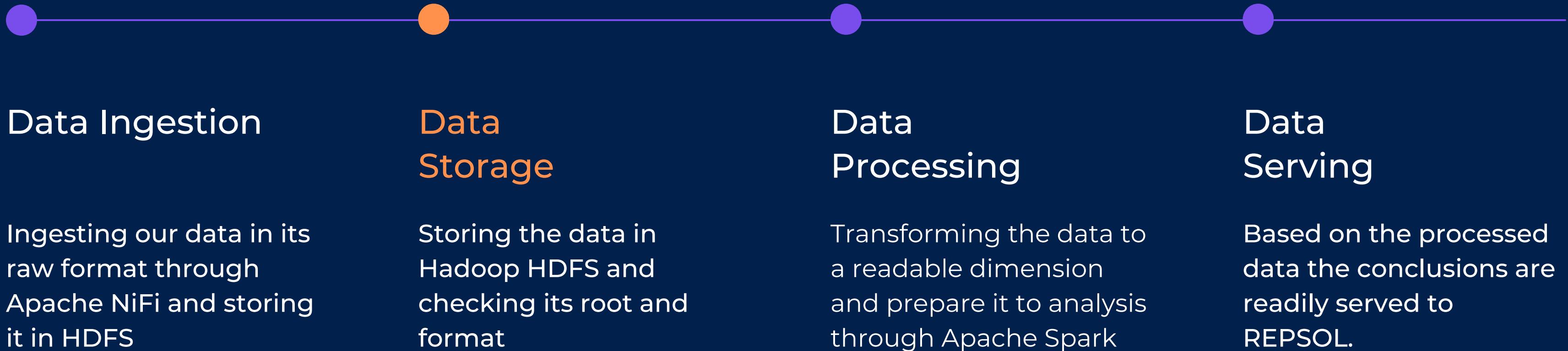
Displaying 17 of 17  
Oldest event available: 12/15/2022 12:05:11 CET

Showing the events that match the specified query. Clear search

Filter

Date/Time	Type	FlowFileUuid	Size	Component Name	Component Type	Actions
12/15/2022 12:05:11.605 CET	CREATE	1f6844dc-ee4d-4551-894f-7ebc9edcad85	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	4dfbbd17-fbb1-4c0f-897f-3ff838adef64	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	1b47616a-6b76-45c1-ae67-1d0768c414df	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	0edf6e9e-3db3-4b09-afdb-d88cca4a36c7	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	b2f2df2a-9d1b-472e-8aa2-a47f17b5ed57	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	47f93bf5-1dfd-49c8-a85f-4c56630b8a53	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	674e3d6b-c54d-43ad-8691-9a84842f9f9f	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	ef8d508e-10da-4551-89e1-008f0e86fa74	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	f943a73c-cd4f-4a8d-8b93-75e27d0529f3	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	734a20d6-ad05-4b55-8fc1-68abaf84f5e8	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	bb85a232-0a68-4a8b-8593-03468afde928	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	42ba7ed3-06c1-4675-8f52-177d8a40b666	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	00ff12b2-8b62-4435-bb04-6b6facddee3	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	3d3e60fb-61bd-4c84-9c3d-5aefd995169b	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	d3453988-8c48-4659-82aa-badac4853219	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.605 CET	CREATE	a66937ff-4730-4cd3-b52e-e09bf83be15c	0 bytes	ListFile	ListFile	 
12/15/2022 12:05:11.601 CET	CREATE	3cff6038-788f-49c3-86b7-440c6d6c6b07	0 bytes	ListFile	ListFile	 

# Data RoadMap

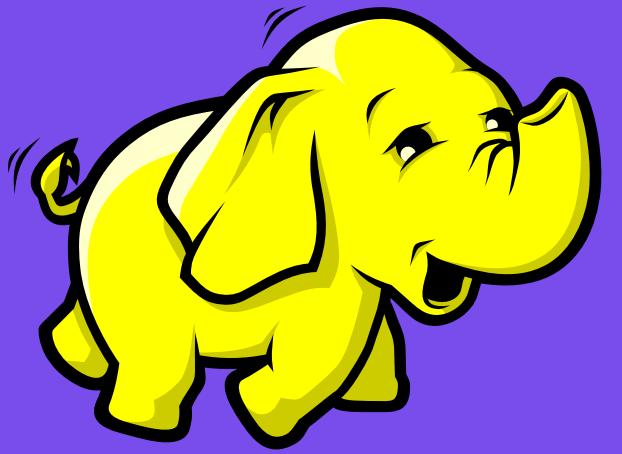


# Batch Storage

## Distributed File System

- Unit of storage: **File**
- Unit of structure: **Folder**
- Structure: **Hierarchical (File System)**
- Data access: **Path**

## Hadoop HDFS



## Core Idea

Run applications (processing) in the nodes that store the data (storage).

VS single machine:  
1. Efficient  
2. Cost-effective

## Main Drawbacks

- All its nodes up and running
- Low-latency

# Batch Storage

## Browse Directory

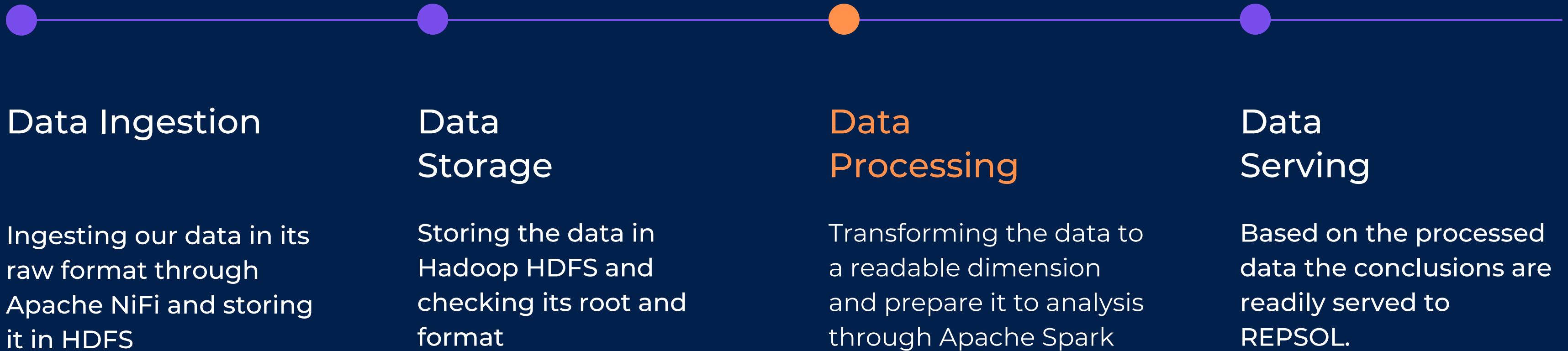
/datalake/raw/gas									Go!				
Show 25 entries									Search:				
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.66 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000000.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.64 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000001.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.72 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000002.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.51 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000003.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.68 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000004.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.59 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000005.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.63 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000006.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.62 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000007.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.63 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000008.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.59 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000009.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.61 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000010.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.51 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000011.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.6 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000012.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.73 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000013.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.6 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000014.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.56 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000015.parquet					
<input type="checkbox"/>	-rw-r--r--	osbdet	hadoop	75.55 MB	Dec 15 12:05	1	128 MB	spanish_gas_prices_000000000016.parquet					

Showing 1 to 17 of 17 entries

Previous 1 Next

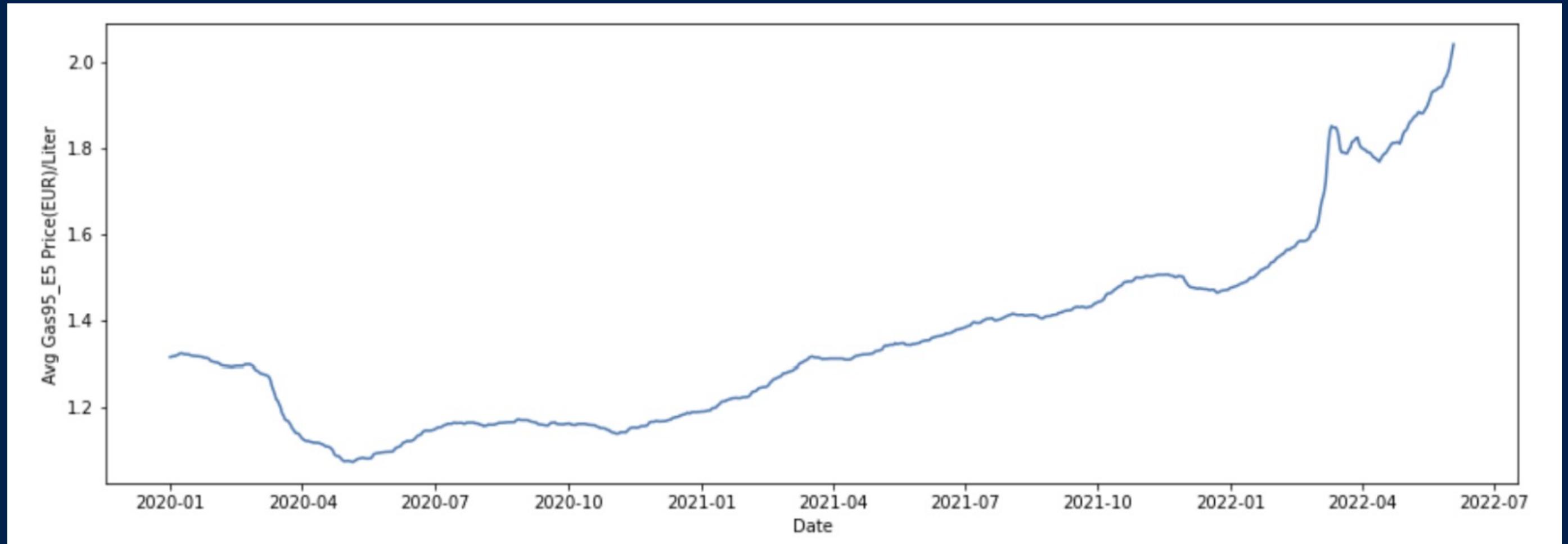


# Data RoadMap



# Batch Processing

Data Analysis: (GAS) Price Development 2020-2022, EUR/Liter



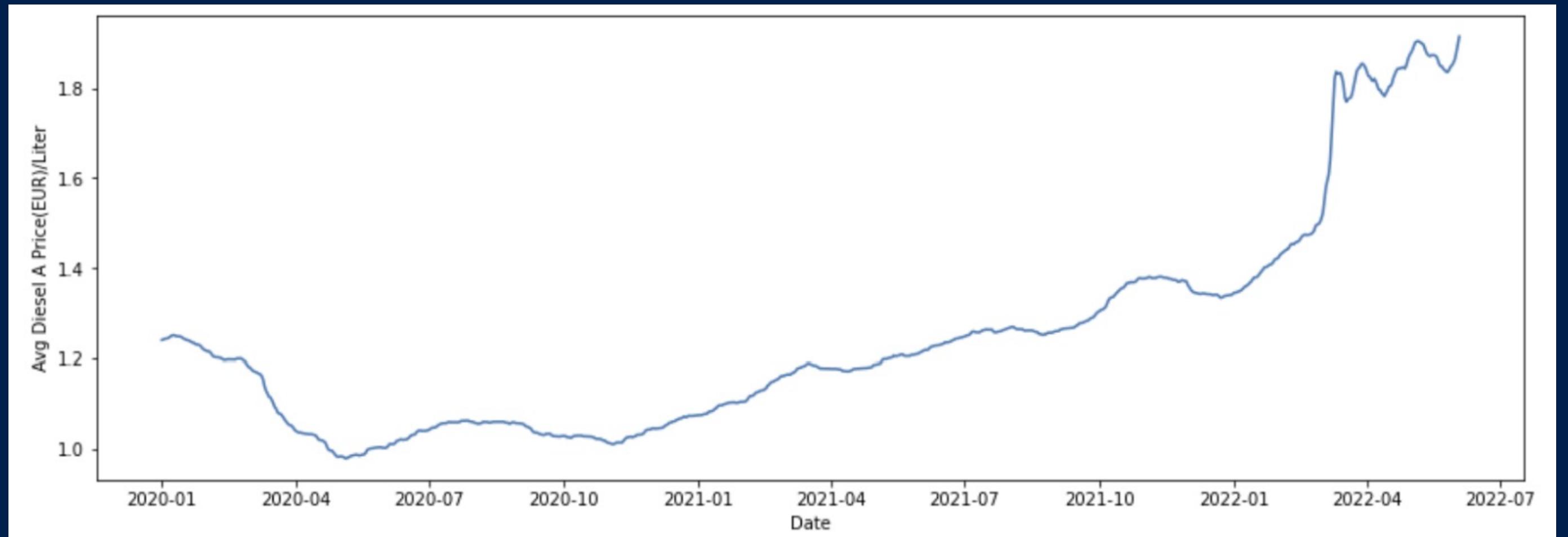
Price drop in Q1-Q2 2020 due to lockdowns

Price increase in Q1-Q2 2022 due to supply chain issues and RU/UA conflict



# Batch Processing

Data Analysis: (DIESEL) Price Development 2020-2022, EUR/Liter



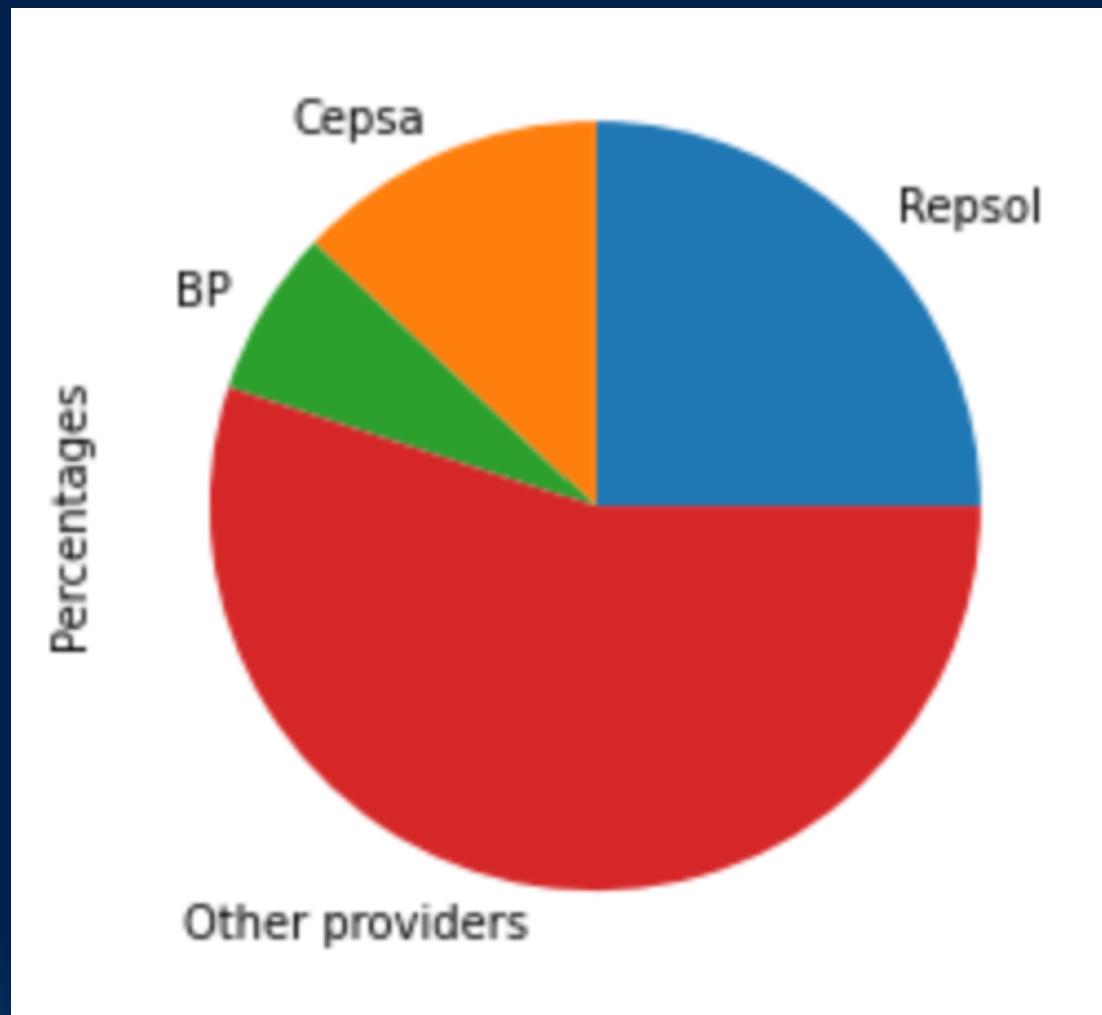
Price drop in Q1-Q2 2020 due to lockdowns

Price increase in Q1-Q2 2022 due to supply chain issues and RU/UA conflict



# Batch Processing

Data Analysis: Repsol has the highest share of stations on the Spanish market (25% of total gas stations)

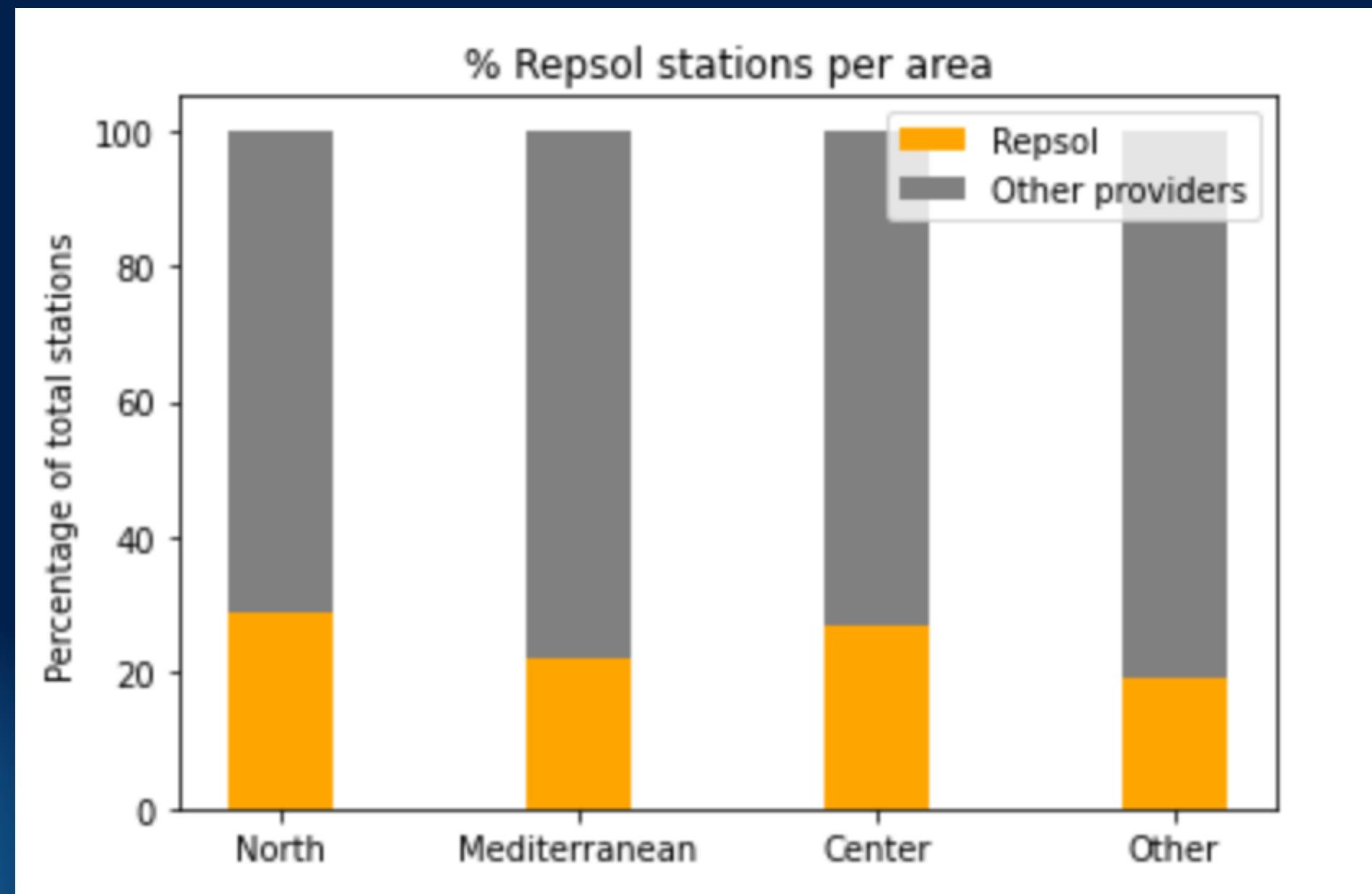


Overview of gas stations,  
% vs. total count of stations in Spain,  
split by company:

- Repsol: 25%
- Cepsa: 13%
- BP: 7%
- Rest of companies: 55%

# Batch Processing

Data Analysis: Repsol is evenly present in all macro-areas of Spain, with no clear geographical outliers (~25% in each).



\*We divided Spain in four macro-areas:  
North regions, Mediterranean ones , Center  
ones and Others (incl. islands, offshore  
territories etc.)

# Batch Processing



(GAS) Ranking: Repsol stations with bottom 10 average prices in 2022

station_id	town	province_name	region_name	Area	average 2022 gas prices
12703	MONTEMAYOR	CÓRDOBA	Andalucia	Mediterranean	1.259000000000
10356	NAVALMORAL DE LA	CÁCERES	Extremadura	Center	1.289000000000
5844	BADAJOZ	BADAJOZ	Extremadura	Center	1.289000000000
1920	ESTIVELLA	VALENCIA / VALÈNCIA	Comunidad Valenciana	Mediterranean	1.289000000000
11184	RIBESALBES	CASTELLÓN / CASTELLÓ	Comunidad Valenciana	Mediterranean	1.290000000000
14742	GIJON	ASTURIAS	Asturias	North	1.293482353000
9824	TORREVIEJA	ALICANTE	Comunidad Valenciana	Mediterranean	1.367000000000
11222	TORREVIEJA	ALICANTE	Comunidad Valenciana	Mediterranean	1.375923077000
10715	NAVALVILLAR DE PE...	BADAJOZ	Extremadura	Center	1.379000000000
4901	HORCAJO DE LOS MO...	CIUDAD REAL	Castilla la Mancha	Center	1.399000000000

Lowest gas prices are concentrated mostly in Mediterranean and Center. We recommend a further cost/revenue deep-dive to establish how viable these gas stations are commercially.

# Batch Processing



(DIESEL) Ranking: Repsol stations with bottom 10 average prices in 2022

station_id	town	province_name	region_name	Area	average 2022 diesel prices
10715	NAVALVILLAR DE PE...	BADAJOZ	Extremadura	Center	1.039000000000000
1920	ESTIVELLA	VALENCIA / VALÈNCIA	Comunidad Valenciana	Mediterranean	1.109000000000000
3848	VALENCIA	VALENCIA / VALÈNCIA	Comunidad Valenciana	Mediterranean	1.109000000000000
12703	MONTEMAYOR	CÓRDOBA	Andalucia	Mediterranean	1.109000000000000
5844	BADAJOZ	BADAJOZ	Extremadura	Center	1.139000000000000
9824	TORREVIEJA	ALICANTE	Comunidad Valenciana	Mediterranean	1.187000000000000
11184	RIBESALBES	CASTELLÓN / CASTELLÓ	Comunidad Valenciana	Mediterranean	1.215000000000000
8842	MIGUEL ESTEBAN	TOLEDO	Castilla la Mancha	Center	1.240000000000000
4901	HORCAJO DE LOS MO...	CIUDAD REAL	Castilla la Mancha	Center	1.249000000000000
14842	BARBATE	CÁDIZ	Andalucia	Mediterranean	1.249000000000000

Lowest diesel prices are concentrated mostly in Mediterranean and Center. We recommend a further cost/revenue deep-dive to establish how viable these gas stations are commercially.

# Batch Processing



(GAS) Ranking: Repsol stations with top 10 average prices in 2022

station_id	town	province_name	region_name	Area	average 2022 gas prices
15522	ALQUEZAR	HUESCA	Aragón	North	2.13000000000000
15541	SANTANDER	CANTABRIA	Cantabria	North	2.10900000000000
15247	VALDEALGORFA	TERUEL	Aragón	North	2.09000000000000
15406	ALHAMA DE MURCIA	MURCIA	Murcia	Mediterranean	2.075555560000
15515	MONFARRACINOS	ZAMORA	Castilla y León	Center	2.06300000000000
15481	CASA PUERTO	MURCIA	Murcia	Mediterranean	2.03900000000000
10498	LORCA	MURCIA	Murcia	Mediterranean	2.02900000000000
15531	TORRES DE ALBANCH...	JAÉN	Andalucia	Mediterranean	2.02450000000000
15217	VALENCIA	VALENCIA / VALÈNCIA	Comunidad Valenciana	Mediterranean	2.016304348000
6488	BAUL	GRANADA	Andalucia	Mediterranean	2.011935065000

The North region has the highest gas prices in the analysed period.

# Batch Processing



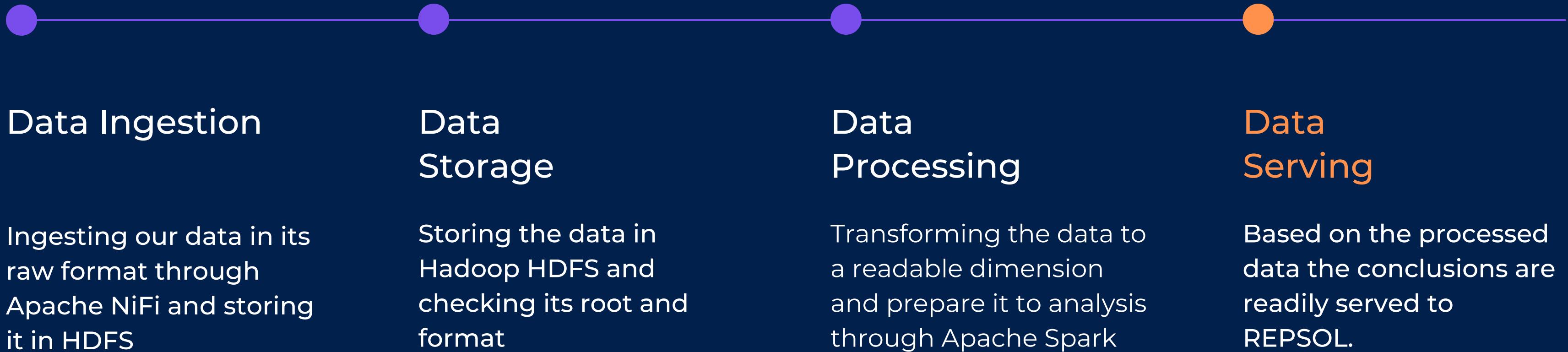
(DIESEL) Ranking: Repsol stations with top 10 average prices in 2022

station_id	town	province_name	region_name	Area	average 2022 diesel prices	
15247	VALDEALGORFA	...	TERUEL	Aragón	North	2.090000000000000
15531	TORRES DE ALBANCH...	...	JAÉN	Andalucía	Mediterranean	1.982500000000000
15492	MADRID	...	MADRID	Madrid	Center	1.9824210530000
3369	AMETLLA DE MAR (L...	...	TARRAGONA	Cataluña	Mediterranean	1.9796296300000
15447	VALDEPEÑAS	...	CIUDAD REAL	Castilla la Mancha	Center	1.9778888890000
15323	VALDEPEÑAS	...	CIUDAD REAL	Castilla la Mancha	Center	1.9778888890000
15522	ALQUEZAR	...	HUESCA	Aragón	North	1.9700000000000
15406	ALHAMA DE MURCIA	...	MURCIA	Murcia	Mediterranean	1.9670740740000
15355	LEPE	...	HUELVA	Andalucía	Mediterranean	1.9650701750000
15471	SANT MARTI D'ALBA...	...	BARCELONA	Cataluña	Mediterranean	1.9627500000000

Highest diesel prices are observed in most areas.

Region Aragón, and town Valdealgorfa in particular, is in Top 3 most expensive fuel types both for gas and diesel.

# Data RoadMap



# Data Serving

Through the exploitation of different instruments from the data value chain, we have been able to ingest and store the data in batch and process it in order to deliver objective data for our client, on which strategic decisions could be easily leveraged.

However, in order to take definitive decisions on the destiny of the highlighted distributors, it would be interesting to gather and analyze internal data regarding the volume of gasoline sold in each of the stations, geo-specific factors should also be taken into account.



# Thank You

Lia Dollison  
Ludovico Gandolfi  
Lakshmi Priyanka Jakka  
Hang Chi Ku  
Ivan Lopez  
Ignacio Pire Rubio  
Roman Zotkin