PROJECT 1

ENERGY DATA ANALYSIS WITH APACHE SPARK

- Alessandro Finocchi - 0340543
- A.A. 2024/2025

AGENDA

1 INTRODUCTION

2 ARCHITECTURE

3 DATA INGESTION

4 STORAGE AND PERSISTENCE

5 DATA PROCESSING

6 RESULTS

6 CONCLUSION

INTRODUCTION

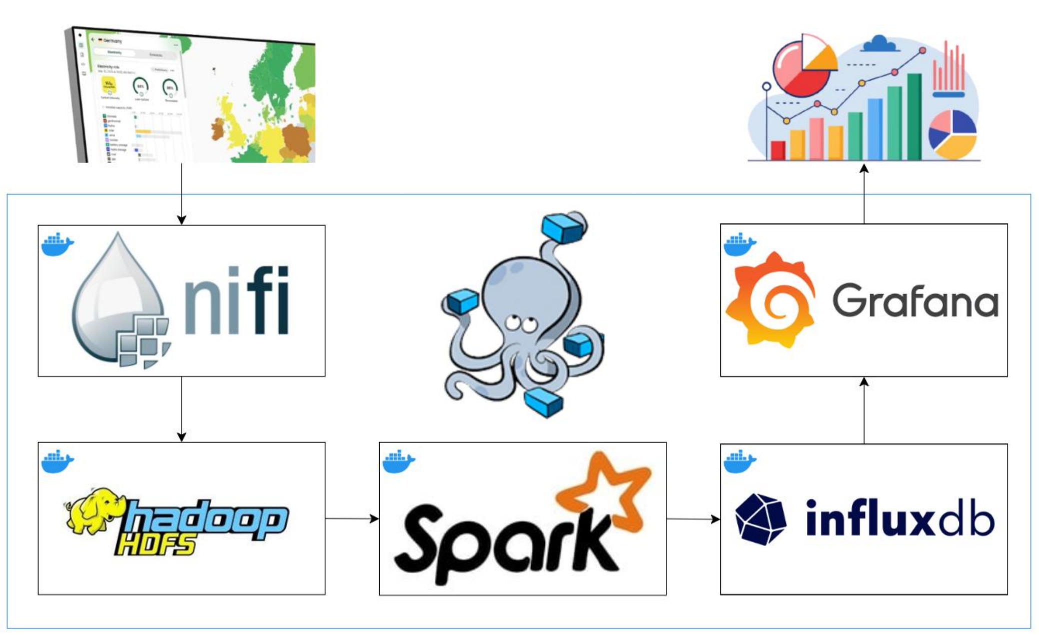
- Deploy analytical platform about energy data
- Focus on Italian and Sweden countries
- Benchmark queries under different configuration
  - Spark APIs (RDD, DataFrame and SQL)
  - File formats (CSV and Parquet)
  - Spark workers number
- Measure and compare query performances with varying configurations

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

ARCHITECTURE

- Orchestrator: Docker compose
- Data ingestion: Apache Nifi
- Storage and Persistence: Hadoop HDFS, InfluxDB
- Data processing: Apache Spark
- Data visualization: Grafana



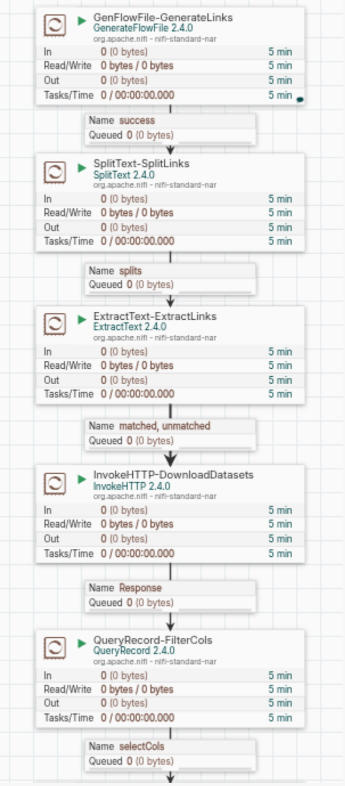
AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

DATA INGESTION

- Nifi flows to manage the ingestion of the dataset from the online repository up to HDFS storage. The datasets have been combined in a single large one.

- Accessible through REST APIs



GenFlowFile-GenerateLinks
GenerateFlowFile 2.4.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

SplitText-SplitLinks
SplitText 2.4.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name splits
Queued 0 (0 bytes)

ExtractText-ExtractLinks
ExtractText 2.4.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name matched, unmatched
Queued 0 (0 bytes)

InvokeHTTP-DownloadDatasets
InvokeHTTP 2.4.0
org.apache.nifi - nifi-standard-nar

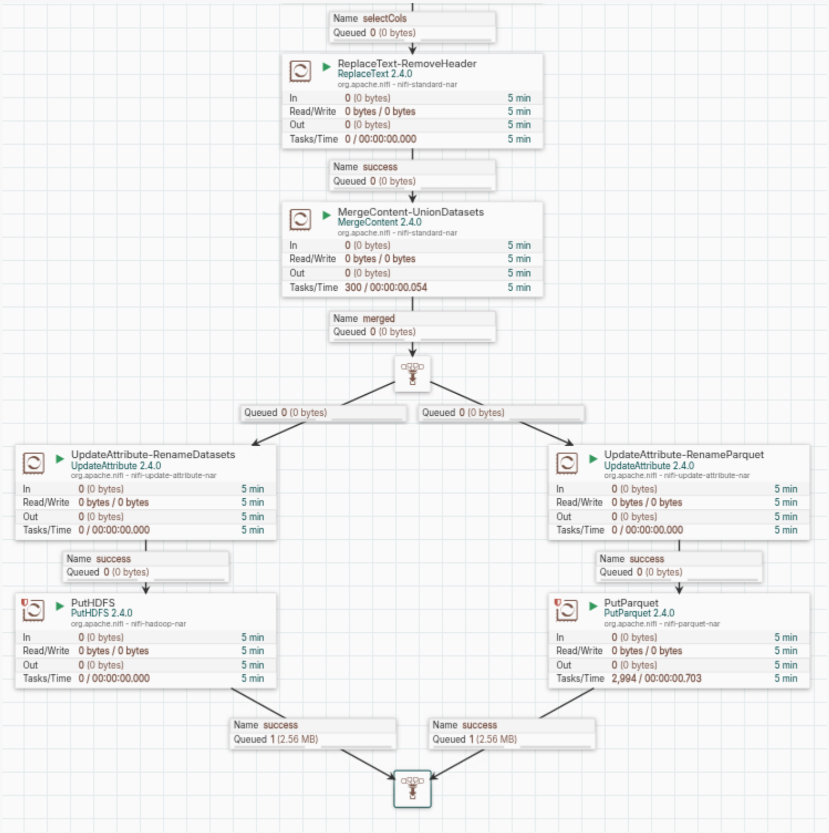
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name Response
Queued 0 (0 bytes)

QueryRecord-FilterCols
QueryRecord 2.4.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name selectCols
Queued 0 (0 bytes)



Name selectCols
Queued 0 (0 bytes)

ReplaceText-RemoveHeader
ReplaceText 2.4.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

MergeContent-UnionDatasets
MergeContent 2.4.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	300 / 00:00:00.054	5 min

Name merged
Queued 0 (0 bytes)

Queued 0 (0 bytes)

Queued 0 (0 bytes)

UpdateAttribute-RenameDatasets
UpdateAttribute 2.4.0
org.apache.nifi - nifi-update-attribute-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

PutHDFS
PutHDFS 2.4.0
org.apache.nifi - nifi-hadoop-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 1 (2.56 MB)

UpdateAttribute-RenameParquet
UpdateAttribute 2.4.0
org.apache.nifi - nifi-update-attribute-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

PutParquet
PutParquet 2.4.0
org.apache.nifi - nifi-parquet-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	2,994 / 00:00:00.703	5 min

Name success
Queued 1 (2.56 MB)

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

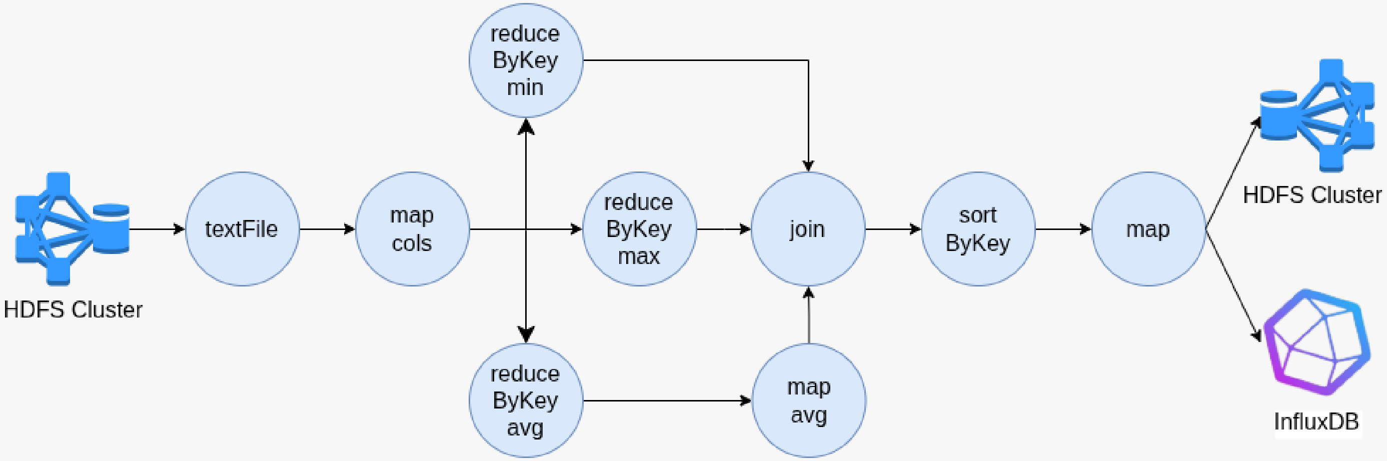
STORAGE AND PERSISTENCE

- Hadoop HDFS Cluster:
  - Number of datanotes configurable
  - Results inside the HDFS directory /data/results
- InfluxDB:
  - Results are time-series data
  - Allows statistical checks before Grafana

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

DATA PROCESSING: QUERY 1

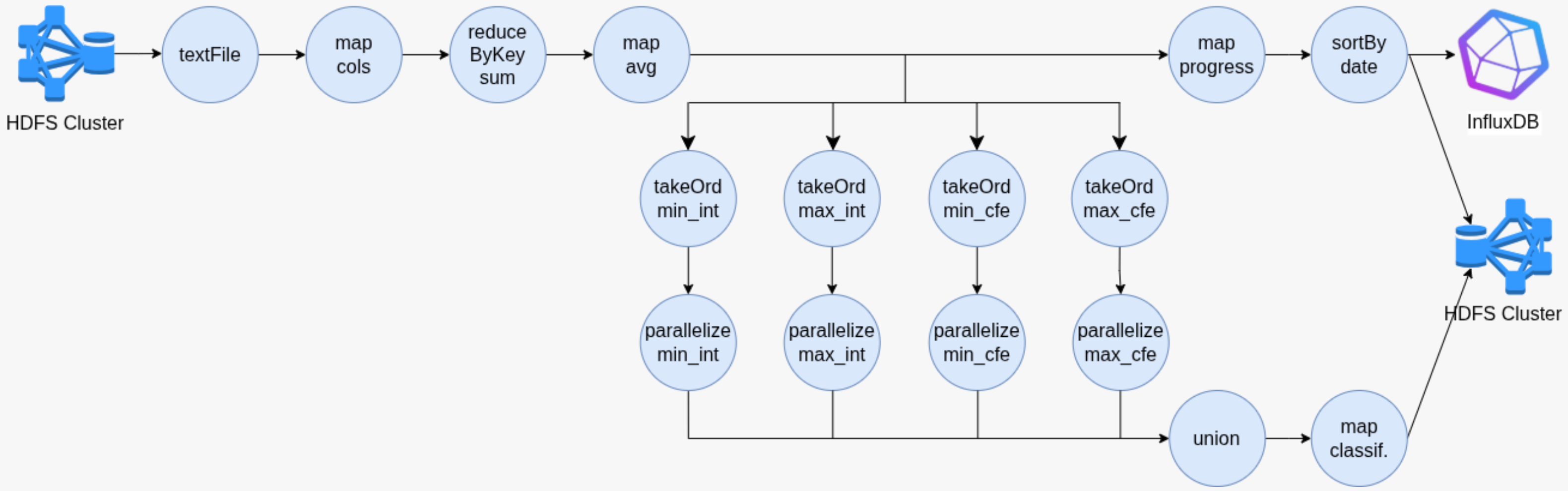


- RDD DAG of query 1
- No caching is needed

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

DATA PROCESSING: QUERY 2

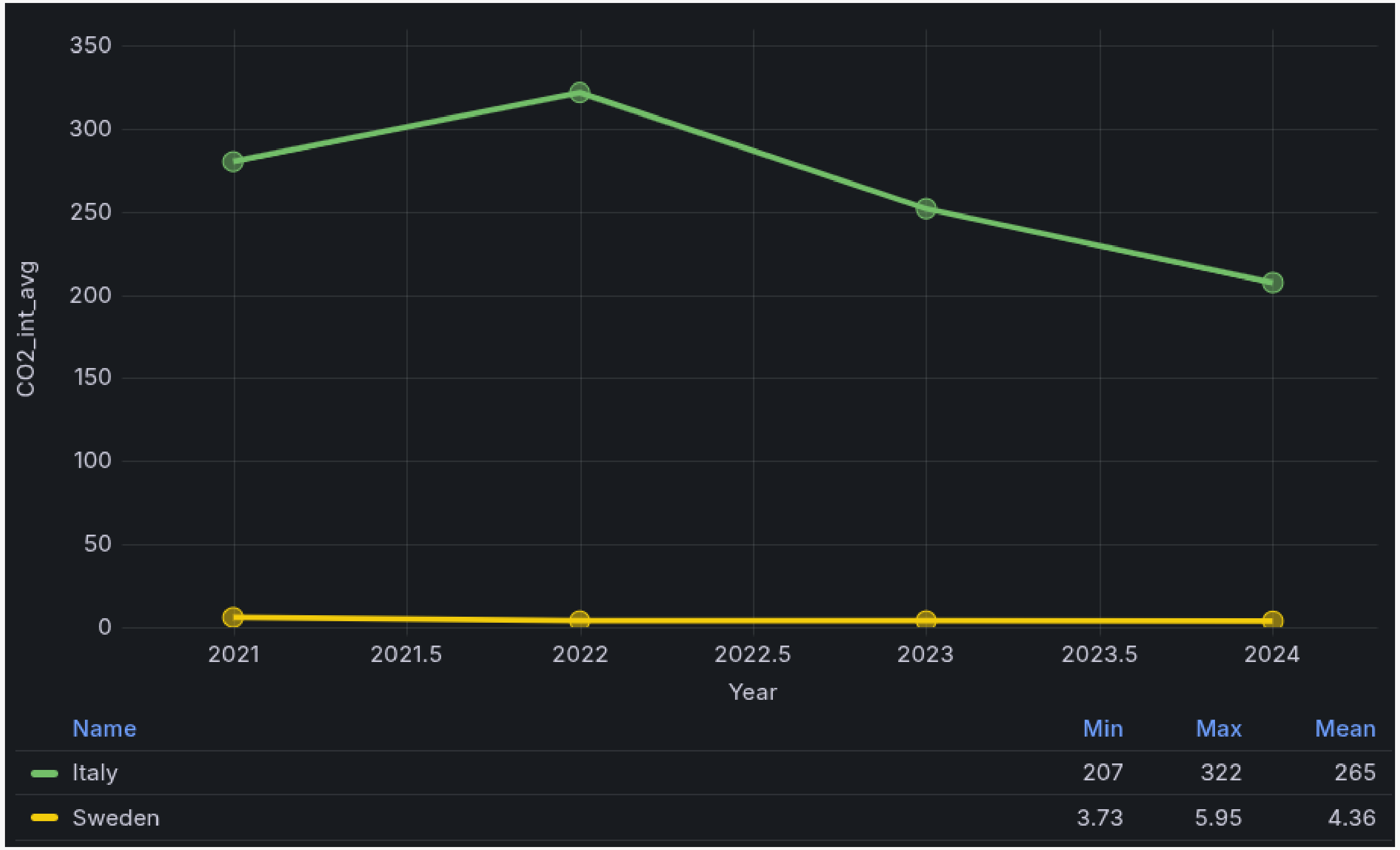


- RDD DAG of query 2
- No caching is needed

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY 1

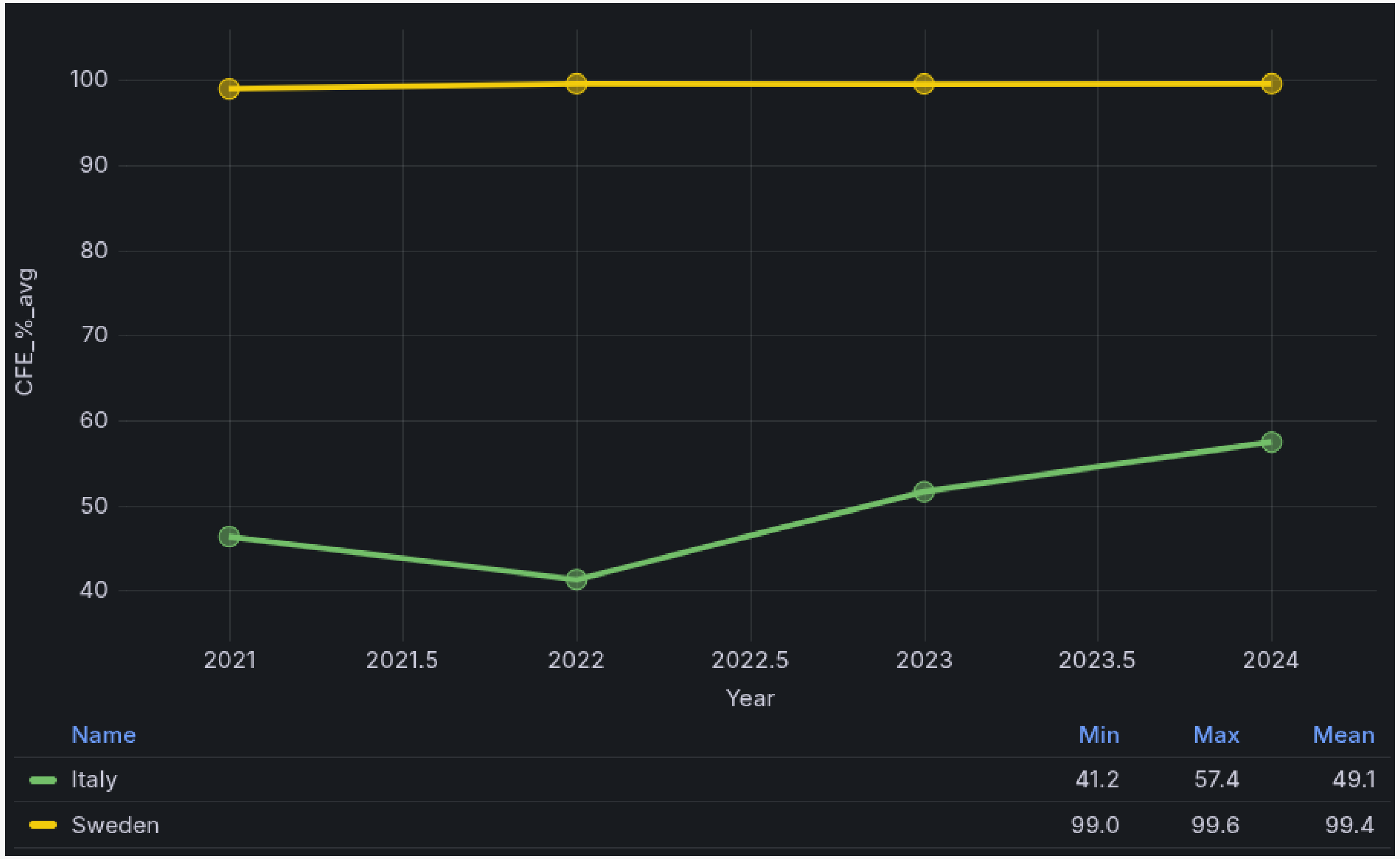


Name	Min	Max	Mean
🟩 Italy	207	322	265
🟨 Sweden	3.73	5.95	4.36

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY 1

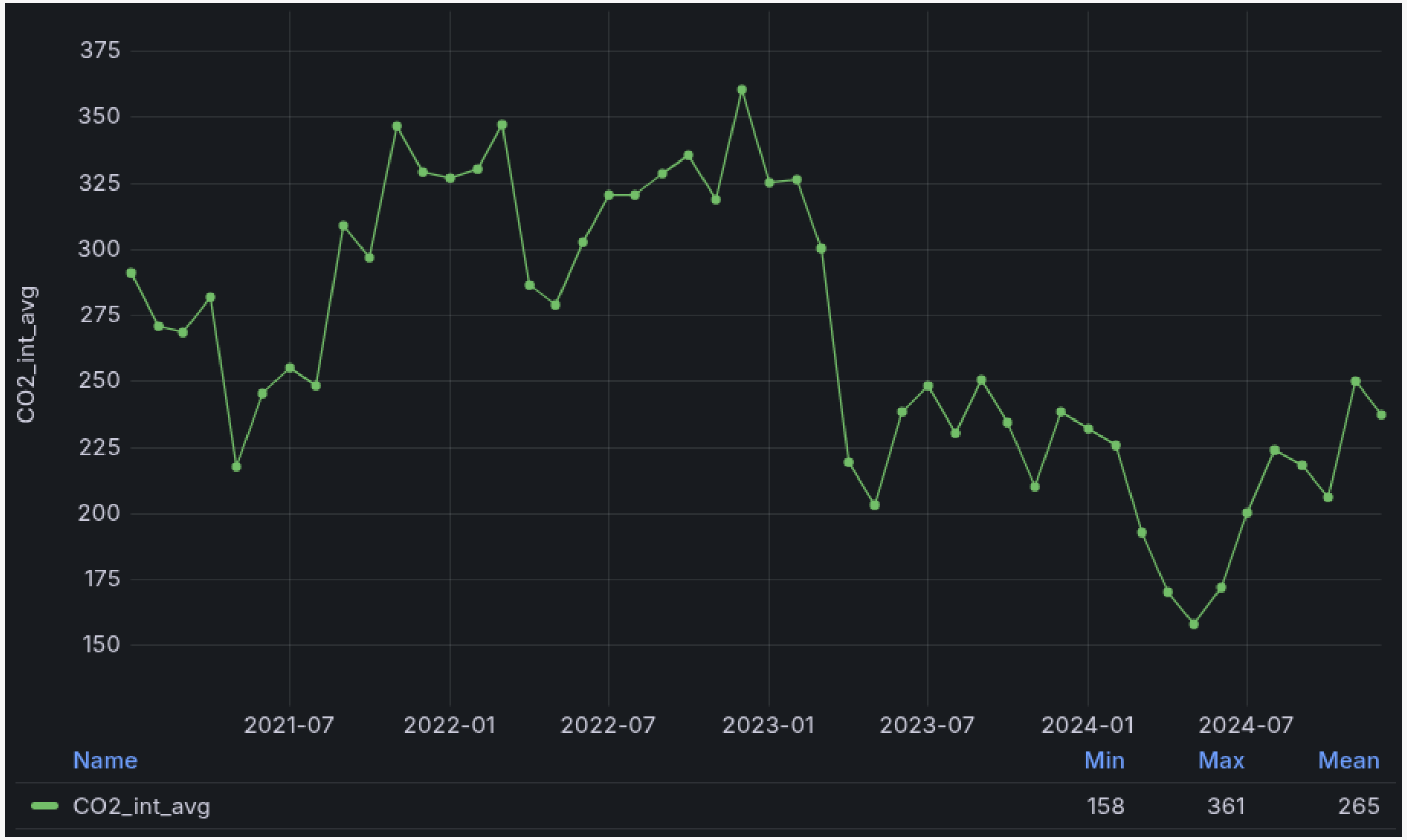


Name	Min	Max	Mean
🟩 Italy	41.2	57.4	49.1
🟨 Sweden	99.0	99.6	99.4

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY 2

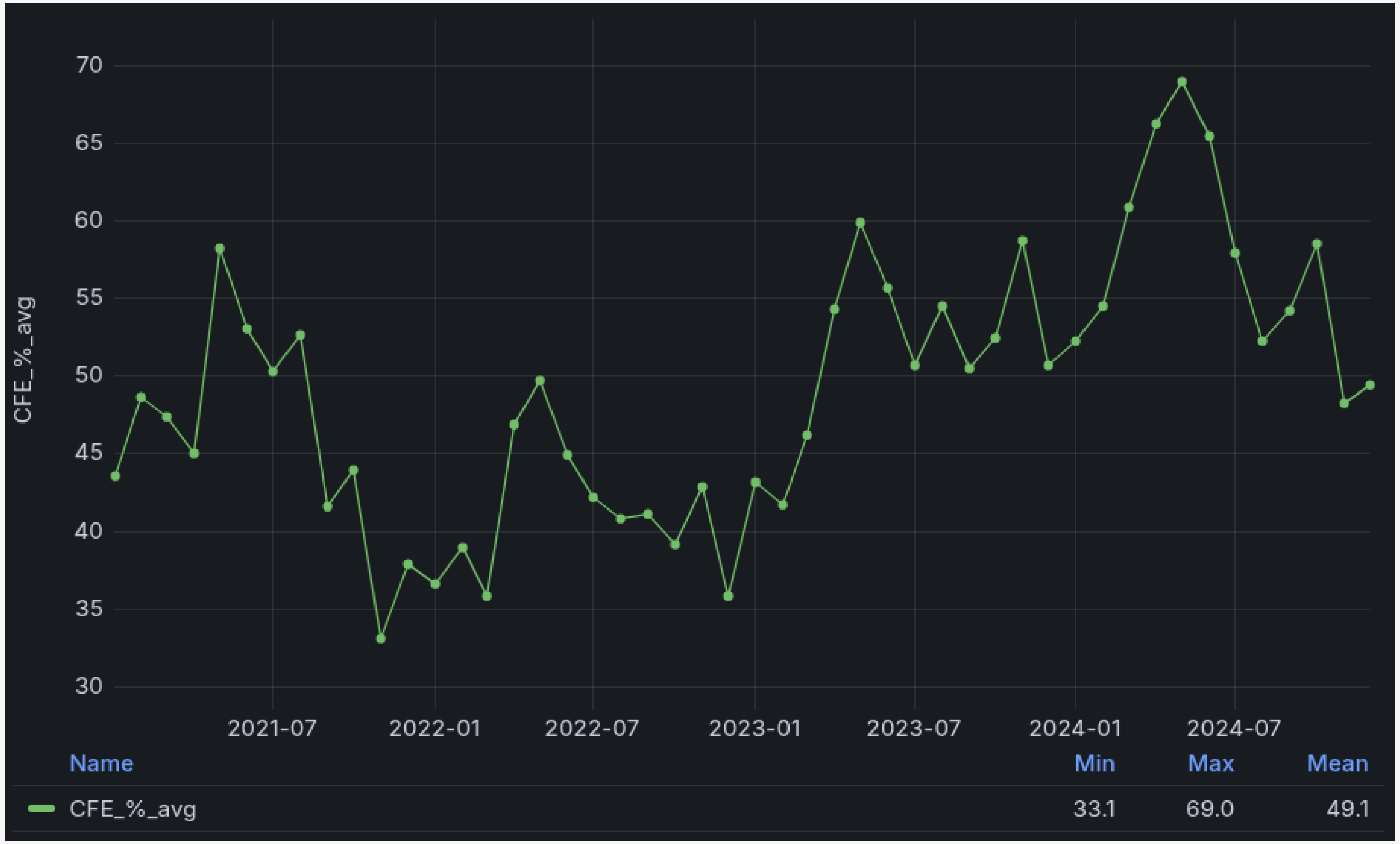


Name	Min	Max	Mean
CO2_int_avg	158	361	265

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY 2

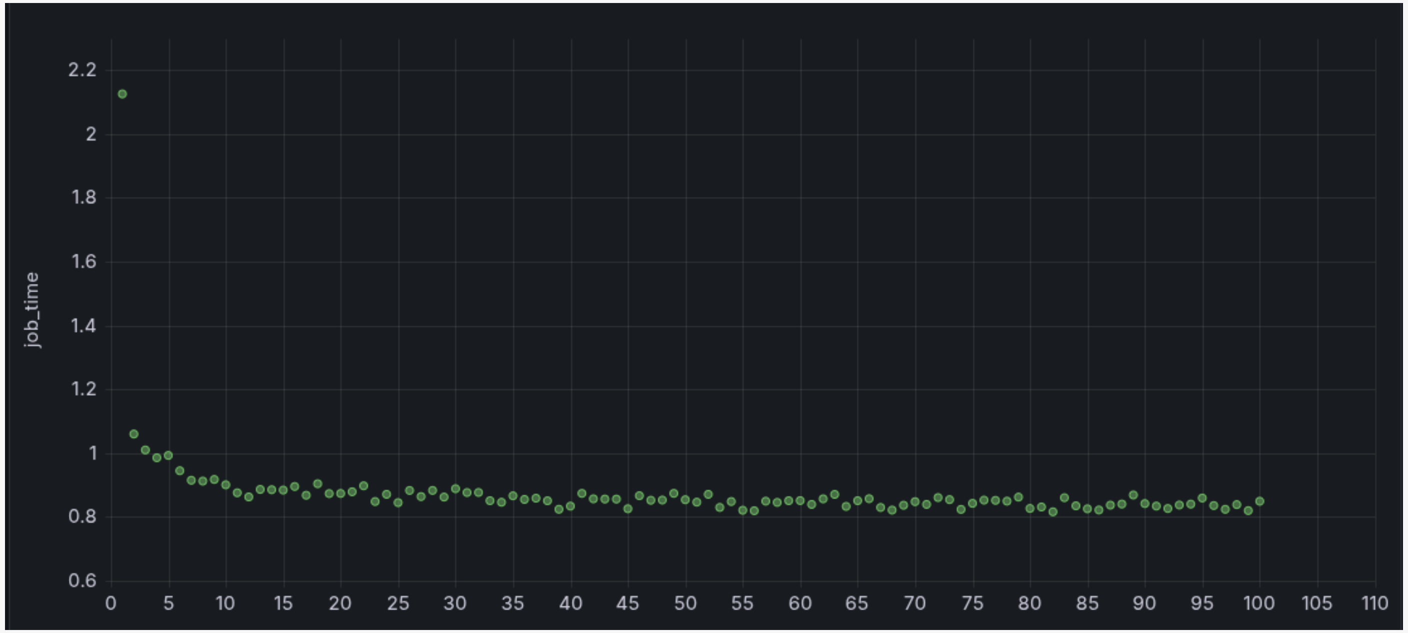


Name	Min	Max	Mean
CFE_%_avg	33.1	69.0	49.1

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY TIMES

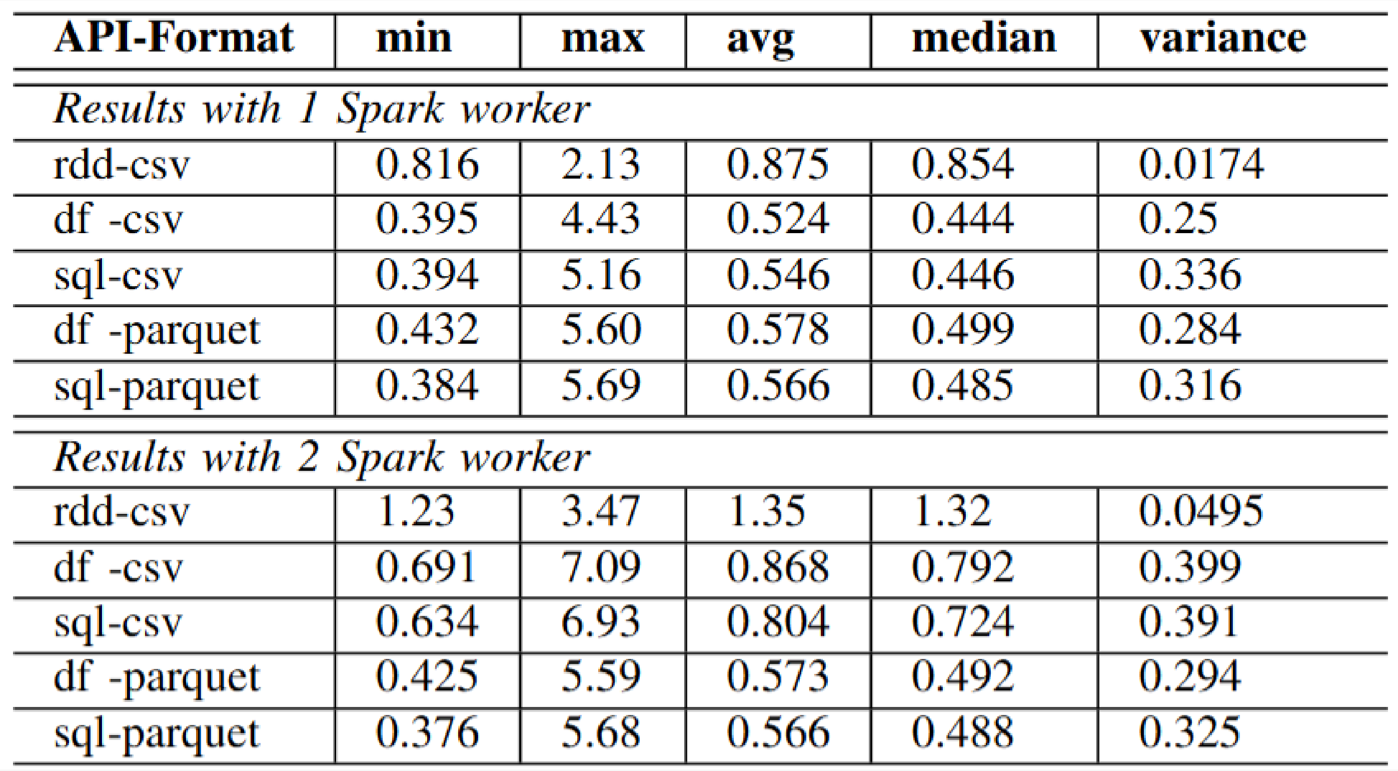


- Typical trend in query execution times (in seconds)
- Spark resource optimization at runtime drastically reduce execution times after the second run

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY 1

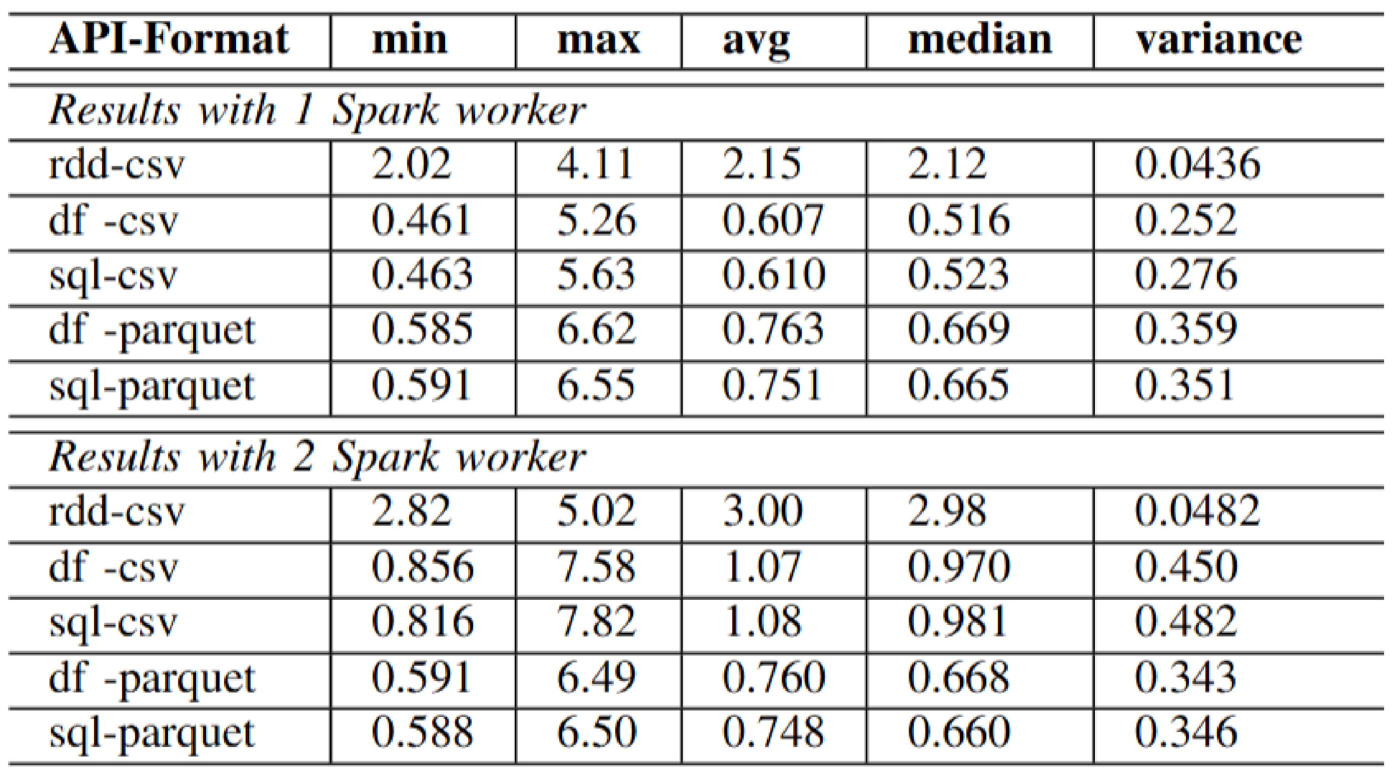


API-Format	min	max	avg	median	variance
Results with 1 Spark worker					
rdd-csv	0.816	2.13	0.875	0.854	0.0174
df -csv	0.395	4.43	0.524	0.444	0.25
sql-csv	0.394	5.16	0.546	0.446	0.336
df -parquet	0.432	5.60	0.578	0.499	0.284
sql-parquet	0.384	5.69	0.566	0.485	0.316
Results with 2 Spark worker					
rdd-csv	1.23	3.47	1.35	1.32	0.0495
df -csv	0.691	7.09	0.868	0.792	0.399
sql-csv	0.634	6.93	0.804	0.724	0.391
df -parquet	0.425	5.59	0.573	0.492	0.294
sql-parquet	0.376	5.68	0.566	0.488	0.325

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

RESULTS: QUERY 2



API-Format	min	max	avg	median	variance
Results with 1 Spark worker					
rdd-csv	2.02	4.11	2.15	2.12	0.0436
df -csv	0.461	5.26	0.607	0.516	0.252
sql-csv	0.463	5.63	0.610	0.523	0.276
df -parquet	0.585	6.62	0.763	0.669	0.359
sql-parquet	0.591	6.55	0.751	0.665	0.351
Results with 2 Spark worker					
rdd-csv	2.82	5.02	3.00	2.98	0.0482
df -csv	0.856	7.58	1.07	0.970	0.450
sql-csv	0.816	7.82	1.08	0.981	0.482
df -parquet	0.591	6.49	0.760	0.668	0.343
sql-parquet	0.588	6.50	0.748	0.660	0.346

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

CONCLUSIONS

Proj.	Query 1	Query 2
Comments	Italy is far behind Sweden in terms of both carbon intensity and carbon-free energy percentage usage.	There appears to be a general downward movement from 2023 onwards in the CO2 intensity trend and a general upward movement from 2023 in the CFE trend, but the overall behaviour seems very unstable, and the global improvement doesn’t look that much significant.
Performances	The maximum execution time is much greater than its mean value: using more Spark workers doesn’t give any consistent advantage, indeed for the csv configuration having 2 Spark workers increases average execution times of about 60%-70%, while in the parquet configuration they are practically the same. Furthermore, DataFrame and SQL APIs show significant increasing of average performance, and even though RDD variance is the lower one, they look like the most preferable	

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

THREATS TO VALIDITY

About query performances:

! **External Validity**: the extent to which one can generalize the findings of a study to other situations, people, settings and measures

- ➢ Larger datasets would certainly lead to different configurations to be preferable

! **Conclusion Validity**: it refers to the correctness of supported conclusions

- ➢ Execution time metric could not be the focus in other contexts
- ➢ Dataset could be too small to reproduce a real case scenario

! **Reliability**: it refers to the consistency and repeatability of the measures

- ➢ Results strongly depend on the machine they are run on
- ➢ Results are not reproducible given the unpredictability of CPUs

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions

END

THANKS FOR YOUR ATTENTION!

Project links:

GitHub page

AGENDA

- Introduction
- Architecture
- Data ingestion
- Storage&Persistence
- Data processing
- Results
- Conclusions