A NETWORK OF PAINTERS

DATA ENGINEERING I. – TERM 2

CEU CENTRAL EUROPEAN UNIVERSITY

Alina Kurmantayeva
Gréta Zsikla
Mihály Hanics
Péter Török

06 DECEMBER 2024

Agenda

1. Executive Summary

Summary of project scope, data source and aim of the analyses

2. KNIME Workflow

Overview of the KNIME workflow implementation

3. Neo4J Operation

Outlining the network analytics carried out in Neo4j and KNIME

4. Analytics Results

Overview of key analytics results and network visualization

Agenda

1. Executive Summary

Summary of project scope, data source and aim of the analyses

2. KNIME Workflow

Overview of the KNIME workflow implementation

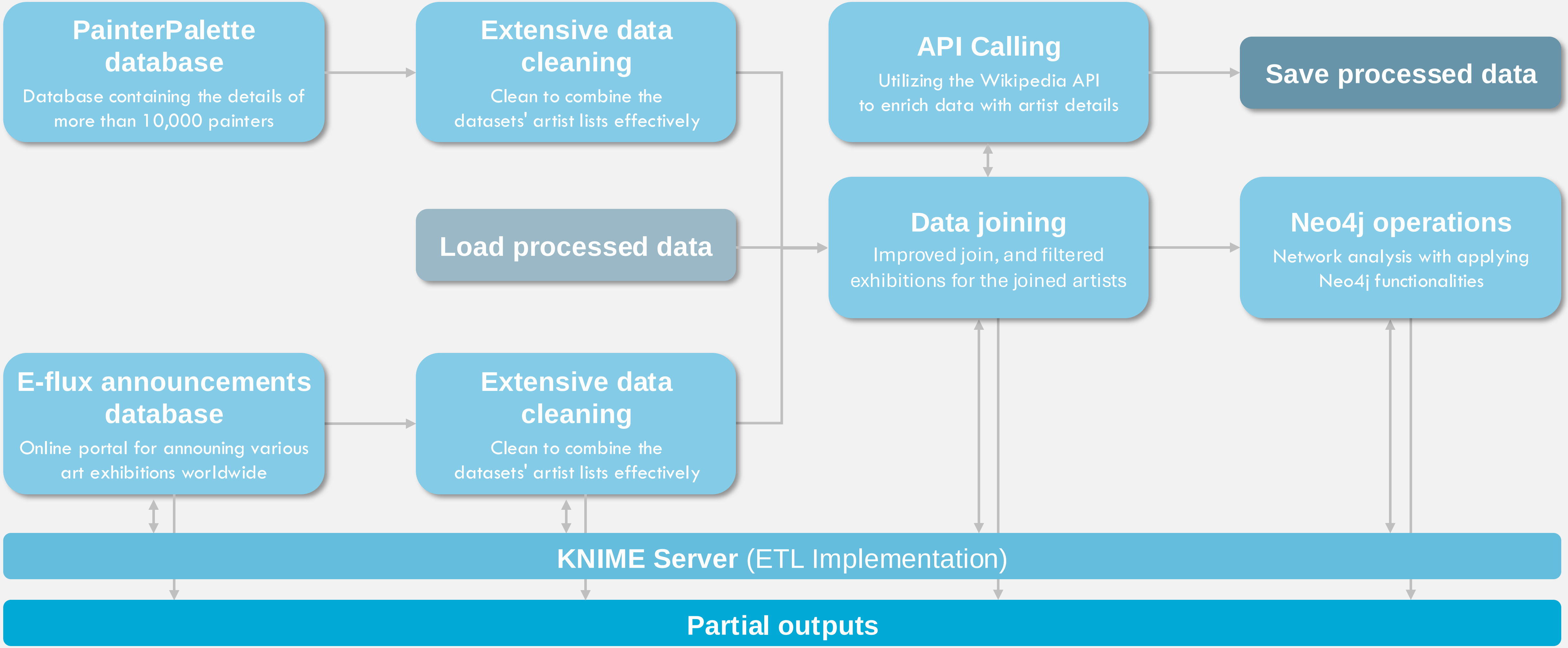
3. Neo4J Operation

Outlining the network analytics carried out in Neo4j and KNIME

4. Analytics Results

Overview of key analytics results and network visualization

Our project work utilized 2 databased, an API and NoSQL solutions to analyse the network of painters in depth



PainterPalette database

Database containing the details of more than 10,000 painters

Extensive data cleaning

Clean to combine the datasets' artist lists effectively

API Calling

Utilizing the Wikipedia API to enrich data with artist details

Save processed data

Load processed data

Data joining

Improved join, and filtered exhibitions for the joined artists

Neo4j operations

Network analysis with applying Neo4j functionalities

E-flux announcements database

Online portal for announing various art exhibitions worldwide

Extensive data cleaning

Clean to combine the datasets' artist lists effectively

KNIME Server (ETL Implementation)

Partial outputs

Agenda

1. Executive Summary

Summary of project scope, data source and aim of the analyses

2. KNIME Workflow

Overview of the KNIME workflow implementation

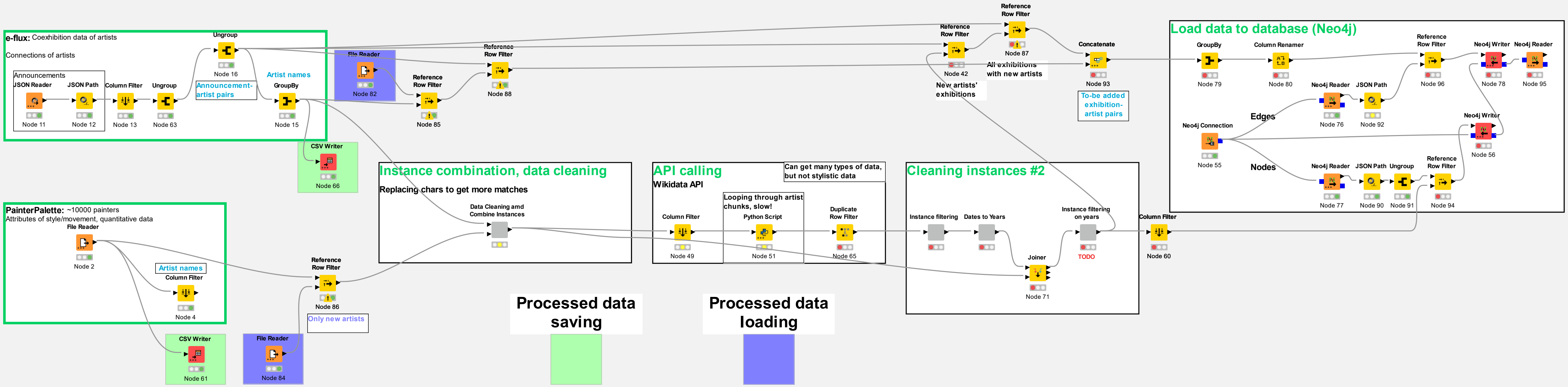
3. Neo4J Operation

Outlining the network analytics carried out in Neo4j and KNIME

4. Analytics Results

Overview of key analytics results and network visualization

Our project work relies on a complex KNIME workflow that reads, cleans, aggregates and analyses data



Data process processing layer

Data cleaning

API Calling

Data cleaning #2

Load to Neo4j DB

Analytics

Each layer of the KNIME workflow has a clearly defined function, intricately interconnected to efficiency

	Data reading & import	Data cleaning & join	API implementation	Neo4j & Analytics
Function	Reading and processing of the PainterPalette and Influx announcements containing artists and exhibitions	Extensive cleaning and standardization of names to implement the joining of the databases more effectively, resulting more matches	Application of API calls to enrich data with further details of artists (e.g., date and place of birth, citizenship etc.)	Connecting the KNIME workflow with Neo4j: load data into database, create and analyze network of artists
Output	Two databases with non-standardized record names and missing values	Joint database of artists and the recent exhibitions with standardized names	Joint database with matching names and enriched data	Analytical insights and visualization of the network
Methods	• File reader • GroupBy • JSON reader • Column filter	• GroupBy • String manipulation	• SparQL querying • Python script (knio.Table) • Retries on failures	• Neo4j reader • Neo4j connector

Each layer of the KNIME workflow has a clearly defined function, intricately interconnected to efficiency

API implementation

Function

Application of API calls to enrich data with further details of artists (e.g., date and place of birth, citizenship etc.)

Output

Joint database with matching names and enriched data

Methods

- SparQL querying
- Python script (knio.Table)
- Retries on failures

```
SELECT ?person ?personLabel ?placeOfBirthLabel ?dateOfBirth...
WHERE { VALUES ?personLabel { {people_string} }
        ?person ?label ?personLabel.
        ?person wdt:P31 wd:Q5.
        ?person wdt:P19 ?placeOfBirth.
        ?person wdt:P569 ?dateOfBirth.
        ...
        SERVICE wikibase:label { bd:serviceParam wikibase:language "en". } }
```

- SparQL: Properties, values (identifiers), string labels
- Query first paralel, retry missing instances, then retry for all languages

Each layer of the KNIME workflow has a clearly defined function, intricately interconnected to efficiency

	Data reading & import	Data cleaning & join	API implementation	Neo4j & Analytics
Function	Reading and processing of the PainterPalette and Influx announcements containing artists and exhibitions	Extensive cleaning and standardization of names to implement the joining of the databases more effectively, resulting more matches	Application of API calls to enrich data with further details of artists (e.g., date and place of birth, citizenship etc.)	Connecting the KNIME workflow with Neo4j: load data into database, create and analyze network of artists
Output	Two databases with non-standardized record names and missing values	Joint database of artists and the recent exhibitions with standardized names	Joint database with matching names and enriched data	Analytical insights and visualization of the network
Methods	• File reader • GroupBy • JSON reader • Column filter	• GroupBy • String manipulation	• SparQL querying • Python script (knio.Table) • Retries on failures	• Neo4j reader • Neo4j connector

Agenda

1. Executive Summary

Summary of project scope, data source and aim of the analyses

2. KNIME Workflow

Overview of the KNIME workflow implementation

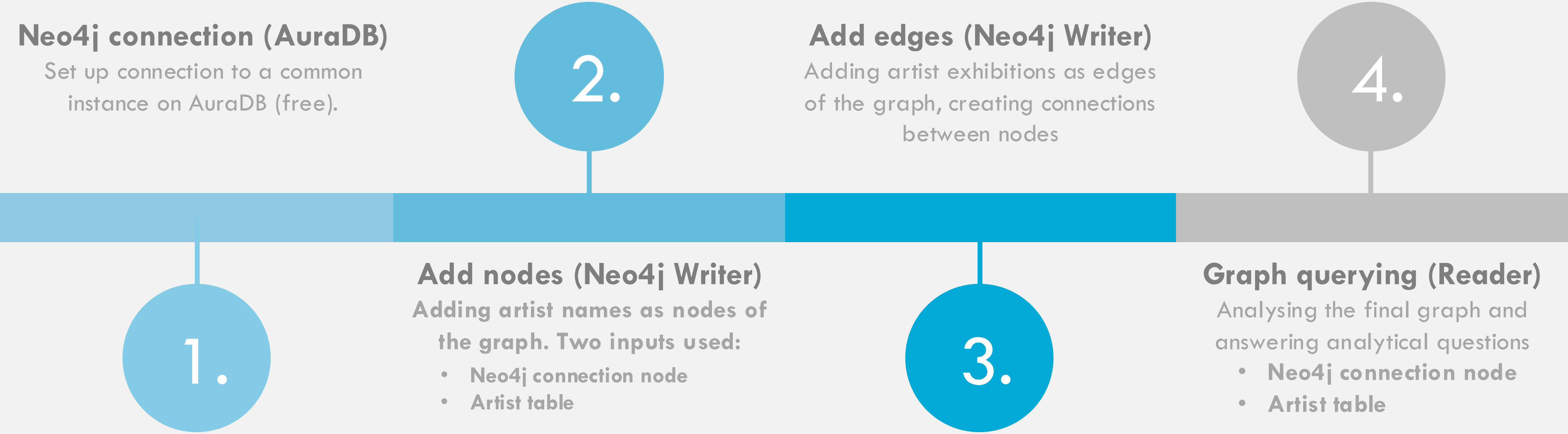
3. Neo4j Operation

Outlining the network analytics carried out in Neo4j and KNIME

4. Analytics Results

Overview of key analytics results and network visualization

Our Neo4j integration is running on a server and uses the standardized datasets as inputs to create a network graph



1.

Neo4j connection (AuraDB)

Set up connection to a common instance on AuraDB (free).

2.

Add nodes (Neo4j Writer)

Adding artist names as nodes of the graph. Two inputs used:

- Neo4j connection node
- Artist table

3.

Add edges (Neo4j Writer)

Adding artist exhibitions as edges of the graph, creating connections between nodes

4.

Graph querying (Reader)

Analysing the final graph and answering analytical questions

- Neo4j connection node
- Artist table

Agenda

1. Executive Summary

Summary of project scope, data source and aim of the analyses

2. KNIME Workflow

Overview of the KNIME workflow implementation

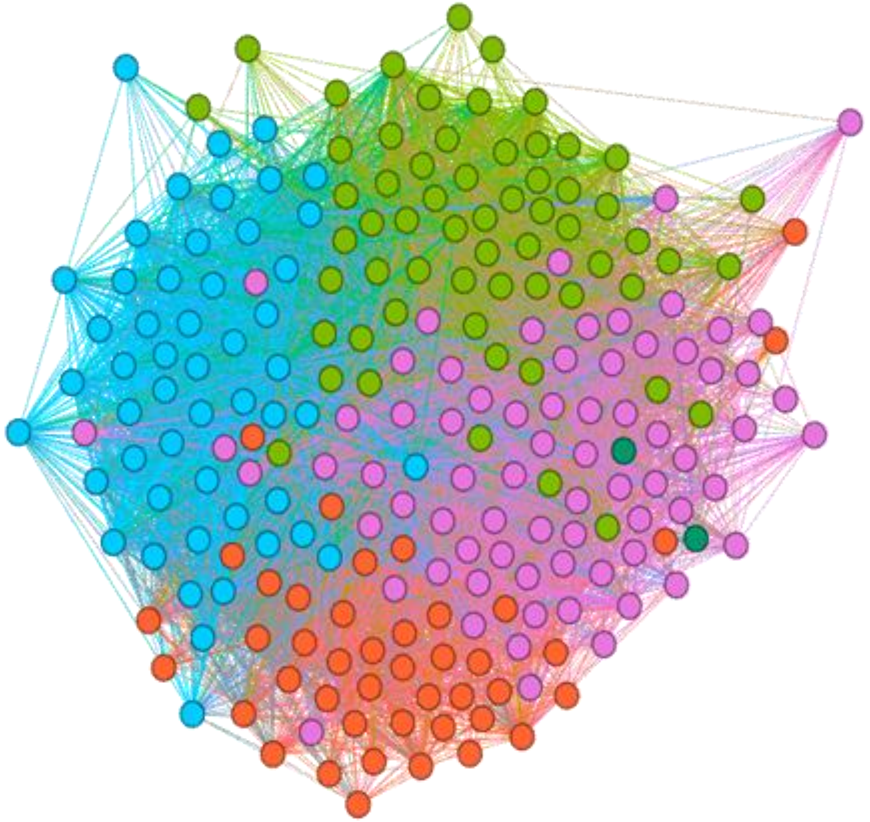
3. Neo4j Operation

Outlining the network analytics carried out in Neo4j and KNIME

4. Analytics Results

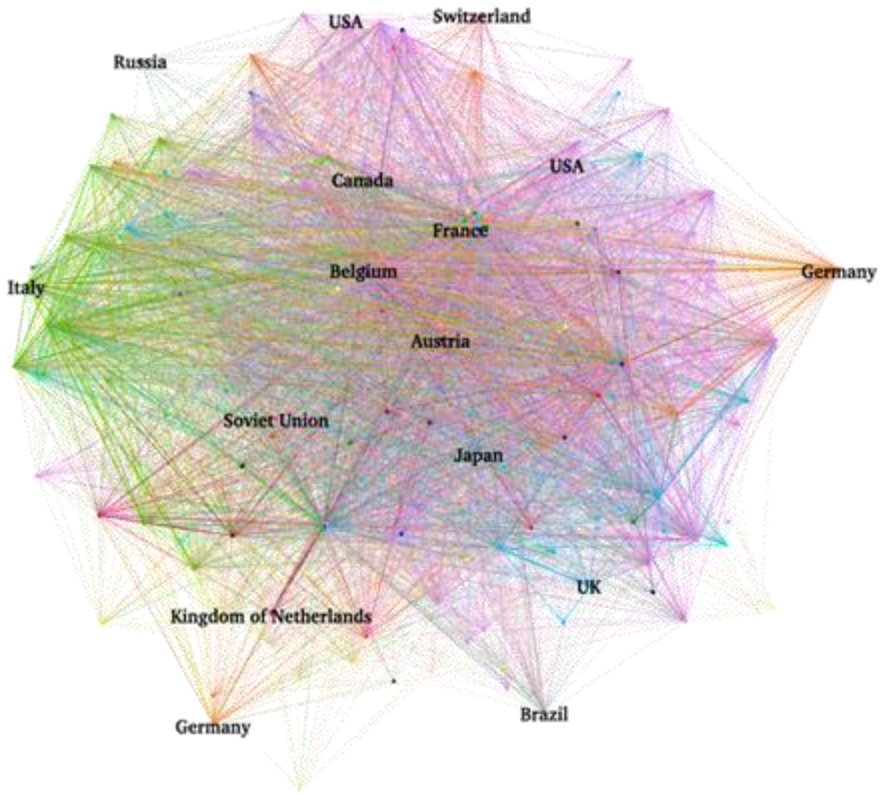
Overview of key analytics results and network visualization

The network visualization in Gephi shows the major communities and networks of artists in a wholistic way



Network visualization of artists based on modularity clustering

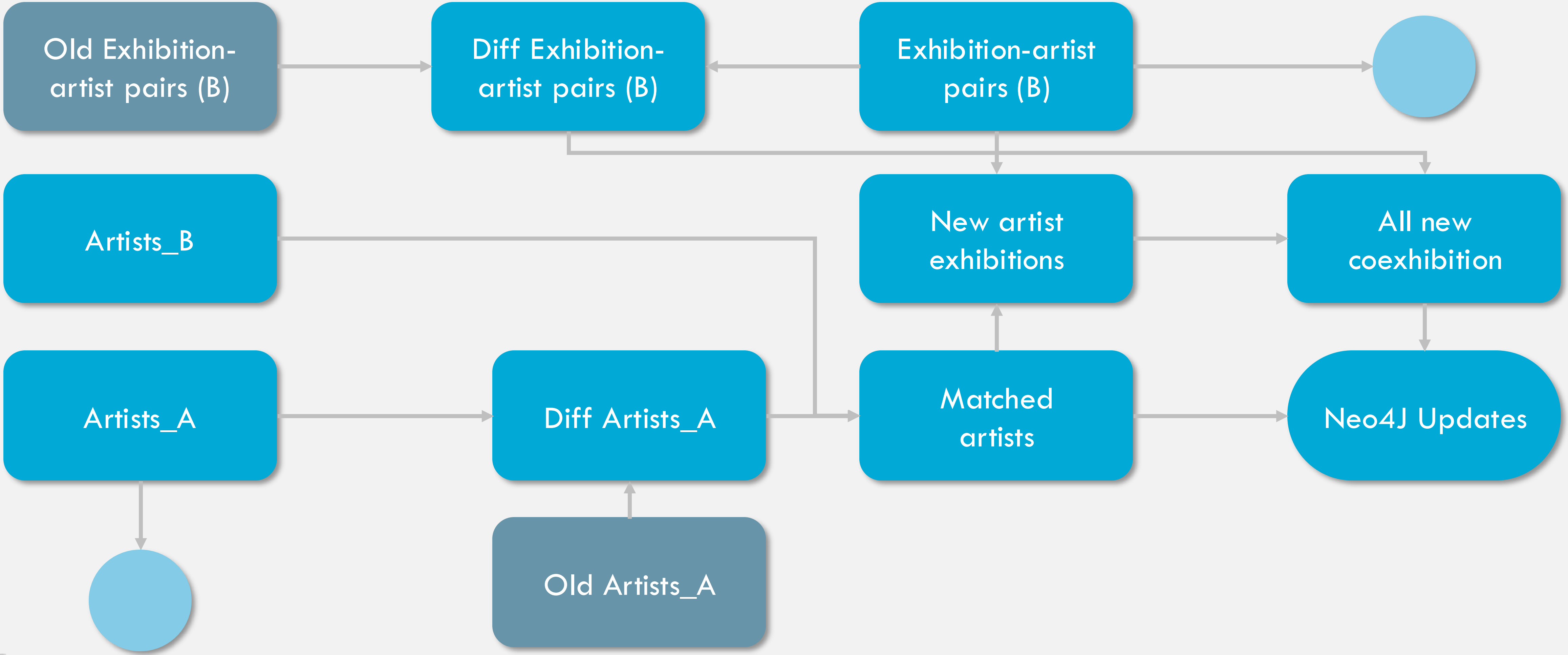
GEPHI



Network visualization of artists based on citizenship centrality

CENTRAL EUROPEAN UNIVERSITY

The KNIME workflow only processes new information and thus it fulfills ETL criterias



CEU CENTRAL EUROPEAN UNIVERSITY

● - Partial outputs

Thank you
for your attention!

Q&A

CEU CENTRAL EUROPEAN UNIVERSITY