Graphical Data Analytic Workflows and Cross-Platform Optimization

David Arnu
RapidMiner

INFORE
Interactive Extreme-Scale Analytics and Forecasting

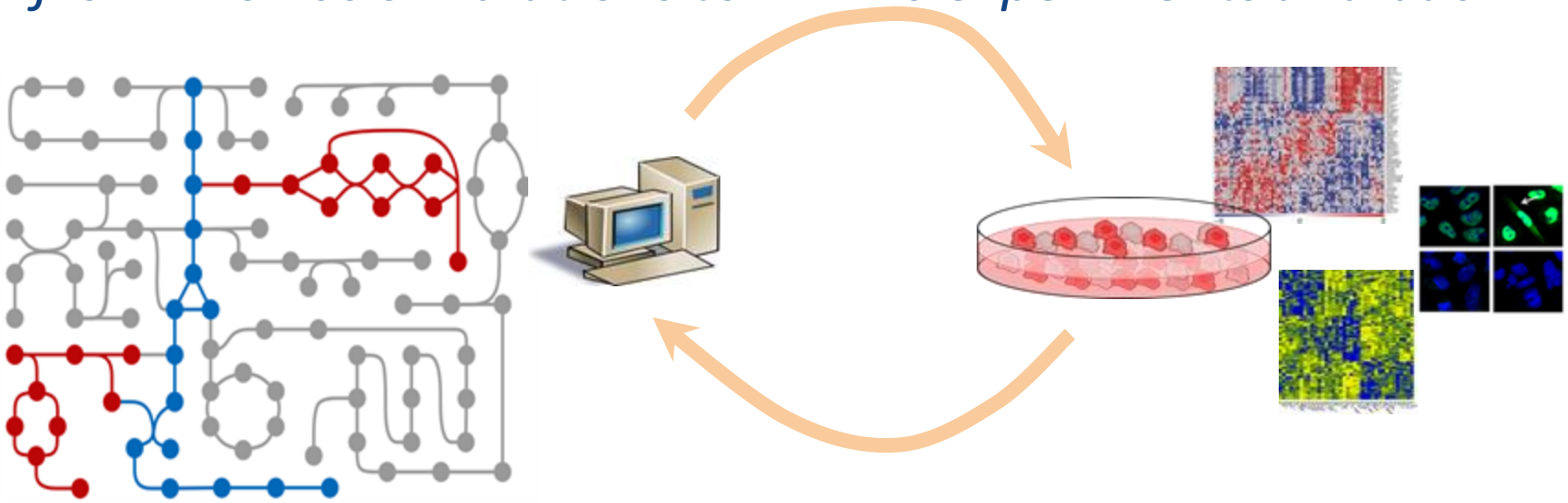
Project Overview

- Graphically design data processing workflows and data analytics tasks with minimal or no programming overhead
- Real-time, interactive machine learning and data mining tools
- Distributed Complex Event Forecasting

Life Sciences Use Case

Studying the effect of drug synergies in cancer

from in-silico simulations to in-vivo experiments and back



- Challenges:
  - Huge CPU, memory requirements + output data to be processed
  - Too many simulations, too few promising ones
  - Train a ML model to classify promising simulations and kill non-promising ones
  - Learn which genes drive evolution of other genes and which to monitor

Financial Use Case

Predicting Price Swings, Systemic Risk and Forecasting Investment Opportunities

Goal: Train ML models that extract valid rules used to perform:

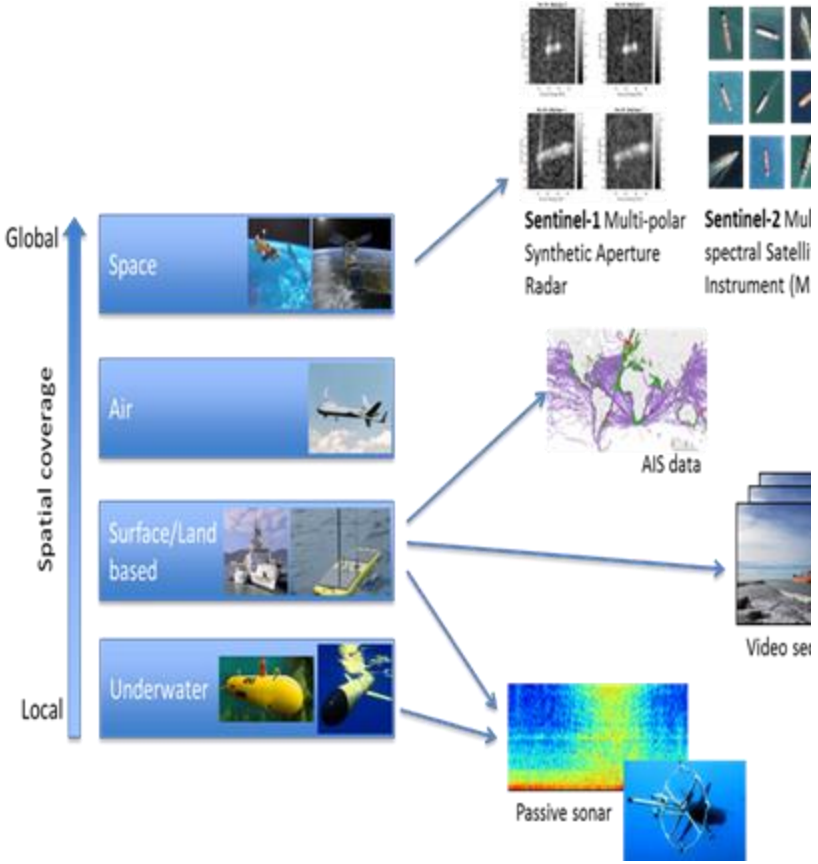
- Real-time suggestion and forecast of investment opportunities
- Systemic risk (i.e., great linkage between major market participants) prediction
- Forecast price swings

Maritime Use Case

Maritime Situational Awareness (MSA), Monitoring Ship Movement and Detecting Illegal Activities

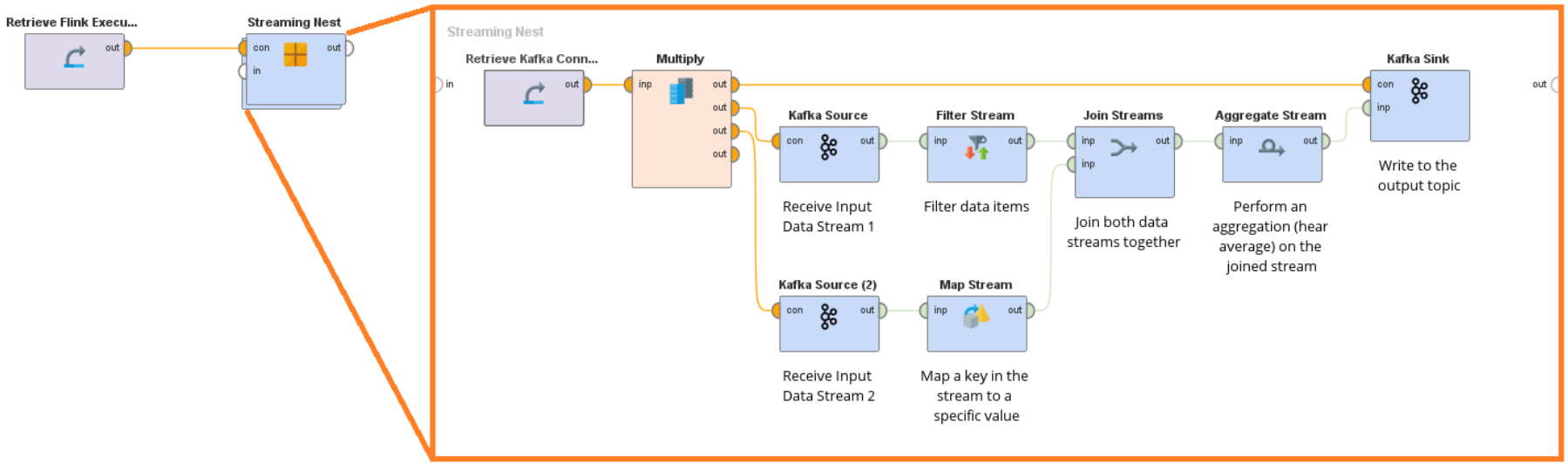
Challenges:

- Large amount of ships to monitor
- Many different data sources are available
- Complex event classification (patterns of movement or other behavior)



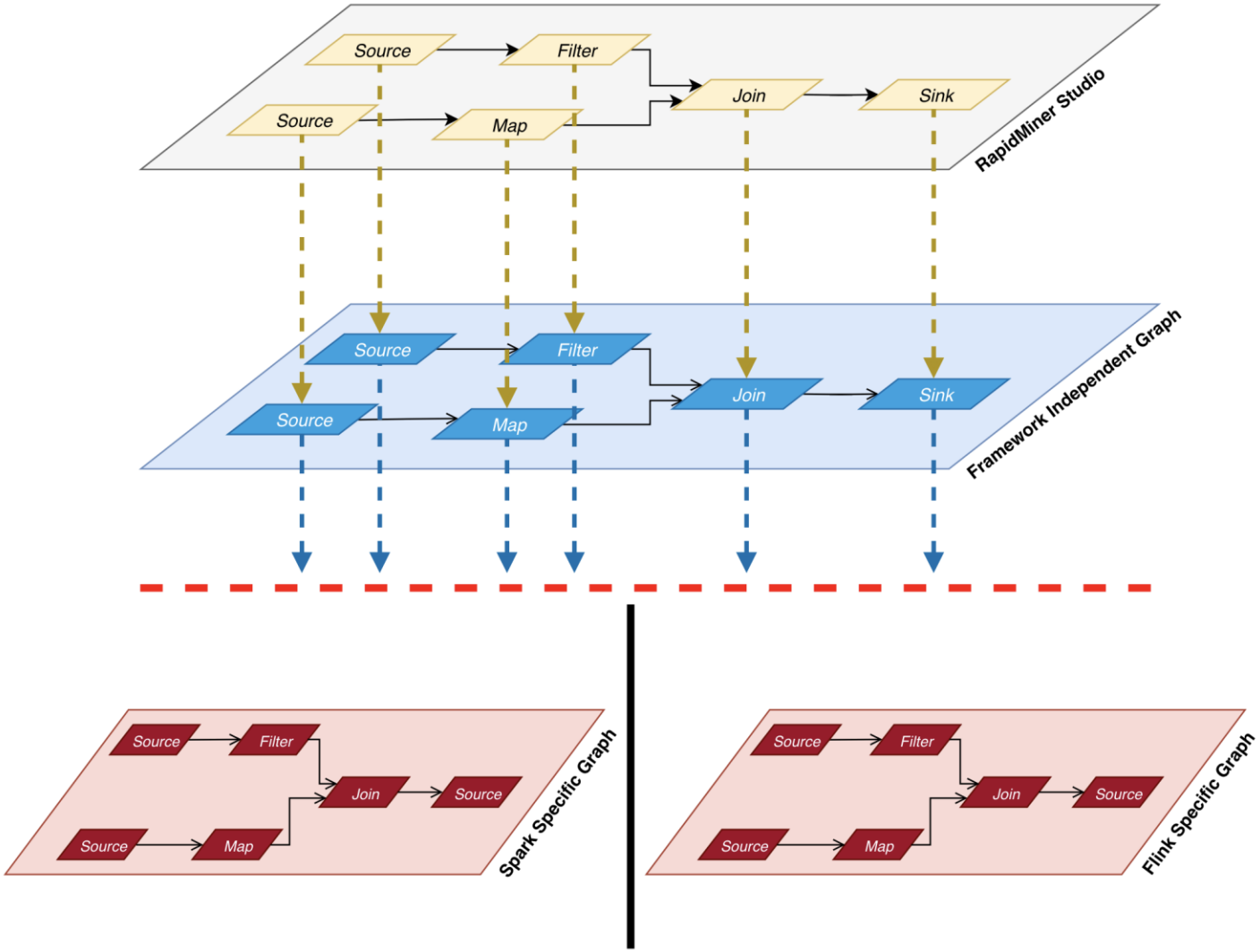
Graphical Data Analytic Workflows

Graphical Editor



- Simply design Streaming Analytics Workflows
- Upon execution **one** job is created and deployed to the connected streaming cluster

Inner Workings



Capabilities

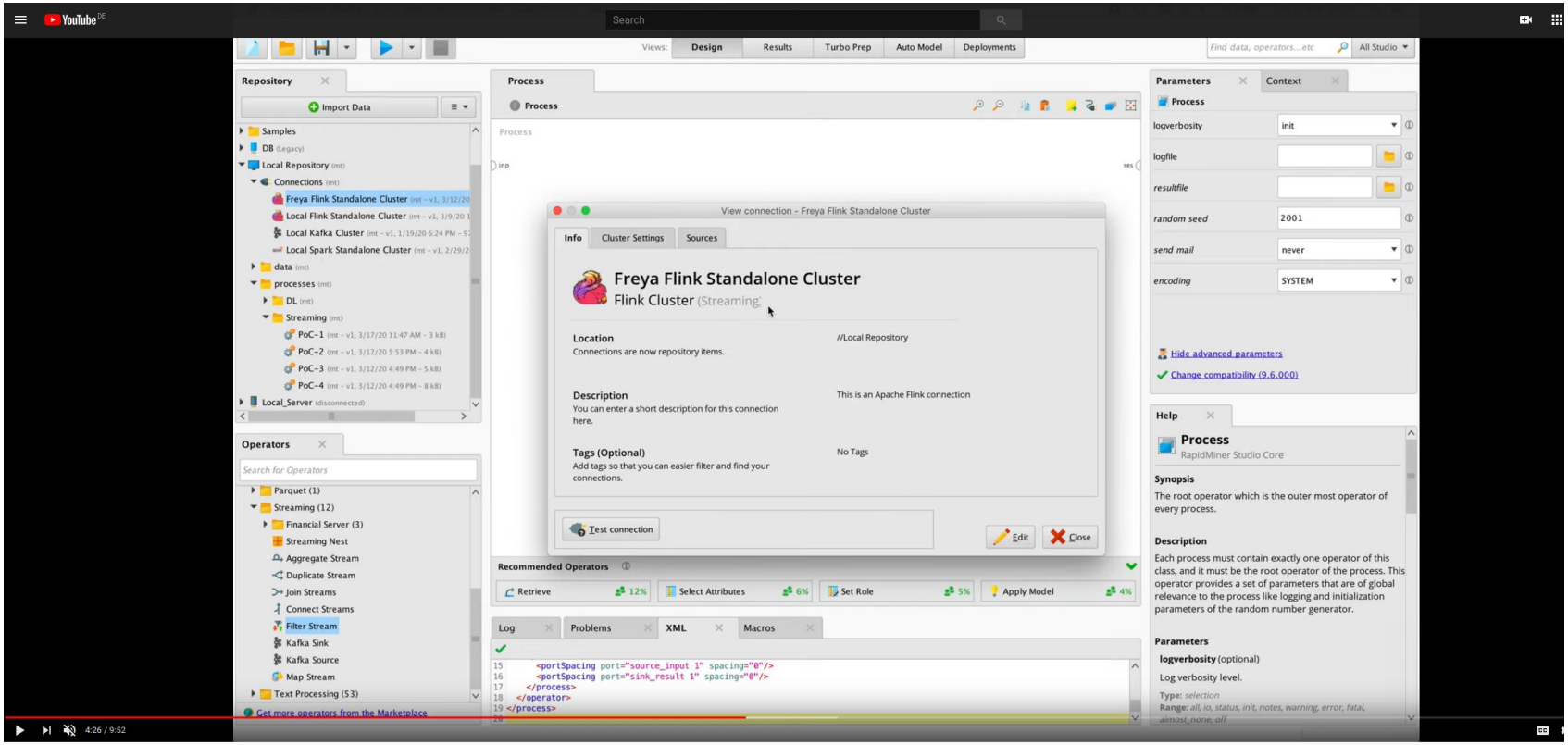
- Supported Clusters:
  - Apache Flink
  - Apache Spark (structured) Streaming
- Available Operations
  - Streaming analytics operations
  - Synopsis Data Engine
  - Custom Online Machine Learning engines (running on Flink and AKKA)
  - Connections to financial service providers

Benefits

- Code-Free development
- Platform and back-end independent
- Pluggable connection management
- Easy to share and collaborate

Link to Demo Video

https://youtu.be/9SKcM70Bi2U



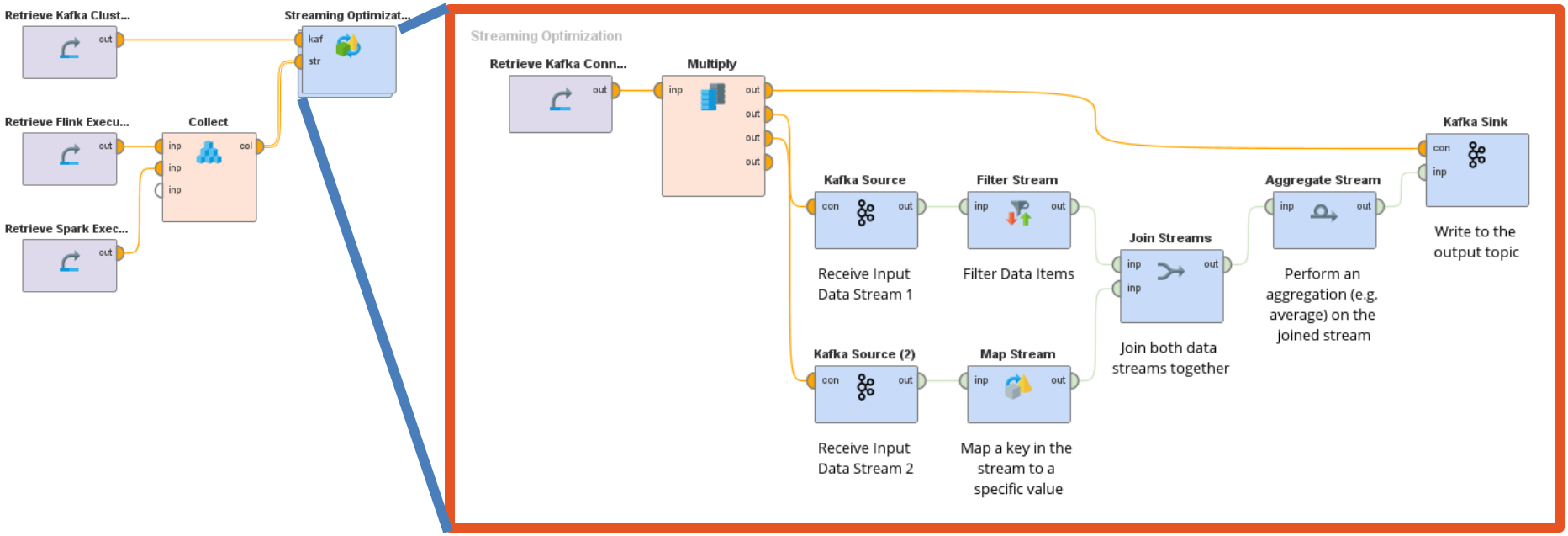
Cross-Platform Optimization

Optimizer Component

- In a multi-cluster set-up the optimal stream execution can depend on
  - Available resources per cluster
  - Data location
  - Software performance and implementation details

- An optimized process layout can greatly enhance the performance

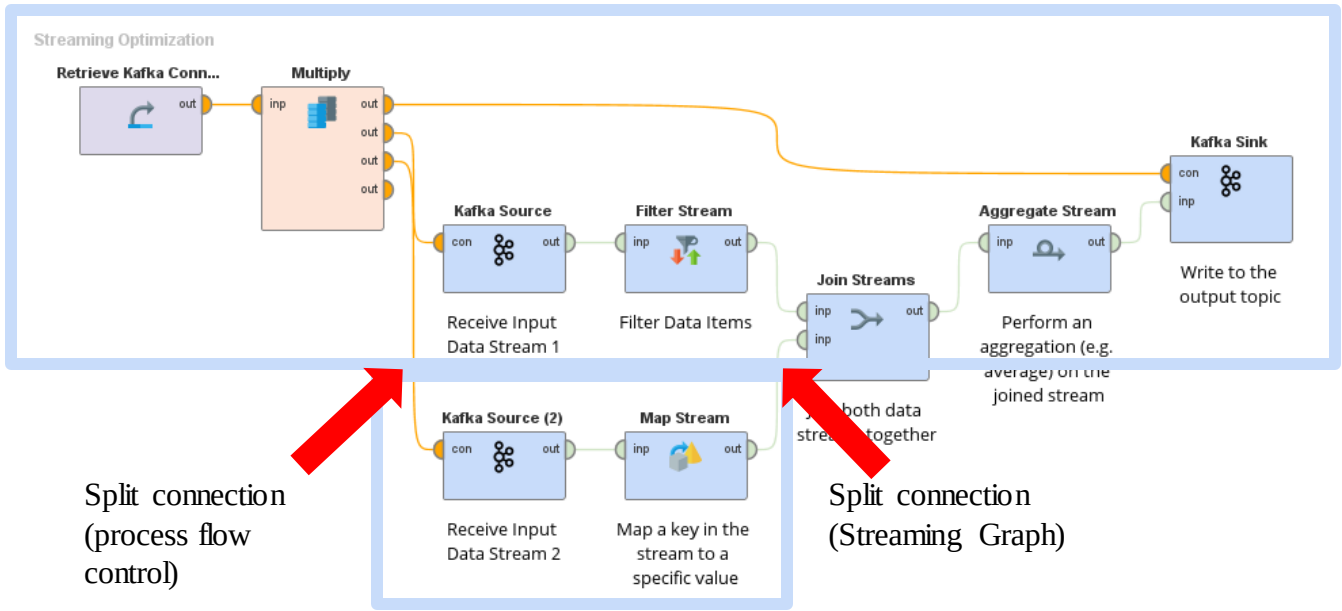
Streaming Optimization Operator



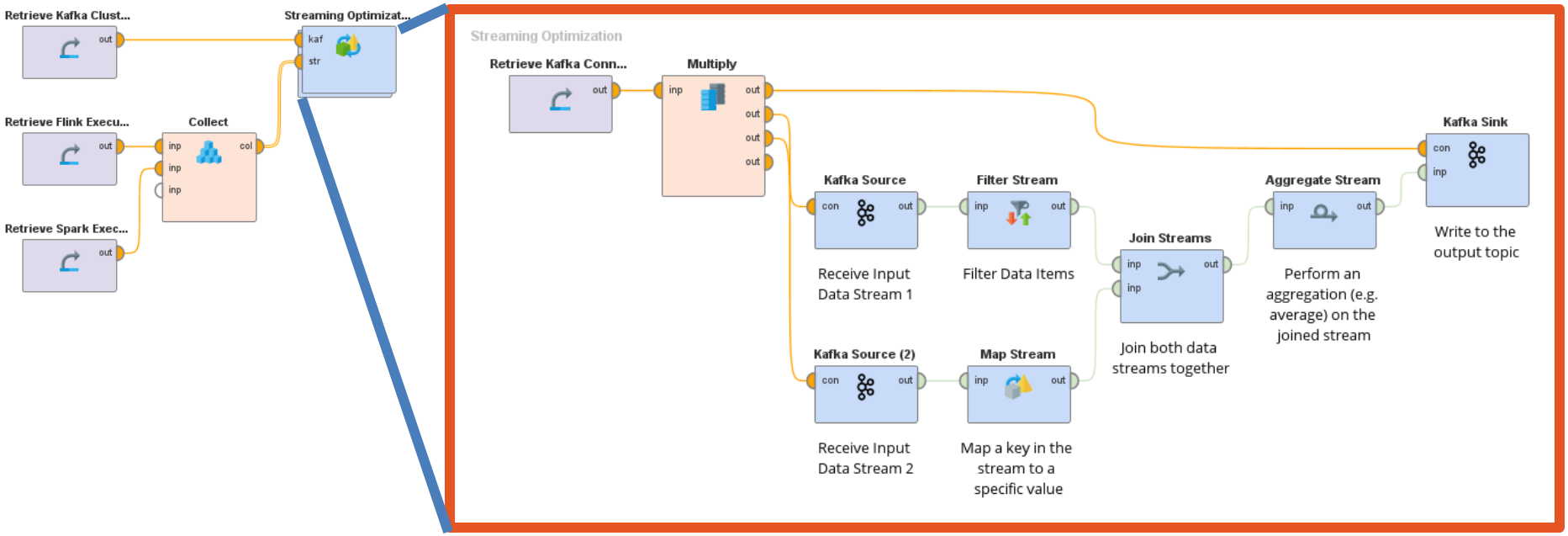
Optimizer Response

optimizer_response.json

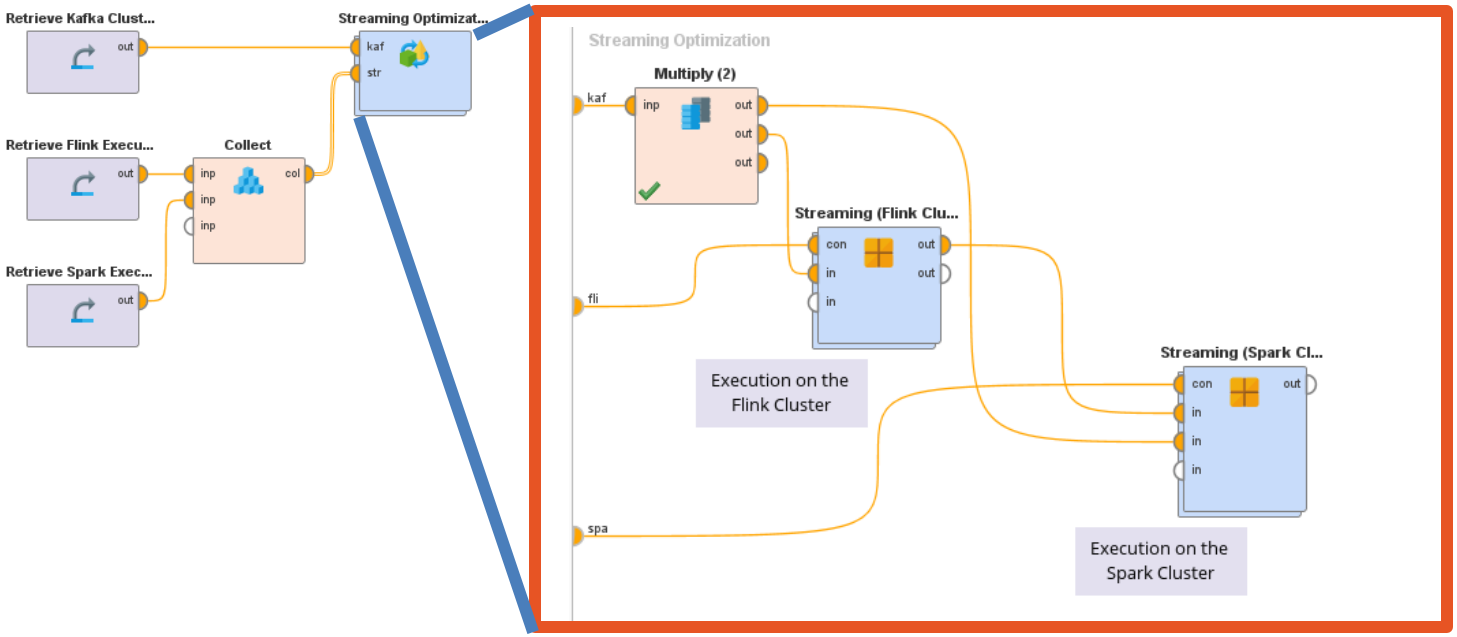
```
{
  "placementSites": [
    {
      "availablePlatforms": [
        {
          "operators": [
            "Retrieve Kafka Connection",
            "Kafka Sink",
            "Aggregate Stream",
            "Join Streams",
            "Kafka Source",
            "Multiply",
            "Filter Stream"
          ],
          "platformName": "flink"
        }
      ],
      "siteName": "flink_barcelona_1"
    },
    {
      "availablePlatforms": [
        {
          "operators": [
            "Kafka Source (2)",
            "Map Stream"
          ],
          "platformName": "spark"
        }
      ],
      "siteName": "spark_barcelona_1"
    }
  ],
  "workflowName": "Streaming"
}
```



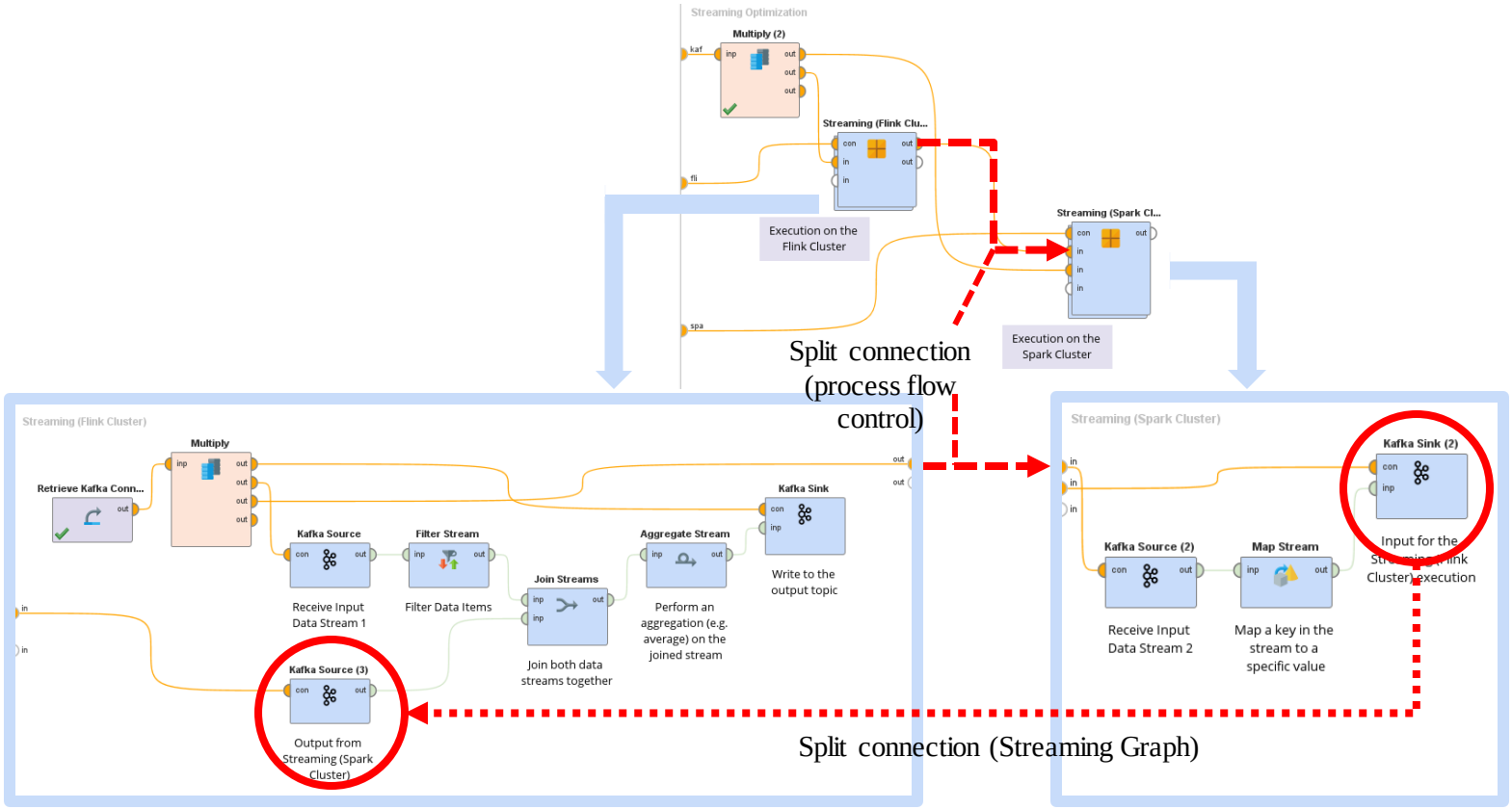
Streaming Optimization Operator



Optimized Workflow



Optimized Workflow



Conclusion

- What we have seen:
  - Project use cases and goals
  - Graphical editor
  - Cross-Platform optimization

- What‘s next:
  - Job and Data Monitoring
  - Better integration of HPC systems
  - Refinements and deployment of the use cases

Thank you!

http://www.infore-project.eu