CoAl : Incremental requirement-driven design and deployment of data intensive flows

Petar Jovanovic, Oscar Romero, Alberto Abelló
Universitat Politècnica de Catalunya, BarcelonaTech
[petar | oromero | aabello]@essi.upc.edu

Alkis Simitsis
HP Labs, Palo Alto, CA, USA
alkis@hp.com

Problem under study

- Incrementally integrating individual data flows into a unified flow satisfying the entire set of inf. requirements
- Maximizing the reuse of the existing data flows
- Lowering overall execution time by sharing data and computation
- Considering execution costs

Defining information requirements

High-level information requirements

Requirement translation [1]

Import plugins [4,5]

SQL, HIVE, VERTICA, APACHE HBASE, pentaho

Challenges & Solution

1) Incremental advancement

- To guarantee semantical overlapping when comparing two operations
- We must ensure that all the predeceasing operations of both operations coincide
- Bottom-up (sources-to-target) algorithm for consolidating data flows

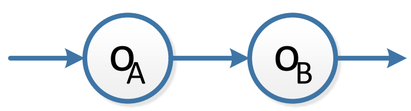
2) Operation reordering

- Swap, distribute/factorize, merge/split, association
- Generic equivalence rules to guarantee the equivalence of reordering transformations
- In terms of properties of data flow operations (i.e., **S**chema, **V**alue, **O**rder)

$o = (\mathbb{I}, \mathbb{O}, \mathbb{S}, \mathbf{Pre}, \mathbf{Post})$

$\mathbf{Pre} = (S_{pre}, V_{pre}, O_{pre})$

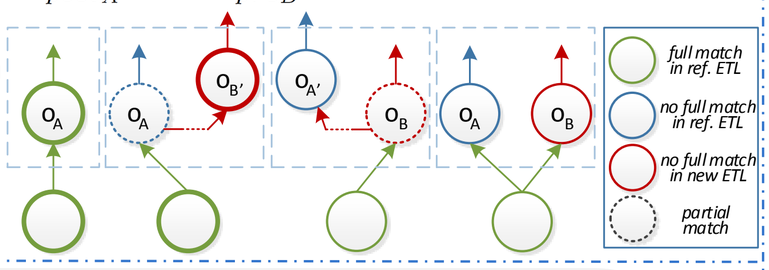
$\mathbf{Post} = (S_{post\_gen}, S_{post\_rem}, V_{post}, O_{post})$



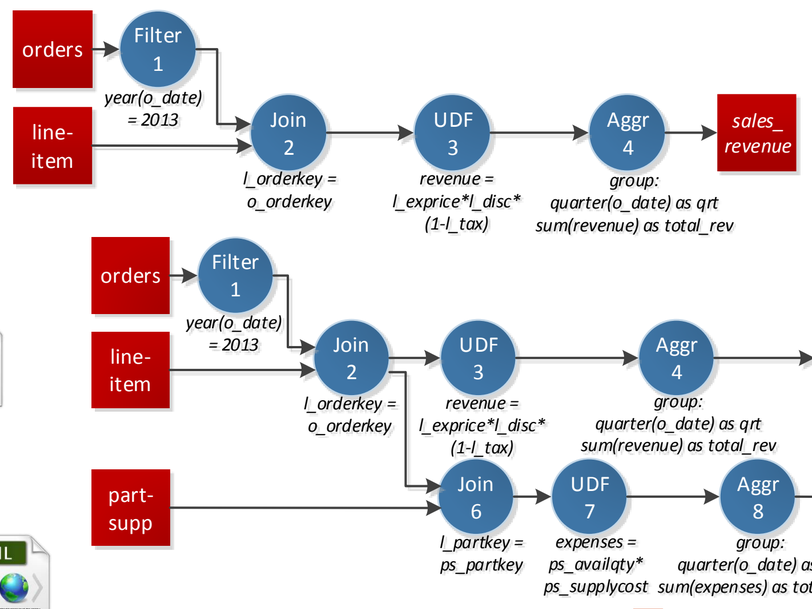
- Check for the conflicts
  - ✓ **S**chema $S_{post\_rem_B} \cap S_{pre_A} = \varnothing \wedge S_{post\_gen_A} \cap S_{pre_B} = \varnothing$
  - ✓ ***V**alue* $V_{post_B} \cap V_{pre_A} = \varnothing \wedge V_{post_A} \cap V_{pre_B} = \varnothing$
  - ✓ ***O**rder* $O_{post_B} \Rightarrow \neg O_{pre_A} \wedge O_{post_A} \Rightarrow \neg O_{pre_B}$

3) Operation comparison

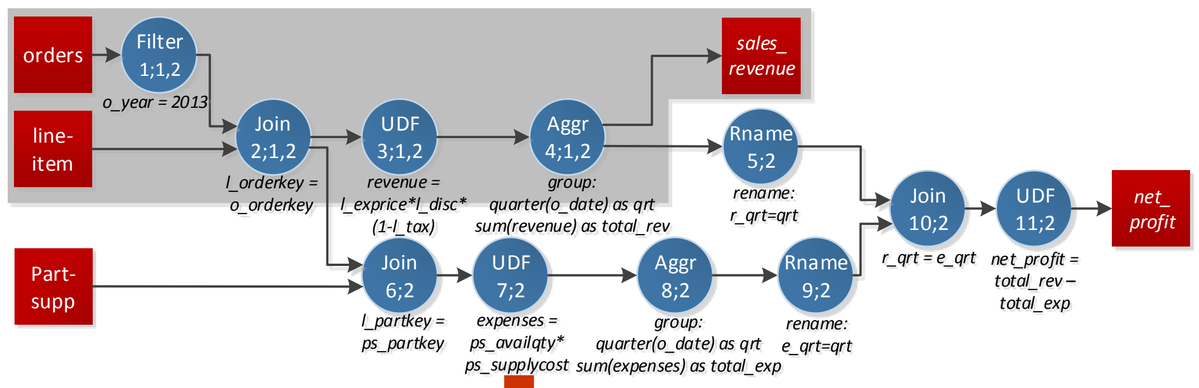
- Generic operation comparison
- Full and partial operation matching



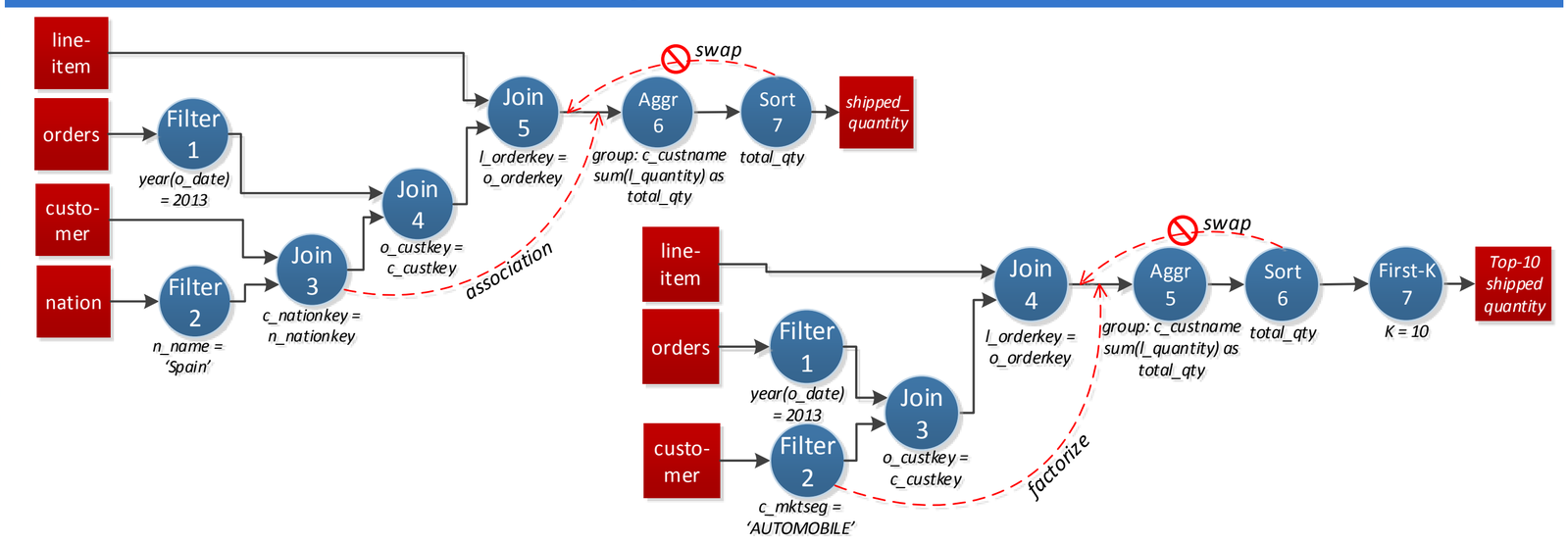
Logical representation of data intensive flows

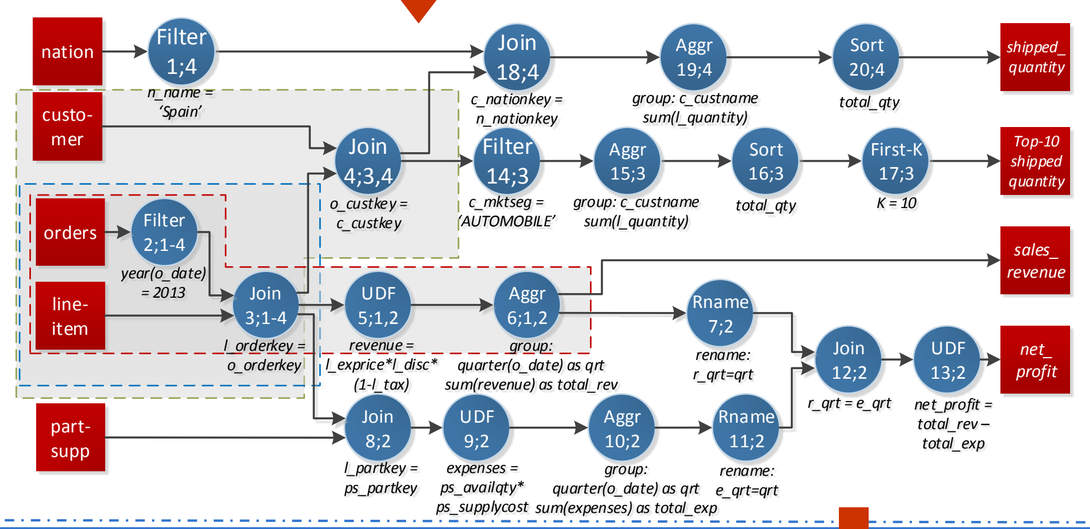


Case 1: complete overlap



Case 2: operation reordering & partial overlap





Deployment of data intensive flows

Export plugins [4,5]

- ✓ Heterogeneous execution environments
- ✓ Distributed flow execution
- ✓ Fragmented data flows
- ✓ Engine-specific parameters
  - ✓ Engine availability
  - ✓ Engine applicability
  - ✓ Data vs. function shipping

SQL, VERTICA, PostgreSQL, MariaDB, HIVE, pentaho, INFORMATICA The Data Integration Company, Cassandra, APACHE HBASE, hadoop

Further reading:

1. Oscar Romero, Alkis Simitsis, Alberto Abelló: GEM: Requirement-Driven Generation of ETL and Multidimensional Conceptual Designs. DaWaK 2011: 80-95
2. Petar Jovanovic, Oscar Romero, Alkis Simitsis, Alberto Abelló: Integrating ETL Processes from Information Requirements. DaWaK 2012: 65-80
3. Petar Jovanovic, Oscar Romero, Alkis Simitsis, Alberto Abelló: Requirement-Driven Creation and Deployment of Multidimensional and ETL Designs. ER Workshops 2012: 391-395
4. Petar Jovanovic, Alkis Simitsis, Kevin Wilkinson: Engine independence for logical analytic flows. ICDE 2014: 1060-1071
5. Petar Jovanovic, Alkis Simitsis, Kevin Wilkinson: BabbleFlow: a translator for analytic data flow programs. SIGMOD Conference 2014: 713-716
6. Petar Jovanovic, Oscar Romero, Alkis Simitsis, Alberto Abelló, Daria Mayorova: A requirement-driven approach to the design and evolution of data warehouses. Inf. Syst. 44: 94-119 (2014)

Fourth European Business Intelligence Summer School (eBISS 2014) – Berlin, Germany