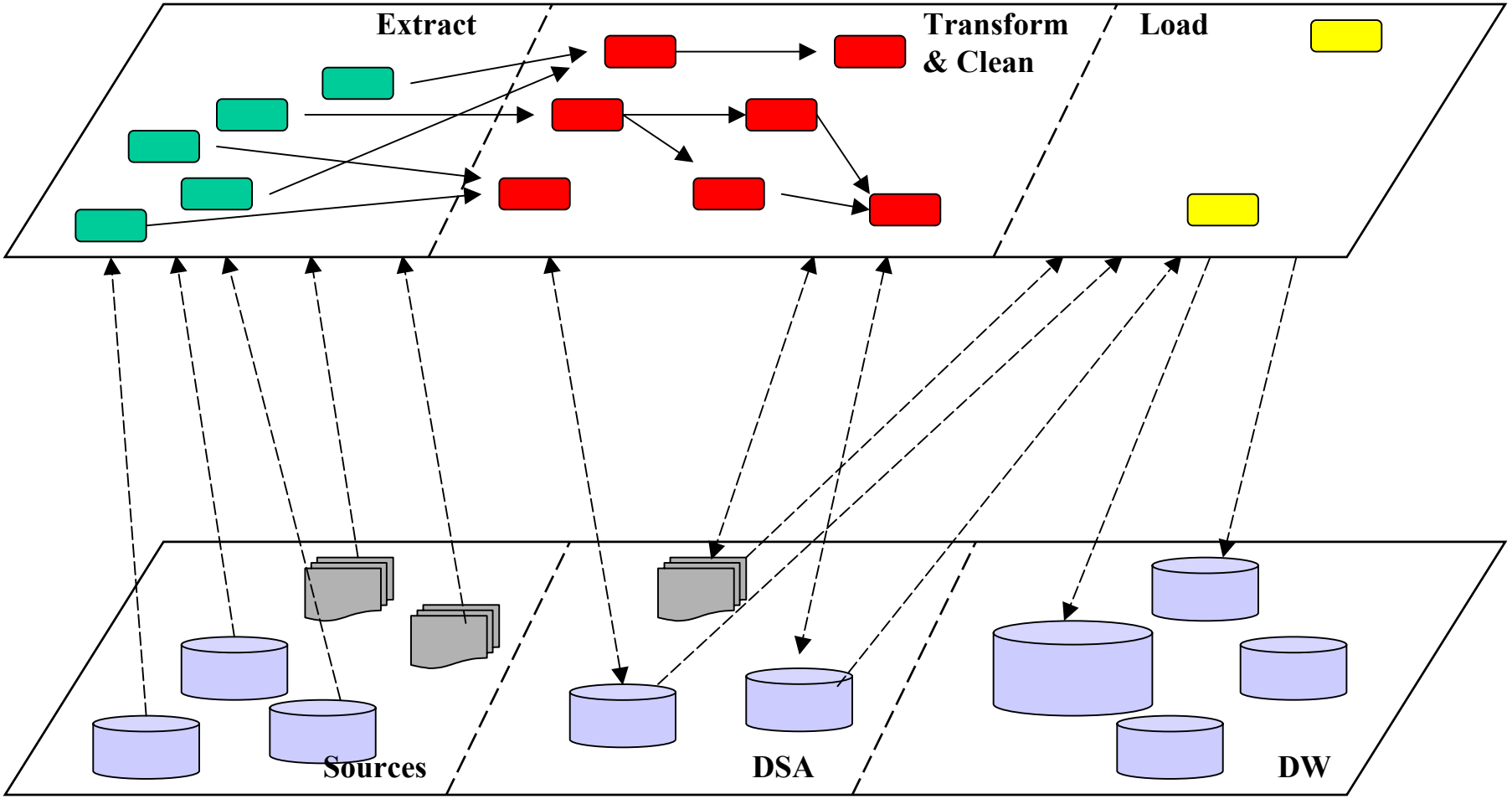
On the Logical Modeling of ETL Processes

Panos Vassiliadis, Alkis Simitsis, Spiros Skiadopoulos

National Technical University of Athens

{pvassil,asimi,spiros}@dbnet.ece.ntua.gr

What are Extract-Transform-Load (ETL) activities?



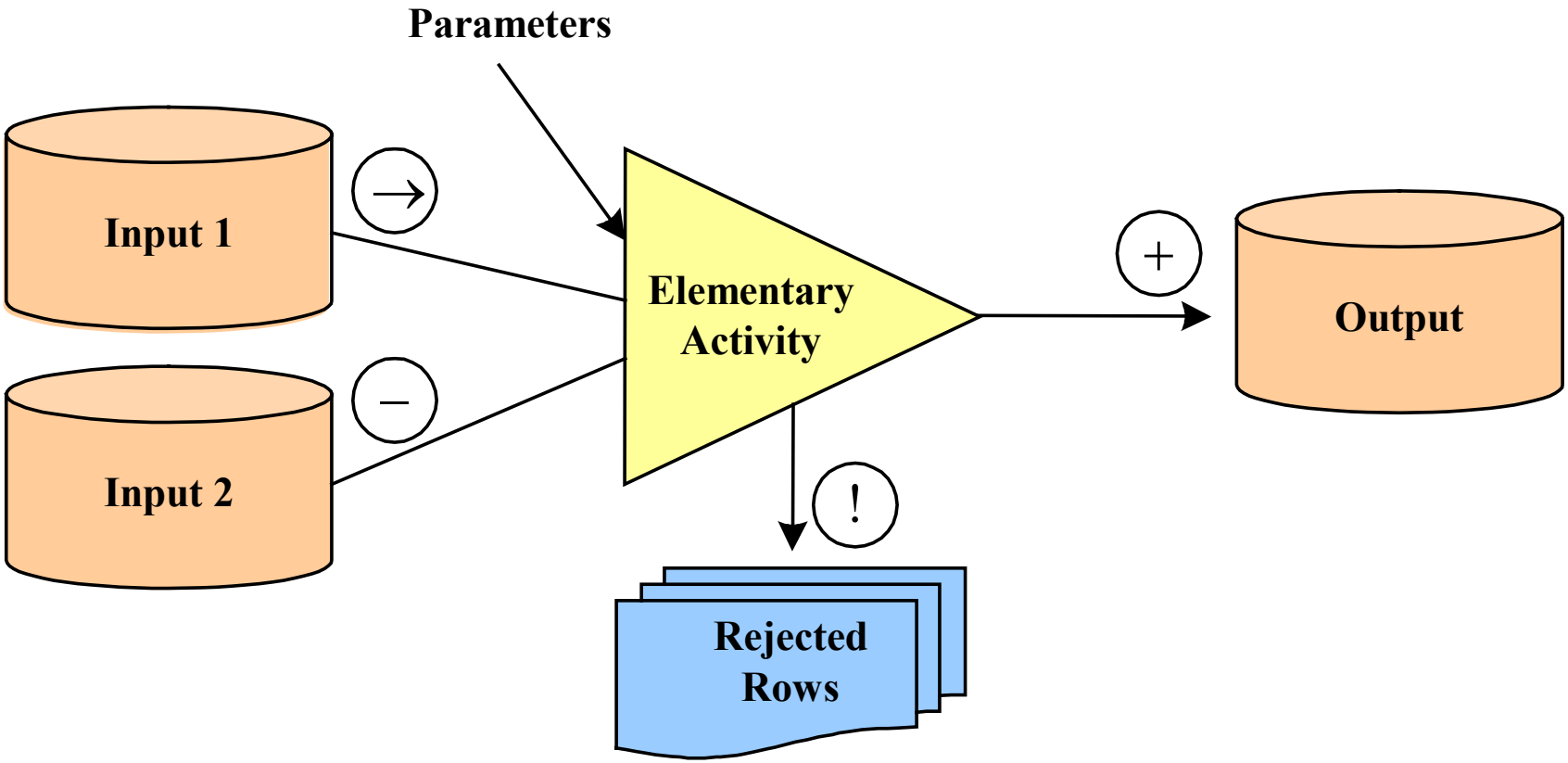
Preliminaries

- **Data types** (name, domain) and **Constants**
- **Attributes**
- **Schemata** (finite lists of attributes)
- **RecordSets** (name, schema, extension) standing for tables and record files
- **Function types** (name, a finite list of parameter data types, and a single return data type). A **function** is an instance of a function type

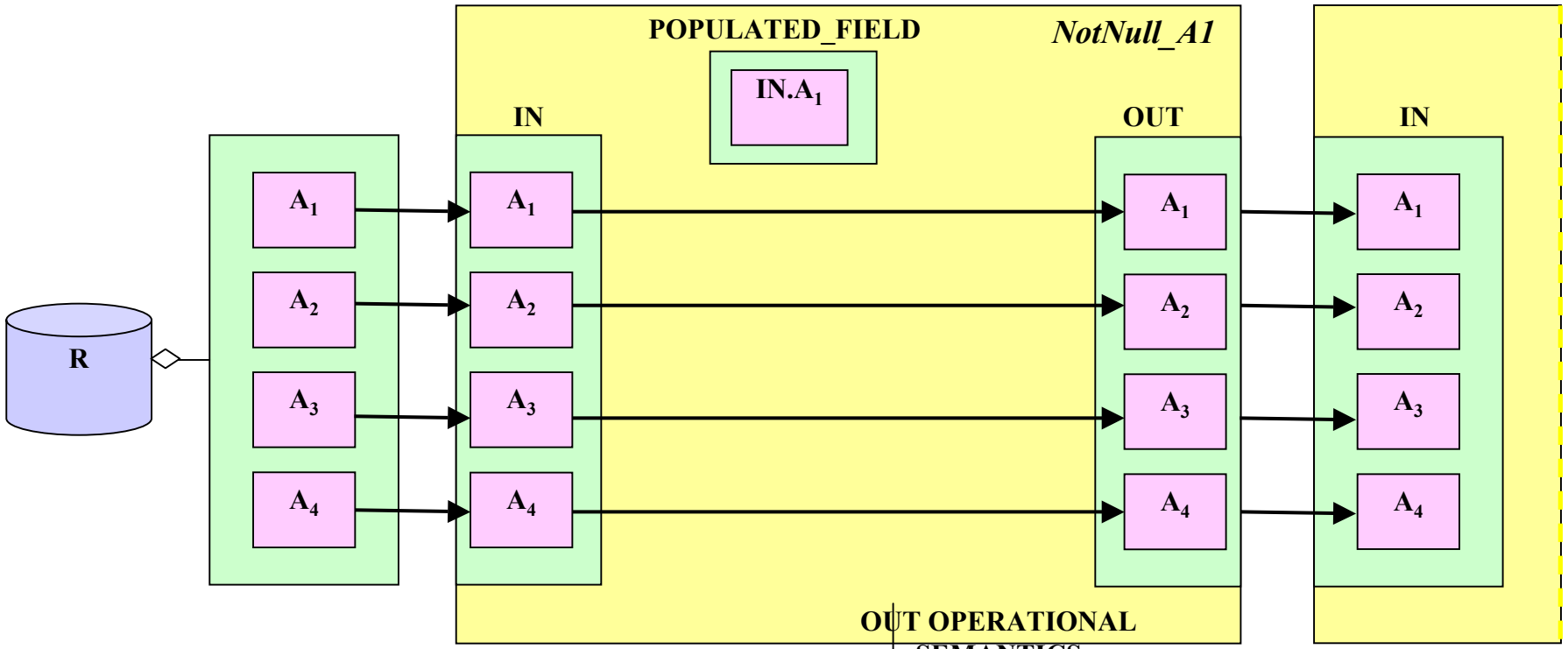
Elementary Activities

- **Name**
- **Input Schemata**
- **Output Schema**
- **Rejections Schema**
- **Parameter List**
- **Output/Rejection Operational Semantics**: an SQL statement describing the content passed to the output of the operation, with respect to its input.
- **Data Provider/Consumer Semantics**: is the output appended to the target, or overwritten? Are the input rows simply read, or removed from the sources as well?

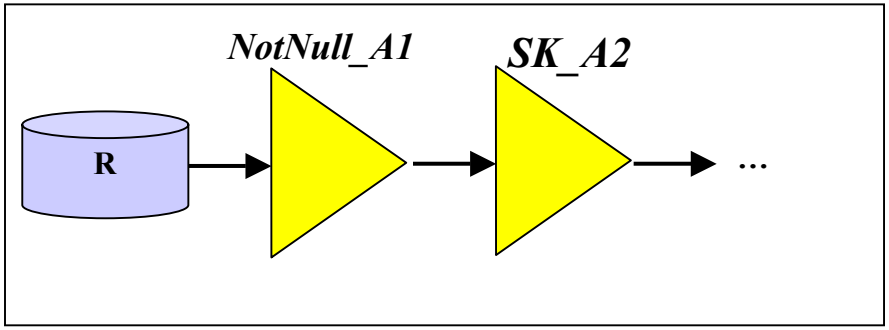
Naïve Example of Elementary Activity



Complex Example

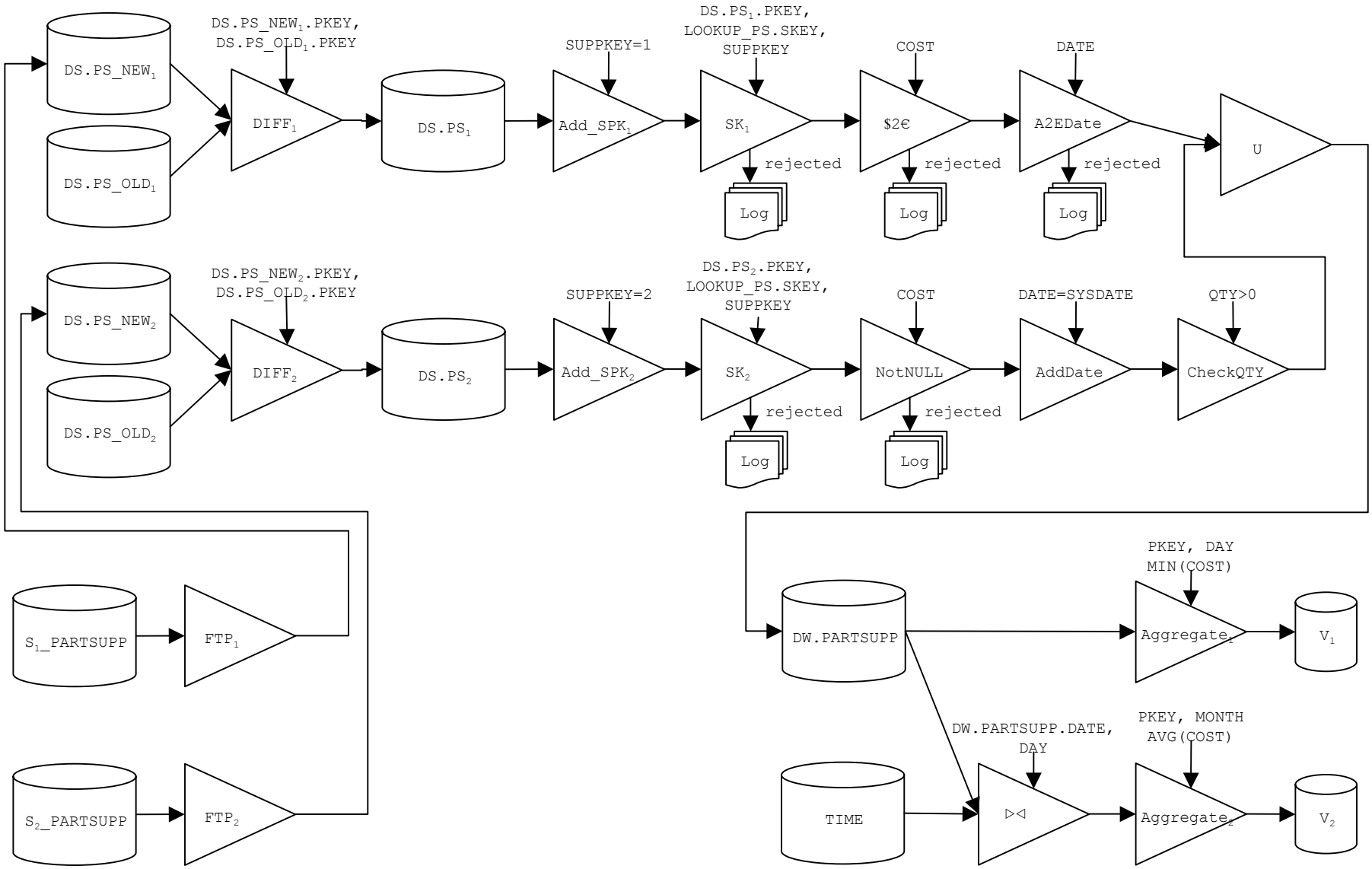


Legend:

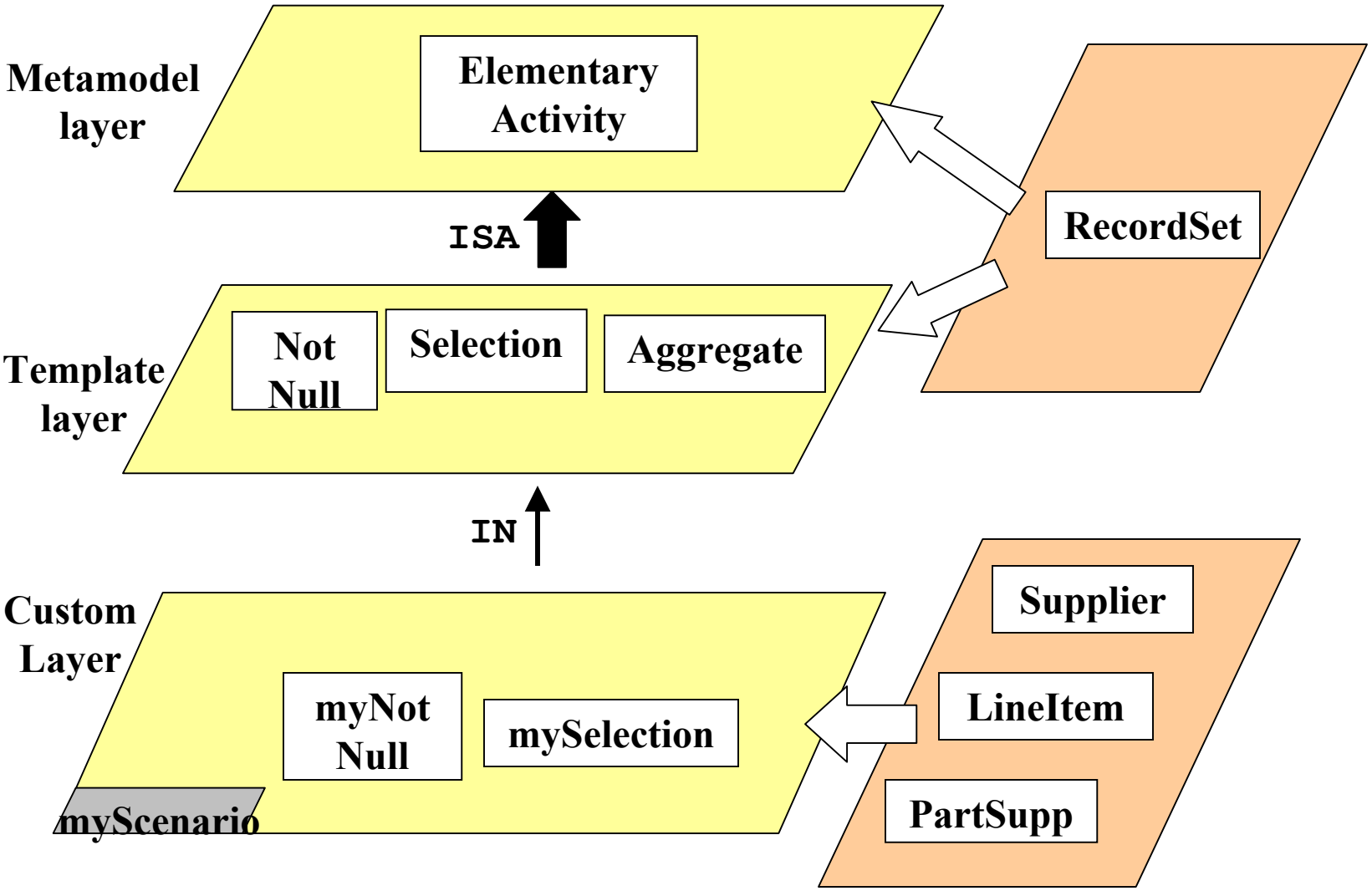



```
SELECT IN.A1 AS OUT.A1, IN.A2 AS
  OUT.A2,IN.A3 AS OUT.A3, IN.A4
  AS OUT.A4
FROM NOTNULL_A1.IN
WHERE IN.A1 NOT NULL
```

A Typical ETL Scenario



Template Activities



Template Activities

Filters

SELECTION ($\varphi(A_{n+1},\ldots,A_{n+k})$)

UNIQUE VALUE (R.A)

NOT NULL (R.A)

DOMAIN MISMATCH (R.A,x_{low},x_{high})

PRIMARY KEY VIOLATION (R.A_1,…,R.A_K)

FOREIGN KEY VIOLATION ([R.A_1,…,R.A_K],[S.A_1,…,S.A_K])

Unary Transformations

PUSH

AGGREGATION ([A_1,…,A_k],[$\gamma_1(A_1)$,…,$\gamma_m(A_m)$])

PROJECTION([A_1,…,A_k])

FUNCTION APPLICATION($f_1(A_1)$,…,$f_k(A_k)$)

SURROGATE KEY ASSIGNMENT (R.PRODKEY,S.SKEY,x)

TUPLE NORMALIZATION (R.A_{n+1},…,R.A_{n+k},A_{CODE},A_{VALUE},h)

TUPLE DENORMALIZATION (A_{CODE},A_{VALUE},R.A_{n+1},…,R.A_{n+k},h')

Binary Transformations

UNION(R,S)

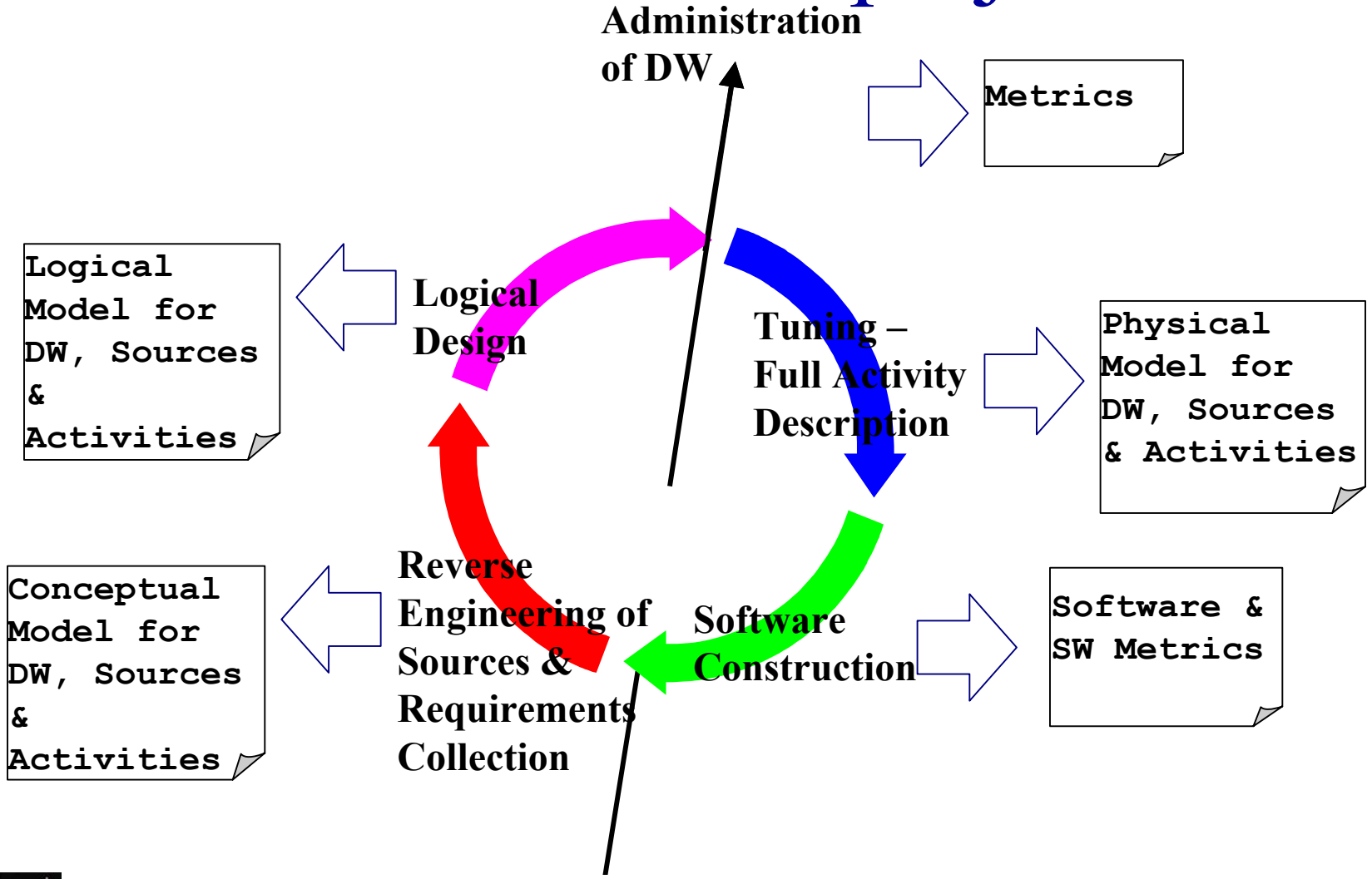
JOIN(R,S,[(A_1,B_1),…,(A_k,B_k)])

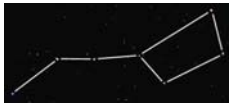
DIFF(R,S,[(A_1,B_1),…,(A_k,B_k)])

Example of Template Activity

Activity type	DOMAIN MISMATCH (R.A,x_{low}, x_{high})
Input schema	[A_1,…,A_n]
Parameters	[$FIELD,R.A] [$LOW,x_{low}] [$HIGH,$x_{high}$]
Output schema	[B_1,…,B_n]
Rejection schema	[C_1,…,C_n]
Equivalent SQL statement for the output of the activity	SELECT A_1 AS B_1,…, A_n AS B_n FROM R WHERE $FIELD IN [$LOW,$HIGH]

The ARKTOS II project

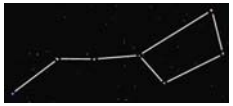




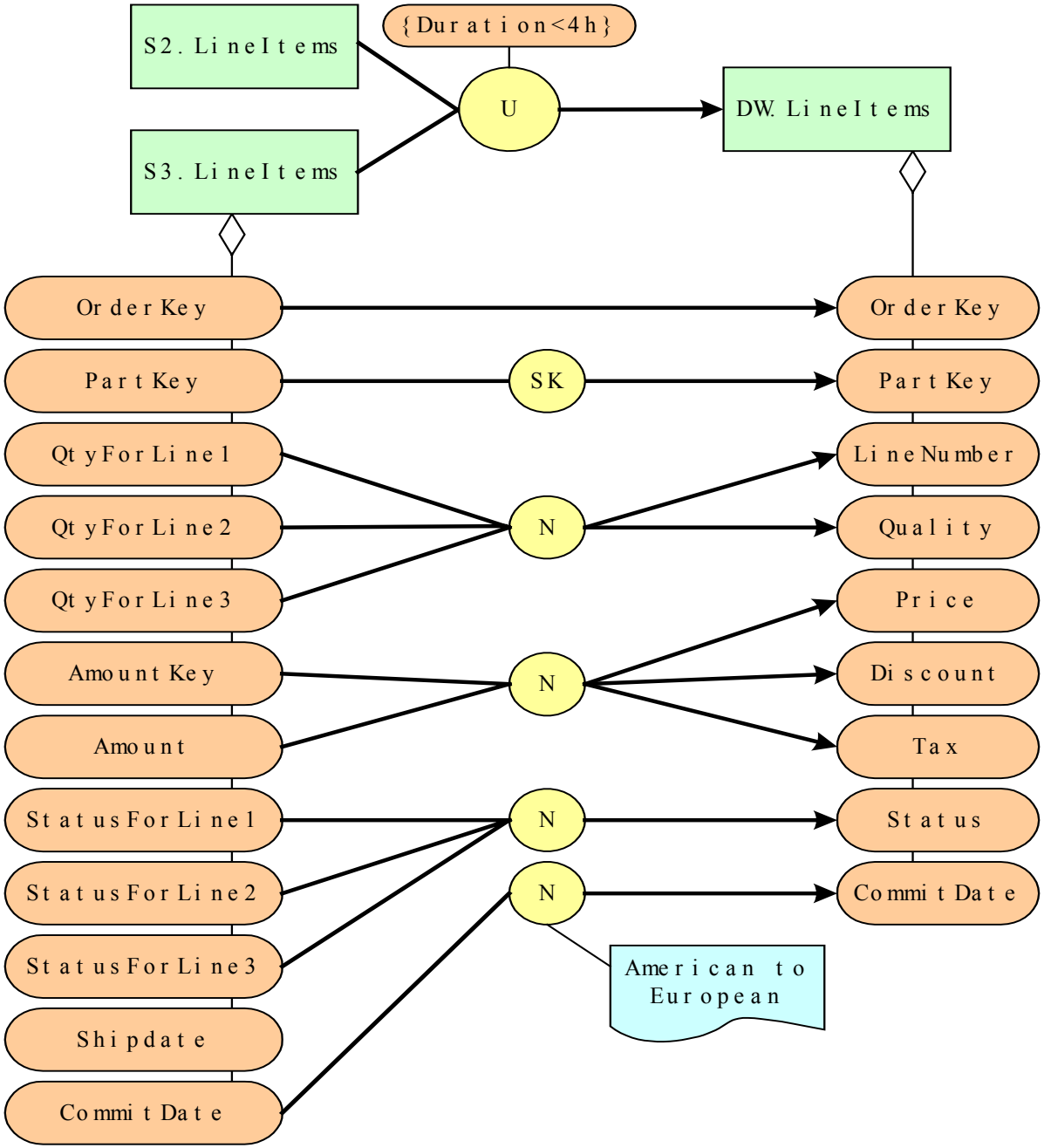
http://www.dblab.ece.ntua.gr/~pvassil/projects/arktos_II/

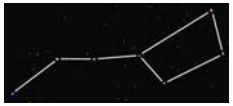
On-going & Future Work

- Conceptual model for the initial phases of ETL projects
- Logical model for activities
- Graph modeling for ETL scenarios [DMDW02]
- Optimization of ETL scenarios
- On-going development of a set of loosely coupled tools



Conceptual Model





Graph Modeling [DMDW'02]