Km4City Technologies: data ingestion and mining - Develop ETL processes –

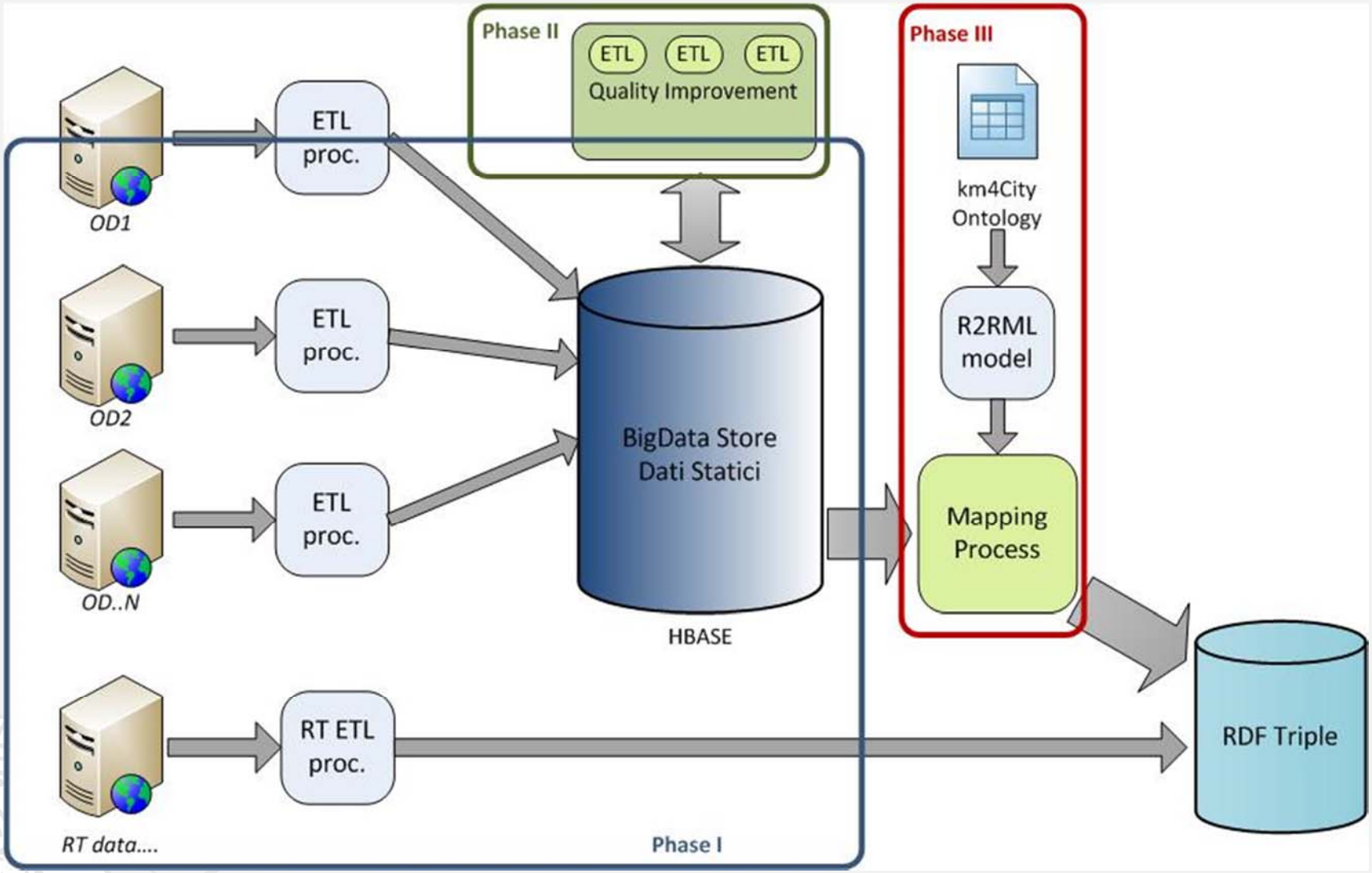
14 Luglio 2016

DISIT Lab, Dipartimento di Ingegneria dell'Informazione, DINFO
Università degli Studi di Firenze
Via S. Marta 3, 50139, Firenze, Italy
Tel: +39-055-2758515, fax: +39-055-2758570
http://www.disit.dinfo.unifi.it ***alias*** http://www.disit.org

Prof. Paolo Nesi, paolo.nesi@unifi.it
Pierfrancesco Bellini, pierfrancesco.bellini@unifi.it
Michela paolucci, michela.paolucci@unifi.it
Simone Panicucci, simone.panicucci@stud.unifi.it

Big Data: from Open Data to Triples

Data Engineering Architecture



ETL Process

The three phases are:

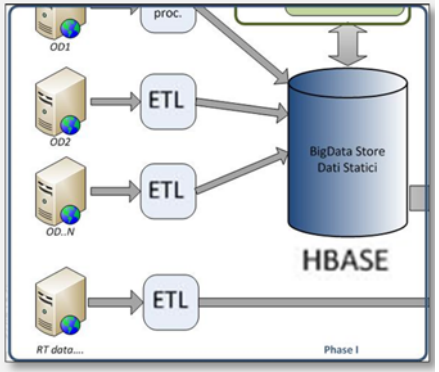
- **Extracting** data from outside sources (**Ingestion** phase).

- **Transforming** data to fit operational needs which may include improvements of quality levels (**Data Quality Improvement** phase).

- **Loading** data into the end target (database, operational data store, data warehouse, data mart, etc.). So the data can be translated in **RDF triples using a specific ontology.**

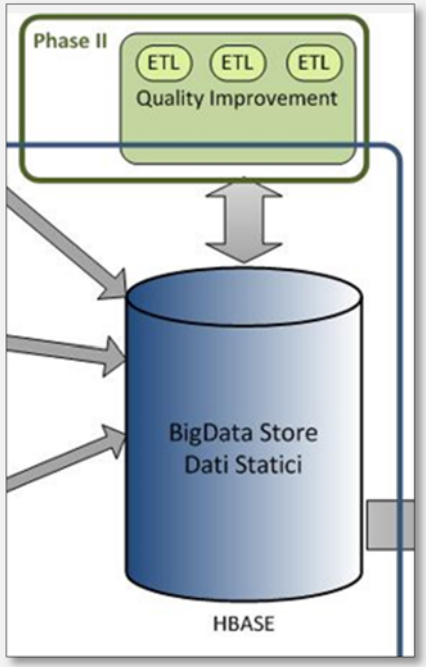
Phase I: Data Ingestion

- **Purpose is to store data in HBase (Big Data NoSQL database)**.
- **Acquisition of wide range of OD/PD**: open and private data, static, quasi static and/or dynamic real time data.
- **Static and semi-static data** include: points of interests, geo-referenced services, maps, accidents statistics, etc.
  - files in several formats (SHP, KML, CVS, ZIP, XML, JSON, etc.)
- **Dynamic data** mainly data coming from sensors
  - parking, weather conditions, pollution measures, bus position, …
  - using Web Services
- Using **Pentaho - Kettle** for data integration (Open Source tool)
  - using specific **ETL** Kettle transformation processes (one or more for each data source)

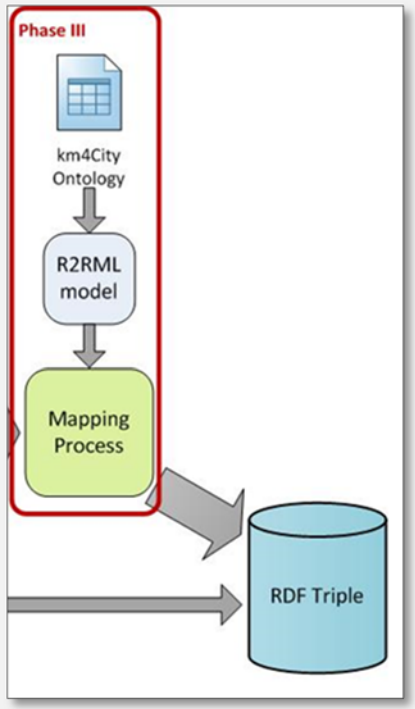


Phase II: Data Quality Improvement

- **Purpose: add more information as possible and normalize data from ingestion**
- **Problems kinds:**
  - Inconsistencies, incompleteness, typos, lack of standards, multiple standards, …
- **Problems on:**
  - Place-name code
  - Street names (e.g., dividing names from numbers, normalize when possible)
  - Dates and Time: normalizing
  - Telephone numbers: normalizing
  - Web links and emails: normalizing



Phase III: Data mapping



- **Purpose is to translate data from QI in RDF triples.**
- **We use triples to do inference on data.**

- Using **Karma Data Integration tool, a mapping model** from SQL to RDF on the basis of the ontology was created.
  - Data to be mapped first temporary passed from HBase to MySQL and then mapped using Karma (in batch mode)
- The mapped data in triples have to be uploaded (and indexed) to the **RDF Store** (Virtuoso).
- Triples are composed by a subject, a predicate and an object.

ETL tool: Pentaho Data Integration (PDI)

KEY CONCEPTS

Pentaho Data Integration (PDI)

- **Pentaho** is a framework that contains several packages integrated to allow complete management:
  - ***Business Intelligence problems;***
  - ***Data Warehouse problems;***
  - ***Big Data problems.***

- **Kettle** is the ETL component Pentaho for data transfer and processing.

Pentaho Data Integration (Kettle) (1)

- Free, **open** ETL tool
  - It is available also in **enterprise version**
- **Developed in Java**, therefore is guaranteed the compatibility and portability with the major operating systems (Windows, Linux, OS X, …)
- **Powerful** Extraction, Transformation and Loading (ETL) capabilities

Pentaho Data Integration (Kettle) (2)

- **Scalable**, standards-based architecture.
- Opportunity to interfacing with the main NoSQL Databases (HBase, Cassandra, MongoDB, CouchDB, …).
- It uses an innovative, **metadata-driven** approach.
- Graphical, **drag and drop** design environment.

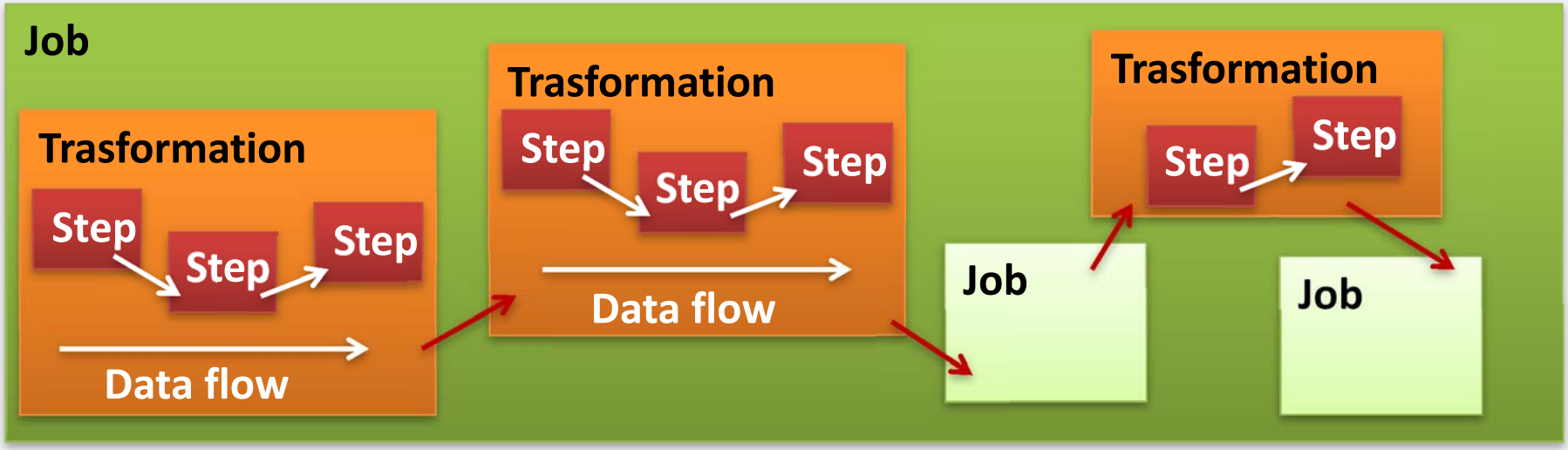
Kettle: Concepts

- Kettle is based on two key concepts (from operating point of view):
  - **Job** (with extension ".kjb");
  - **Transformation** (with extension ".ktr"), composed of several **steps**.

- Kettle's key components are:
  - **Spoon** for ETL process modeling;
  - **Pan** to execute the transformations from command line;
  - **Kitchen** to execute the Job from command line.

Kettle: Operational structure

Kettle operating components are organized as follows:



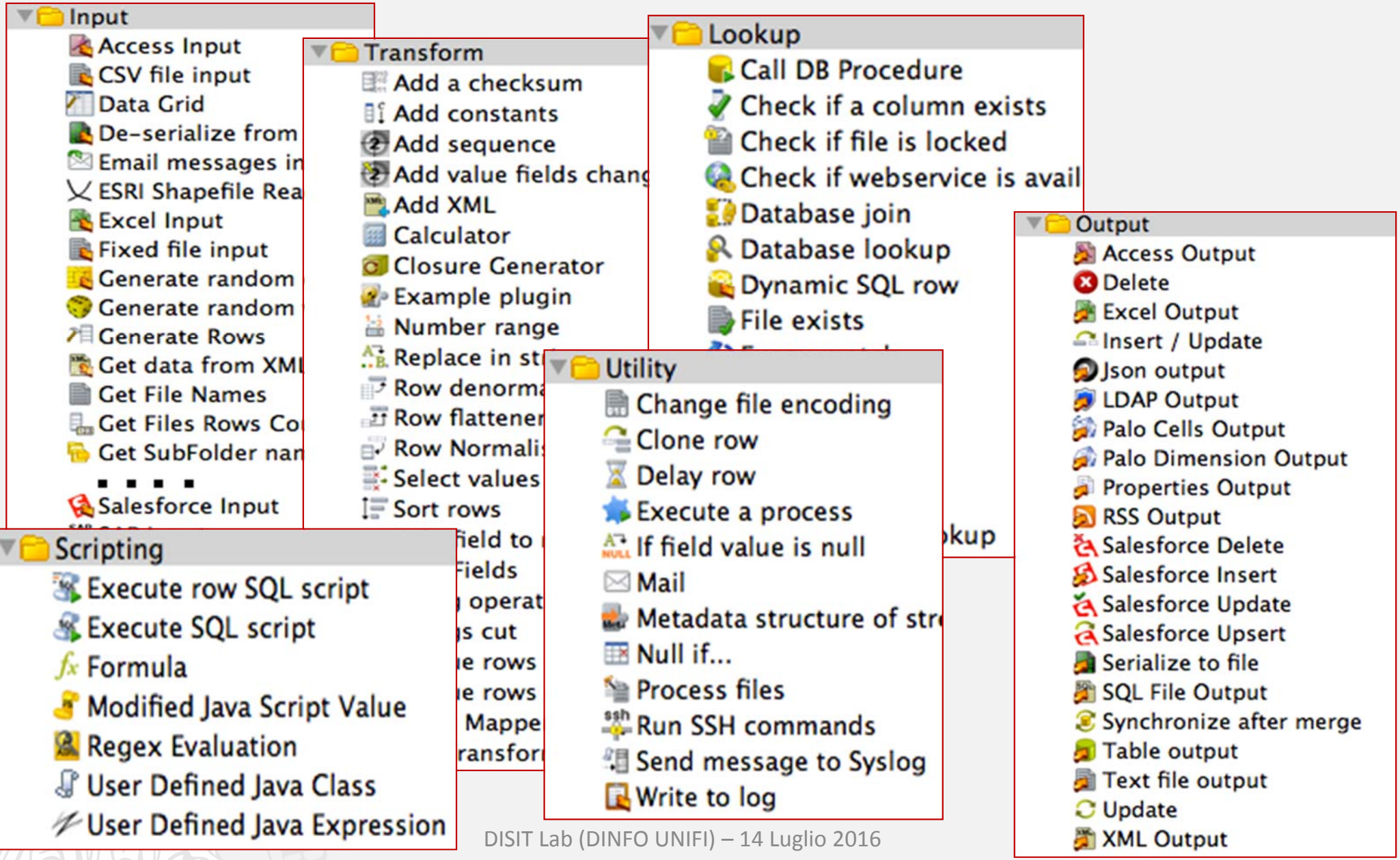
- The data are seen as rows flow from one step to another one.
- The steps are executed in parallel on separated threads and there is no necessarily a beginning or end point of the transformation.
- A job manages the sequential execution of lower-level entities: transformations or other jobs.

Type of Steps in Spoon (1)

Three different kinds of steps:

- **Input:** process some kind of 'raw' resource (file, database query or system variables) and create an output stream of records from it.
- **Output:** (the reverse of input steps): accept records, and store them in some external resource (file, database table, etc.).
- **Transforming:** process input streams and perform particular actions on them (adding new fields/new records); these actions produce one or more output streams.
- **NOTE: 'Main.kjb'** is usually the primary job.
- REF: http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps

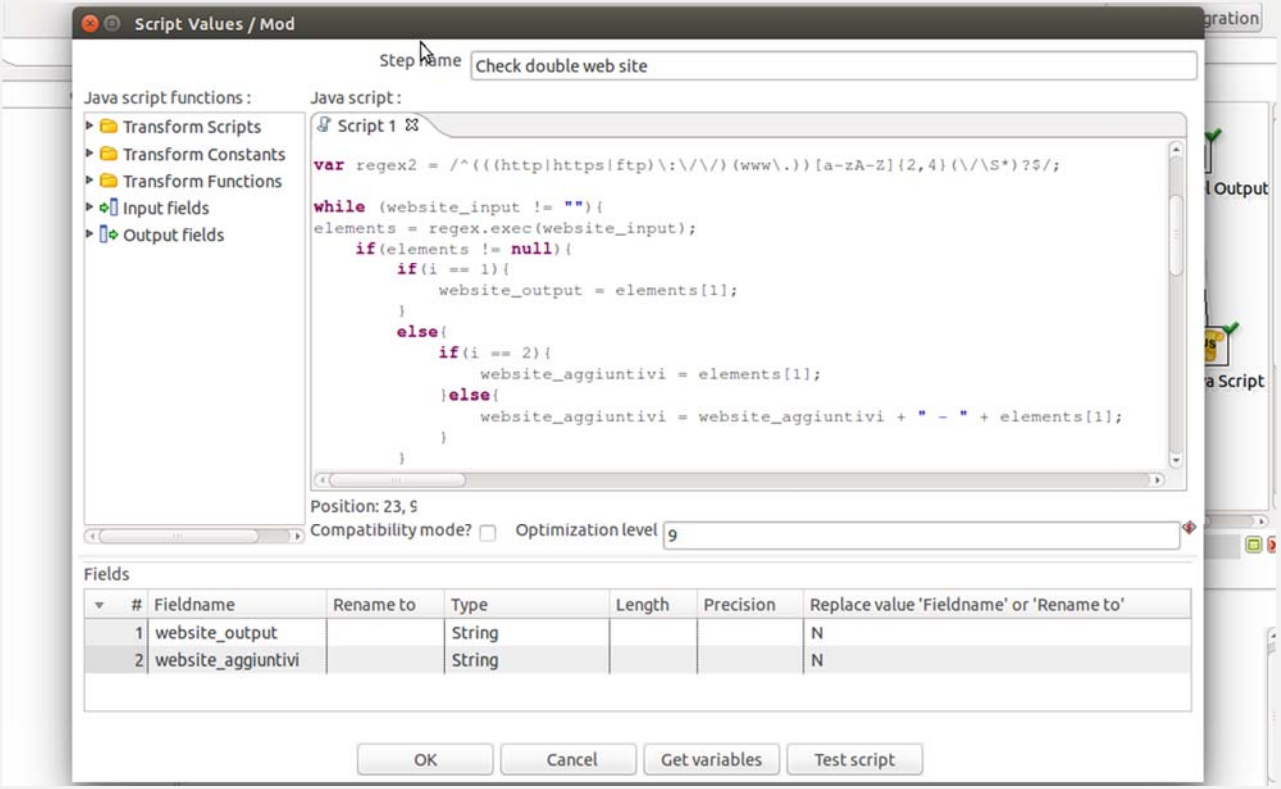
Type of Transformations in Spoon



Kettle: Transformations

Kettle offers many types of steps in order to execute various data operations, also it offers:

- ***possibility of add some JavaScript code;***
- ***possibility of use regular expressions.***

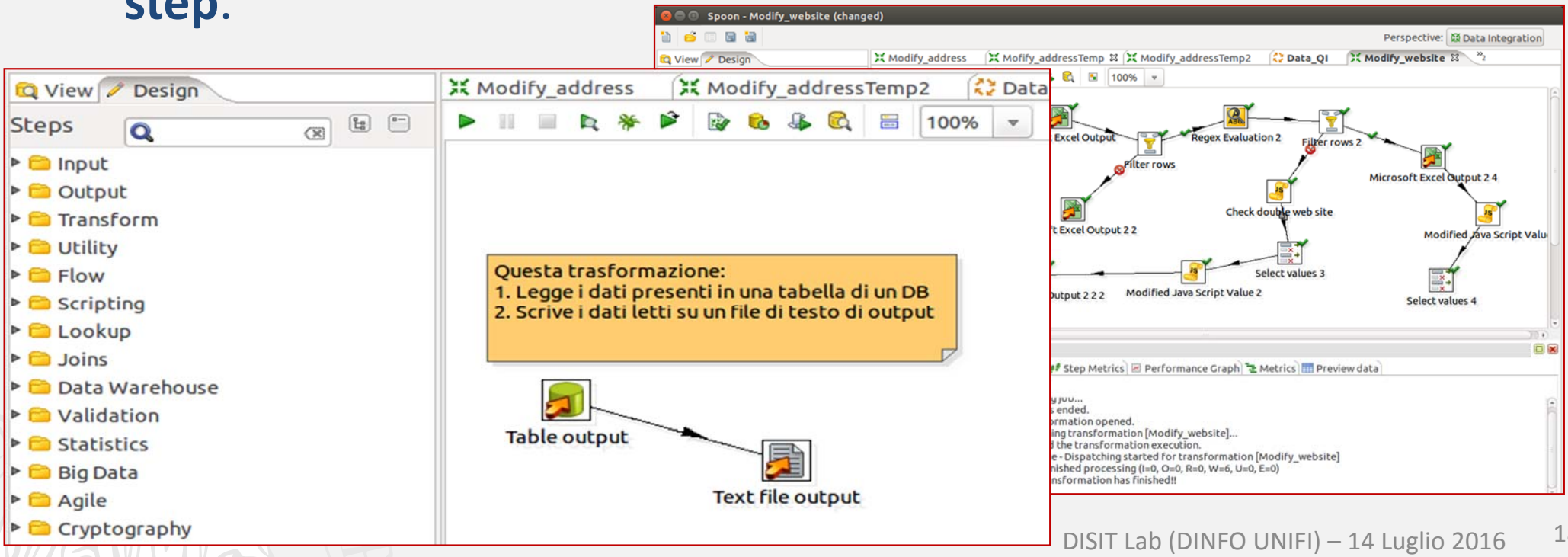


Spoon Concepts: Steps and hoops

- One **step** denotes a particular kind of **action** that is performed **on data**.
- **Hops** are links to connect steps together and allow data to pass from one step to another.
- Steps can be easily "created" by **dragging** the icon from the treeview **and dropping** them on the graphical model view.
- Kettle provides a lot of different step types and can be **extended with plugin**.

Kettle: Transformations (1)

- Transformations define how the data must be collected, processed and reloaded.
- Consist of a series of steps connected by links called Hop.
- Typically a transformation has one **input step**, one or multiple **transformation steps** and one or more **output step**.



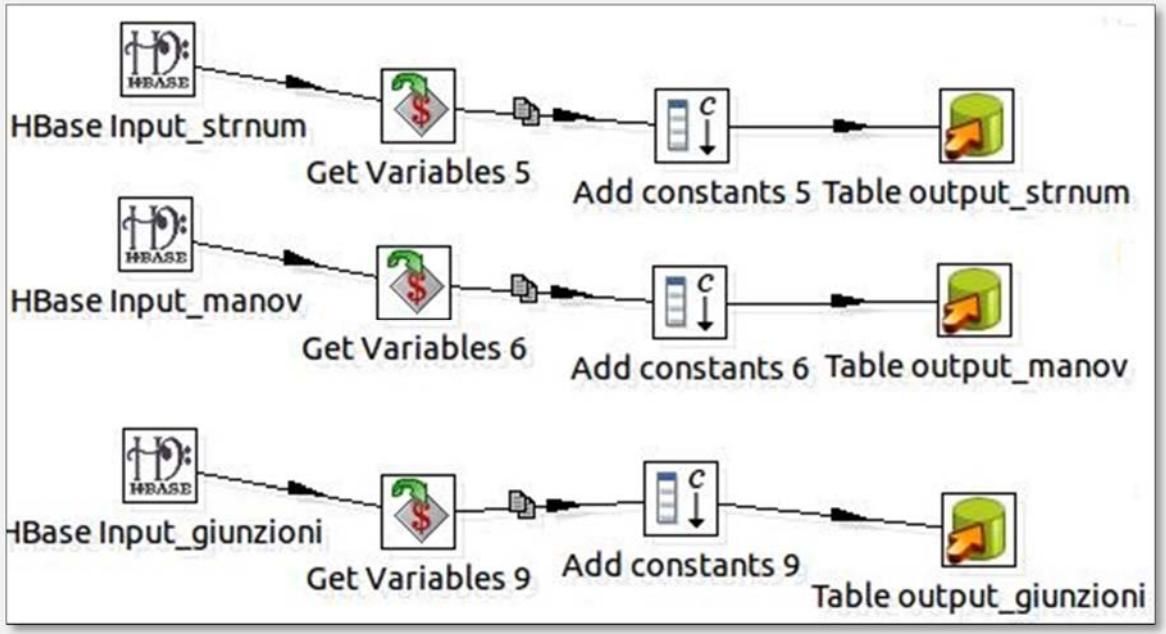
Sequential Execution

These steps (transformations) are executed sequentially (there is a single flow execution).

```
void main(){
        int a;
        f1(a);
        f2(a+2);
}
```

The statements are executed sequentially.

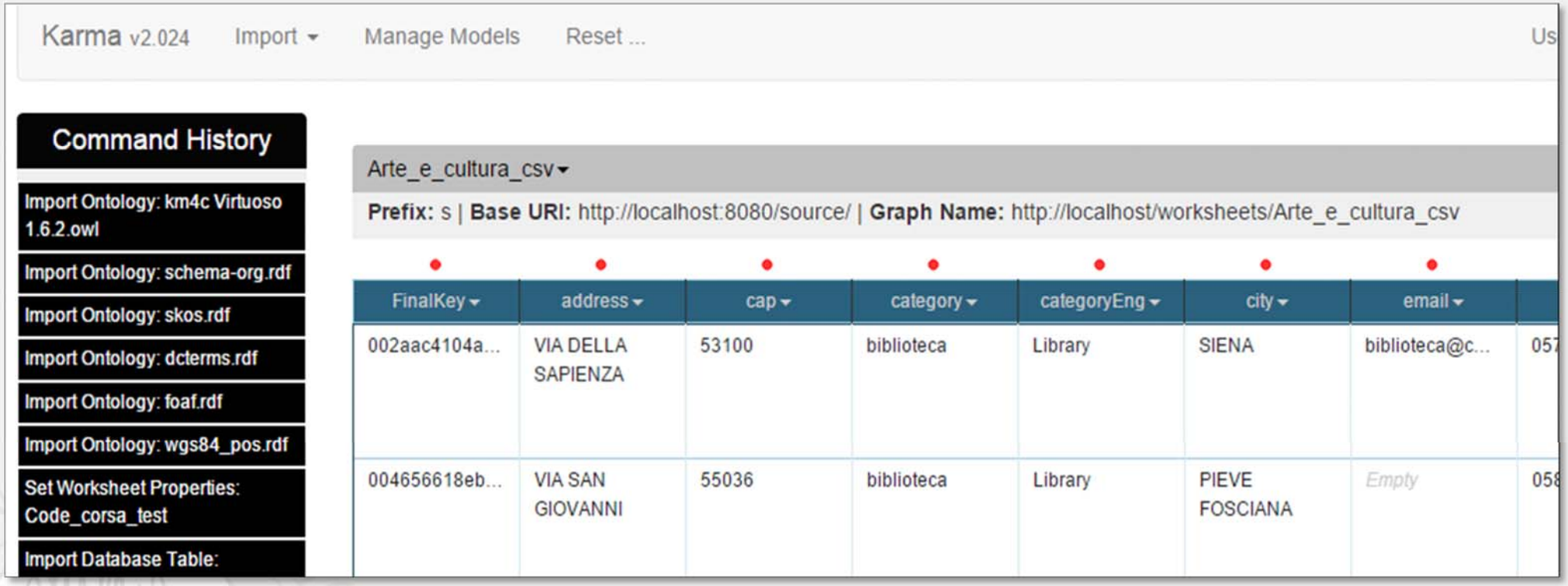
Parallel Execution



- Unlike before there are multiple streams of execution that are executed in parallel (simultaneously).
- Like in a multi-threading programming, multiple thread (portions of the running program) can virtually run independently and in parallel.

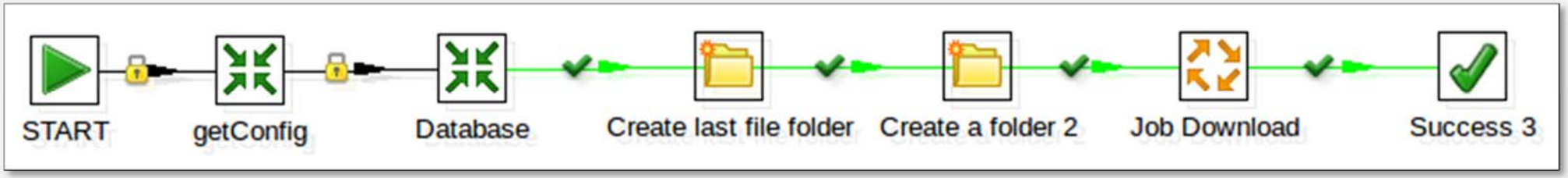
Karma

- Karma is a mapping model based on ontology (**km4city**) from MySQL tables to RDF.
- Triples are uploaded to **Virtuoso**, an RDF Store.
- It can import MySQL tables but no HBase ones.



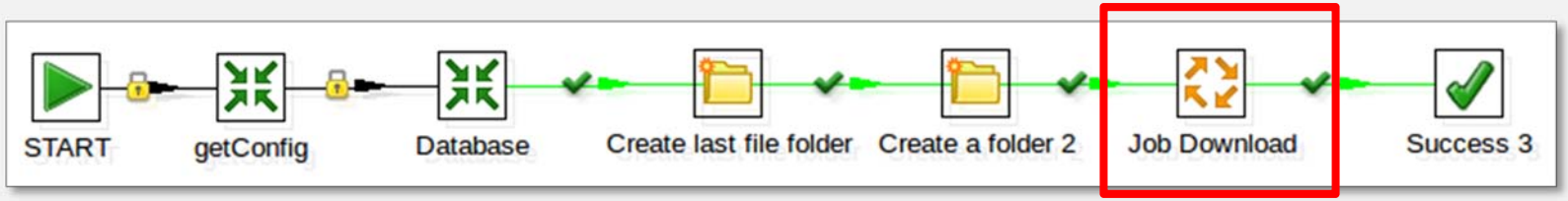
ETL Example

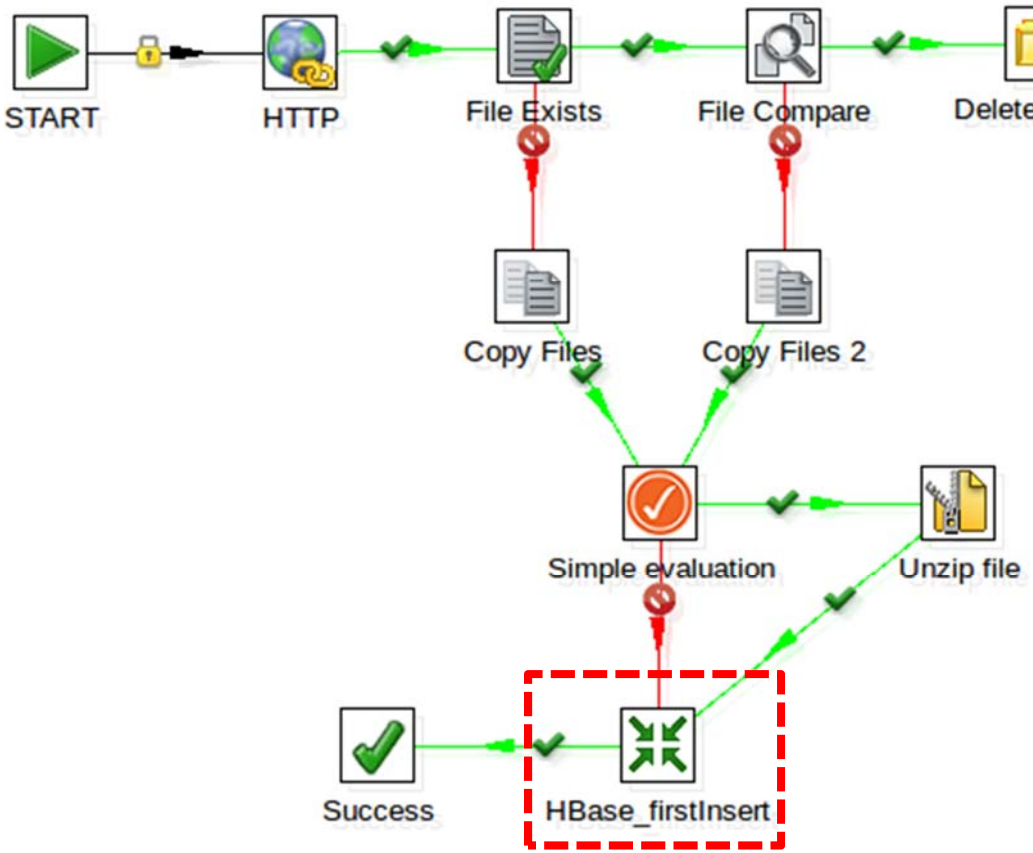
Ingestion



- '**getConfig**' loads database connection parameters (usernames, passwords, …) from a *.csv* file.
- '**Database**' transformation loads some process information.
  - This transformation uses a MySQL table (*process_manager2*) to get some process information (URL, file format, description, …).
- '**Job Download**' effectively downloads the dataset.

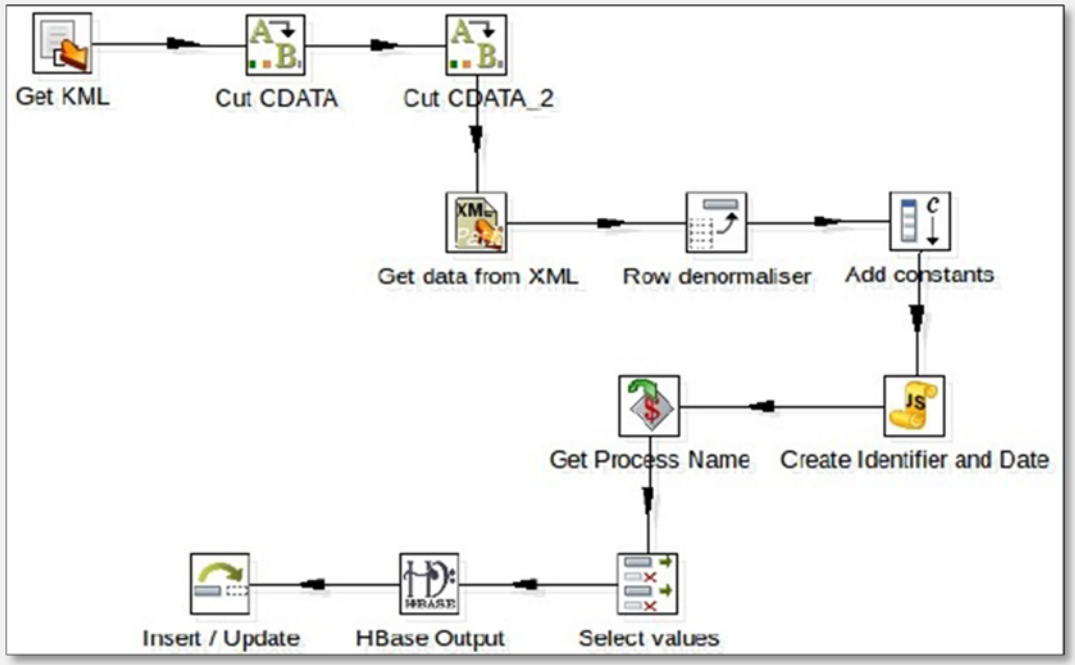
Job Download





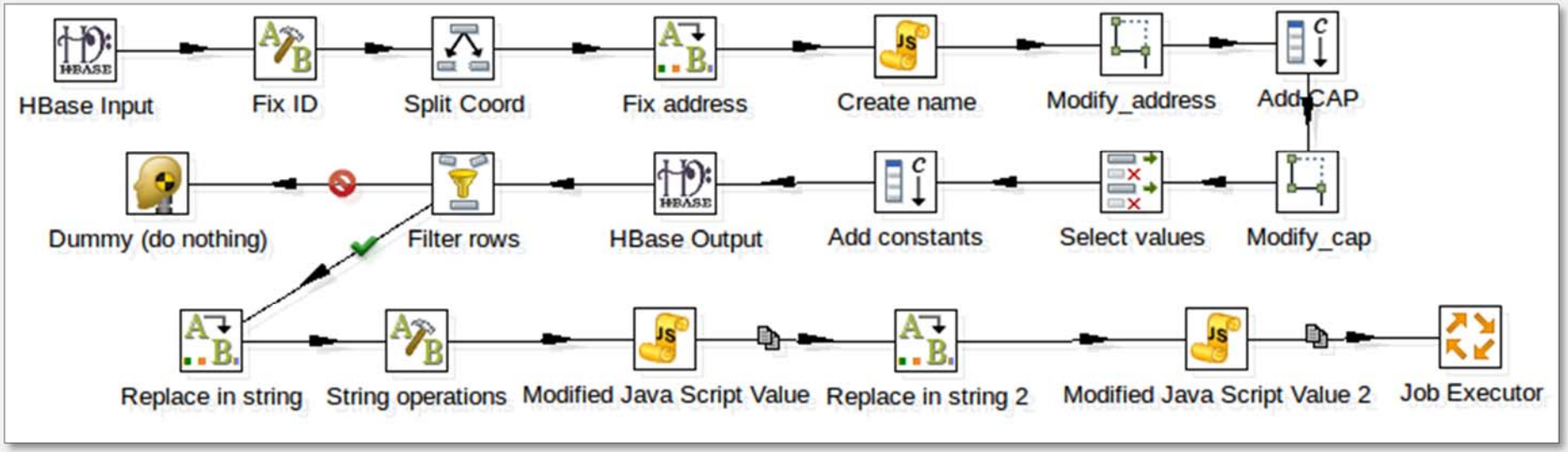
- '**HTTP**' step downloads the dataset defined in *PARAM* (URL) field.
- If a file with the same name already exists, we compare them.
  - If they are the same, we delete the folders created before.
  - Otherwise we unzip the file and '**HBase_firstInsert**' is called.

HBase_firstInsert



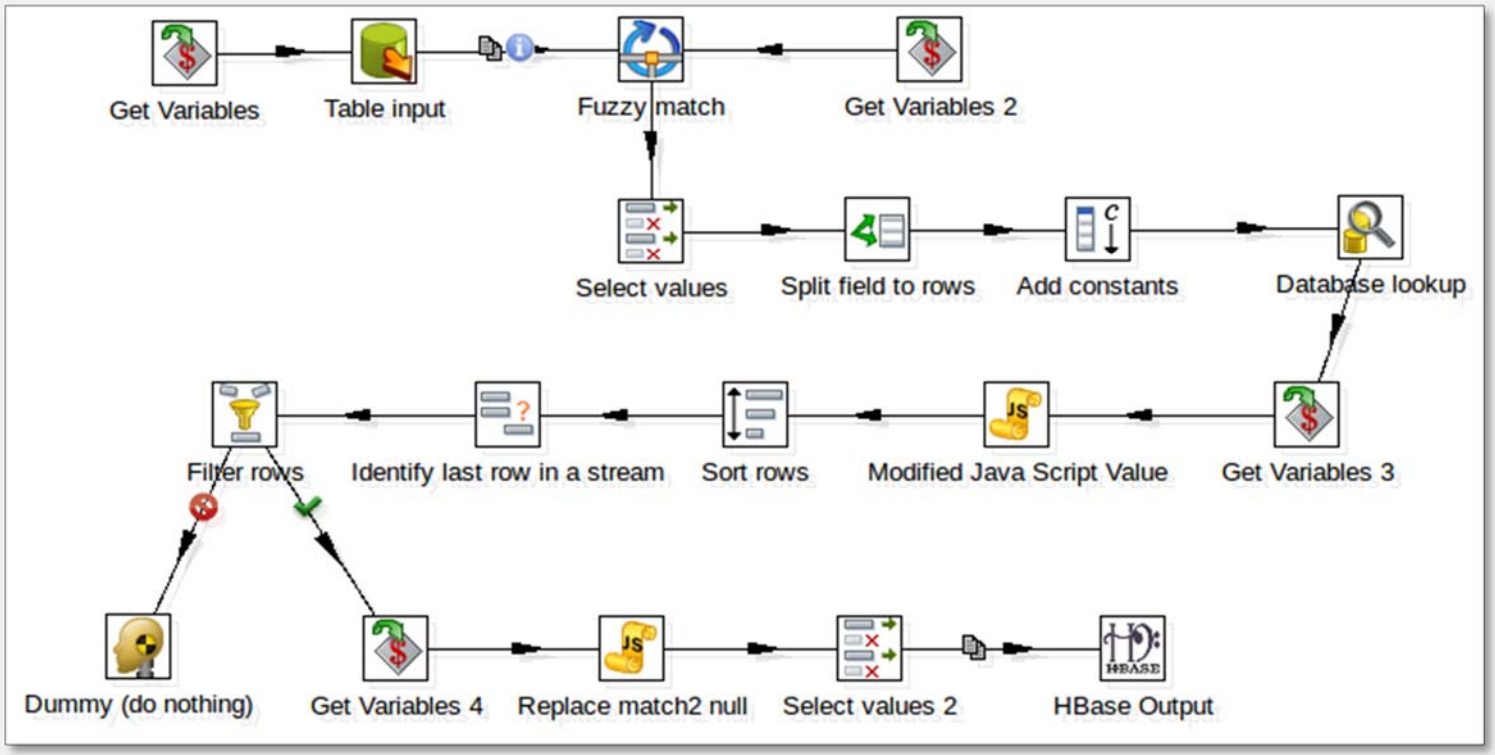
- '**Get KML**' loads the file just downloaded.
- '**Get data from XML**' and '**Row denormaliser**' extract fields from source file.
- In the JS step, we create an identifier (it will be use as key in HBase).
- '**HBase Output**' saves the information in a HBase table.
- '**Insert/Update**' updates the last ingestion date in MySQL table *process_manager2.*

Quality Improvement



- '**HBase Input**' gets back data saved at the ingestion end.
- '**Fix address**' is used to correct typing error (e.g. Giovambattista instead of Giovanbattista) or to simplify search of right toponym code.
- '**Modify_***' transformations normalize address, CAP, website, e-mail, phone number, …
- '**Add constants**' adds two fields (address_syn, codice_toponimo) which we will use in the next job.
- '**HBase Output**' saves in a new HBase table the quality improvement result.
- For the rows which have an un-empty *streetAddress*, the steps below extract a word from *streetAddress* which we will use to find the right toponym code.

Job Executor

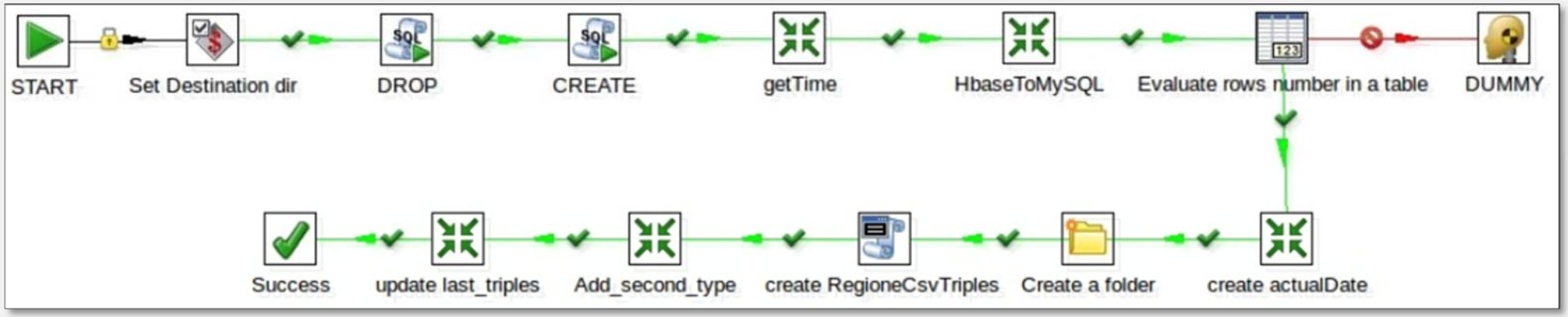


- To get toponym codes we use *tbl_toponimo_bis* MySQL table. It contains road names, toponym codes, town codes, …. Each road has his center coordinates (lat, long).
- '**Table input**' does the query on that table using the word created before.
- '**Fuzzy match**' calculates similarity (from 0.2 up to 1) between the query result and the address acquired during ingestion phase.
- Following steps calculate distance from ingestion coordinates and query ones and select the closest.
- '**HBase Output**' fills *address_syn* and *codice_toponimo* in QI HBase table.

QI Problems

- The method used for toponym code extraction sometimes produces wrong results.
- Given a word, might be impossible to determine right toponym code (e.g., query using '*Brunelleschi*' returns '*Via dei Brunelleschi*', '*Piazza Brunelleschi*').
- Summarize roads with their centers and calculate distances could not return right results (if the point of interest is far from his road center, it could be closer to another road center).

Triplification



- Since we use Karma to generate models, we have to move data from HBase to MySQL.
- '**DROP**' and '**CREATE**' respectively deletes MySQL table if already exists and creates a new one.
- '**getTime**' returns last triple generation timestamp for a specific process (using MySQL *process_manager2* table).
- '**HBaseToMySQL**' moves data from HBase to MySQL only if timestamp just calculated is older than the date of last ingestion (it prevents to generate triples based on the same data).
- '**create RegioneCSVTriples**' calls Karma script to generate triples based on the model.
- '**update last_triples**' updates last triple generation timestamp in *process_manager2*.

Training

Si ricordano i requisiti

- PC con almeno 6 GB di RAM (suggeriti: 8GB)
- Azioni che devono essere state effettuate prima del training:
  - Scaricare la macchina virtuale seguendo il LINK: http://www.disit.org/vmsdetl/VMSDETL-2016-v0-3.rar
  - Decomprimere la Macchina Virtuale in una directory e metterla in esecuzione con VMware player o workstation
- Approfondimenti: http://www.disit.org/6690

MySQL

The MySQL database provides the following features:

- High Performance and Scalability.
- Performance Schema for monitoring user- and application-level performance and resource consumption.
- SQL and NoSQL Access for performing complex queries and simple, fast Key Value operations.
- Platform Independence giving you the flexibility to develop and deploy on multiple operating systems.

- It provides phpMyAdmin (**localhost/phpmyadmin** in a browser) to do table operations, queries, …

HBase

- Apache HBase is the **Hadoop database**, a **distributed**, **scalable**, **big data store**.
- We can use it for random, real-time read/write access to your Big Data.
- This project's goal is the hosting of very large tables (billions of rows to millions of columns).
- Apache HBase is an open-source, distributed, versioned, non-relational database modeled after Google's Bigtable.

- We will create and modify HBase tables using **H-Rider**.

Virtual Machine DISIT

Comandi da terminale: http://www.disit.org/drupal/?q=node/6690

- Lanciare **Spoon**: (file system: home/programs/data-integration)
  - cd programs/data-integration ./spoon.sh
- Lanciare **MySQL**: (file system: home/programs/data-integration)
  - start: sudo /opt/lampp/lampp start
  - stop: sudo /opt/lampp/lampp stop
  - username=disit e password=ubuntu
- Lanciare **HBase** (propedeutico per HRider)
  - cd programs/data-integration
    - avvio: ./start-hbase.sh
    - stop: ./stop-hbase.sh
- Lanciare **HRider** (file system: home/programs)
  - Aprire: h-rider-1.0.3.0.jar con Java7
  - Inserire 'localhost' o IP macchina virtuale

Useful Elements

- MySQL table '*process_manager2*' contains for each process some information (link, file format, last ingestion date, …).
- MySQL table '*tbl_toponimo_BIS*' contains toponym codes for the most roads in Tuscany.

- When we save files, we create some folders (year, month, date, hour, minutes). It is important especially for real-time data to check differences between datasets.

1. Static Dataset (Taxi park)

- Run Ingestion phase (*Main.kjb*).
- Check the output in H-Rider ('**Taxi**' table).

key	Family1:D...	Family1:ID	Family1:P...	Family1:S...	Family1:a...	Family1:c...	Family1:c...	Family1:lo...	Family1:p...	Family1:si...	Family1:ti...
taxi1	ACCIAIUOLI	1	2	Piazza Nic...	Fri Jul 10 ...	11.22479...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi10	DONATELLO	10	8	Piazzale ...	Fri Jul 10 ...	11.26783...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi11	SASSO	11	6	Piazza del...	Fri Jul 10 ...	11.25772...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi12	FERRUCCI	12	5	Piazza Fr...	Fri Jul 10 ...	11.27190...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi13	FRANCIA	13	8	Piazza Fr...	Fri Jul 10 ...	11.29916...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi14	GIORGINI	14	4	Piazza Gi...	Fri Jul 10 ...	11.25021...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi15	INDIPEND...	15	5	Piazza del...	Fri Jul 10 ...	11.25263...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi16	LIBERTA'	16	4	Piazza del...	Fri Jul 10 ...	11.26219...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi17	PARTERRE	17	6	Via Mafal...	Fri Jul 10 ...	11.26390...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi18	PRATESE	18	3	Piazza Va...	Fri Jul 10 ...	11.19363...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi19	MANNELLI	19	7	Via Mann...	Fri Jul 10 ...	11.27831...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi2	ALBERTI	2	4	Piazza Le...	Fri Jul 10 ...	11.28082...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi20	MICHELA...	20	3	Piazzale ...	Fri Jul 10 ...	11.26421...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi21		21	8	Viale Giov...	Fri Jul 10 ...	11.24554...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi22	MUGELLO	22	4	Via Mugello	Fri Jul 10 ...	11.21827...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi23	CARISSIMI	23	6	Via di No...	Fri Jul 10 ...	11.22298...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi24	SAN GIOV...	24	7	Via dei P...	Fri Jul 10 ...	11.25430...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi25	ROMANA	25	8	Piazzale d...	Fri Jul 10 ...	11.24208...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi26	PUCCINI	26	4	Piazza Gi...	Fri Jul 10 ...	11.22825...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi27	SAN IACO...	27	3	Piazza di ...	Fri Jul 10 ...	11.23987...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi28	SAN MARCO	28	10	Piazza di ...	Fri Jul 10 ...	11.25852...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi29	SANTA CR...	29	6	Piazza di ...	Fri Jul 10 ...	11.26067...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi3	PIO FEDI	3	2	Via Pio Fedi	Fri Jul 10 ...	11.20450...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi30		30	3	Lungarno...	Fri Jul 10 ...	11.25340...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi31		31	3	Borgo Sa...	Fri Jul 10 ...	11.25268...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi32		32	2	Piazza dei...	Fri Jul 10 ...	11.25836...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi33	NOVELLA	33	6	Piazza di ...	Fri Jul 10 ...	11.24943...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi3339	AEROPOR...	3339	30	Via del T...	Fri Jul 10 ...	11.20115...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi3340		3340	6	Viale Ales...	Fri Jul 10 ...	11.22743...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi34	MAZZINI	34	8	Viale Ber...	Fri Jul 10 ...	11.27251...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi35	STAZIONE	35	34	Piazza del...	Fri Jul 10 ...	11.24906...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi36	VALFONDA	36	6	Viale Filip...	Fri Jul 10 ...	11.25243...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi37	TERZOLLE	37	6	Piazza del...	Fri Jul 10 ...	11.23446...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi38	UNITA'	38	3	Piazza del...	Fri Jul 10 ...	11.25071...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi39	VERGA	39	3	Viale Giov...	Fri Jul 10 ...	11.30002...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi4	BECCARIA	4	7	Piazza Ce...	Fri Jul 10 ...	11.27047...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi40	SESTESE	40	2	Piazza Ro...	Fri Jul 10 ...	11.21873...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi41	RIFREDI	41	6	Via dello ...	Fri Jul 10 ...	11.23675...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi42	VESPUCCI	42	8	Il Prato	Fri Jul 10 ...	11.24263...	FI	FIRENZE	Taxi_kmz	taxi	1436540...
taxi43	VETTORI	43	5	Piazza Pie...	Fri Jul 10 ...	11.23393...	FI	FIRENZE	Taxi_kmz	taxi	1436540...

Via dei Pecori

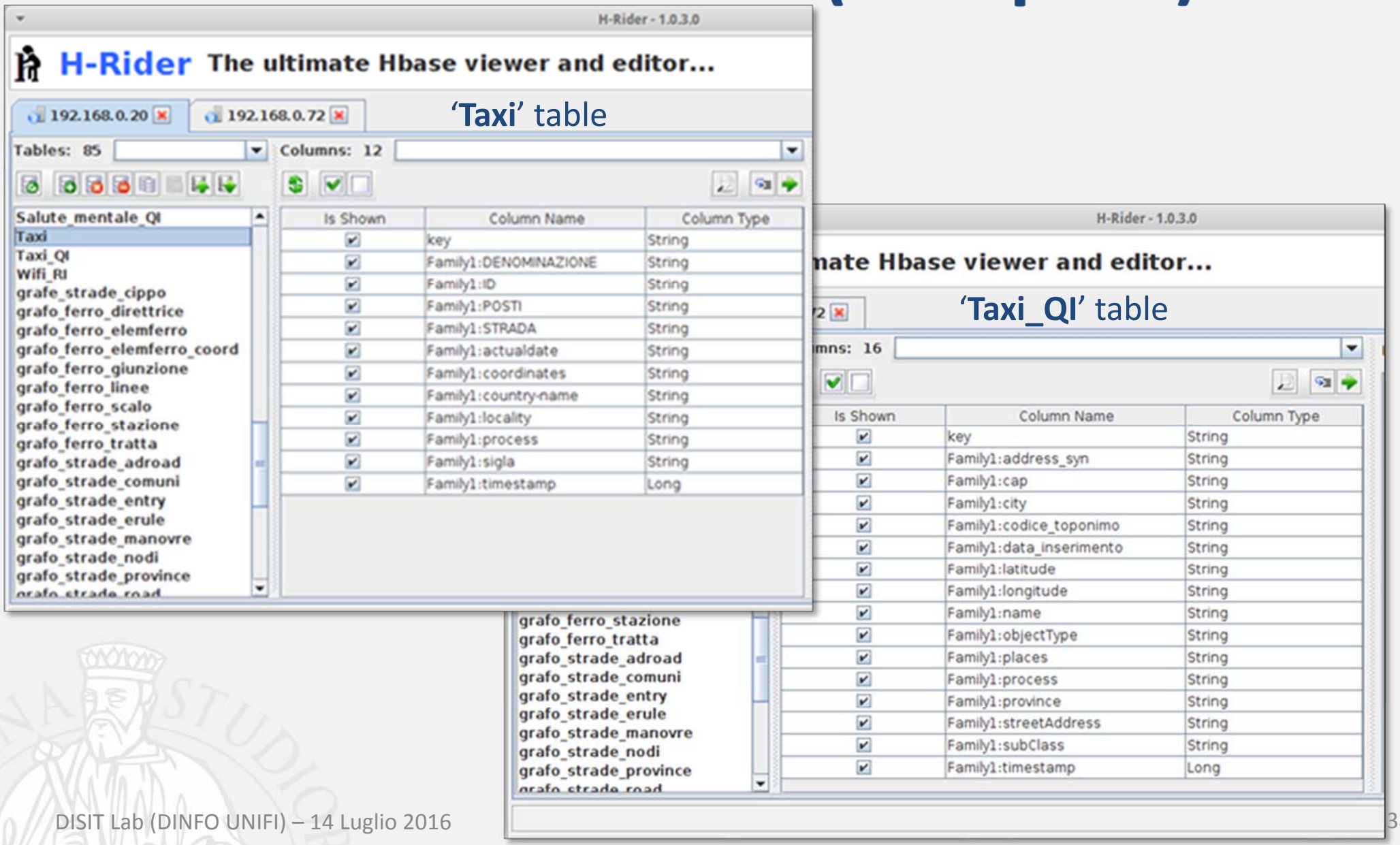
1. Static Dataset (Taxi park)

- Run Quality Improvement code (*Data_QI.kjb*).
- Check the output in '**Taxi_QI**' table (fields added, data normalization, …).

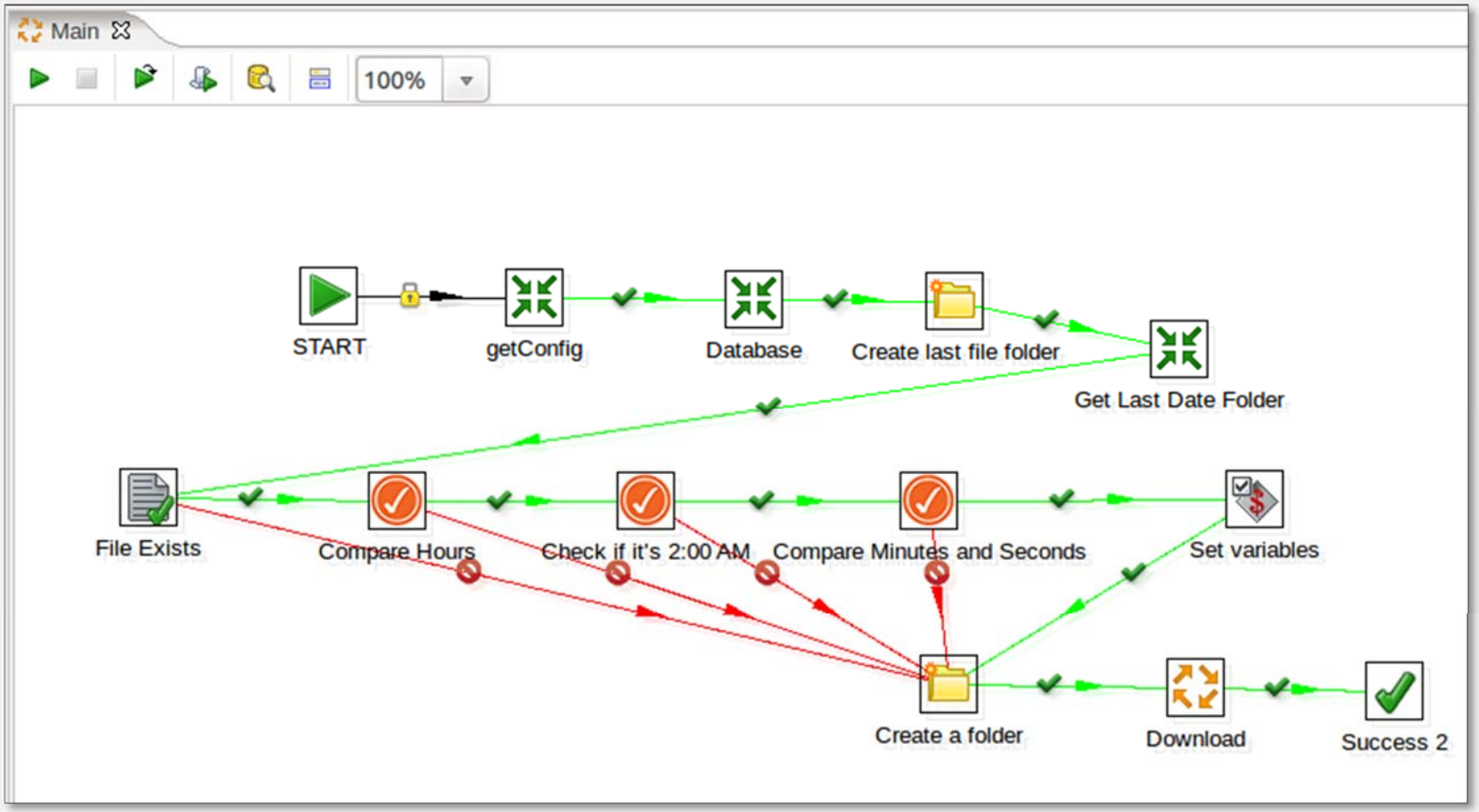
key	Family1:address_syn	Family1:cap	Family1:city	Family1:c...	Family1:d...	Family1:la...	Family1:lo...	Family1:n...	Family1:o...	Family1:pl...	Family1:p...
taxi1	PIAZZA NICCOLÒ ACCIAIOLI	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.73572...	11.22479...	Piazzola t...	taxi	Posti: 2	Taxi_kmz
taxi10	PIAZZALE DONATELLO	50132	FIRENZE	RT04801...	Fri Jul 10 ...	43.77833...	11.26783...	Piazzola t...	taxi	Posti: 8	Taxi_kmz
taxi11	PIAZZA DEL DUOMO	50122	FIRENZE	RT04801...	Fri Jul 10 ...	43.77258...	11.25772...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi12	PIAZZA FRANCESCO FER...	50126	FIRENZE	RT04801...	Fri Jul 10 ...	43.76329...	11.27190...	Piazzola t...	taxi	Posti: 5	Taxi_kmz
taxi13	PIAZZA FRANCIA	50126	FIRENZE	RT04801...	Fri Jul 10 ...	43.75806...	11.29916...	Piazzola t...	taxi	Posti: 8	Taxi_kmz
taxi14	PIAZZA GIOVANBATTISTA...	50134	FIRENZE	RT04801...	Fri Jul 10 ...	43.79213...	11.25021...	Piazzola t...	taxi	Posti: 4	Taxi_kmz
taxi15	PIAZZA DELL'INDIPENDE...	50129	FIRENZE	RT04801...	Fri Jul 10 ...	43.77946...	11.25263...	Piazzola t...	taxi	Posti: 5	Taxi_kmz
taxi16	PIAZZA DELLA LIBERTA'	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.78323...	11.26219...	Piazzola t...	taxi	Posti: 4	Taxi_kmz
taxi17	VIA MAFALDA DI SAVOIA	50129	FIRENZE	RT04801...	Fri Jul 10 ...	43.78630...	11.26390...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi18	PIAZZA VASCO MAGRINI	50145	FIRENZE	RT04801...	Fri Jul 10 ...	43.79908...	11.19363...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi19	VIA MANNELLI	50132	FIRENZE	RT04801...	Fri Jul 10 ...	43.77520...	11.27831...	Piazzola t...	taxi	Posti: 7	Taxi_kmz
taxi2	PIAZZA LEON BATTISTA ...	50136	FIRENZE	RT04801...	Fri Jul 10 ...	43.76948...	11.28082...	Piazzola t...	taxi	Posti: 4	Taxi_kmz
taxi20	PIAZZALE MICHELANGIOLO	50125	FIRENZE	RT04801...	Fri Jul 10 ...	43.76241...	11.26421...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi21	VIALE GIOVAN BATTISTA ...	50134	FIRENZE	RT04801...	Fri Jul 10 ...	43.80321...	11.24554...	Piazzola t...	taxi	Posti: 8	Taxi_kmz
taxi22	VIA MUGELLO	50127	FIRENZE	RT04801...	Fri Jul 10 ...	43.79752...	11.21827...	Piazzola t...	taxi	Posti: 4	Taxi_kmz
taxi23	VIA DI NOVOLI	50127	FIRENZE	RT04801...	Fri Jul 10 ...	43.79206...	11.22298...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi24	VIA DEI PECORI	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.77275...	11.25430...	Piazzola t...	taxi	Posti: 7	Taxi_kmz
taxi25	PIAZZALE DI PORTA ROM...	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.76010...	11.24208...	Piazzola t...	taxi	Posti: 8	Taxi_kmz
taxi26	PIAZZA GIACOMO PUCCINI	50144	FIRENZE	RT04801...	Fri Jul 10 ...	43.78556...	11.22825...	Piazzola t...	taxi	Posti: 4	Taxi_kmz
taxi27	PIAZZA DI SAN IACOPINO	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.78382...	11.23987...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi28	PIAZZA DI SAN MARCO	50121	FIRENZE	RT04801...	Fri Jul 10 ...	43.77800...	11.25852...	Piazzola t...	taxi	Posti: 10	Taxi_kmz
taxi29	PIAZZA DI SANTA CROCE	50122	FIRENZE	RT04801...	Fri Jul 10 ...	43.76907...	11.26067...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi3	VIA PIO FEDI	50142	FIRENZE	RT04801...	Fri Jul 10 ...	43.78168...	11.20450...	Piazzola t...	taxi	Posti: 2	Taxi_kmz
taxi30	LUNGARNO DEGLI ACCIAI...	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.76852...	11.25340...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi31	BORGO SAN IACOPO	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.76748...	11.25268...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi32	PIAZZA DEI MOZZI	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.76530...	11.25836...	Piazzola t...	taxi	Posti: 2	Taxi_kmz
taxi33	PIAZZA DI SANTA MARIA ...	50123	FIRENZE	RT04801...	Fri Jul 10 ...	43.77303...	11.24943...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi3339	VIA DEL TERMINE	50127	FIRENZE	RT04801...	Fri Jul 10 ...	43.80177...	11.20115...	Piazzola t...	taxi	Posti: 30	Taxi_kmz
taxi3340	VIALE ALESSANDRO GUID...	50127	FIRENZE	RT04801...	Fri Jul 10 ...	43.79554...	11.22743...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi34	VIALE BERNARDO SEGNI	50132	FIRENZE	RT04801...	Fri Jul 10 ...	43.77495...	11.27251...	Piazzola t...	taxi	Posti: 8	Taxi_kmz
taxi35	PIAZZA DELLA STAZIONE	50123	FIRENZE	RT04801...	Fri Jul 10 ...	43.77625...	11.24906...	Piazzola t...	taxi	Posti: 34	Taxi_kmz
taxi36	VIALE FILIPPO STROZZI	50129	FIRENZE	RT04801...	Fri Jul 10 ...	43.78306...	11.25243...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi37	PIAZZA DEL TERZOLLE	50127	FIRENZE	RT04801...	Fri Jul 10 ...	43.79406...	11.23446...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi38	PIAZZA DELL'UNITA' ITALI...	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.77516...	11.25071...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi39	VIALE GIOVANNI VERGA	50135	FIRENZE	RT04801...	Fri Jul 10 ...	43.77859...	11.30002...	Piazzola t...	taxi	Posti: 3	Taxi_kmz
taxi4	PIAZZA CESARE BECCARIA	50121	FIRENZE	RT04801...	Fri Jul 10 ...	43.77039...	11.27047...	Piazzola t...	taxi	Posti: 7	Taxi_kmz
taxi40	PIAZZA ROBERTO DAVID...	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.82008...	11.21873...	Piazzola t...	taxi	Posti: 2	Taxi_kmz
taxi41	VIA DELLO STECCUTO	50141	FIRENZE	RT04801...	Fri Jul 10 ...	43.80032...	11.23675...	Piazzola t...	taxi	Posti: 6	Taxi_kmz
taxi42	PIAZZALE DI PORTA AL P...	50100	FIRENZE	RT04801...	Fri Jul 10 ...	43.77471...	11.24263...	Piazzola t...	taxi	Posti: 8	Taxi_kmz
taxi43	PIAZZA PIER VETTORI	50143	FIRENZE	RT04801...	Fri Jul 10 ...	43.77158...	11.23393...	Piazzola t...	taxi	Posti: 5	Taxi_kmz

PIAZZA LEON BATTISTA ALBERTI

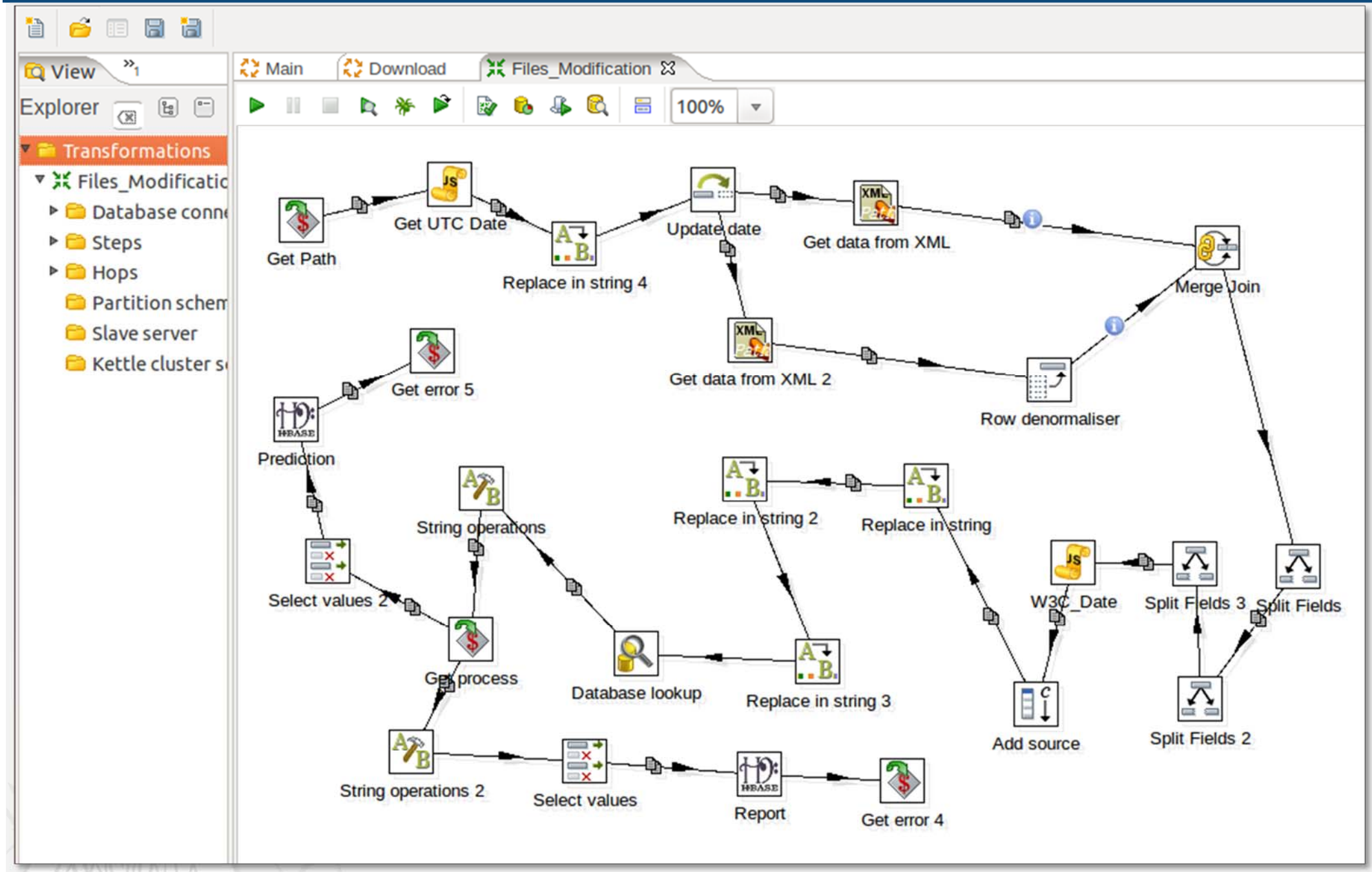
1. Static Dataset (Taxi park)



2. Real-Time Dataset (Weather)



This Job is the same seen before for ingestion. The only difference is that it checks if there are two different datasets at the same hour.


View
Explorer
Transformations
Files_Modificatic
Database conn
Steps
Hops
Partition schem
Slave server
Kettle cluster s
Main
Download
Files_Modification
100%
Get Path
Get UTC Date
Replace in string 4
Update date
Get data from XML
Merge Join
Get data from XML 2
Row denormaliser
Get error 5
Prediction
String operations
Replace in string 2
Replace in string
Select values 2
Get process
Database lookup
Replace in string 3
W3C_Date
Split Fields 3
Split Fields
Add source
Split Fields 2
String operations 2
Select values
Report
Get error 4

2. Real Time Dataset (Weather)

- Create two HBase tables (and their mappings in Spoon) to complete ingestion phase
  - *table name*: '**weatherPrediction**', *key*: '**predictionKey**' field
  - *table name*: '**weatherReport**', *key*: '**reportKey**' field

#	Alias	Key	Column family	Column name	Type	Indexed values
1	predictionKey	Y			String	
2	data_inserimento	N	Family1	data_inserimento	String	
3	descr	N	Family1	descr	String	
4	hour	N	Family1	hour	String	
5	process	N	Family1	process	String	
6	quota_neve	N	Family1	quota_neve	String	
7	reportKey	N	Family1	reportKey	String	
8	temp	N	Family1	temp	String	
9	temp_max	N	Family1	temp_max	String	
10	temp_min	N	Family1	temp_min	String	
11	temp_perc	N	Family1	temp_perc	String	
12	um	N	Family1	um	String	
13	uv	N	Family1	uv	String	
14	week_day	N	Family1	week_day	String	
15	wind	N	Family1	wind	String	

#	Alias	Key	Column family	Column name	Type	Indexed values
1	reportKey	Y			String	
2	Istat	N	Family1	Istat	String	
3	W3C_Date	N	Family1	W3C_Date	String	
4	comune	N	Family1	comune	String	
5	data_inserimento	N	Family1	data_inserimento	String	
6	fase	N	Family1	fase	String	
7	luna_sorge	N	Family1	luna_sorge	String	
8	luna_tramonta	N	Family1	luna_tramonta	String	
9	ora_altezza	N	Family1	ora_altezza	String	
10	process	N	Family1	process	String	
11	sole_altezza	N	Family1	sole_altezza	String	
12	sole_sorge	N	Family1	sole_sorge	String	
13	sole_tramonta	N	Family1	sole_tramonta	String	
14	source	N	Family1	source	String	
15	time_ms	N	Family1	time_ms	String	

Km4City Technologies: data ingestion and mining - Develop ETL processes –

14 Luglio 2016

DISIT Lab, Dipartimento di Ingegneria dell'Informazione, DINFO
Università degli Studi di Firenze
Via S. Marta 3, 50139, Firenze, Italy
Tel: +39-055-2758515, fax: +39-055-2758570
http://www.disit.dinfo.unifi.it ***alias*** http://www.disit.org

Prof. Paolo Nesi, paolo.nesi@unifi.it
Pierfrancesco Bellini, pierfrancesco.bellini@unifi.it
Michela paolucci, michela.paolucci@unifi.it
Simone Panicucci, simone.panicucci@stud.unifi.it