Why Table Understanding Matters

Practical solutions to real-life problems

Vincenzo Cutrona

IT Consultant | R&D OpenLab @ Corvallis SRL (Tinexta Spa Group)
Ph.D. Student | University of Milano - Bicocca

corvallis PROCESS & SOLUTIONS

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Why Table Understanding?

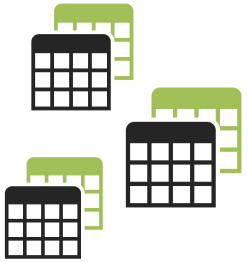
Glossary

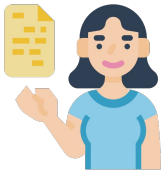
CEA: Cell to Entity
CTA: Column to Type
CPA: Column Pair to Property

STI: CEA + CTA + CPA

- ★ Sales Reports
- ★ Pricing Databases
- ★ and more...

Tables Extraction





- ★ Data Practitioners
- ★ Business People

KG construction

Tabular Data | Ontology | New KG

Data Enrichment

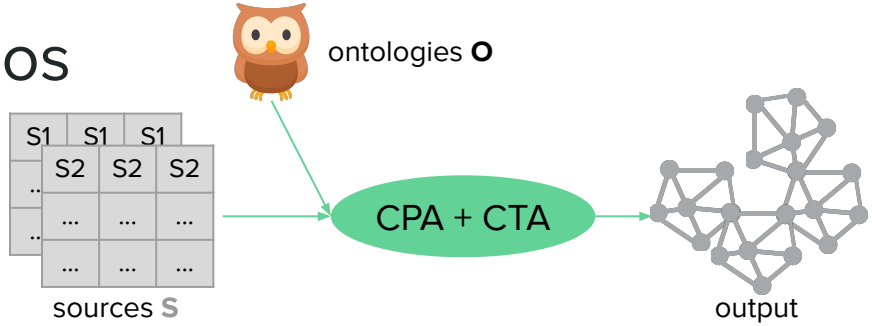
Tabular Data | Existing KGs | Extended Tabular Data

KG population

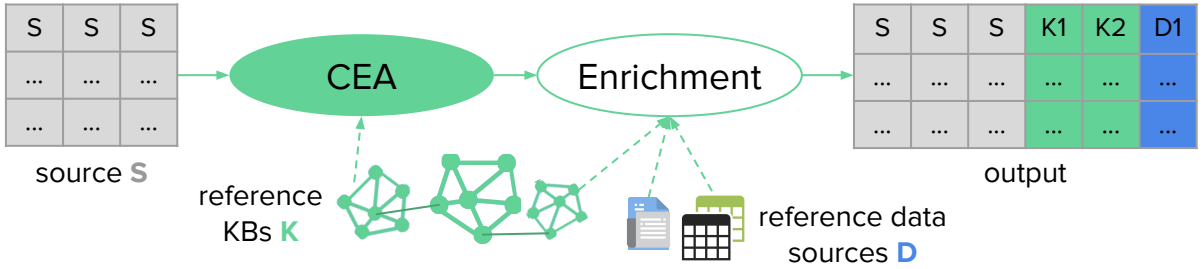
Existing KG | Tabular Data | Extended KG

3 Common Application Scenarios

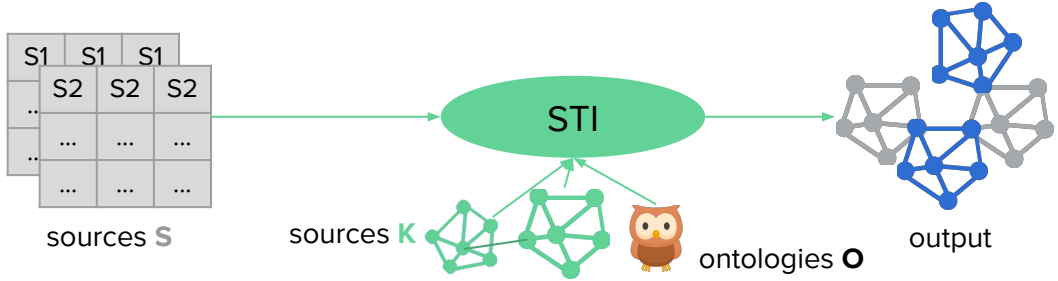
KG construction #1



(Tabular) Data Enrichment #2



KG Population #3



Discover Business Data

#1

Challenge Collecting and **aggregating** information about a **business entity** from **public sources**

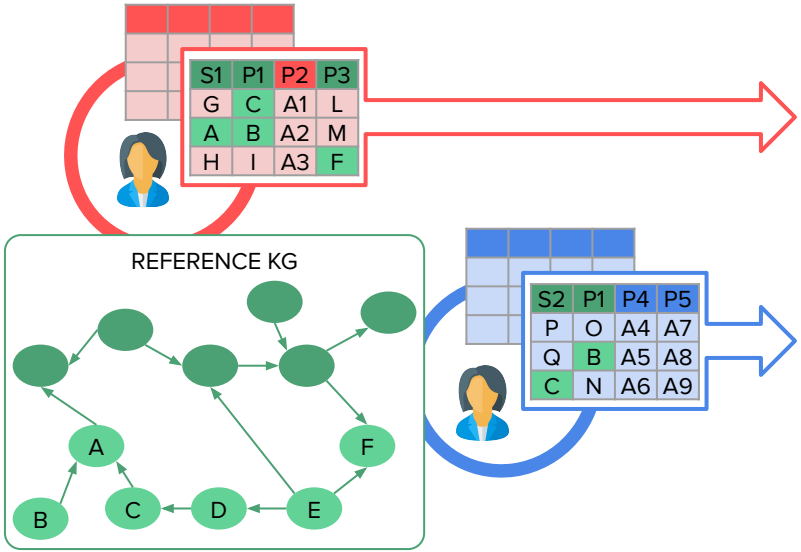
Goal Providing a **data marketplace** based on a **highly interconnected graph** of company-related information (a.k.a. *the business KG*)

Scenario #1: KG Construction

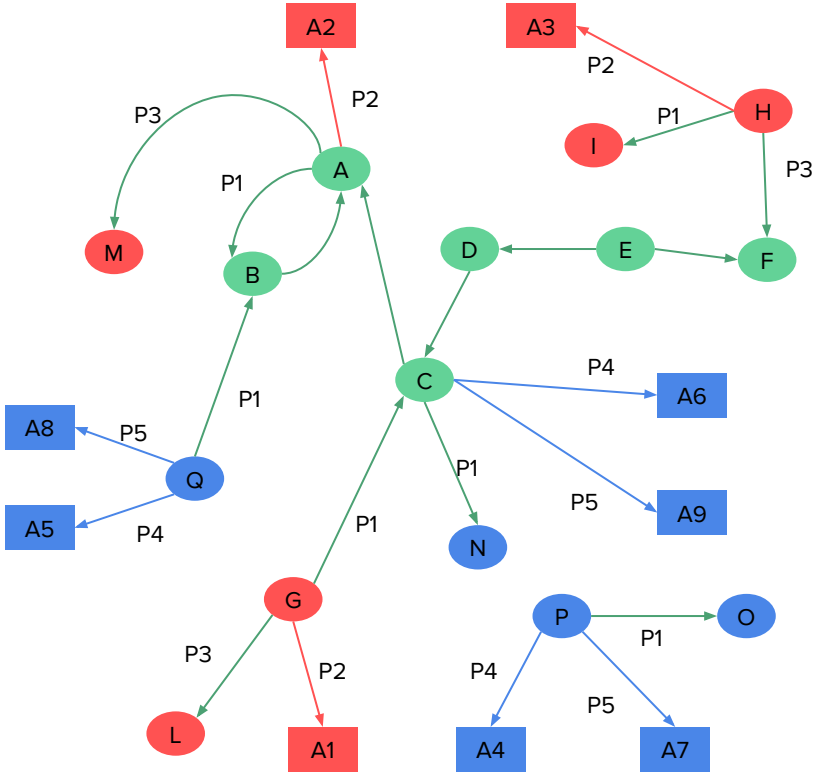
- Table schema to KG schema
  - Inferred with CTA (e.g., ColNet) and CPA
  - Manually provided (mapping languages like RML)
    - Human in the loop is acceptable, because the number of columns is usually under control
- Records assumed to be unique
  - **NO deduplication**: rows representing the same object (i.e., same ID), are **collapsed into a single entity** (i.e., same URI)
- Main challenges:
  - **Help the user** in defining the schema mapping (e.g., type and property suggestion)
  - Speed up the **triple construction process** (separable problem -> high parallelism)

KG Construction in euBusinessGraph

- CTA and CPA with user in the loop
- Optional: CEA to reuse existing entities

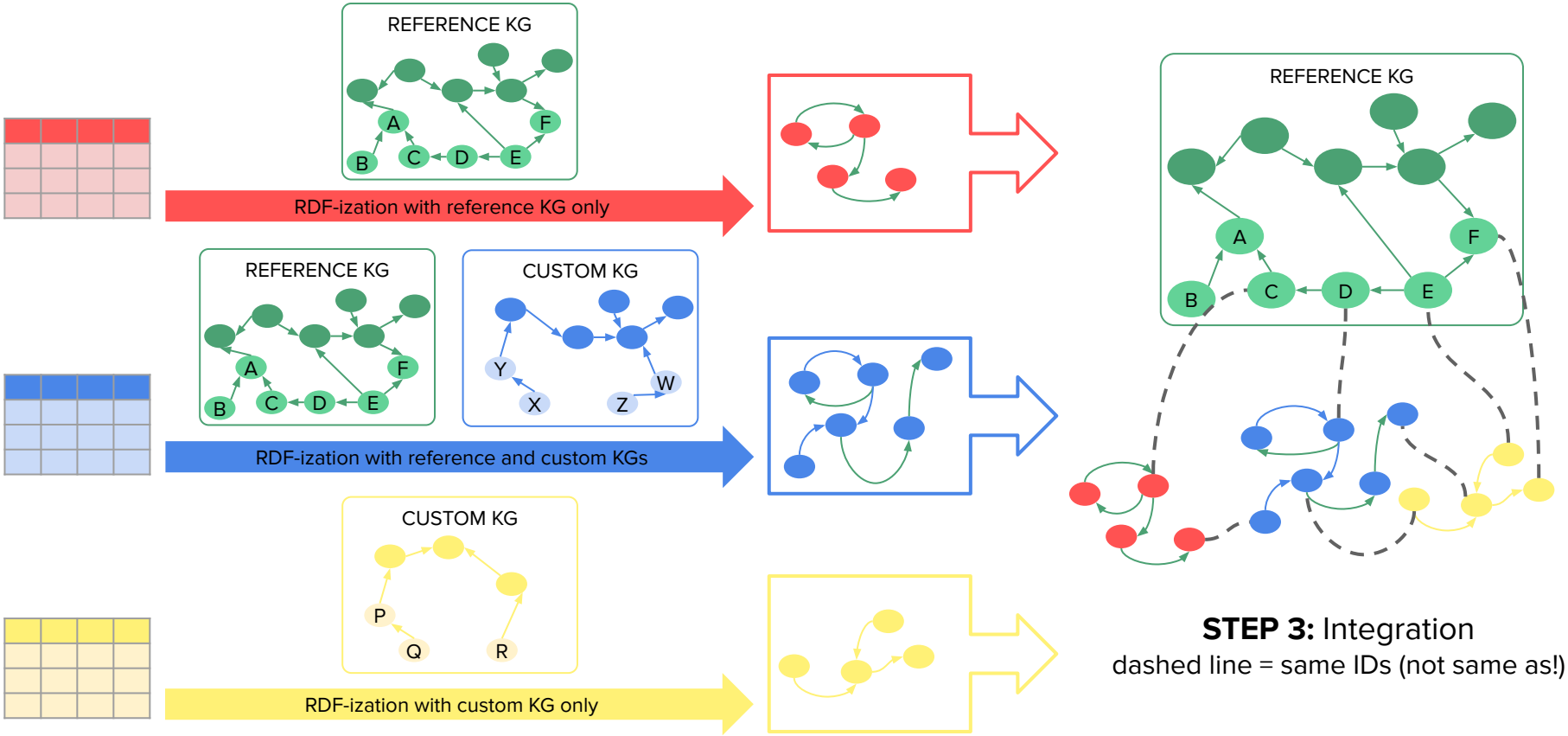


STEP 1: CTA + CPA (+ CEA, optional)



STEP 2: RDF-ization

KG Construction in euBusinessGraph (cont.)

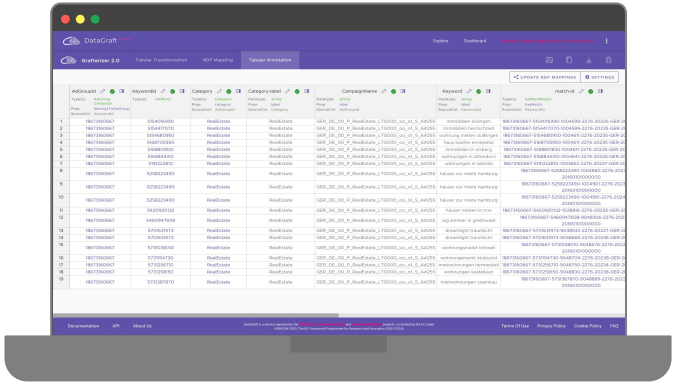


STEP 3: Integration

dashed line = same IDs (not same as!)

Supporting KG Construction with DataGraft

- **Common data model**: the euBusinessGraph ontology
- **DataGraft**: a “one-stop-shop for *hosted data management*” with two integrated tools:
  - **Grafterizer 2.0**
    - Tabular Transformation
    - RDF Mapping (with **Grafter**)
  - **ASIA**
    - Tabular annotation (CTA and CPA)
    - Vocabulary suggestion with **ABSTAT**
- **Big Data Environment**: to execute the RDF mapping on massive amount of data

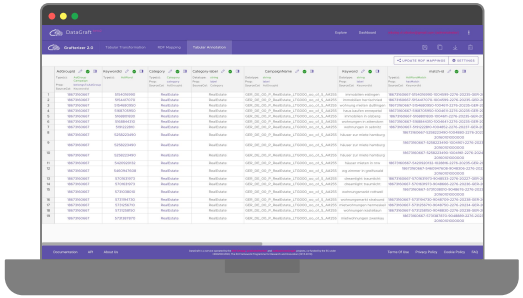


Different outputs:

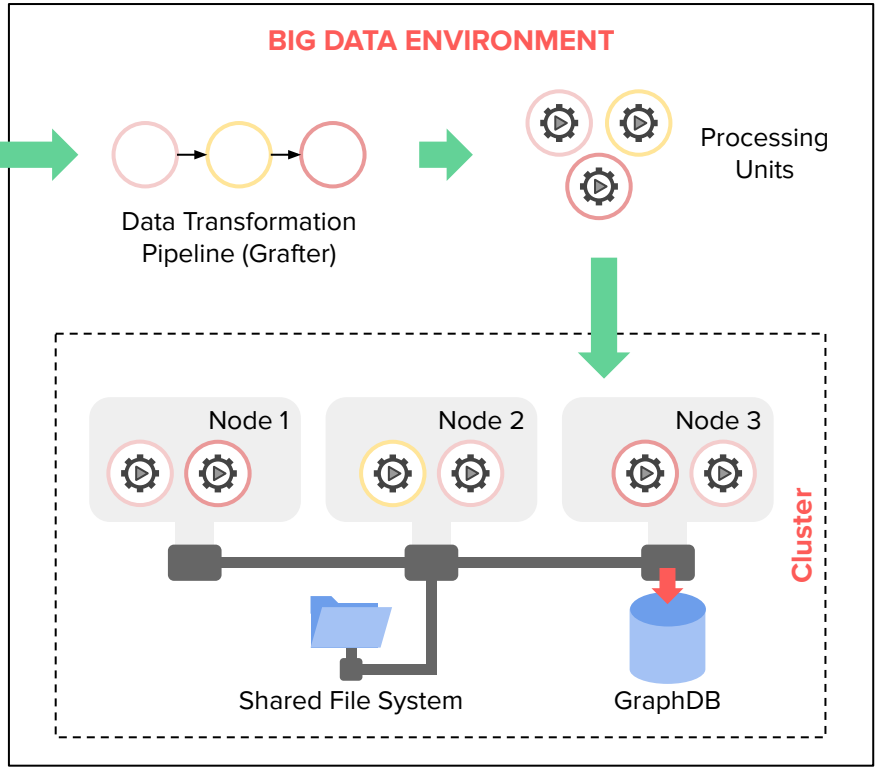
RDF data (small data)

Executable file (massive data)

Supporting KG Construction with DataGraft (cont.)



Executable File
(RDF Mapping)



- RDF Mapping to Transformation Pipeline (using **Grafter**)
- Pipeline split into **Processing Units**
- Processing Units are deployed to **execution nodes** (parallel execution)

KG Construction - Results from euBusinessGraph

Project results

- Release a **marketplace** to publish company data
  http://marketplace.businessgraph.io
  - 4 data providers (**~1.6 billion triples**)
- **6 business cases** supported

DBpedia

DBpedia 2016-10 (en) contains 1.7 billion triples

Additional results

- SIRENE challenge: publishing the official database of French enterprises and establishments
  - More than 10 million units (**16 GB)**
  - The resulting RDF dataset counts ~**3 billion triples** (.nt)

Supporting Event and Weather-based Data Analytics

#2

Challenge Data collected by individual companies provide a **partial view** on the customer journey

Goal Provide tools to develop **analytical services** that **consider events** that impact on customer decisions (a.k.a. *the EW-Shopp Toolkit*)

Scenario #2: Enrich Tabular Data

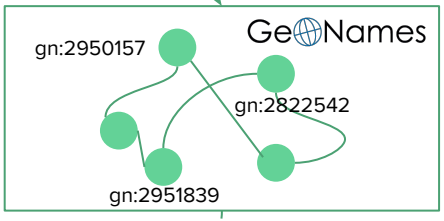
- CEA is the only key task
  - CTA and CPA may support CEA (e.g., CTA used to filter CEA results)
- Main challenges:
  - Deal with **multi-valued properties** in the enrichment
    - Tables are not ready to include such data
  - CEA on “custom” KGs
    - A few properties are assumed to exist in STI algorithms (e.g., rdfs:label)
  - **Speed up CEA**
    - Property usually overlooked (not evaluated in recent challenges)
    - Crucial in industry (how to enrich a large dataset with **millions of records**)

Tabular Data Enrichment in EW-Shopp

Step 1
Business Data

KEYWORD	#im	REGION	Date
194906	64	Thuringia	11/03/2017
517827	50	Bavaria	12/03/2017
459143	42	Berlin	12/03/2017
891139	36	Bavaria	11/03/2017

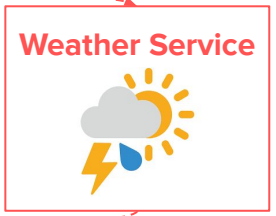
CEA



Step 2
CEA (from AdWords labels to GeoNames IDs)

KEYWORD	#im	REGION	Date	geoId.
194906	64	Thuringia	11/03/2017	gn:2822542
517827	50	Bavaria	12/03/2017	gn:2951839
459143	42	Berlin	12/03/2017	gn:2950157
891139	36	Bavaria	11/03/2017	gn:2951839

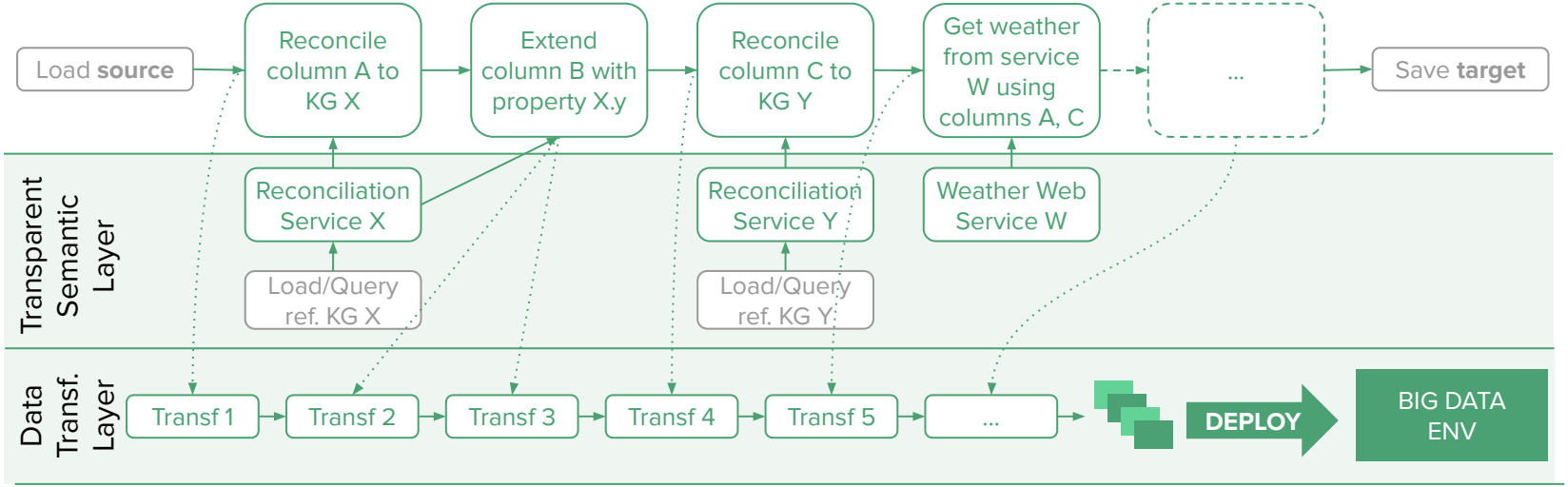
EXTENSION



Step 3
Extension from service (from <GeoNames ID, Date> to temperature)

KEYWORD	#im	REGION	Date	geoId.	C°/+0	C°/+1
194906	64	Thuringia	11/03/2017	gn:2822542	18	20
517827	50	Bavaria	12/03/2017	gn:2951839	17	19
459143	42	Berlin	12/03/2017	gn:2950157	17	20
891139	36	Bavaria	11/03/2017	gn:2951839	19	23

Table Annotation for Data Enrichment

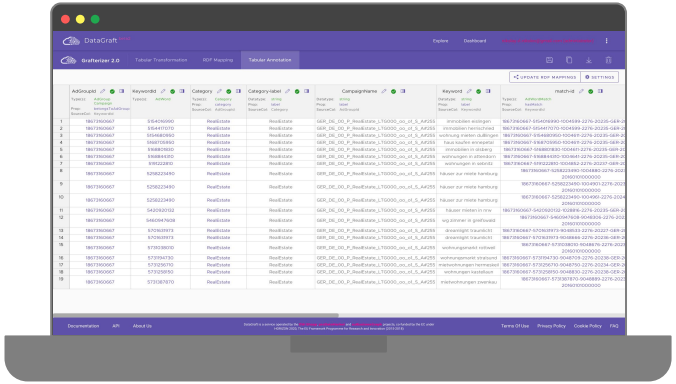


Declarative and interactive approach

- Each step **annotates the table** and enables new actions (Data Discovery)
- Annotations steps are converted **to data transformation steps**
- Integration under user's control (intermediate results)

Supporting Tabular Data Enrichment in EW-Shopp

- **ASIA**
  - Tabular annotation (CEA, CTA and CPA)
  - Different reconciliation service for CEA
  - Different extension service (weather and events)
    - Different aggregation strategies!
- **STI to Data Transformation Pipeline**
  - Allow us to **reuse DataGraft/Grafterizer 2.0**
- **Big Data Environment**: to execute the data transformations on massive amount of data
  - Custom deployment to support several reconciliation services and speed up the process

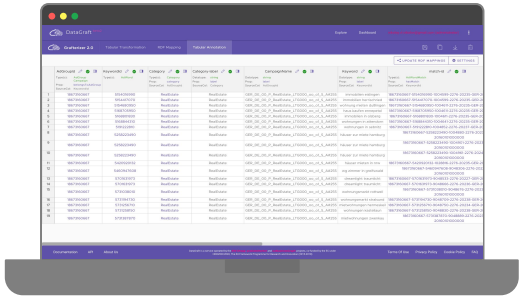


Different outputs:

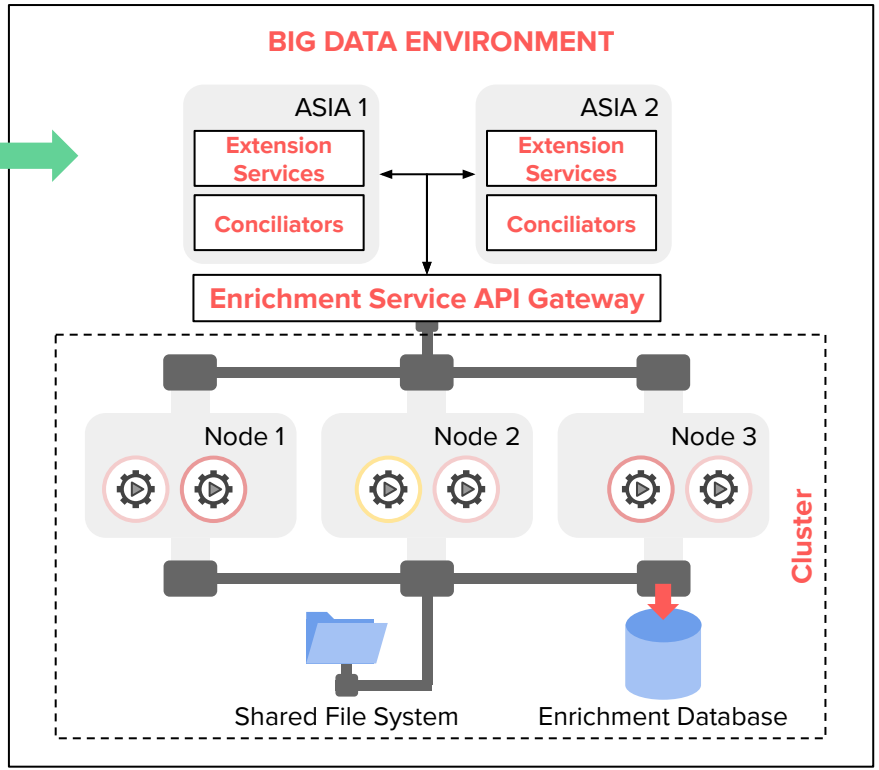
CSV data (small data)

Executable file (massive data)

Supporting Tabular Data Enrichment in EW-Shopp (cont.)



Executable File
(Data Transformations)



- Enrichment steps to Transformation Pipeline (using **Grafter**)
- Pipeline split into **Processing Units,** deployed to **execution nodes**
- ASIA Services (for CEA and extensions) deployed next to the data

Tabular Data Enrichment - Results from EW-Shopp

Project results

- Release of the EW-Shopp Data Preparation and Enrichment tool (part of the EW-Shopp Toolkit), with comes with different features:
  - Tabular Data Transformation (Grafterizer 2.0)
  - Tabular Data Annotation and Extension (ASIA)
- **4 business cases** supported
  - The EW-Shopp Data Preparation and Enrichment tool used to transform and enrich digital marketing campaign data with geographical and weather-related data (~100 GB)

Journey Experience
Data Improve

#3

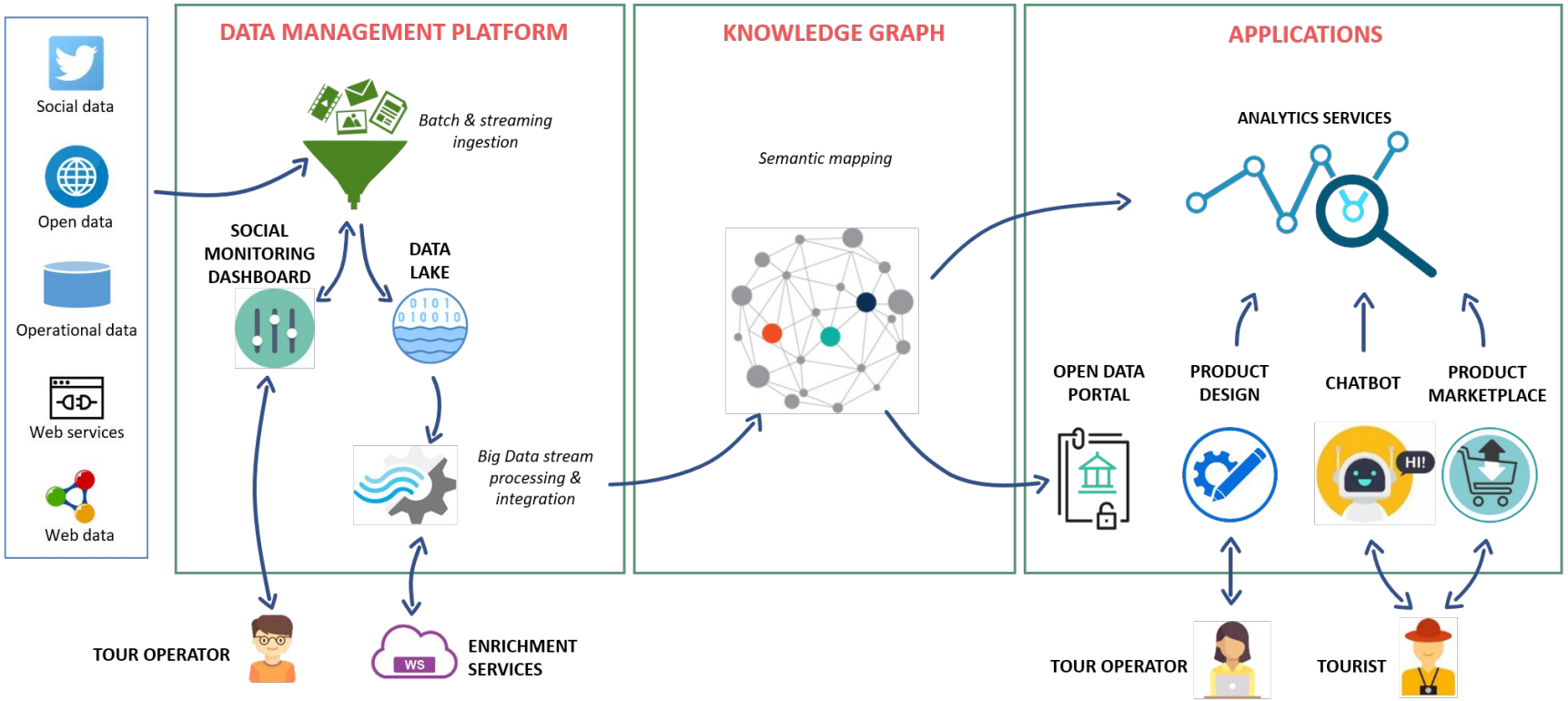
Challenge Tourism relevant data are scattered (high variety, due to different providers, different formats) and continuously updated

Goal Provide a Data Manipulation Platform to support the **Knowledge Extraction** from **Tourism Big Data** (a.k.a. *the DMP*)

Scenario #3: KG Population

- Deduplication **is fundamental**!
  - The same object may have different IDs in different sources
- Incremental update?
  - Typical approach: creating a new versioned KG from scratch (e.g., DBpedia 2014, 2016)
  - Not suitable in industry (imagine a graph containing daily bank transactions)
- Deal with “different” sources, e.g., **streams** (infinite sequence of rows)
  - “Record” understanding? (a table with only one row)

KG Population in JEDI

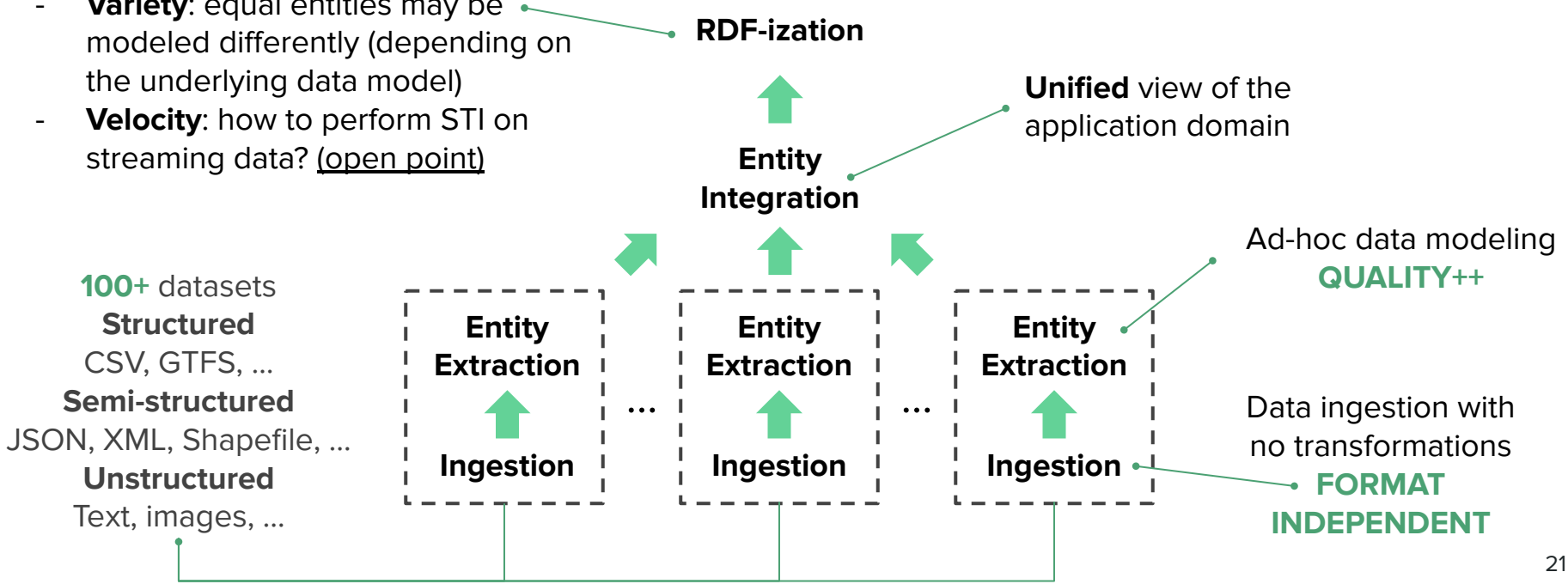


KG Population in JEDI (cont.)

STI for KG Population

Done on “well-formed” entities but:

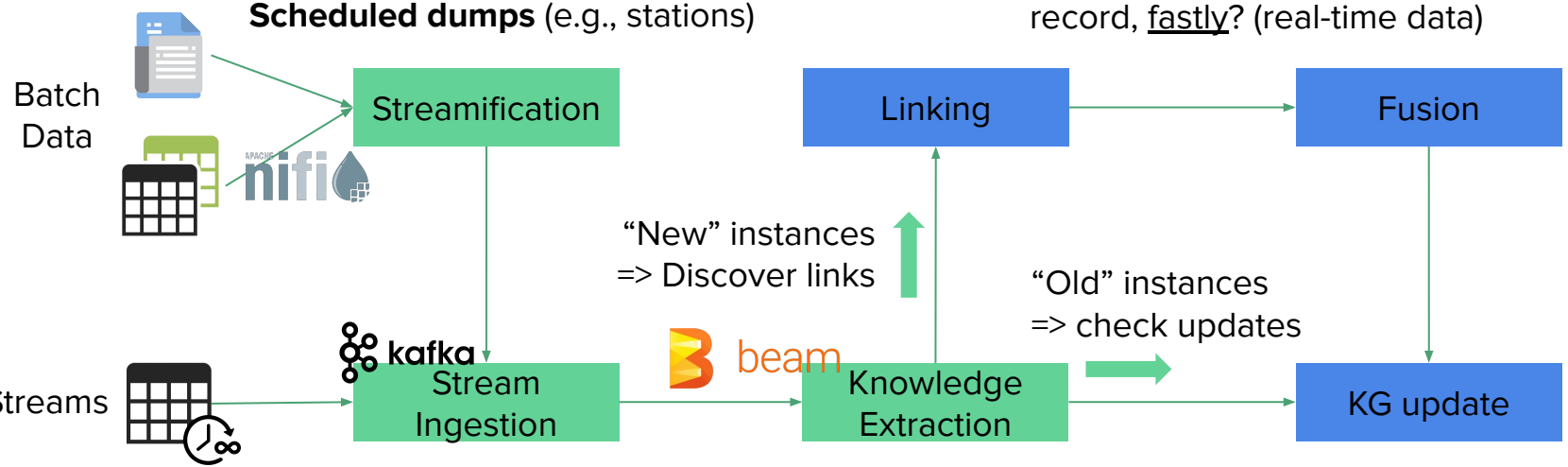
- **Variety**: equal entities may be modeled differently (depending on the underlying data model)
- **Velocity**: how to perform STI on streaming data? (open point)



Supporting KG population in JEDI

Batch data converted into streams
High-frequency data (e.g., weather)
Scheduled dumps (e.g., stations)

CHALLENGES (#1)
One (record) vs. many (entities)
=> how to perform STI on a single record, fastly? (real-time data)



Single ingestion point
=> only streams
(**sequences of records**)

CHALLENGES (#2)
Streaming vs. static (KG)
=> How to populate the KG consistently?
(e.g., remove links)

= Next steps

Q&A

vincenzo.cutrona@corvallis.it
vincenzo.cutrona@unimib.it

vcutrona9 vincenzo-cutrona

Find out more!

EW-Shopp
https://www.ew-shopp.eu/

euBusinessGraph
https://www.eubusinessgraph.eu/

JEDI
https://corvallis.it/open-lab/progetti/ (ITA only)

Icons made by ultimatearm, prettycons, and Freepik from Flaticon