Australian National University

COMP3430 / 8430
Data wrangling

Lecture 10: Overview of data integration
(Lecturer: Peter Christen)

Lecture outline

- What is data integration
- Example of data integration
- Main data integration tasks
- Example application: Woo by Yahoo!

What is data integration? (1)

- In many organisations there is an increasing quest to integrate data from different parts of the organisation, or to enrich data with external data
  - Many applications require data from various sources

- Data integration is the process of integrating data from multiple sources to obtain a single view over all sources
  - To enable answering queries using the combined information
  - Integration can be **virtual** (keep data at sources) or **physical** (copying the integrated data into a database or data warehouse)

What is data integration? (2)

- Main reasons for data integration:
  - Reuse data from various legacy databases and systems (often held as data silos)
  - Reconcile the different points of views adopted by different systems in an organisation
  - Integrate external data (such as social network data, information from statistical agencies, etc.)

- A major challenge of data integration is heterogeneity
  - At different levels: source type, schemas, data types, data values, semantics, etc.

Example (1)

Name	Address	Phone	Age	Gender
John Smith	26 Miller St, O'Conner A.C.T.	6127 8042	42	M
Miss Mary Miller	4 Main Road Dixon ACT 2060	01 2345 6789	21	F
Dr Meyer, Paul	5/42 MillAve, Sydeny 2000	61 (0) 4 643 765	57	U

Title	FName	LName	Street	Suburb	Postcode	State	Sex	DoB
Mr	John	Smith	26 Miller Street	O'Connor	2602	ACT	0	12/03/1975
Ms	Marie	Miller	4 Main Road	Dickson	2602	ACT	1	23/12/1995
Dr	Paul	Meyer	5 Mill Avenue	Ryde	2112	NSW	0	4/10/1957
Mr	Paul	Meier	42 Miller Avenue	Manly	2095	NSW	0	10/08/1960

Example (2) – Schema mapping

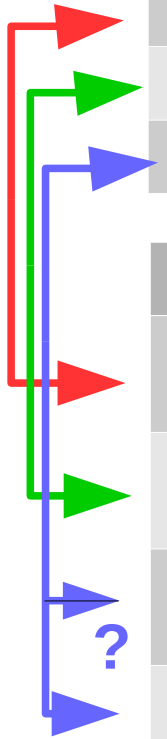
Name	Address	Phone	Age	Gender
John Smith	26 Miller St, O'Conner A.C.T.	6127 8042	42	M
Miss Mary Miller	4 Main Road Dixon ACT 2060	01 2345 6789	21	F
Dr Meyer, Paul	5/42 MillAve, Sydeny 2000	61 (0) 4 643 765	57	U

Title	FName	LName	Street	Suburb	Postcode	State	Sex	DoB
Mr	John	Smith	26 Miller Street	O'Connor	2602	ACT	0	12/03/1975
Ms	Marie	Miller	4 Main Road	Dickson	2602	ACT	1	23/12/1995
Dr	Paul	Meyer	5 Mill Avenue	Ryde	2112	NSW	0	4/10/1957
Mr	Paul	Meier	42 Miller Avenue	Manly	2095	NSW	0	10/08/1960

Example (3) – Record linkage

Name	Address	Phone	Age	Gender
John Smith	26 Miller St, O'Conner A.C.T.	6127 8042	42	M
Miss Mary Miller	4 Main Road Dixon ACT 2060	01 2345 6789	21	F
Dr Meyer, Paul	5/42 MillAve, Sydeny 2000	61 (0) 4 643 765	57	U

Title	FName	LName	Street	Suburb	Postcode	State	Sex	DoB
Mr	John	Smith	26 Miller Street	O'Connor	2602	ACT	0	12/03/1975
Ms	Marie	Miller	4 Main Road	Dickson	2602	ACT	1	23/12/1995
Dr	Paul	Meyer	5 Mill Avenue	Ryde	2112	NSW	0	4/10/1957
Mr	Paul	Meier	42 Miller Avenue	Manly	2095	NSW	0	10/08/1960



Example (4) – Data fusion

Name	Address	Phone	Age	Gender
John Smith	26 Miller St, O'Conner A.C.T.	6127 8042	42	M
Miss Mary Miller	4 Main Road Dixon ACT 2060	01 2345 6789	21	F
Dr Meyer, Paul	5/42 MillAve, Sydeny 2000	61 (0) 4 643 765	57	U

?

?

?

Title	FName	LName	Street	Suburb	Postcode	State	Sex	DoB
Mr	John	Smith	26 Miller Street	O'Connor	2602	ACT	0	12/03/1975
Ms	Marie	Miller	4 Main Road	Dickson	2602	ACT	1	23/12/1995
Dr	Paul	Meyer	5 Mill Avenue	Ryde	2112	NSW	0	4/10/1957
Mr	Paul	Meier	42 Miller Avenue	Manly	2095	NSW	0	10/08/1960

Three main tasks of data integration

- Schema mapping and matching
  - Identify which attributes or attribute sets across database tables contain the same type of information
- Record linkage / data matching / entity resolution
  - Identify which records in one or more databases correspond to the same real-world entity (person, business, product, etc.)
  - A special case is deduplication (or duplicate detection) in a single database
- Data fusion
  - Merge pairs or groups of records that correspond to the same entity into one clean, up-to-date, and consistent record that represents the entity

Example application: World of Objects

- Goal: To enable various products in Yahoo! to synthesise knowledge-bases of entities relevant to their domains *(Bellare et al., VLDB, 2013)*
- Desiderata:
  - **Coverage**: the fraction of real-world entities
  - **Accuracy**: information must be accurate
  - **Linkage**: the level of connectivity of entities
  - **Identifiability**: one and only one identifier for a real-world entity
  - **Persistence/content continuity**: variants of the same entity across time must be linked
  - **Multi-tenant**: be useful to multiple portals

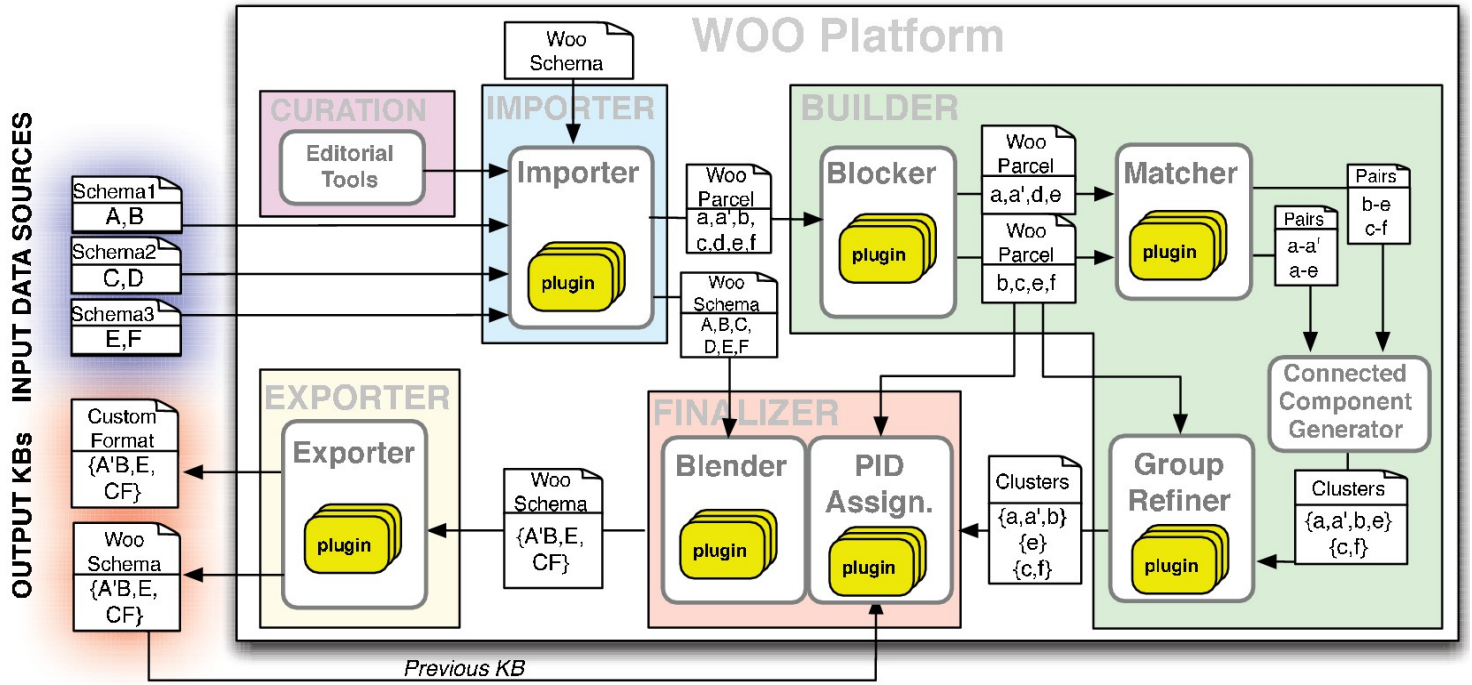
(based on slides by Hye-Chung Kum, Texas A&M)

Woo: Knowledge-base synthesis

- **Knowledge-base synthesis** is the process of ingestion, disambiguation, and enrichment of entities from a variety of structured and unstructured data sources
  - Sheer scale of the data ⇒ Hundreds of millions of entities daily
  - Diverse domains ⇒ From hundreds of data sources
  - Diverse requirements ⇒ Multiple tenants, such as Locals, Movies, Deals, and Events in (for example) the Yahoo! website

Woo architecture (1)

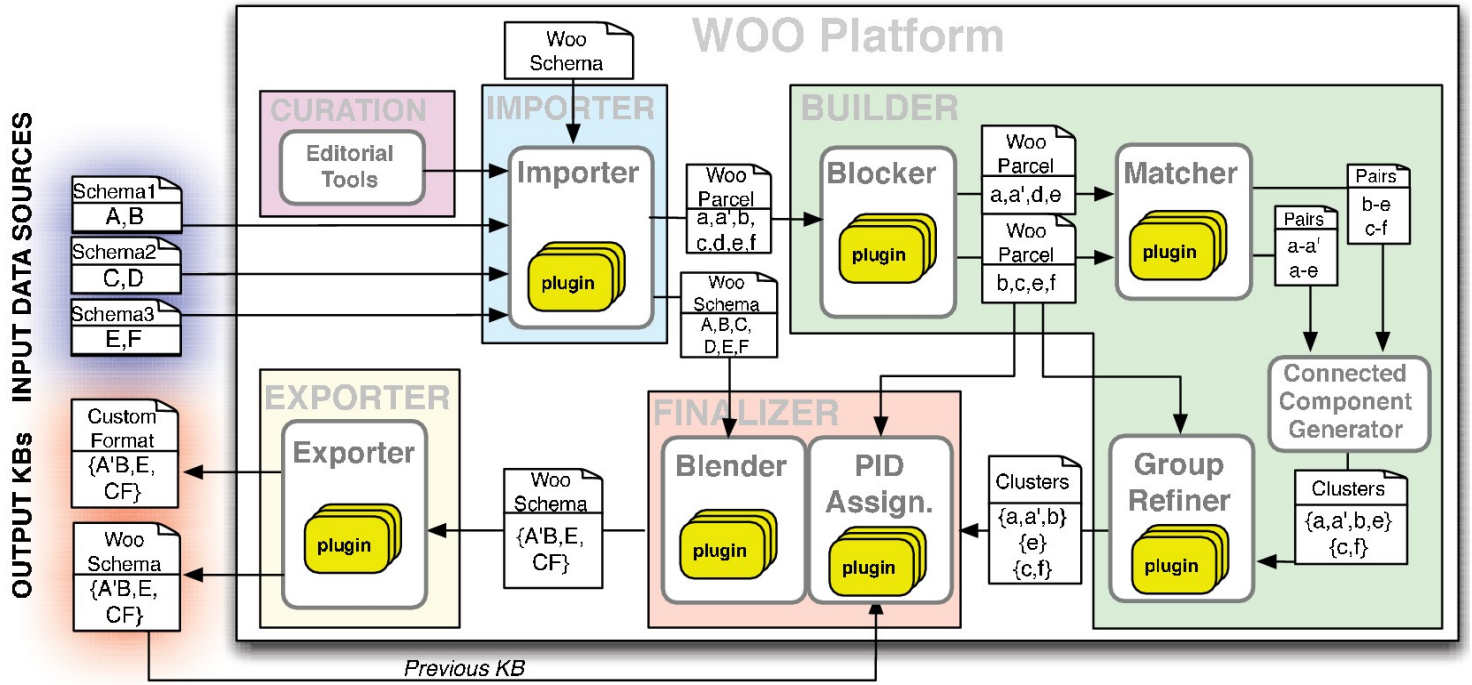
(source: *Bellare et al., VLDB, 2013)*



- **Importer** takes a collection of data sources as input (like XML feeds, RDF content, relational databases, or other custom formats)

Woo architecture (2)

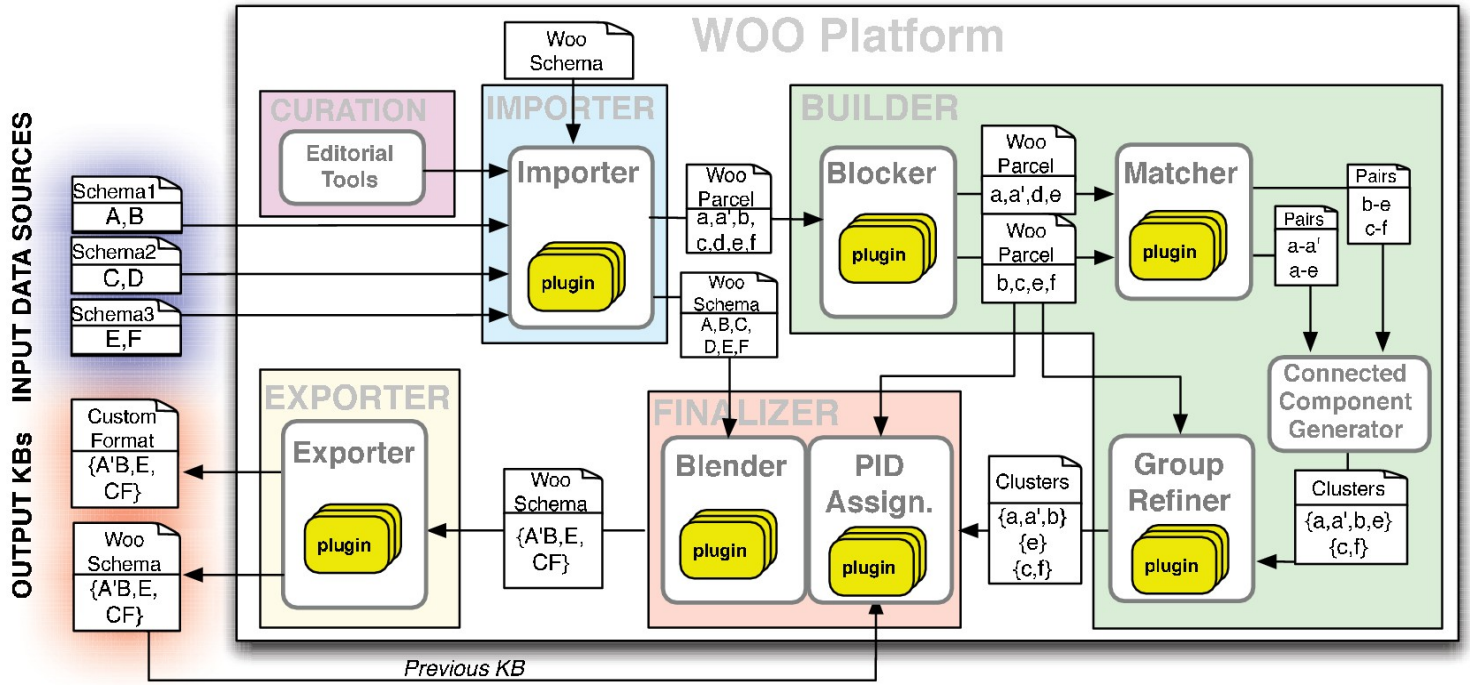
(source: *Bellare et al., VLDB, 2013)*



- Each data source is converted into a common format called *WOO schema*
- The *WOO Parcel*, containing only the attributes needed for matching, is pushed to the **Builder**

Woo architecture (3)

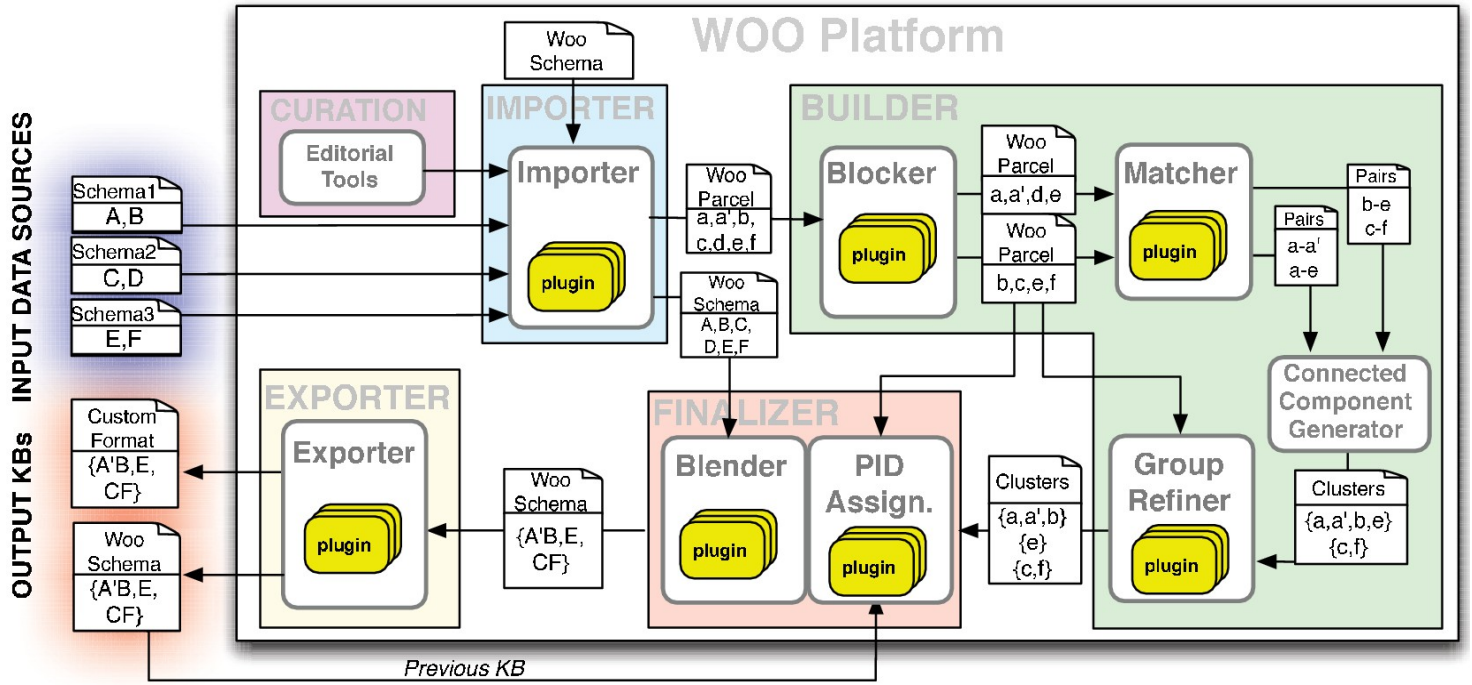
(source: *Bellare et al., VLDB, 2013)*



- **Builder** performs the entity deduplication and produces a clustering decision.
- It includes: (1) *Blocker*, (2) *Matcher*, (3) *Connected Component Generator*, and (4) *Group Refiner*

Woo architecture (4)

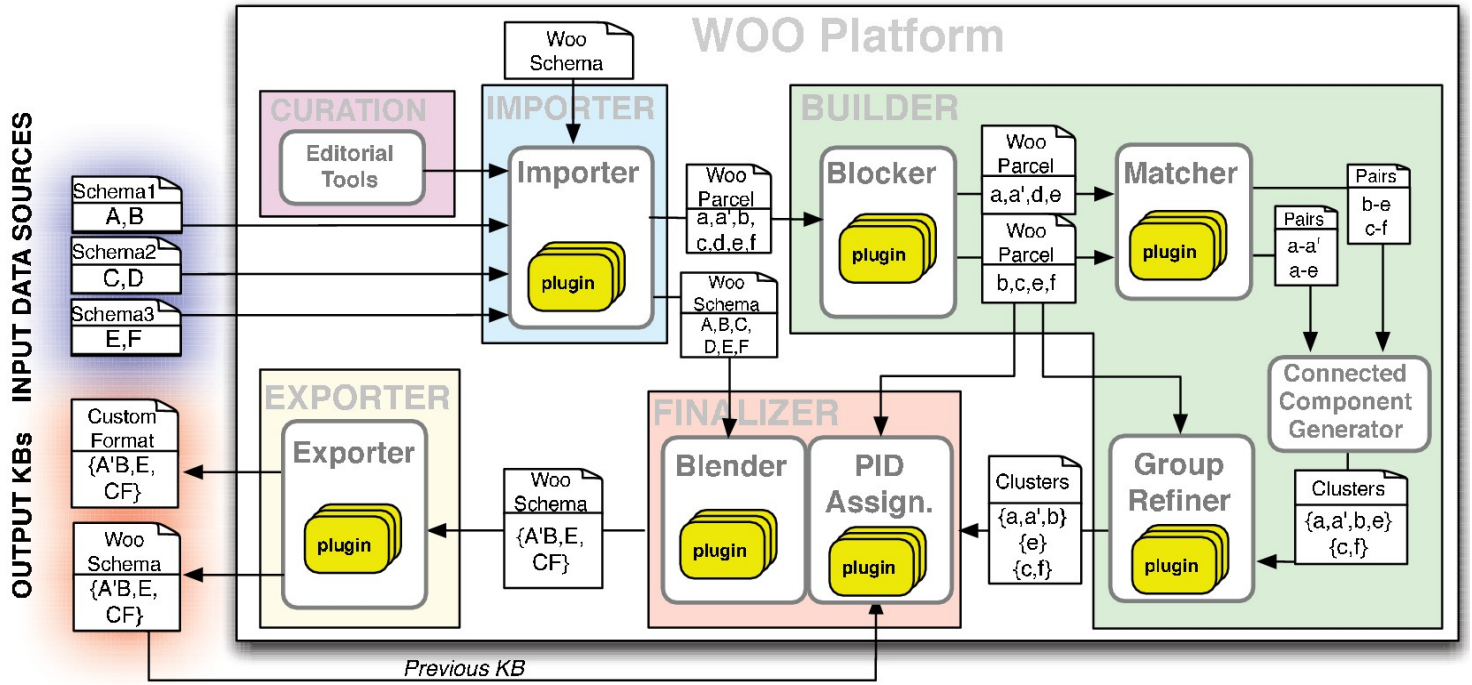
(source: *Bellare et al., VLDB, 2013)*



- **Finalizer** is responsible for handling the persistence of object identifiers and the blending (fusion) of the attributes of the (potentially many) entities that are being merged

Woo architecture (5)

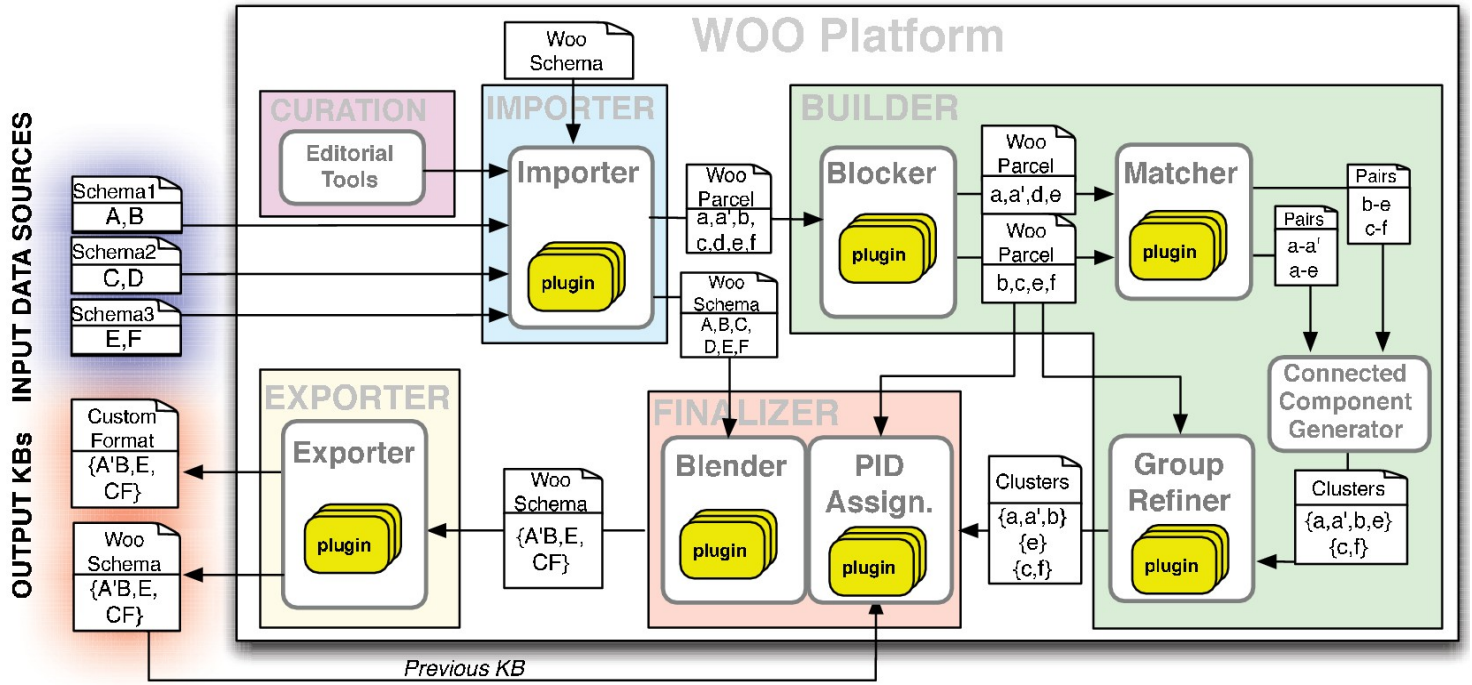
(source: *Bellare et al., VLDB, 2013)*



- **Exporter** generates a fully integrated and de-duplicated knowledge-base, either in a format consistent with the WOO schema or in any custom format

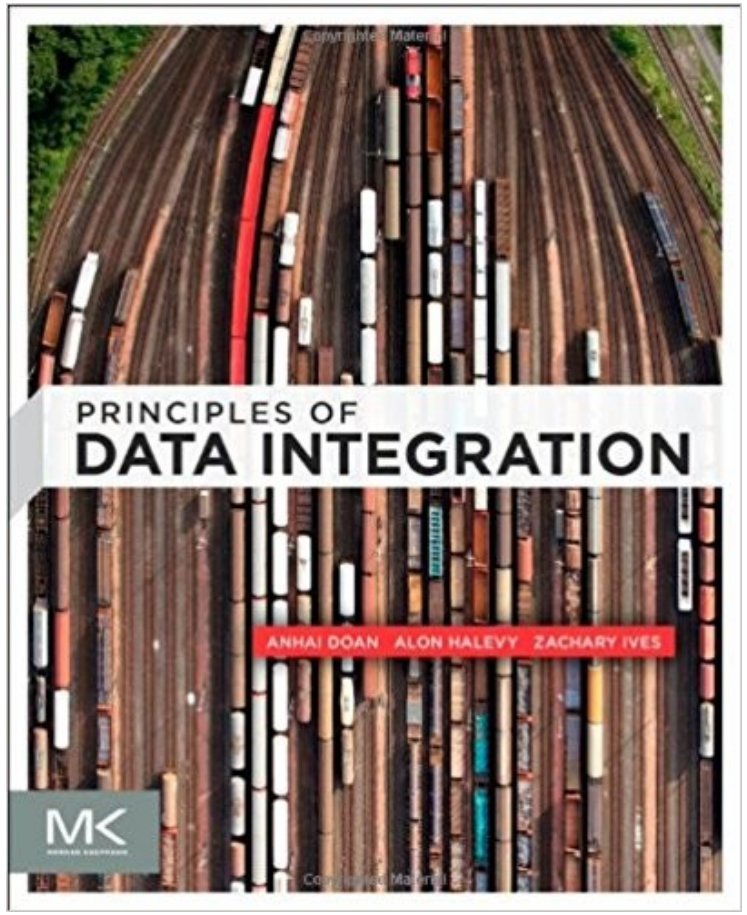
Woo architecture (6)

(source: *Bellare et al., VLDB, 2013)*



- **Curation** enables domain experts to influence the system behaviour through a set of GUIs, such as forcing or disallowing certain matches between entities, or by editing attribute values

Recommended book on data integration



Principles of Data Integration
by AnHai Doan, Alon Halevy and Zachary Ives; Morgan Kaufmann, 2012

"Principles of Data Integration is the first comprehensive textbook of data integration, covering theoretical principles and implementation issues as well as current challenges raised by the semantic web and cloud computing. The book offers a range of data integration solutions enabling you to focus on what is most relevant to the problem at hand. Readers will also learn how to build their own algorithms and implement their own data integration application."