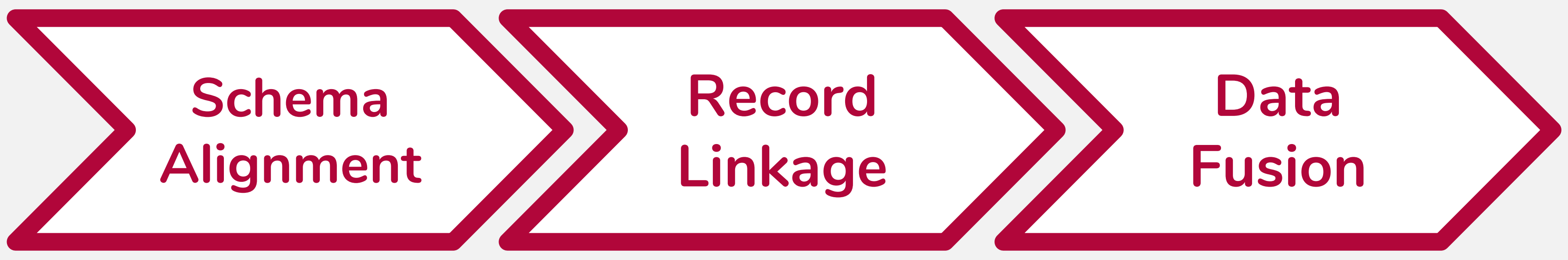
Emerging Topics in Data Integration

Traditional Data Integration[1]



Usually done using pipeline architecture with three major steps:

1. **Schema Alignment:** Find attributes with same meaning.
2. **Record Linkage:** Find records that refer to the same distinct entity.
3. **Data Fusion:** Decide the true value for an item with multiple sources.

Challenges

Volume
Huge volume of data and large number of data sources.

Velocity
Dynamic data sources with frequently changing information.

Variety
Heterogeneous data sources and evolving schemas and representations.

Veracity
Significant differences in data quality e.g. in coverage, accuracy, and timeliness.

Emerging Topics

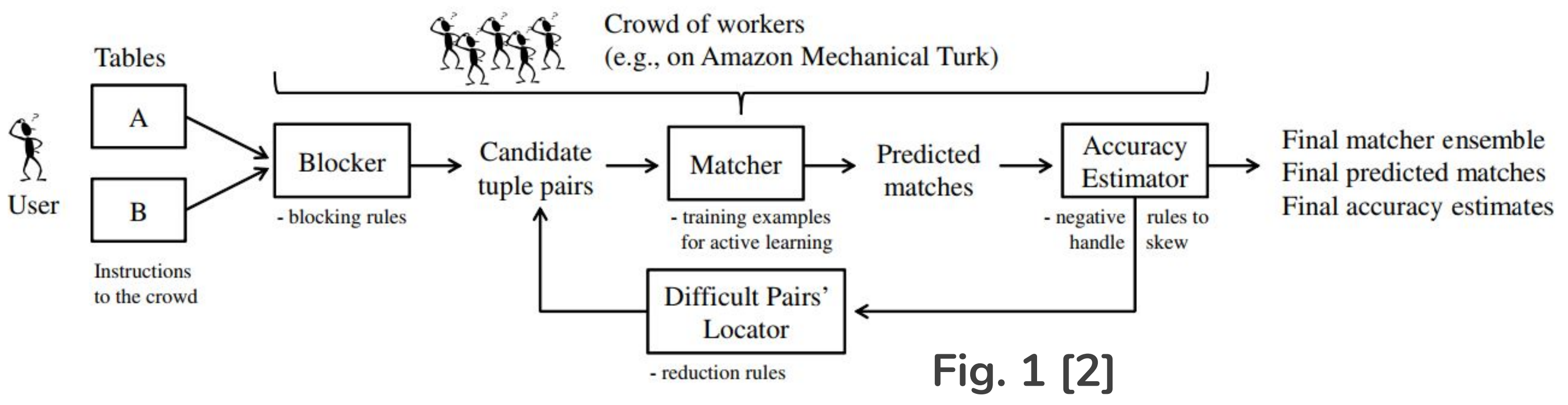


Fig. 1 [2]

Crowdsourcing

Hands-off Crowdsourcing[2]

- End-to-end workflow for record linkage without external intervention.
- Achieves high accuracy with low costs.

Future Work

- Impact of data quality on crowdsourcing results.
- How to apply crowdsourcing to recent algorithmic innovations.

Source Selection

Goal: Balance cost and benefit of integration. Not always worthwhile to integrate all sources.

Static Sources[6]

- Select subset of sources with highest profit.
- Estimate accuracy of data fusion.

Dynamic Sources[7]

- Time-dependent definition of quality metrics.
- Statistical model for describing evolution of the world.

Future Work

- Handle dependent data sources.
- Existing work only considers data fusion.

Best Effort Schema Alignment

Goal: Start with best effort solution with pay-as-you-go improvements[3]:

- **Probabilistic Schema[4]:** Clustering of mapped attributes annotated with probability of them being true (p-schema).
- **Best Effort Queries[4]:** Queries return approximate answers based on p-schema.
- **Pay-as-you-go User Feedback[5]:** Improve mapping using user feedback. Maximize benefit by finding best candidates for users to decide on.

Source Profiling

Goal: Discover sources that are relevant and have sufficient quality.

Bellmann System[8]: Surface data quality issues, find linked attributes, discover join paths, ...

Database Summarizing[9]: Identify domains and main tables. Cluster tables based on strength and importance of a table.

Future Work:

- Incremental profiling.
- Profiling for non-relational sources.

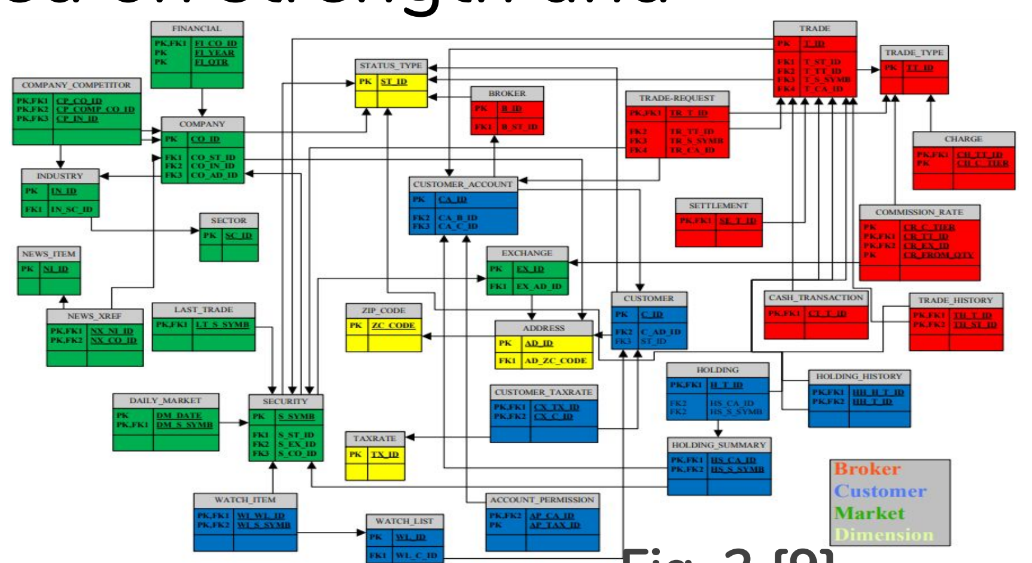


Fig. 2 [9]

References

[1] Dong et al. *Big data integration.* Synthesis Lectures on Data Management 7.1, 2015

[2] Gokhale et al. *Corleone: hands-off crowdsourcing for entity matching.* In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2014

[3] Franklin et al. *From databases to dataspaces: a new abstraction for information management.* ACM SIGMOD Rec., 34 (4), 2005

[4] Das Sarma et al. *Bootstrapping pay-as-you-go data integration systems.* In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2008

[5] Jeffery et al. *Pay-as-you-go user feedback for dataspace systems.* In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2008

[6] Dong et al. *Less is more: Selecting sources wisely for integration.* Proc. VLDB Endowment, 2012

[7] Rekatsinas et al. *Characterizing and selecting fresh data sources.* In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2014

[8] Dasu et al. *Mining database structure; or, how to build a data quality browser.* In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2002

[9] Yang et al. *Summarizing relational databases.* Proc. VLDB Endowment, 2009

HPI Hasso Plattner Institut

Fabian Windheuser
Data Engineering in der Praxis (M.Sc.)
Hasso Plattner Institute, Potsdam, Germany
E-Mail: fabian.windheuser@student.hpi.de