Data integration: evolution and challenges in recent years

Gabriela Martínez
(airamgabriela17@gmail.com)

Sara Díaz
(sara.diaz052@gmail.com)

Introduction

The concept of **data integration** started being theoretically explored in the early 1980s when the computer science community started doing research on how to effectively combine information silos.

in 1991, the **IPUMS project** launched by the University of Minnesota to integrate various samples, surveys and census around the world; led to the **first data integration system.**

From there on, the commercial arena recognizes the industry under the **Enterprise Information Integration (EII) concept.** and it envisions to provide tools to integrate heterogeneous data sources without having to create a central data warehouse repository as in the beginning.

This research aims to show the evolution and main challenges in the last two decades of the data integration concept over three main stages:

Main idea

During the **earlier years** of the development of this field, the main challenges it faced were **associated with the complexity of generating mappings and processing queries** and having to deal with uncertainty, while **latter times have been dedicated to thinking more about semantic-oriented issues** that arise under the same uncertain scenario, where a new player has been set: the *big data* and its variety of structured and unstructured formats

Infancy

GAV, LAV, and GLAV

In the late 1990s, two additional problems associated with the use of the data warehousing:

1. ETLs were harsh to be executed on frequently updated data sources
2. Query interfaces only on summarized data sources

This gives birth to two new different approaches of integration architectures:

GAV (Global-as-view): the mappings model the global schema as a set of view definitions over the schemas of the data sources.

- PRO: offer a simple unfolding strategy to execute queries
- CON: Doesn't deal well with rapid increase of data sources

LAV (Local-as-view): the contents of data sources are modeled as views over the global schema

- PRO: offer more flexibility when a variety of data sources is frequently on the table -> leading data integration in the big data era
- CON: Query processing in LAV is a very difficult task

Alternatives have been proposed to take advantages of both, namely Global-Local-As-View (**GLAV**) and Target-based Integration Query System

Teenage years

Earlier Challenges

This era (early 00's) focuses on *provide a uniform query interface to a multitude of data sources,* which leads to the following challenges:

Generating schema mappings: initially focused on generating semi-automated schema mappings and later explored as a machine learning problem. The XML format contributed but still lacked semantic on its tags.

Adaptive query processing: The problem lies on existing database optimizers and execution engines being not appropriate.

Model management: algebra approach to metadata management that offers a higher level programming interface

Peer-To-Peer data management (PDMS): decentralized, easily extensible data management architecture in which any user can contribute new data.

Dataspaces: aim to offer pay-as-you-go data management with no startup delay

Uncertainty and lineage: Use lineage (e.g. text snippets and URLs) to reduce uncertainty (best search results)

All the challenges converge under a set of assumptions that make up the **classical DI paradigm,** which became outdated with the arrival of the big data era and its new requirements:

1. The global schema has a reasonable size and can be built with modest effort
2. The data sources are structured and have well-defined schemas
3. There is a need to integrate all the data sources at hand
4. All data integration functionality should be part of an end-to-end system
5. The data in the data sources is mostly correct and consistent across them

Adulthood

Today's landscape

Nowadays, the **some earlier challenges have evolved and are still under research.** Dataspaces, for instance, still hold technical complexities as architectures move to be loosely coupled and service-oriented.

However, it can be said that Integration deepest challenge today is not merely related with the architecture style, but with **the underlying logics to express communication** amongst the data ecosystem. Model management and PDMS have been very important in this.

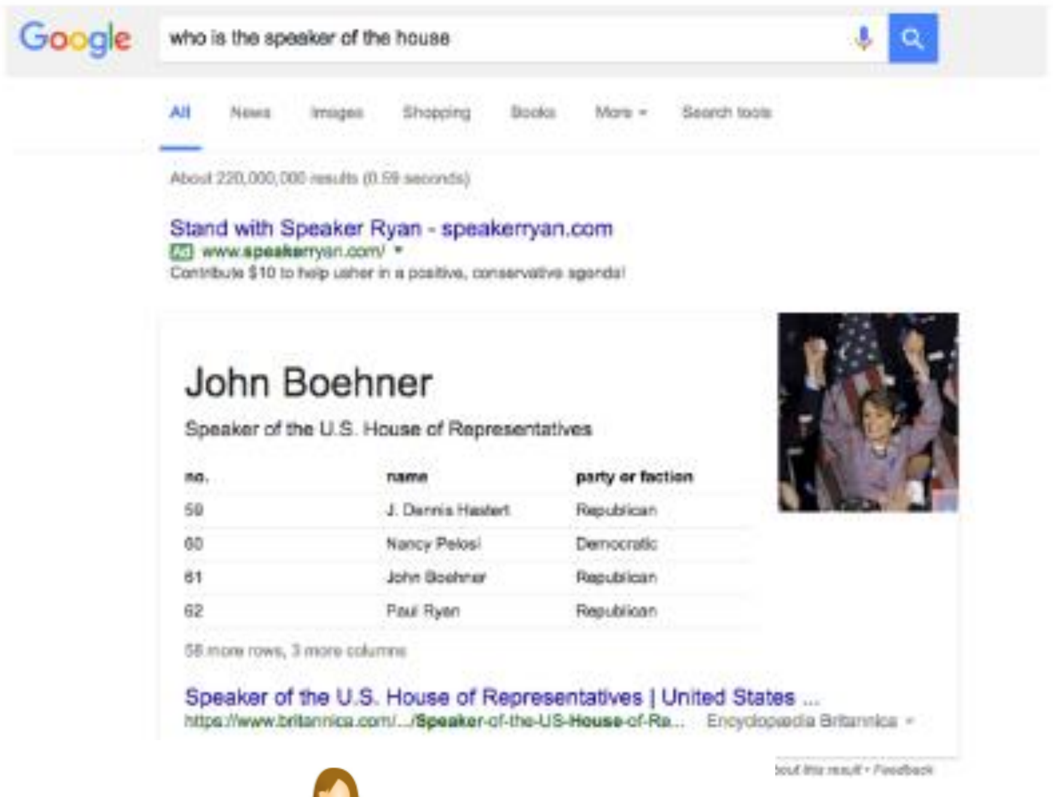
Additionally, other challenges have risen:

Open source tools: many systems created by the EII are not properly adopted as data integration complexities are not recognized as such by normal practitioners -> there's a need for open source tools that can replace these in a more independent way. An example is *BigGorilla*.

Combining structured and unstructured data: combining different nature of data formats seems to be an intuitive challenge. potential solutions should be guided by:

- A proper declarative language
- Mappings should allow original data sources access

An example of this can be seen with the google search: *who is the speaker of the house:*



About 220,000,000 results (0.59 seconds)

Stand with Speaker Ryan - speakerryan.com
Ad www.speakerryan.com/
Contribute $10 to help usher in a positive, conservative agenda!

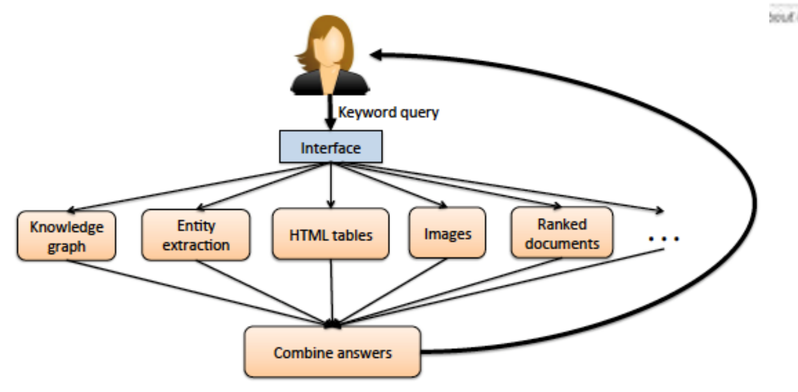
John Boehner
Speaker of the U.S. House of Representatives

no.	name	party or faction
59	J. Dennis Hastert	Republican
60	Nancy Pelosi	Democratic
61	John Boehner	Republican
62	Paul Ryan	Republican

58 more rows, 3 more columns

Speaker of the U.S. House of Representatives | United States ...
https://www.britannica.com/.../Speaker-of-the-US-House-of-Re... Encyclopaedia Britannica

bout this result • Feedback



Conclusion

The big data era has created previously unseen new realities **where data integration is required with increasing urgency.** Although not a research field, it is now growing towards a more mature development path where identifying the correct **data context and semantics** has become mandatory.

Literature cited (main sources)

Halevy, Alon & Rajaraman, Anand & J. Ordille, Joann. (2006). Data Integration: The Teenage Years. VLDB 2006 - Proceedings of the 32nd International Conference on Very Large Data Bases. 9-16.

Golshan, B., Halevy, A.Y., Mihaila, G.A., & Tan, W.C. (2017). Data Integration: After the Teenage Years. PODS.

AnHai Doan, Pedro Domingos, and Alon Y. Halevy. Reconciling Schemas of Disparate Data Sources: A Machine Learning Approach. In Proceedings of the ACM SIGMOD Conference, 2001.

Summer School
eBiss 2019
Berlin, Germany
June 30th - July 5th