MarkLogic

Semantics Driven Data Integration

Jen Shorten
Architect, MarkLogic

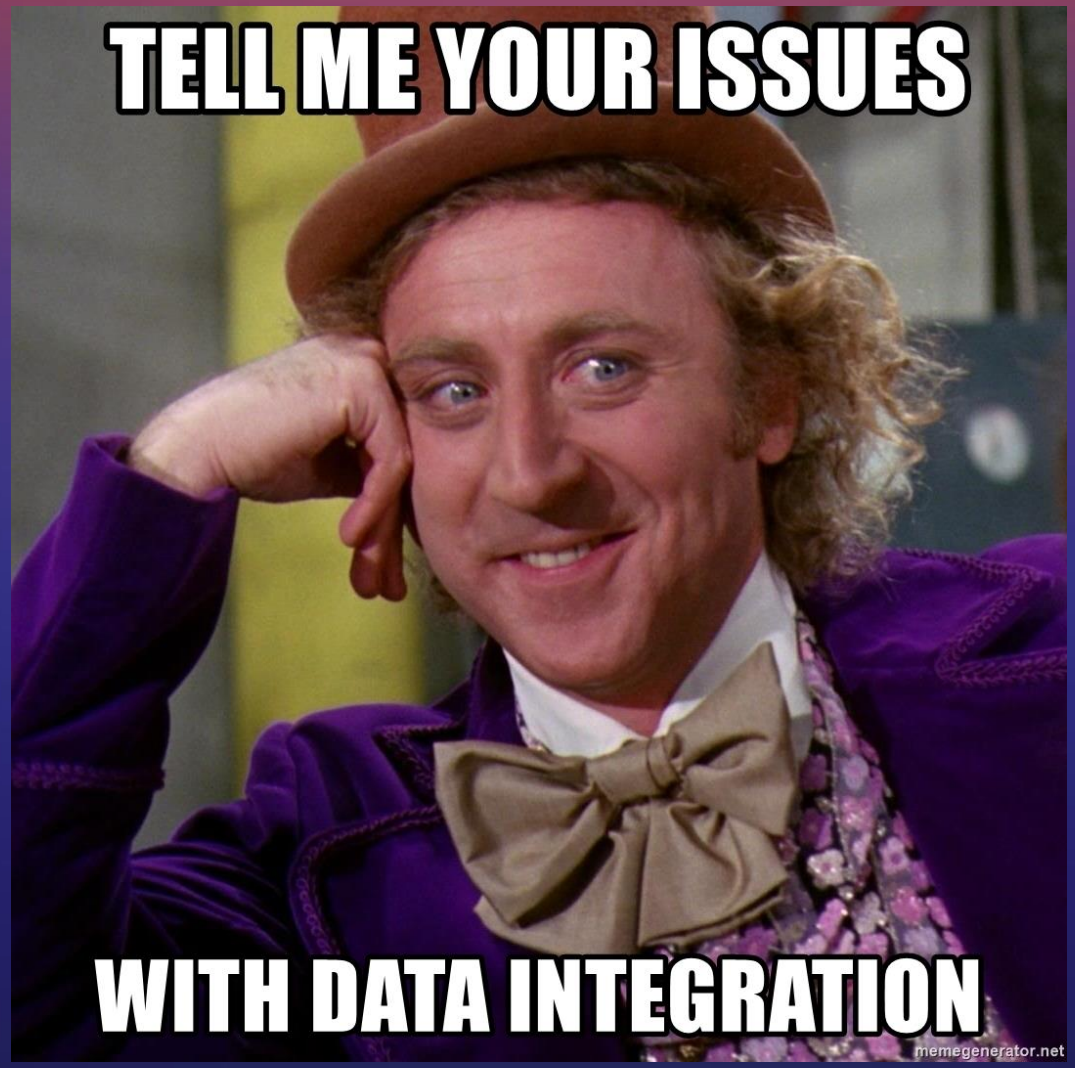
Edward Thomas
Consultant, MarkLogic

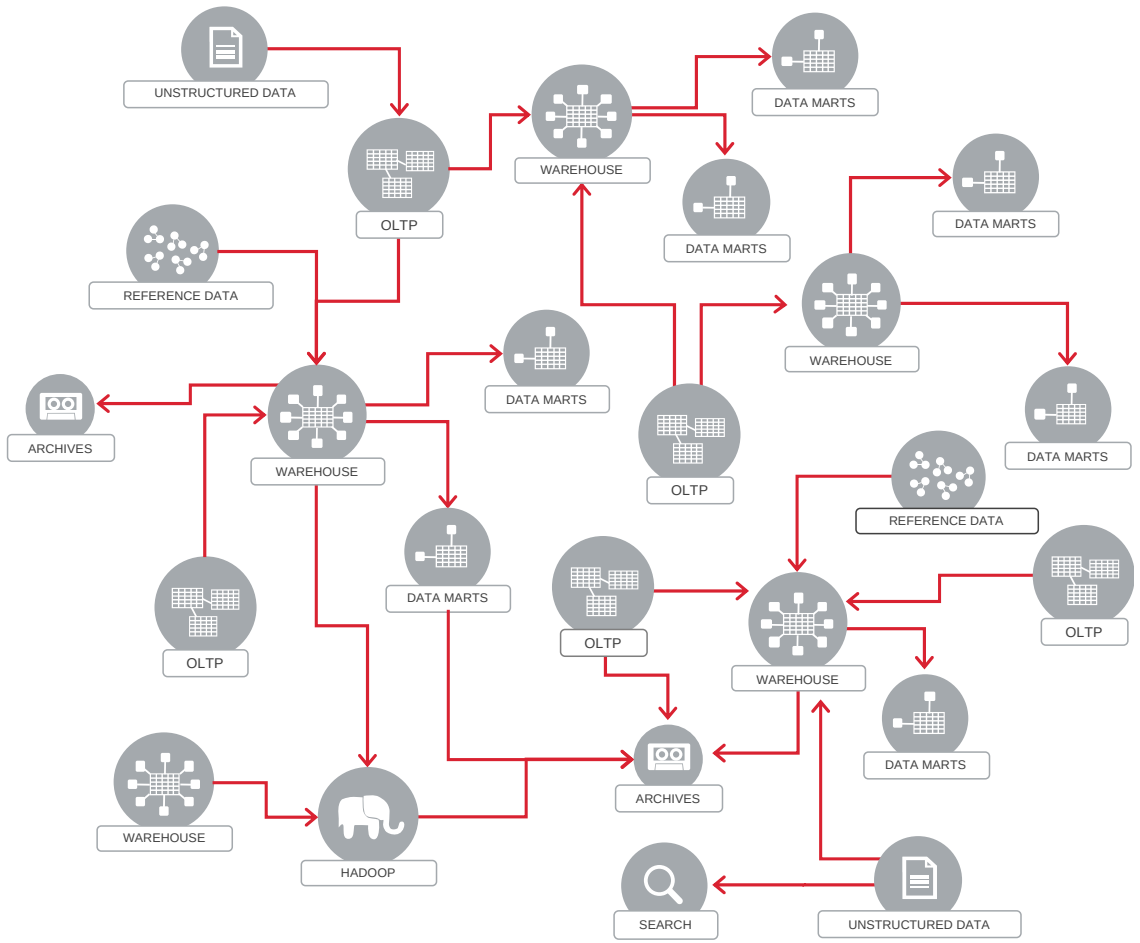
MarkLogic

Why Is Data Integration Important?

- Organisations are busy creating vast amounts of interesting and useful data
- Operational "run the business" data is needed in real-time more than ever as organisations undergo digital transformations.
- At the same time, analytical “observe the business” functions are becoming as important as the operational data for learning and developing new products, services and advancements in knowledge
- There is so much knowledge trapped in legacy systems, and that knowledge is so valuable that simply throwing it away would be a profound loss

MarkLogic


TELL ME YOUR ISSUES
WITH DATA INTEGRATION
memegenerator.net



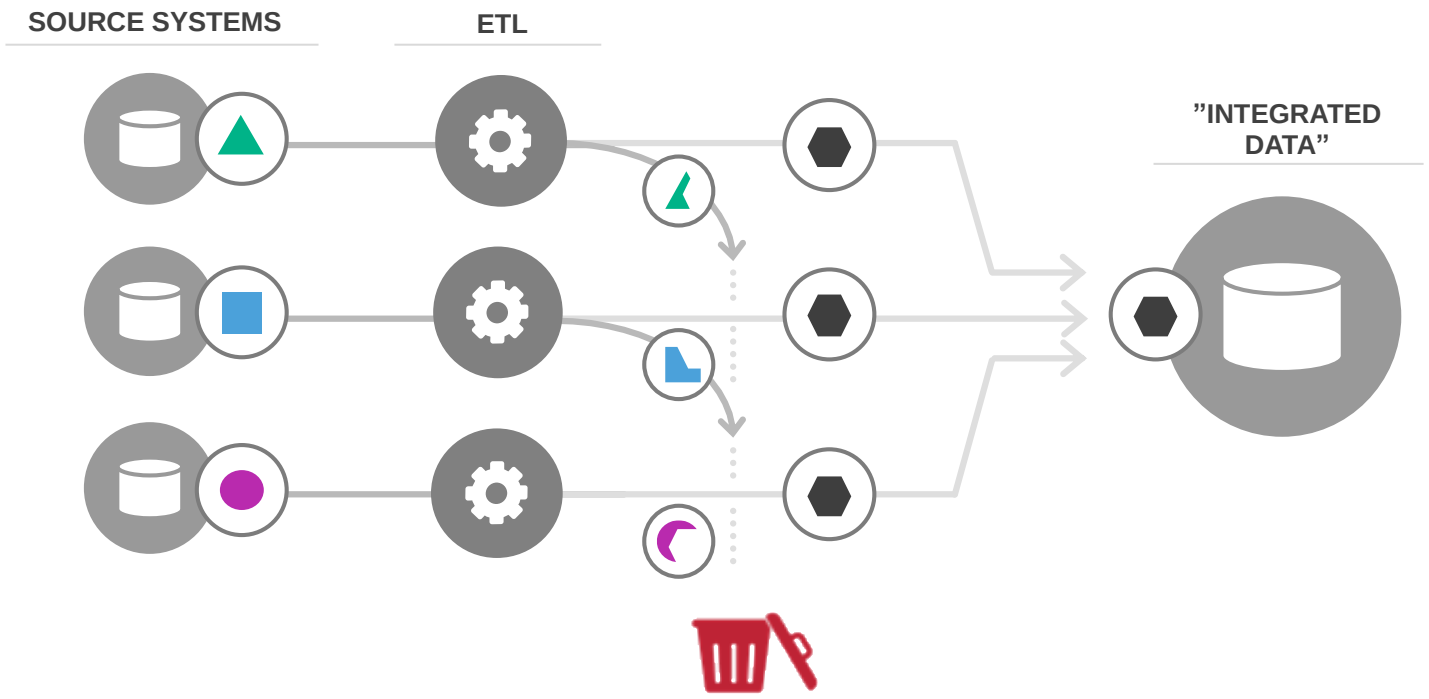
DATA INTEGRATION

Why Is It So Difficult?

- Data it is most often created in isolated silos across the organisation
- Even with the proliferation of next generation databases, most organisations are still creating volumes of data in rigid relational schemas
- Merging data across silos requires ETL
- Data is constantly changing and the pace of change will only increase

Traditional Approaches

- Sharepoint
- File stores
- Master Data Management
- Data Warehouse
- ERP
- Data Lakes

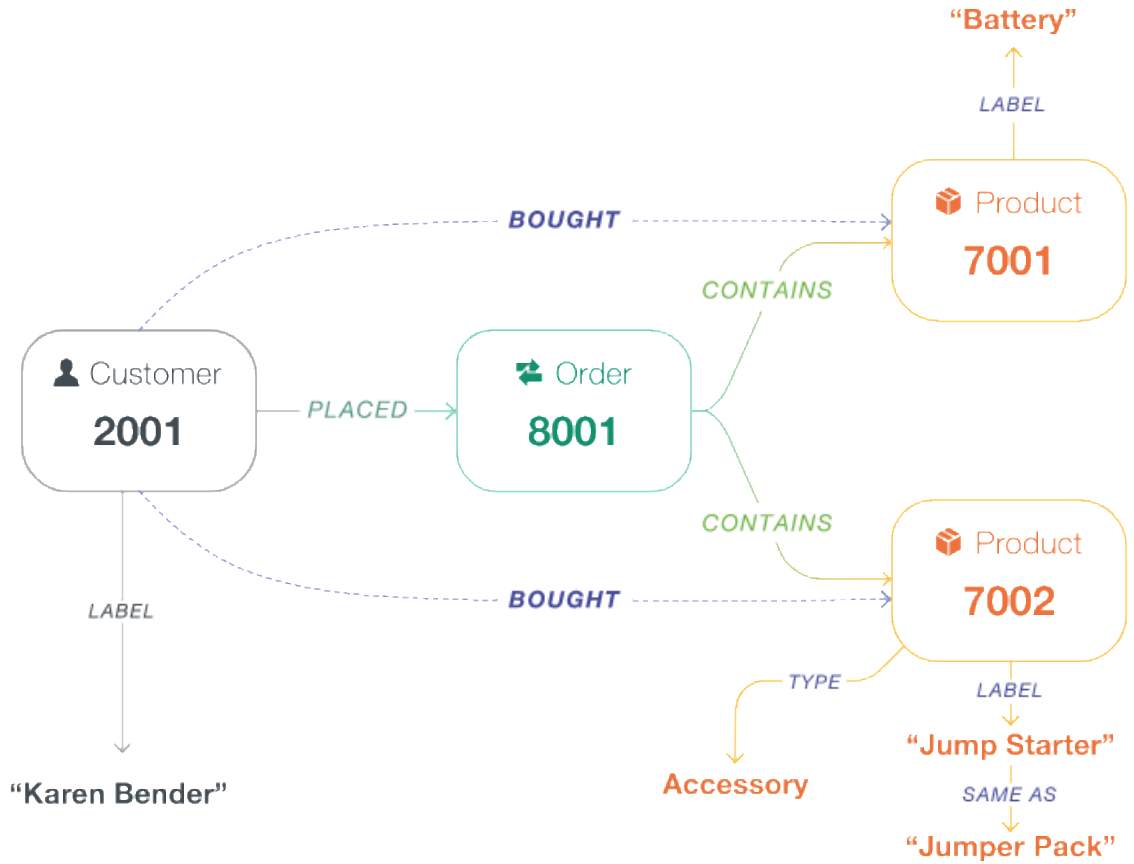


Things Go Missing In Lakes!

MarkLogic

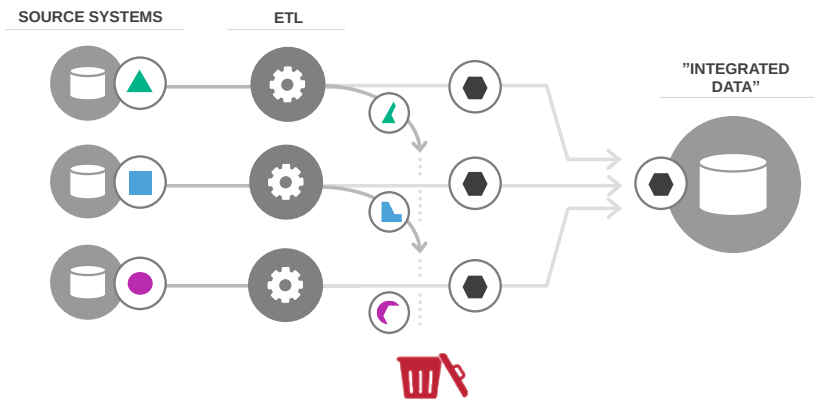
Semantics Based Approaches

- Triple/Graph Stores
- Multi-model



Triple/Graph Stores

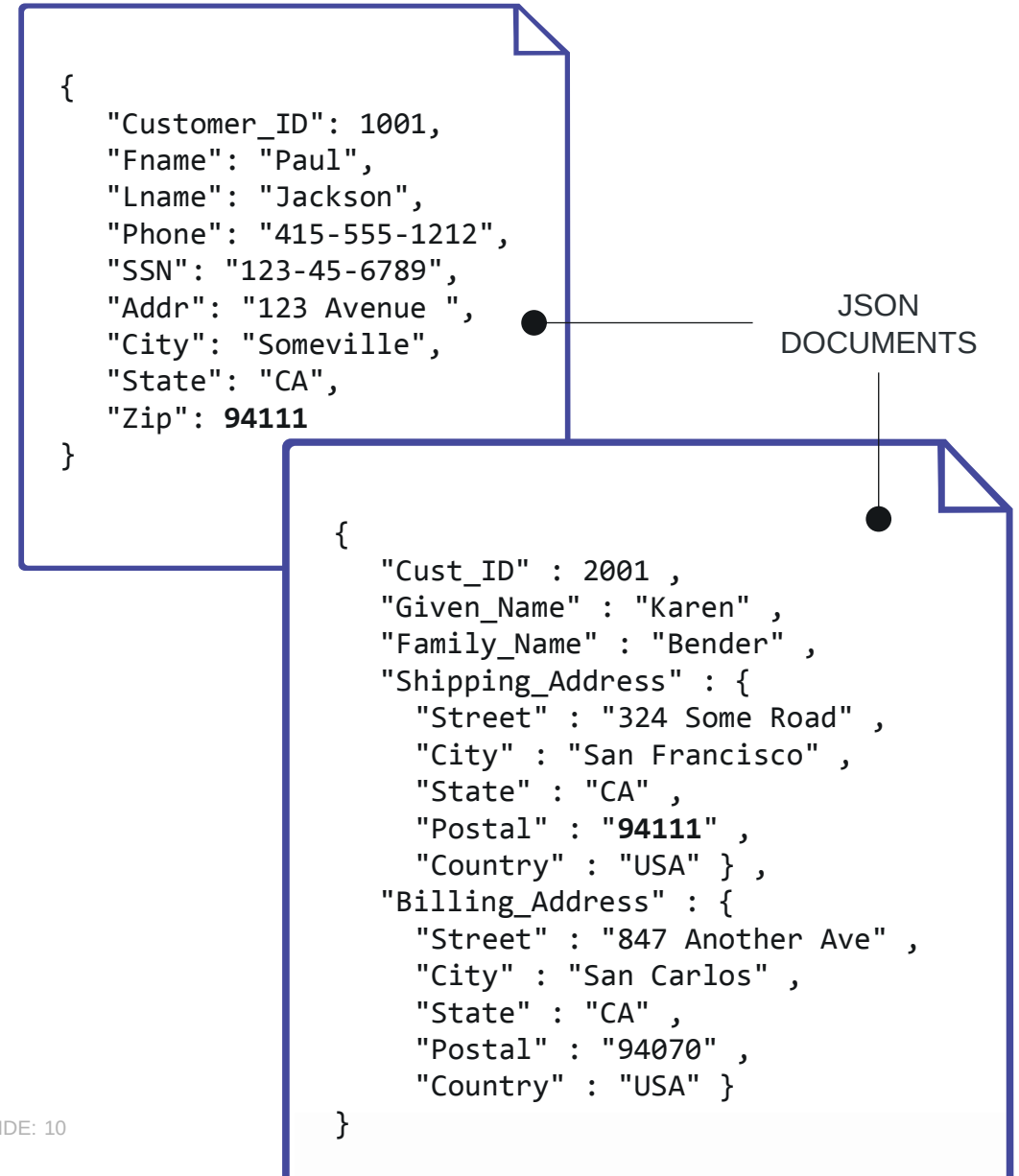
- Chop up all your data into triples
- Integrate data from different sources using mappings between ontologies and schemas
- Very flexible, but hard to scale
- Queries can get very weird, very quickly
- Some pieces of data naturally go together
  - Common access patterns, common security, created and deleted together
- Why not keep them together in the database
  - Call this a document



MarkLogic

Multi-Model

- Keep the data that looks like a document as a document
- Enrich the document with triples where it can improve functionality
  - store and version along side the document
- Domain knowledge and reference data as triples
  - let the database manage these
  - SKOS/OWL
- Identify common entities across data and harmonize only what you need when you need it
  - Just In Time data integration



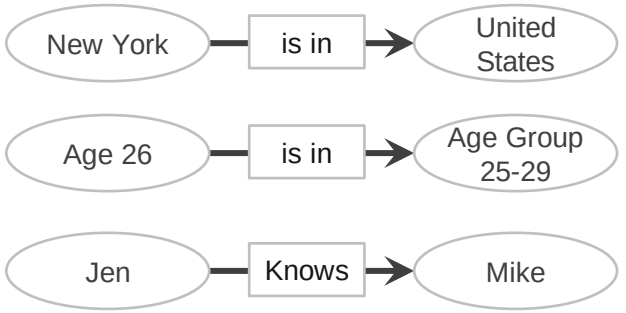
MODELING ENTITIES

The Document Model

- Natural and human-readable
- Heterogeneous data is okay (schema-agnostic)
- Query across data harmoniously (e.g., search for zip code, “94111”, returns both records)
- Partition documents using *collections* (e.g., create a collection for each source system)
- Insert/update/delete documents in a single transaction – *even if it changes the schema*

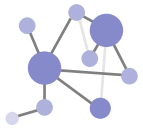
Multi-Model Example

```
{
  "Place": 2,
  "Bib": 481,
  "First name": "Jen",
  "Last name": "Kross",
  "Distance": "10k",
  "City": "New York",
  "Age": 26,
  "Gender": "Female",
  "Time": "1:29:52.2",
  "Sponsors": ["Nike", "Gatorade"],
  "Quote": "Just run it."
}
```

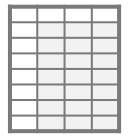


ID	Name	Address	City	State	…
1	…	…			…
2	…	…			…
3	Jen Kross	123 1st St.	New York	NY	…
4	…	…			…
5	…	…			…
6	…	…			…
…	…	…			…

DOCUMENTS
(JSON/XML, JavaScript/XQuery)



SEMANTIC DATA
(RDF Triples, SPARQL)



RELATIONAL
(Tables, SQL)

Multi-Model Use Cases

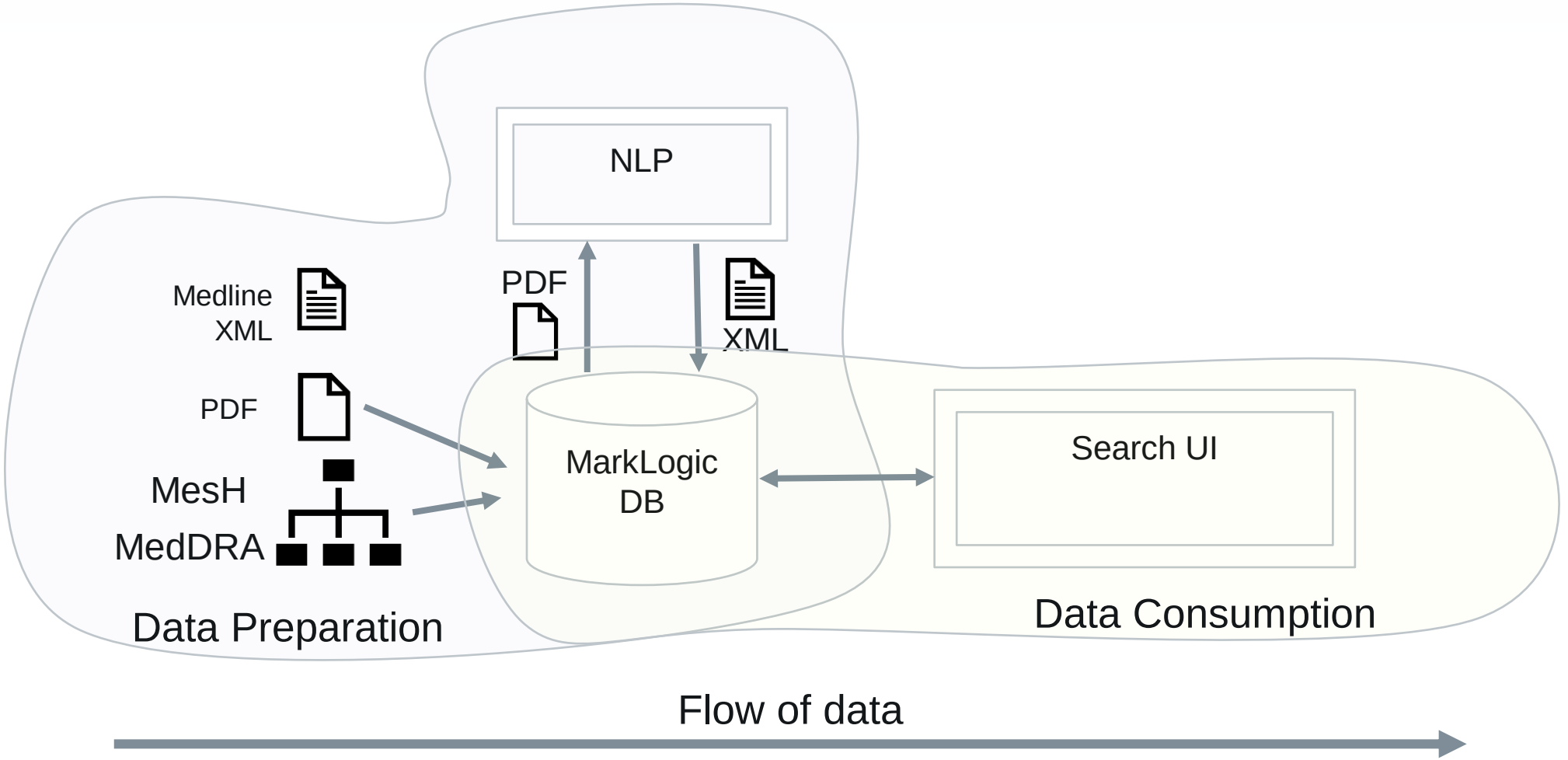
- Asking Meaningful Questions -Legacy Data Exploitation
- Intelligence Applications - Dynamic Data Consolidation
- 360º views of knowledge assets - Operational Data Hub

MarkLogic

Use Case: Legacy Data Exploitation

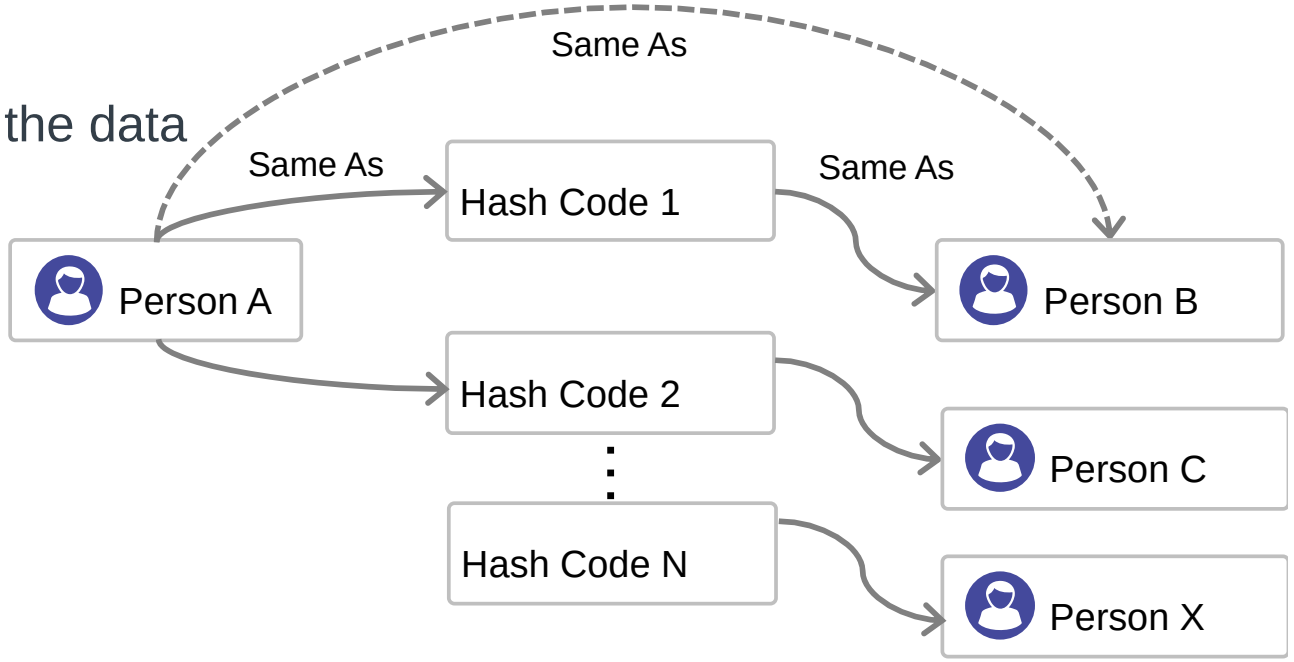
- Vast majority of legacy data is stored in documents (e.g. PDF, Word, XML)
- Answer difficult questions like
  - "Have we run this experiment before and if so what were the outcomes?"
  - "What questions has the regulator asked about similar compounds?"
  - "Where there any significant side effects at any dosage?"
- Where entity extraction and natural language processing techniques are used to enrich documents, the inclusion of semantics based search is very powerful.
- For this use case to succeed combined queries i.e. semantics + full-text are essential

Example: Pharmaceutical Research Discovery

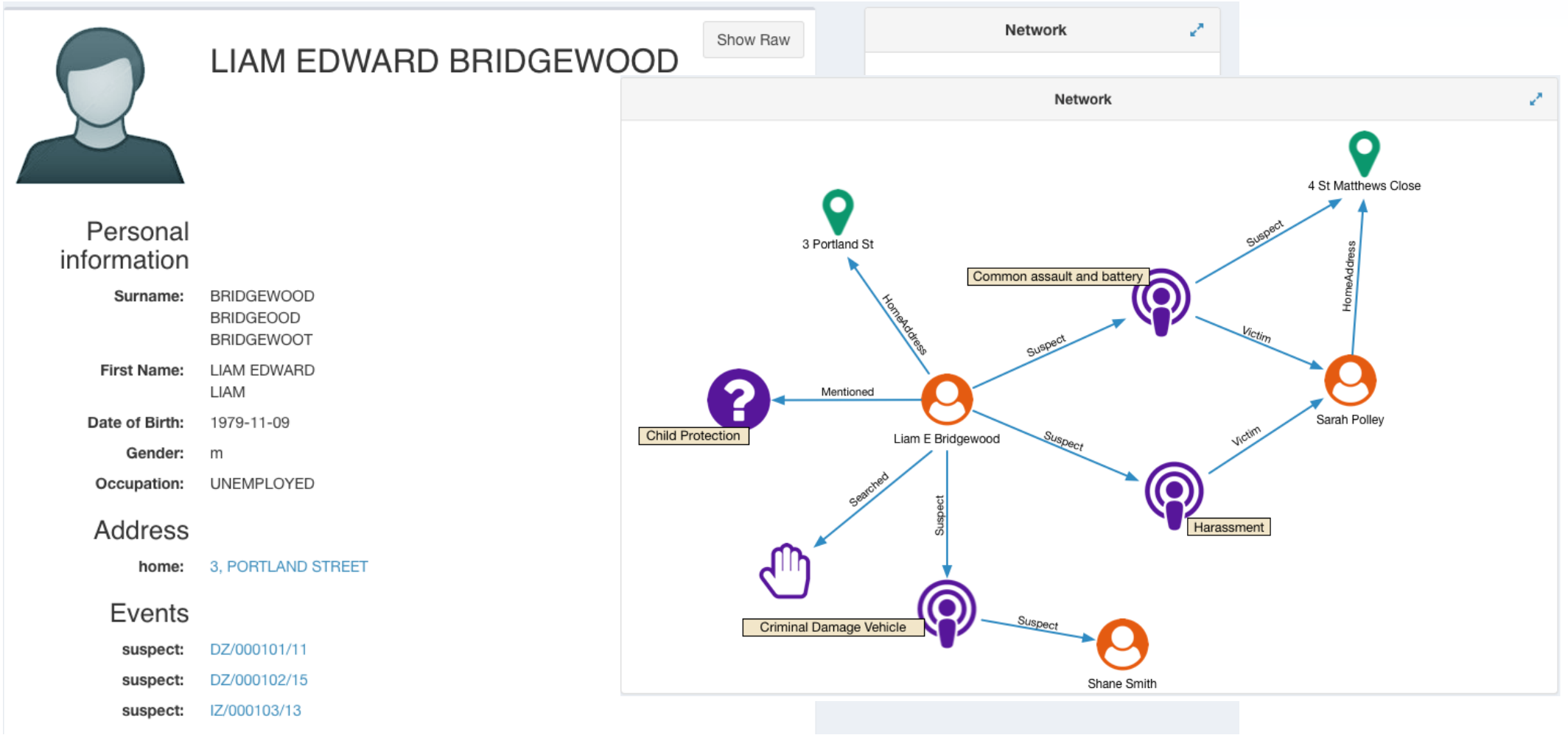


Use Case: Dynamic Data Harmonisation

- In fast moving environments where the system needs to rapidly ingest and analyse over which there is no control at source
- Data quality is spotty at best
- Data is in multiple formats
- Canonical model design is not an option
- Semantics key for identifying relationships in the data



Example: Police Intelligence Application



Use Case: Data Security

- Semantic Data derived from a source usually has the same security requirements as the source document
- Aligning security between different databases adds complexity
- Storing and managing triples with documents means that only the people who can see the documents can see the triples

MarkLogic

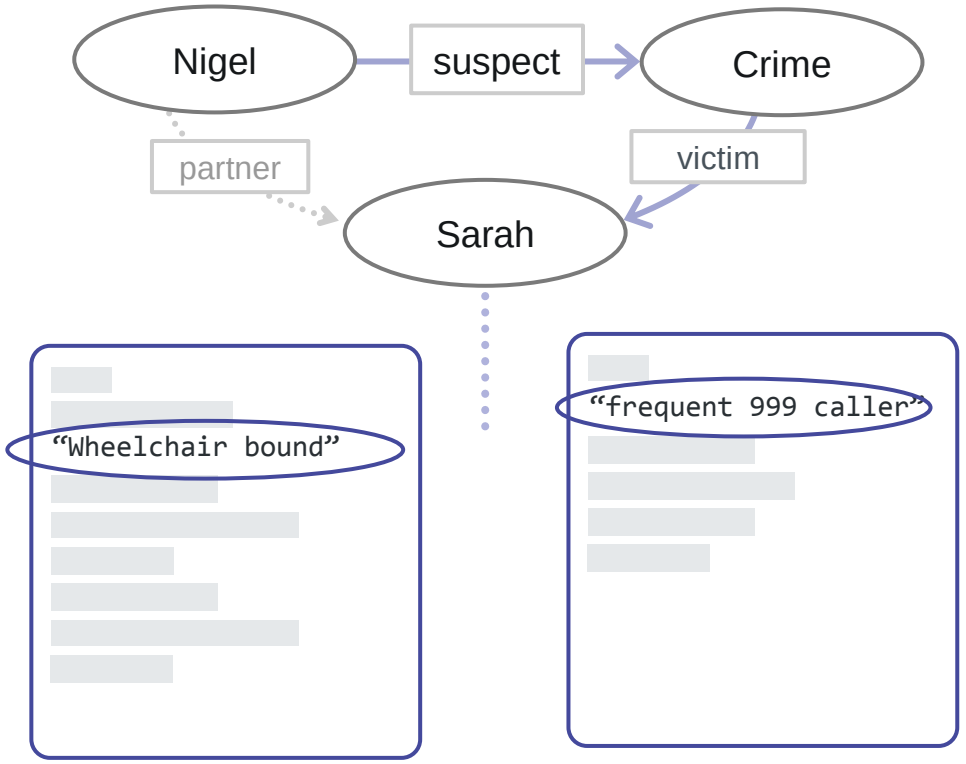
Techniques

- Linking models
- Load as-is, query as is
- Semantic expansion of terms in the query
- Harmonize the data in the document where needed, preserving the source data for traceability
- Use templates to create triples in a standard schema
- Semantic search and query using domain specific reference data (skos/ontology)

MarkLogic

Load As-Is

- Why?
- Immediately load and index all kinds of data
  - RDF triples
  - CSV files
  - XML/JSON documents
  - Binary files
- Start with basic full text search
- SELECT * FROM *all tables*
  WHERE *any column* = 'string'



MULTI-MODEL: DOCUMENTS & TRIPLES TOGETHER
JSON, XML, & RDF

THE IDEAL SOLUTION

Use All of the Data

- Semantic linking to see relationships between people, locations, events and objects
- Extract context from narrative text
- Build a complete picture by exploiting the value in all of the data

MarkLogic

Thank You!