DATAVERSITY

The Great Escape: Liberating 20+ Years of Legacy Enterprise Data

Demo Day
July 19, 2023

About Today's Speaker

Eliud Polanco

- Former Head of Analytics at HSBC, DeutscheBank, ScotiaBank
- Former Head of Data Innovation @ Citi
- Current President at Fluree PBC, a start-up focused on secure data-sharing and collaboration at scale using Zero-Trust and Semantics

What Fluree Offers

fluree CORE

Intelligent Knowledge Graph

Fluree combines cryptography and semantic data standards to enable secure data collaboration

fluree sense

AI/ML Data Cleansing Pipeline

With **Fluree** it is faster and easier to clean raw data and make it consumable for users and algorithms via semantic Knowledge Graph

Agenda

I. **The Knowledge Graph opportunity.** Why Knowledge Graphs are emerging as more critical for Enterprise use cases.

II. **The Legacy Data Problem.** Why Knowledge Graphs have historically been difficult to adopt at scale.

III. **The Innovations.** How JSON-LD and AI/ML are helping bridge the gap between relational and semantic worlds.

IV. **The Demo.** An How Fluree converts scale legacy data into data ready for semantic Knowledge Graphs.

V. **Getting Started With Knowledge Graphs In Your Organization.** A BluePrint for socializing KG use cases in your company.

I. The Knowledge Graph Opportunity

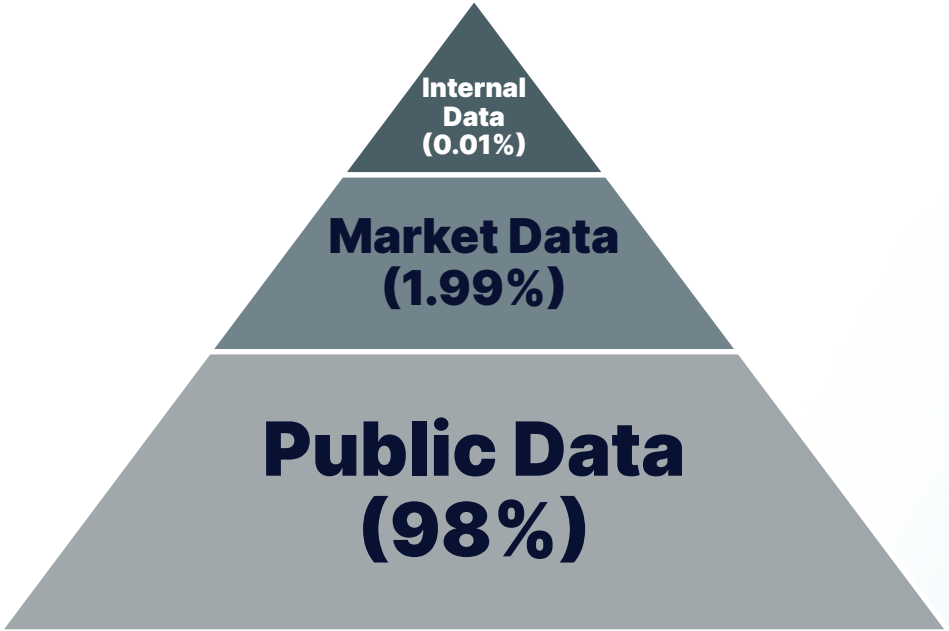
Why Look At Knowledge Graphs?

- **Data sharing** is becoming an imperative for competitive differentiation
  - Inside a company's four walls
  - Between companies and regulatory agencies who have adopted semantic standards for how they want to receive data
  - Between companies and suppliers, partners, even competitors using common industry ontologies

- **AI, Algos and Machines** are becoming the **primary consumer** of vast stores of data inside a company
  1. Generative AI and **Large Language Models** (LLM) shook the world
  2. **Internet of Things**: devices create data, and devices consume data

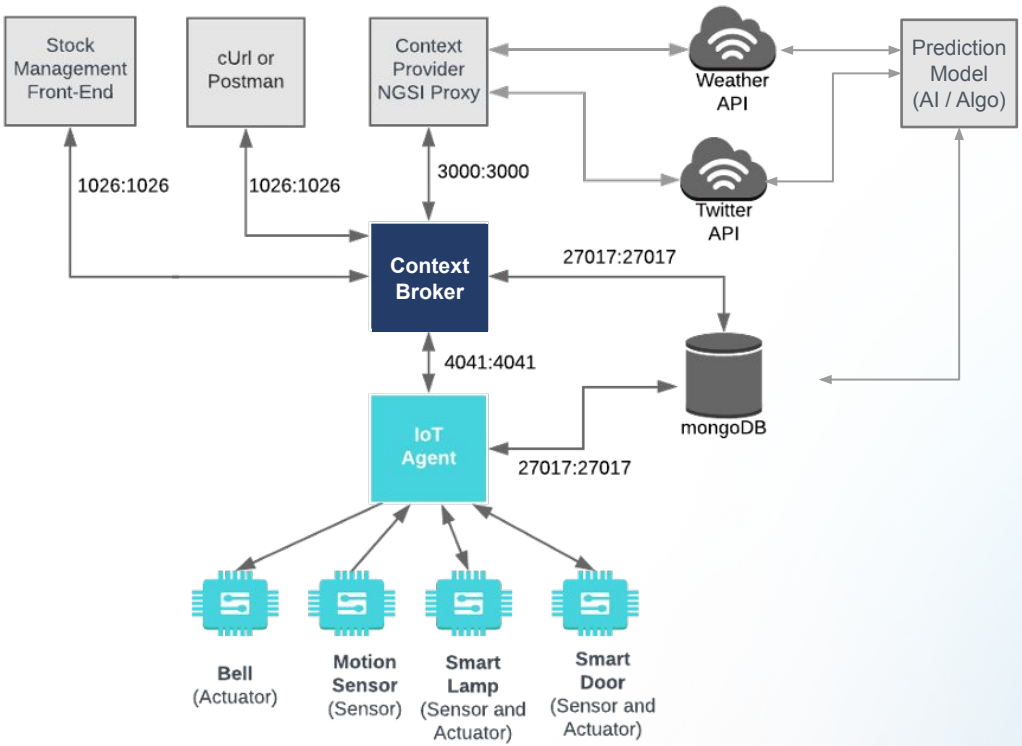
Semantic interoperability is no longer a niche use case… it is an economic imperative

1. Large Language Models (LLMs)



- The vast quantity of data consumed into LLMs are public data that is either unstructured, or using semantic Web standards (RDF)
- Some market data (data available for a price) is opportunistically being integrated into LLMs (at a price)
- Most internal data is out of the reach, because:
  - It is proprietary.
  - It lacks semantic context.
- Large enterprises today are experimenting in how to **build their own internal LLMs** using a mix of their own structured and unstructured corporate internal data.
- This will drive **competitive differentiation inside companies** in how they create products and optimize their own business processes

Internet of Things (IoT)

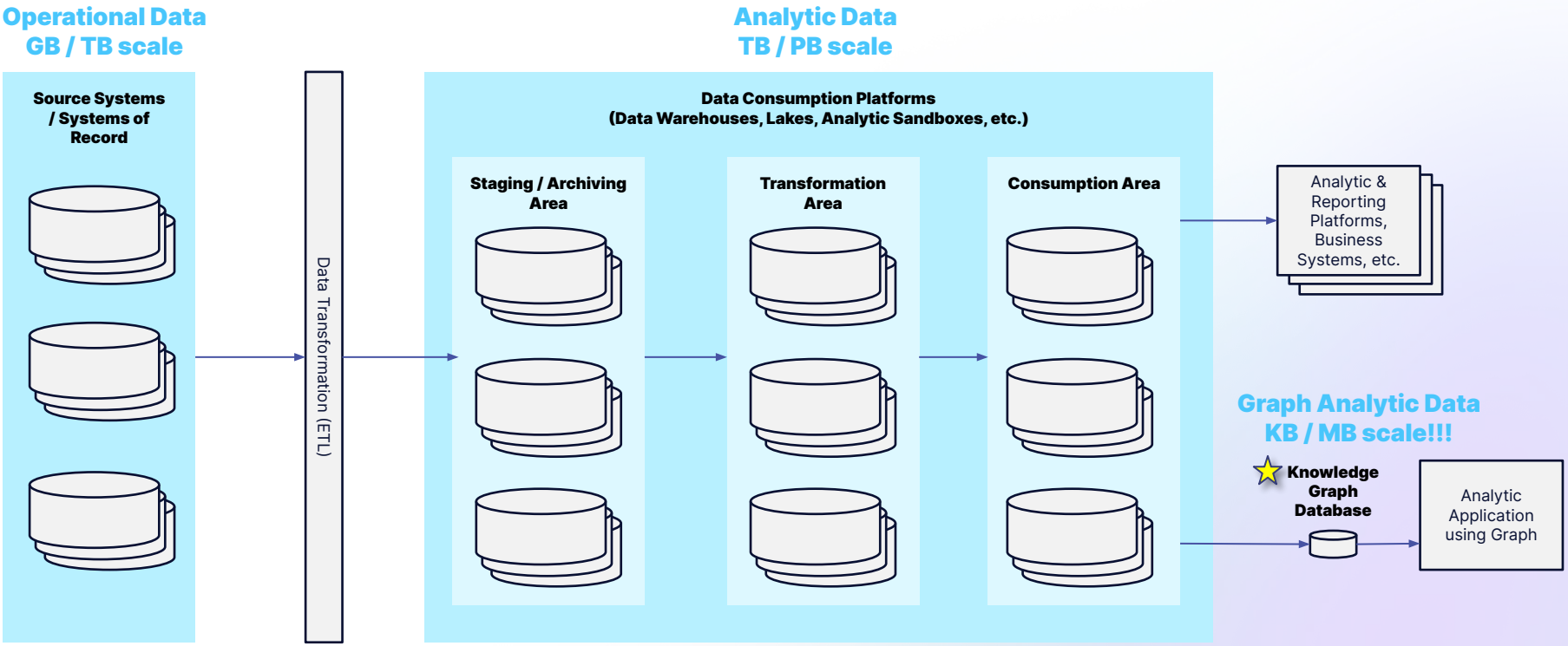


- Data being created by sensors and actuators...
- ...and being distributed via APIs...
- ...to be consumed by algorithmic models (e.g., predictive maintenance, or personalization recommendation)...
- ...with little man-in-the-middle intervention.
- Many IoT devices use JavaScript Object Notation (JSON) as the defacto lightweight data interchange format. It is easy for machines to parse and generate.
- This means, that over time, a vast amount of Operational Technology (OT) data will be created in JSON.
- There is a desire to link Information Technology (IT) or relational world data to OT data to create even better models.

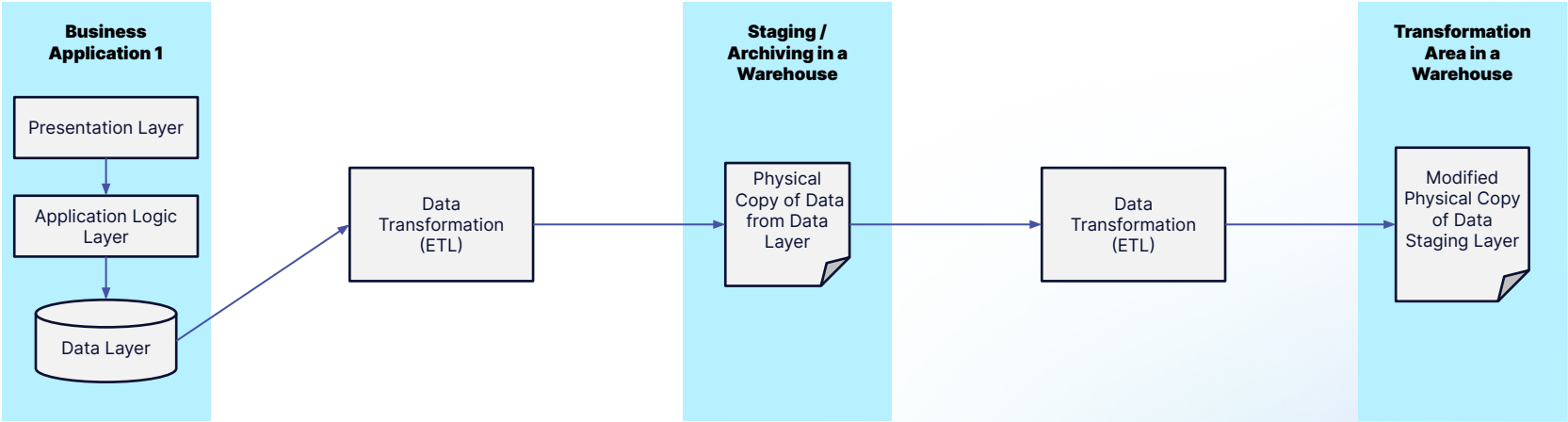
II. The Legacy Data Problem

The Problem In A Nutshell

Common Conceptual Data Architecture Pattern in Most Large Enterprises



Barrier 1: Contextless Data Proliferation



- Data created by app, for the app. Mostly for Operational use cases
- Semantic context for Data Layer embedded inside the application logic layer

- Data saved for a consumer other than the originating app
- But data is flat and has no context!!

- The copy has a different schema than the original
- Any consumer can create their own consumption schema on a use case by use case basis!

Example: Data from an ERP System

mandt	matnr	ersda	ernam	laeda	aenam	vpsta	pstat	lvorm	mtart	mbrsh	matkl	bism
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2020-05-21T00:00:0...	PHOU	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1005	22
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-06-20T00:00:0...	MSMITH3	KDELBSVQGCZX	KDELBSVQGC	X	Z1H	P	1000	103
400	000000000000000013	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	101
400	000000000000000072	2001-06-04T00:00:0...	DNOBLE	2020-02-04T00:00:0...	RJACKAN	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	106
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-11-29T00:00:00...	LSANTACR	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	107
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-01-25T00:00:0...	MSMITH3	KDELBSVQGCZX	KDELBSVQGC	X	Z1H	P	1000	110
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1005	22
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2020-05-21T00:00:0...	PHOU	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1005	22
400	000000000000000001	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	101
400	000000000000000019	2001-06-04T00:00:0...	DNOBLE	2019-06-20T00:00:0...	MSMITH3	KDELBSVQGCZX	KDELBSVQGC	X	Z1H	P	1000	104
400	000000000000000061	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	110
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2020-05-21T00:00:0...	PHOU	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1005	22
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	105
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	110
400	000000000000000103	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1005	22
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1005	22
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	107
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	null	Z1H	P	1000	106
400	0000000000000000...	2001-06-04T00:00:0...	DNOBLE	2019-07-03T00:00:0...	BSAFINEJ	KDELBSVQGCZX	KDELBSVQGC	X	ZOH	P	1100	103

- What is this...?!!
- What do these columns mean...?

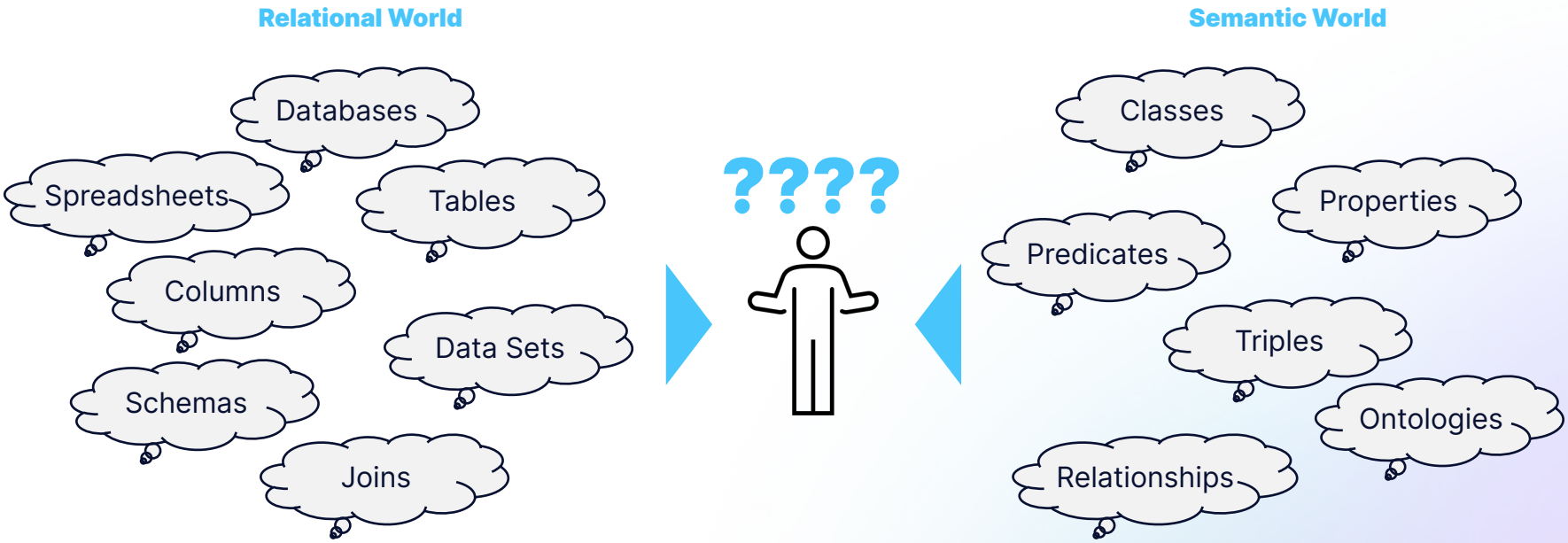
ERP System Data Dictionary

CONCEPTS (239)

Concepts	Definitions
Add Concept...	Add definition...
MANDT	Client
MATNR	Material Number
ERSDA	Created On
ERNAM	Name of Person who Created the Object
LAEDA	Date of Last Change
AENAM	Name of Person Who Changed Object
VPSTA	Maintenance status of complete material
PSTAT	Maintenance status
LVORM	Flag Material for Deletion at Client Level
MTART	Material Type
MBRSH	Industry Sector
MATKL	Material Group
BISMT	Old material number
MEINS	Base Unit of Measure
BSTME	Purchase Order Unit of Measure
ZEINR	Document number (without document management system)
ZEIAR	Document type (without Document Management system)

- OK, this table does have a meaning…
- It's just that the context resides in some document somewhere else **apart from the data** (e.g., some catalog system or document repository)
- Now multiply this problem for hundreds of apps and tens of thousands of database tables

Barrier 2: Mental Model and Syntax



- Two completely different worlds
- Very few people in an organization can operate proficiently in both
- Years of legacy investment, time, training in Relational World

III. The Innovations

Two Technical Innovations

- **JSON Linked Data Standard (JSON-LD)**: a nice go-between Relational World and Semantic World

- **Robust data classification AI Models**: a method for layering semantics to flat data at the massive scale required

What is JSON Linked Data? (JSON-LD)

- A method for encoding linked data ("semantic context") using JSON
- Easy to link ontologies to JSON documents and map to RDF

```
{ "@context": {
      "name": "http://xmlns.com/foaf/0.1/name",
      "homepage": {
            "@id": "http://xmlns.com/foaf/0.1/workplaceHomepage",
            "@type": "@id"
      },
      "Person": http://xmlns.com/foaf/0.1/Person
 },
 "@id": "https://me.example.com",
 "@type": "Person",
 "name": "John Smith",
 "homepage": "https://www.example.com/"
}
```

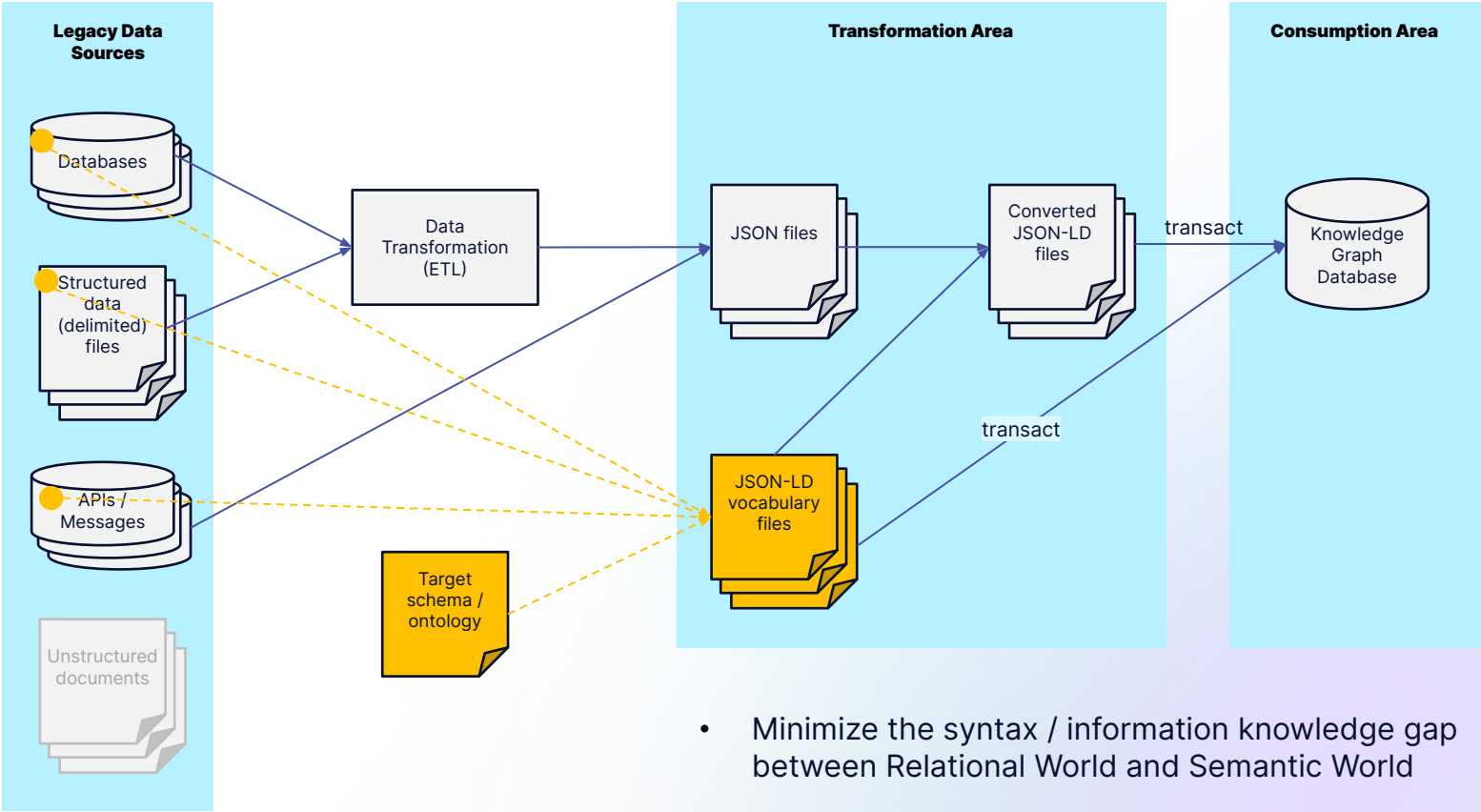
Describes the vocabulary being used to describe a Person entity

Why does JSON-LD matter?

- IoT transacts in JSON
- Lots of APIs transact in JSON
- Very easy to ETL from CSV to JSON
- JSON-LD is not a huge leap from JSON
- Lots of websites use JSON-LD for Google Search Engine Optimization
- Generative AI uses it a lot!!! Easy fit for LLMs

JSON-LD as a bridge to RDF / Triple Stores

- Represents a DTD, DDL, schema or data dictionary that describes the file located somewhere (inside the app logic layer, catalog, functional requirements doc, etc.)



- Minimize the syntax / information knowledge gap between Relational World and Semantic World

Using Data Classification Models to add semantics at scale

- Machine learning models can be trained how to tag columns to vocabulary terms and then infer vocabulary to vocabulary synonym terms
- Once the model is trained, you can run an automated process to scan all columns and then classify them to their respective terms.
- Once complete, you can convert the discovered mappings and semantic relationships into json-ld files that can be transacted into a Knowledge Graph.
- There are many ML classification models; for this talk we will go through an example using a Neural Network

AI/ML Data Classification Methodology

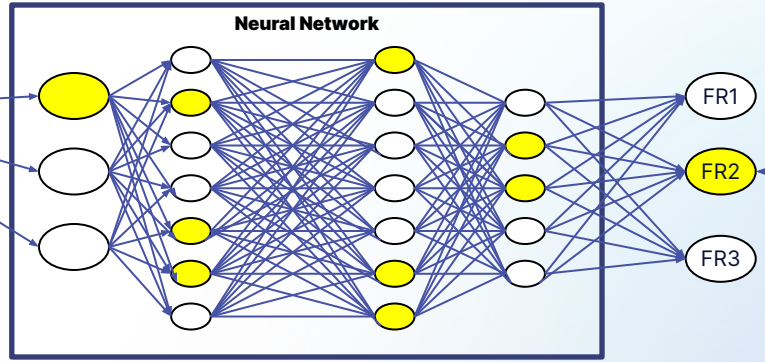
Training Data

Col A	Col B	Col C	Col D
00001	Apple	Red	Shiny
00002	Banana	Yellow	Long
00003	Grape	Purple	Squishy

Classifiers (Vocabulary for an Entity Called "Fruit")

ID	Term	Definition
FR1	Fruit Name	The name of the fruit
FR2	Fruit Color	The hue or color of the fruit
FR3	Characteristic	Defining feature whether shape or texture used to best describe the fruit

Col C
Red
Yellow
Purple



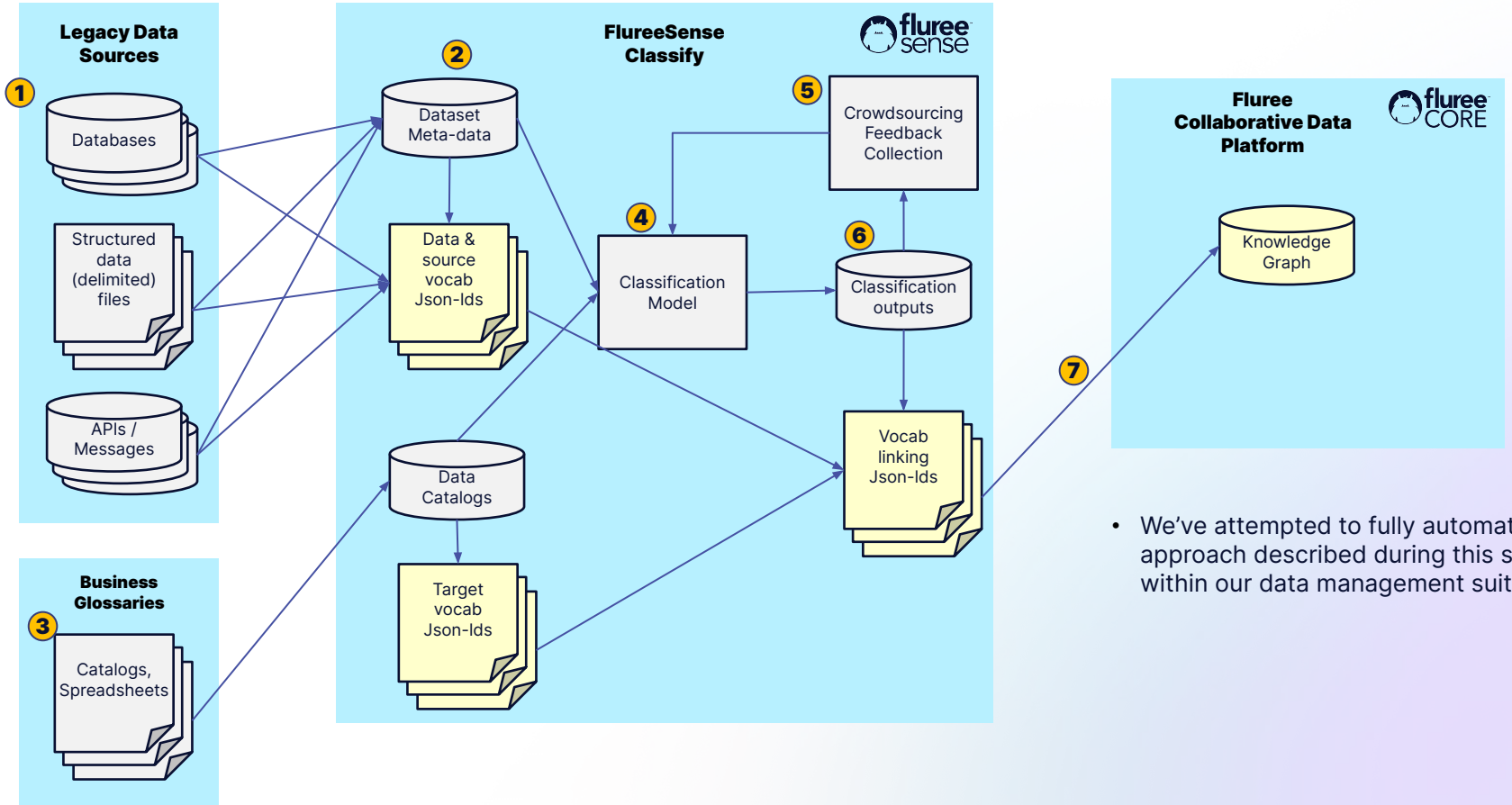
Col	Mapping
Col B	FR1
Col C	FR2
Col D	FR3

Scaling the Process Through Crowdsourcing

- The power of the approach is to distribute the training to as many people as possible
- Tens of people can individually classify 1-2 tables, or validate the machine prediction and give it reinforcement
- This will generate thousands or tens of thousands of new features to further refine the model

IV. The Demo

Fluree's Approach to Legacy Data Conversion



- We've attempted to fully automate the approach described during this session within our data management suite

Step 1. Connect to Data Sources

fluree sense Classify Start by Default search

Pharma Demo User
Sign Out

Data Sources

Create New Data Source

DATA SOURCES (10)

Data Source Name	Data Source Type	Created By	Last Modified	# of Datasets	# of Group Entitlements	# of User Entitlements
IDMP Mapping Full	Hadoop	pharma_user	04/11/2023	10	1	6
IDMP Mapping	Hadoop	pharma_user	04/11/2023	14	1	6
SAP ECC	Hadoop	pharma_user	03/29/2023	2	1	6
GSRS GSRS registration system for the ingredients in medicinal products	Hadoop	pharma_user	04/7/2023	2	1	6
IQVIA Market intelligence of pharmaceutical customers, products and clinical inf...	Hadoop	pharma_user	04/11/2023	5	1	6
GEN Pharma Default Datasource Default Datasource created while creating a tenant	Hadoop	59_user	04/7/2023	13	1	6

Version: v0.2.1 Environment: fsqa

Step 2. Generate Data Set Profiles

fluree sense Classify | Start by Default | search | Pharma Demo User Sign Out

Data Sets > MARA

Data Set Description
SAP Material data used to retain pharma drug mate...
Object Tags
MARA | Medicinal Products | MedicinalProduct | ...

Data Source
SAP ECC
Data Set Type
Parquet

Total Records
273,105
Total of Columns
22

Last Modified On
03/29/2023
Classify Set

Overall quality score 100%

Data Set Attributes | Data Set Sample | Data Set Relationships | Data Refresh | Data Entitlements | Related Projects | Data Quality

Attribute	Data Type	Classification
ManufacturerCode	STRING	
ManufacturerDesc	STRING	
MAKTX	STRING	
mara_ZZ_PLANT	LONG	
STRU_DESC	STRING	
FAMI_DESC	STRING	
SUBF_DESC	STRING	
STRE_DESC	STRING	
PAFM_DESC	STRING	
PASI_DESC	LONG	
MATNR	LONG	
mara_MATNR	LONG	
WRKST	STRING	
ZZ_STRUCTURE	LONG	
ZZ_PFAMGRP	STRING	
mara_ZZ_PFAMGRP	STRING	
ZZ_PRODFM	STRING	
MTART	STRING	
UMREN	STRING	
LAEDA	DATE	

MedicinalProductName

Profiler Result
Total Size for Profiling: 273,105 Rows
Total Rows: 273,105
Total Columns: 22
Total Sampled Rows: 273,105
Last Extract on 03/29/2023

Results

uniques	13 (0.0%)
Nulls	252974.0 (92.63%)

Data Types

NUMBER	100.0%

Distribution Of Values

Min. Value	2
Max. Value	99

Regex Patterns

9	72.89%
99	27.11%

Top 10 Most Frequent

	252974
2	8794
8	4694
18	2408
99	1240
9	1186
95	536
98	396
16	357
	168

Top 10 Least Frequent

94	15
12	96
97	105
96	136
11	168
16	357
98	396
95	536
9	1186
99	1240

Classification Result

Run Model

Version: v0.2.1 Environment: fsqa

Step 3. Import Ontologies

fluree sense Classify Start by Default search

Pharma Demo User Sign Out

Data Catalogs

Create New Data Catalog

DATA CATALOGS (6)

Data Catalog	Data Catalog Description	Created By	Last Modified On	Semantic Objects	Concepts	Mapped Data Sources	Mapped Data Sets	Mapped Data Set Columns	# of Data Quality Rules	Measured Data Quality
Automation_catalogtest	null	Pharma Demo User	05/2/2023	0	0	0	0	0	Add	n/a
GEN Pharma Catalog	Business information model for Generics Pharmaceutical company	Pharma Demo User	03/30/2023	3	32	3	3	51	Add	n/a
Global Substance Registration System (GSRS)	Registration system for the ingredients in medicinal products. It makes it easier for regulators and other stak.	Pharma Demo User	02/27/2023	34	786	2	2	25	2	100%
Identification of Medicinal Products (IDMP)	IDMP is a suite of five standards developed within the ISO to facilitate the unique identification of medicinal .	Pharma Demo User	03/9/2023	3	90	3	3	32	2	100%
SAP ECC Data Dictionary	Data glossary for SAP ECC R3 system	Pharma Demo User	04/7/2023	75	4784	1	1	11	Add	n/a
zFlureeSense System Catalog	Out of the box system catalog and semantic entities	Fluree Sense	02/27/2023	2	262	1	1	2	Add	n/a

fluree | | Version: v0.2.1 Environment: fsqa

Step 3. Import Ontologies (cont'd)

fluree sense Classify | Start by Default | search | Pharma Demo User Sign Out

Data Catalogs > Identification of Medicinal Products (IDMP) > **Pharmaceutical Products**

Technical View | Business View

Pharmaceutical Products
Data elements and structures for unique identification and exchange of regulated pharmaceutical product information

3 Data Sources As of 5/8/2023 | 4 Data Sets As of 5/8/2023 | 0 Data Quality Rules As of 5/8/2023 | n/a Measured Data Quality As of 5/8/2023

CONCEPTS (19)

Concepts	Definitions	Tags	Actions	Synonyms	Mapped Data Columns	Train Model	# of Data Quality Rules	Measured Data Quality
Add Concept...	Add definition...		+					
Reference Strength	strength of an active substance(s) and/or specified substance(s) used as a reference from which the strength of an inv...	+ Add Tag		1	0		Add	n/a
Presentation Strength	quantity or range of quantities of the substance/specified substance present per unitary volume (or mass) expressed in...	+ Add Tag		1	1		Add	n/a
Pharmaceutical Product Identifier	unique identifier for a pharmaceutical product	+ Add Tag		2	3		Add	n/a
Placebo	inactive substance, treatment or procedure that is intended to provide baseline measurements for the experimental pro...	+ Add Tag		1	0		Add	n/a
Packaged Medicinal Product	medicinal product in a container being part of a package, representing the entirety that has been packaged for sale or s...	+ Add Tag		1	0		Add	n/a
Investigational Medicinal Product I...	unique identifier allocated to an investigational medicinal product supplementary to any existing identifier as ascribed b...	+ Add Tag		0	0		Add	n/a
Clinical Trial	investigation in human subjects intended to discover or verify the clinical, pharmacological and/or other pharmacodyna...	+ Add Tag		2	0		Add	n/a
Package Item		+ Add Tag		1	1		Add	n/a
Outer Packaging		+ Add Tag		1	0		Add	n/a
Intermediate Packaging		+ Add Tag		1	0		Add	n/a
Immediate Container		+ Add Tag		1	0		Add	n/a
Investigational Medicinal Product		+ Add Tag		0	0		Add	n/a
Medicinal Product Identifier	unique identifier allocated to a medicinal product supplementary to any existing authorization number as ascribed by a ...	+ Add Tag		0	0		Add	n/a
Medicines Regulatory Agency	institutional body that, according to the legal system under which it has been established, is responsible for the grantin...	+ Add Tag		1	0		Add	n/a
Mass Based Strength		+ Add Tag		1	0		Add	n/a
Concentration		+ Add Tag		1	0		Add	n/a
Activity Based Strength		+ Add Tag		1	0		Add	n/a

Version: v0.2.1 Environment: fsqa

Step 4. Train Classification Model

fluree sense Classify | Start by Default | search | Pharma Demo User Sign Out

Data Catalogs > Identification of Medicinal Products (IDMP) > Pharmaceutical Products > Pharmaceutical Product Identifier > **Train Model**

Review Prior Training Data — Select Data Sets — Map Data Columns — Finalize Training Data

Map the column from the data sets selected that accurately represent the Concept that is being trained.

TM TOTAL DATA ATTRIBUTES (8)

Data set	Attribute	Mapping(0)
IDMP Medicinal Products - IMS	Manufacturer	
IDMP Medicinal Products - IMS	MedicinalProductIdentifier	
IDMP Medicinal Products - IMS	MedicinalProductName	
IDMP Medicinal Products - IMS	PackageItem	
IDMP Medicinal Products - IMS	PharmaceuticalProductIdentifier	
IDMP Medicinal Products - IMS	PresentationStrength	
IDMP Medicinal Products - IMS	RegistrationNumber	
IDMP Medicinal Products - IMS	id	

PHARMACEUTICAL PRODUCTS SEMANTIC OBJECT MODEL MAPPINGS (0 FIELDS MAPPED)

Concept	Mapping
Pharmaceutical Product Identifier	Drag & drop data set attibute here.

IDMP Medicinal Products - IMS
PharmaceuticalProductIdentifier

Cancel | Previous Step | Next Step

Version: v0.2.1 Environment: fsqa

Step 5. Review results and calibrate model

fluree sense Classify | Start by Default | search | Pharma Demo User Sign Out

Data Sets > MARA

Data Set Description: SAP Material data used to retain pharma drug mate...
Object Tags: MARA, Medicinal Products, MedicinalProduct ...
Data Source: SAP ECC
Data Set Type: Parquet
Total Records: 273,105
Total of Columns: 22
Last Modified On: 03/29/2023
Classify Set
Overall quality score: 100%

Data Set Attributes | Data Set Sample | Data Set Relationships | Data Refresh | Data Entitlements | Related Projects | Data Quality

Attribute	Data Type	Classification
ManufacturerCode	STRING	ManufacturerId
ManufacturerDesc	STRING	Manufacturer, ManufacturerName
MAKTX	STRING	Medicinal Product Name ...
mara_ZZ_PLANT	LONG	ZZ_PLANT
STRU_DESC	STRING	
FAMI_DESC	STRING	Agent Substance Name ...
SUBF_DESC	STRING	Substance Name Classifier-... ...
STRE_DESC	STRING	Strength
PAFM_DESC	STRING	
PASI_DESC	LONG	
MATNR	LONG	MATNR ...
mara_MATNR	LONG	MATNR
WRKST	STRING	Medicinal Product Name ...
ZZ_STRUCTURE	LONG	ZZ_STRUCTURE
ZZ_PFAMGRP	STRING	Medicinal Product Name ...
mara_ZZ_PFAMGRP	STRING	Medicinal Product Name ...
ZZ_PRODFM	STRING	MedicinalProductHierarchyG... ...
MTART	STRING	MTART
UMREN	STRING	
LAEDA	DATE	LAEDA

Profiler Result
Total Size for Profiling: 273,105 Rows
Total Rows: 273,105
Total Columns: 22
Total Sampled Rows: 273,105
Last Extract on 03/29/2023

Results

uniques	13 (0.0%)
Nulls	252974.0 (92.63%)

Data Types

NUMBER	100.0%

Distribution Of Values

Min. Value	2
Max. Value	99

Regex Patterns

9	72.89%
99	27.11%

Top 10 Most Frequent

	252974
2	8794
8	4694
18	2408
99	1240
9	1186
95	536
98	396
16	357
	168

Top 10 Least Frequent

94	15
12	96
97	105
96	136
11	168
16	357
98	396
95	536
9	1186
99	1240

MedicinalProductName

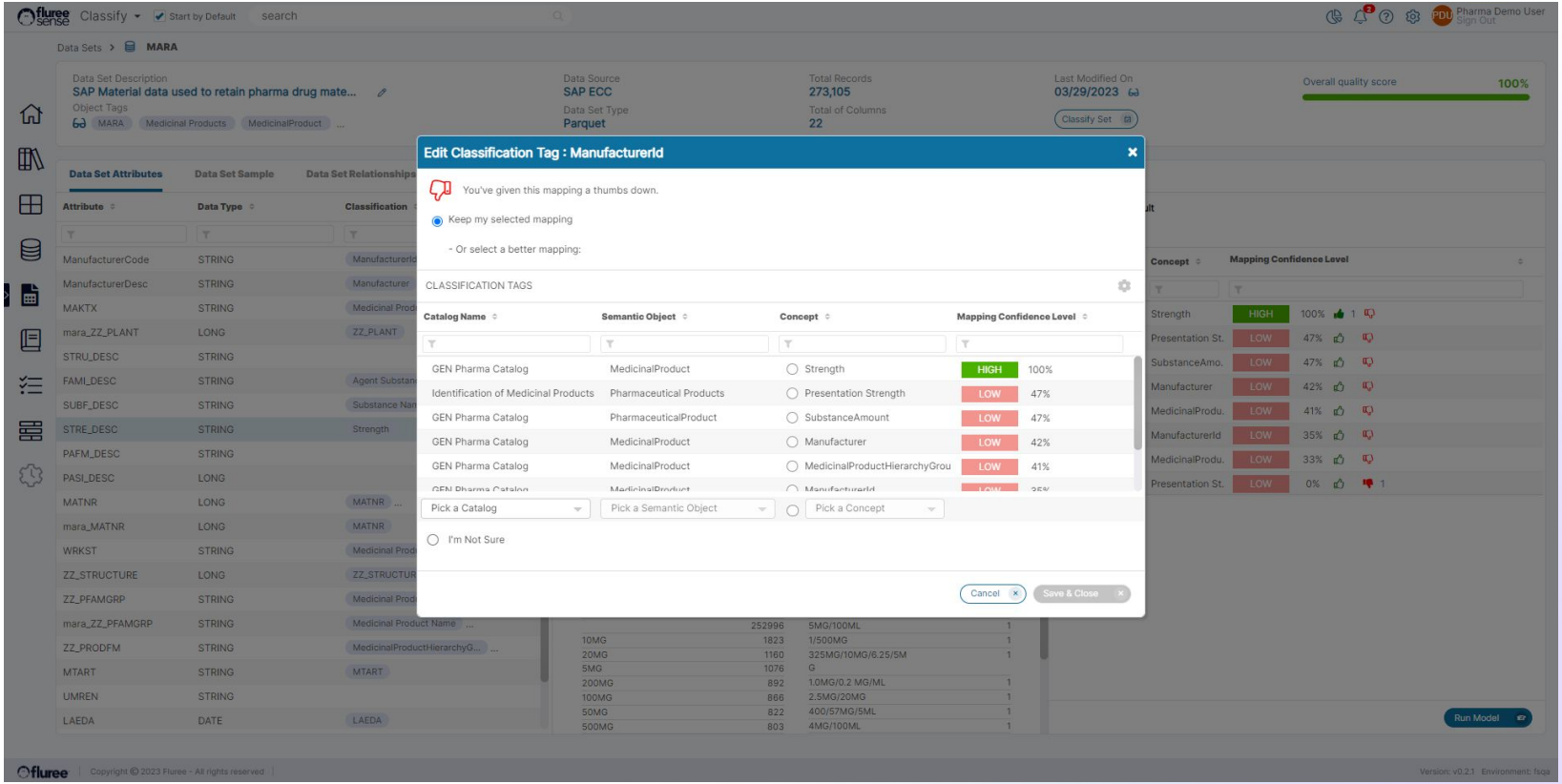
Classification Result

Semantic Object	Concept	Mapping Confidence Level		
MARA	ZZ_PLANT	HIGH	100%	1
MedicinalProduct	Pharmaceutical.	LOW	35%	
Medicinal Products	Pharmaceutical.	LOW	31%	
MedicinalProduct	MedicinalProdu.	LOW	0%	1
MedicinalProduct	MedicinalProdu.	LOW	0%	1

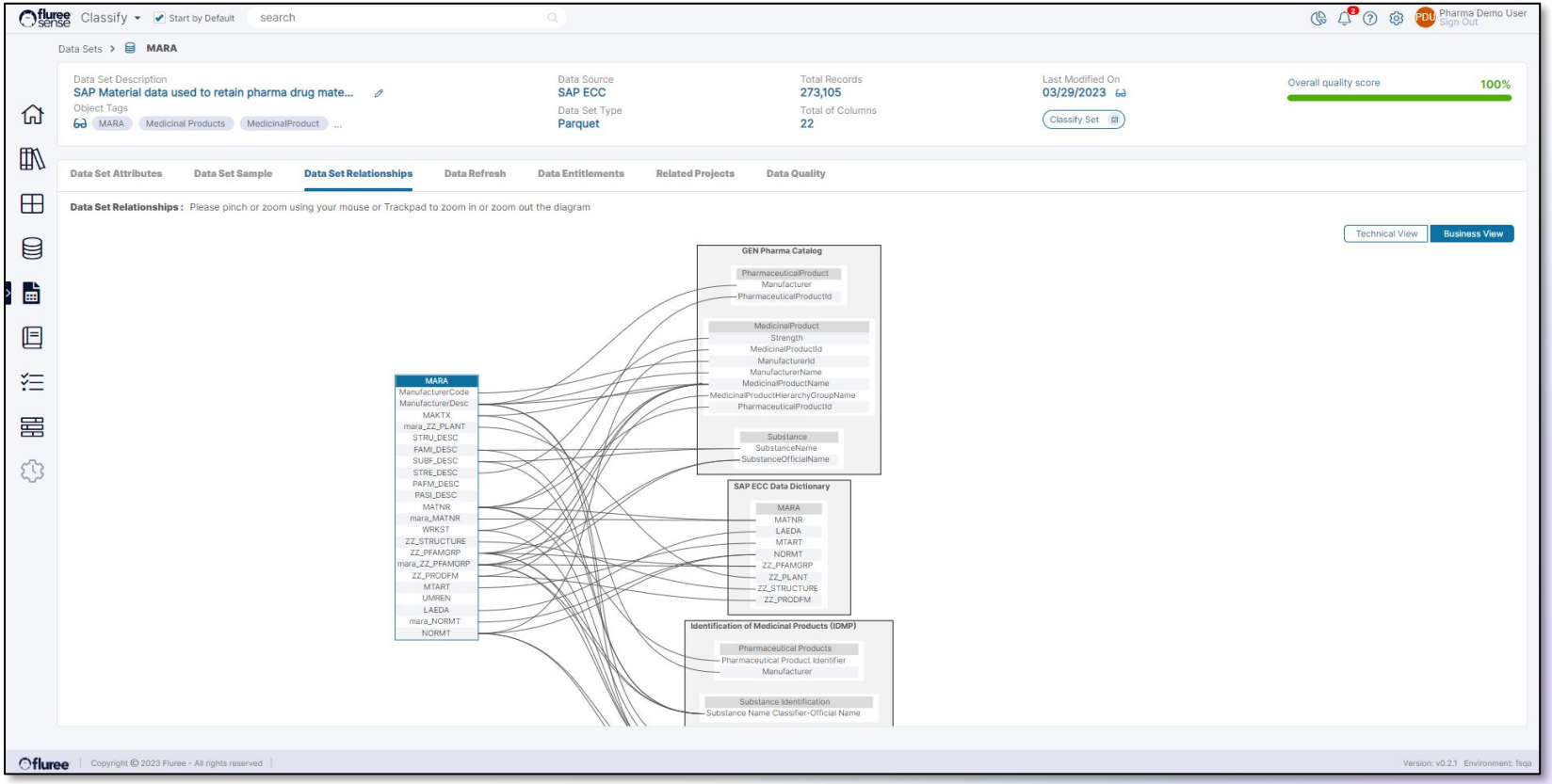
Run Model

fluree | Copyright © 2023 Fluree - All rights reserved | Version: v0.2.1 Environment: fsqa

Step 5. Review results and calibrate model (cont'd)



Step 5. Review results and calibrate model (cont'd)



Step 6. Validate Discovered Synonyms

fluree sense Classify Start by Default search Pharma Demo User Sign Out

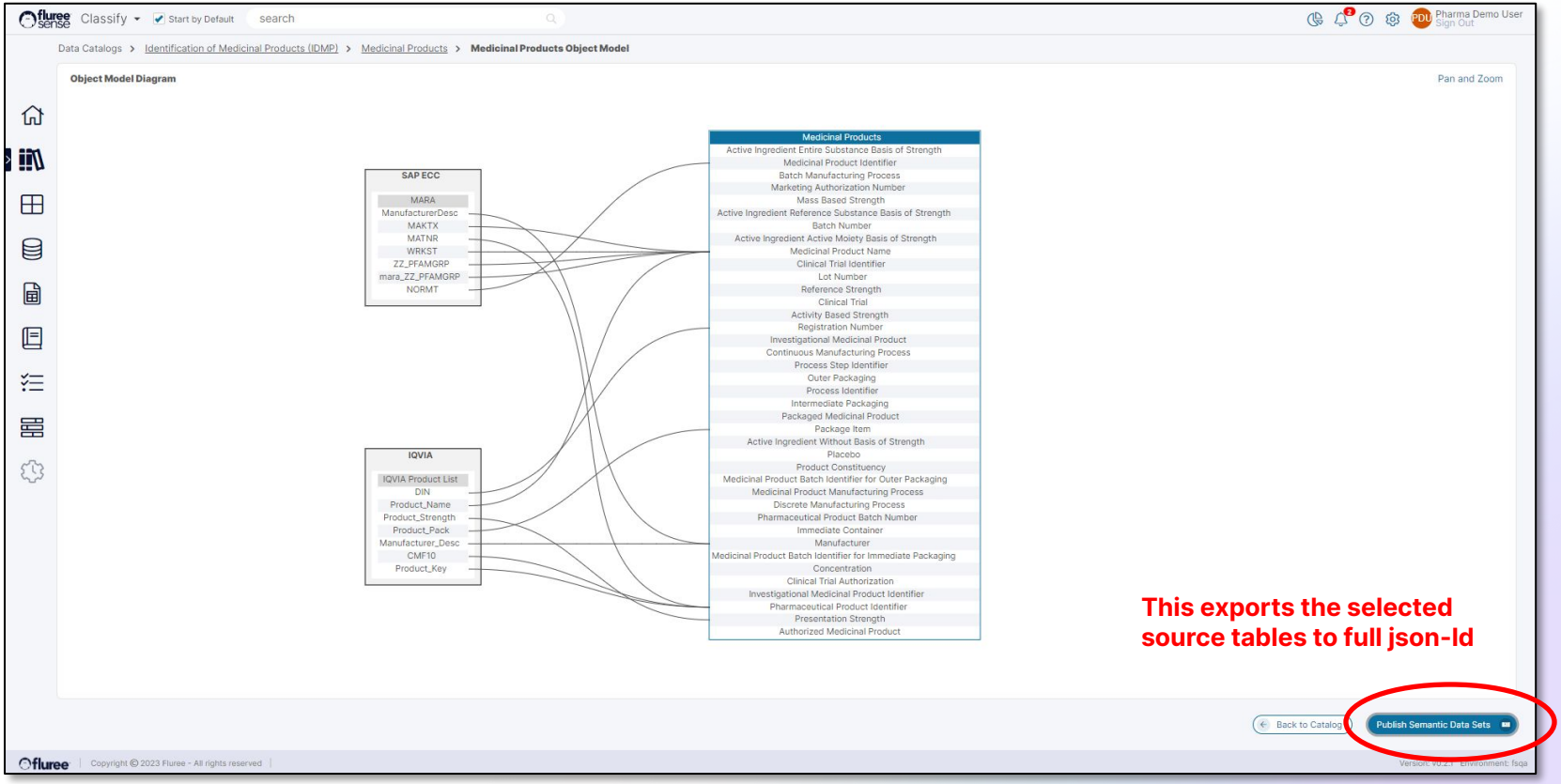
Synonyms

SYNONYM NAME: SUBSTANCE IDENTIFICATION:(15)

From Original Catalog: Identification of Medicinal Products (IDMP)			From Other Catalogs : Global Substance Registration System (GSRS)			
Catalog	Semantic Object	Concept	Synonym Catalog	Synonym Object	Synonym Concept	Synonym Confidence
Identification of Me...	Substance Identification	Moiety Identifier	Global Substance Registration System (GSRS)	Substance	Approval ID	HIGH 100% 1
Identification of Me...	Substance Identification	Molecular Structure	Global Substance Registration System (GSRS)	Structure	Moiety Formula	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Code	Global Substance Registration System (GSRS)	Substance	Approval ID	HIGH 100% 1
Identification of Me...	Substance Identification	Nucleic Acid Substance	Global Substance Registration System (GSRS)	NucleicAcid	Nucleic Acid UUID	HIGH 100% 1
Identification of Me...	Substance Identification	Polymer Substance	Global Substance Registration System (GSRS)	Polymer	Polymer UUID	HIGH 100% 1
Identification of Me...	Substance Identification	Protein Substance	Global Substance Registration System (GSRS)	Protein	Protein UUID	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Name Clas...	Global Substance Registration System (GSRS)	SubstanceReference	Agent Substance Name	HIGH 100% 1
Identification of Me...	Substance Identification	Excipient	GEN Pharma Catalog	Substance	SubstanceId	HIGH 100% 1
Identification of Me...	Substance Identification	Moiety	GEN Pharma Catalog	Substance	SubstanceId	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Code	GEN Pharma Catalog	Substance	SubstanceId	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Name Clas...	GEN Pharma Catalog	Substance	SubstanceName	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Name Clas...	GEN Pharma Catalog	Substance	SubstanceSynonymName	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Code	GEN Pharma Catalog	PharmaceuticalProduct	SubstanceId	HIGH 100% 1
Identification of Me...	Substance Identification	Substance Name Clas...	GEN Pharma Catalog	Substance	SubstanceOfficialName	HIGH 100% 1
Identification of Me...	Substance Identification	Molecular Structure	GEN Pharma Catalog	Substance	SubstanceMolecularStructureId	HIGH 100% 1

Save & Close Save & Run

Step 7. Publish Semantic Data Sets



This exports the selected source tables to full json-ld

V. Getting Started In Your Organization

Define Your Use Case

- Which business use case?
- Who are the users?
- Which ontology?
- Which data sources to start with?

Driver	Common Use Cases	Users	Ontology	Data Sources
LLM	• Internal knowledge mgmt • Business process optimization	• Internal transformation teams	• Business process models	• KM systems • Project management systems • Accounting systems
IT / OT Data Integration	• Supply Chain Optimization • Smart Factory Enablement	• Factory Operations	• Standardized Bill of Materials	• Sensors • ERP • SCADA • PLCs
Internal Data Sharing	• Marketing (Upsell / Cross-sell) • Corporate Functions data sharing (Risk, Finance, Compliance)	• Business End Users • Corporate Functions End Users	• Internal Business Information Models • Upper Ontologies	• Customer • Product • Account • Transaction
External Data Sharing	• Supply Chain Optimization • Industry Data Sharing	• Operations across external Partnerships	• Industry Standard Ontologies	• Manufacturing data

Run a FailFast PoC

Phase I: Data Acquisition	Phase II: Classification Training	Phase III: Data Consumption
• Connect to existing systems that you can easily access • Start with existing data warehouses and data lakes • Stand-up the analytic infrastructure	• Collect business information models and ontologies • Collect as much Source to Target mapping information, data catalogs, documented ETL logic • Import as training material for classification model • Conduct focused 2-week training with Human Subject Matter Experts	• Ingest data into Knowledge Graph • Generate test queries and results that demonstrate the business use case objectives
3-6 weeks	**4 weeks**	**3 weeks**

In Closing

- We’re at this interesting juxtaposition where the world is both waking up to the need for semantics, AND the tools and capabilities for creating it at scale are better than ever
- We hope we’ve shared practical concepts on how it can be done
- Please do not hesitate to reach out to me if you are interested further in pursuing this journey!!
- **Eliud Polanco – epolanco@flur.ee**

Questions?

Keep in touch with us

flur.ee/sense