# Building the Knowledge Graph Pipeline for Business Intelligence

**Dr. Bambang Purnomosidi D. P.**
Magister Teknologi Informasi
Intelligent Software Systems Research Group
Universitas Teknologi Digital Indonesia (https://www.utdi.ac.id/)

5th ISRITI

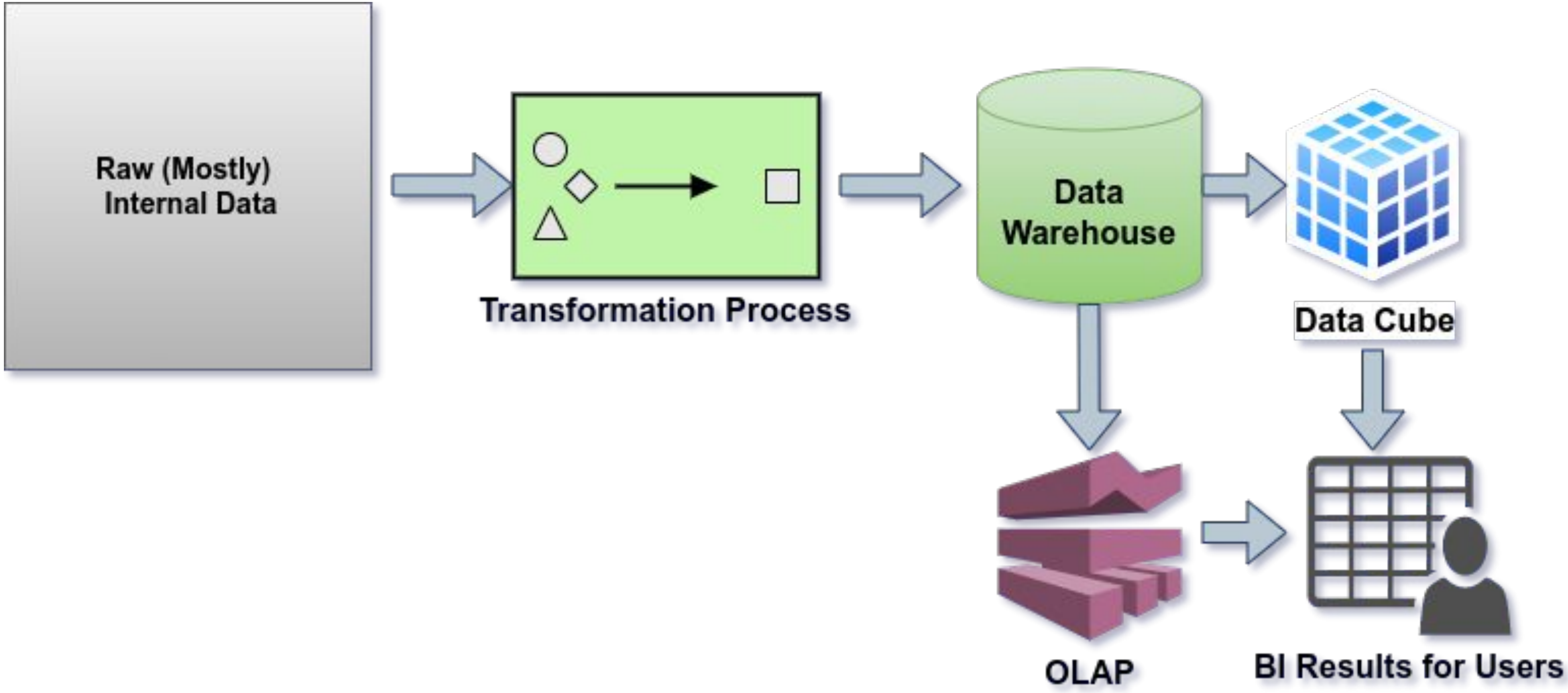
# Agenda

1. On Business and Competitive Intelligence
2. Components of (Traditional) BI
3. Data and BI
4. Data Warehouse, Data Lake, and Data Lakehouse.
5. Graph Data Model
6. Graph Database
7. Knowledge Graph
8. Knowledge Graph and BI
9. Knowledge Graph Pipeline

# On Business and Competitive Intelligence

- Harrison et. al. (2015): A business intelligence (BI) system is commonly known as a suite of technological solutions that facilitates organizations to amass, integrate and analyse vast stocks of data in order to understand their opportunities, strengths and weaknesses.
- A CI focuses on company's industry and industry rivals for better business decisions. CI is a subset of BI.
- BI == Decision Support System
- Primary purpose: to reduce uncertainties in decision making process. - ranging from operational to strategic.

# Components of (Traditional) BI



# Data and BI

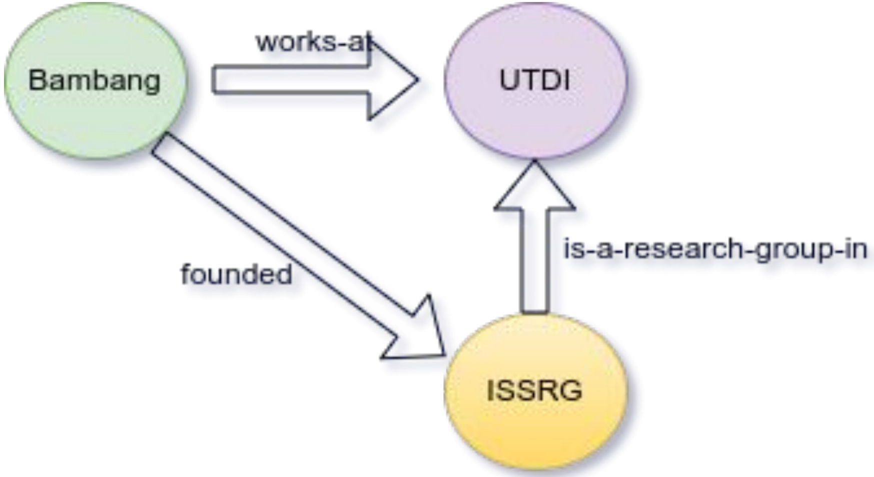
- Data consists of internal and external data
- Common data infrastructure:
  - Data Warehouse
  - Data Mart
- Problems:
  - Scattered data with so many formats (open specification - like XML, JSON, text file - or closed / proprietary - like MS Office file formats, PDF, etc).
  - Data models: SQL, NOSQL, etc.

# Data Warehouse, Data Lake, and Data Lakehouse

- **Data Warehouse**: mostly internal structured data, stored in multidimensional data.
    - => ETL
- **Data Lake**: structured, semi-structured, and unstructured (big) data from internal and external.
    - => ELT
- **Data Lakehouse**: combination of DW and DL.

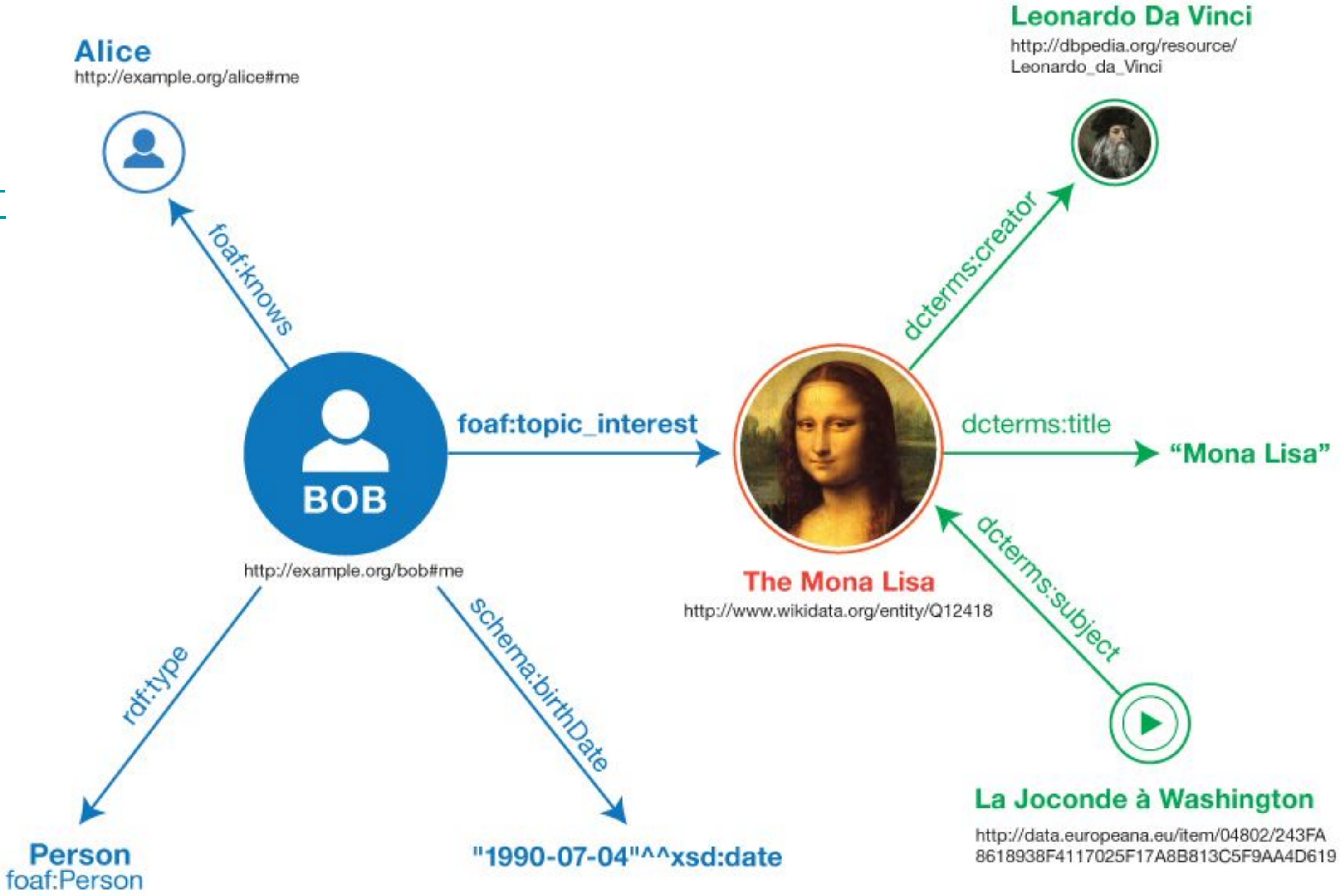
# Graph Data Model

- Triple: RDF (Resource Description Framework)
- Property Graph



**RDF**

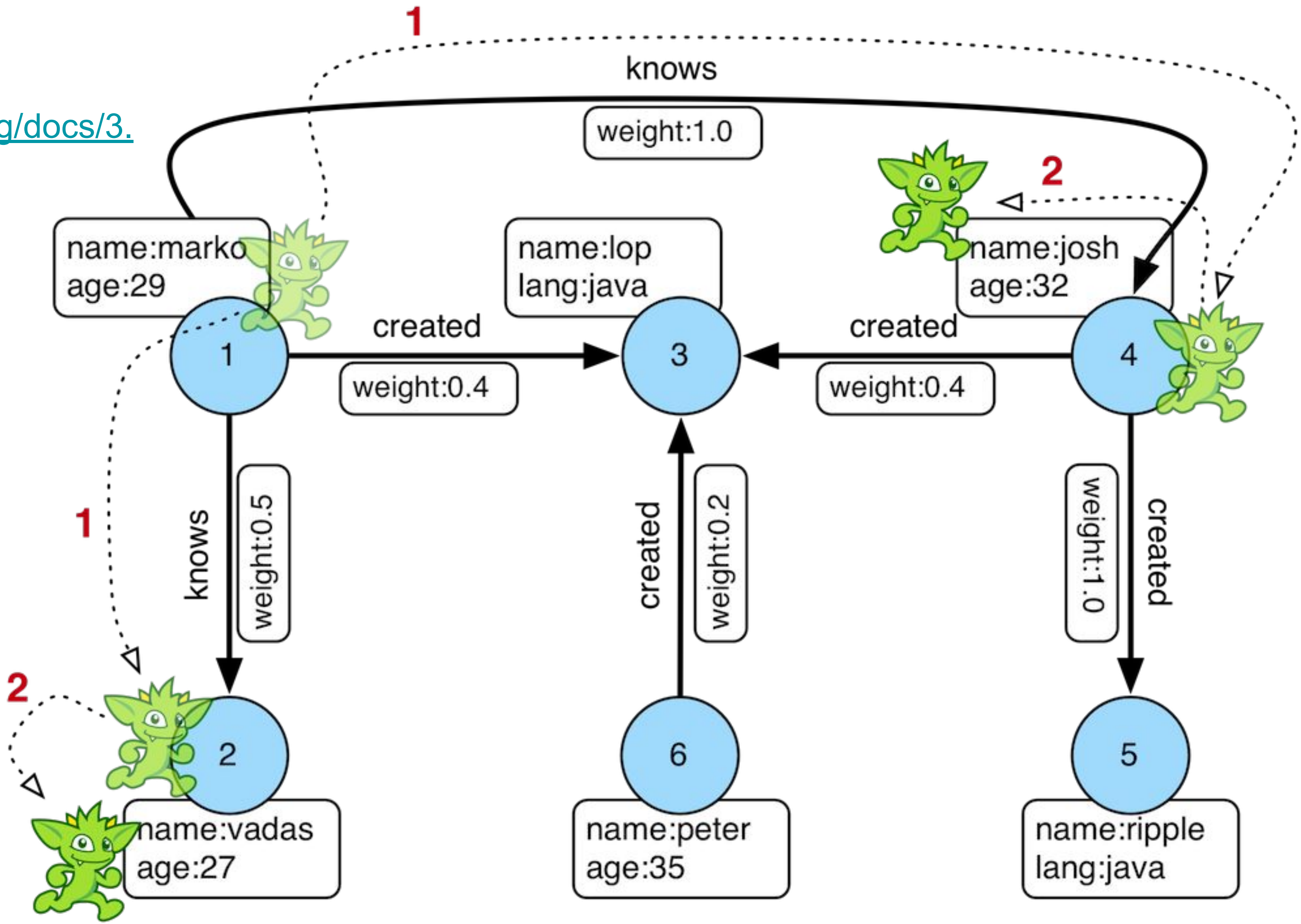
Source: https://www.w3.org/TR/rdf11-primer/


Alice
http://example.org/alice#me
foaf:knows
BOB
http://example.org/bob#me
foaf:topic_interest
The Mona Lisa
http://www.wikidata.org/entity/Q12418
Leonardo Da Vinci
http://dbpedia.org/resource/Leonardo_da_Vinci
dcterms:creator
dcterms:title
“Mona Lisa”
dcterms:subject
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619
rdf:type
Person
foaf:Person
schema:birthDate
"1990-07-04"^^xsd:date

```xml
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
         xmlns:dcterms="http://purl.org/dc/terms/"
         xmlns:foaf="http://xmlns.com/foaf/0.1/"
         xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:schema="http://schema.org/">
   <rdf:Description rdf:about="http://example.org/bob#me">
      <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
      <schema:birthDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1990-07-04</schema:birthDate>
      <foaf:knows rdf:resource="http://example.org/alice#me"/>
      <foaf:topic_interest rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://www.wikidata.org/entity/Q12418">
      <dcterms:title>Mona Lisa</dcterms:title>
      <dcterms:creator rdf:resource="http://dbpedia.org/resource/Leonardo_da_Vinci"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">
      <dcterms:subject rdf:resource="http://www.wikidata.org/entity/Q12418"/>
   </rdf:Description>
</rdf:RDF>
```

**Property Graph**

Source:
[https://tinkerpop.apache.org/docs/3.6.1/reference/](https://tinkerpop.apache.org/docs/3.6.1/reference/)



# Graph Database

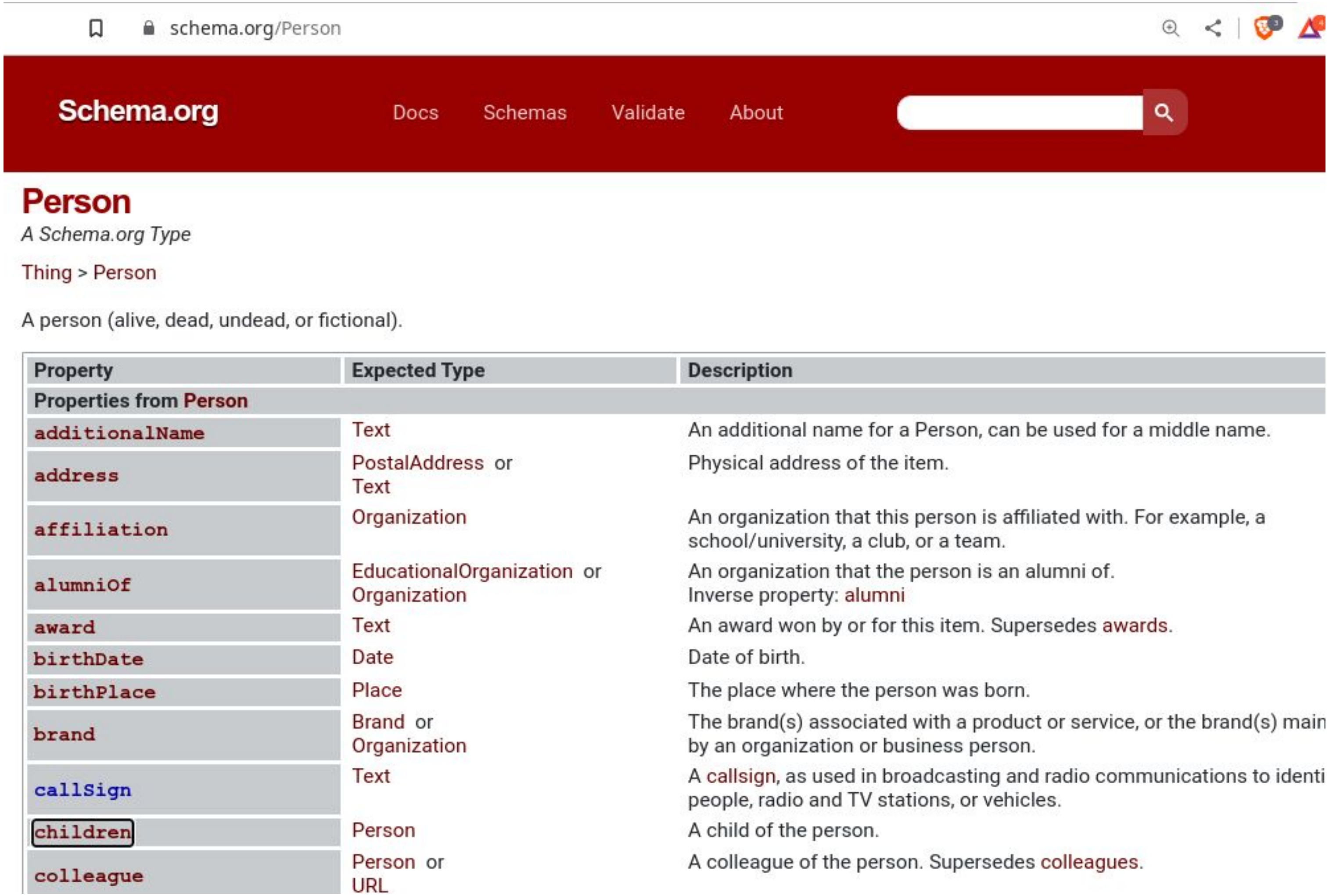
- RDF Database: TripleStore. Query language: SPARQL.
- Property Graph Database:
  - Neo4J: Cypher - OpenCypher
  - TinkerPop: Gremlin
  - Nebula Graph: nGQL
  - ArangoDB: AQL
  - Standard: GQL, evolved from OpenCypher

# Knowledge Graph

- KG term has been used since 1972. Next: WordNet (semantic relationship between words and meanings - 1985), DbPedia (general purpose knowledge - 2007), Google Knowledge Graph (2012).
- Also known as *Semantic Network*.
- KG: graph model to store **interlinked** descriptions of **entities** – objects, events, situations or abstract concepts – while also encoding the **semantics** underlying the used terminology.

## OpenCypher

```
CREATE (n:Person {name: 'Bambang Purnomosidi D. P.’, jobTitle: 'Researcher', worksFor: ‘UTDI’})
```



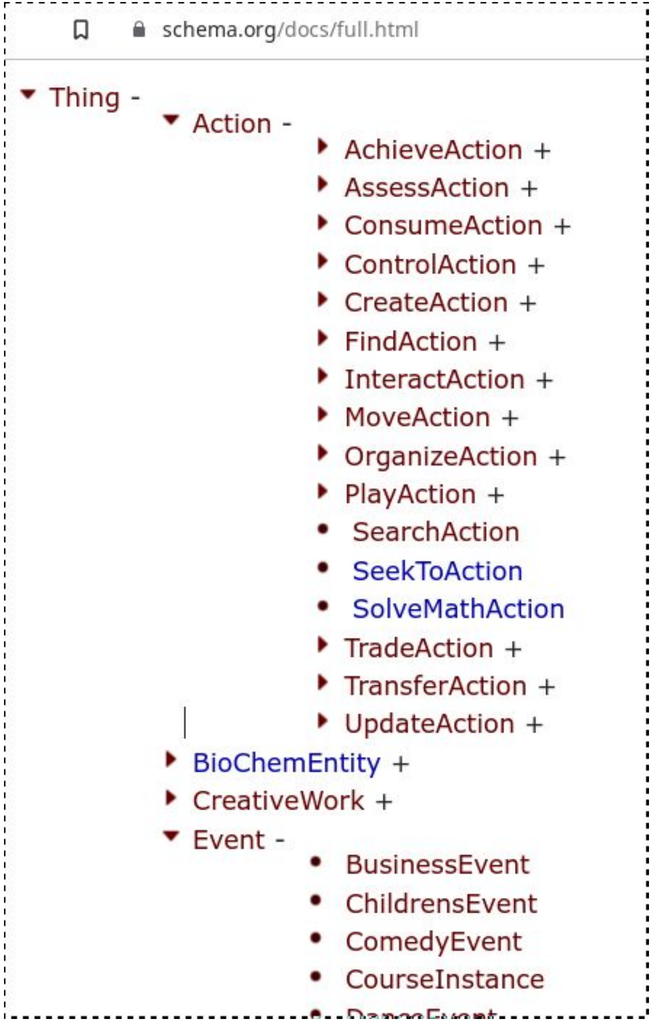
# Person

*A Schema.org Type*

Thing > Person

A person (alive, dead, undead, or fictional).

| Property | Expected Type | Description |
|---|---|---|
| Properties from Person | | |
| additionalName | Text | An additional name for a Person, can be used for a middle name. |
| address | PostalAddress or<br>Text | Physical address of the item. |
| affiliation | Organization | An organization that this person is affiliated with. For example, a school/university, a club, or a team. |
| alumniOf | EducationalOrganization or<br>Organization | An organization that the person is an alumni of.<br>Inverse property: alumni |
| award | Text | An award won by or for this item. Supersedes awards. |
| birthDate | Date | Date of birth. |
| birthPlace | Place | The place where the person was born. |
| brand | Brand or<br>Organization | The brand(s) associated with a product or service, or the brand(s) main by an organization or business person. |
| callSign | Text | A callsign, as used in broadcasting and radio communications to identi people, radio and TV stations, or vehicles. |
| children | Person | A child of the person. |
| colleague | Person or<br>URL | A colleague of the person. Supersedes colleagues. |



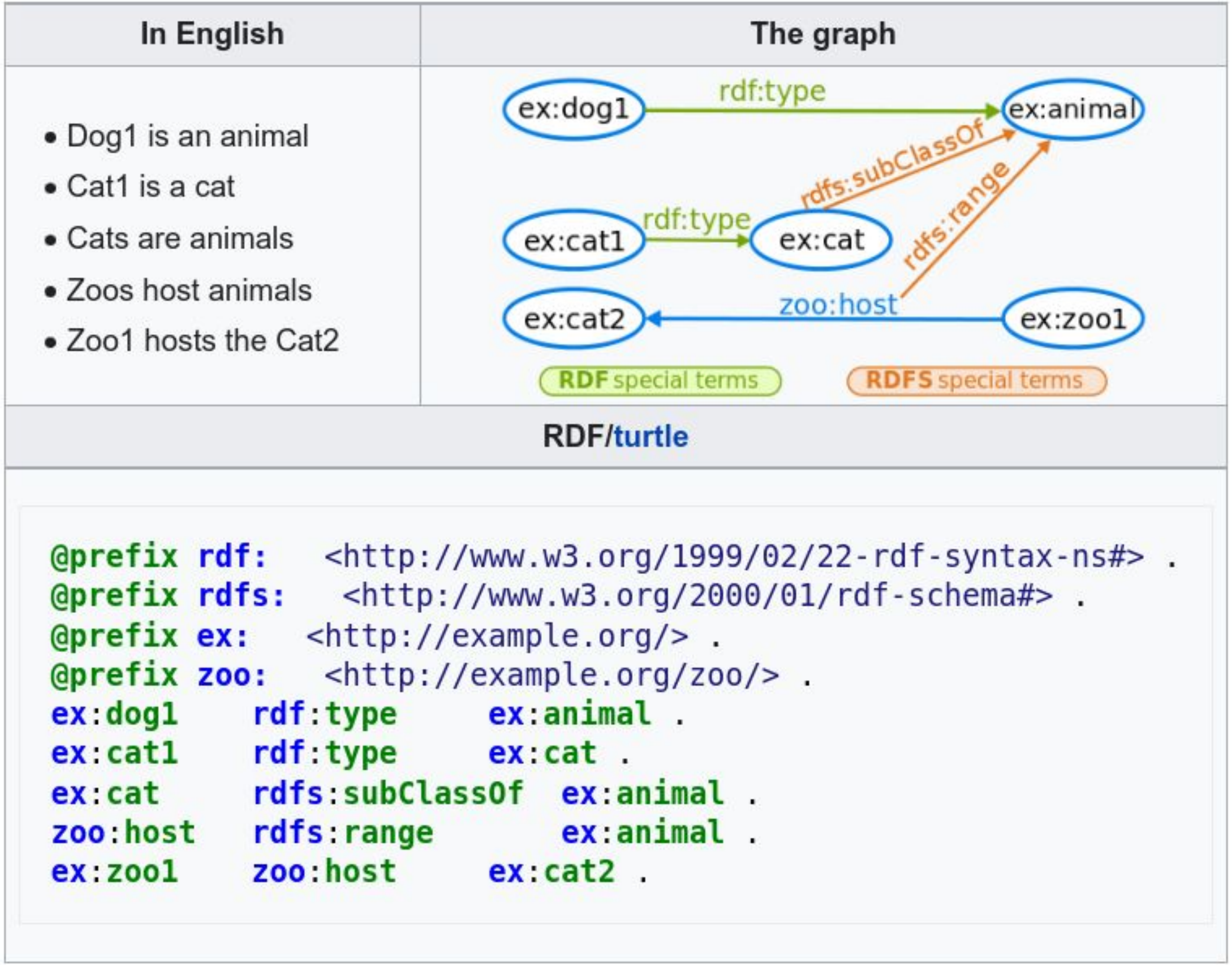
schema.org/docs/full.html

- Thing -
  - Action -
    - AchieveAction +
    - AssessAction +
    - ConsumeAction +
    - ControlAction +
    - CreateAction +
    - FindAction +
    - InteractAction +
    - MoveAction +
    - OrganizeAction +
    - PlayAction +
    - SearchAction
    - SeekToAction
    - SolveMathAction
    - TradeAction +
    - TransferAction +
    - UpdateAction +
  - BioChemEntity +
  - CreativeWork +
  - Event -
    - BusinessEvent
    - ChildrensEvent
    - ComedyEvent
    - CourseInstance
    - DanceEvent

Many schema available so that we can create entities, relationship between entities, and semantic for those entities.

# Triple

- Data in RDF
- Semantic / Ontology in RDFS / OWL



| In English | The graph |
| --- | --- |
| • Dog1 is an animal<br>• Cat1 is a cat<br>• Cats are animals<br>• Zoos host animals<br>• Zoo1 hosts the Cat2 | |

RDF/turtle

```
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex:   <http://example.org/> .
@prefix zoo:   <http://example.org/zoo/> .
ex:dog1    rdf:type        ex:animal .
ex:cat1    rdf:type        ex:cat .
ex:cat     rdfs:subClassOf    ex:animal .
zoo:host   rdfs:range         ex:animal .
ex:zoo1    zoo:host        ex:cat2 .
```

Source: https://en.wikipedia.org/wiki/RDF_Schema

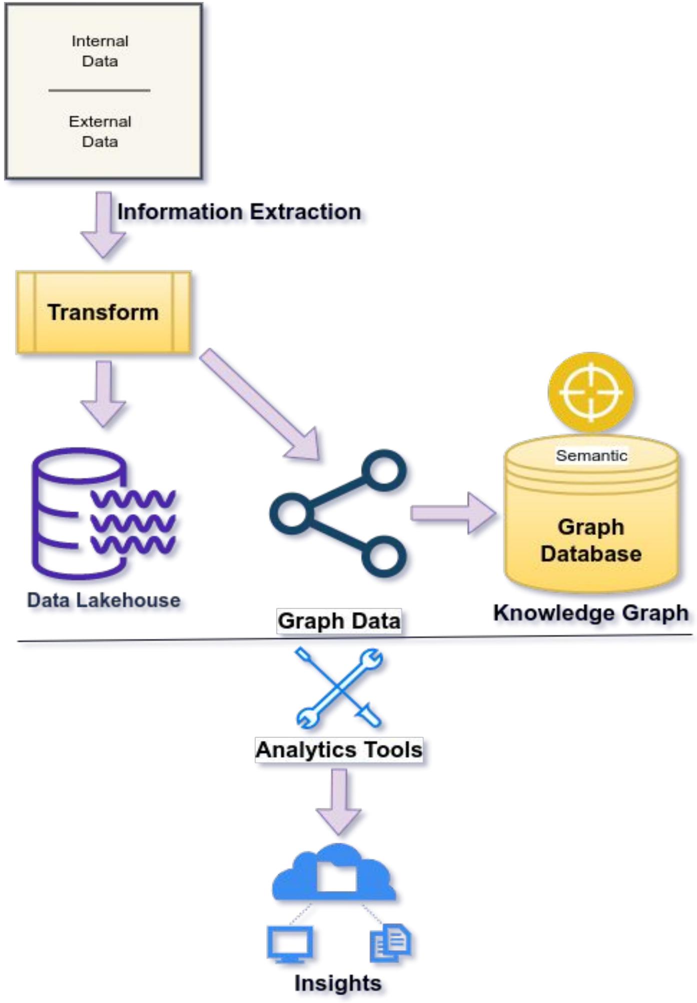
# Knowledge Graph and BI

- Many use cases on how business can benefit from knowledge graph + graph analytics..
- In **marketing**: recommender system  for social media influencer using PageRank - useful to choose social media influencer in a specific product domain.
- Related products-services recommendation.
- Path analysis for lowest expenses product delivery.
- Supply chain optimization using shortest path analysis to optimize routes.
- Current position in competition
- Finding potential customers
- Finding financial fraud inside an organization
- /etc

# Knowledge Graph Pipeline

Bird Eye View

Data Sources -> Transform Into Graph Data -> Combined with Semantic -> Knowledge Graph



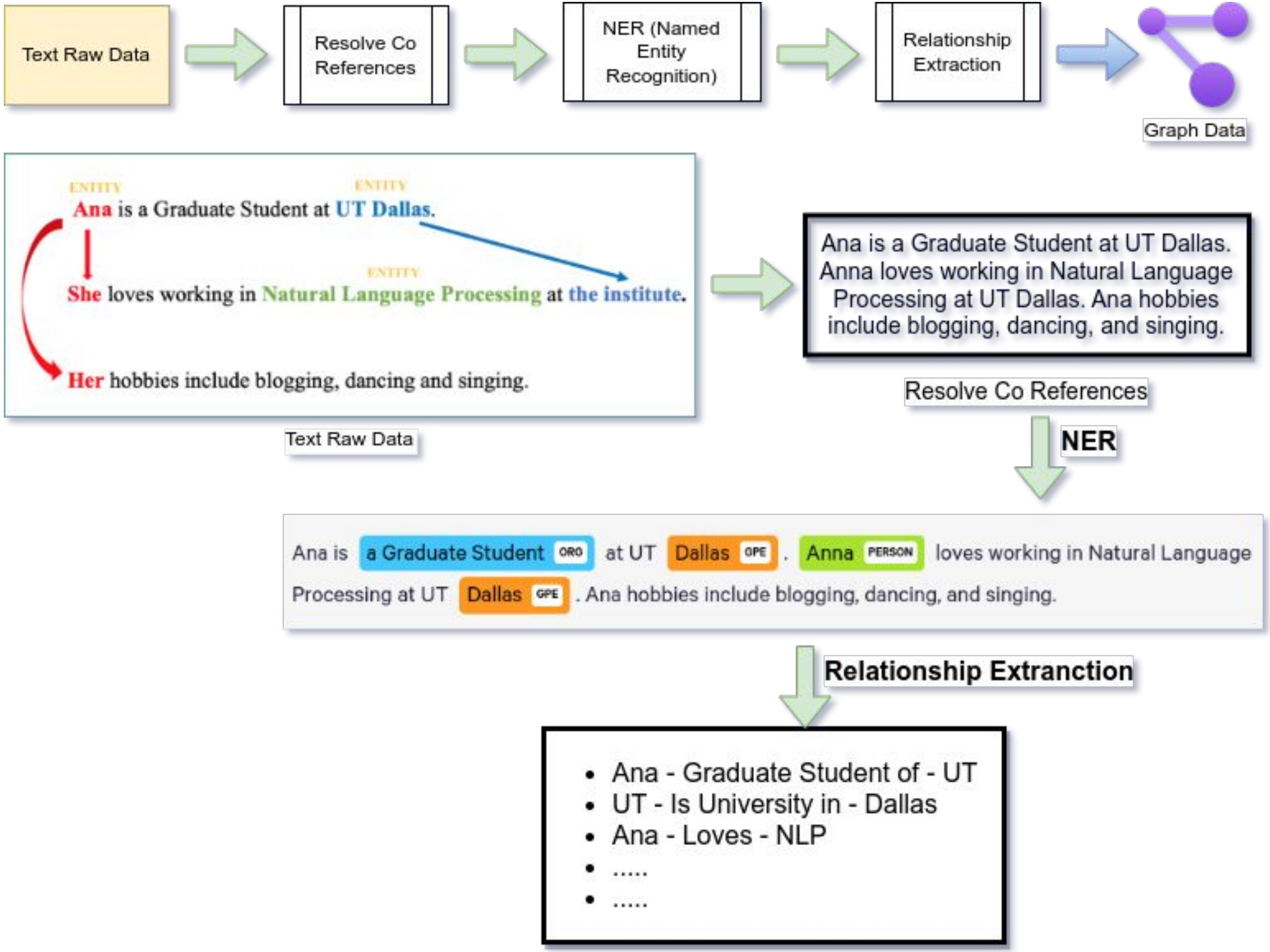
## Internal Data:

- Spreadsheet
- TPS (*Transaction Processing System*)
- SQL, NOSQL

## External Data

- Endpoint data
- HTML page
- Text

The most complex information extraction: text to graph.



Ana is a Graduate Student at UT Dallas.

She loves working in Natural Language Processing at the institute.

Her hobbies include blogging, dancing and singing.

Ana is a Graduate Student at UT Dallas. Anna loves working in Natural Language Processing at UT Dallas. Ana hobbies include blogging, dancing, and singing.

Ana is a Graduate Student at UT Dallas . Anna loves working in Natural Language Processing at UT Dallas . Ana hobbies include blogging, dancing, and singing.

- Ana - Graduate Student of - UT
- UT - Is University in - Dallas
- Ana - Loves - NLP
- .....
- .....