Istat's Reference Architecture for Big Data: Internet as a Data Source

Monica Scannapieco | Directorate for Methodology and Statistical Process Design, Istat

Joint work with

Antonino Virgillito
Directorate for Information and Communication Technology
Istat

Donato Summa, **Diego Zardetto**
Directorate for Methodology and Statistical Process Design
Istat

Istat | Istituto Nazionale di Statistica

Outline

- Why a Reference Architecture
- Logical Architectural Schemes for *Internet As a Data Source*
  - Web sites
  - Twitter data
- Conclusions

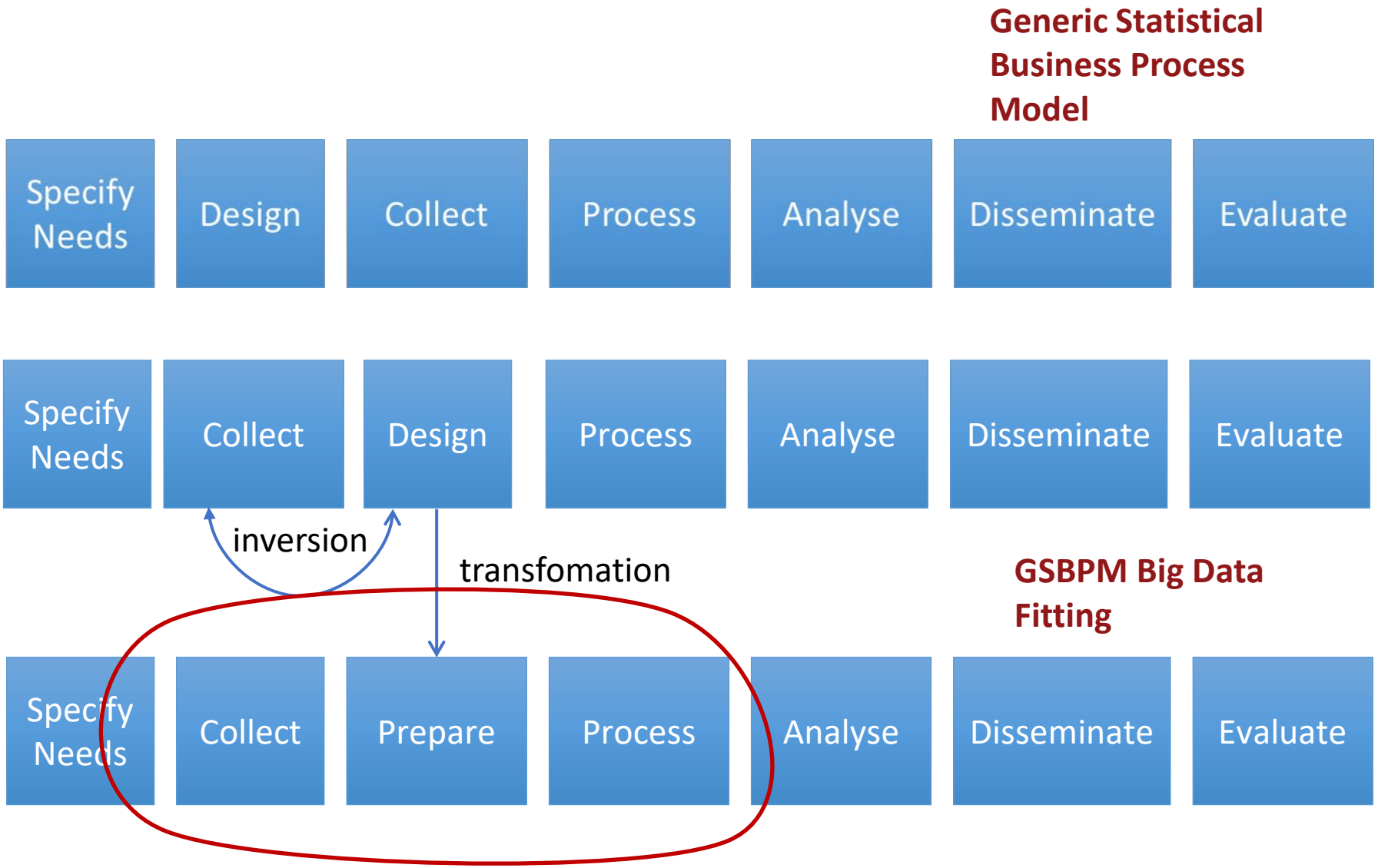
Why a Reference Architecture

- Few concrete results in terms of available architectural standards to rely on
  - Possibly due to the relatively recent attention that the Big Data phenomenon received by National Statistical Office (NSOs)
- Need of architectural standards
  - NSOs are starting investing now
  - Sharing and reuse easier if Big Data investments follow common guidelines
- Purpose of this contribution: to pose some initial bases towards an Istat's reference architecture for Big Data
  - Focus on Internet as a Data Source
  - Some findings achieved within the European project ESSnet on Big Data Pilots-I , already shared at EU level

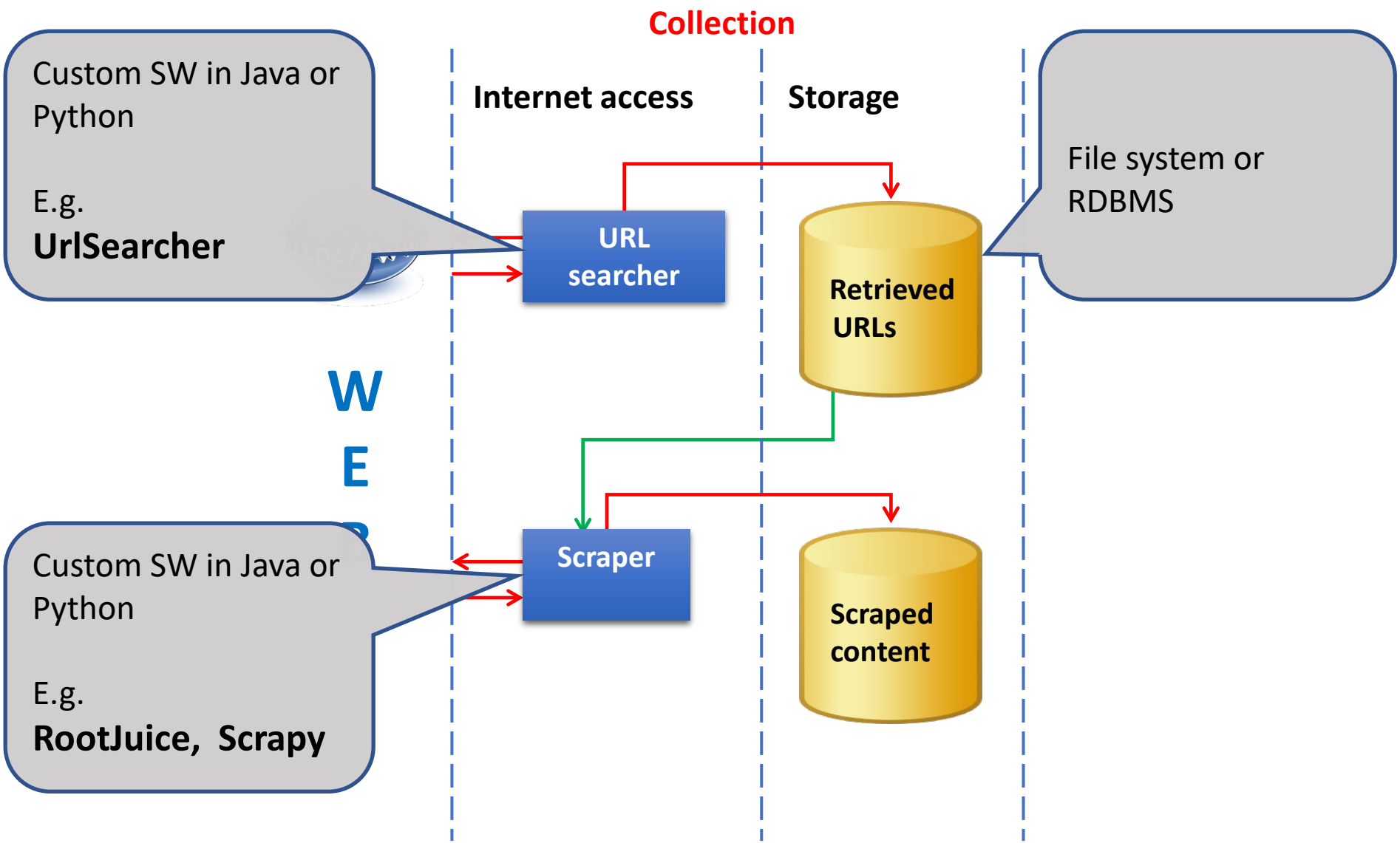
Internet as a Data Source

- Internet as a Data Source: set of Internet-accessible sources that can be used for collecting data considered as relevant for Official Statistical purposes
  - Web sites
  - Social media, specifically Twitter
- Applications (as described today)
  - Replacement/integration surveys and registers: ICT Usage Survey, Business Register, Online price statistics, etc.
  - New indicators: daily social mood on economy from tweets

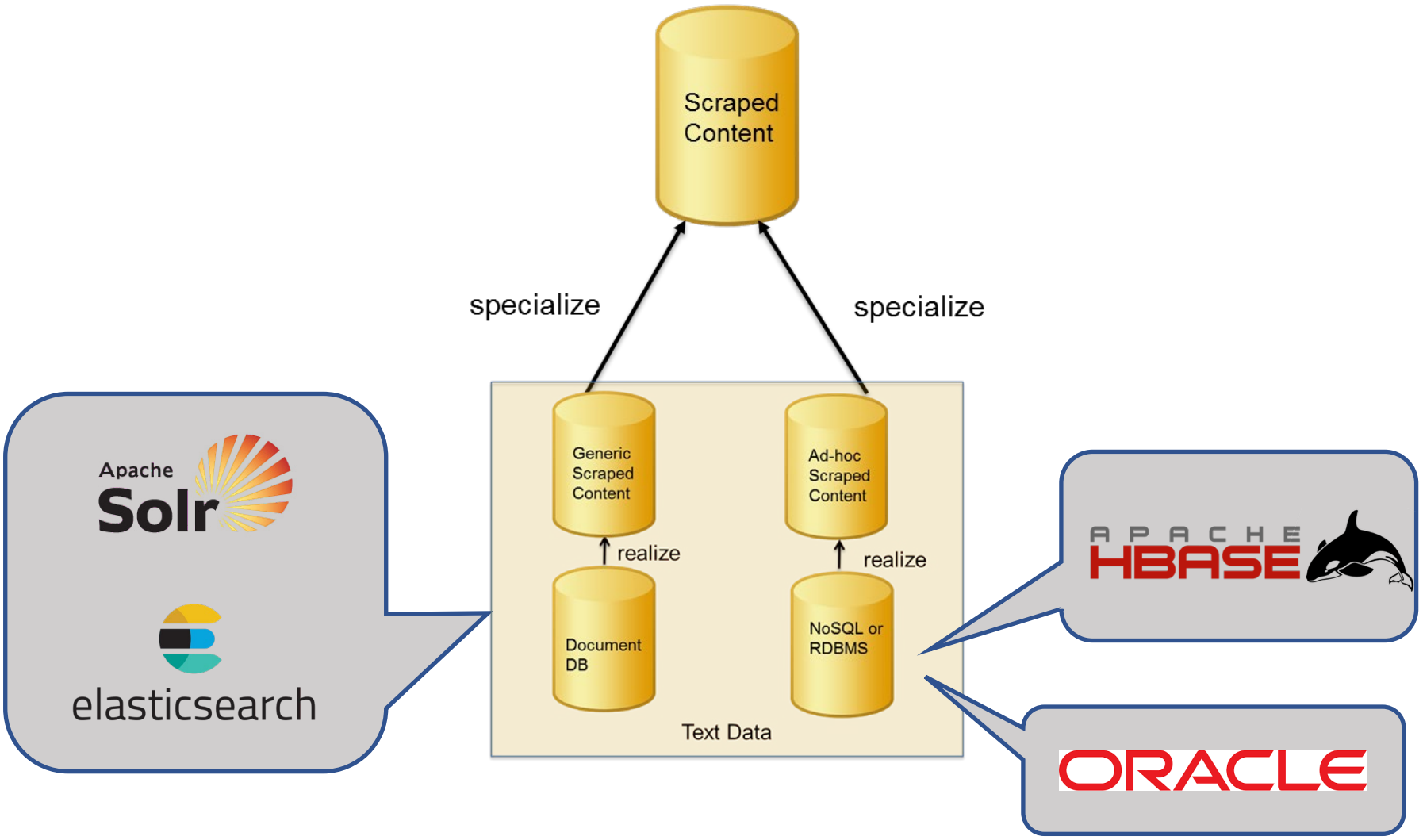
GSBPM Fitting for Big Data



Internet as a Data Source: Web Sites



Internet as a Data Source: Web Sites



Internet as a Data Source: Web Sites

Java or Python libraries

E.g.
NLTK (Python)
JSOUP (Java)
BeautifulSOUP (Python)

Preparation

Generic Preparation

- Special Chars Removal
- Tokenization
- Word filters (e.g. Stopwords removal)
- Basic orthographic repair
- Lemmatization
- Stemming

Task specific Preparation

- Data labelling and annotation (e.g. for supervised learning and entity extraction)
- Dimensionality reduction
- Text encoding
- Language modelling
- Knowledge representation
- Validation

Java,R or Python libraries E.g. **caret** (R)

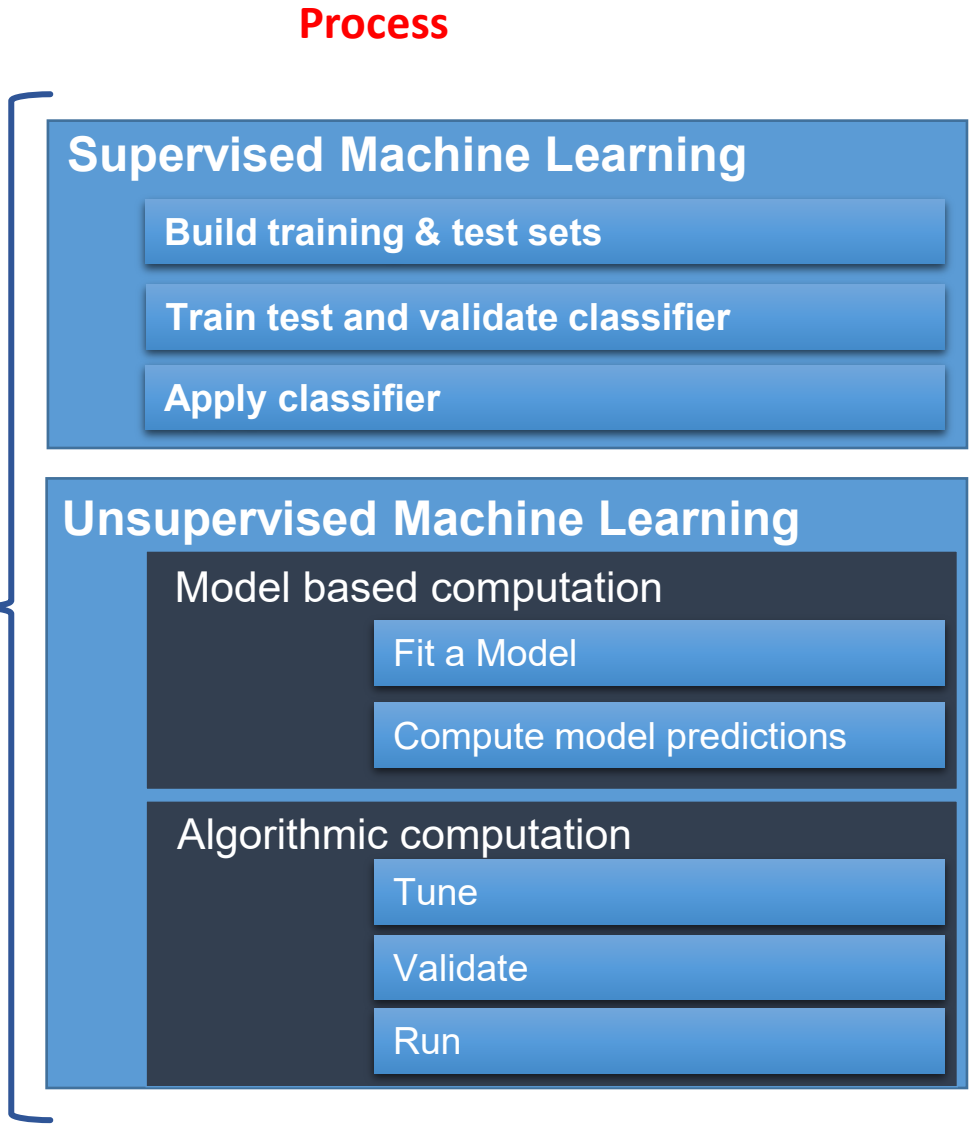
Word embeddings framework E.g. **GloVE** or **Word2vec**

Semantic Web languages E.g. **OWL**

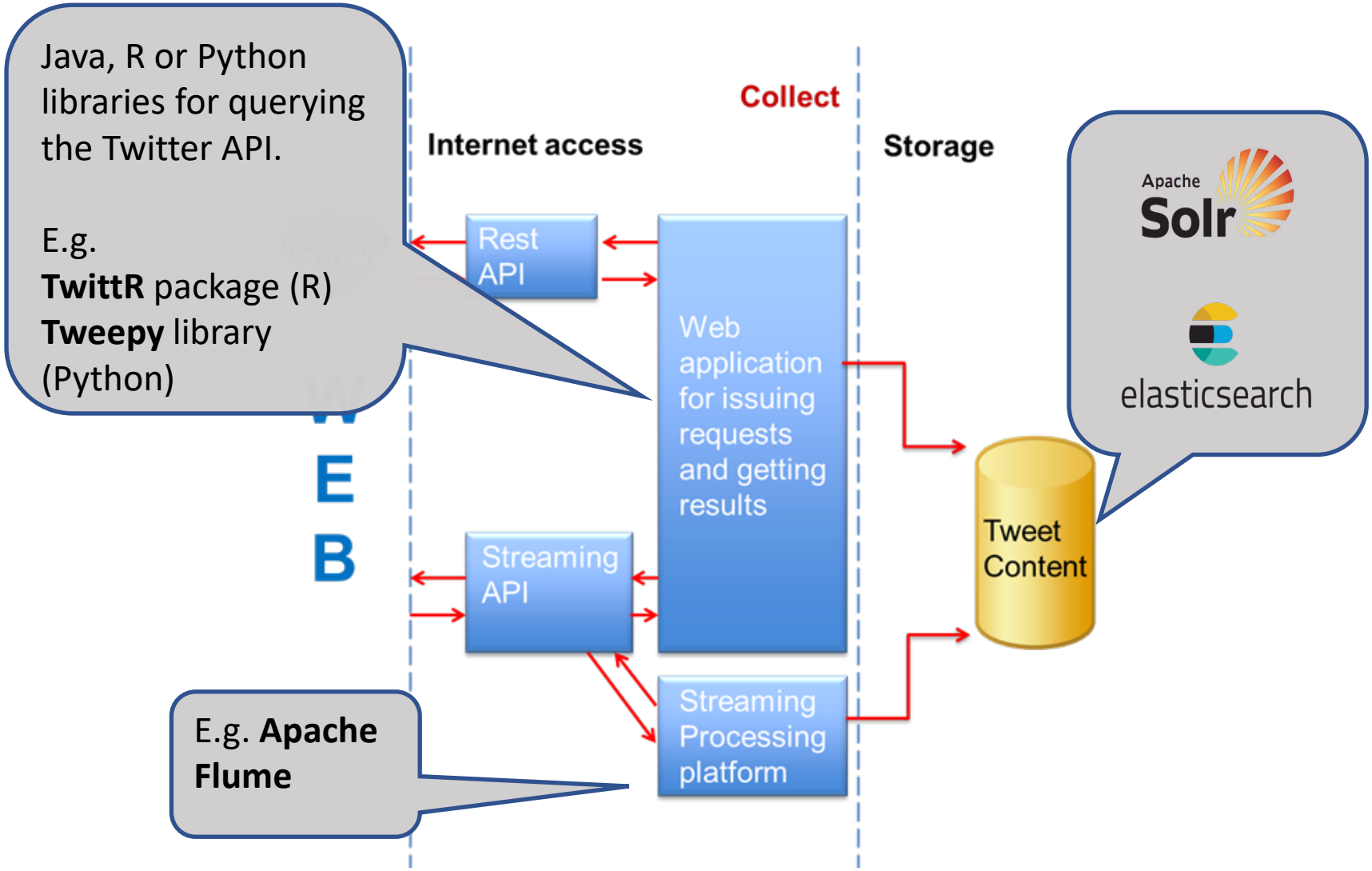
Internet as a Data Source: Web Sites

R or Python libraries

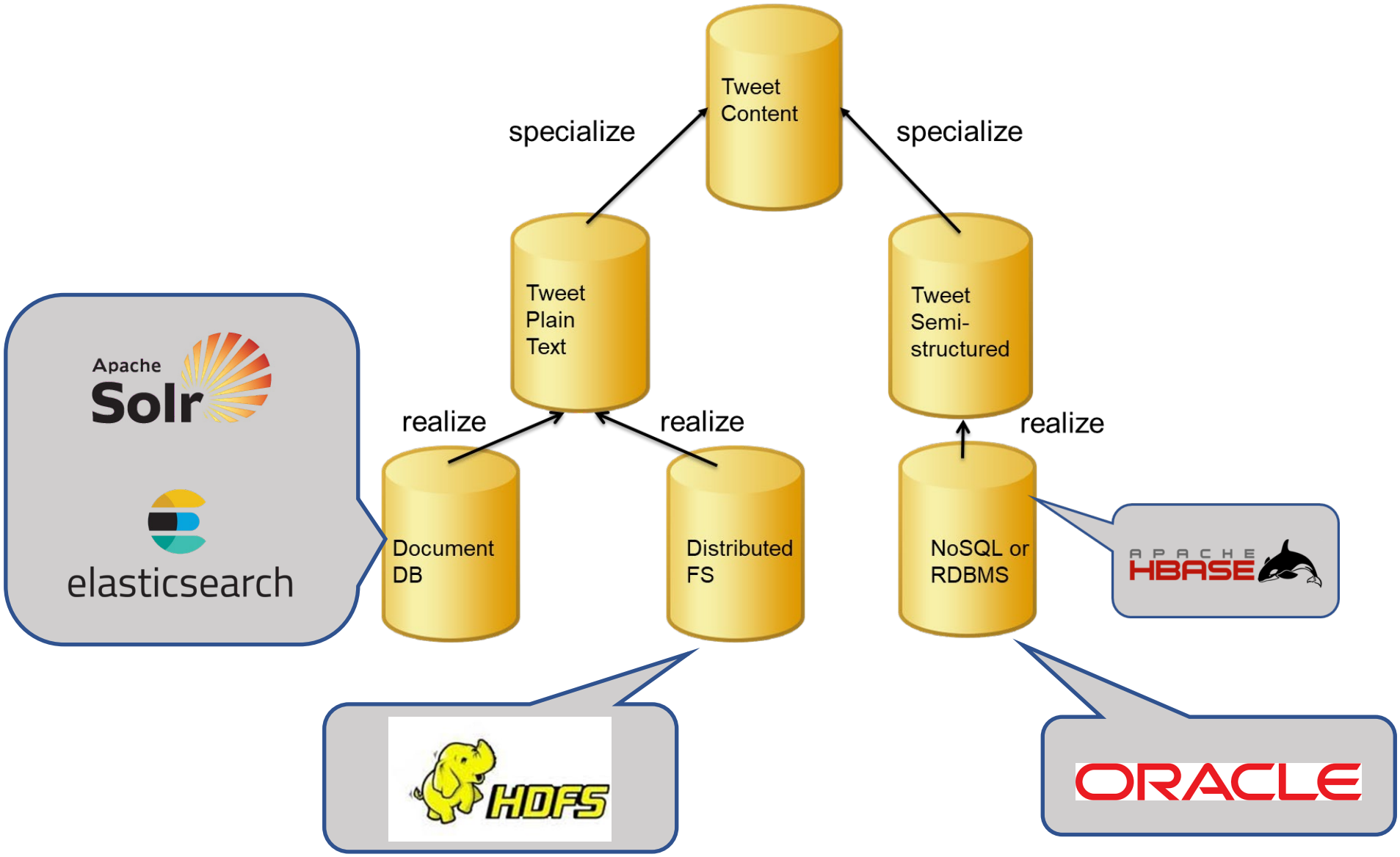
E.g.
caret (R)
scikit-learn (Python)



Internet as a Data Source: Twitter Data



Internet as a Data Source: Twitter Data



Conclusions

- Reference architectures are useful for:
  - Standardizing: vendor/technology independence, support to governance and capability planning
  - Sharing: reuse and cost reduction

- Next steps:
  - Extension to: Process-mediated data and Machine-generated data
  - Methodology architecture needed as well: work to do!