Horizon 2020

Understanding Europe's Fashion Data Universe

Survey Document of Existing Datasets and Data Integration Solutions

Deliverable number: D1.1

Version 4.0

Funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 732328

Project Acronym:	FashionBrain
Project Full Title:	Understanding Europe's Fashion Data Universe
Call:	H2020-ICT-2016-1
Topic:	ICT-14-2016-2017, Big Data PPP: Cross-sectorial and cross-lingual data integration and experimentation
Project URL:	https://fashionbrain-project.eu

Deliverable type	Report (R)
Dissemination level	Public (PU)
Contractual Delivery Date	30 June 2017
Resubmission Delivery Date	27 February 2019
Number of pages	30, the last one being no. 23
Authors	Alessandro Checco - USFD
Peer review	Jennifer Dick - USFD Jennie Zhang - MDBS

Change Log

Version	Date	Status	Partner	Remarks
1.0	03/07/2017	Final	USFD	Rejected 15/03/2018
2.0	20/04/2018	Resubmitted Final	USFD	Rejected 26/07/2018
3.0	21/05/2018	Resubmitted Final	USFD	Rejected 15/10/2018
4.0	27/02/2019	Resubmitted Final	USFD	

Deliverable Description

An overview of existing state-of-the-art solutions for data integration including infrastructures, algorithms, and datasets covering both academic research as well as industry solutions resulting from Task 1.1.

Abstract

This document provides an overview of existing state-of-the-art solutions for data integration, including infrastructures, algorithms, and datasets covering both academic research as well as industry solutions. In particular, the focus is on the datasets and techniques that are more suited to address the particular challenges of the FashionBrain project.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AIE	Active Intelligence Engine
API	Application Programming Interface
ASIN	Amazon Standard Identification Number
B2B	Business to Business
CDC	Change Data Capture
CPU	Central Processing Unit
CRM	Customer Relationship Management
CSV	Comma Separated Values
CT	Core Technology
CWM	Common Warehouse Metamodel
EDI	Electronic Data Interchange
EFR	Enterprise File Replication
EII	Enterprise Information Integration
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
FDWH	Fashion Data Warehouse
FTP	File Transfer Protocol
GUI	Graphical User Interface
HDFS	Hadoop Distributed File System
HTTP	HyperText Transfer Protocol
ID	Identifier
I/O	Input/Output
JDBC	Java Database Connectivity
JMS	Java Message Service
JSON	JavaScript Object Notation
LDAP	Lightweight Directory Access Protocol
LSTM	Long Short-Term Memory
MFT	Managed File Transfer
ODBC	Open Database Connectivity
OLAP	Online Analytical Processing
Q/A	Question and Answer

RDBMS	Relational Database Management System
RL	Record Linkage
SOA	Service-Oriented Architecture
SQL	Structured Query Language
SNMP	Simple Network Management Protocol
TCP	Transmission Control Protocol
UCI	University of California, Irvine
UI	User Interface
URL	Uniform Resource Locator
WP	Work Package
XBRL	eXtensible Business Reporting Language
XML	eXtensible Markup Language

1 Introduction

Data integration is the process of combining data residing in different sources and providing the data consumer with a unified view. The collection of heterogeneous data, particularly its integration and curation is a problem broadly studied in the literature [21, 14, 8, 12]. However, in the world of online fashion retail, high performance ad-hoc solutions are required in order to meet the near real-time large-scale requirements of the industry. Two of the main issues in this area are: (i) the integration amongst different data infrastructures and sources (e.g. from retailers, manufacturers, social media, logistics, website, customer care, etc.), and (ii) the complexity of the workflow needed to enable advanced queries over the integrated data. Moreover, the practical implementation of high performance, state-of-the-art data management solutions is a challenge in the world of Big Data due to the ever increasing amount of data accumulating from social media, retailers and other sources.

The main data integration tasks[1] focused on in the FashionBrain project are:

Aspect based Opinion Mining: given a set of reviews associated to a catalog, it is fundamental to analyse them to complement product catalogues and to improve search functionalities. This is related to Scenario 3 and Challenge 5 described in detail in Deliverable D1.2.

Entity Linkage: this technique solves the problem of linking a product catalog item to (i) text or (ii) images. This is related to Challenges 3 and 5.

Entity Recognition: the technique that allows parts of the text to be associated with an existing ontology. This is related to Scenario 3.

Ontology mapping: refers to the construction of a fashion ontology and the mapping of different ontologies into a unique one. This is related to Challenge 4.

Integration with social network data: where input from blogs and social networks needs to be integrated to the existing infrastructure.

Integration with crowdsourced data: where additional datasets are created using a crowd of paid workers or where the tasks are delegated to the crowd. This is related to Core Technology 3.

The remainder of this deliverable is structured as follows: Section 1.2 presents a brief history of data integration. Section 1.3 provides the requirements for datasets and techniques needed in the fashion industry. Section 2 provides an overview of the existing, publicly available datasets. Section 3 is a summary of the commercial

[1] Refer to Deliverable D1.2 for more details on the mentioned challenges and scenarios.

solutions for data integration. Finally, Section 4 provides a survey of the related academic work, with specific focus on the solutions resulting from scenarios and challenges dealt with in the FashionBrain project.

1.1 Scope of This Deliverable

This deliverable is related to the Core Technologies CT1 (Semantic Integration), CT2 (Infrastructures for scalable cross-domain data integration and management), and CT4 (In-Database Named Entity Recognition and Linking methods). Moreover, it addresses the requirements of the following challenges (refer to Deliverable D1.2 for more details):

Challenge 4: Linking Entities to Product Catalogue.

Challenge 5: Opinion Mining on Fashion Reviews.

Challenge 6: OLAP Queries over Text and Catalogue Data.

Related work with other deliverables:

Scope This deliverable is strictly related to Deliverable D1.2, and complements it by providing a survey of the existing datasets and solutions that could be used to address the challenges described there.

Dependencies We refer to Deliverable D1.2 for more details on the described challenges.

Contribution This deliverable will contribute to WP1 in its entirety, as it will help all partners to identify the best solutions and the shortcomings of the state-of-the-art.

1.2 A Brief History of Data Integration

Combining heterogeneous data sources (information silos) has been studied since the early 1980s, with a **data warehousing** approach where the objective is to create a single-view schema to make different sources compatible.

The data warehouse approach has the advantage of providing a fast and simple architecture, as the data are already physically stored in a single queryable repository, so it only takes little time to resolve queries [3, 27].

When datasets are frequently updated, the extract, transform, load (ETL) process needs to be continuously executed for synchronization, and thus this approach often becomes unfeasible.

An alternative to data warehousing (to address the above problem) is to provide a unified query-interface to access real time data over a mediated schema, retrieving information directly from the original databases. This approach is based on schema matching and in corresponding query transformation. By following this approach, a

key problem involving the resolution of the semantic conflicts between data sources is avoided [25].

Another common problem in data integration is data isolation. The main technique used to avoid this is to enrich the information with structural metadata (data entities).

Another predominant topic in the field of data integration is the level of data structuring in the solutions used. Currently, data hubs, data vaults and data lakes approaches have surpassed the more traditional structured Enterprise Data Warehouses [23, 16]. These techniques combine various kinds of data into a unique location, without the need of a complex relational schema to structure the data, allowing for a more agile development.

1.3 Data Integration Requirements

The scenarios in the fashion industry where data integration is more problematic, pointing out the requirements for data integration that arise from them, both in terms of infrastructure as well as datasets are described below. We will refer to the core technologies as well as the scenario and research challenges described in Deliverable D1.2, and we will focus on analysing one of the main workflows that FashionBrain project aims to consider in terms of data integration.

In Figure 1.1, the main data processing workflow related to FashionBrain is illustrated which will be used as motivation for the description of the requirements for data integration. Each step of the workflow is described in the following subsections, together with the corresponding requirements for that step. Further details on the scenarios and challenges mentioned can be found in Deliverable D1.2.

FashionBrain Taxonomy (Challenge 4)

To support data integration across sources and data types, and to have a central data schema, a **fashion taxonomy** needs to be built [26], which is based on redundancy-minimizing shopping categories (unlike the traditional fashion ontologies in which e.g. clothes are divided depending on the gender). This is necessary when integration of images and text are done against a structured catalog (as shown in Figure 1.1). The FashionBrain project is also exploring alternative solutions when integration between images and text is done without the need of a schema, as in the case of deep neural networks [4] where long short-term memory (LSTM) is an artificial neural network architecture that is recurrent, and allows data to flow both forwards and backwards within the network. This technique will be used for many of the of the tasks related to CT1 and CT4.

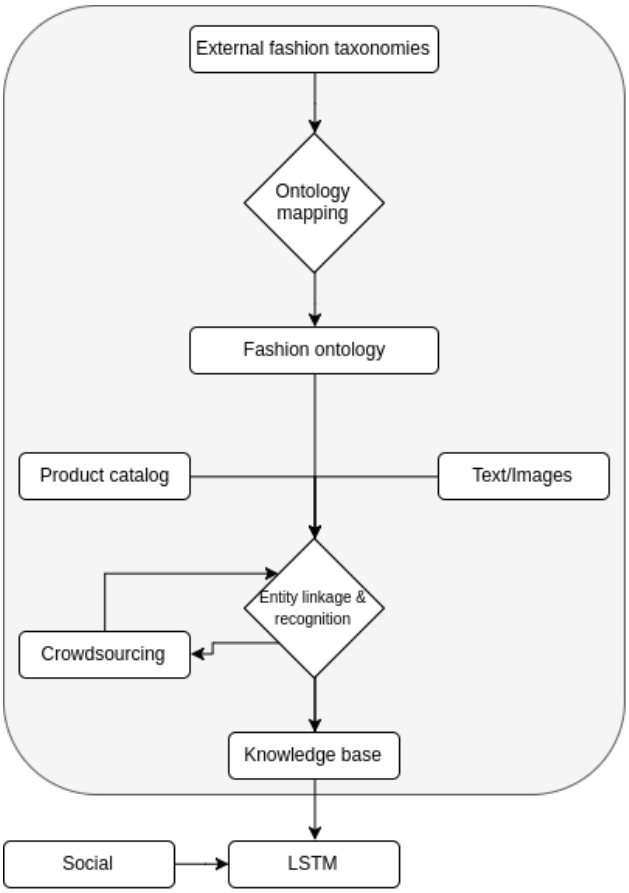

Figure 1.1: FashionBrain data processing workflow.

Product Taxonomy Linking (Challenge 4)

The obtained hierarchical structure needs to be fine-grained at a level that allows any item from any dataset to be linked to some layers of the hierarchy and subsequently, to link fashion entities to a certain layer in the hierarchy. As a result, it will be possible to define the relationships between the categories for each item and classify each of them as: e.g. "clothing"/"shoes"+ "material"+ "colour" + "pattern". The definition of relationships requires a prior mapping between instances of different classes. This is a fundamental step that will enable more fine-grained data integration and a richer search experience. A typical problem when creating ontologies from different datasets is **ontology mapping** [6] (the need to combine heterogeneous ontologies). This problem arises in the FashionBrain project when different taxonomies from the web (e.g. ebay, amazon etc.) need to be integrated in the fashion ontology. In Section 4 we will discuss existing solutions for ontology mapping.

Linking Entities to Product Catalogue (Challenge 4)

One of the main technological developments needed in the fashion industry is **entity linking** [26, 11] from text and from images. The goal is to be able to identify and disambiguate fashion products against a product catalog. Moreover, effective information extraction from unstructured content (e.g. twitter, blogs, as well as multimedia sources such as Instagram) is critical for trends prediction. A typical application is the entity extraction process from text and image, and mapping them as instances in an already developed ontology. In Section 3, we will present solutions for data integration regarding this problem. In Section 4 we will show the current state-of-the art in academia regarding entity linking.

Opinion Mining and Knowledge Base Creation (Challenges 2 and 5)

Given a set of reviews associated to a catalog, it is fundamental to analyse them so as to complement product catalogues and improve search functionalities. To achieve that, we will integrate our results from previous tasks to obtain a global view of the knowledge base, which will allow a clear mapping of the opinions in a review, in an appropriate schema. Being able to realise such analysis directly in the database is fundamental to improve performances and to simplify the workflow [18]. This is related to Challenge 2 as End-To-End Learning is only possible with an integrated view of the data.

Crowdsourcing (Core Technology 3)

Task 6: To solve the problem regarding lack of training data, **crowdsourcing** is often used. Current uses of crowdsourcing for fashion data include the entity linking from images to product catalogs and processing of product reviews (e.g. extraction and classification of sizing issue mentions). Crowdsourcing will also be used to train the entity recognition algorithms used. Section 2 will describe a series of datasets that have been obtained in this way.

Infrastructures for scalable cross-domain data integration and management (Core Technology 2)

Finally, retailers manage data on fashion products and transactions in a fashion data warehouse (FDWH), which is often a relational database management system. Recombining relational data from a FDWH with text data from **social networks** is therefore an important operation for learning about their users, monitoring trends and predicting new brands. A typical problem is the join of relational data (entities and relationships) to text data and vice versa. Section 3 details the existing commercial solution able to perform such operations.

2 Existing Datasets

This section collects the existing publicly available datasets that are related and can potentially be used for the FashionBrain project. At the end of this section, how these datasets will be used in the FashionBrain project, and which datasets are missing or require additional data collection will be discussed. Table 2.1 summarizes the requirements of datasets for the fashion industry.

Dataset type	**Publicly available dataset**	**Assessment**
Shop Inventory	Fashionista, DressesAttributeSales	Needs integration with bigger inventory
Social Media	None	One of the FashionBrain objectives is to collect and integrate social media data
Ontologies	None	One of the FashionBrain objective is to build a fashion ontology
User reviews	Amazon Product	other reviews are needed (e.g., large dataset on clothes for sizing issues)
Product catalog data	DressesAttributeSales	more recent and granular sales data
Product Images	Fashion-focused Creative Commons Social, Apparel Classification with Style, Fashion 10.000, DeepFashion	More data are needed.

Table 2.1: Current available datasets in the fashion industry.

In the following section, each dataset is described in detail.

2.1 Amazon Product Dataset

This dataset contains product reviews and metadata from Amazon and includes 142.8 million reviews spanning May 1996 - July 2014. This dataset includes

reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Source

http://jmcauley.ucsd.edu/data/amazon/

References

R. He, J. McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering". WWW, 2016 http://cseweb.ucsd.edu/~jmcauley/pdfs/www16a.pdf

J. McAuley, C. Targett, J. Shi, A. van den Hengel. "Image-based recommendations on styles and substitutes". SIGIR, 2015 http://cseweb.ucsd.edu/~jmcauley/pdfs/sigir15.pdf

2.2 Amazon Questions and Answers Dataset

This dataset contains Question and Answer (Q/A) data from Amazon, totaling around 1.4 million answered questions. This dataset can be combined with Amazon product review data (Section 2.1) by matching ASINs in the Q/A dataset with ASINs in the review data. The review data also includes product metadata (product titles etc.).

Source

http://jmcauley.ucsd.edu/data/amazon/qa/

References

Mengting Wan, Julian McAuley. "Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems". ICDM 2016. http://cseweb.ucsd.edu/~jmcauley/pdfs/icdm16c.pdf

Julian McAuley, Alex Yang. "Addressing complex and subjective product-related queries with customer reviews". WWW 2016. http://cseweb.ucsd.edu/~jmcauley/pdfs/www16b.pdf.

2.3 DeepFashion Dataset

DeepFashion database is a large-scale clothes database and contains over 800,000 diverse fashion images ranging from well-posed shop images to unconstrained consumer photos. It is annotated with rich information of clothing items. Each image in this dataset is labeled with 50 categories, 1,000 descriptive attributes, bounding box and clothing landmarks. DeepFashion contains over 300,000 cross-pose/cross-domain image pairs.

Source

http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html

References

Liu, Ziwei, et al. "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

2.4 Fashion 10.000 Dataset

This dataset is composed of a set of Creative Common images collected from Flickr. It contains 32 398 images distributed in 262 fashion and clothing categories. The dataset comes with a set of annotations that are generated using Amazon Mechanical Turk (AMT). The annotations target 6 different aspects of the images which are obtained by asking 6 questions from AMT workers.

Source

http://www.st.ewi.tudelft.nl/~bozzon/fashion10000dataset/

References

http://dl.acm.org/citation.cfm?doid=2557642.2563675

2.5 DressesAttributeSales Dataset

This dataset contain attributes of dresses and their recommendations according to their sales. Sales are monitored on alternate days.

Characteristics from UCI Machine Learning repository:

Data Set Characteristics:	Text	Number of Instances:	501	Area:	Computer
Attribute Characteristics:	N/A	Number of Attributes:	13	Date Donated	2014-02-19
Associated Tasks:	Classification, Clustering	Missing Values?	Yes	Number of Web Hits:	49645

Source

https://archive.ics.uci.edu/ml/machine-learning-databases/00289/

2.6 Fashionista Dataset

The fashionista package, from Kota Yamaguchi, contains the Fashionista dataset without annotation, which was collected from chictopia.com in 2011.

Dataset Information

The data is stored in a tab-delimited text files in the following format. Text files are split into chunks. Concatenating them will recover the full data records in one table.

posts/xxx.txt

Post(ID, URL) Post represents a blog post in chictopia.com. In Chictopia, bloggers upload up to 5 photos to a single post. This table keeps the URL of these posts and a unique identifier in the dataset.

Note that a blogger might have deleted some of the posts since the dataset was collected. There is no guarantee that all posts are available.

photos/xxx.txt

Photo(post_ID, URL) Photo represents a picture associated to each post. In the table, one row keeps a URL and the ID of the associated post.

garments/xxx.txt

Garment(post_ID, name) Garment is meta-data extracted from the post. In each post, bloggers list their clothing items from the pre-defined set of clothing types. This garment table keeps pairs of the ID of the post and extracted garment name.

Source

https://github.com/grahamar/fashion_dataset

References

Parsing Clothing in Fashion Photographs http://www.cs.unc.edu/~hadi/cvpr_2012.pdf

2.7 Apparel Classification with Style Dataset

In this dataset, clothing classes are defined. The dataset consists of over 80 000 images and bounding, obtained by reorganizing some of ImageNet's synsets.

Dataset Information

Attributes: Image, bounding box. **Classes:** Polo shirt, Vest, Blouses, T-shirt, Shirt, Short dress, Sweater, Uniform, Undergarment, Suit, Robe, Cloak, Jacket, Coat, Long dress/

Source

http://data.vision.ee.ethz.ch/cvl/fashion-data.tar.bz2

References

http://people.ee.ethz.ch/~lbossard/projects/accv12/accv12_apparel-classification-with-style.pdf Bossard, Lukas, et al. "Apparel classification with style." Asian conference on computer vision. Springer Berlin Heidelberg, 2012.

2.8 Fashion-focused Creative Commons Social Dataset

This dataset is a fashion-focused Creative Commons dataset, designed to contain a mix of general images as well as a large component of images that are relevant to particular clothing items or fashion accessories. The dataset contains 4810 images and related metadata, with ground truth on images' tags obtained with Mechanical Turk (AMT). Two different groups of annotators (i.e., trusted annotators known to the authors and crowdsourcing workers on AMT) participated in the ground truth creation.

Source

http://skuld.cs.umass.edu/traces/mmsys/2013/fashion/

References

http://dl.acm.org/citation.cfm?id=2483984

3 Data Integration - Industry Solutions

An overview of the leading commercial solution for data integration is provided below. In the FashionBrain project we will mostly compare and integrate our work against open source software. Therefore, in section 3.1, we will first review existing free software implementation for data integration. After that, a small summary of existing paid software is given.

3.1 Free Software Solutions

Oracle

Oracle Data Integrator is a platform that provides out-of-the-box integration with ERPs, CRMs, B2B systems, flat files, XML data, LDAP, JDBC, and ODBC. Oracle Data integrator generates native code for many RDBMS engines and thus, usually does not need a conventional ETL transformation server. It is also fully integrated with Oracle Fusion Middleware, Oracle Database, and Exadata.

Pentaho

Pentaho Data Integration offers solutions to extract and integrate data from heterogeneous sources. It integrates very well with open source solutions like MonetDB. It provides high performance ETL solutions and supports multi-threaded engines, data partitioning and clustered execution. It can execute jobs on Pentaho Data Integration servers or Hadoop. It supports in-memory caching, dynamic lookups and parallel bulk-loaders.

Talend

Talend provides more than 450 connectors to integrate data sources. It uses a set of open source tools to provide real time and batch solutions for NoSQL, data warehousing, data synchronization as well as migration and data sharing. It is able to connect natively to many Cloud applications and Web services. Talend integration services is also able to connect data marts, Online Analytical Processing systems and software as a service system.

Apache Kafka

Apache Kafka is a distributed streaming platform running on multiple servers, that allows storage and processing streams of records in a fault tolerant way. It is ideal to build real-time streaming data processes to transfer data between systems or to process them in real-time in an online application. Kafka has four core APIs:

- The Producer API to publish streams of records.
- The Consumer API allows an application to process the stream of records.
- The Streams API allows an application to ingest one or more input streams and also allows it to produce a transformed output stream.
- The Connector API connects Kafka streams to existing applications or systems.

In Kafka the communication is handled by a high-performance, language agnostic TCP protocol.

Apache NiFi

Apache NiFi implements data routing, transformation and system mediation logic. Its main characteristics are:

- Web-based user interface
- Loss tolerant vs guaranteed delivery
- Low latency vs high throughput
- Dynamic prioritization
- Data Provenance
- Track dataflow from beginning to end
- Multi-tenant authorization and internal authorization/policy management

Gobblin

Gobblin is a data ingestion framework for extracting and transforming large volumes of data from a myriad of data sources, like databases, APIs, FTP servers, etc. onto Hadoop. It handles the ETLs tasks, including the additional tasks of scheduling, task partitioning, error handling and data quality checking.

Skool

Skool tries to provide answers to the limitations of the aforementioned tools: Gobblin is more focused on data flow scheduling than on ingestion and extraction; Apache Nifi does not cover end-to-end flow very well; and Oracle Data Integrator has limited support for big data. Skool has the following features:

- Seamless data transfer to/from a relational database or flat files and HDFS.

- Automatic Hive tables generation.
- Automatic generation of file-creation scripts and jobs from Hadoop tables
- Automatic regression testing.

3.2 Paid Software Solutions

Actian

http://www.actian.com/

Actian provides solutions for the design of integration processes for data warehouse loading, the converting of data between formats, the deployment of application integration scenarios, and the integration of applications in the cloud and on-premise. The main technologies provided are Lifecycle Management interfaces, Service Oriented Architecture Platform, Cloud-to-Cloud Computing and Reusable Metadata.

Alooma

http://www.alooma.com/

Alooma main focus is to provide a Data Pipeline as a Service, with attention to security. The objective is to provide solutions for integration, cleansing and integration of data, with the capability of connecting the following data warehouses: Amazon Redshift, Google BigQuery, Snowflake, Salesforce, MySQL, Oracle, Microsoft Azure, Looker and Tableau.

Adeptia

http://www.adeptia.com/

Adeptia provides solution for Business to Business integration, Application Integration and Business Process Management, as well as Data Integration. The main technical solution provided are "Any-to-Any Conversion, Graphical Data Mapper, Human Workflow, SOA, Metadata-Driven, Web-Based UI, Code-Free, Business User Access, EDI & Trading Partner Management, Preconfigured Connections, Web Service API Publishing, Web Portals and Customer Onboarding".

Altova

https://www.altova.com/

Altova focuses on XML solution development tools to assist developers with data integration, data management and software development. It provides drag and drop GUIs to map XML, databases, flat file, JSON, EDI, Excel, XBRL, and/or Web

services data between each other. A high performance server allows the execution of the dataflow described with the GUI.

Attivio

http://www.attivio.com/

Attivio's Active Intelligence Engine (AIE) can process both structured and semi-structured data, Big Data and unstructured content from a wide variety of databases, document repositories, content management systems, email systems, websites, social media and file servers, providing out-of-the-box connectors to many systems such as relational databases, file content, XML and CSV data. Some of the connectors available are designed for Microsoft SharePoint, Microsoft Exchange, Active Directory, EMC Documentum Content Server, website harvesting, and Hadoop.

Attunity

http://www.attunity.com/

Attunity is a provider of information availability software that focuses on data replication, change data capture (CDC), data connectivity, enterprise file replication (EFR) and managed-file-transfer (MFT). It provides data replication controlled by a simple GUI and focuses on "Zero Footprint" architecture, meaning that no agents must be placed on the source or target, eliminating overhead for mission-critical systems.

Denodo

http://www.denodo.com/

Denodo is a company that mainly focuses on Data Virtualization. It offers high performance Data Integration and abstraction for Big Data and real-time data services. The Denodo Platform provides a virtual "logical data lake" for accessing the data, stored in potentially many heterogeneous systems.

It provides an API that makes the data lake appear as a single unified version of the data.

Dell Boomi

http://www.boomi.com/

Dell Boomi is an integration solution focused on cloud solutions for data quality services, data management and data integration.

Centralized user and role management and single sign-on are provided to simplify the data management and data integration process.

HVR

http://www.hvr-software.com/

HVR provides real-time data replication for Business Intelligence, Big Data, and hybrid cloud, allowing data connection between many sources including SQL databases, Hadoop, data warehousing, as well as the most commonly used file systems. HVR provides solutions for data migrations, Data Lake consolidation, geographic replication, database replication, and cloud integrations.

HVR offers faster log-based capture from SQL Server, improved loads into Teradata and Amazon Redshift, and full support for log-based Change Data Capture on open source PostgreSQL.

IBM InfoSphere

http://www-01.ibm.com/software/data/integration/

IBM InfoSphere Information Server provides a rich set of information integration and governance capabilities to integrate big data with a traditional enterprise platform. It focuses on the ability of understanding the data through visualization, monitoring and cleansing tools. IBM InfoSphere can integrate to databases such as its own InfoSphere Warehouse Enterprise (based on IBM DB2), IBM's Big Data Analytics solution Netezza, SQL, Oracle and other databases or the Enterprise Service Bus and message brokers such as MQSeries.

Informatica

http://www.informatica.com/

Informatica is a data integration provider for data governance, data migration, data quality, data synchronization and data warehousing.

Informatica's mainframe data integration solutions are multi-platform and connects to a wide variety of on-premise and cloud-based applications—including enterprise applications, databases, flat files, file feeds, and social networking sites.

Information Builders

http://www.informationbuilders.com/

Information Builders provides solutions for real-time data integration, data delivery and data federation. It provides tool for extract, transform, and load, enterprise information integration (EII) initiatives and big data integration. Information Builders enables users to issue distributed queries that correlate and manipulate data from many different relational databases, packaged applications, structured data files such as XML and EDI documents, as well as legacy databases and files.

Jitterbit

http://www.jitterbit.com/

Jitterbit provides data integration solutions based on graphical "No-Coding" approaches, to simplify the configuration and management of complex integration projects. Available in the cloud or on-premise, Jitterbit automatically discovers system configurations and allows non-technical users to point and click to define source & target systems, drag and drop to map data transformations, and run integration operations on batch, trickle or real-time schedules.

Liaison

http://www.liaison.com/

The Liaison dPaaS Platform consists of three modules:

Data Orchestration: provides functionalities to integrate applications and data in the cloud and as an enterprise system;

Data Persistence: allows data management with APIs that support the schema-on-read approach;

Data Visualization: provides customizable interfaces for data visualization and data flow analysis.

Microsoft SSIS

https://docs.microsoft.com/en-us/sql/integration-services/sql-server-integration-services

SQL Server Integration Services (SSIS) is Microsoft's Extract, Transform, Load (ETL) tool and is integrated with SQL Server. SSIS provides a set of built-in tasks, containers, transformations, and data adapters that support the development of business applications. The main focus of SSIS is designing complex workflows without writing any code. It provides GUIs to manage SQL Server databases. It is composed of 2 engines: the runtime engine and the dataflow engine.

Mulesoft

http://www.mulesoft.com/

Mulesoft is a Business to Business application delivery network focused on providing APIs for data and application connection. It supports a variety of on premise and cloud based applications and systems. Besides APIs, it provides a browser and a command line tool for configuration and control.

Sap

http://www.sap.com/

SAP BusinessObjects Integration software allows to integrate many data formats to standard protocols such as CWM, XML, HTTP/HTTPS, JMS, and SNMP. It provides a system to generate reports, visualizations or dashboards and focuses on preserving data security. The technology used supports parallel processing, grid computing and real-time data movement for many hardware architectures.

SAS

http://www.sas.com/

SAS Data Management provides point-and-click interfaces to build pipeline models for data integration. The data flow process can integrate data marts, data streams and data warehouses. SAS natively supports Oracle, DB2, Teradata, Netezza, SQL Server, Aster and Hadoop. It is able to perform multithreaded, load-balanced operations on multiple hardware architectures.

Stone Bond

http://www.stonebond.com/

Stone Bond provides a platform for configuration, deployment and monitoring of data integration solutions. It provides an intuitive configuration interface for data transformation and for data visualization.

Striim

http://www.striim.com/ Striim focuses on real time data with services for alerts, visualization and triggers. It supports an SQL-like language and claims to scale linearly while also being able to handle hundreds of millions of messages per second.

Syncsort

http://www.syncsort.com/

Syncsort's main focus is compression and join algorithms, with particular attention to speed. It is, in practice, a data integration acceleration suite focused on reducing CPU time and disk I/O on commodity hardware.

4 Data Integration - Academic Solutions

The problem of data integration has been broadly studied and the main technologies developed are referred to in the literature [14, 20]. In this section, the data integration problems that are more relevant to the fashion industry are focused on.

4.1 Record Linkage - Entity Resolution

Record Linkage (RL) is the process of locating records in a dataset referring to the same entity across different data sources. RL is necessary when joining datasets based on entities that may or may not share a common identifier [7, 15].

In data warehousing, RL is fundamental; each source system may have a different way of representing same entities. Usually, RL occurs in the second phase of the ETL process.

Entity resolution [7] is a process executed by a middleware, where non-obvious relationships across different data silos can be exposed, allowing the connection of such sources. Entity resolution engines apply rules based on deterministic or probabilistic rules to identify hidden relationships across the data.

The simplest kind of record linkage, called deterministic or rules-based record linkage, generates links based on the number of individual identifiers matching among the available datasets [24]. When all, or some identifiers (above a threshold), are identical, two records are considered matching. This approach is ideal when entities share common identifiers and when quality of data is relatively high. More complex deterministic rules have been devised to infer more complicated connections [7, 12, 15]. When data quality decreases or data complexity increases, the number of deterministic rules needed grows rapidly, making the usage of specialized software tools (see Section 3) a fundamental necessity. Moreover, when new data, with characteristics that are different than expected, enter the system, it may be necessary to completely restructure the deterministic rules.

Probabilistic record linkage [3], otherwise known as fuzzy matching or probabilistic merging, has a different approach. It considers a large range of identifiers, estimates the ability to identify a match or non-match for each identifier, builds a weighted set and uses these weights to estimate the probability that two records refers to the same entity. This approach usually requires a training phase that uses a set of gold standard examples manually entered into the system (in modern systems, this phase can be carried out via crowdsourcing).

Correctly configuring the parameters of a fuzzy system is not simple and it is paramount as it heavily affects the balance between precision and recall. A key technique to obtain such configuration is blocking [13, 17]. Currently, state-of the art of integrated entity linking while editing can be used in crowdsourced solutions, e.g. TASTY [4]. When performance is important, a solution that implements theses tasks directly in the database is fundamental, e.g. using INDREX[18].

4.2 Ontology Mapping

Ontology mapping (ontology alignment or ontology matching) is the process of determining correspondences between elements in ontologies. A set of correspondences in a mapping is also called an alignment. This alignment can be syntactic, external, or semantic [10]. Ontology alignment tools have been developed to process database schemas [10, 2], XML schemas [5], and other frameworks. A typical approach is to first convert the ontology to a graph representation before the match [19]. Such graphs can be represented with the triples <subject, predicate, object>. Automatic systems have been proposed and they have been proven to obtain satisfactory results [22, 9, 1]. Currently the most promising solution, in terms of performance, appears to be COMA++ [1].

4.3 Crowdsourcing

In order to build large datasets and train machine learning algorithms, one of the most promising tools is LSUN [28]: a partially automated labeling scheme, leveraging deep learning with humans in the loop. Starting from a large set of candidate images for each category, LSUN iteratively samples a subset, asks people to label them, classifies the others with a trained model, splits the set into positives, negatives, and unlabeled based on the classification confidence, and then iterates with the unlabeled set.

4.4 Integration with social network data

A promising solution for complex text/image spaces, like those obtained from social network, is one where integration between images and text is done without the need of a schema, as is the case with Fashion DNA [4]. In this example, coordinate vectors locating fashion items in an abstract space are built through a deep neural network architecture that ingests curated article information such as tags and images, and is trained to predict sales for a large set of frequent customers. In the process, a dual space of customer style preferences naturally arises. Interpretation of the metric of these spaces is straightforward: the product of Fashion DNA and customer style

vectors yields a forecast purchase likelihood for the customer-item pair, while the angle between Fashion DNA vectors is a measure of item similarity.

4.5 Aspect based Opinion Mining

To solve such a complex task, a mix of the solutions described in Sections 4.1 and 4.3 are needed, where the workflow includes text extraction, entity linkage and the use of crowdsourcing to train an appropriate classifier.

4.6 Summary of Academic Solutions

Table 4.1 summarises the free software solutions considered as state-of-the-art for the problems presented in Section 1. For the remainder of the FashionBrain project, the members of the consortium will refer to this table when developing new solutions for data integration problems.

Task	**Solution**
Aspect based Opinion Mining	TASTY, LSUN, INDREX
Entity Linkage	TASTY
Entity Recognition	INDREX
Ontology Mapping	COMA++
Integration with Social Network Data	Fashion DNA
Integration with Crowdsourced Data	LSUN

Table 4.1: State-of-the-art data integration solutions.

Bibliography

[1] David Aumueller, Hong-Hai Do, Sabine Massmann, and Erhard Rahm. Schema and ontology matching with coma++. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 906–908. Acm, 2005.

[2] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm. *Schema Matching and Mapping*. Springer Science & Business Media, February 2011.

[3] T Blakely. Probabilistic record linkage and a method to calculate the positive predictive value. *Int. J. Epidemiol.*, 31(6):1246–1252, 2002.

[4] Christian Bracher, Sebastian Heinz, and Roland Vollgraf. Fashion dna: Merging content and sales data for recommendation and article mapping. *arXiv preprint arXiv:1609.02489*, 2016.

[5] Akmal Chaudhri, Mario Jeckle, Erhard Rahm, and Rainer Unland. *Web, Web-Services, and Database Systems: NODe 2002 Web and Database-Related Workshops, Erfurt, Germany, October 7-10, 2002, Revised Papers*. Springer, July 2003.

[6] Namyoun Choi, Il-Yeol Song, and Hyoil Han. A survey on ontology mapping. *ACM SIGMOD Record*, 35(3):34–41, 2006.

[7] Peter Christen. *Towards Parameter-free Blocking for Scalable Record Linkage*. 2007.

[8] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science & Business Media, July 2012.

[9] Carlo Curino, Giorgio Orsi, and Letizia Tanca. X-som: A flexible ontology mapper. In *Database and Expert Systems Applications, 2007. DEXA'07. 18th International Workshop on*, pages 424–428. IEEE, 2007.

[10] Marc Ehrig. *Ontology Alignment: Bridging the Semantic Gap*. Springer Science & Business Media, December 2006.

[11] Pablo Gamallo and Marcos Garcia. Entity linking with distributional semantics. In *Lecture Notes in Computer Science*, pages 177–188. 2016.

[12] Lise Getoor and Ashwin Machanavajjhala. Entity resolution for big data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, 2013.

[13] Phan H Giang. A machine learning approach to create blocking criteria for record linkage. *Health Care Manag. Sci.*, 18(1):93–105, March 2015.

[14] Behzad Golshan, Alon Halevy, George Mihaila, and Wang-Chiew Tan. Data integration. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '17*, 2017.

[15] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data Quality and Record Linkage Techniques*. Springer Science & Business Media, May 2007.

[16] Bill Inmon. *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications, April 2016.

[17] Robert Patrick Kelley and United States. Bureau of the Census. Statistical Research Division. *Blocking Considerations for Record Linkage Under Conditions of Uncertainty*. 1984.

[18] Torsten Kilias, Alexander Löser, and Periklis Andritsos. Indrex: In-database relation extraction. *Information Systems*, 53:124–144, 2015.

[19] Simon Kocbek and Jin-Dong Kim. Exploring biomedical ontology mappings with graph theory methods. *PeerJ*, 5:e2990, March 2017.

[20] Maurizio Lenzerini. Data integration: a theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02*, page 233, New York, New York, USA, 2002. ACM Press.

[21] Matteo Magnani and Danilo Montesi. A survey on uncertainty management in data integration. *J. Data and Information Quality*, 2(1):1–33, July 2010.

[22] Miklos Nagy, Maria Vargas-Vera, and Enrico Motta. Dssim-managing uncertainty on the semantic web. 2007.

[23] Pradeep Pasupuleti and Beulah Salome Purra. *Data Lake Development with Big Data*. Packt Publishing Ltd, November 2015.

[24] L L Roos and A Wajda. Record linkage strategies. part i: Estimating information and evaluating approaches. *Methods Inf. Med.*, 30(2):117–123, April 1991.

[25] K Saranya, M S Hema, and S Chandramathi. Data fusion in ontology based data integration. In *International Conference on Information Communication and Embedded Systems (ICICES2014)*, 2014.

[26] Xiao Yue Wang, Ru Jiang Bai, and Xiao Fan Yu. Comparison of the fashion ontology integration models. *Key Eng. Mater.*, 480-481:397–401, 2011.

[27] Jennifer Widom. Research problems in data warehousing. In *Proceedings of the fourth international conference on Information and knowledge management - CIKM '95*, 1995.

[28] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.