Data Wrangling at Scale

The experience of EW-Shopp

Nikolay Nikolov  
SINTEF  
Oslo, Norway  
nikolay.nikolov@sintef.no

Michele Ciavotta  
University of Milan-Bicocca  
Milan, Italy  
michele.ciavotta@unimib.it

Flavio De Paoli  
University of Milan-Bicocca  
Milan, Italy  
flavio.depaoli@unimib.it
EW-Shopp: Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey

• Creation of a platform to integrate
  – Customer decision journey data
  – Weather and Events data
  – Multi-lingual

• to support
  – Analytics
  – Business Intelligence
Customer Decision Journey

Recognition of Need

Information

Buy

Customer Care

Ads

Social interactions

Online searches

Retails (online)

Retails (offline)

Call centers

Example

• Search a product online
• Get prompted with ads
• Go to a comparison shopping platform
• Go to a retail website
• Visit a store
• Buy
• Call customer care
Objectives

Why Weather and Events in Customer Decision Journey?

• Weather and events impact on customer decision journey

• Help companies that operate in the **eCommerce, Retail and Marketing industries**
  
  - embed **events** and **weather** in their daily **data analyses**
  
  - so as to **gain deep customers’ insights** and **optimize the services** offered to their customers.
Technical challenges

- Different domains
  - Corporate data vs. event data vs. weather data
- Generalizing and scaling up analyses, business questions and services
  - Tool-supported data integration
  - Bridging across different identifiers used in corporate data and weather/event data
- Volume and variety of data
  - E.g., 300K campaigns in 74 different countries bidding over 2B keywords vs dozens of weather variables

Figure 11: Example of column generation in DataGraft by data addition

Once the datasets are completed, QMiner will develop the analytics required in each case, from the wide variety of algorithms included in libraries like Spark MLlib.

To conclude, a particular example, during the last HandBall World Championship that took place in Germany, we monitored the correlation between the German Team matches and the online digital campaigns indicators under the "SportFitness" category.

Figure 12: Correlation between German national handball team matches and online digital indicators in the "SportFitness" category.
EW-Shopp at glance

ANALYTICS
QMiner

VISUALIZATION
Knowage

DATA INTEGRATION LAYER
hosting + data transformations & quality + semantic table annotation + data summaries
DataGraft++ ASIA

CONSUMER JOURNEY ANALYTICS

LOCATION INTELLIGENCE

DIGITAL MARKETING CAMPAIGN OPTIMIZATION

EVENT-BASED SALES ANALYSIS
Data sources - Core Data
- Events
- Weather
- Products
- Geospatial

Corporate services
- Data Hosting
- Processing
- Data Enrichment
- Orchestration
- Management

Platform services
- Data Wrangler
- Data Analyzer
- Data Reporter

Crosscutting Services
- Security
- Monitoring
Considering most of all data science pipelines:

- 80% time is spent to clean and enrich data
- Only the 20% is actual data analysis

Big Data and Business Analytics (BDA) revenues will reach $210 billion in 2020*

*Worldwide Semiannual Big Data and Analytics Spending Guide from International Data Corporation (IDC), 2017
Data Wrangler
• Validation
• Cleaning
• Filtering
• Guided Annotation
• Enrichment (instance and schema level)

A typical Data Science Project
Data Enrichment

Reconciliation
Link values to shared identifiers (used to bridge with additional data to be fetched)

Extension
Use the reconciled values for fetching additional data, based on specific properties

For each row, retrieve the selected parameters in that region, on that date

<table>
<thead>
<tr>
<th>GRUPO</th>
<th>KEYWORD</th>
<th>REGION</th>
<th>Geonames ID</th>
<th>Clicks</th>
<th>Category</th>
<th>Date</th>
<th>WF-temp ++0</th>
<th>WF-temp ++1</th>
<th>WF-cloud ++0</th>
<th>WF-cloud ++1</th>
<th>WF-precip ++0</th>
<th>WF-precip ++1</th>
</tr>
</thead>
<tbody>
<tr>
<td>36874389</td>
<td>194906</td>
<td>Thuringia</td>
<td>2822542</td>
<td>64</td>
<td>BusinessManagement</td>
<td>11/03/2018</td>
<td>13</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>37850338</td>
<td>517827</td>
<td>Bavaria</td>
<td>2951839</td>
<td>50</td>
<td>NewsMediaPublications</td>
<td>12/03/2018</td>
<td>15</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>36874385</td>
<td>459143</td>
<td>Berlin</td>
<td>2950157</td>
<td>42</td>
<td>TravelTourism</td>
<td>12/03/2018</td>
<td>18</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>0,122</td>
<td>0,153</td>
</tr>
<tr>
<td>36874388</td>
<td>891139</td>
<td>Bavaria</td>
<td>2951839</td>
<td>36</td>
<td>Vehicles</td>
<td>11/03/2018</td>
<td>14</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>0,136</td>
<td>0,122</td>
</tr>
<tr>
<td>36874385</td>
<td>459143</td>
<td>Bavaria</td>
<td>2951839</td>
<td>30</td>
<td>HomeGarden</td>
<td>10/03/2018</td>
<td>14</td>
<td>14</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0,136</td>
</tr>
</tbody>
</table>
Reference Dataflow

Data set management
- Data set dimension reduction
- Platform vs. Corporate component

Workflow
- Design vs. Production
- Interactive vs. Batch
Reference Dataflow

Data set management
- Data set dimension reduction
- Platform vs. Corporate component

Workflow
- Design vs. Production
- Interactive vs. Batch
DatGraft: Data Transformation and RDF Publication Process

- Interactive design of transformations
- Repeatable transformations
- Reuse/share transformations (user-based access)
- Cloud-based deployment of transformations
- Self-serviced process
- Data and Transformation as-a-Service
ASIA: Assisted Semantic Interpretation and Annotation of tabular data

Support the semantic annotation of tabular data

- **Schema-level annotations**
  - reconcile columns against shared vocabularies

- **Instance-level annotations**
  - reconcile values against shared systems of identifiers

**Applications**

- RDF-ization: generate knowledge graphs represented in RDF
- Content Extension: fetch additional data from the KG
DW architecture and workflow
DW architecture and workflow
• **System Orchestration**
  – Defining/configuring data workflow on a high level

• **Processing**
  – Cluster of host machines – Docker, Rancher agents
  – Shared file system
  – Enrichment database accessible over the network
  – Rancher stack configuration – derived from the **System Orchestration** component
Workflow generic elements: file-based data sharing

- **input** folder: **input data** for the entire workflow; input for the first step
- Each step has a **work** folder for **intermediate results** (if necessary)
- Each step has an **out** folder that holds the **results** of the process
- Each step (except 1st) uses the previous step's **out** folder as its input folder
- The last step's **out** folder provides the output of the workflow (if necessary)
  - could be a step to upload data to the enrichment db
Conclusions

• First release of a platform for end-to-end semantic data enrichment
• Preliminary sets of experiments are promising
• A first set of interlinking services is available (weather, GeoNames, GeoTargets, products...)
• Tons of work to do
  – Closing the wrangling loop allowing the user to improve the reconciliation
  – Improving scalability of reconciliation matchers.
  – Implementing enrichment with semantic annotated events
Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey

Discuss your event and weather-based analytics with the LinkedIn group

Weather & Event-based Analytics in Business

Get in touch with us!

Matteo Palmonari
palmonari@disco.unimib.it
www.ew-shopp.eu
@EwShopp

EW-Shopp H2020 Project

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement n. 732590
Examples in Digital Marketing (JOT INTERNET MEDIA)

- Corporate data
  - Performance of keywords in digital ads, e.g., # impressions

- Pattern examples
  - After/during heavy rainy days in Madrid (Feb 11th 2017)

- Business questions and service design:
  - Best moment to launch a campaign for category “DiningNightLife” / keyword “hamburguesa en casa”?