ETL Process Plan

# Phone Data Harvest; Scraping and Analysing Mobile Device Information

## Group 2 (EN):

### Daniel Herrera - [daniel.herrerarussert@stud.hslu.ch](mailto:daniel.herrerarussert@stud.hslu.ch)

### Ramon Burkhard – [alainramon.burkhard@stud.hslu.ch](mailto:alainramon.burkhard@stud.hslu.ch)

### Jack Brown – [jack.brown@stud.hslu.ch](mailto:jack.brown@stud.hslu.ch)

## March 14, 2024

**Project ideation**

In the initial stages of defining our project topic, we explored various ideas. We considered extracting data from retail companies such as "Amazon," "Mediamarkt," and "Galaxus," delved into social networks like "X" and "Instagram" to analyze trends, and explored data related to sports activity from sources like "Wikipedia," "Olympics.com," and "laliga.com."

After some consideration, we decided that delving into data associated with retail companies would be more interesting for the entire group. At this point, the focus shifted to deciding where and what to scrape. Initially, we contemplated utilizing the "Mediamarkt" group as our project source, motivated by its widespread presence across different European markets. This approach would enable us to analyze diverse country markets, including Spain, Germany, and Switzerland.

An alternative idea within the same theme emerged: exclusively scraping data from the Swiss market using platforms like "Galaxus.ch," "Mediamarkt.ch," and "Interdiscount.ch." Ultimately, we opted for the latter, specifically sourcing data from Switzerland domains. The main reason behind this decision lay in the similar structures of data presentation across domains of the "Mediamarkt" group. This would influence the group to adopt a uniform approach, consequently, the data manipulation process would be nearly identical.

Opting for domains from various companies proved more intriguing, as each student would encounter distinct structures, potentially leading to varied challenges. This diversity is expected to enhance the overall learning experience for the group.

Regarding the "What?" aspect, the group unanimously decided to focus on scraping data related to notebooks or smartphones. This decision was straightforward, driven by the observation that smartphones, across all pages, offered more diversity in models, quantity, ratings, and reviews.

**1. Introduction and motivations**

*A brief overview of our project and its purpose. We must mention the importance of understanding and comparing electronic product prices across different markets. Discuss the motivation behind the project, such as identifying pricing trends, market competitiveness, and potential cost savings for consumers.*

*Because we are planning on comparing sources from different countries, we must explain why we have chosen the ones we have chosen.*

**Contextualization of the Project**

To provide context for our project and facilitate its development, give meaning to our inquiries, and justify our decisions, we present a story below:

The "TipTopClub" is an application and platform that charges a monthly or yearly fee from its users. It offers registered users the convenience of searching for products in various categories such as sports equipment, books, construction tools, office materials, and electronics. The platform then presents search results, already highlighting the best options based on specified parameters (price, specifications, delivery time, ratings, and reviews). Users can select a product from the results, complete their request, and the "TipTopClub" takes charge by placing the order on behalf of the user. This eliminates the need for users to register on multiple websites and manage an address book. Another advantage of the club is the option to participate in collective orders with other members, potentially reducing costs. The club then handles the separation and dispatches the products to the user.

In this context, the "TipTopClub" has tasked its data science team with delving into the next product category listed in their database: smartphones. To fulfil this request, the board has asked for a preliminary study from the data science team, focusing on extracting data from three different websites to evaluate the project's viability.

**Context diagram**

3 Sources -> Extraction -> Transformation -> Storage

Source 1

galaxus.ch

*Jack Brown*

Source 2

mediamarkt.ch

*Ramon Burkhard*

Source 3

interdiscount.ch

*Daniel Herrera*

**Extraction**

Fields: title, price, rating

Scraping with Python in PyCharm

Packages: Selenium

**Extraction**

Fields: title, price, rating

Scraping with Python in PyCharm

Packages: Selenium

**Extraction**

Fields: title, price, rating

Scraping with Python in PyCharm

Packages: Selenium

**Transformation**

Data cleaning  
Data enrichment

Python in PyCharm  
Packages: Pandas

**Merge of the data**

Based on model, storage, source, etc.

**Storage on MariaDB**

Merged data sets

Dataset 1

Dataset 2

Dataset 3

**Sources**

We plan on scraping data from three Switzerland based websites (dynamic web pages): galaxus.ch, mediamarkt.ch, and interdiscount.ch. These platforms offer an immense variety of phones, with information like make, model specifications, delivery time, ratings, and type of sale (new or refurbished). With this data, we aim to enable users to find the best product given their prioritized specifications (e.g. best price and largest phone) while incorporating aspects like review ratings and delivery time. The data can be merged by phone maker and corresponding attributes.

**galaxus.ch:**

For our analysis, we use <https://www.galaxus.ch/en/s1/producttype/smartphones-24?take=>. This provides necessary filters to set us onto the phones page. Galaxus is one of the biggest online electronic sellers in Switzerland which means we will have a wide variety of phone data to extract. To be exact, as of March 14, 2024, there are currently 2,087 products that can be scraped. After inspecting Galaxus’s robot.txt file, crawlers like sogou spider, Yandex, eventmachine httpclient, and niki-bot are not allowed to scrape any url with “/” in it (does not apply to us). We also wouldn’t be allowed to access things like /cert, /Files, or /management (which also does not apply to us), so nothing will stop us from scraping this website.

**mediamarkt.ch:**

**interdiscount.ch:**

* Electronic product lists (iPhone 14, 15, Samsung, etc.), model description (colour, storage, size, etc), delivery time, rating, new or preowned, and image of phone.
* Three different market reseller websites.

**3.1. Source 1 – galaxus.ch (maybe incorporate scraping price trend)**

**3.2. Source 2 – mediamarkt.ch**

**3.3. Source 3 – interdiscount.ch**

**Expected end results/ Attributes**

In this section we present our preliminar expected structure of the csv uploaded from each group member in mariadb. The structure will be used as a parameter for all group members. the idea behind it is that the group members after scraping and the data manipulation can achieve a dataframe structured in this way.

“Id” - Unique value that identify the row;

“Brand” - Brand attributed to the product;

Model - Model of the product with version;

“Category” - Category that the product belong;

“Condition” - If the product is new or used;

“Size” - Size of the screen of the product in inches;

“Space capacity” - Storage capacity of the product in GB;

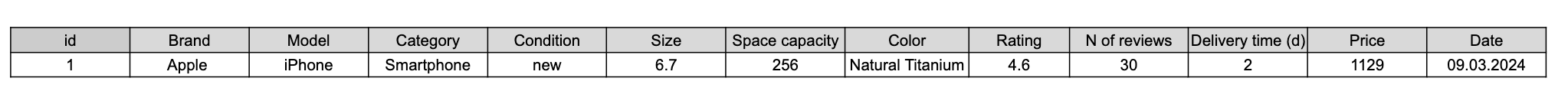
“Color” - Main color of the product;

“Rating” - Average rating from 0 - 5 given by buyers;

“N of reviews” - Number of reviews given by buyers;

“Price” - Price of the product;

“Date” - Date of the scrapping.



**2. Research questions**

* How do electronic product prices vary across different platforms?
* What impact do promotions have on product prices?
* Can we identify patterns in consumer ratings and preferences based on reseller of choice?

**5. Risks and Challenges**

We have outlined the proposed ETL (Extract, Transform, Load) workflow and undertaken a preliminary assessment of the websites chosen. It is foreseeable that challenges may arise during the scraping procedure. Since the main objective of our project is to collect data pertaining to the topics at hand, a solution in the event of unexpected issues would be to switch to a different source of data. A consequent additional challenge would be the identification of an adequate replacement website, which considering the broad topic of the project and the abundance of mobile phone resellers, should be straightforward.

Some additional challenges have been identified in the dynamic nature of these online resellers, with fluctuations of price data and ratings over time. Also, the consideration of foreign websites adds another layer of dynamic behavior, with variations in currency exchange data, and differences in pricing between countries. This is however a predictable circumstance and partly the focus of our research questions.

As far as the characteristics of the data, naming conventions of variables and strings could differ within one dataset, as well as between sets from different portals and countries. With some thorough and systematic data cleaning, all three sources could be set to an equivalent standard of quality. This will also provide a significant aid for the next stage of the project, when all three sources are merged into one larger data set. An advisable approach would be to establish standardizations for variable naming and model-specific contents and rating codes, which will be elaborately described in a chapter of the final report. Only with a systematic review and standardization of the data can all three sources be acceptably combined and compared on equal terms.

**6. Reference list**

**7. List of tables and figures**