Data Governance @ SneakerPark

Prepared by: Hong Tran

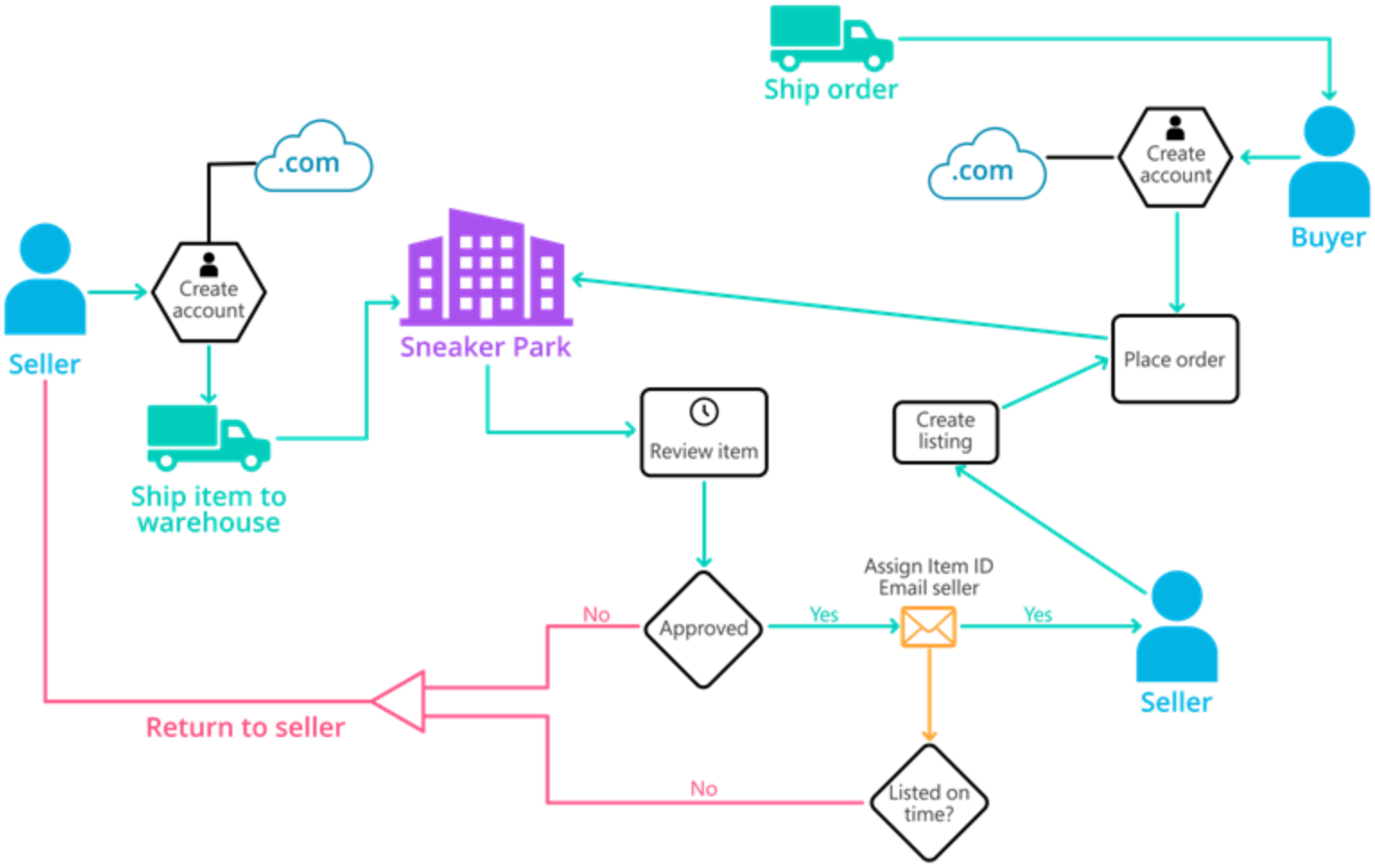
Submitted on:

Background

- **SneakerPark** is an online shoe reseller that allows people to buy and sell used and new shoes. Buyers can bid for shoes or buy them outright, and sellers can set a price or sell to the highest bidder.
- Each buyer and seller must have an active account in order to sell, bid, or purchase sneakers using SneakerPark's website.
- SneakerPark authenticates the shoes before shipping them to the buyer, so before listing an item, the seller must ship it to SneakerPark's warehouse. Upon receipt, SneakerPark assigns an item number to each pair of sneakers and notifies the seller that they are now free to list their item. If the item is not listed within 45 days, SneakerPark returns it to the seller and sends an invoice to the seller for the shipping cost.
- If the item is found to be inauthentic or in an unacceptable condition, it is also returned back to the seller in a similar fashion.
- When the item sells, the buyer's account is credited with the purchase price minus the SneakerPark service fee and shipping fees to deliver the item to the buyer.
- Currently, SneakerPark only supports sales within the United States.

Background (cont'd)

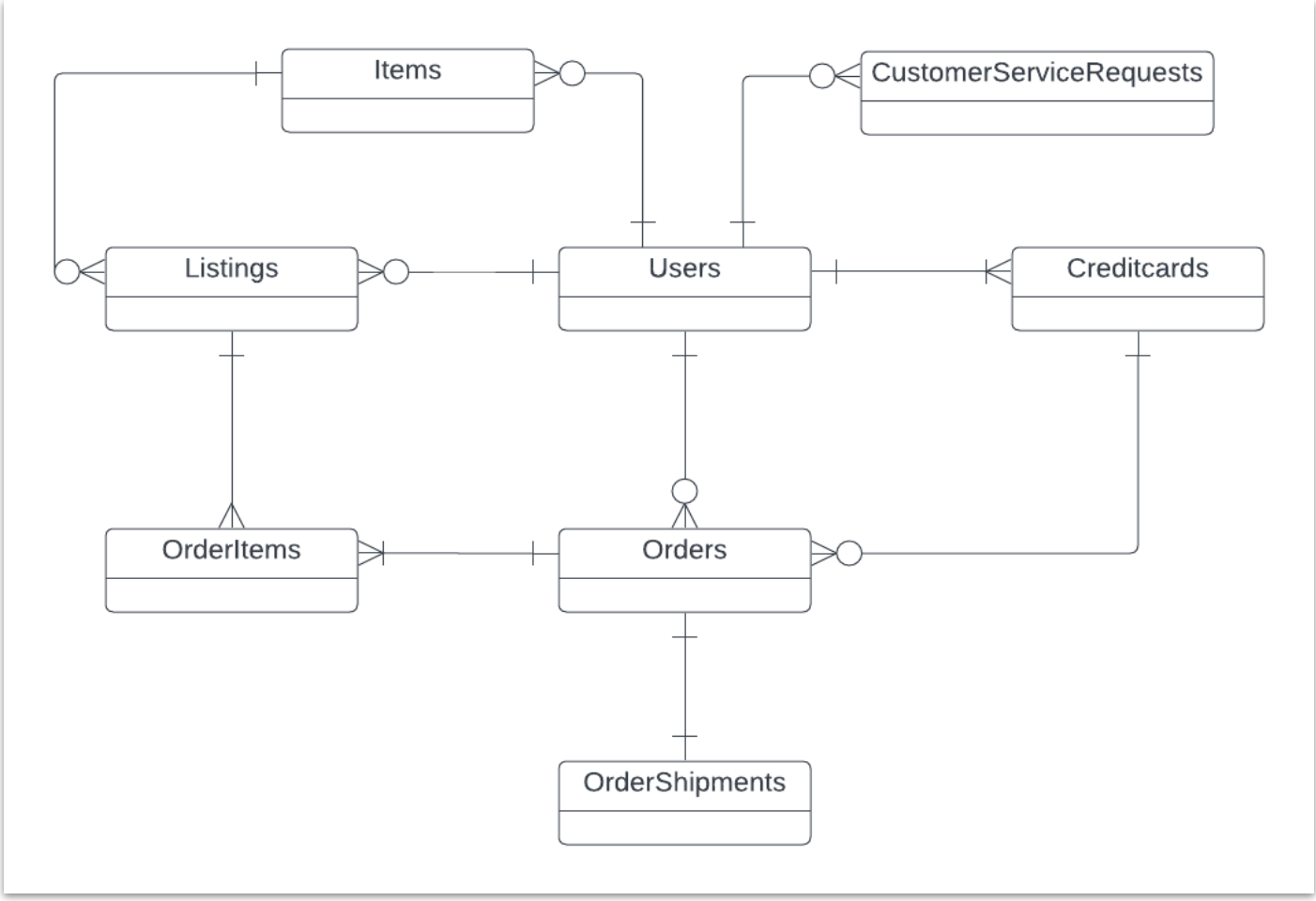
- Below you can see a diagram that will hopefully help you visualize some of SneakerPark's business processes. Keep in mind that it does not capture ALL processes and every nuance, but simply serves as another artifact to use in your project.



Step 1

Enterprise Data Catalog

Part 1: Enterprise Data Model


Items
CustomerServiceRequests
Listings
Users
Creditcards
OrderItems
Orders
OrderShipments

Step 2

Enterprise Data Catalog

Part 2: Metadata

Create the first version of the Metadata Catalog by documenting the metadata from all systems in the "Data Dictionary" and the “Enterprise Data Catalog” tabs of the provided Sheets template.

Step 3

Data Quality

Part 1: Profiling and Cleansing

Profile the data to identify **data quality issues** saw in the data. Also provide **at least 1 data quality issue that haven't been seen** in the data, but can foresee occurring in the future. Then come up with the data quality rule for each data quality issue, including for the one that is foresee.

Step 4

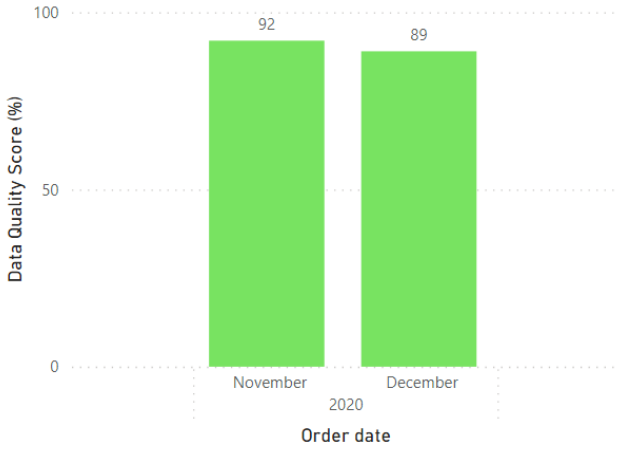
Data Quality

Part 2: Monitoring

Data Quality Threshold

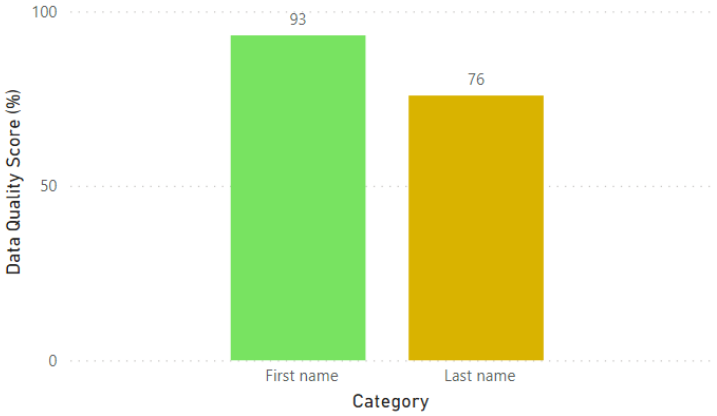
< 70% 70-80% >= 80%

Percentage of records with TotalAmount greater than zero



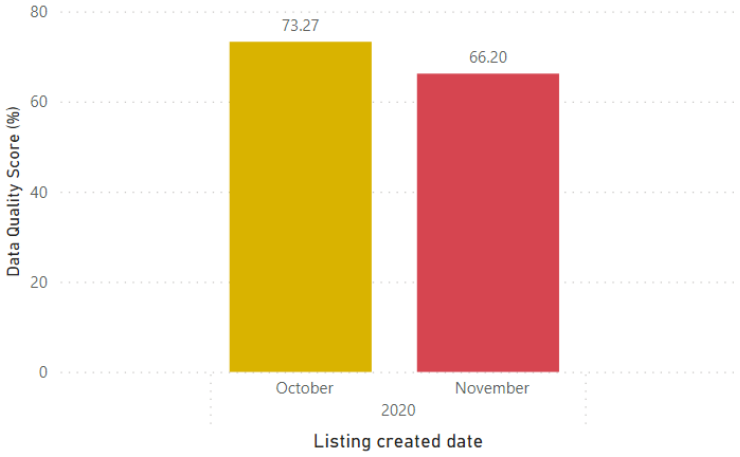
Year	Month	Qualified records	Total records	Percentage of qualified records
2020	November	46	50	92
2020	December	43	48	89

Percentage of records having FIRST NAME/LASTNAME in User Service different from Customer Service Application for the same person



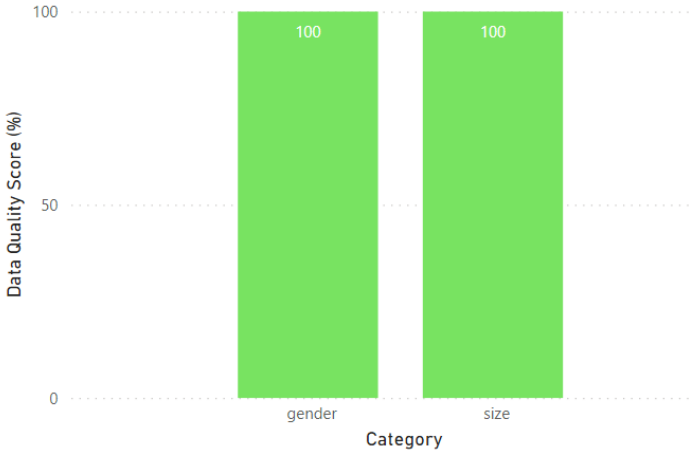
Category	Qualified records	Total records	Percentage of qualified records
First name	27	29	93.10
Last name	22	29	75.86

Percentage of records without variant in brand in Listing Service system



Year	Month	Qualified records	Total records	Percentage of qualified records
2020	October	159	217	73.27
2020	November	141	213	66.20

Percentage of listing records linked to gender or size

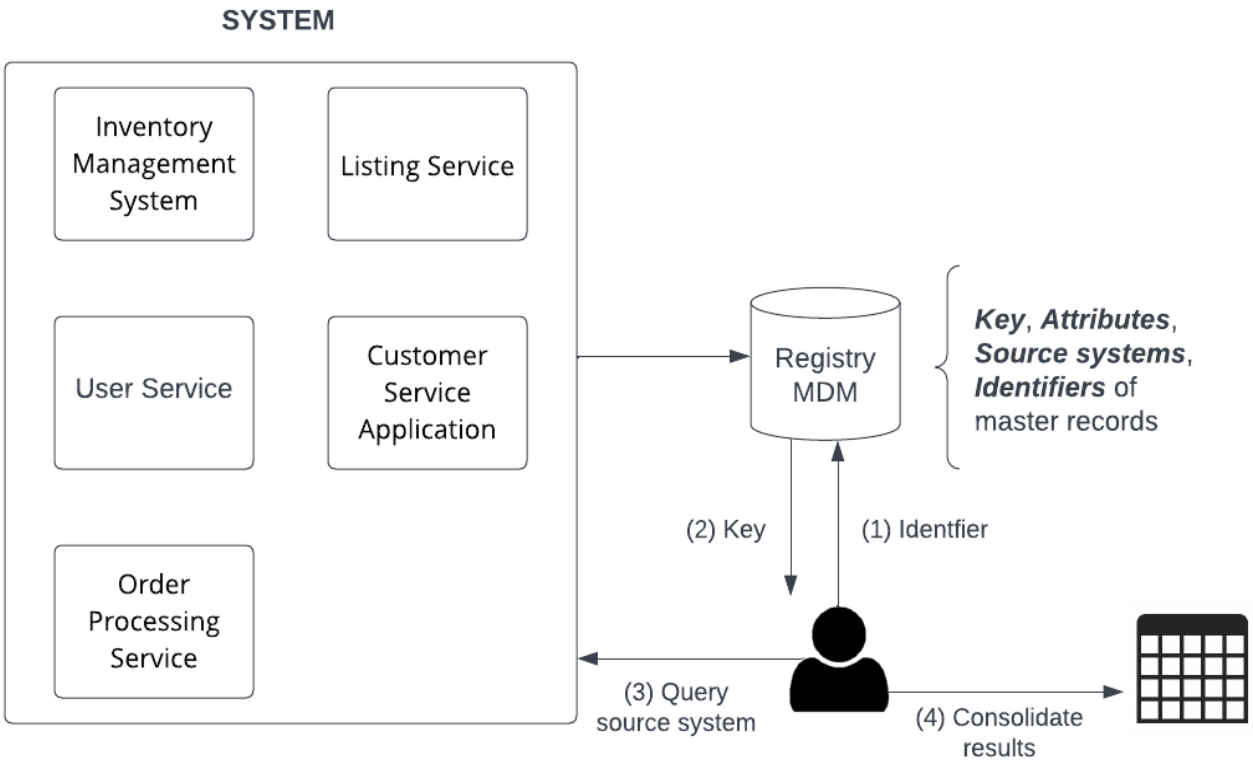


Category	Qualified records	Total records	Percentage of qualified records
gender	430	430	100
size	430	430	100

Step 5

Master Data Management

Part 1: MDM Architecture



Explanation

After examine SneakerPark's systems and business model, ***Registry MDM*** is chosen for MDM architecture of SneakerPark. The implementation of this style will have minimal intrusion to the operation of existing system. This is because Registry MDM only store list of master data ***key***, ***attribute***, and ***source system name/locations*** in a central data for users to query consolidate that master data.
This benefit aligns with requirement from SneakerPark as they want the MDM architecture minimally disruptive to existing system (especially the Order Processing Service). This architecture is also suitable for SneakerPark as it never had a MDM architecture before.

Step 6

Master Data Management

Part 2: Master Record

This step will define a set of **matching rules** that will be used by the SneakerPark's MDM Hub to match item and customer entities between the company's different systems as follow:

- Customer:

(1) Same ***Email***, ***CreditCardNumber*** and ***CreditCardExpirationDate***. Match the UserID records on Email of Users, CreditCardNumber and CreditCardExpirationDate of CreditCards.

(2) Same ***Email*** and ***OrderID***. Match the UserID records on Email of Users and OrderID of CustomerServiceRequests.

- Item:

(1) Same ***ItemName*** and ***SellerID*** . Match the ItemID records on the ItemName of Items and the SellerID of Listings.

(2) Same ***BrandName***, ***ArrivalDate*** and ***SellerID***. Match the ItemID records on the Brand Name and Arrival Date of Items and Seller ID of Listings.

Step 7

Data Governance:

Roles and Responsibilities

This section discusses what **data governance roles and responsibilities** will be necessary to oversee this new Data Management initiative and whether SneakerPark should make new hires.

Aspect	Responsibility	Role	Person
Data Quality	- Perform data profiling - Measure data quality - Monitor data quality dashboard	Data Steward	Jessica
Data Quality	Remediation option for bad data quality	Data architect	Me
Metadata	Create and update Enterprise Data Model and Enterprise Data Catalog	Data Steward	Jessica
Master Data	Create Registry MDM and perform ongoing maintenance	Technical developer	Jake
Master Data	- Remediation strategy to improve data quality for Enterprise Data Warehouse - Define match rules for golden records	Data architect	Me
Master Data	Create and automate golden records to populate MDM	Technical Developer	Jake
Master Data	- Review match rule for golden records - Approve/Reject golden records	Data Steward	Jessica

From above discussion, we need extra skill of data steward and technical developer for SneakerPark.

But for now it's not necessary to make new hires because those responsibilities can be managed by me as a new hired data architect, Jake and Jessica.

- ***Jake*** is good at database administration, fixing data issues and he's supporting all data systems at SneakerPark. Although the skill of technical developer is not really his profession, it's not difficult for him to reskill as he's a highly technical person.
- ***Jessica*** has been instrumental in diagnosing data issues and finding solutions. She knows SneakerPark's data extremely well so others can look to her for an answer to understand the data. She will take the role of data steward
- For ***me***, as a data architect, I will focus on creating rules and suggest remediation option.

Standout Suggestions

1. Create a Business Glossary for SneakerPark and define common terms such as Item, Buyer, etc. Think and discuss how SneakerPark can improve on the consistency of the terms that its systems currently use. (You can use the “Business Glossary” tab of the same Sheets template you’ve been using for the other parts of this project to get you started.)

1. Document SneakerPark’s current naming conventions. Can you think of any improvements? (You can use the “Standard Naming Conventions” tab of the same Sheets template you’ve been using for the other parts of this project to get you started.) Some examples of Naming Conventions include;

    - Do not use spaces or special characters.
    - Use only LOWERCASE.
    - All identifier fields should end in “_id”.
    - Avoid acronyms and abbreviations.

1. Write SQL scripts for the matching rules that you’ve created in Step 6.