Entity Resolution at Scale: Mapping 100+ million customer's account to customer's identity

Ashwini Maurya
Albertsons Companies

Abstract

Customer identity is at core of many businesses. This becomes very complex in presence of over 100 million customers, especially when most of Personally Identifiable Information (PII) are missing or have invalid values. At Albertsons, we use PII info along with engagement data, item level transactions, payments, and third party data to map customer accounts at both individual shopper and household levels. We develop various methodologies and pipelines for data processing , feature engineering, predictive modeling, id elevation and stamping. These algorithms are scalable and quite robust in the sense that the generated shopper id (customer level mapping) i) uniquely maps to individual customer, ii) is time invariant, iii) is invariant to customer's additional accounts, and iv) invariant to updated PII at a future point in time. Model is packaged as python package and deployed in production. The analysis of about 100+ million card numbers shows reduction of approximately 9% from customers accounts to shopper id, and reduction of approximately 16% from customer accounts to household ids.

1 Introduction and Scope

In north America, an individual usually engages with 4 devices on average (a significant increase from 2 devices in 2018). Among digitally engaged users, over 50% have 3 or more active emails. These statistics are much higher at household level. When digitally engaging or even when registering at store, these customers often don't disclose the correct personally identifiable information (PII). At Albertsons we found that very low coverage on PIIs such as first and last names, phones, emails, addresses. The statistics on missing and invalid pii varies over time, more recent registrations have better coverage than previous years. Also the missing values across PIIs is highly skewed for store only customers than ecom customers.

Moreover, the ecom users leave a rich trail of information such as cookies, browsing history, device info, ip address (location), clicks, likes of content or product they are engaging with. In addition, use of CRM software, subscription info, payments systems used in point of sale transactions (such as credit card info), first and third party delivery info (such as doordash, uber eats etc), social media data, customer surveys, and non PII data (such as transactions) adds to the complexity that makes it quite challenging for businesses when it comes to mapping these data points to customers data silos.

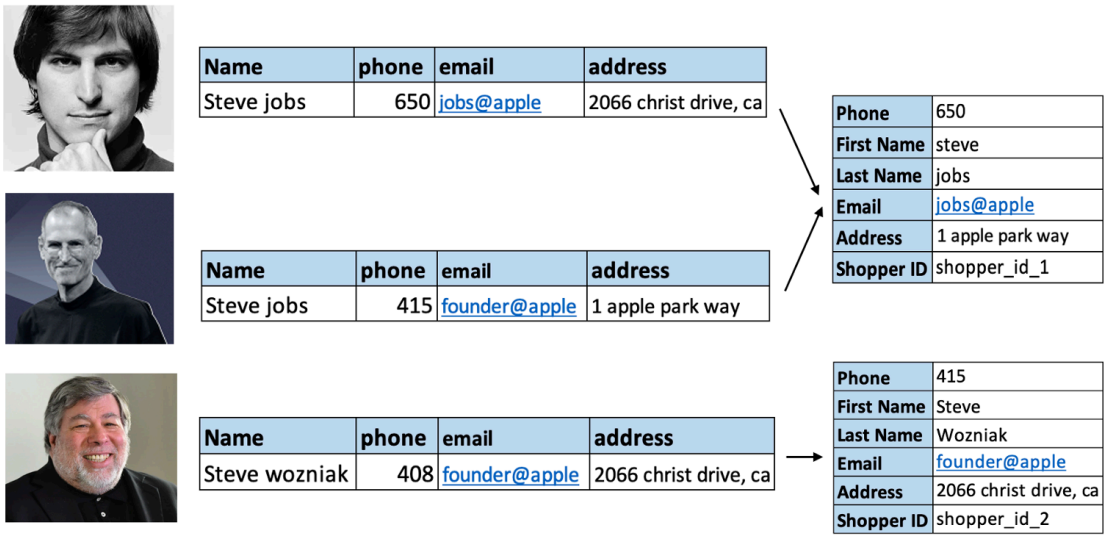


Figure 1: An example of entity resolution

Entity Resolution refers to process of stitching various customer info into single entity. The entity here can be an individual or a set of individual sharing certain PIIs

or core attributes of the underlying set. For example in personalization such as music recommendations the entity could be an individual. An example of household level entity is where a group of individuals want to be put together in one entity such as to avail aggregated reward or loyalty points at household level. In a house, husband and wife are two different shoppers but they are part of same household. Consequently both husband and wife will have separate shopper ids but will have common household (unless they want to be put in different household).

Figure 1 gives an illustration of complexity in the entity resolution process. In this example, we know that first two records are same individual and third record is a different individual. Although this is an example of clean and complete data on all PII fields, it is often not the case in real data sets. In this example, all records share same first name. First and third record share same address whereas second and third record share same email address. The goal of entity resolution is to map first two records to same entity, and third record to a different entity.

At Albertsons there are numerous IDs used across many applications. Some of these are described below.

UUID:	Universally Unique Identifier (UUIDs) are 128 character unique identifier. UUID follows specific structrue as defined in RFC 4122.
GUID:	Globally Uniquely Identifier are also 128 bit unique reference numbers generated by algorithms. There is no real difference between UUID and GUID except that GUIDs were originally used by Microsoft for the Windows O/S.
Club Card:	Albertsons club card number is assigned to customers at time of account sign up.

Household ID: A household level ID is assigned to a household that maps one or more individuals to a household id. There are instances when individuals in same household want to be assigned to have different hoousehold id assignements within Albertsons ecosystem. Such an altercation in householding is often done via Albertsons call center calls, which is time consuming, expensive and difficult to scale.

While all of these IDs may have some scope depending upon the use case, none of these IDs uniquely map to a customer. An ID mapping uniquely to a customer is of fundamental importance when it comes to personalization, marketing & promotion, providing better customer service, early detection of churning customers among many. At Albertsons we refer these individual customer IDs as shopper IDs. As shopper IDs have many use cases, these IDs must have some nice properties that makes them easy to use across many systems, applications and over time. Below we list such desired properties of a good shopper id.

1.1 Properties of a good Entity ID

We may refer these as oracle properties of good entity id. These properties are applicable to broader context of Entity IDs (outside the scope of this paper and Albertsons companies ecosystem).

1. Each Entity ID should uniquely map to an individual customer (or set of individual when entity is group of individuals).
2. Entity ID should be time invariant, i.e. it does not change over time.
3. Entity ID should be invariant to customer getting new club cards. In this case new card should be assigned the same entity id as the original one.
4. Entity ID should be invariant to customer's updated PII at later point of time.

At Albertsons we developed the entity IDs that satisfy all of the properties outlined above.

2 Entity Resolution as building block of customer 360

Entity Resolution is key element in building customer 360. Customer 360 is single and comprehensive view of customer's info and their journey (that can happen through various channels), their engagement through various products and services. Customer 360 can also be thought of providing the following views of customers at one place:

Who are they?

The customer PII such as name, phone, email, physical address etc may be collected during sign up process. Having knowledge of name, one can deduce gender of customer to have better understanding of their products needs. Using their purchase history business can also learn if they are ecom or store only customers thus enriching info about customer's identity.

What are they doing?

Knowledge of customers engagement such as their browsing and purchasing histories, use of services such as delivery enables business to understand customer and their needs at deeper level. This can be used to enhance recommendations, send better promotional offers such as sending coupons of items they might need in a near future. Business can also use these insights and the third party (engagement with other social media apps such as twitter, facebook etc) to target potential customers.

How are they doing?

Knowledge of how a customer engages with products and services is very useful information. For example knowing a customer uses third party app (e.g. instacart) to order groceries, business can uses this info to improve in house delivery services. Knowledge of customers engagement with different devices can be very helpful too. For instance if in a given household, different shoppers use different devices to order groceries, it may be good idea to generate device specific recommendations using device customer's engagement than aggregated household level.

Having 360 view of customers can enable business quickly learn about customers personal preferences, having a better understanding of their products and services, offer new products (such as Albertsons Fresh Pass) and lower churn rates. For more details refer to section on use cases of customer 360.

3 Customer data flow across various systems

The diversified nature of customers engagement across various channels/devices generates rich engagement and transactions data. Aggregating data from various sources is very time consuming, susceptible to errors, and requires significant resources. Because of this many business despite having made huge investments to collect, process and store the rich customer data, often fail to gather collective insights and use it towards their various needs. Figure 2 gives a high level view of customer data flow across various systems at any typical retail business.

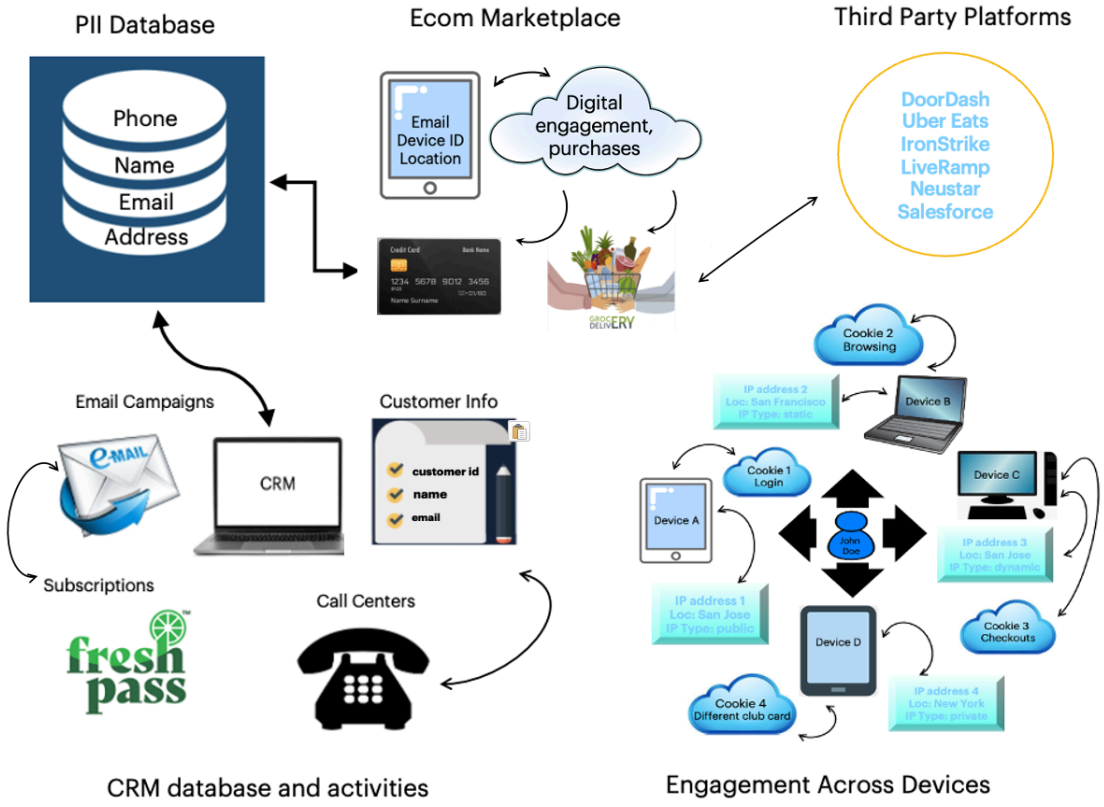


Figure 2: Customer Data Flow

To further illustrate the importance of unified customer view such as customer 360, consider the following example. Lets say a customer is using anonymous IP address to engage with company website such as searching products but uses different apps to logins and item purchases. Not having access to combined data from both devices provides little clue about customers non-purchased products that might be interesting to them. On the other hand having a complete view of customer engagement and other relevant history, the businesses can send coupons for products that the customer viewed but did not purchase, thus increasing sales. Another interesting aspect of data

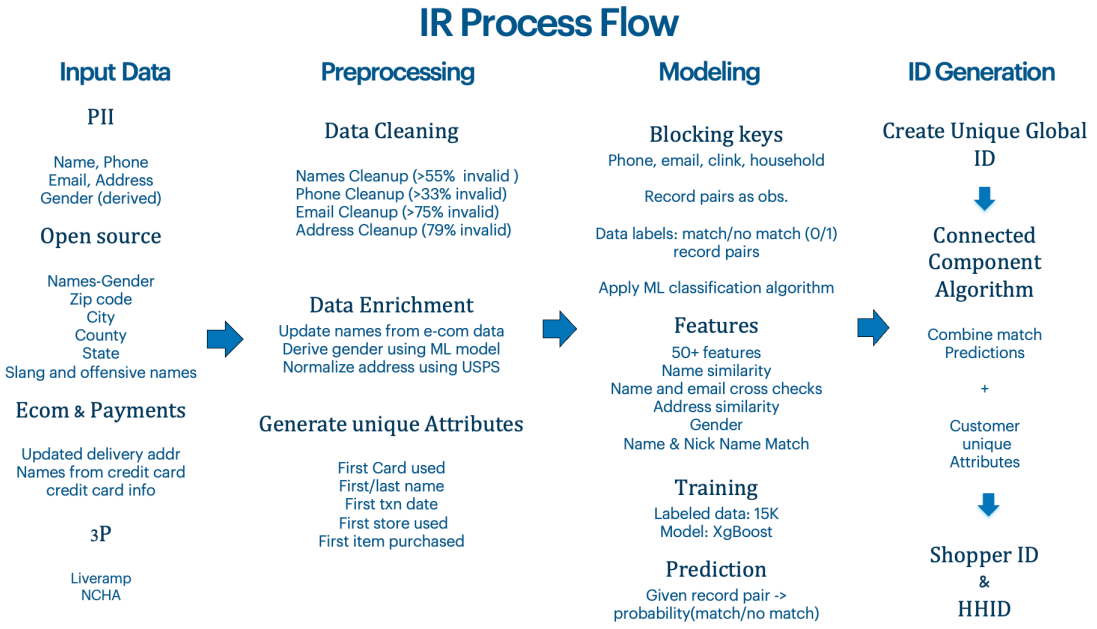
integration is from third party services such as doordash, uber eats etc. Such an integration provides data enrichment that is critical to building correct and update customer 360. Additionally it can also aid insights into identifying potential customers for other organizations such as delivery services subscriptions.

4 Identity Resolution: Technical Details

Mapping from card numbers (PII) to shopper id and household id consists of number of steps. While technical details is omitted here but the following is the architecture used to generate customers shopper id.

Entity resolution involves running number of steps in sequential manner. Not doing so may result in different shopper id and household id assignments. Here is the flow diagram from raw PII data to generating shopper id hhid.

Figure 3: Identity Resolution Process flow: from raw data to ID stamping



5 Some of customer 360 Use Cases

5.1 Creating Amazing Customer Experience

A recent research survey[1] by McKinsey & Company found that 'many organizations fail to see how improving the customer experience can create value. And more importantly despite their efforts, many struggle to understand what really drives the customer experience. The research revealed that improving the customer experience increased sales revenues by 2 to 7 percent, profitability by 1 to 2 percent, and the overall shareholder return by 7 to 10 percent.' Customers experience is backbone of business especially the ecommerce one. Ecom business can have many challenges from incorrect product description, slow delivery, poor accessibility, and suffer from long wait for customer service. This makes customer online shopping an unpleasant one. And in fact, for many customers their frustration can lead to bad reviews among their friends and on social media channels reaching hundreds of users in no time.

Among others one of the reason Albertsons customer reach out to call center is to update their household assignment. An example of this is to merge a group of card numbers to one household to avail accumulated reward points. With over 100+ million customers, this requires significant customer service personals and thus significant cost to company. While it is desired to have a convenient and time efficient process (such as automatic algorithmic process that can run in backend) to handle shopper and household id assignment, there is none at this time. Also the call center may not have access to other source of information such as third party data, payments data extracts PII, or even other first party data at Albertsons. Customer 360 is big win in such cases.

5.2 Accelerating Businesses Digital Growth

Online shopping have been on the rise for all sort of businesses in past few decades and Amazon has been a leading example of this. Online businesses are successful for many reasons such as low operating costs, easier scalability, flexible returns, quality customer support services, easier marketing and better security among many. Albertsons has made strategic investments such as building and enhancing micro fullfilment centers, fresh assortments, automation technologies, building own picking app among others. The company also made strong emphasis to online deliveries viz : 'lovable product',

[1] https://www.mckinsey.com/tr/our-insights/prediction-the-future-of-customer-experience

'fast shipping, and 'differentiated experiences'.

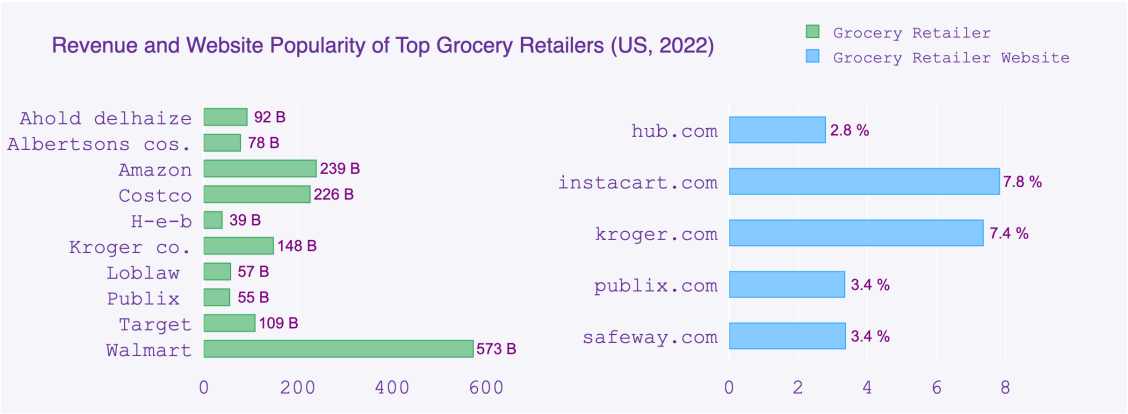


Figure 4: Top Retailers by Sales

Explosion of social media applications (for instance reddit.com) offers better interactions among customers where they can learn about other custimer's opinion and product reviews. Such customer interactions offers challenges as well as opportunities for digital businesses. For instance by leveraging customer 360, business can quickly learn about decline in sales and address the underlying product issues. In case of boost in sales, businesses can focus on product supply, and logistics to meet product's demand.

5.3 Marketing

The core of a great marketing strategy is to promote the brand/product through a meaningful story that connects people to its core values. With increased used of digital tools, consumers generate rich engagement history. Businesses can use this rich trail of information to create personalized marketing campaigns[2] for its customers. For instance given recent purchase history of a customer, business can choose to show specific ads to the customer that has maximum opportunity of a cross sell or use to engage them with favorite influencers video/post for brand building and increased product awareness. Businesses can design better A/B testing experiments such as analyze how different segments of population would respond to new business initiatives and such.

[2] Some of very successful campaigns: Nike's 1988 ad 'Just Do It':featuring a 80 year old marathoner Walt Stack that inspiring people, California Milk Processor Board: "Got Milk?": how life without milk so bland?, Apple's "Creativity Goes On": demonstrating use of Apple devices in creating simple joyous moments

The Albertsons companies spend $498 million in advertising cost being lowest among some top grocery retailers.

Table 1: Top Grocery Retailer's Revenue and Ads Budget (US, 2022)

Retailer	Revenue (billion $)	Ads Budget (billion $)	Revenue per $ ad spend
Albertsons	77.65	0.50	155.92
Amazon	514	20.6	24.95
Costco	227	-	-
Kroger	148	1.03	143.69
Walmart	567	3.90	145.38

The ratio of revenue to ads expenditure for Albertsons, Kroger, and Walmart are quite similar. These stores are quite similar in the sense that they have strong store as well as growing ecom presence. Since Amazon does not have as many physical store as other retailers, the company spends large sum of money on ads expenditure for brand building and for to remain a competitive business. **Note that all of the sales are not driven by ads and therefore this requires careful interpretation**. With customer 360, businesses can optimize the ads spending to generate higher returns and create brand value for their customers.

5.4 Personalized Recommendation

Imagine Netflix without its recommendation system, it is just a database of movies, not visible to naked eye. To watch a movie, you would need to precisely know its details such as title and manually search from a large database and then play. Quite inconvenient, right? Same with ecom grocery, without having a good recommendation system, the users would have to search every product they want from the pool of over half a million products. Needless to say, recommendation systems are backbone of successful ecom businesses. Top retailers like amazon attribute over 35% of revenue to successful recommendation system and uses it in every step: from product discovery to checkout. In fact for streaming businesses such as Netflix[3] (over

[3] Gomez-Uribe and Neil Hunt, Netflix's employees in 2016 asserted that "the combined effect of personalization and recommendations saved Netflix more than $1B per year."

75%), Youtube, Spotify, the majority of these company's revenue can be tied to a very successful recommendation systems.

Recommendation systems (recsys) are also a powerful tool for product discovery: showing customers something they would like and something they won't be able to find themselves. At the core, these recommendation algorithms are statistical machine learning[4] models trained on customer's past purchases, product views, likes, comments, PIIs (such as gender) and many other derived features. Armed with comprehensive view of customer experiences, customer 360 can enable business to deliver better personalized recommendations.

5.5 Early Identification of churning customers

For business of any size, higher churn rates can have devastating effect on its growth. With help of customer 360, bsinesses can consolidate customer data from various touchpoints including past interactions, support tickets, chat history, social media engagements and apply predictive models to infer likelihood of churn, well ahead they actually churn. This provides an early opportunity for business to make an intervention such as address any issues, offer promotional deals and incentives thus increasing customer satisfaction and increasing their lifetime value.

Above I have highlighted few of important use cases focusing on customer experience, digital growth, brand building, persoanlizaton and customer retention. Some of other use cases that business can benefit from customer 360 are aiding in regularity compliance and data security, leveraging unified data to build better products, making easier for many business and data science professionals within company to enhance their productively, among others.

[4] many of the state of art recommendation systems now use modern deep deep learning frameworks such as transformers

6 In House ID Graph's Performance to Live Ramp's ID Graph

LiveRamp is a data collaboration platform. Liveramp takes Albertsons PII data as input to their ID graph framework so called "Portrait Engine" to generate portrait ID. A portrait ID is synonymous to shopper id in scope as both of these map to individual customers. In the past, Liveramp has used their earlier ID graph engine called Abilitec graph (uses deterministic matching algorithm) to provide clink and hlink. Clinks and hlinks map to customer level id and household level IDs respectively. To measure the effectiveness of the in house built (Albertsons) ID graph output to LiveRamp's ID graph outputs, we compute the accuracy, precision score, recall score, area under roc curve.

6.1 Comparison Methodology

We manually went over many record pairs to create a validation benchmark dataset. This is different labeled data used in probabilistic model training (not described here due to copyright issues). The benchmark dataset has each record pair labeled as match or no match. This is assumed to be ground truth and used to compute the various metrics to assess the performance of in house ID graph and Live Ramp's ID graph outputs.

Live Ramp has access to various other data sources as customers credit card info in addition to Albertsons PII data. To make an apple to apple comparison, we limit the comparative analysis to only those record pairs sharing same blocking keys.

6.2 Independent Validation Dataset Results

Metric	Albertsons ID Graph	LiveRamp`s AbiliTec ID Graph	LiveRamp`s Portrait Engine ID Graph
Precision	97.93%	96.73%	97.41%
Recall	89.41%	83.93%	85.44%
Accuracy	89.32%	83.82%	85.60%
Area Under ROC curve	89.09%	83.54%	85.98%

Figure 5: Validation Data Metrics

For all the metrics, in house built ID graph model performs better than Live Ramp's ID graph model outputs for both AbiliTec ID graph and Portrait Engine ID graph model outputs.