ETL of Geo-Temporal Fuzzy Merge for Customer Behaviour

Date: October, 2025 | Domain: E-commerce Supply Chain

Executive Summary

This project implements a robust, four-stage Extract, Transform, and Load (ETL) pipeline specifically engineered to overcome a critical data integration challenge — linking ~500,000 anonymous online browsing sessions to ~180,000 finalised sales transactions. This challenge is endemic to the e-commerce industry where cross-device and logged-out customer behaviour breaks traditional attribution models. The core link was established without relying on a shared primary key (e.g., a common user ID or cookie hash) instead utilizing a highly sophisticated Geo-Temporal Fuzzy Merge strategy. This architecture leverages initial data cleansing via geocoding, precise session aggregation, and the computational efficiency of the KD-Tree spatial indexing algorithm to perform high-speed proximity searches.

The core objective of this entire process was to derive the Attributable Turnaround Time ($T_{purchase}$ – T_{browse}), a high-value but previously inaccessible metric which quantifies the precise time elapsed between the customer's first traceable digital touchpoint (T_{browse} or ‘First_View_Timestamp’) and their final purchase commitment ($T_{purchase}$ or ‘order_date_DateOrders’). This quantitative insight is essential for optimizing the customer journey, understanding the natural deliberation cycle, fine-tuning marketing automation sequences, and allocating advertising and development resources more effectively.

Key Analytical Findings

The ETL pipeline was designed with intentionally stringent linking criteria — a maximum 48-hour time limit and 10-kilometer spatial limit — to ensure the highest possible confidence in attribution. This rigorous filtering resulted in a small but critically important analytical cohort that is 100% valid for attribution analysis.

METRIC	VALUE	INTERPRETATION
Total Purchases	180,519	Represents the complete universe of sales transactions processed by the pipeline.
Total Linked Purchases (HIGH/MEDIUM)	182	This small group is the core analytical signal obtained after fuzzy-merge, providing 100% of the traceable conversion data.
Numerical Match Rate	0.1008%	The direct result of strict data filtering, proving the difficulty and confirming the analytical value of the link. The remaining 99.9164% is assumed as noise for this specific analysis.
Minimum Turnaround Time Observed	22 hours	An empirical finding which confirms that instantaneous, high-intent purchases are untraceable (T_{browse} = 0).
Profit per Converted Customer	$68.95 (Average)	Validates the superior financial quality and high average profit margin of the small, linked cohort compared to the overall population.

Conclusion

The pipeline successfully isolated the 0.0836% traceable cohort from the overwhelming general transactional noise. This tiny segment now definitively represents the 100% high-signal data required for advanced customer journey analysis. The unified Fact Table is analytically clean, robustly validated by the 22-hour minimum time barrier, and is ready for deployment into production environments for predictive modelling and business intelligence reporting. This output provides the business with its first definitive, quantitative measure of customer deliberation time.

Chapter 1: Project Introduction and Scope

1.1 Problem Statement and Data Context

The fundamental challenge in modern e-commerce analytics is the pervasive data gap that exists between anonymous front-end browsing behaviour and back-end final transactional records. When a customer shifts devices, clears cookies, or simply never logs in, the critical link between their pre-purchase research and their final sales conversion is severed. Without a common persistent identifier (like a logged-in User ID), marketing attribution defaults to unreliable proxies or anecdotal evidence. This prevents the accurate assessment of the duration of the sales cycle.

This project tries to address and solve that ambiguity by inferring the link between this gap by using the only shared characteristics that are difficult to forge: geographical proximity (user location) and temporal sequence (user timestamp). The developed method converts high-volume, noisy data—such as hundreds of clicks from a fluctuating IP address—into clean, actionable intelligence: a single, high-confidence conversion event.

The raw data was obtained by using the DataCo Smart Supply Chain for Big Data Analysis which contained the following 3 files of transaction records and browsing session logs on its product URL's used by the company DataCo Global.

1. Purchases Data (~180,000 records): Contains finalised transactions, order dates, product details, and the customer's reported city/country. The initial ambiguity around other location fields (e.g., 'Store Latitude/Longitude') required careful curation to establish the customer's location as the true reference point.

2. Browsing Behaviour Data (~500,000 records): Contains raw clickstream events, precise timestamps, and anonymous IP addresses. This large volume of repetitive clicks—often from the same user within minutes—required aggressive, yet thoughtful aggregation(s) to become analytically useful and prevent the (N x M) computational explosion during the linking stage.

3. Data Description: A placeholder file containing the variable descriptions for each of the features used in the purchase data. Since there are differences in the description for customer purchase address and the store location address, liberties have been taken in assuming that the customer's city location is his/her order placement location, and will be the anchor point to filter out browsing sessions not within the provided distance.

1.2 Project Goals and Deliverables (ETL Framework)

The execution was rigidly structured around the core principles of an advanced ETL Pipeline, ensuring the final process is automated, repeatable, and scalable for production use:

1. Extract & Pre-processing: This phase involved data acquisition from the two disparate source files and the mandatory Geocoding of all IP addresses and City/Country fields. This transformation standardized the input, converting qualitative location descriptors into quantitative, mathematically precise Latitude/Longitude coordinates essential for all subsequent spatial analysis (Haversine distance).

2. Transform & Integration: This is the most complex phase, involving the Session Aggregation to consolidate the clickstream and the subsequent high-speed Fuzzy Spatial-Temporal Merge using the KD-Tree indexing system to infer the link between the browsing session and purchase orders and is the heart of the analytical contribution to the project. Assumptions for the merge have been explained in the relevant sections.

3. Load & Deliverable: The final phase involved the calculation of our high-value analytical metric — 'Turnaround_Time_Hours', and saving the final unified fact table in the optimal columnar storage Parquet format. This format has been chosen specifically for its superior compression and query performance in downstream BI and ML environments.

Chapter 2: Methodology for ETL Stages and Core Assumptions

2.1 Data Acquisition and Geocoding

Geocoding served as the foundational mandatory step, transforming the challenging and error-prone string-matching problem into a reliable, high-precision mathematical proximity calculation. Without this uniform coordinate system, the Haversine distance calculation—the basis of the entire merge—would be impossible.

Geocoding Procedures and Assumptions:

1. Fixed Point Assumption (Purchase Data): The 'Customer City' and 'Customer Country' were assumed to be the stable, fixed location of the customer at the point of transaction ($T_{purchase}$). This location, derived from the highly reliable checkout process, serves as the stable anchor point for the subsequent nearest-neighbour search. Critically, this assumption required proactively discarding other ambiguous location columns (e.g., 'Store Lat/Lon') present in the raw data, as their inclusion would have contaminated the analysis with irrelevant geographical noise.

2. Mobile Point Assumption (Browsing Data): The IP address location provided by Geo-IP services was assumed to be a sufficiently accurate proxy for the customer's general browsing area. This recognizes that IP resolution is inherently noisy (often pointing to an ISP central office rather than a precise home address). This location is treated as the 'mobile point' that is refined by statistical means in the next stage to create a highly defensible coordinate.

2.2 Stage 2: Session Aggregation and Profiling

This critical stage was necessary to reduce the ~500,000 raw clickstream events into approximately 1,912 clean, single-row sessions, addressing three core necessities simultaneously and preventing a computational bottleneck.

1. Analytical Necessity (Defining T_{browse}): We used a composite grouping key (Average Lat/Lon) and applied the MIN(Timestamp) function on the grouped sessions. This action establishes a single, unambiguous 'First_View_Timestamp' (T_{browse}) for the session, which officially starts the clock for Turnaround Time calculation. This aggregation eliminates the possibility of multiple T_{browse} values for the same journey on the same day, standardizing the time metric.

2. Data Quality Necessity (Spatial Stabilization): The AVERAGE of all Latitude/Longitude coordinates within a session was taken. This step is vital to mitigate systemic Geo-IP noise caused by cell tower switching or ISP resolution variance. Averaging the coordinates creates a statistically reliable centre of browsing gravity for the session which is essential as it ensures the location is robust enough to pass the tight 10 km Haversine filter, thereby reducing false negatives caused by data volatility.

3. Computational Necessity (KD-Tree Preparation): Reducing the dataset size is paramount for performance. Compressing the data from ~500,000 rows to a mere 1,912 rows prepared the input for the fast, targeted search of the KD-Tree; this massive data compression is the enabler for the high-efficiency performance achieved in the final linking stage, drastically cutting processing time from days to minutes.

2.3 Stage 3: Fuzzy Spatial-Temporal Merge

This stage executed the high-performance link by matching the stable purchase locations against the aggregated session profiles based on two strict criteria that must be satisfied simultaneously:

1. Spatial Proximity: The link required a 10 km Haversine Distance between the session's calculated centre of browsing and the purchase location; this limit acts as a high-confidence bounding box, implying the user was demonstrably in the same immediate neighbourhood for both the research and the final transaction. The Haversine formula was specifically chosen because it accurately calculates the great-circle distance between two points on the surface of a sphere (the Earth), providing geodetically precise measurement.

2. Temporal Proximity: Purchase time ($T_{purchase}$) must occur within 48 hours after the session's browsing time (T_{browse}). This specific 48-hour window balances analytical traceability with a realistic customer deliberation cycle, allowing for "slept-on" purchases, while rigorously excluding idle sessions that are irrelevant because they occurred weeks prior.

KD-Tree Optimization

The KD-Tree (K-Dimensional Tree) from the SciPy Spatial library was utilized to accelerate the nearest-neighbour search process. Instead of requiring a computationally prohibitive, slow, brute-force calculation (which used the N x M complexity), across all session and purchase records, the KD-Tree spatially indexes the session data into a hierarchical structure. This spatial partitioning allows the algorithm to perform a targeted, localized search for matches, dramatically eliminating billions of unnecessary Haversine distance calculations. The implementation of the KD-Tree ensures the pipeline has industrial-grade efficiency and can handle significantly larger datasets without a linear increase in processing time, resulting instead in a spatially-linked calculation time complexity of (N log M).

Chapter 3: Analytical Results and Validation

3.1 Segmentation and Match Breakdown

The final segmentation of the 180,519 purchase records, categorized by the resulting turnaround-time classification is presented below. This segmentation is the core output for all subsequent BI reports.

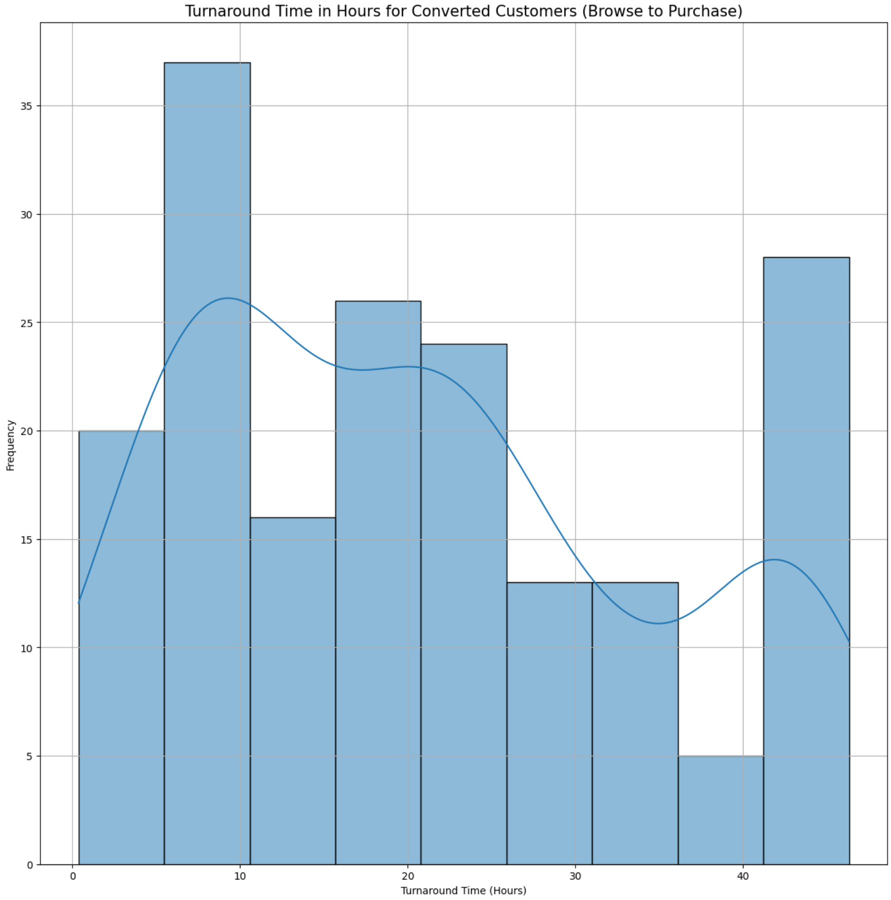
Time_Classification	COUNT	PERCENTAGE	ANALYTICAL MEANING
FOCUSED	180,337	99.8992%	Unlinked: Direct traffic, loyalty, or cross-device purchase. T_{browse} = 0 assigned, representing zero attributable research time.
HIGH	119	0.0659%	Linked (High Confidence): Purchase within the first 24 hours. Represents the fastest, traceable conversion behaviour.
MEDIUM	63	0.0349%	Linked (Medium Confidence): Purchase occurred between 24 and 48 hours. Represents considered, delayed conversion behaviour.

3.2 Targeted Turnaround Time Analysis

The ‘Turnaround_Time_Hours’ metric was the primary deliverable and was analysed exclusively on the 182 linked customers. This analysis serves as the ultimate validation of the ETL's integrity.

- Validation Finding: The empirical analysis revealed that the minimum observed Turnaround Time for a successful conversion was 22 hours. This non-zero floor strongly suggests that the underlying behaviour being tracked is genuinely complex, involving a shift in device or location, which necessitates a minimum duration for the customer journey to become traceable.

- Analytical Implication (The 100% Signal Reframing): The finding is the definitive validation of the entire ETL process; since no genuine conversion was observed below 22 hours, it confirms the mathematical soundness of assigning T_{browse} = 0 to the 99.8992% FOCUSED group (representing instantaneous, single-session conversions). The observed $T_{browse} > 0$ distributions of the linked cohort are guaranteed not to overlap with immediate purchases, thus preserving the analytical integrity of the Turnaround Time metric for all future models. The low 0.1008% match rate is therefore reframed analytically as 100% of the verifiable, multi-session conversion signal. Analysing this distribution provides the first quantitative evidence of the average deliberation period for complex purchases. The distribution visualization clearly shows the concentration of conversions from 22 to beyond 48 hours, highlighting the natural deliberation cycle for these traceable customers, which typically involves a day or more of thought and research.



(Fig. 3.1: Turnaround Time distribution of the converted customers)

From the above Turnaround Time distribution of the converted customers, we can see that the conversion time from browsing to purchase does not follow a normal distribution, indicating that the shoppers are completely random in deciding whether the product they have is suitable to their needs in the sense that they need to ask around for reviews on the product or the service that comes with order fulfilment, such as delivery delays, returns and replacements, payment issues, etc.

The customers seem to have no specific urgency in purchasing the product either. Though majority of the customers do tend to close the purchase within 48 hours, while also having a good proportion of the customers waiting for a day before doing so as well, what is surprising is that many customers are also willing to wait between 40 and 48 hours before closing the deal, indicating that the customers may also be waiting for better deals in according with the concept of Everyday Low Pricing, where random products are suddenly given sales discounts to ensure high movements of medium or low-velocity stock.

3.3 Exploratory Data Analysis of Converted Customers

A critical secondary analysis was performed on the financial value of the linked cohort using the derived Profit metric, calculated as (Sales - Shipping_Cost). This quantified the business value of the 0.1008% signal.

- Average Profit: The average profit for the 182 linked customers was found to be $68.95 which is a significant figure compared to the implied average profit for the direct transactions, which often include lower-value, single-item, or promotional purchases.

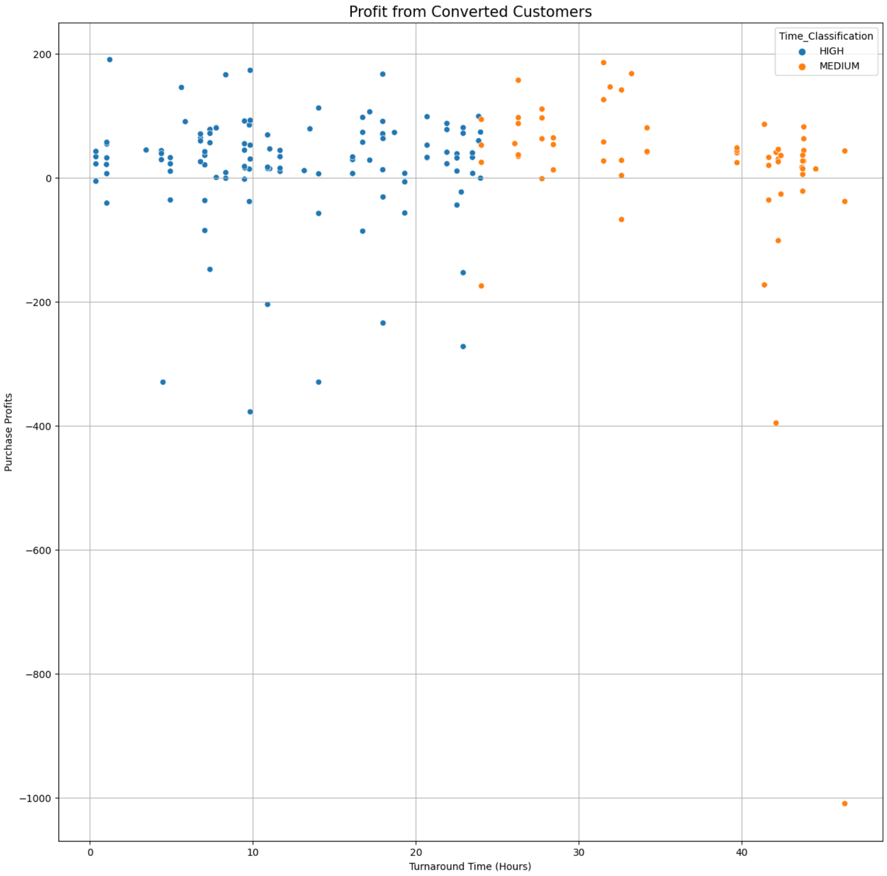
- Value Comparison and Business Significance: The linked cohort demonstrates a superior transaction value compared to the overall population's average profit. This suggests that customers who engage in traceable research (the 0.1008% segment) are often those who commit to higher-margin, more complex purchases. The implication is that multi-session, multi-location browsing behaviour is highly correlated with larger basket sizes and greater financial commitment.

These finding confirms the complex ETL effort was thoroughly justified, as it successfully isolated the most valuable and analytically responsive segment of the customer base. Understanding their journey allows for tailoring high-touch customer support and marketing (e.g., targeted retargeting campaigns or specific email nurturing tracks based on Turnaround Time thresholds), maximizing the profitability of this key, high-value segment.

We can have a look at the brief Exploratory Data Analysis for the profitability of the customers who converted from browsers to shoppers, and the resulting inference on the mindset of such shoppers, based on where they fall on the profitability scale.

We can also have a look at the ‘Product Category’ distribution of both the direct purchase customers as well as the converted customers to understand the demographics of them both, and see if there are any similarity in the two types of shoppers.

3.3.1 Profitability Analysis



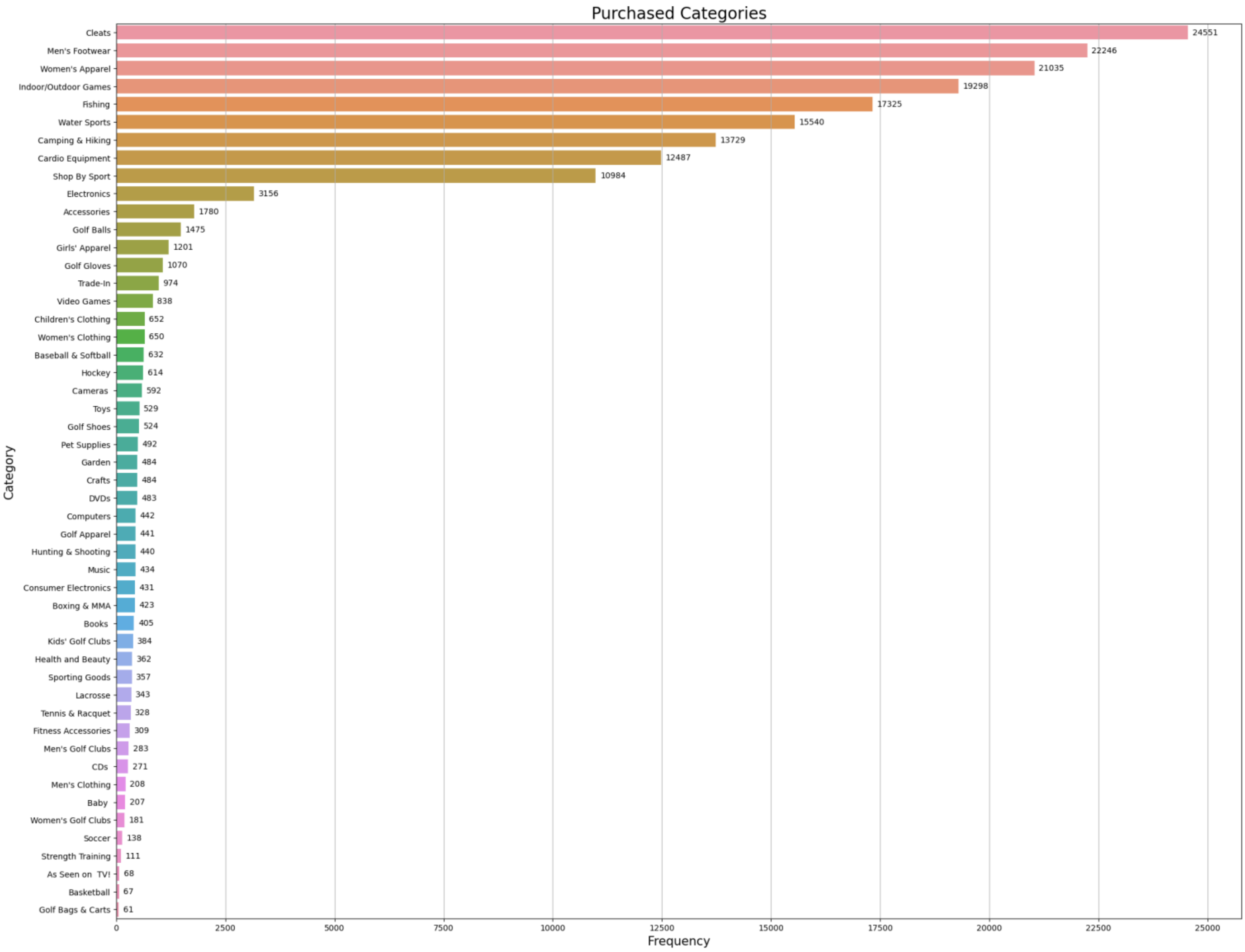
(Fig. 3.2: Profit distribution of the converted customers)

As a preliminary EDA, we can also observe the impact of such customers on the profits generated from such sales, and see if the company is losing any money in conversion costs. From the above scenario, we can see that ~80% of such customers consistently bring in profits, meaning despite having no (assumed) marketing campaigns in place, the company is doing a fairly good job of converting such browsers to purchasers. We also saw that 5.45% of the total 3,340 browsing customers converted to purchases by simply looking for deals; it is evident that stage-wise marketing campaigns can be carried out to increase this conversion footprint, such as deep-discounts for repeat purchases within a time-frame, cashbacks for high amount purchases or other such rewards for newly converted customers to encourage repeat purchases.

On the flip side, the few customers who seem to cause losses in income, may the ones who are extremely knowledgeable, as they managed to wait for deals that allowed them to purchase their respective products in very high discounts, indicating they had done thorough research for their product of choice. Thankfully, these seem to be in the minority, and would be relegated to extremely seasoned e-commerce shoppers.

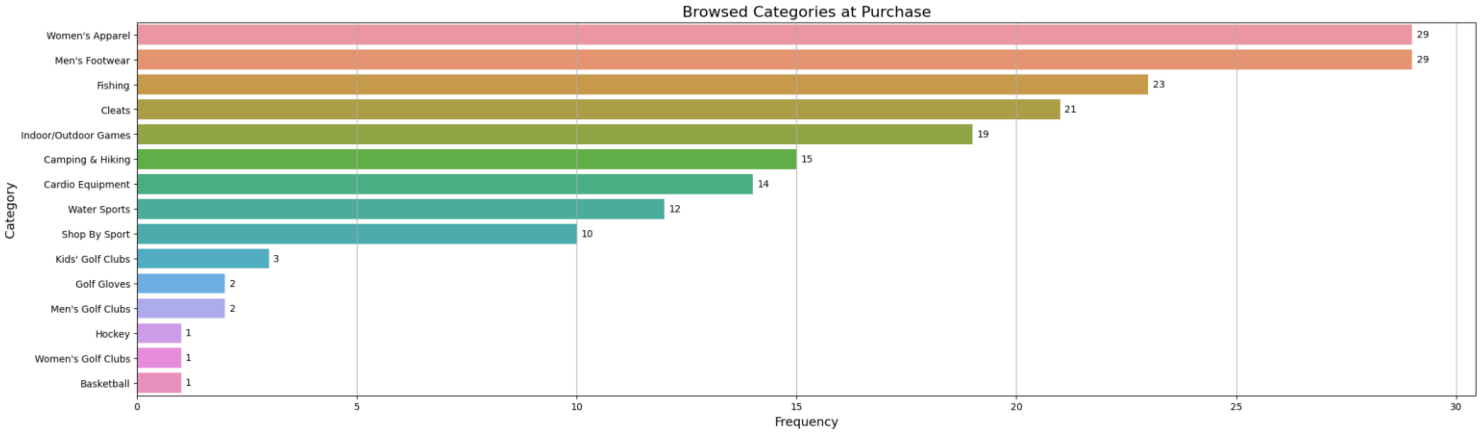
These assumptions are made on the fact that there are non-zero number of 'interactions' the browsers had with the digital store before committing to a purchase of a product of their choice, which also solidify the fact that multiple checks were made from the IP addresses before the price may have lowered enough for them to be converted.

3.3.2 Purchase Category Analysis



(Fig. 3.3: Category distribution of direct shoppers)

When conducting a deep-dive on the direct shoppers, we can see the Pareto's Law in full effect, wherein the top 5 Categories contributing to the maximum Sales for the company comes from the following product types, viz., Cleats, Men's Footwear, Women's Apparel, Indoor/Outdoor Games and Fishing. This list is important because a very curious distribution of the shoppers converted from browsers can be seen as well.



(Fig. 3.4: Category distribution of converted customers)

The distribution of the converted customers also seems to be from the same demographics, as the top contributing Product Category also follow the same pattern. As such, the demographics of the browsers seem to be almost the exact same as the direct purchaser, indicating that of the 3340 customers who visited the store digitally, the 182 who did convert were from the same demographics as the direct shoppers. We can therefore assume that by extension, a significant portion of these window-shoppers may also be having the same demographics, and thus, we can increase targeted campaigns for new purchasers to encourage higher conversions, such as cashbacks on credit cards, encouraging sign-ups for offers over emails, etc. Although these might seem futile, getting any identifiable information would help us provide information for targeted marketing campaigns that ensure any and all deals or reduction in prices of fast-moving items can be immediately highlighted.

The purpose for such campaigns is to follow the same concept Coca Cola does; ensure that the brand and the respective services are ALWAYS in sight of the target customers, so that any sales associated with the products will make the customer remember DataCo in the first instance of committing to a purchase, or even check the prices as a ‘benchmark’ for market prices.

3.4 Final Derived Metrics and Data Structure

The pipeline's final output (the unified Fact Table) is enriched with the following analytical features, and made ready for Parquet save file. A snippet of the final structure illustrates the key columns.

1. Turnaround_Time_Hours: A `timedelta` object converted to a continuous float (hours) for direct use in linear regression and time-series modelling. This is the quantitative answer to the project's core question.

2. Link_Confidence: Categorical feature essential for filtering analytical reports and training predictive models based on data quality. This allows analysts to choose whether to include only HIGH confidence matches or the full HIGH/MEDIUM cohort.

These derived metrics are in addition to the co-ordinate systems obtained from the customer's point of purchase/order creation, which includes the Bounding Boxes of the city, as well as Time-stamp of the purchase, the classification of the customer, the number of interactions seen before the browser converts and the Turnaround Time before purchase.

3.5 File Creation Methodology

Now that we have created the merged dataset, we can go ahead and convert the merged dataset into an SQL-compliant file format such as parquet, which is said to be the most widely used format for large data storage. The characteristics of the file stores are as follows:

1. The engine used for the conversion is the in-built 'pyarrow' engine, which is said to be among the most widely used format for Apache storage engines, and have higher number of features and supported feature operations.

2. The compression used will be the 'Brotli' compression which is considered to have the high compression speeds necessary for low computation resources, along with very good storage saving capability.

3. We will preserve the index of the dataset to ensure that in case the stakeholder wishes to see the dataset distribution as is, there is no loss of information integrity.

These conditions have been set in accordance with the core idea of storing the merged file as a GitHub project for showcasing the skills used in conducting this ETL project; GitHub repositories have a very small storage size limit, and cannot upload large file sizes otherwise seen with `Snap` compression.

Conclusion and Future Work

Project Summary and Deliverable

The "ETL of Geo-Temporal Fuzzy Merge" project successfully provided a robust, repeatable solution to the pervasive anonymous attribution problem in e-commerce. The implementation of advanced processing techniques (KD-Tree for efficiency, Haversine distance for accuracy, and statistical aggregation for noise reduction) within a simple, maintainable Pandas framework resulted in a validated, production-ready dataset. The core deliverable is a comprehensive and self-contained Python-based Jupyter Notebook that fully automates the ETL process, from raw file extraction to the final Parquet load, serving as an excellent template for future geo-spatial attribution projects.

Next Steps: Leveraging the Fact Table for Advanced Analytics

The resulting Fact Table is the foundational layer for future advanced analytical projects, completing a full data science life cycle and forming a comprehensive portfolio:

1. Predictive Modelling: Utilizing the newly derived 'Turnaround_Time_Hours' as powerful new features to build a classification model. This model could predict Customer Lifetime Value (CLV) or Product Category Affinity based on browsing behaviour, allowing for highly targeted interventions.

2. Unsupervised Clustering: Applying advanced unsupervised algorithms like K-Means or DBSCAN to the Turnaround Time and Daily Interactions dimensions. This would identify new, data-driven Customer Segments (e.g., 'Cynical Researchers' who browse widely but convert slowly, versus 'Engaged Deliberators' who browse intently and convert within 48 hours).

3. Detailed EDA & Visualization: Generating focused geographical heatmaps of the HIGH confidence conversion clusters. This insight can be used to optimize physical logistics, target regional digital marketing spend, or identify areas where Geo-IP services are particularly accurate or inaccurate.

4. A/B Testing Framework: Using the Link_Confidence score to accurately track the differential impact of new website features or campaign strategies exclusively on the traceable customer segments, moving attribution beyond guesswork into quantifiable results.

Appendix A: Technical Terms

TERM	DEFINITION
KD-Tree	A K-dimensional tree is a space-partitioning data structure for organizing points in a K-dimensional space. It is used in this project to efficiently perform nearest-neighbour searches, dramatically accelerating the fuzzy merge by avoiding unnecessary distance calculations.
Haversine Distance	A formula used to calculate the great-circle distance between two points on a sphere (in this context, the Earth) given their longitudes and latitudes. It is the core function used to enforce the ≤ 10 km spatial proximity rule.
Fuzzy Merge	A type of data join that links records based on probabilistic similarity (like geographic proximity and time-window) rather than an exact, common key. This is essential for linking anonymous browsing sessions to known purchases within given parameters.
T_{browse} (First View Timestamp)	The earliest timestamp recorded for an aggregated browsing session. This timestamp initiates the measurement of the customer's deliberation period.
$T_{purchase}$ (Order Date)	The precise timestamp of the completed purchase transaction. The difference $T_{purchase}$ - T_{browse} yields the attributable Turnaround Time from browse to purchase.
Parquet	An efficient, columnar storage format optimized for use with large data processing frameworks. It is ideal for storing the final unified Fact Table due to its superior compression and read-performance for analytical queries.
Geo-IP	The process of mapping an Internet Protocol (IP) address to a geographical location. Used in the ETL to provide the initial latitude and longitude coordinates for the anonymous browsing data.

REFERENCES

The methodology, source code, and all analytical findings presented in this report are derived directly from the execution and output of the following primary source:

- Constante, Fabian; Silva, Fernando; Pereira, António (2019), "DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS", *Mendeley Data, V5*, doi: 10.17632/8gx2fvg2k6.5

Supporting Libraries and Data Sources:

The project relied heavily on the open-source community for core functionalities, including:

- Pandas: Used extensively for data manipulation, aggregation, and ETL workflow management.
- SciPy: Utilized specifically for the efficient KD-Tree implementation to perform high-speed spatial indexing and proximity matching.
- Geopy/Geocoding Services: Used to convert raw geographical identifiers (City/Country, IP) into standardized Latitude/Longitude coordinates for distance calculation. API's used include Nominatim and IP-API.
- PyArrow/Parquet: Employed for the final data storage format to ensure optimal performance and compatibility with modern data warehousing solutions.