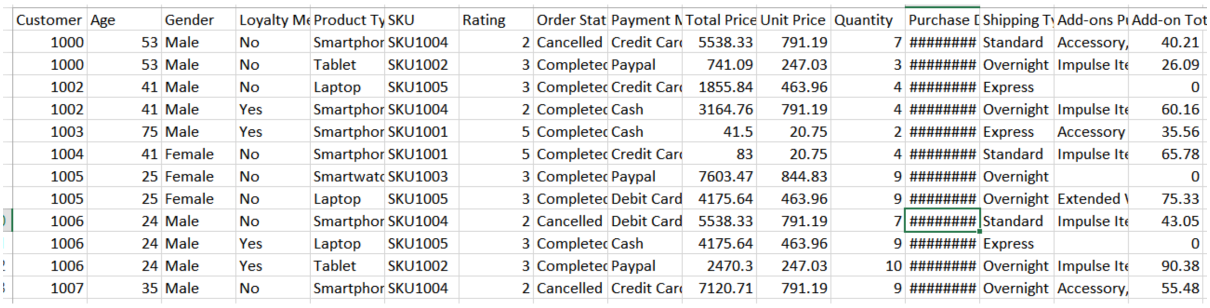
Insight Horizon: A Salesforce Data Cloud Project

Problem Statement

The goal of this project is to transform this scattered data into well-organized, actionable lists that enable businesses to identify trends, streamline workflows, and make informed decisions with greater precision and speed. By leveraging the power of a modern data cloud, we aim to bridge the gap between raw data and actionable intelligence**.**



Customer	Age	Gender	Loyalty Me	Product Ty	SKU	Rating	Order Stat	Payment M	Total Price	Unit Price	Quantity	Purchase D	Shipping Ty	Add-ons P	Add-on Tot
1000	53	Male	No	Smartphor	SKU1004	2	Cancelled	Credit Car	5538.33	791.19	7	########	Standard	Accessory,	40.21
1000	53	Male	No	Tablet	SKU1002	3	Completec	Paypal	741.09	247.03	3	########	Overnight	Impulse Ite	26.09
1002	41	Male	No	Laptop	SKU1005	3	Completec	Credit Car	1855.84	463.96	4	########	Express		0
1002	41	Male	Yes	Smartphor	SKU1004	2	Completec	Cash	3164.76	791.19	4	########	Overnight	Impulse Ite	60.16
1003	75	Male	Yes	Smartphor	SKU1001	5	Completec	Cash	41.5	20.75	2	########	Express	Accessory	35.56
1004	41	Female	No	Smartphor	SKU1001	5	Completec	Credit Car	83	20.75	4	########	Standard	Impulse Ite	65.78
1005	25	Female	No	Smartwatc	SKU1003	3	Completec	Paypal	7603.47	844.83	9	########	Overnight		0
1005	25	Female	No	Laptop	SKU1005	3	Completec	Debit Card	4175.64	463.96	9	########	Overnight	Extended \	75.33
1006	24	Male	No	Smartphor	SKU1004	2	Cancelled	Debit Card	5538.33	791.19	7	########	Standard	Impulse Ite	43.05
1006	24	Male	Yes	Laptop	SKU1005	3	Completec	Cash	4175.64	463.96	9	########	Express		0
1006	24	Male	Yes	Tablet	SKU1002	3	Completec	Paypal	2470.3	247.03	10	########	Overnight	Impulse Ite	90.38
1007	35	Male	No	Smartphor	SKU1004	2	Cancelled	Credit Car	7120.71	791.19	9	########	Overnight	Accessory,	55.48

The dataset includes:

- **Customer IDs**
- **Product names customer purchased**
- **Total product quantities or items purchased**

Unfortunately, the dataset is far from perfect. It includes duplicate records and lacks a direct unique identifier—a critical issue for generating accurate insights.

Project Objective

Our goal is to:

1. Analyze the dataset for correctness and completeness.
2. Process the data through a series of steps including cleaning, transforming, and enriching until it is ready for analysis.
3. Produce a list of customers who have spent more than **$20,000 across all purchases**.

The sourcing dataset is available here.

Steps to Transform and Activate Data

Step 1 - Analyze the Dataset

The first step involves gaining familiarity with the dataset. Key tasks include:

- Checking if all necessary fields are present (e.g., Customer ID, purchase amounts, product details).
- Understanding the dataset structure (relationships between products and customer purchases).
- Highlighting any immediately visible data issues, such as missing fields, duplicate rows, or inconsistent data types.

Outcome

A general understanding of dataset completeness and issues that need to be addressed.

Step 2 - Create the Architecture Diagram

Before jumping into action, you need to design the data architecture. A flow or architecture diagram will visually represent:

- **Data Lake Objects** housing raw data, transformed data, and processed outputs.
- **Data Model Object relationships**, such as defining linkages between customer identifiers and total purchase values.

Use tools like Lucidchart, Visio, or even a simple whiteboard to map this out. Example relationships:

- `Customer -> Product Purchases -> Aggregated Purchase Behavior.`

Outcome

A clear structure that outlines how data moves from ingestion to segmentation and activation.

Step 3 - Check Data Quality & Resolve Missing Information

Data quality is critical for trustworthy results. Key actions include:

- Ensuring all required fields (IDs, purchase amounts) are populated.
- Verifying data types match their expected formats (numeric for purchase amounts, string for product names).
- Handling missing values through imputation, default values, or logical assumptions.

Use data profiling tools or frameworks like Talend, Great Expectations, or manual SQL queries.

Outcome

An enriched dataset free from missing or improperly typed fields, ready for transformation.

Step 4 - Prepare the Data Lake Objects and Data Model Objects

The next step is to define:

- **Raw Data Lake Objects:** Store the original, untouched dataset.
- **Processed Data Lake Objects:** Prepare a layer where data starts undergoing cleanup and transformation.
- **Data Model Objects:** Create structured tables or views linking cleaned customer data to aggregated purchase data.

Outcome

A layered storage system within the data lake designed to track progress from raw data transformations to final outputs.

Step 5 - Perform Data Cleanup and Deduplication

Begin the transformation process by targeting duplicates and cleaning noisy data using **Identity Resolution.**

- Deduplicate based on customer IDs and purchase entries.
- Merge related entries (e.g., where customer IDs appear with similar names but different formats).

Popular tools for such tasks include Apache Spark, PySpark, or SQL scripts.

Outcome

A granular, deduplicated dataset that correctly represents unique customer purchase history per ID.

Step 6 - Apply Calculated Insights

After cleanup, compute relevant metrics, such as:

1. **Total Purchase Value by Customer ID:** Sum up all purchases grouped by Customer ID.
2. **Average Purchase Frequency per Customer:** Calculate how often each customer purchases.

Store these actionable insights as part of a new dataset (e.g., a "Calculated Insights Table").

Outcome

A set of ready-to-use calculated data that highlights customer purchase habits.

Step 7 - Perform Segmentation

Segmentation involves targeting customers that match a specific condition—such as aggregate spending above $20,000.

Use SQL or Python scripts to segment the dataset, creating groups like:

- "High-value spenders" (>$20,000)
- "Casual customers" (<$20,000)

Example SQL Query:

```
SELECT customer_id, SUM(total_spent) AS total_spent

FROM transactions

GROUP BY customer_id

HAVING total_spent > 20000;
```

Outcome

An exportable list of high-value customers ready for activation.

Step 8 - Set Up Activation Targets

Once your segmentation is complete:

1. Load the high-value customer list into activation targets, such as CRM tools (Salesforce, HubSpot) or marketing automation software.
2. Validate the exported data to ensure accuracy.

Outcome

A clean, activated target list set up for engagement campaigns or further analytics.

Wrapping It All Together

By following this comprehensive process, we've transformed a messy dataset into an actionable list of high-value customers spending more than $20,000.

Your Next Steps

Turning raw data into meaningful insights is a skill every data professional should master. Got a messy dataset of your own? Use this method or adapt it to fit your goals.

Need expert guidance? Collaborate with your peers or data teams, and remember—every great data story begins with a structured approach.