CUSTOMER ENTITY RESOLUTION -

PRODUCT RECOMMENDATION

Personalized product recommendations that understand you better than anyone else

April 11, 2023

Submitted by Team: **APPS**

ARUNITA SARKAR

PRIYADHARSHINI KASI

PUNAM SHET

SOUMYATA JENA

1. Abstract

In the era of data-driven decision-making, businesses are increasingly leveraging customer data to gain insights and improve their marketing strategies. One powerful application of customer data is in identifying products that are likely to be of interest to specific customers and providing personalized recommendations. This can significantly improve customer satisfaction, and engagement, and drive sales for the business. However, businesses must also develop an effective customer entity resolution strategy that accurately identifies and links data records to the correct real-world entity while minimizing errors and false positives. This paper aims to explore the importance of using customer data for personalized recommendations and the challenges and strategies associated with customer entity resolution in the context of e-commerce businesses.

2. Introduction

In today's highly competitive e-commerce landscape, businesses need to harness the power of customer data to gain a competitive edge. One effective way to do this is by identifying products that are likely to be of interest to specific customers and providing personalized recommendations. Personalized recommendations can enhance customer satisfaction, increase engagement, and drive sales by offering relevant and timely product suggestions. However, to achieve accurate and reliable personalized recommendations, businesses must first develop an effective customer entity resolution strategy that can accurately identify and link data records to the correct real-world entity.

3. Importance of Personalized Recommendations

Personalized recommendations are critical for businesses to improve customer satisfaction and engagement. Businesses can identify patterns and trends that inform personalized recommendations by analyzing customer data, including browsing history, purchase behavior, and preferences. For example, if a customer frequently purchases running shoes and sports apparel, a recommendation engine can suggest related products, such as running socks or fitness accessories. Such personalized recommendations can enhance the customer's shopping experience, increase the likelihood of repeat purchases, and foster customer loyalty.

4. Goal

To improve customer satisfaction and engagement and drive sales, businesses must ensure personalized recommendations' accuracy and relevance. This requires an effective customer entity resolution strategy.

5. Proposed Solutions

I. Customer Entity Resolution:

Customer entity resolution involves accurately identifying and linking data records to the correct real-world entity in a database or system, such as a customer or a company. This process helps eliminate duplicate or redundant data and ensures that customer information is accurate and up to date. Customer entity resolution is critical for building a comprehensive and accurate customer 360-degree view, which can help businesses gain insights into customer behavior, preferences, and interactions, and facilitate personalized marketing, customer service, and decision-making.

II. Recommendation Systems:

Recommendation systems use algorithms and data analysis techniques to provide personalized recommendations to customers for products or services they are likely to be interested in. These systems analyze historical customer data, such as browsing behavior, purchase history, and preferences, to make relevant and timely recommendations. Recommendation systems can help businesses improve customer engagement, increase sales, and enhance customer satisfaction by providing personalized and convenient shopping experiences.

E.g., **Collaborative filtering** is a technique used in recommendation systems to provide personalized recommendations based on user behavior and item preferences.

- User-based collaborative filtering relies on users with similar past behavior or preferences.
- Item-based collaborative filtering relies on items that are similar in terms of user interactions.

III. Customer Segmentation:

Customer segmentation involves dividing customers into distinct groups based on common characteristics or behaviors. This can help businesses better understand their customers and tailor their marketing efforts to specific groups. By analyzing customer data, businesses can identify patterns, trends, and preferences among different segments, which can be used to optimize marketing strategies, product offerings, and customer experiences.

Two common types of customer segmentation:

- **Based on Behavior:** This involves grouping customers based on past behaviors, such as purchasing behavior, browsing behavior, engagement with marketing campaigns, loyalty program participation, and other relevant actions. By analyzing customer behavior, businesses can identify patterns and characteristics among customers to create segments for targeted marketing efforts. For example, customers who frequently purchase high-end products or engage with email campaigns may be grouped into a segment to receive personalized promotions or offers.

- **Based on Preferences:** This involves grouping customers based on their stated preferences, interests, or demographic information such as age, gender, location, interests, hobbies, or other relevant preferences. By understanding customer preferences, businesses can create segments for targeted marketing messages, product recommendations, or offers. For example, customers who have expressed an interest in eco-friendly products or a specific type of cuisine may be grouped into a segment to receive relevant promotions or recommendations.

6. Data Source

The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing orders from multiple dimensions: from order status, price, payment, and freight performance to customer location, product attributes, and finally reviews written by customers.

Kaggle Dataset Link - https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

7. Data Schema

The data is divided into multiple datasets for better understanding and organization.

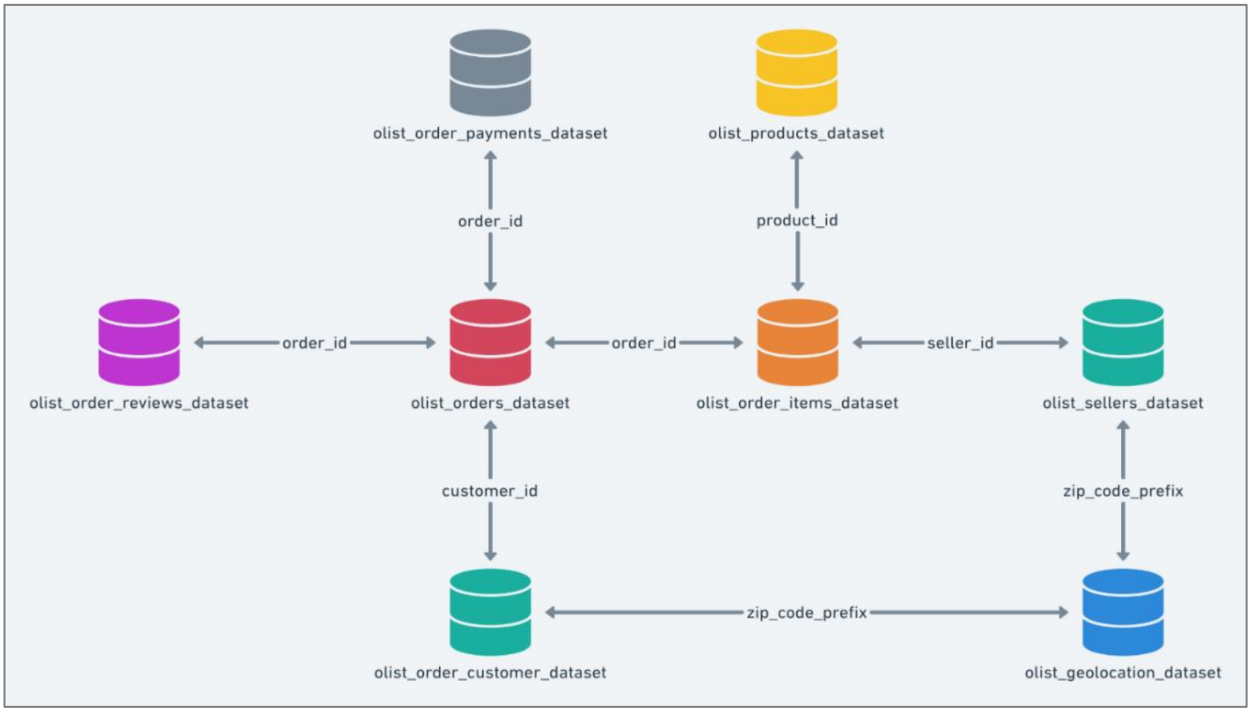

FIG 1. Data Schema of Customer 360-degree View

8. Input Data Overview

- **Customers Dataset:** This dataset has information about the customer and its location. One can use it to identify unique customers in the orders dataset and to find the orders' delivery location.

    - Each order in the system is assigned a unique customer_id, which means that the same customer may have different ids for different orders. The customer_unique_id field in the dataset is used to identify customers who have made repurchases at the store, as without it, each order would have a different customer associated with it.

- **Geolocation Dataset:** This dataset has information on Brazilian zip codes and their lat/lng coordinates. One can use it to plot maps and find distances between sellers and customers.

- **Order Items Dataset:** This dataset includes data about the items purchased within each order.

- **Payments Dataset:** This dataset includes data about the order payment options.

- **Order Reviews Dataset:** This dataset includes data about the reviews made by the customers.

    - After a customer purchases at Olist Store, the seller is notified to fulfill the order. Once the customer receives the product or the estimated delivery date is due, the customer is sent a satisfaction survey via email to provide feedback on the purchase experience and leave comments.

- **Order Dataset:** This is the core dataset. From each order, one will find all the other information.

- **Products Dataset:** This dataset includes data about the products sold by Olist.

- **Sellers Dataset:** This dataset includes data about the sellers that fulfilled orders made at Olist. One can use it to find the seller's location and to identify which seller fulfilled each product.

- **Category Name Translation:** Translates the product_category_name to English.

9. Architecture

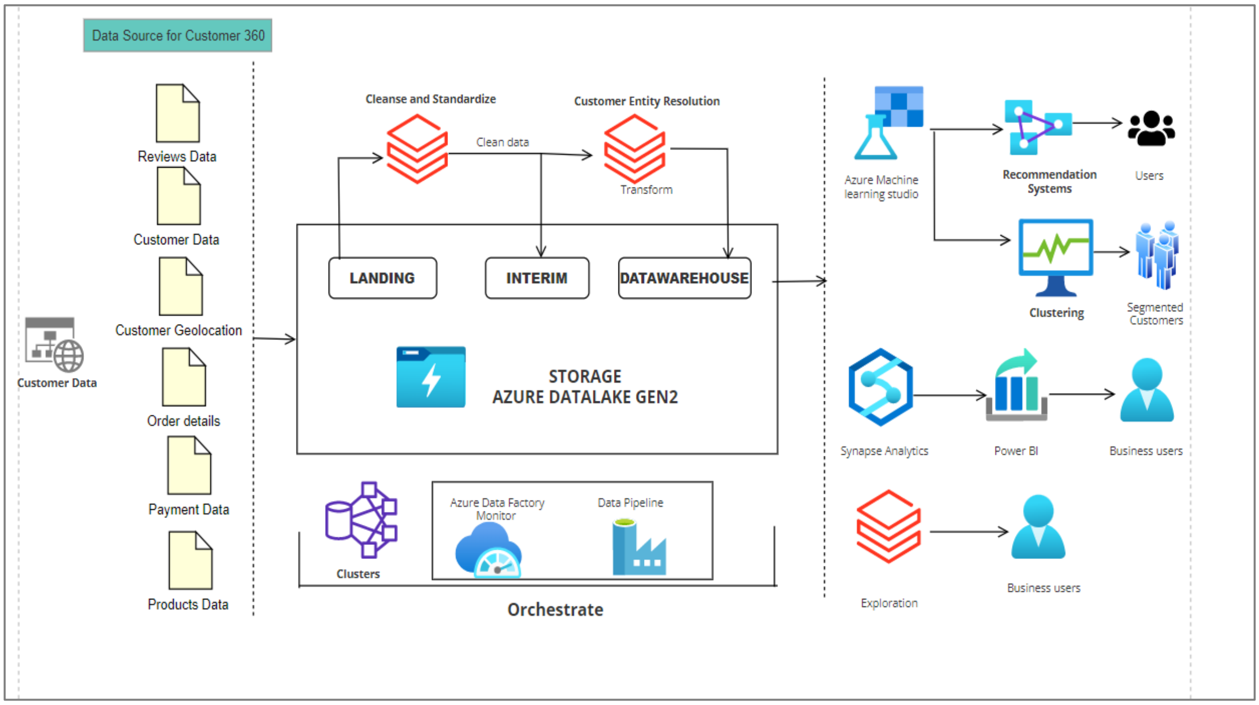

FIG 2. Architecture

We will be getting data from the Kaggle website. This dataset consists of Customer related data, which are Customer profile data, Customer Geolocation, Order details of the customers, Payment data, Product related data and Review data. Once we get the data from Kaggle we will be storing the data in an MS SQL server as it is an automated process, and this helps in failover scenarios like when the Kaggle website is down. Once the data is in the MS SQL server, we will be loading the data into the Azure Data Lake Gen 2. This will be in the form of blob storage. Once the data has landed in the Data Lake, we will be pulling the data into the Azure Databricks.

- In Azure Databricks we will be creating clusters and performing compute part of Data engineering. The data is cleaned and standardized. Once the data is preprocessed, we will be storing the data again in the storage. The clean data is retrieved, and we will be performing Customer Entity resolution on the customer data where we will be creating the customer 360-degree view.

- This will enable us to identify the customer by a unique ID. Once the customers are identified as a part of Customer Entity Resolution, we will be storing the data in the Data Lake. Then the Recommendation system is built on the customers’ review data. Then the data will be used for customer clustering purposes and later it can be analyzed from the business perspective. All these pipelines are built on the Data Factory which helps in the automation and monitoring process.

1. Data Extract from Kaggle

I. Data is collected from Kaggle - This is Brazilian eCommerce data which has all the relevant data to perform customer entity and recommendation systems.
II. Download the data from Kaggle.
III. Pipeline creation - Here we will be creating an automated pipeline for pulling the data from Kaggle to MS SQL Server.
IV. Automate/Schedule - This process will be automated in the Data Factory running in the scheduled fashion.

2. ETL Process

I. Extract - Data is extracted from Kaggle and loaded into the SQL server. From the SQL server data is transferred to Azure Storage Datalake Gen2. Data is stored here in BLOB format.
II. Transform - Source data is cleaned and transformed using Azure Databricks to check Data Quality, create Data Catalog, and obtain Data Lineage. Here the customer entity resolution is obtained for each customer.
III. Load - Data is loaded into the Azure storage Data Lake Gen2 and analytics is performed in Synapse Analytics.

3. Customer Entity Resolution

The customer entity is a fundamental aspect of any business, as it represents the individuals or organizations who purchase goods or services from the business. A common business problem related to the customer entity is how to attract and retain customers while maximizing profits.

One challenge businesses face is identifying their target customer base and understanding their needs and preferences. By doing so, businesses can tailor their products or services to meet the specific needs of their customers, which can lead to increased customer satisfaction and loyalty.

Another issue businesses may face is customer churn, which occurs when customers stop using a business's products or services. This can result from a variety of factors, such as poor customer service, competition from other businesses, or changes in the market. To combat customer churn, businesses may need to improve their customer service, offer incentives for customer loyalty, or differentiate themselves from competitors.

The idea behind building personalized experiences is deriving actionable insights from all bits of information that can be gathered about a customer. Data generated from many sources like sales transactions, website browsing, product ratings, customer surveys, support center calls, third-party data purchased from data aggregators, online trackers, and more can come together to form a 360-degree view of the customer. To build a true customer 360 leveraging all that data, organizations must establish common customer identities, linking together customer records across disparate data sets so they can collect all the common information about one customer – this is called **Customer Entity Resolution**.

The process of matching records to one another is known as entity resolution. It is a process used to identify and consolidate customer data from various sources to create a unified view of each customer. Several algorithms can be used for customer entity resolution, depending on the specific requirements and characteristics of the data. When dealing with entities such as customers' data, the process often requires the comparison of first name/last name and address information which is subject to inconsistencies and errors. In such scenarios, we often rely on probabilistic **(fuzzy) matching techniques** that identify matches based on degrees of similarity between these elements. There is a wide range of techniques that can be employed to perform such matching. The challenge is not just to identify which of these techniques provides the best matches but how to compare one record to all the other records that make up the dataset in an efficient manner.

- **Fuzzy matching** is a type of algorithm used in data matching and record linkage, which is used to identify pairs of records that may represent the same real-world entity even if the records are not identical. This is useful when dealing with data that may contain errors or inconsistencies, such as misspelled names or incomplete addresses.
- The basic idea behind fuzzy matching is to compare the similarity between two strings or records using a similarity score, which considers the differences and similarities between the strings. The similarity score can then be used to determine whether two strings or records are a match or not.

Overall, fuzzy matching is a useful technique for identifying similar records and can be used in a wide range of applications, including customer entity resolution, fraud detection, and data cleansing. However, it is important to choose the right algorithm for the specific application and to carefully tune the parameters to achieve the best possible results.

10. Data Pipeline

Azure Data Factory is a cloud-based data integration service offered by Microsoft Azure that enables businesses to create, schedule, and manage data pipelines to move and transform data between various sources and destinations. It allows organizations to process and transform data from a wide range of sources, including on-premises data sources, cloud-based data sources, and software-as-a-service (SaaS) applications.

Azure Data Factory provides a range of features to help organizations build data integration solutions, such as:

I. Data Integration - Azure Data Factory enables businesses to ingest data from various sources, including structured, unstructured, and semi-structured data, and prepare it for analytics or processing.

II. Data Transformation - With Azure Data Factory, businesses can transform data as per their business requirements, using a wide range of transformation activities, such as mapping, filtering, and aggregation.

III. Workflow Orchestration - It enables users to orchestrate complex data processing workflows with dependencies and trigger-based executions, allowing businesses to automate their data pipelines.

IV. Monitoring And Management - Azure Data Factory provides monitoring and management features to enable businesses to track the performance of data pipelines, monitor data flows, and ensure the timely execution of workflows.

V. Integration with Azure Services - Azure Data Factory integrates with other Azure services such as Azure Synapse Analytics, Azure Databricks, and Azure HDInsight to provide a comprehensive data integration solution.

Overall, Azure Data Factory simplifies the process of creating and managing data integration pipelines, allowing businesses to ingest, transform, and integrate data from various sources to deliver insights and drive business value.

11. Storage

Azure Data Lake Gen2 is a cloud-based service offered by Microsoft as part of the Azure cloud platform. It provides a highly scalable and cost-effective storage solution for big data and analytics workloads.

The architecture of Azure Data Lake Gen2 is built on top of Azure Blob Storage, which provides highly scalable and durable object storage. Data is organized into containers, and each container can contain an unlimited number of blobs. These blobs can be secured using shared access signatures. Azure Data Lake Gen2 provides a hierarchical namespace that enables users to organize data into directories and subdirectories. This helps to organize data and make it easily accessible for processing and analysis.

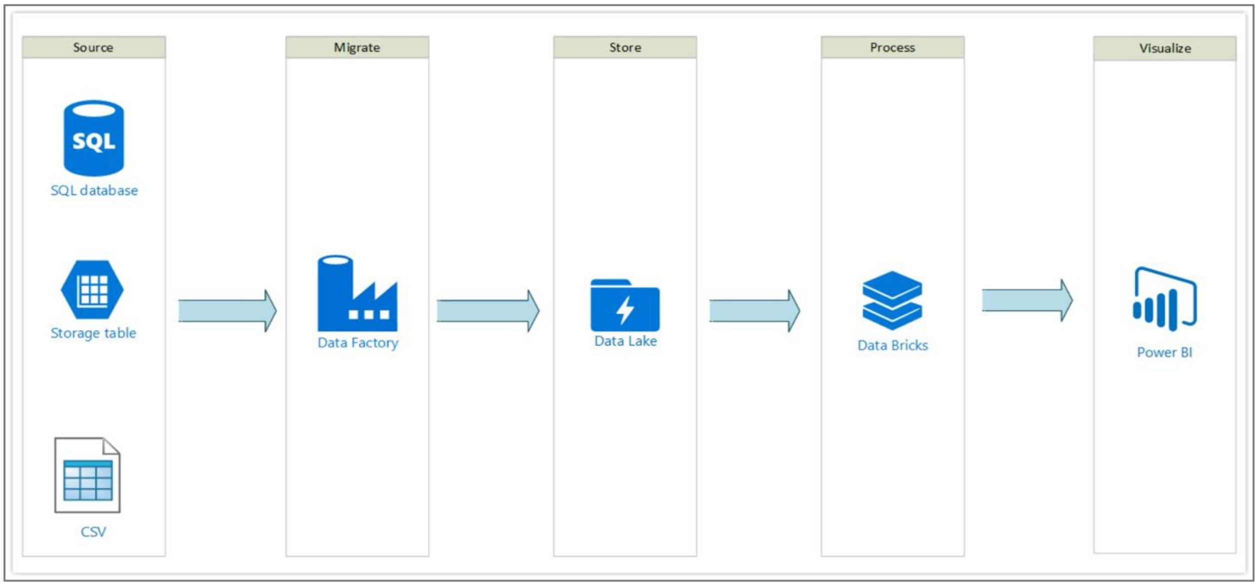

FIG 3. Connect, Ingest, and Transform Data

Pre-Requisites: To connect MS SQL Server database to Azure Data Lake Gen2 using Azure Data Factory (ADF), we need to have certain steps to be fulfilled first:

I. An Azure subscription.
II. An Azure Data Factory instance.
III. A self-hosted integration runtime to connect to the on-premises SQL Server database.
IV. A connection string for the on-premises SQL Server database.
V. An Azure Storage account to store the data in Azure Data Lake Gen2.
VI. An Azure Data Lake Gen2 linked service to connect to the Azure Storage account.

Once we have these pre-requisites, we can follow the steps below to connect the MSSQL database to Azure Data Lake Gen2 using ADF:

I. Create an Azure Data Factory instance - we need to create an Azure Data Factory instance in our Azure portal.
II. Create a linked service for your MSSQL database - In the Azure Data Factory portal, create a linked service for the MSSQL database by providing the required connection information such as server name, database name, username, and password.
III. Create a linked service for Azure Data Lake Gen2 - Next, we need to create a linked service for Azure Data Lake Gen2 by providing the required connection information such as storage account name, file system name, and authentication method.
IV. Create a dataset for your MSSQL database - Create a dataset for the MSSQL database that defines the source data for the pipeline. We need to specify the table from which we want to extract data.
V. Create a dataset for your Azure Data Lake Gen2 - Create a dataset for Azure Data Lake Gen2 that defines the destination data for the pipeline. We need to specify the path where we want to store the data in your data lake.
VI. Create a pipeline - Create a pipeline that defines the flow of data from the MSSQL database to Azure Data Lake Gen2. We need to specify the source and destination datasets.
VII. Trigger the pipeline - Once we have created the pipeline, we can trigger it to run either manually or on a schedule.

By following these steps, we can move data from the MSSQL database to Azure Data Lake Gen2 using Azure Data Factory. Once the data is in the data lake, we can perform various data processing and analysis tasks such as data cleaning, customer entity resolution, recommendation systems, and customer segmentation using Azure services such as Azure Databricks and Azure Synapse Analytics.

End-to-End analytics steps in Azure Data Lake Gen2

Steps we need to perform in Azure Data Lake Gen2 after data ingestion for building Customer Entity Resolution, Recommendation Systems, and Customer Segmentation -

I. Data Processing - Once data is ingested from MSSQL Database to Azure Data Lake Gen2, we need to perform data processing tasks such as data cleansing and transformation. This will be done using Azure Databricks.

II. Data storage - The processed data will then be stored in Azure Data Lake Gen2 in formats such as Parquet or CSV depending on our requirements.

III. Customer Entity Resolution - Next we need to create a dataset of customers by performing entity resolution. This involves identifying and merging customer records from multiple data sources and creating a single, accurate view of the customer. This will be done using Azure Databricks.

IV. Recommendation Systems - Using customer data, we need to perform recommendations to customers on products they might be interested in using collaborative filtering. This will be done using Azure Databricks.

V. Customer Segmentation - Next we need to perform K-Means clustering to segment customers into distinct groups based on their behavior, demographics, or any other relevant criteria. This will be done using Azure Databricks or Azure Machine Learning Studio.

VI. Data Visualization - We must present the insights in interactive visualization using the Power BI tool.

12. Data Cleaning

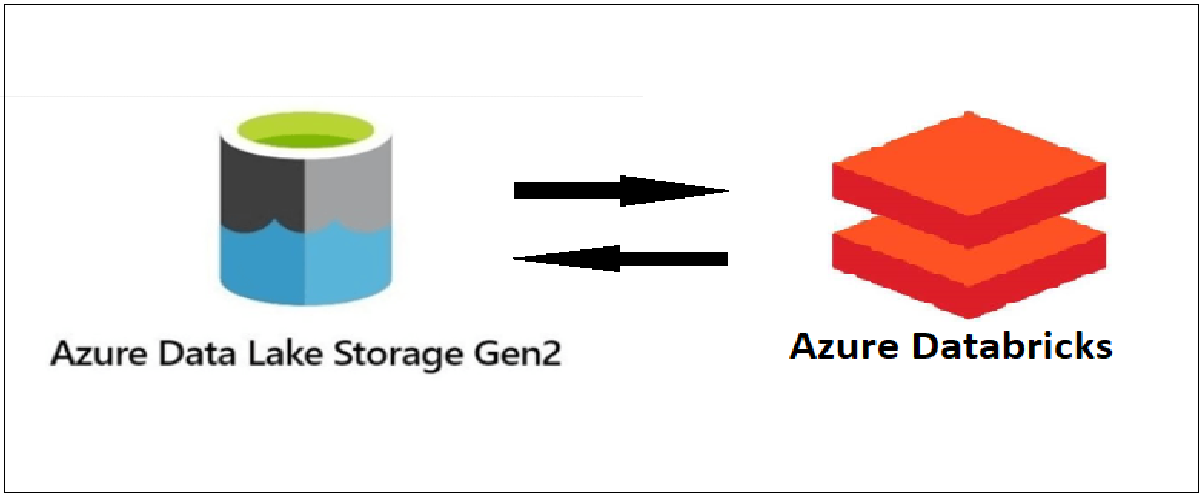

FIG 4. Connect, Ingest, and Transform Data

The data stored in Azure storage Data Lake Gen 2 will be fetched into Azure Databricks to perform the cleaning steps.

- Data cleaning is one of the unavoidable steps in any machine learning project. It identifies and fixes errors in the data to make it more useful to generate a model. Cleaning the data accurately gives us reliable outcomes which can then be used for business preparation.

We will initially do an Exploratory Data Analysis to understand the flaws in our data and then take steps to prepare the data for further processing, the data cleaning step is majorly divided into three steps:

- **Records**

We would remove any duplicate data present in our datasets. If the duplicate data are not removed, our machine learning models are going to be at risk of overfitting. The same parameters will be passed more than one time, but our models will learn nothing new. Duplicate entries can also ruin the train test split thus resulting in deficient performance estimates.

- **Outliers**

We will be removing outliers from our data set as part of cleaning the data. Outliers are the exception values in a dataset that is significantly distant or different from others. Outliers could be caused by errors in capturing, processing, or manipulating data. The presence of outliers in the dataset will result in higher error metrics which negatively affects the model performance.

- **Structure**

As part of the data cleaning process, we would be removing unwanted characters from our input. Unwanted characters do not add value to our machine-learning models. These include HTML tags, special characters, line breaks, and numbers from our data.

We would also make sure that all the data types are correct concerning the attribute values. If not, we will rectify the same to ensure correct processing. Incorrect data types may not allow us to process the data the way we want to and hence will result in inaccurate outcomes.

Missing or NULL values in the model inputs can result in an error while generating the model. To avoid that, we are going to analyze our data and handle them in one of two ways:

- We will drop the records containing NULL or missing values.
- We will impute the missing values with the mean or mode value of the attribute. Since during imputation, the same value will be assigned to all the missing values, it could distort the relationship between variables which means we have to be careful to decide which method to proceed with.

We will break down complex data values in simple forms to avoid violation of tidy data principles. This step will be done depending on the type of column we are having. Splitting the data might uncover vital information which adds value to our model performance and increase accuracy.

In addition to the above steps, we would also perform two more steps as below to make sure that our data is compliant with ‘Data Quality’ features.

I. Removing irrelevant information - We would be removing all that information from our input that is not related to or does not add value to our project. This way our models do not have to do extra work on processing that data while learning nothing useful from them.

II. Ensuring data completeness - A thorough analysis of our data by EDA would help us understand if our data is complete and it has all the required information needed to generate our machine learning model.

Once we finish cleaning and preparing the data, we will re-do the EDA to check if our data is uniform and if the cleaning process is successful or not. We will ensure that our input is in line with the six main checks for data quality.

I. Completeness - The data provides a complete understanding of the information we are trying to capture.
II. Consistency - The data is coordinated throughout the system in different units.
III. Conformity - The data is compliant with the standard definitions.
IV. Accuracy - The data is relevant to real-world applications.
V. Integrity - The data is valid and can be traced back across systems.
VI. Timeliness - The data is available when required and expected.

13. Algorithm

- **Collaborative Filtering**

After performing customer entity resolution on the cleaned and merged data, we need to build a recommendation engine using Collaborative Filtering.
Collaborative Filtering is a technique used in recommender systems to predict user preferences and generate recommendations based on the preferences of similar users or items. There are two main types of collaborative filtering - user-based and item-based.

- User-based collaborative filtering predicts what a user will like based on the preferences of users who like them.
- Item-based collaborative filtering, on the other hand, predicts what a user will like based on the similarity of items they have liked in the past.

A high-level analytical approach to perform user-based and item-based collaborative filtering -

I. Data Preprocessing - We need to perform data preprocessing which involves cleaning and transforming data to make it ready for modeling. This may include handling missing values, normalizing data, and removing duplicates.
II. Data Exploration - Next, we need to explore data in understanding the characteristics and identifying patterns and relationships. This may involve analyzing the distribution of data, identifying correlations, and creating visualizations.
III. Data Splitting - We need to split the data into training and testing sets for model validation. This may include 80% of the data for training and 20% for testing.
IV. Model Development - Next, we need to develop a collaborative filtering model using the training data. We will develop both user-based and item-based collaborative filtering.

V. Model Evaluation - We need to perform a model evaluation to measure the accuracy of the models using metrics such as mean squared error (MSE) or root mean squared error (RMSE).

- **Customer Segmentation**

On the merged data we will also perform customer segmentation using the K-Means clustering technique.

Customer segmentation is the process of dividing customers into smaller groups based on similar characteristics such as demographics, behavior, and preferences. This enables businesses to tailor marketing strategies and product offerings to better meet the needs of each segment and improve customer engagement and loyalty.

A high-level analytical approach to perform customer segmentation -

I. Data Preprocessing - We need to perform data preprocessing which involves cleaning and transforming data to make it ready for modeling. This may include handling missing values, normalizing data, and removing duplicates.
II. Feature selection - Next, we need to identify the features that are most important for customer segmentation. This can be done using feature importance.
III. Clustering - Once the features have been identified, we need to use the K-Means clustering algorithm to group customers with similar characteristics together.
IV. Interpretation - Finally, we need to interpret the results of the segmentation analysis to understand the different customer segments.

14. Challenges

I. **Data Source** - Obtaining a complete and reliable customer 360 view dataset can be challenging. This may involve integrating data from various sources, dealing with data quality issues, and ensuring data accuracy and relevance. Incomplete or inconsistent data may impact the effectiveness of customer segmentation, recommendation, and personalization efforts.

II. **Cold-start Problem** - When a new user is introduced to the system, there may be insufficient data available to generate accurate recommendations. Without historical data on user preferences or behaviors, it can be challenging to provide personalized recommendations. New users may receive generic or less relevant recommendations, which can impact their initial experience and engagement with the system.

III. **Customer Entity Id** - Managing data from multiple silos can pose challenges, including the issue of duplicate customer records and the need to use fuzzy mapping techniques for entity resolution. Duplicate customer records can lead to inaccurate segmentation, recommendations, and personalization efforts. Fuzzy mapping techniques may be required to identify and merge duplicate records, which can be complex and time-consuming.

IV. **Evaluation Metrics** - Selecting appropriate evaluation metrics to effectively measure the performance of the recommendation system can be a challenging task. There are various metrics available, such as precision, recall, F1-score, and conversion rate, which may have different interpretations and implications. Choosing the right metrics that align with business goals and accurately reflect the performance of the recommendation system can be crucial for evaluating its effectiveness and making data-driven improvements.

15. Conclusion

In conclusion, the proposed project that involves using customer entity resolution, recommendation systems, and customer segmentation can greatly benefit businesses in several ways. By leveraging customer entity resolution, businesses can ensure accurate and complete customer data, which can improve decision-making, customer experience, and operational efficiency. Recommendation systems can provide personalized product recommendations to customers, which can enhance customer satisfaction and loyalty, increase sales, and improve the bottom line. Customer segmentation can help businesses identify and target specific customer groups with relevant messaging and offers, leading to more effective marketing campaigns and customer retention.

In combination, these three technologies can offer businesses a powerful toolset to gain insights into customer behavior, preferences, and needs, enabling them to provide more personalized, relevant, and valuable experiences. As a result, businesses can enhance customer satisfaction, improve retention, and increase revenue. The proposed project can be a significant investment for any business looking to enhance its customer experience and achieve sustainable growth in today's competitive market.