# Customer 360 V2 (Technical Documentation)

**Introduction**

**Overview of the Data Product**

The Customer 360 Data Product represents a transformative solution designed to empower organizations with a holistic and unified view of their customers. By consolidating data from various touchpoints and sources, this product aims to provide deep insights into customer behaviour, preferences, and interactions. The comprehensive customer profiles generated enable businesses to enhance customer experiences, optimize marketing strategies, and make informed, data-driven decisions.

**Purpose and Objectives**

The primary purpose of the Customer 360 Data Product is to equip organizations with a powerful tool that facilitates a nuanced understanding of their customer base. By aggregating and analysing diverse datasets, the product aims to achieve the following objectives:

Enhance Customer Engagement: Enable personalized and targeted customer interactions by leveraging insights derived from a 360-degree view of customer profiles.

Optimize Marketing Strategies: Facilitate data-driven marketing campaigns, segmentation, and targeting, resulting in improved campaign effectiveness and customer conversion rates.

Drive Operational Efficiency: Support decision-making processes by providing actionable insights to various business functions, ultimately improving operational efficiency and resource allocation.

Increase Customer Retention and Loyalty: Ide ntify at-risk customers, implement proactive retention strategies, and enhance overall customer satisfaction, fostering loyalty and repeat business.

**Target Audience**

The Customer 360 Data Product is tailored to meet the needs of diverse professionals involved in data analysis, engineering, and science. The primary target audience includes:

Business Analysts: Individuals responsible for interpreting and translating business requirements into data-driven insights. Business analysts will leverage the product to derive actionable recommendations for enhancing customer experiences and driving business outcomes.

Data Engineers: Professionals tasked with designing, implementing, and maintaining the data infrastructure. Data engineers will play a crucial role in ensuring seamless data integration, reliability, and real-time updates within the Customer 360 Data Product.

Data Scientists: Experts in statistical analysis, machine learning, and predictive modeling who will utilize the product to uncover hidden patterns, forecast customer behaviors, and contribute to the development of advanced analytics solutions.

Document Version and Change History

Document Version: 1.0

Last Updated: 12-12-2023

**Product Description**

**High-Level Description of the Data Product**

The Customer 360 Data Product is a cutting-edge solution designed to provide organizations with a unified and dynamic view of their customers. At its core, the product consolidates data from diverse sources, including CRM systems, sales transactions, customer support interactions, and online engagements, to create a comprehensive 360-degree view of each customer. This high-level description encapsulates the essence of a transformative tool that enables businesses to gain profound insights into customer behavior, preferences, and interactions.

**Key Features and Capabilities**

1. Data Integration:

Seamless integration of data from multiple sources, ensuring a unified and enriched customer profile.

2. 360-Degree Customer Profile:

Comprehensive customer profiles encompassing demographic information, purchase history, communication preferences, and more.

3. Real-time Updates:

Continuous and real-time updates to keep customer profiles current, facilitating prompt responses to changing customer behaviors.

4. Advanced Analytics and Reporting:

Powerful analytics tools and customizable reporting dashboards for deriving actionable insights from customer data.

5. Segmentation and Targeting:

Tools for segmenting customers based on various criteria, enabling personalized marketing and communication strategies.

6. Predictive Analytics:

Utilization of predictive modeling and machine learning algorithms to forecast customer behaviors and recommend personalized offerings.

7. Security and Compliance:

Stringent security measures and data governance protocols to ensure the accuracy, reliability, and privacy of customer information.

**Business Context and Application**

The Customer 360 Data Product finds application across diverse business contexts, including:

Marketing Optimization: Tailoring marketing campaigns based on customer insights to enhance targeting and campaign effectiveness.

Customer Retention: Implementing proactive retention strategies by identifying at-risk customers and addressing their needs.

Operational Efficiency: Supporting decision-making processes and resource allocation through data-driven insights.

Personalized Customer Experiences: Enabling personalized interactions and experiences by leveraging a deep understanding of individual customer preferences.

**Expected Outcomes and Benefits**

Implementing the Customer 360 Data Product is expected to yield the following outcomes and benefits:

Improved Customer Satisfaction: Enhanced customer experiences through personalized interactions and targeted communication.

Increased Revenue: Identification of cross-selling and upselling opportunities, leading to increased revenue generation.

Reduced Churn: Proactive identification and addressing of at-risk customers, resulting in reduced churn rates.

Data-Driven Decision-Making: Empowering organizations with data-driven insights for strategic decision-making.

Enhanced Marketing ROI: Optimizing marketing strategies and campaigns for improved return on investment.

This section provides a detailed overview of the Customer 360 Data Product, highlighting its key features, capabilities, business applications, and the expected outcomes and benefits for organizations leveraging this transformative solution.

**Technical Specifications**

**Customer 360 Lens**

**Recency Frequency and Monitoring (RMF) Analysis:**

The integration of RFM (Recency, Frequency, Monetary Value) analysis within Customer 360 (C360) lenses has empowered businesses to deeply understand customer behaviour and tailor marketing strategies effectively. By analyzing recent interactions (Recency), frequency of engagement (Frequency), and monetary contributions (Monetary Value), C360 segments customers into groups. This segmentation helps identify high-value, loyal customers for personalized engagement strategies. It also highlights dormant segments requiring re-engagement tactics to prevent potential churn. Utilizing RFM within C360 lenses elevates targeted marketing, enhances customer experiences, and fosters long-term loyalty, driving revenue growth and informed decision-making.

**Recency:** Recency in RFM analysis within C360 lenses illuminates recent customer engagements, indicating their current activity levels and potential ongoing interactions. This metric enables businesses to identify active customers more likely to respond to immediate marketing strategies or personalized offers. Leveraging recency insights allows for targeted engagement, potentially boosting immediate sales or eliciting prompt customer responses, thus elevating overall customer satisfaction and engagement levels.

**Frequency:** Frequency analysis in C360 lenses delves into customer engagement consistency, signifying customer loyalty and habitual interactions. This aspect identifies highly engaged customers, forming the foundation for strategies aimed at retaining loyalty and encouraging repeat interactions. Segmentation and understanding of high-frequency customers empower businesses to design targeted marketing efforts and loyalty programs tailored to maintain and potentially increase these customers' lifetime value, ensuring sustained revenue streams.

**Monitoring:** within C360 lenses' Monetary Value assessment, involves continual tracking of customer transactions and their impact on overall revenue. This process identifies high-value customers whose spending significantly influences business success. Through ongoing monitoring, tailored strategies like exclusive offers or personalized services are applied to retain and increase spending from these customers. Prioritizing enhanced experiences for such clientele maximizes their contribution to the company's financial success.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column Name** | | **Description** | | **Type** | | | | **Dimension** |
| customer\_id | | unique identifier of the customer | | DIMENSION | | | | Customer |
| first\_name | | Business first name the customer operates under | | DIMENSION | | | | Customer |
| last\_name | | Business last name the customer operates under | | DIMENSION | | | | Customer |
| birth\_date | | Birth date of the customer | | DIMENSION | | | | Customer |
| gender | | gender of the customer | | DIMENSION | | | | Customer |
| marital\_status | | marital status of the customer | | DIMENSION | | | | Customer |
| hobbies | | hobbies of the customer | | DIMENSION | | | | Customer |
| annual\_income | | annual income of the customer | | METRIC | | | | Customer |
| email\_id | | email address of the customer | | DIMENSION | | | | Customer |
| phone\_number | | contact number of the customer | | DIMENSION | | | | Customer |
| mailing\_street | | postal address of the customer | | DIMENSION | | | | Customer |
| country | | name of the county | | DIMENSION | | | | Customer |
| state | | state code where the customer physical address is located | | DIMENSION | | | | Customer |
| city | | name of the city | | DIMENSION | | | | Customer |
| zip\_code | | ZIP code associate with the customer physical address | | DIMENSION | | | | Customer |
| register\_date | | registered date of customer | | DIMENSION | | | | Customer |
| ${customer.customer\_id} | | count of total customers | | METRIC | | | | Customer |
| session\_id | | session identifier of the activity event | | DIMENSION | | | | **Product views** |
| activity\_uuid | | unique identifier of the activity event | | DIMENSION | | | | **Product views** |
| visit\_start\_ts | | timestamp of the moment when visit occured | | DIMENSION | | | | **Product views** |
| entity\_id | | customer identifier | | DIMENSION | | | | **Product views** |
| activity\_year | | year of the moment when activity occured | | DIMENSION | | | | **Product views** |
| activity\_ts | | timestamp of the moment when activity occured | | DIMENSION | | | | **Product views** |
| product\_sku | | product identifier | | DIMENSION | | | | **Product views** |
| activity\_month | | month of the moment when activity occured | | DIMENSION | | | | **Product views** |
| traffic\_channel | | user acquisition channel name | | DIMENSION | | | | **Product views** |
| event\_action | | event action name for activity | | DIMENSION | | | | **Product views** |
| occurence | | number of occurence for activity | | METRIC | | | | **Product views** |
| traffic\_source | | user acquisition source name | | DIMENSION | | | | **Product views** |
| ${product\_views.session\_id} | | product view session count | | METRIC | | | | **Product views** |
| activity\_repeated\_at | | recurring activity timestamp | | DIMENSION | | | | **Product views** |
| ${product\_views.product\_sku} | | total number of product that viewed | | METRIC | | | | **Product views** |
| ROUND(DATE\_DIFF('second', MIN(${product\_views.visit\_start\_ts}), MAX (${product\_views.activity\_ts}))/cast(60 as double),2) | | session engagement time in minute | | METRIC | | | | **Product views** |
| day(current\_date - date(${product\_views.activity\_ts})) | | number of days of activity from current date | | METRIC | | | | **Product views** |
| day(current\_date - date(${product\_views.activity\_ts})) | | days since last product was viewed | | METRIC | | | | **Product views** |
| activity\_uuid | | unique identifier of the activity event | | | DIMENSION | | | Items added to favourite |
| session\_id | | session identifier of the activity event | | | DIMENSION | | | Items added to favourite |
| entity\_id | | customer identifier | | | DIMENSION | | | Items added to favourite |
| activity\_ts | | timestamp of the moment when activity\_occured | | | DIMENSION | | | Items added to favourite |
| product\_sku | | product identifier | | | DIMENSION | | | Items added to favourite |
| product\_name | | product name | | | DIMENSION | | | Items added to favourite |
| category\_name | | product's category name | | | DIMENSION | | | Items added to favourite |
| subcategory\_name | | product's subcategory name | | | DIMENSION | | | Items added to favourite |
| brand\_name | | product's brand name | | | DIMENSION | | | Items added to favourite |
| parent\_company | | brand's parent company name | | | DIMENSION | | | Items added to favourite |
| product\_price | | price of the product | | | METRIC | | | Items added to favourite |
| quantity | | quantity of the order was placed for | | | METRIC | | | Items added to favourite |
| order\_value | | sales of order was placed for | | | METRIC | | | Items added to favourite |
| event\_action | | event action name for activity | | | DIMENSION | | | Items added to favourite |
| traffic\_channel | | user acquisition channel name | | | DIMENSION | | | Items added to favourite |
| traffic\_source | | user acquisition source name | | | DIMENSION | | | Items added to favourite |
| occurence | | number of occurence for activity | | | METRIC | | | Items added to favourite |
| activity\_repeated\_at | | recurring activity timestamp | | | DIMENSION | | | Items added to favourite |
| ${items\_added\_to\_favorite.session\_id} | | count of unique sessions | | | METRIC | | | Items added to favourite |
| ${items\_added\_to\_favorite.product\_sku} | | total number of items was favourited | | | METRIC | | | Items added to favourite |
| day(current\_date - date(${items\_added\_to\_favorite.activity\_ts})) | | number of days of activity from current date | | | METRIC | | | Items added to favourite |
| **Column Name** | **Description** | | | **Type** | | | | **Dimension** |
| activity\_uuid | unique identifier of the activity event | | | DIMENSION | | | | **Items added to cart** |
| session\_id | session identifier of the activity event | | | DIMENSION | | | | **Items added to cart** |
| entity\_id | customer identifier | | | DIMENSION | | | | **Items added to cart** |
| activity\_ts | timestamp of the moment when activity\_occured | | | DIMENSION | | | | **Items added to cart** |
| product\_sku | product identifier | | | DIMENSION | | | | **Items added to cart** |
| product\_name | product name | | | DIMENSION | | | | **Items added to cart** |
| category\_name | product's category name | | | DIMENSION | | | | **Items added to cart** |
| subcategory\_name | product's subcategory name | | | DIMENSION | | | | **Items added to cart** |
| brand\_name | product's brand name | | | DIMENSION | | | | **Items added to cart** |
| parent\_company | brand's parent company name | | | DIMENSION | | | | **Items added to cart** |
| product\_price | price of the product | | | METRIC | | | | **Items added to cart** |
| quantity | quantity of the order was placed for | | | METRIC | | | | **Items added to cart** |
| order\_value | sales of order was placed for | | | METRIC | | | | **Items added to cart** |
| event\_action | event action name for activity | | | DIMENSION | | | | **Items added to cart** |
| traffic\_channel | user acquisition channel name | | | DIMENSION | | | | **Items added to cart** |
| traffic\_source | user acquisition source name | | | DIMENSION | | | | **Items added to cart** |
| occurence | number of occurence for activity | | | METRIC | | | | **Items added to cart** |
| activity\_repeated\_at | recurring activity timestamp | | | DIMENSION | | | | **Items added to cart** |
| ${items\_added\_to\_cart.session\_id} | count of unique sessions | | | METRIC | | | | **Items added to cart** |
| ${items\_added\_to\_cart.product\_sku} | total number of items was added to cart | | | METRIC | | | | **Items added to cart** |
| ${items\_added\_to\_cart.quantity} | total quantity of items was added to cart | | | METRIC | | | | **Items added to cart** |
| day(current\_date - date(${items\_added\_to\_cart.activity\_ts})) | number of days of activity from current date | | | METRIC | | | | **Items added to cart** |
| activity\_uuid | | unique identifier of the activity event | | | | DIMENSION | | **Order placed** |
| session\_id | | session identifier of the activity event | | | | DIMENSION | | **Order placed** |
| entity\_id | | customer identifier | | | | DIMENSION | | **Order placed** |
| activity\_ts | | timestamp of the moment when activity\_occured | | | | DIMENSION | | **Order placed** |
| activity\_year | | year of the moment when activity\_occured | | | | DIMENSION | | **Order placed** |
| activity\_month | | month of the moment when activity\_occured | | | | DIMENSION | | **Order placed** |
| transaction\_id | | transaction identifier | | | | DIMENSION | | **Order placed** |
| product\_sku | | product identifier | | | | DIMENSION | | **Order placed** |
| product\_name | | product name | | | | DIMENSION | | **Order placed** |
| category\_name | | product's category name | | | | DIMENSION | | **Order placed** |
| subcategory\_name | | product's subcategory name | | | | DIMENSION | | **Order placed** |
| brand\_name | | product's brand name | | | | DIMENSION | | **Order placed** |
| parent\_company | | brand's parent company name | | | | DIMENSION | | **Order placed** |
| product\_price | | price of the product | | | | METRIC | | **Order placed** |
| quantity | | quantity of the order was placed for | | | | METRIC | | **Order placed** |
| order\_value | | sales of order was placed for | | | | METRIC | | **Order placed** |
| event\_action | | event action name for activity | | | | DIMENSION | | **Order placed** |
| traffic\_channel | | user acquisition channel name | | | | DIMENSION | | **Order placed** |
| traffic\_source | | user acquisition source name | | | | DIMENSION | | **Order placed** |
| occurence | | number of occurence for activity | | | | METRIC | | **Order placed** |
| activity\_repeated\_at | | recurring activity timestamp | | | | DIMENSION | | **Order placed** |
| ${order\_placed.session\_id} | | count of unique sessions | | | | METRIC | | **Order placed** |
| day(current\_date - date(${order\_placed.activity\_ts})) | | days since last order was placed | | | | METRIC | | **Order placed** |
| cast(count(distinct(${order\_placed.transaction\_id})) as decimal(20,0 )) | | count of total activities. | | | | METRIC | | **Order placed** |
| cast(count(distinct(${order\_placed.transaction\_id})) FILTER (WHERE ${order\_placed.duration\_in\_days} between {{start\_day:0}} and {{end\_day:365}}) as decimal(20,0)) | | count of total activities. | | | | METRIC | | **Order placed** |
| round(sum(${order\_placed.order\_value}),2) | | total order value | | | | METRIC | | **Order placed** |
| ${order\_placed.transaction\_id} | | unique order\_no for a customer | | | | DIMENSION | | **Order placed** |
| day(current\_date - date(${order\_placed.activity\_ts})) | | number of days of activity from current date | | | | METRIC | | **Order placed** |
| segment\_name | name of the audience | | DIMENSION | | | | **Segment** | |
| customer\_id | unique identifier of the customer | | DIMENSION | | | | **Segment** | |
| ${segment.customer\_id} | count of customers categorized into segments | | METRIC | | | | **Segment** | |

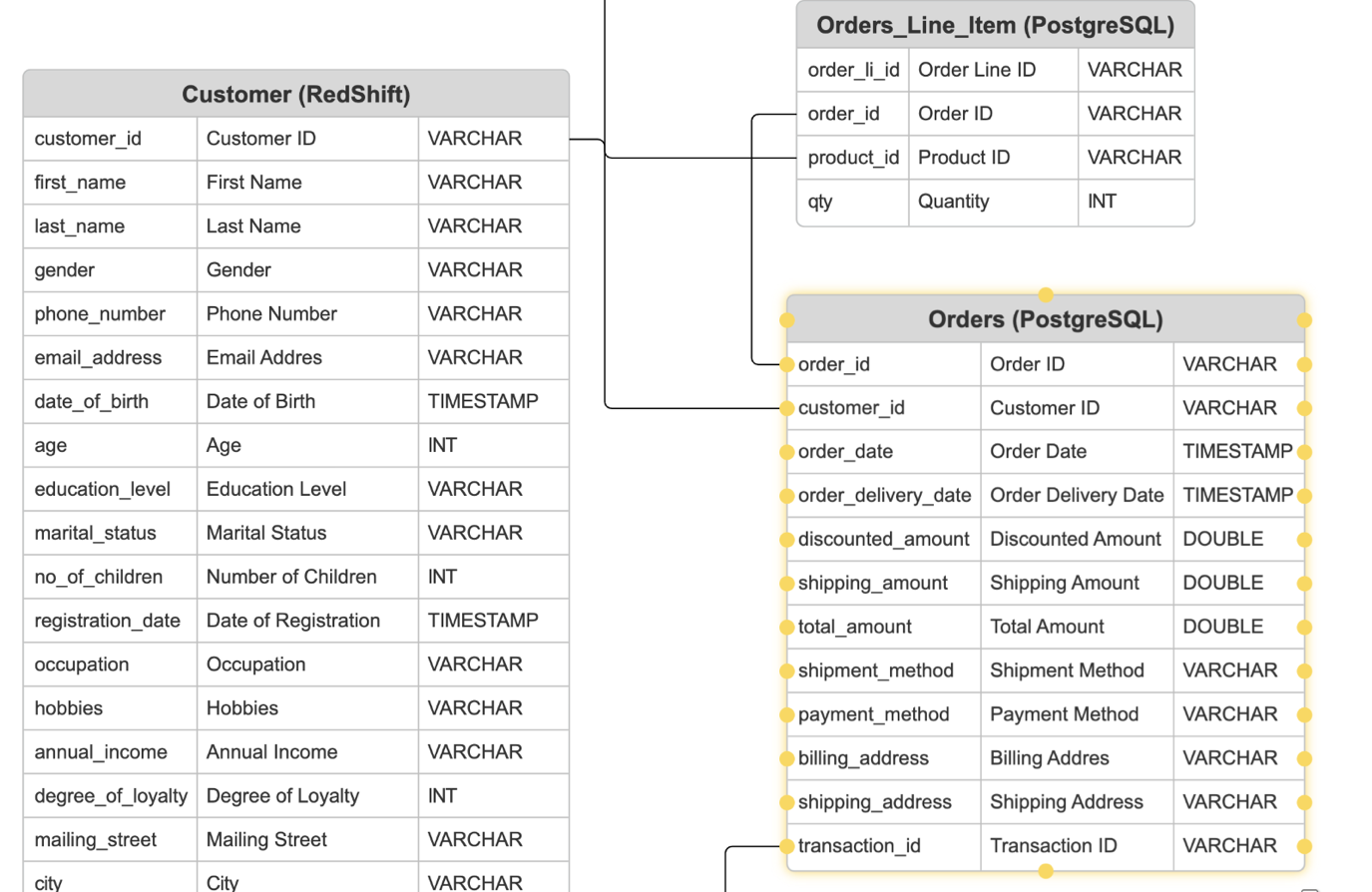
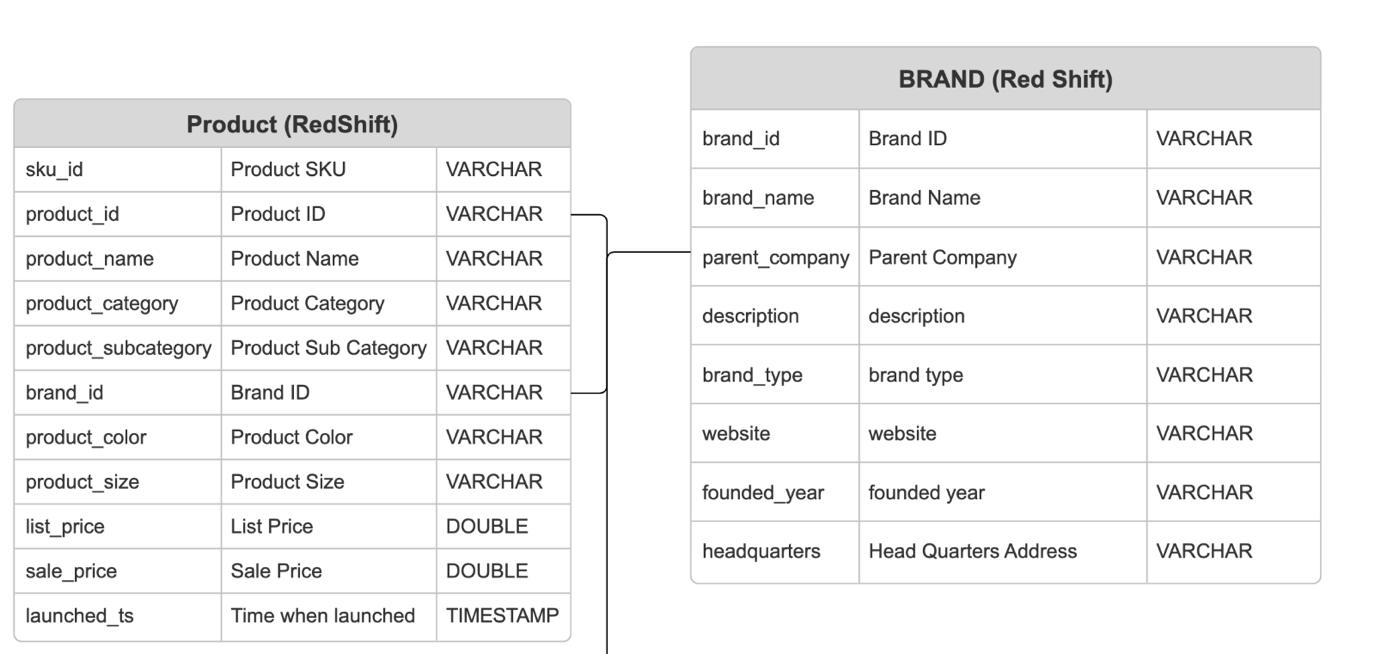
**Data Sources and Formats**

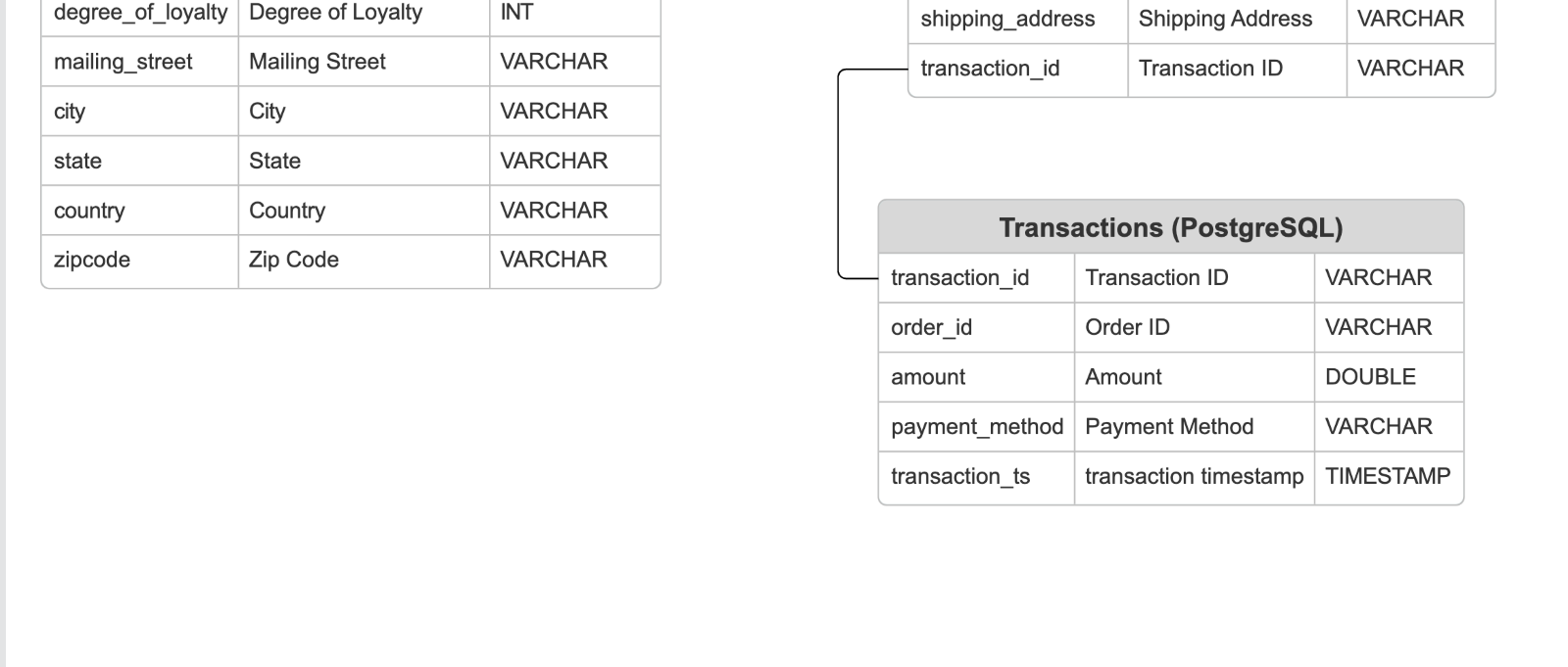
**Datasets**

Customer

1. Product
2. Orders
3. Transactions
4. Order line item
5. Web Analytics
6. Brands

Entity Relationship Diagram





**Data Sources**

1. AWS Redshift
2. PostgreSQL
3. Web Analytics (Google)

**Data Volume and Update Frequency** (Balaji)

The volume of data processed by the Customer 360 Data Product depends on the scale and scope of the organization. Factors influencing data volume include the customer base size, transaction frequency, and engagement levels. Update frequencies vary based on the source but are optimized for real-time updates to ensure the most current customer profiles.

**System Requirements and Dependencies**

Compute Cluster: Apache Spark

Number of Nodes: 3 nodes

Node Type: High-memory instances (e.g., r5.4xlarge) with at least 16 CPU cores and 64 GB RAM per node.

Total CPU Cores: 48 cores

Total RAM: 192 GB

Apache Kafka Cluster

Number of Brokers: 3 brokers (for redundancy and fault tolerance)

Broker Instance Type: Instance type with at least 8 CPU cores and 32 GB RAM per broker.

Total CPU Cores: 24 cores

Total RAM: 96 GB

PostgreSQL Database

Instance Type: High-memory instance (e.g., r5.xlarge) with at least 4 CPU cores and 16 GB RAM.

Storage: Enough storage capacity to handle the "orders\_enriched\_data" dataset.

**Data Flow and Architecture Diagram**

**Security and Compliance Considerations**

Security Measures

Encryption: Data encryption in transit and at rest to ensure the confidentiality and integrity of customer information.

Access Control: Role-based access control mechanisms to restrict data access based on user roles.

Audit Trails: Logging and monitoring features to track user activities and changes to customer profiles.

Compliance Considerations

Data Privacy Regulations: Adherence to data protection regulations such as GDPR, HIPAA, or industry-specific compliance standards.

Data Retention Policies: Implementation of policies for the secure storage and responsible disposal of customer data.

**Built Using**

Python: For data processing, transformation, and integration tasks.

Apache Kafka: To enable real-time streaming of customer order events.

Apache Spark: For handling large-scale data processing efficiently.

PostgreSQL: As the primary database to store the enriched order data.

**Data Models and Algorithms**

**Description of Data Models Used**

The Customer 360 Data Product employs a sophisticated blend of data models to construct a comprehensive and unified representation of customer profiles:

1. Entity-Relationship Model:

Illustrates the relationships and connections between entities, such as customers, transactions, and interactions.

2. Graph Model:

Captures the intricate connections and dependencies between various elements, providing a visual map of customer relationships and network structures.

3. Predictive Models:

Harnesses machine learning algorithms, including regression analysis, decision trees, and neural networks, to anticipate customer behaviours and uncover patterns.

4. Segmentation Models:

Utilizes algorithms like k-means clustering to categorize customers based on shared characteristics, enabling targeted marketing and personalized engagement.

**Algorithms and Analytical Methods**

1. Predictive Analytics:

Applies advanced algorithms to predict customer behaviours, preferences, and trends, enabling proactive decision-making.

1. Predictive Sales
2. Predictive Churn

2. Clustering Algorithms:

Utilizes k-means clustering and other algorithms to identify patterns and group customers based on similarities, facilitating tailored strategies.

3. Recommendation Algorithms:

Implements collaborative filtering and content-based recommendation algorithms to suggest personalized products and services.

1. Market Basket Analysis Based product recommendation
2. Marketing Channel Optimisation and recommendation

4. Descriptive Analytics:

Leverages statistical methods and analytics tools to derive meaningful insights from historical customer data, aiding in comprehensive understanding.

**Assumptions and Limitations**

Assumptions

Quality of Input Data: Assumes that the accuracy and reliability of models depend on the quality and completeness of the input data.

Predictability: Assumes a degree of predictability in customer behavior based on historical patterns.

Limitations

Probabilistic Nature: Acknowledges that predictive models are inherently probabilistic and may not always accurately forecast customer behaviors.

Data Constraints: Recognizes that data models may not capture all nuances of customer interactions or external factors due to data limitations.

**Model Validation and Performance Metrics**

Model Validation

Cross-Validation: Utilizes techniques like k-fold cross-validation to assess model performance across different subsets of the dataset.

Holdout Validation: Reserves a portion of the dataset for validation to evaluate the model's ability to generalize to new data.

Performance Metrics

Accuracy: Measures the overall correctness of predictions.

Precision and Recall: Evaluates the model's ability to correctly identify positive instances and avoid false positives.

F1 Score: Strikes a balance between precision and recall, providing a holistic performance metric.

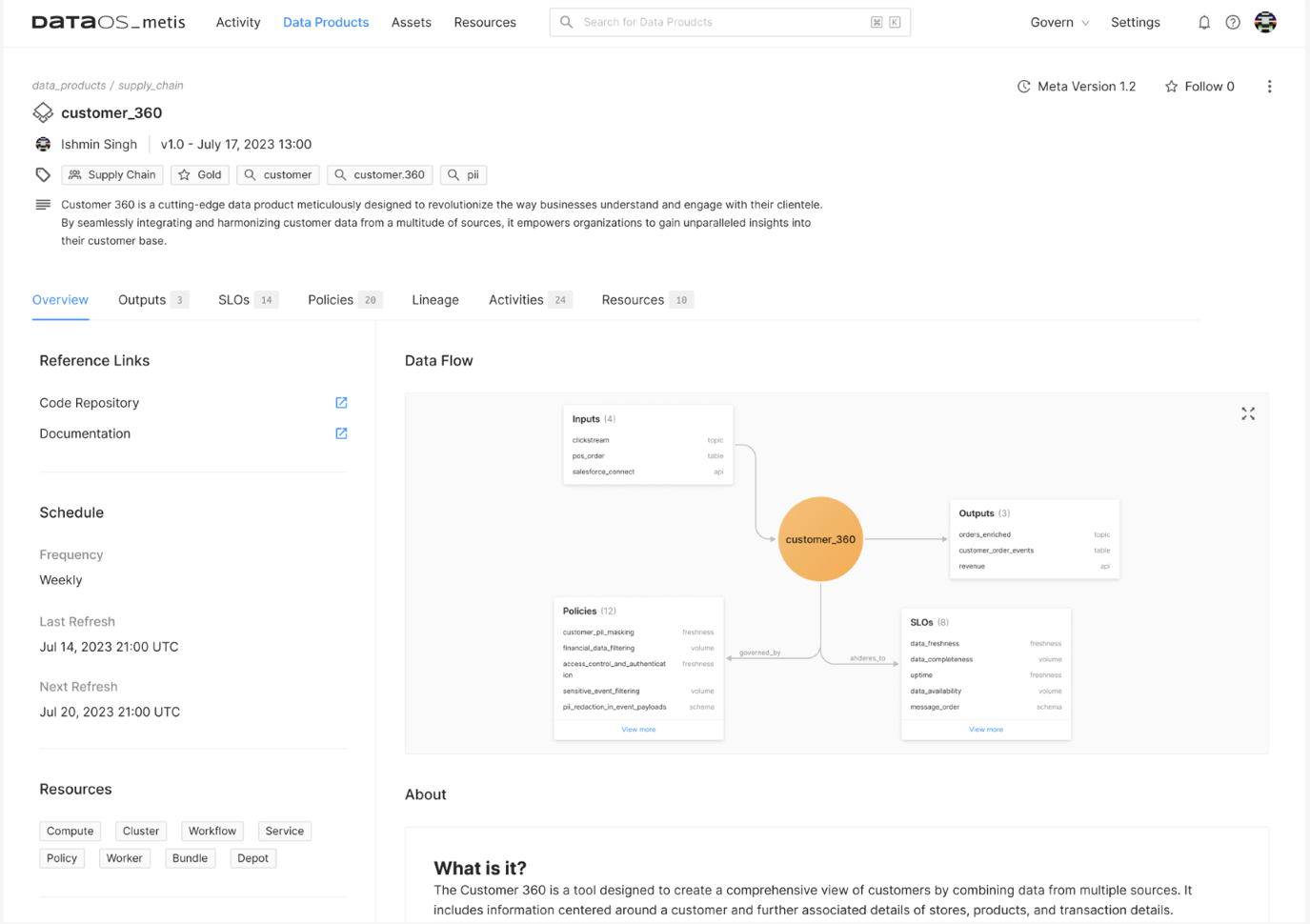
Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC): Assesses the trade-off between sensitivity and specificity, offering insights into the model's discriminatory power.

This section unveils the intricacies of the data models and algorithms powering the Customer 360 Data Product, shedding light on the methods employed to create a unified customer perspective, predict behaviours, and segment

**User Guide**

**Accessing the Data Product**

The data product can be accessed from the Data Product hub on the corresponding DataOS instance.



**Advanced Features (for Data Scientists)**

Data Scientists can enjoy majority of their EDA through METIS and Odin which has all the major Univariate and Multivariate analysis in place.

**Troubleshooting and Support**

If they need any support on the data product, they can reach out to Balaji at [balaji.ext@tmdc.io](mailto:balaji.ext@tmdc.io), Aravind at [aravind@tmdc.io](mailto:aravind@tmdc.io)

**Data Quality and Governance**

**Data Quality Measures and Standards**

Data quality in the Customer 360 Data Product will be rigorously measured through a multi-faceted approach encompassing various dimensions. Firstly, accuracy checks will be conducted to ensure that the information within customer profiles aligns with the actual interactions and transactions recorded across different sources. Completeness assessments will verify the presence of essential data points, promoting a holistic customer view. Timeliness will be monitored to ensure that customer profiles reflect real-time updates, allowing for timely and relevant insights. Consistency checks will be employed to identify any discrepancies or anomalies in data patterns, while uniqueness measures will ensure the absence of duplicate records. Additionally, data quality audits will be conducted regularly to address any discrepancies, and feedback loops will be established to continuously enhance data accuracy and reliability. Through these comprehensive measures, the Customer 360 Data Product aims to maintain a high standard of data quality, fostering trust in the information provided for strategic decision-making and operational efficiency.

**Data Cleansing and Preprocessing Steps**

Data cleaning and preprocessing in the Customer 360 Data Product involve a meticulous process to ensure the quality and integrity of the information. Initially, raw data from diverse sources undergoes cleansing procedures, addressing missing values, outliers, and inconsistencies. Standardization and normalization techniques are applied to ensure uniformity, while duplicate records are identified and eliminated. Data preprocessing involves transforming and aggregating information to create a cohesive dataset. Advanced algorithms are employed to handle complex tasks such as imputing missing values and encoding categorical variables. Through this thorough cleaning and preprocessing pipeline, the Customer 360 Data Product guarantees a reliable foundation for robust analytics and actionable insights.

**Data Governance Policies and Data Stewardship**

There are two types of Data Governance policies:

* **Role-Based Access Control (RBAC):**
  + Users see only corresponding rows and datasets based on their access role.
* **Discretionary Access Control (DAC):**
  + Currently, DAC (full access) is granted to:

|  |
| --- |
| **Users** |
| Data Stewards |
| Data Producers (of each Data Product) |

For RBAC:

* Certain Personally Identifiable Information (PII) is masked for each category.

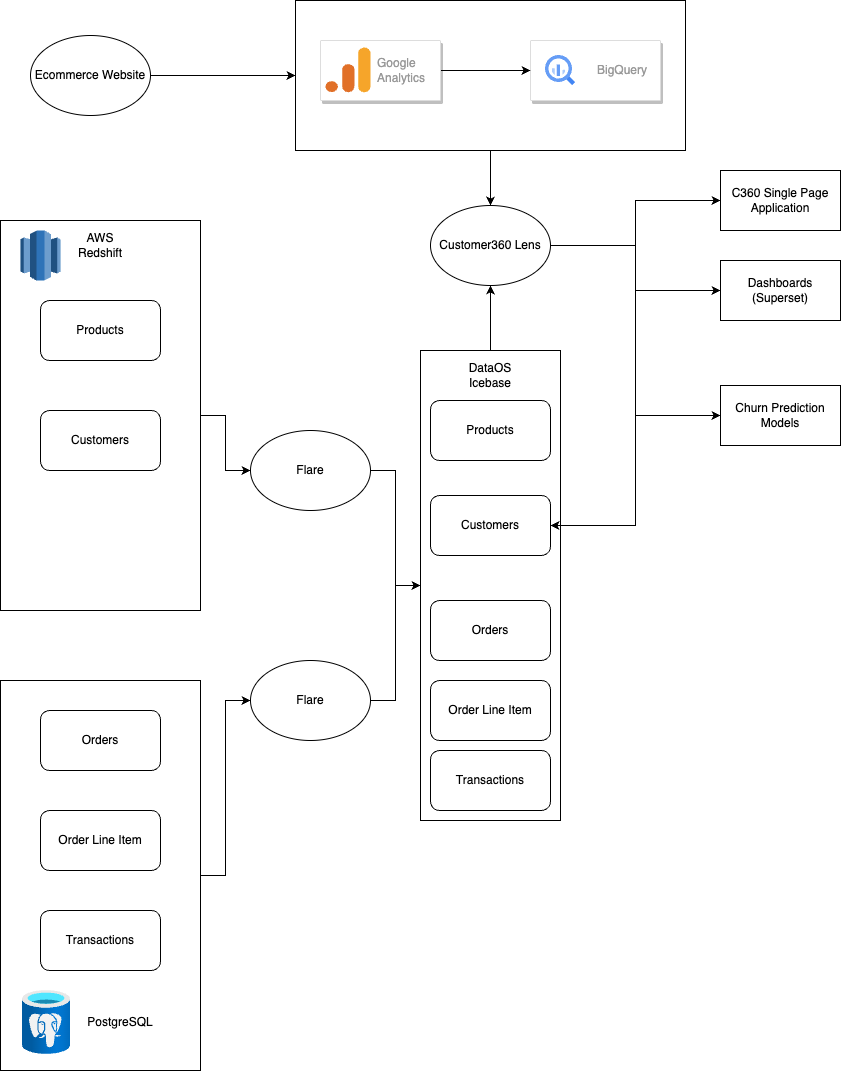
|  |  |
| --- | --- |
| **Category** | **Masked Attributes** |
| Marketing Team | Age (Range), Salary (Range) |
| Data Consumer | All PII information masked through SHA256 |

**Data Steward Responsibilities:**

* Clearly define roles and responsibilities for data ownership.
* Appoint data stewards responsible for overseeing data quality and integrity.

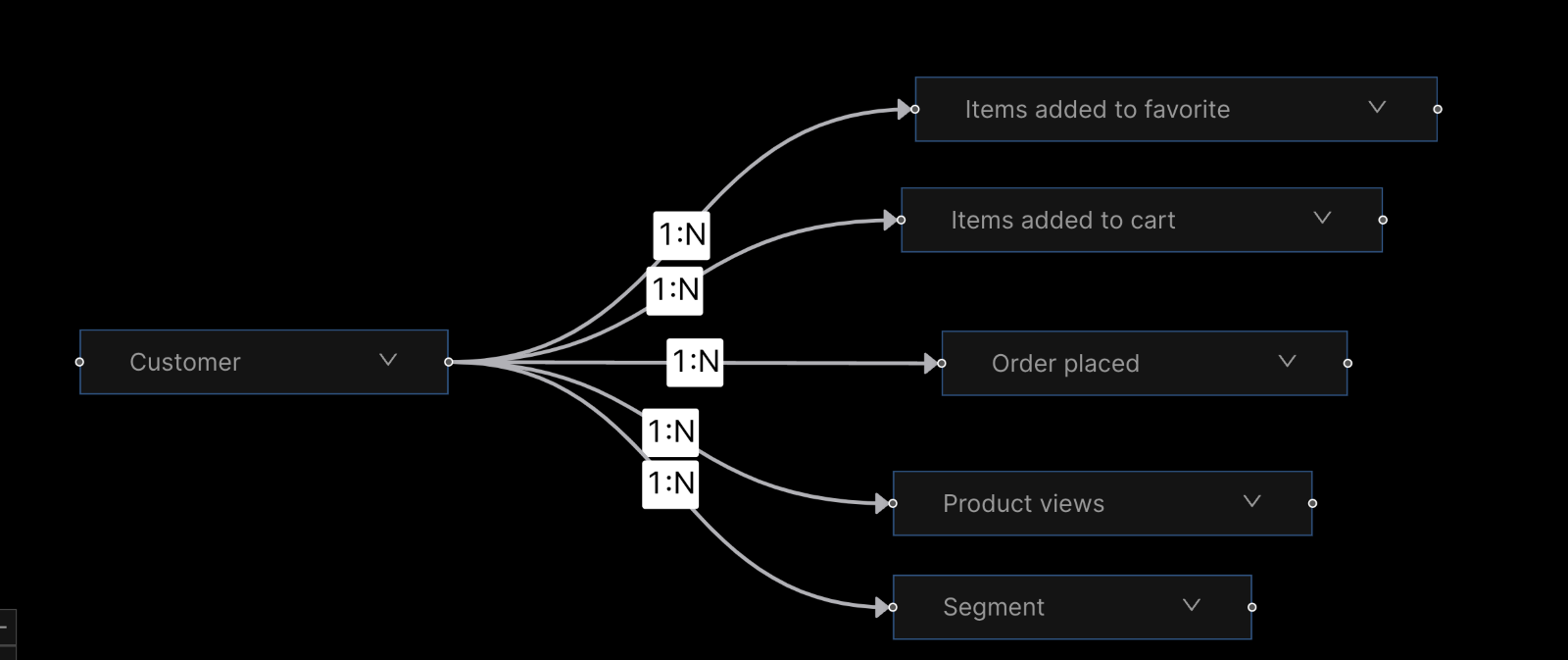
**Data Lineage and History Tracking**

In DataOS, the data lineage of the Custome360 product can be accessed through the Hera application within DataOS. The change history and history tracking is possible through METIS.



**Customer 360 Lens**

The customer 360 Lens is built by connecting the entities mentioned above as shown in the diagram below,



The definition of the lens can be found in the following URL

<https://github.com/StacknexusDatanators/retail-accelerator/blob/main/dataos_yamls/lens/c360_solution_accelerator.yaml>

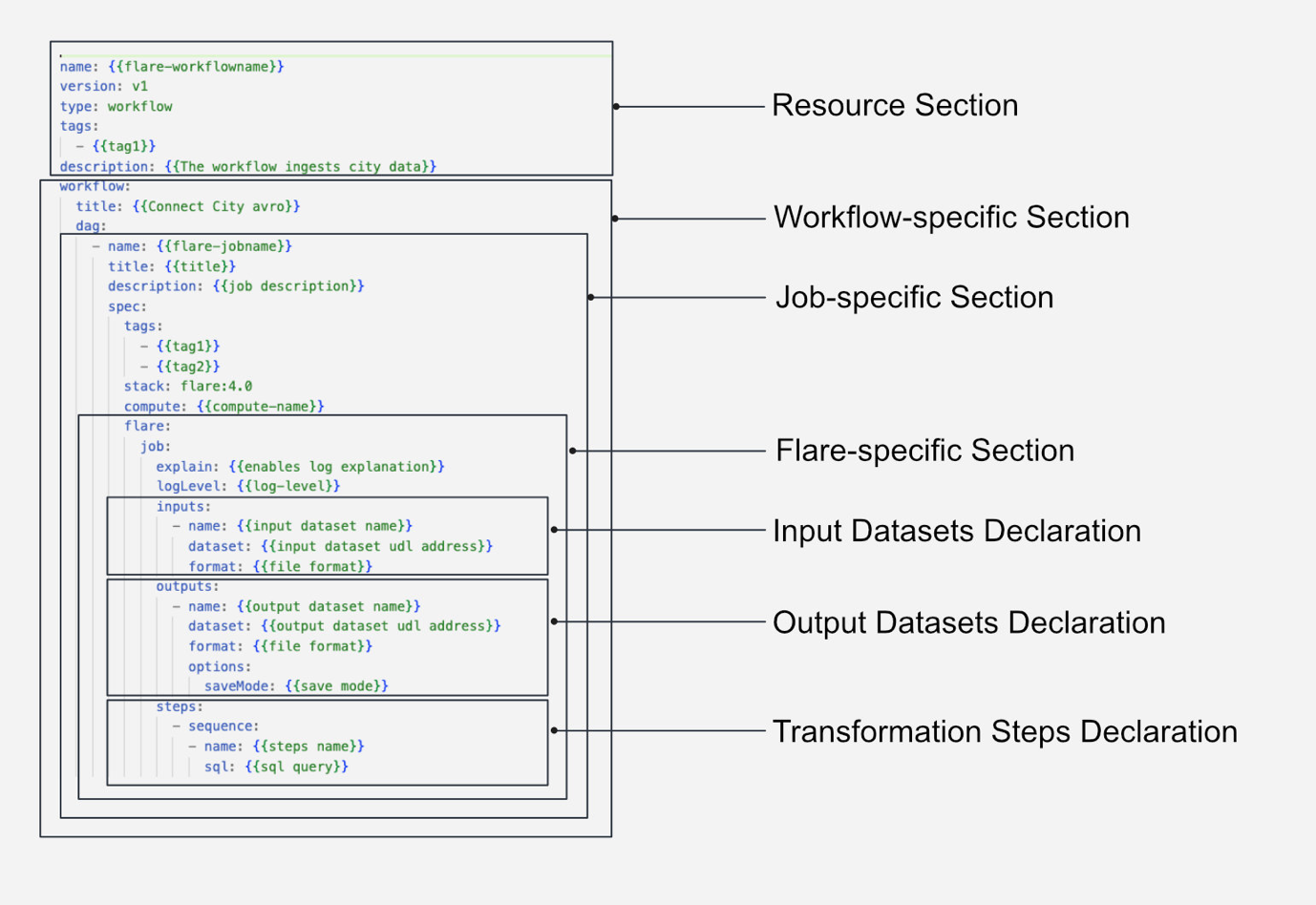
**Integration and Deployment**

**Integration with Other Systems (for Data Engineers)**

The data can be integrated with any other data system through various methods,

1. Spark based Flare jobs where the raw datasets can be integrated or migrated to their destination

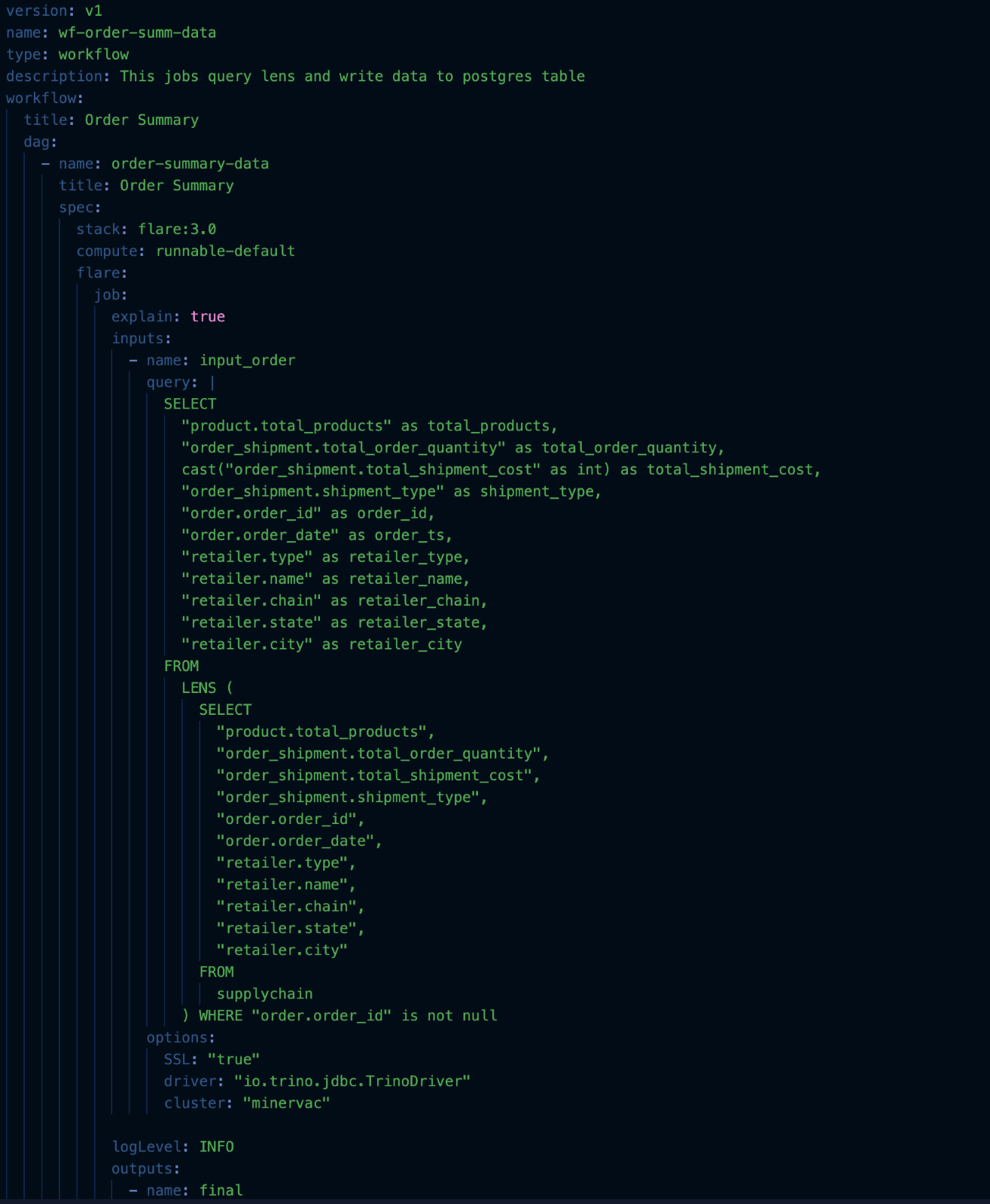
Flare is a Data Migration Engine Built on top of Apache Spark. A typical Flare YAML has the following Structure,

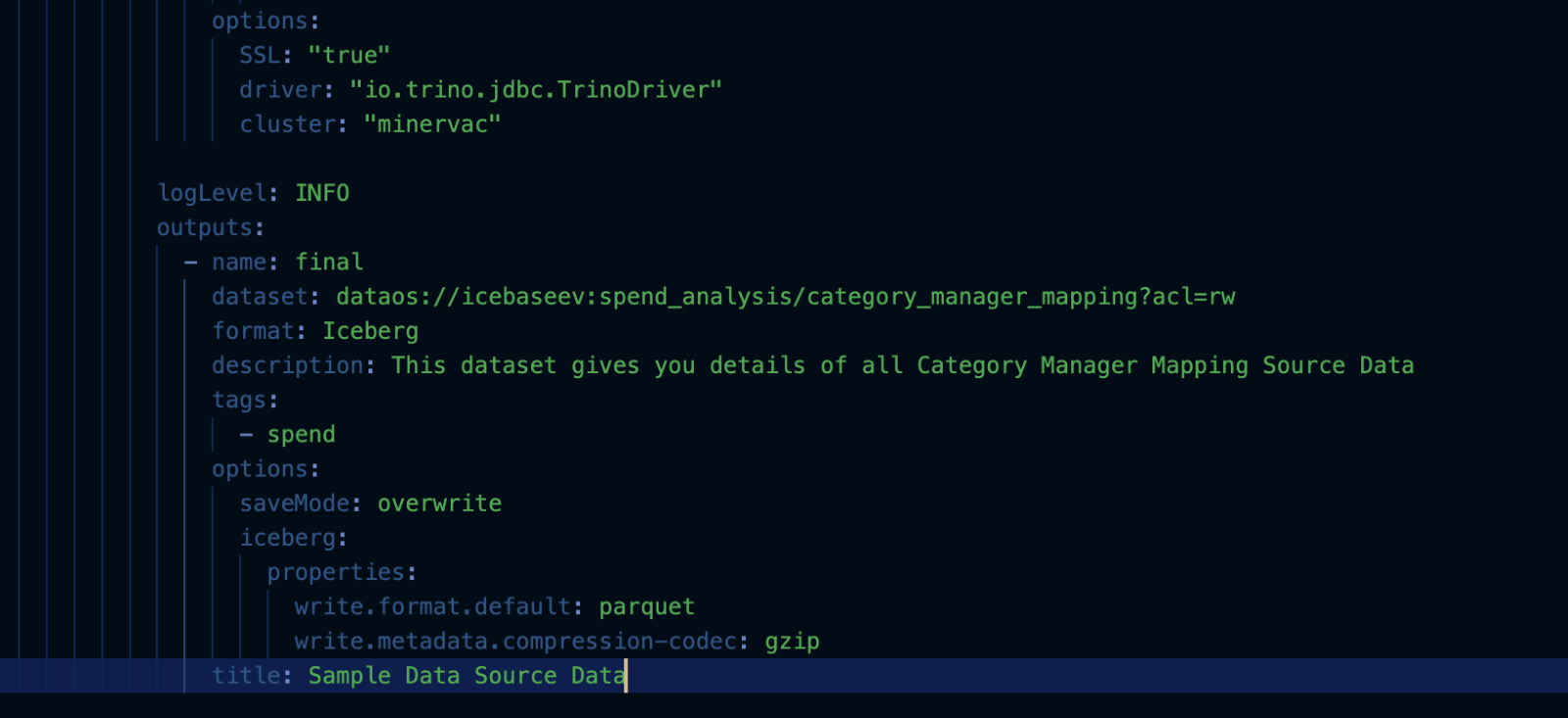


You can find more information about flare in the following link

(<https://dataos.info/resources/stacks/flare/#syntax-of-flare-yaml-configuration>)

Here is an example flare job for an enriched orders dataset that can be built from the Customer360 Data Product





You can access the source YAMLs in the following github repository.

<https://github.com/StacknexusDatanators/retail-accelerator>

1. TrinoSQL based query connection to all the datasets and lenses avaxtilable

(NEED TO COMPLETE)

**Scalability and Performance Considerations**

Scalability and performance considerations in the Customer 360 Data Product involve optimizing infrastructure to accommodate growing datasets and user demands. Implementing scalable architecture ensures responsiveness, allowing the platform to efficiently handle increased data volume, user interactions, and evolving analytics needs, fostering a seamless and high-performance experience.

**Backup and Disaster Recovery Procedures**

Backup and disaster recovery strategies for the Customer 360 Data Product are pivotal for safeguarding against unforeseen events. Regular backups of critical data are conducted to secure against accidental deletions, system failures, or cyber threats. An effective disaster recovery plan outlines procedures for swift data restoration, minimizing downtime. Off-site data storage and redundancy measures are implemented to enhance resilience. Periodic testing of recovery procedures ensures their efficacy, guaranteeing that the Customer 360 Data Product can swiftly recover from disruptions, maintain data integrity, and resume operations, thereby safeguarding against potential data loss or system outages.

**Use Cases and Scenarios**

**Typical Use Cases for Business Analysis**

1. Personalized Marketing Campaigns:

Analyze customer preferences and behavior to tailor marketing campaigns for increased engagement and conversion rates.

2. Customer Segmentation:

Utilize segmentation models to categorize customers based on demographics and behaviors, optimizing targeted marketing strategies.

3. Churn Prediction and Retention Strategies:

Implement predictive analytics to identify at-risk customers and devise proactive retention strategies, reducing churn rates.

4. Sales Performance Analysis:

Evaluate sales data to identify top-performing products, optimize pricing strategies, and enhance overall sales performance.

5. Customer Satisfaction Analysis:

Analyze customer feedback and support interactions to assess satisfaction levels and improve customer service strategies.

**Advanced Analytical Scenarios for Data Scientists**

1. Predictive Modeling for Revenue Forecasting:

Develop and deploy predictive models to forecast revenue based on historical data and market trends.

2. Anomaly Detection for Fraud Prevention:

Implement advanced anomaly detection algorithms to identify unusual patterns indicative of potential fraud in transactions.

3. Natural Language Processing (NLP) for Customer Sentiment Analysis:

Apply NLP techniques to analyze customer reviews and social media data for sentiment analysis, informing brand perception strategies.

4. Machine Learning Recommender Systems:

Develop personalized recommender systems using machine learning algorithms for accurate product or content recommendations.

5. Time Series Analysis for Demand Forecasting:

Utilize time series analysis to forecast customer demand, optimizing inventory management and supply chain operations.

What is a Data Product?

A data product is a package of data, metadata, characteristics of the data, means of use, Governance, Quality.

Data Products Spec sheet:

A data product spec sheet will also include the usecases, scenarios and other information..

Data product spec sheet encapsulates the entirety of information that users seek and consume about the product, including exhaustive details on its sources, technologies employed, use cases, and generated insights. This encompasses all relevant data-driven insights, aims, and functionalities, serving as a holistic repository of knowledge vital for understanding the product's utility from a business perspective. Data product spec sheet details use cases such as: spanning software applications, platforms, algorithms, or systems, serve to transform raw data into actionable insights, optimizing processes, enhancing user experiences, and fostering informed decision-making across diverse industries and domains.

Data Product platform:

It supports ingress data interoperability and egress data interoperability.

### **Interoperability:**

The Customer 360 Data Product demonstrates strong interoperability by seamlessly integrating and consolidating data from disparate sources, including AWS Redshift, PostgreSQL, and Google Web Analytics. This interoperability ensures that data from various formats and platforms can be harmoniously processed and utilized within the system. Additionally, the use of technologies such as Apache Spark, Apache Kafka, and PostgreSQL databases in the system requirements guarantees that these components interact efficiently, allowing for smooth data flow and processing across the entire ecosystem. The architectural diagrams visually represent the interconnections, showcasing the system's interoperability by illustrating how different components communicate and work together to create a unified view of customer data.

### **Reliability:**

Reliability in the Customer 360 Data Product is evident through its emphasis on real-time updates and robust system configurations. The continuous and real-time updates to customer profiles ensure the accuracy and responsiveness of the system, reflecting its reliability in reflecting current customer behaviors and preferences. Specifications detailing the Compute Cluster, Apache Kafka Cluster, and PostgreSQL Database, with specific node configurations and redundancy measures, reinforce the system's reliability by ensuring high availability and fault tolerance. The architecture's fault-tolerant design, along with redundancy in clusters, contributes to the system's resilience against failures, further enhancing reliability.

### **Governance:**

Governance within the Customer 360 Data Product encompasses comprehensive security measures, compliance considerations, and structured data stewardship. The product incorporates robust security measures, including encryption for data in transit and at rest, role-based access controls (RBAC), and stringent audit trails. These measures ensure data integrity, confidentiality, and regulatory compliance (such as GDPR, HIPAA) while mitigating risks associated with unauthorized access or data breaches. The delineation of governance policies like Data Lineage and History Tracking through applications like Hera and METIS showcases the system's adherence to governance principles, facilitating traceability and accountability across data lifecycles. The clear definition of data ownership roles, appointment of data stewards, and data lineage mechanisms underline the product's commitment to governance best practices.

### **Discoverability:**

Discoverability within the Customer 360 Data Product is underscored by its detailed technical specifications, comprehensive documentation, and traceability mechanisms. The product's exhaustive documentation encompasses detailed information about features, data sources, system requirements, and dependencies, facilitating easy access and understanding for users involved in data analysis, engineering, and science. The inclusion of tools like Hera and METIS for data lineage and history tracking amplifies the system's discoverability, enabling users to navigate through the data ecosystem, trace data origins, and comprehend the historical evolution of customer profiles. Additionally, the RBAC and DAC policies ensure that users can access and explore data within their prescribed access policies, further enhancing discoverability while safeguarding sensitive information.

Definition for C360  
The Customer 360 is an innovative data solution designed to provide organizations with a unified, in-depth understanding of their customer base. By consolidating data from various sources like CRM systems, sales transactions, and online interactions, the C360 creates comprehensive customer profiles, detailing demographics, purchase history, and preferences. This holistic view enables businesses to enhance customer experiences, refine marketing strategies, and make informed decisions. Tailored for data professionals, it enables seamless integration and advanced analytics, fostering personalized interactions, operational efficiency, and lasting customer loyalty. The C360 stands as a powerful tool leveraging data for precise insights and strategic advantage.

DATA SCHEMAS

|  |  |  |
| --- | --- | --- |
| **Product (Redshift) - Updated Weekly** | | |
| **Column Name** | **Description** | **Type** |
| sku\_id | Product SKU | VARCHAR |
| product\_id | Product ID | VARCHAR |
| product\_name | Product Name | VARCHAR |
| product\_category | Product Category | VARCHAR |
| product\_subcategory | Product Subcategory | VARCHAR |
| brand\_id | Brand ID | VARCHAR |
| product\_colour | Product Colour | VARCHAR |
| product\_size | Product Size | VARCHAR |
| list\_price | List Price | DOUBLE |
| sale\_price | Sale Price | DOUBLE |
| launched\_ts | Time when launched | TIMESTAMP |

|  |  |  |
| --- | --- | --- |
| **Brand (Redshift) - Updated Monthly** | | |
| **Column Name** | **Description** | **Type** |
| brand\_id | Brand ID | VARCHAR |
| brand\_name | Brand Name | VARCHAR |
| parent\_company | Parent Company | VARCHAR |
| description | Description | VARCHAR |
| brand\_type | Brand Type | VARCHAR |
| website | Website | VARCHAR |
| founded\_year | Founded Year | VARCHAR |
| headquarters | Head Quarters Address | VARCHAR |

|  |  |  |
| --- | --- | --- |
| **Order Line Item (PostgreSQL) - Updated in real time** | | |
| **Column Name** | **Description** | **Type** |
| order\_li\_id | Order LIne ID | VARCHAR |
| order\_id | Order ID | VARCHAR |
| product\_id | Product ID | VARCHAR |
| qty | Quantity | INT |
| **Orders (PostgreSQL)** | | |
| **Column Name** | **Description** | **Type** |
| Order\_id | Order ID | VARCHAR |
| Customer\_id | Customer ID | VARCHAR |
| Order\_state | Order Date | TIMESTAMP |
| Order\_delivery\_date | Order Delivery Date | TIMESTAMP |
| Discounted\_amount | Discounted Amount | DOUBLE |
| Shipping\_amount | Shipping Amount | DOUBLE |
| Total\_amount | Total Amount | DOUBLE |
| Payment\_method | Payment Method | VARCHAR |
| Billing\_address | Billing Address | VARCHAR |
| Shipping\_address | Shipping Address | VARCHAR |
| Transaction\_id | Transaction ID | VARCHAR |

|  |  |  |
| --- | --- | --- |
| **Transactions (PostgreSQL) - Updated in Real time** | | |
| **Column Name** | **Description** | **Type** |
| transaction\_id | Transaction ID | VARCHAR |
| order\_id | Order ID | VARCHAR |
| amount | Amount | DOUBLE |
| payment\_method | Payment Method | VARCHAR |
| transaction\_ts | Transaction Time Stamp | TIMESTAMP |

|  |  |  |
| --- | --- | --- |
| **Customer (Red Shift) - Updated Dail** | | |
| **Column Name** | **Description** | **Type** |
| customer\_id | Customer ID | VARCHAR |
| first\_name | First Name | VARCHAR |
| last\_name | Last Name | VARCHAR |
| gender | Gender | VARCHAR |
| phone\_number | Phone Number | VARCHAR |
| email\_address | Email Address | VARCHAR |
| date\_of\_birth | Date of Birth | TIME STAMP |
| age | Age | INT |
| education\_level | Education Level | VARCHAR |
| marital\_status | Marital Status | VARCHAR |
| no.\_of\_children | Number of Children | INT |
| registration\_date | Registration Date | TIME STAMP |
| occupation | Occupation | VARCHAR |
| hobbies | Hobbies | VARCHAR |
| annual\_income | Annual Income | VARCHAR |
| degree\_of\_loyalty | Degree of Loyalty | INT |
| mailing\_street | Mailing Street | VARCHAR |
| city | City | VARCHAR |
| state | State | VARCHAR |
| country | Country | VARCHAR |
| zipcode | Zip Code | VARCHAR |

**Security:** C360 employs a security framework encompassing encryption, access control, and audit trails. Through advanced encryption techniques, the product secures data both during transit and at rest, shielding sensitive customer information from unauthorized access or breaches. Access control mechanisms, such as Role-Based Access Control (RBAC) and Discretionary Access Control (DAC), limit data access based on user roles. This ensures that only authorized personnel, like Data Stewards or Data Producers, can access specific information while employing data masking techniques for sensitive Personally Identifiable Information (PII). Detailed audit trails track and monitor user activities and changes made to customer profiles, fostering accountability and enabling rapid identification of any potential security threats or unauthorized alterations. These robust security measures collectively fortify the platform against vulnerabilities and ensure the protection of customer data from various threats.

**Compliance:** C360 aligns with pertinent compliance standards and regulations, underscoring its commitment to data privacy, protection, and responsible data management. Adhering to stringent data privacy regulations such as GDPR and HIPAA, the product implements stringent measures to safeguard customer data, ensuring lawful and ethical handling practices. Additionally, the establishment of data retention policies within C360 ensures secure data storage and responsible disposal, in compliance with stipulated retention periods and privacy requirements. These compliance efforts not only foster customer trust by upholding data integrity and privacy but also mitigate risks associated with legal and regulatory implications.

By integrating security measures and aligning with stringent compliance standards, the Customer 360 data product not only guarantees data confidentiality and integrity but also ensures ethical data handling practices. This comprehensive approach to security and compliance within C360 instils confidence in users and organizations, fostering trust in the platform's capability to handle sensitive customer information responsibly and securely across its data ecosystem.

**Example how a lens query works:**

The query exemplifies a Customer 360 lens in action, showcasing its ability to extract specific insights. By focusing on Californian customers, it retrieves details such as the total customer count and segmented profiles like "High Spending Customers" or "Expecting Parents." This demonstrates the lens's capability to distill vast datasets into actionable customer attributes, facilitating targeted decision-making and personalized engagement strategies.

**Sample Query:**

|  |
| --- |
| SELECT  \*  FROM  LENS (  SELECT  "customer.total\_customers",  "segment.segment\_name"  FROM  c360\_solution\_accelerator  WHERE  "customer.state"='California'  LIMIT  50000  ) |

**Sample Output:**

|  |  |
| --- | --- |
| customer/total\_customers  COUNT | segment/segment\_name  STRING |
| 106 | High Spending Customers |
| 4 | Expecting Parents |
| 1 | Frequent Cart Abandoner |
| 1 | Frequent Buyers |

**Sample of a TrinoSQL based query connection:**

The Python script connects to a TrinoSQL server and queries ten records from the 'clickstream' table in the 'retail\_acclerator' database. This connection uses HTTPS authentication with specific headers for access, allowing easy data retrieval from TrinoSQL servers using Python.

**Sample of TrinoSQL connection in python and query:**

|  |
| --- |
| conn = connect(  host="tcp.cheerful-maggot.dataos.app",  port="7432",  auth=BasicAuthentication("user\_name","YXRsYXNfZTc1YzFhNjZhZTQwNmRiN2QyZjQ1MWI"),  http\_scheme="https",  http\_headers={"cluster-name": "minervac"}  ) |
| Query = ‘’’SELECT \* FROM icebasedev.retail\_acclerator.clickstream LIMIT 10’’’ |

**Beacon:**

The Beacon API endpoint, serves a crucial role in the realm of retail analytics. Specifically designed for churn prediction, this API retrieves data from a PostgreSQL database, a robust open-source relational database system. Operating within the DataOS platform, the API is integral to forecasting customer churn in retail scenarios. By leveraging the capabilities of PostgreSQL and the DataOS environment, it plays a key role in providing valuable insights for businesses aiming to anticipate and address customer attrition effectively.

|  |
| --- |
| **URI = "https://cheerful-maggot.dataos.app/churnpred/api/v2/retail\_churn\_preds\_data"** |

**Data Quality Metrics:**

Ensuring the reliability and accuracy of data is paramount for the effectiveness of the Customer 360 Data Product. To maintain high data quality, regular checks and assessments are conducted on each dataset. Key quality metrics include:

* + **Null Values:** Regular checks are performed to identify and address any null values within the datasets. Null values can impact analysis and decision-making, so their presence is meticulously monitored.
  + **Completeness:** Assessments are made to ensure that each dataset is complete, with all necessary fields populated. This metric helps guarantee that there are no missing components crucial for a comprehensive understanding of customer profiles.
  + **Consistency:** Data consistency is vital for accurate analytics. Checks are conducted to identify and rectify any inconsistencies or anomalies within the datasets, maintaining a coherent and reliable dataset.
  + **Timeliness:** Timely updates are crucial for reflecting real-time changes in customer profiles. The frequency of data updates aligns with the dynamic nature of customer interactions, ensuring that the information remains current.

**Frequency of Quality Checks:** Regular data quality checks are integrated into the operational workflow of the Customer 360 Data Product. The frequency of these checks is determined by several factors:

* + **Source Data Updates:** Quality checks align with the update frequency of source data. For datasets with more dynamic sources, such as real-time transactions, checks may be more frequent.
  + **Business Needs:** The frequency is also tailored to meet specific business requirements and decision-making cycles. Critical datasets may undergo more frequent assessments to support timely and accurate decision-making.
  + **Data Volume and Complexity:** Larger datasets or those with intricate relationships may require more frequent checks to maintain a high level of data quality.

By incorporating these quality metrics and aligning checks with relevant factors, the Customer 360 Data Product ensures that the information driving business decisions remains accurate, reliable, and up-to-date.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Quality Checks | Frequency | Latest Quality Score |
| Customer | Value Range Check – Age  Completeness Check - Contact Information  Accuracy Check – Gender | Weekly | 80% |
| Product | Missing Values Check – Description  Consistency Check – Category  Uniqness Check – Product ID | Weekly | 95% |
| Orders | Consistency Check - Order Status  Timeliness Check – Delivery Date  Accuracy Check – Payment Method | Daily | 87% |
| Web Analytics | Outliers Check – Page Views  Completness Check – User Sessions  Trend Analysis – Bounce Rate | Daily | 92% |
| Brands | Uniqueness Check – Brand ID  Accuracy Check – Brand Description  Consistency Check - Brand Category | Monthly | 97% |
| Transactions | Completeness Check – Transaction Details  Accuracy Check – Transaction Amount  Timeliness Check – Transaction Processing Time | Daily | 91% |

**Data Update Frequency:**

**Customer Dataset:**

* + **Frequency:** Daily
  + **Explanation:** Customer profiles are dynamic and subject to frequent changes. Daily updates ensure that the system reflects the most recent customer interactions, preferences, and behaviours.

**Product Dataset:**

* + **Frequency:** Weekly
  + **Explanation:** Product information, including inventory levels, features, and pricing, may not change as rapidly as customer profiles. Weekly updates strike a balance between maintaining accuracy and minimizing unnecessary data processing.

**Orders Dataset:**

* + **Frequency:** Real-time
  + **Explanation:** Given the critical nature of order data, updates occur in real-time to capture transactions as they happen. This enables prompt responses to changes in customer behaviour and facilitates accurate order fulfilment.

**Transactions Dataset:**

* + **Frequency:** Daily
  + **Explanation:** Daily updates ensure that transactional data, including payment information and order details, are current. This frequency aligns with the need for up-to-date insights into customer purchasing behaviour.

**Order Line Item Dataset:**

* + **Frequency:** Real-time
  + **Explanation:** Real-time updates are crucial for granular details related to order line items. This ensures that any changes or modifications are immediately reflected in the dataset.

**Web Analytics Dataset:**

* + **Frequency:** Hourly
  + **Explanantion:** Web analytics data can offer insights into customer interactions in near real-time. Hourly updates allow for quick adaptation to changing online behaviours and trends.

**Brands Dataset:**

* + **Frequency:** Weekly
  + **Explanation:** Brand-related information, such as partnerships or product launches, may not change as frequently. Weekly updates strike a balance between accuracy and computational efficiency.

Questioning the data

1. What is a data product – Clear definition
2. What is customer 360
3. Interoperability, etc.
   1. Interoperability – how c360 is interoperable
   2. Govern...
4. What are the datasets in there
5. How are they useful.

Joby’s points

1. Where does the data come from? Some brief on the diverse sources
2. Reliability of the data and accuracy
3. Addon in the target audiences.
4. In key feature & capabilities In 5 segmentation and targeting we should specify the criteria.
5. In data volume and update frequency
   1. How regularly updates in the data any specific time span?
   2. Minimum and maximum data volume?
   3. Max data which server can handle
   4. Criteria which we are using for the most active and current customers and the filter we are using.
6. In security and compliance consideration
   1. Network security
   2. Third party data source risk management
   3. Backup plan elaborated if needed
7. In assumption and limitation in limitation data constrains what are the types of data limitations we will be going to face.
8. In Business context and application we can add Marketing Optimisation
9. In advanced analytics scenario for Datascientist we can include customer service enhancement by reviewing feedback.

Note : Lens Metrics and dimensions and TrinoSQL needs to complete

Sections to add

1. What is a data product – Clear definition
2. Interoperability, Reliability, Governance, Discoverability
3. C360 - definition
4. Data Sources
5. Data Schema <TABLE , source, dataset, format, update frequency, volume>
6. What are the quality metrics , Frequency
7. Now in the customer 360 Lens has 53 columns , they are, facts about that, definitions, basic EDA
8. Security and Compliance
9. How to query from the lens – Trino
   1. Example Queries
   2. Sample Results