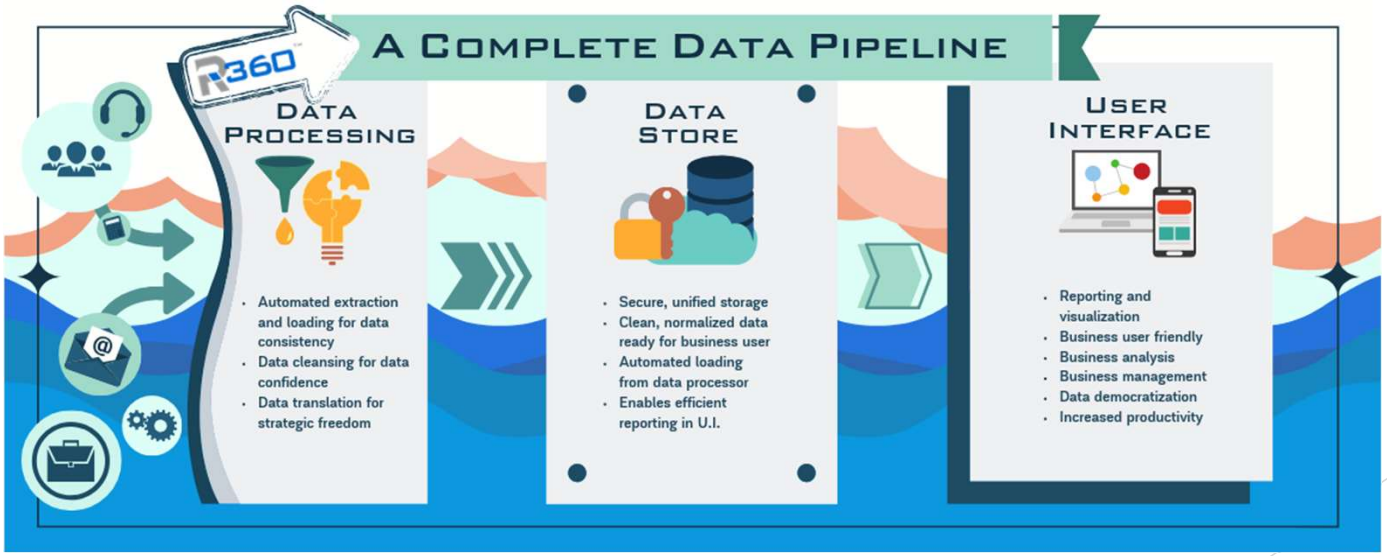
Data Pipeline Solutions for Modern E-commerce

Driving Business Insights for B2B & B2C Online Retail

Members:
- Gaurav Chugh (Gaurav.gauravchugh@aivancity)
- Sivasankar Bhusaram (sivasankar.bhusaram@aivancity)
- AkshayKumar.Kashyap

Organization & Business Domain

- Client uses B2B and B2C models in e-commerce.
- **B2B:** Bulk purchasing, supply chain complexity, larger order sizes.
- **B2C:** Personalized experiences, dynamic pricing, high volume of small transactions.
- Digital channels drive retail value by personalizing experiences, expanding reach, and optimizing operations for customer and business growth



Problem Definition : Real Time Customer Challenges

- Delayed Order Tracking
- Inventory Mismatch (Data Quality issue, inconsistent data formats, missing OR incomplete data)
- Slow responses to Customer behavior. (Network bottlenecks due to inefficient data processing frameworks)
- Managing High Data volumes. (Handling this surge with scaling horizontally)
- Personalization issues
- Stress problems such as data fragmentation, late insights, and missed opportunities to upscale customer experience through reactive business practices

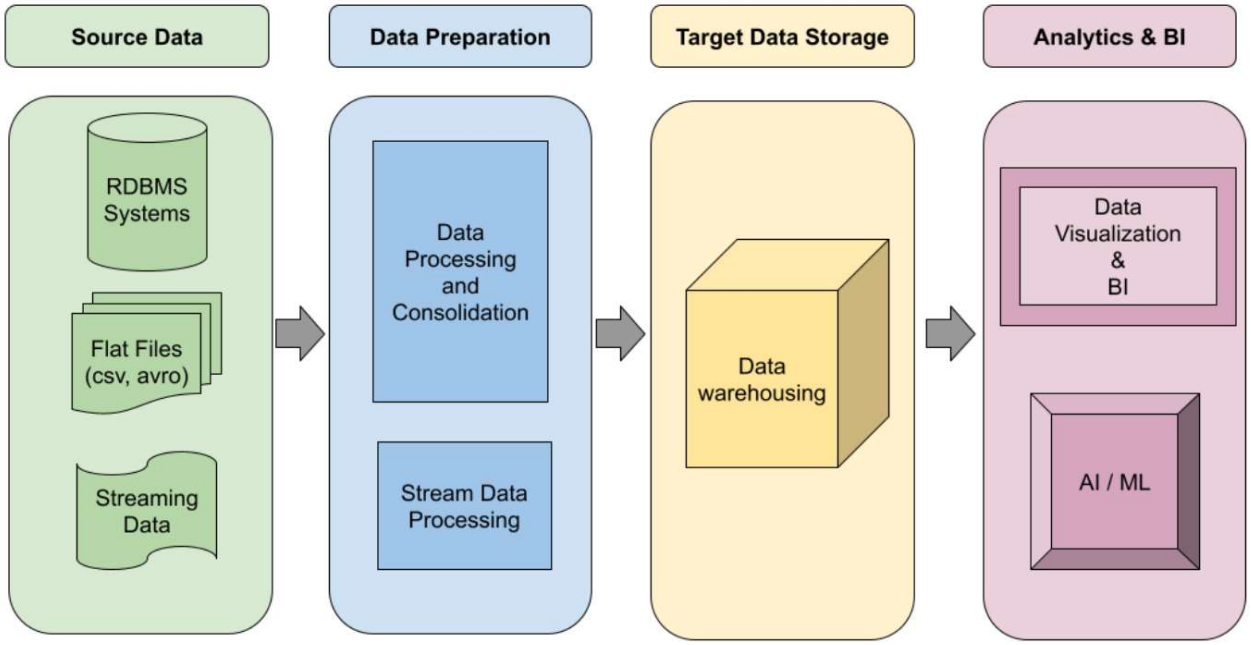
Personalized Product Recommendations

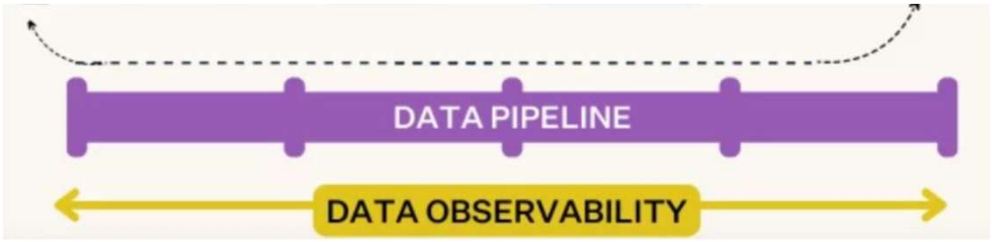
- **Issue:** Showing generic or irrelevant products.
- **Solution with Data Pipeline:**
  - **Data Ingestion & Storage:** User interaction data (clicks, views, add-to-carts, purchases, search queries, time spent on page) is captured from the e-commerce platform and stored in MinIO (for raw logs/events) or landed in PostgreSQL staging tables. Product catalog data is also ingested.
  - **Processing & Modeling (dbt & Airflow):**
    - Airflow orchestrates dbt jobs that run frequently (e.g., every 5-15 minutes for near real-time).
    - dbt transforms this raw data into user profiles, item embeddings, user-item interaction matrices, and calculates product similarity ("users who viewed X also viewed Y," "frequently bought together") or collaborative/content-based filtering scores. Incremental dbt models update these features efficiently.
- **Serving (PostgreSQL):** Pre-computed recommendations or user/item features are stored in PostgreSQL, accessible with low latency by the website/app.
- **Real-time effect:** As a user browses, their recent actions are fed into the pipeline. Within minutes, their personalized recommendations on the homepage, product pages, or in-cart suggestions are updated.

Dynamic Content & Offer Personalization:

- **Issue:** Displaying the same banners, promotions, or homepage layout to everyone.
- **Solution with Data Pipeline:**
  - **Data Ingestion & Segmentation (MinIO, PostgreSQL, dbt, Airflow):**
    - Collect behavioral data, purchase history, and demographic data (if available).
    - dbt models, orchestrated by Airflow, segment users based on this data (e.g., "high-value customers," "deal seekers," "new users," "users interested in X category"). Segments are updated frequently.
  - **Real-time Decisioning Logic:** The e-commerce frontend can query PostgreSQL for a user's segment or specific real-time behavioral triggers (e.g., viewed 3+ items in a specific category in the current session).
  - **Serving Personalized Content:** Based on the segment or real-time triggers, the website dynamically displays personalized banners, tailored offers, or even custom landing pages.

Data Pipeline Architecture flow





Data Pipeline - Technology Stack Overview

- **Apache Airflow:** For orchestrating workflows.
- **DBT:** For SQL-based data transformation.
- **PostgreSQL:** As a robust, scalable storage solution.
- **MinIO:** For object storage that mimics AWS S3.
- **OpenMetadata:** Opensource data catalog to collects, organizes, and indexes metadata from multiple sources.

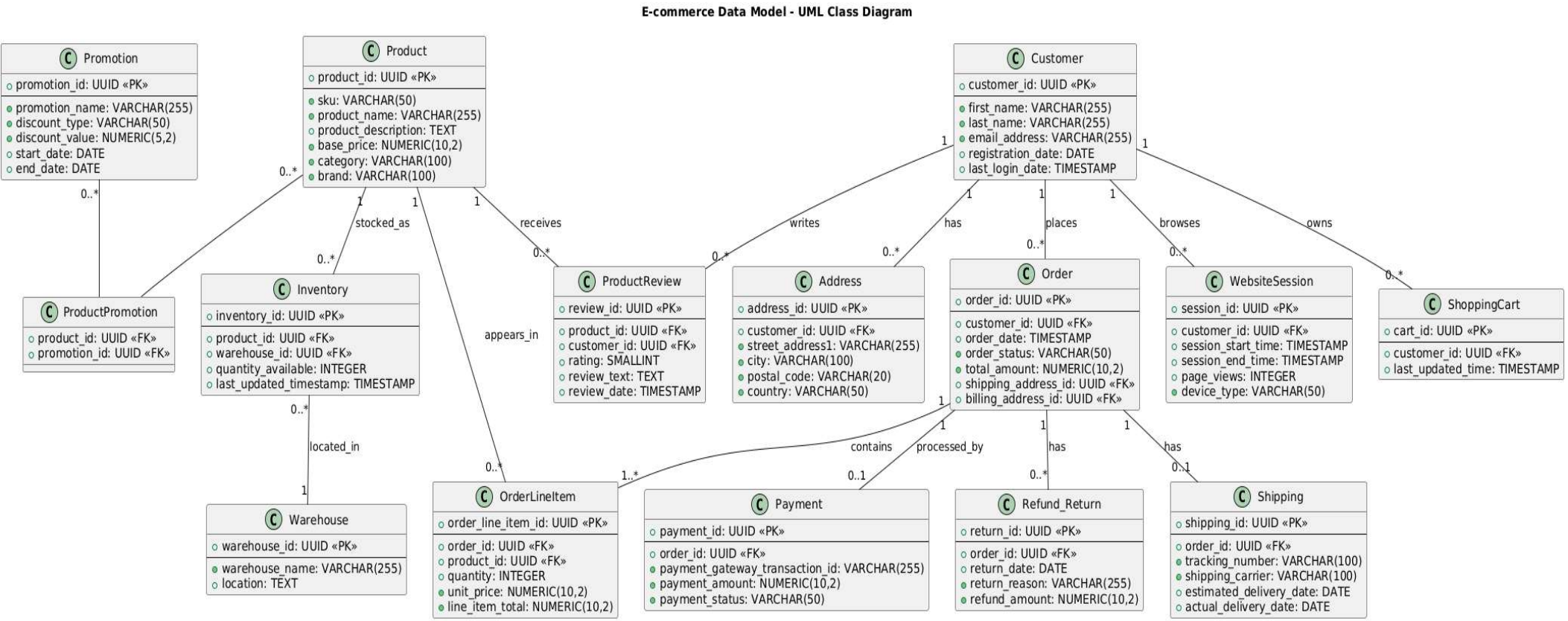
Data Governance & Catalog

- **Definition:** Data Governance is the framework of rules, policies, standards, processes, and controls for managing and using an organization's data assets.
  - **Growing Importance:** Essential for ensuring **data quality** (accuracy, completeness, consistency for reliable analytics and personalization), **data security** (protecting sensitive customer and business data), and **compliance** (meeting regulations like GDPR, CCPA, etc.). In e-commerce, this builds customer trust and avoids costly penalties.
- **Power of Data Catalogs & Glossaries:**
  - **Data Catalogs (like OpenMetadata):** Provide a centralized, searchable inventory of all data assets (databases, tables, dashboards, dbt models, etc.). They enable **data discovery**, allowing teams to easily find and understand relevant data.
  - **Business Glossaries:** Standardize data definitions and business terms (e.g., "Active Customer," "Gross Merchandise Value," "Conversion Rate") across the organization, ensuring everyone speaks the same data language. This resolves ambiguity and improves communication.

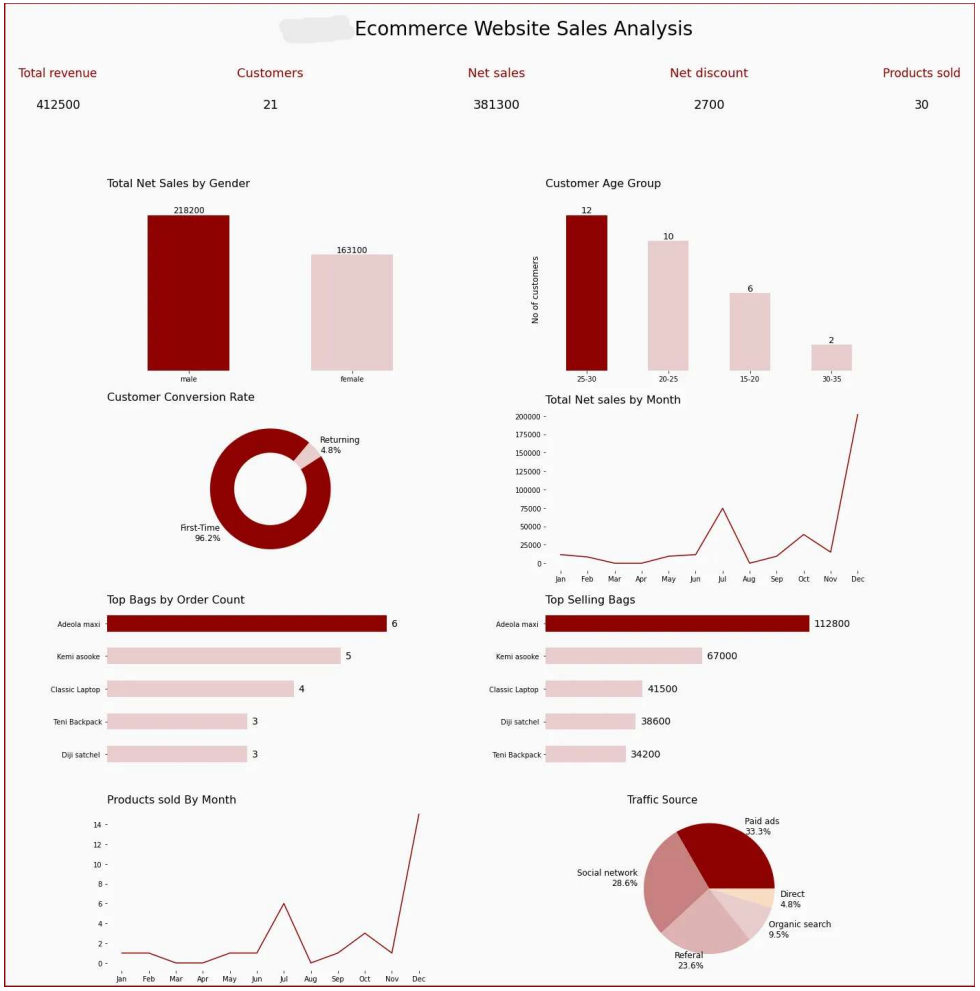
Data Glossary & Metadata Management

•**Customer:** A unique individual or entity that has made at least one purchase on the e-commerce platform.
•**Order:** A confirmed transaction including one or more products, shipping details, and payment information. A 'valid order' requires successful payment authorization and inventory confirmation.
•**Cart Abandonment:** The event where a customer adds items to their online shopping cart but does not complete the purchase within a specified timeframe (e.g., 30 minutes).
•**Conversion Rate:** The percentage of website visitors who make a purchase. Calculated as (Number of Transactions / Number of Website Visitors) * 100.
•**Inventory Turnover:** The number of times a company's inventory is sold and replaced over a period. Calculated as (Cost of Goods Sold / Average Inventory).

Data Modeling & Associations & Entity Relationship



Ecommerce Dashboards KPIs For In-Depth Insights



Why Data Governance Matters

- Data governance establishes clear standards, processes, and responsibilities for data creation, storage, and usage. This leads to more accurate, consistent, and reliable data across all systems.
- **Improved Data Quality:**
  - **Ensures Accuracy:** Correct product info, reliable customer profiles, clean transactions.
  - **Reduces Errors:** Minimizes manual corrections, streamlines operations.
  - **Builds Trust:** Leads to better customer experience and fewer returns.
- **Regulatory Compliance:**
  - **Ensures Privacy:** Adheres to GDPR, CCPA, PCI DSS for customer data.
  - **Avoids Fines:** Protects against penalties and reputational damage.
  - **Audit Ready:** Provides clear data handling documentation.
- **Enhanced Decision-Making:**
  - **Actionable Insights:** Fuels effective marketing, optimized inventory, and personalized experiences.
  - **Strategic Confidence:** Enables data-driven decisions for growth and profitability.

Future Improvements & Scopes in Data Pipelines

- **Emerging Trends:**

•**Real-time Ingestion & Processing:** Moving beyond batch to immediate data availability for instant insights.
•**Cloud-Native Architectures:** Leveraging scalable and flexible cloud services (e.g., AWS, Azure, GCP) for infrastructure.
•**Serverless Computing:** Reducing operational overhead with services that automatically scale and manage infrastructure.

- **Technological Advancements:**

•**Increased Automation:** From setup to scaling, minimizing manual intervention.
•**AI/ML Integrations:** Embedding machine learning for intelligent data quality, anomaly detection, and predictive processing within pipelines.
•**Self-Healing & Adaptive Pipelines:** Systems that automatically detect and resolve issues, or adapt to changing data volumes/types.

Advanced Monitoring & Automation

•**Real-time Monitoring for Proactive Issue Resolution:**
•Leverage real-time monitoring tools (logs, metrics, dashboards) to gain immediate insights into data pipeline health and performance.
•Identify bottlenecks, data quality issues, or resource constraints **before** they escalate into critical failures.
•Example: Tracking data volume processed, latency, error rates, and resource utilization in real-time.
•**Automated Alerts & Self-Healing Capabilities:**
•Implement intelligent alert systems that trigger notifications for predefined thresholds or anomalies.
•Develop self-healing mechanisms (e.g., auto-restarts for failed tasks, dynamic resource scaling) to minimize downtime and manual intervention.
•Focus on **end-to-end pipeline visibility** to trace data lineage and pinpoint exact failure points rapidly.

Feedback & Iterative Improvement Process

- Need for (CI-CD) continuous improvement in pipeline performance.
- Recapitulate the benefits: data quality, compliance, faster analytics.
- Encourage an action-oriented approach to data-driven transformation

Conclusions & Key Takeaways

- **Solving the Business Problem with a Robust Data Pipeline:**
  - We addressed the critical need for [**timely customer insights, efficient inventory management**] through a comprehensive data pipeline solution.
  - Our chosen tech stack – utilizing [**PostgreSQL for storage, Spark for processing, Kafka for streaming**] – provides a scalable, resilient, and performant foundation.
  - The defined pipeline architecture, from ingestion to consumption, ensures data flows efficiently and reliably.
  - Strong data governance principles are embedded throughout, guaranteeing data quality, security, and compliance.

- **Realizing Tangible Benefits:**
  - **Enhanced Data Quality:** By implementing robust validation and transformation steps, we ensure data is clean, accurate, and ready for use.
  - **Ensured Compliance & Security:** Our approach adheres to regulatory requirements (e.g., GDPR, local privacy laws) and incorporates strong security measures to protect sensitive information.
  - **Faster & More Reliable Analytics:** Business users and analysts now have timely access to high-quality data, enabling quicker insights and more informed decision-making.

- **Driving Data-Driven Transformation:**
  - This data pipeline is a cornerstone for [**company name**] journey towards becoming a truly data-driven organization.
  - It empowers real-time operations, supports advanced analytics, and opens doors for future AI/ML initiatives.
  - We encourage an action-oriented approach, fostering collaboration between business and technical teams to continuously leverage and evolve these data capabilities for strategic advantage.

THANK YOU

- Gaurav Chugh
- Sivasankar Bhusaram
- Akshaykumar Kashyap