AetherMart Semester Project

Fall 2025

Project Overview

This project aims to design, implement, and manage a scalable and intelligent data infrastructure for an e-commerce platform. Beyond just product catalog management, this system will encompass critical data for customer profiles, order processing, and product reviews, supporting AetherMart's multi-channel sales strategy. You will simulate a growing business, incorporating various data management techniques, including advanced SQL, distributed databases, ETL pipelines, NoSQL integration, and vector database capabilities for AI-driven features like semantic search and recommendations. The project will emphasize data integrity, security, performance, and high availability across all data domains.

Business Case: AetherMart

About AetherMart

Founded five years ago by a group of tech enthusiasts in a garage, AetherMart has rapidly evolved from a niche online store selling custom-built gaming PCs to a leading e-commerce platform for cutting-edge technology. Our product catalog now spans high-tech gadgets, sophisticated smart home devices, premium electronics, and an expanding suite of bespoke digital services, including AI-powered home automation consultations. We pride ourselves on being early adopters of innovation, offering a meticulously curated selection of products that push the boundaries of technology.

AetherMart operates on a multi-channel sales strategy, reaching customers through our primary e-commerce website, a rapidly growing mobile application, and strategic pop-up retail partnerships in major tech hubs. This diversified approach means we collect vast amounts of data not only on products but also on customer interactions, purchase histories, and valuable product reviews. Our culture is fast-paced, highly collaborative, and driven by a relentless pursuit of customer satisfaction and technological excellence. We believe in empowering our customers with the latest innovations, and our growth trajectory demands a data infrastructure that is as forward-thinking and robust as our product offerings, capable of seamlessly managing product, customer, order, and review data across all channels.

Key Personnel

Alex Kim (CTO): Alex is the co-founder and technological visionary behind AetherMart. With a background in distributed systems and a keen eye for emerging technologies, Alex is the strategic force driving AetherMart's technical roadmap. He's known for his demanding but fair leadership, always pushing the team to innovate and build resilient, scalable solutions.

He spends his days balancing long-term architectural planning with urgent operational demands, constantly evaluating how technology can give AetherMart a competitive edge. He values stability and efficiency above all else, knowing that any technical hiccup directly impacts the bottom line and AetherMart's reputation for reliability, especially as transaction volumes grow across multiple sales channels.

Sarah Chen (Head of Product Management): Sarah joined AetherMart two years ago, bringing a wealth of experience from a larger retail tech giant. She's the voice of the customer and the market, meticulously curating AetherMart's product offerings and defining the user experience across web and mobile. Sarah is highly analytical and data-driven, constantly seeking insights to optimize product discovery, merchandising, and sales performance. Her biggest frustrations often stem from data inconsistencies or delays that prevent her team from making quick, informed decisions about product launches, inventory adjustments, or understanding which products are performing best on which sales channel. She's a strong advocate for clean, accessible data that directly supports business growth and a seamless multi-channel experience.

David Miller (Lead Database Architect): David is AetherMart's quiet but indispensable architect of data. He was one of the earliest hires, brought in to bring order to the company's fledgling data systems. With a deep expertise in relational databases and a growing interest in distributed systems, David is responsible for the design, implementation, and maintenance of AetherMart's entire data infrastructure. He's meticulous, pragmatic, and often the first to identify potential scalability or performance bottlenecks. David is constantly challenged to balance the need for immediate solutions with building a future-proof, robust data backbone that can support not just product data, but also the rapidly expanding volumes of customer, order, and review data from all sales channels. He reports directly to Alex and is the go-to person for all things data.

Maria Rodriguez (Head of Marketing & Analytics): Maria leads AetherMart's efforts to understand its customers and market its products effectively across all channels. She's a dynamic leader with a strong grasp of data analytics, always looking for new ways to leverage insights for targeted campaigns, personalized recommendations, and competitive analysis. Maria is particularly passionate about AI and machine learning, seeing them as the future of customer engagement and cross-channel personalization. Her team relies heavily on clean, integrated data—from customer demographics and purchase history to product review sentiment—to build sophisticated models and measure campaign effectiveness. She often pushes David's team for faster access to richer, more diverse datasets that can inform AetherMart's multi-channel strategy.

Project Goals

- A well-structured MariaDB database schema for product catalog, customer, order, and review data.

- Implemented advanced SQL features for data manipulation, security, and integrity across these new data domains.
- A functional multi-node MariaDB Galera Cluster demonstrating high availability for all critical business data.
- Designed and partially implemented robust ETL pipelines for data ingestion and transformation from various sources, supporting a multi-channel strategy.
- Explored and integrated a complementary NoSQL database for specific use cases (e.g., unstructured product reviews, flexible customer attributes).
- Leveraged MariaDB's vector capabilities for semantic search on product data and potentially for customer insights.
- Analyzed and presented the architectural choices, performance considerations, and security measures implemented for a holistic e-commerce data ecosystem.

General Milestone Requirements

Scripting Requirement

All setup, configuration, and database operations should be **scripted as much as possible**. Prioritize using SQL scripts for database-specific tasks (schema creation, stored procedures, triggers, partitioning, etc.) and shell scripts (Bash) or Python scripts for environment setup, data loading, and orchestration on your Ubuntu EC2 instances. Manual steps should be minimized and clearly documented if unavoidable.

Video Submission

For each milestone, you will create a video demonstration. Save your video to YouTube with an **unlisted privacy level**. You will submit the YouTube URL through Peerceptiv on the course website.

AI Usage Disclosure

If you use AI tools (e.g., large language models, code generators) to assist with any part of the milestone (e.g., writing SQL queries, debugging code), you must explicitly address the following in your PowerPoint slide deck:

- How AI was used: Describe the specific AI tool(s) and how they were applied.
- Corrections made: Detail any significant corrections, modifications, or refinements you had to make to the AI-generated output.
- Lessons Learned: Reflect on what you learned from using AI for that specific task, including its benefits, limitations, and how it impacted your understanding or efficiency.

Milestones

Milestone 1: Foundational MariaDB Setup, Advanced SQL & Initial Data Management (Weeks 1-2)

Objective

Establish the core relational database infrastructure, implement foundational data management practices, and apply advanced SQL features for data preparation and security, now encompassing product, customer, and order data.

Business Case Narrative

It was a tough Monday morning for Sarah Chen, AetherMart's Head of Product Management. She glared at the latest product report, a chaotic mess of inconsistent data. Every new gadget added seemed to introduce another formatting issue. Sarah walked directly to David Miller's office, AetherMart's Lead Database Architect. "David, I'm pulling my hair out trying to get a consistent list of our new smart home devices. Some product names are abbreviated, others have full descriptions, and the pricing data is all over the place. How can we even think about launching new features if our core product data isn't reliable?"

David leaned back, a familiar weariness in his eyes. "You're absolutely right, Sarah. Our current manual processes for data entry and updates just aren't scalable. We're onboarding suppliers faster than we can manually clean their data. But it's not just products. Maria's team is struggling with fragmented customer data from our website and mobile app, and our order system is barely holding together. We need to implement a robust, standardized database schema, automate data ingestion for products, customers, and orders, and enforce strict data quality rules right from the source. It's the only way to get this under control across all our sales channels."

Later that day, in their weekly tech leadership sync, Alex Kim, the CTO, shifted the conversation to a more pressing strategic concern. "David, I've also been reviewing our internal audit findings, and frankly, our current access controls for all our databases are far too lax. We're handling sensitive product information, supplier contracts, and now, increasingly detailed customer and order data. A data breach, or even just inconsistent product listings or incorrect order fulfillment caused by unauthorized changes, can severely damage AetherMart's reputation and lead to compliance issues. Security and data integrity are non-negotiable as we continue to scale aggressively." David knew he had a critical mandate: lay a secure, reliable foundation for AetherMart's core operational data in the cloud, covering products, customers, and orders.

Key Tasks

- Environment Setup (AWS EC2 & Ubuntu): Deploy and manage an AWS EC2 instance (Ubuntu OS) to host your MariaDB database. Ensure proper network configuration (security groups, inbound/outbound rules) for secure access. Create a dedicated database for the e-commerce platform (e.g., `aethermart_db`).

- Schema Design & Data Ingestion Preparation: Design a robust relational schema for core e-commerce data (`products`, `categories`, `suppliers`, `customers`, `orders`, `order_items`, `reviews`). You will be provided with initial datasets for these tables, which include some intentional errors. Your task will be to prepare for their ingestion and subsequent cleansing.

- Initial Data Ingestion (ETL/ELT): Implement a basic ETL/ELT process (e.g., Python script, SQL `LOAD DATA INFILE`) to ingest the provided data into MariaDB for all designed tables. Handle basic loading requirements and initial data type conversions.

- Advanced SQL Implementation: Create SQL `VIEWS`, `VIRTUAL COLUMNS`, `STORED PROCEDURES`, and `USER-DEFINED FUNCTIONS (UDFs)` for common data operations and transformations.

- Data Cleansing & Security Basics: Implement basic data cleansing techniques using SQL to address intentional errors in the provided data. Design and implement user roles and permissions for basic access control. Apply SQL queries to identify data quality issues and potential security vulnerabilities.

Deliverables

- SQL scripts for database creation, schema definition, view, virtual column, stored procedure, and UDF creation.

- SQL scripts or Python code for initial data ingestion.

- SQL scripts for user role and permission setup.

- A PowerPoint slide deck documenting the schema, data ingestion methodology, ETL process, and initial data quality/security findings, recorded as part of your demo.

- A video presentation (Technical & Feature Walkthrough) demonstrating your progress and explaining your PowerPoint slides.

Rubric

Criteria	Beginning	Developing	Proficient	Mastery
Technical Implementation				
MariaDB Environment Setup (AWS EC2)	**1** EC2 instance not deployed or MariaDB not installed/accessible.	**2** EC2 instance deployed, but MariaDB has connectivity/configuration issues.	**4** EC2 instance and MariaDB mostly configured; minor network/security group issues.	**5** AWS EC2 instance (Ubuntu) correctly deployed and configured; MariaDB installed and fully accessible; secure network setup.
Schema Design & Data Ingestion Preparation	**3** Schema is absent or fundamentally flawed; no preparation for data ingestion.	**7** Schema exists but lacks structure/normalization for new data types; minimal preparation for data ingestion.	**11** Logical schema for all core tables with minor improvements needed; sufficient preparation for data ingestion.	**15** Well-structured, logical, and normalized schema for `products`, `customers`, `orders`, `order_items`, and `reviews`; effective preparation for ingesting provided data.
Initial Data Ingestion	**2** ETL/ELT process fails or ingests data incorrectly for primary tables.	**4** ETL/ELT process has significant errors or requires substantial manual intervention for new data types.	**7** ETL/ELT process works for most tables but could be more efficient or handle edge cases better.	**10** ETL/ELT process effectively and correctly ingests provided data for all designed tables; basic loading requirements handled.
Advanced SQL Features	**3** Few or no advanced SQL features implemented; existing features are non-functional.	**7** Some advanced SQL features attempted but with major errors or incomplete functionality, especially for new data types.	**11** Most advanced SQL features implemented across data domains, but some have minor bugs or limited scope.	**15** All required `VIEWS`, `VIRTUAL COLUMNS`, `STORED PROCEDURES`, and `UDFs` correctly implemented and demonstrate intended functionality across `products`, `customers`, `orders`, and `reviews`.

Criteria	Beginning	Developing	Proficient	Mastery
Data Cleansing & Security Basics	**1** No data cleansing or security measures implemented.	**2** Attempted basic data cleansing or security, but ineffective or incomplete for new data types.	**4** Basic data cleansing applied; user roles/permissions designed but may have minor flaws across domains.	**5** Basic data cleansing techniques effectively applied for products, customers, and orders to address provided errors; user roles and permissions correctly designed and implemented for basic access control across different data domains.
PowerPoint Slide Deck Quality				
Content & Organization	**4** Slide deck is absent or poorly organized; content is minimal or unclear.	**8** Slide deck is present but lacks organization or clarity; content is incomplete.	**14** Slide deck is organized and clear; content is mostly complete and accurate.	**20** Slide deck is well-structured, clear, and comprehensive; all required content is accurate and professionally presented.
Video Presentation Quality				
Clarity & Demonstration	**4** Video is absent or very poor quality; demonstrations are unclear or missing.	**8** Video is present but hard to follow; demonstrations are incomplete or confusing.	**14** Video is clear and mostly concise; demonstrations are present but could be more effective.	**20** Video is clear, concise, and engaging; demonstrations are effective and clearly showcase the work.
AI Usage Disclosure & Reflection				

Criteria	Beginning	Developing	Proficient	Mastery
Disclosure & Lessons Learned	**2** No disclosure of AI usage or superficial mention.	**4** Mentions AI usage but lacks detail on application, corrections, or lessons learned.	**7** Adequately describes AI usage and corrections; provides some reflection on lessons learned.	**10** Clearly states AI tools used, details specific corrections/refinements, and provides insightful reflection on benefits, limitations, and impact of AI usage.

Milestone 2: NoSQL Concepts, Data Distribution Strategies & Partitioning Implementation (Weeks 3-4)

Objective

Understand the motivations behind NoSQL databases, explore conceptual strategies for distributing data, and implement MariaDB data partitioning to address scalability for AetherMart's growing and diverse data.

Business Case Narrative

The initial data consolidation brought a sense of order to AetherMart's core operational data, but its rapid growth and multi-channel strategy presented new, complex challenges. In a mid-week strategy session, Maria Rodriguez, Head of Marketing & Analytics, pointed to a slide filled with customer feedback. "David, our marketing team needs to quickly analyze thousands of customer reviews to understand sentiment about our new 'Quantum Echo' smart speaker. Trying to parse all that text from a relational table is a nightmare. Can't we store this kind of unstructured data more effectively? We need to react faster to customer feedback, whether it comes from the website or our mobile app."

Sarah Chen, always thinking about the product roadmap and customer experience, added, "And from a product management perspective, our new bespoke digital services have completely different attributes than our physical gadgets. Trying to fit everything into a single, highly normalized table is becoming incredibly complex. We need more flexibility in how we define and store product attributes, and even customer preferences, without constantly altering the main schema. Our current rigid structure is slowing down new product introductions and personalized customer experiences."

David listened intently, already anticipating these issues. "I hear you both. The relational model excels at structured, transactional data like core orders and products, but for highly variable or unstructured data like reviews, or for massive analytical queries on our rapidly growing customer and product catalogs, it can become a bottleneck. We need to start thinking beyond just SQL. Also, our `products`, `customers`, and `orders` tables are projected to hit tens of millions of records by next year. We need a concrete plan for how to scale them without grinding everything to a halt when a single database server can't handle the load anymore, especially with traffic coming from multiple channels. This means we need to start implementing partitioning."

Alex Kim, sensing the shift needed, concluded, "Performance and adaptability are absolutely critical. We can't afford to be limited by our database architecture as we expand our product lines, user base, and sales channels. Explore all options, David, especially how we can prepare for massive analytical workloads and integrate these diverse data types efficiently. We need to be agile and ready for whatever data comes next, and that includes making our current MariaDB setup more scalable."

Key Tasks

- MariaDB Triggers: Design and implement at least two `TRIGGERS` in MariaDB to enforce complex data integrity rules for products, customers, and orders. Analyze their operational impact.
- NoSQL Conceptual Analysis: Research and explain NoSQL motivations, CAP theorem, and characteristics of key-value/document models relevant to AetherMart.
- Data Partitioning Implementation (MariaDB SQL): Implement horizontal partitioning for at least one large table. Justify and demonstrate.
- MariaDB ColumnStore Exploration: Research and evaluate ColumnStore for AetherMart's analytical workloads.

Deliverables

- SQL scripts for trigger implementation.
- SQL scripts for implementing data partitioning.
- A PowerPoint slide deck detailing trigger analysis, NoSQL concepts, partitioning implementation, and ColumnStore analysis, recorded as part of your demo.
- A video presentation (Technical & Feature Walkthrough) demonstrating your progress and explaining your PowerPoint slides.

Rubric

Criteria	Beginning	Developing	Proficient	Mastery
Technical Implementation				
MariaDB Triggers	**4** No triggers implemented or they fail to function.	**9** Triggers attempted but contain major errors or do not enforce rules correctly.	**15** Triggers mostly functional, but operational impact analysis is limited.	**20** Two or more triggers correctly designed and implemented for products, customers, or orders; operational impact thoroughly analyzed, considering performance implications.
NoSQL Conceptual Analysis	**3** Explanation of NoSQL is inaccurate or very limited; CAP theorem not addressed.	**7** Basic explanation of NoSQL motivations and models, but with inaccuracies or missing key concepts.	**11** Clear explanation of NoSQL motivations, CAP theorem, and model characteristics; some relevant use cases for AetherMart's data.	**15** Clear, accurate, and comprehensive explanation of NoSQL motivations, CAP theorem, and characteristics of key-value/document models with relevant e-commerce use cases for products, customers, orders, and reviews.
Data Partitioning Implementation	**2** No partitioning implemented or implementation is fundamentally incorrect.	**4** Partitioning attempted but with major errors, incorrect strategy, or no clear justification.	**7** Partitioning implemented and functional, but strategy or column choice could be better justified; minor issues in demonstration.	**10** Horizontal partitioning correctly implemented on at least one large table using SQL; choice of strategy and column is well-justified; effects of partitioning are clearly demonstrated.

Criteria	Beginning	Developing	Proficient	Mastery
MariaDB ColumnStore Exploration	**1** No exploration of ColumnStore or inaccurate information.	**2** Limited exploration of ColumnStore; benefits/scenarios are unclear.	**4** Basic understanding of ColumnStore architecture and some benefits/scenarios.	**5** Clear and accurate evaluation of MariaDB ColumnStore's architecture, benefits, and suitable analytical scenarios for AetherMart's diverse e-commerce data. **Note: ColumnStore is for evaluation/exploration, not required implementation.**
PowerPoint Slide Deck Quality				
Content & Organization	**4** Slide deck is absent or poorly organized; content is minimal or unclear.	**8** Slide deck is present but lacks organization or clarity; content is incomplete.	**14** Slide deck is organized and clear; content is mostly complete and accurate.	**20** Slide deck is well-structured, clear, and comprehensive; all required content is accurate and professionally presented.
Video Presentation Quality				
Clarity & Demonstration	**4** Video is absent or very poor quality; demonstrations are unclear or missing.	**8** Video is present but hard to follow; demonstrations are incomplete or confusing.	**14** Video is clear and mostly concise; demonstrations are present but could be more effective.	**20** Video is clear, concise, and engaging; demonstrations are effective and clearly showcase the work and explain conceptual topics.
AI Usage Disclosure & Reflection				

Criteria	Beginning	Developing	Proficient	Mastery
Disclosure & Lessons Learned	**2** No disclosure of AI usage or superficial mention.	**4** Mentions AI usage but lacks detail on application, corrections, or lessons learned.	**7** Adequately describes AI usage and corrections; provides some reflection on lessons learned.	**10** Clearly states AI tools used, details specific corrections/refinements, and provides insightful reflection on benefits, limitations, and impact of AI usage.

Milestone 3: MariaDB Replication & Clustering Fundamentals (Weeks 5-6)

Objective

Implement both standard MariaDB replication and a highly available Galera Cluster, and begin exploring MariaDB's vector capabilities for future AI-driven features, now for all critical e-commerce data.

Business Case Narrative

The tension in the air was palpable during the executive meeting. AetherMart had experienced a brief but impactful database hiccup last week, causing a temporary halt to product listings and sales across all channels. Alex Kim started the meeting with a stern tone. "David, that brief database hiccup last week cost us thousands in potential sales and a lot of customer goodwill. We're an online-first business, relying on our website, mobile app, and partners; 24/7 availability isn't a luxury, it's a fundamental requirement. What's our plan to ensure we never have a single point of failure again for our core operational data, including products, customers, and orders?"

Sarah Chen immediately agreed. "Exactly. When the product catalog isn't accessible, customers can't browse. When the order system is down, they can't buy. It directly impacts our bottom line and brand perception across every channel. We need a system that can withstand failures without interruption."

David, having anticipated this, presented his findings. "I've been looking into various replication and clustering solutions. Standard primary-replica replication offers some redundancy, but failover isn't always automatic, which means some downtime. For true high availability and near-zero data loss on node failure, especially for our critical customer and order data, we need to explore synchronous multi-primary solutions like Galera Cluster. This milestone will focus on setting up and rigorously testing both, so we can compare their suitability for AetherMart's needs."

Maria Rodriguez, ever focused on innovation, then interjected, "And while you're at it, can we start thinking about how we'll power more intelligent search? Our current keyword search is okay, but customers often search for concepts, not just exact terms. 'Smart home security for apartments' should bring up relevant cameras and sensors, not just products with 'smart' in the name. And can we also think about personalized recommendations based on customer behavior?" David nodded, adding, "That's where vector capabilities come in. We can start by adding a placeholder for product embeddings, and even for customer preferences, which will be crucial for semantic search and personalized recommendations later. But first, let's nail down this high availability."

Key Tasks

- Standard MariaDB Replication Setup (AWS EC2 & Ubuntu): Deploy and manage at least two AWS EC2 instances (Ubuntu OS) to construct and configure a basic

MariaDB primary-replica replication setup. Ensure proper network configuration (security groups, inbound/outbound rules) for replication traffic. Perform basic replication operations, including verifying data synchronization for all core tables (products, customers, orders, reviews) and testing failover to the replica (conceptual). Document the advantages and disadvantages of asynchronous replication in an e-commerce context.

- MariaDB Galera Cluster Setup (AWS EC2 & Ubuntu): Deploy and manage at least three AWS EC2 instances (Ubuntu OS) to construct and configure a multi-node MariaDB Galera Peer-to-Peer Cluster. Ensure proper network configuration (security groups, inbound/outbound rules) for synchronous replication between cluster nodes.
- Cluster Operations & Testing: Perform basic cluster operations: Add a new node to the cluster. Gracefully remove a node. Test failover scenarios (e.g., stopping a node and verifying application connectivity to remaining nodes) for all critical data operations (product lookups, customer logins, order placements). Validate data consistency across all nodes after various operations and data modifications.
- Initial MariaDB Vector Capabilities: Explore MariaDB's native vector capabilities (e.g., using `VECTOR` data type or appropriate extensions/plugins if available). Demonstrate basic vector storage and retrieval by adding a simple vector column to the `products` table (e.g., for a placeholder "product description embedding") and potentially a `customers` table (e.g., for "customer preference embedding"). Ingest a small set of synthetic vector data into these columns.
- Data Integration into Clusters: Integrate a portion of your provided dataset (from Milestone 1) into both the standard replication setup and the newly formed MariaDB Galera Cluster. Validate data consistency for all data types (products, customers, orders, reviews) across both replication types after ingestion. Verify that the initial vector data is also replicated correctly in the Galera Cluster.

Deliverables

- Documentation (e.g., README or Wiki) detailing replication and cluster setup.
- Screenshots or video demonstrating Galera cluster operations.
- SQL scripts for basic vector storage and retrieval.
- A PowerPoint slide deck summarizing setup experiences, challenges, HA verification, and replication comparison, recorded as part of your demo.
- A video presentation (Technical & Feature Walkthrough) demonstrating your progress and explaining your PowerPoint slides.

Rubric

Criteria	Beginning	Developing	Proficient	Mastery
Technical Implementation				
Standard MariaDB Replication	**3** Standard replication setup is absent or fails to function.	**7** Standard replication attempted but has major configuration errors or data synchronization issues.	**11** Primary-replica replication mostly configured; some data synchronization issues or limited failover testing.	**15** Primary-replica replication correctly configured and verified for all core tables; advantages/disadvantages thoroughly documented.
MariaDB Galera Cluster Setup (AWS EC2)	**4** Galera Cluster setup is absent or fails to form.	**9** Galera Cluster attempted but has significant configuration errors or fails to maintain quorum.	**15** Multi-node Galera Cluster mostly configured; minor connectivity or consistency issues.	**20** Multi-node MariaDB Galera Peer-to-Peer Cluster correctly constructed and configured on EC2 for synchronous replication, supporting all core data.
Cluster Operations & Testing	**2** No demonstration of cluster operations or testing.	**4** Attempted cluster operations but with significant errors or lack of validation.	**7** Basic cluster operations demonstrated, but testing is limited or validation is incomplete.	**10** Successful demonstration of node addition, removal, failover testing for critical data operations, and thorough data consistency validation across the cluster.
Initial MariaDB Vector Capabilities	**1** Vector capabilities not explored or demonstrated.	**2** Attempted vector storage/retrieval but with errors or limited functionality.	**4** Basic vector storage and retrieval demonstrated, but integration or understanding is limited.	**5** Basic vector storage and retrieval effectively demonstrated by adding vector columns to relevant tables (e.g., products, customers) and ingesting/retrieving synthetic data.
PowerPoint Slide Deck Quality				

Criteria	Beginning	Developing	Proficient	Mastery
Content & Organization	**4** Slide deck is absent or poorly organized; content is minimal or unclear.	**8** Slide deck is present but lacks organization or clarity; content is incomplete.	**14** Slide deck is organized and clear; content is mostly complete and accurate.	**20** Slide deck is well-structured, clear, and comprehensive; all required content is accurate and professionally presented.
Video Presentation Quality				
Clarity & Demonstration	**4** Video is absent or very poor quality; demonstrations are unclear or missing.	**8** Video is present but hard to follow; demonstrations are incomplete or confusing.	**14** Video is clear and mostly concise; demonstrations are present but could be more effective.	**20** Video is clear, concise, and engaging; demonstrations are effective and clearly showcase the work.
AI Usage Disclosure & Reflection				
Disclosure & Lessons Learned	**2** No disclosure of AI usage or superficial mention.	**4** Mentions AI usage but lacks detail on application, corrections, or lessons learned.	**7** Adequately describes AI usage and corrections; provides some reflection on lessons learned.	**10** Clearly states AI tools used, details specific corrections/refinements, and provides insightful reflection on benefits, limitations, and impact of AI usage.

Milestone 4: Optimizing Clustered MariaDB & Advanced AI Data Preparation (Weeks 7-8)

Objective

Optimize the performance of the clustered MariaDB environment, design a more robust ETL pipeline, and generate/query vector embeddings for AI-driven features, now for all critical e-commerce data.

Business Case Narrative

The Galera Cluster was a significant step forward for AetherMart's availability, but new challenges quickly emerged. In a weekly stand-up, Maria Rodriguez addressed David directly. "David, the marketing team is ready to roll out semantic search for the 'AetherMart Discover' feature, and we're pushing for personalized recommendations. We need to ensure the product data is perfectly clean and the vector embeddings are accurate. Can we really get 'smart' search and recommendations if our data isn't perfectly prepped, including customer behavior data?"

David nodded grimly. "You're hitting on a critical point, Maria. Even with the cluster, we're seeing some query slowdowns, especially during peak traffic for order processing and customer lookups. And our ETL pipeline, while functional, still requires too much manual oversight for quality issues across products, customers, and orders. It's slowing down updates to the product catalog and delaying insights into customer behavior. We can't feed messy data to AI models and expect good results. We need to optimize the cluster's performance and automate the data preparation for vector generation to make it truly hands-off."

Sarah Chen added, "It's not just for AI. If a new product launch is delayed because of data quality issues in the pipeline, or if we can't quickly update customer profiles, we lose market advantage and customer trust. We need this entire process, from data ingestion to transformation, to be as smooth and hands-off as possible across all our operational data."

Later, Alex Kim sent a blunt message on Teams: "Performance and data quality are paramount across all our data domains. We've invested in this infrastructure; now we need to ensure it's running optimally and delivering the clean, high-quality data necessary for our next generation of AI-driven features and seamless multi-channel operations. Show me the numbers on those query optimizations, David, and a plan for a truly robust ETL that handles all our core data." The pressure was on to refine their data processes.

Key Tasks

- MariaDB Cluster Troubleshooting & Performance Analysis (AWS EC2 & Ubuntu): Identify and troubleshoot cluster issues. Analyze query performance. ✨ Expand Task

- Advanced ETL/Data Pipeline Design & Implementation: Design and implement a robust ETL pipeline for all core e-commerce data. ✨ Expand Task
- MariaDB Vector Embedding Generation & Similarity Search: Generate vector embeddings and perform similarity searches. ✨ Expand Task
- Refined AI Data Preparation: Refine feature engineering and normalization techniques. ✨ Expand Task

Deliverables

- A PowerPoint slide deck detailing cluster troubleshooting and performance analysis.
- Code for the advanced ETL/data pipeline.
- SQL scripts for vector embedding generation and similarity search.
- A PowerPoint slide deck on vector embedding methodology and results, recorded as part of your demo.
- A video presentation (Technical & Feature Walkthrough) demonstrating your progress and explaining your PowerPoint slides.

Rubric

Criteria	Beginning	Developing	Proficient	Mastery
Technical Implementation				
MariaDB Cluster Troubleshooting & Performance	**3** No troubleshooting or performance analysis performed.	**7** Identified some issues but analysis is superficial; no clear bottlenecks identified.	**11** Identified and analyzed most cluster issues; basic performance analysis with some tools.	**15** Identified and thoroughly troubleshooted common cluster issues on EC2; comprehensive performance analysis using appropriate tools (`EXPLAIN`, `SHOW STATUS`, `MariaDB Monitor`) to identify bottlenecks for product, customer, and order operations.
Advanced ETL/Data Pipeline Design & Implementation	**4** ETL pipeline is basic or dysfunctional; no enhanced data quality steps.	**9** ETL pipeline has major errors or lacks robust error handling/transformation; limited data quality steps for new data types.	**15** ETL pipeline is functional and includes some enhanced quality steps for most data types, but could be more robust or efficient.	**20** Robust ETL/data pipeline solution designed and implemented, leveraging advanced SQL features and including enhanced data quality and transformation steps for all core e-commerce data (products, customers, orders, reviews).
MariaDB Vector Embedding Generation & Similarity Search	**2** No vector embedding generation or similarity search.	**4** Attempted vector generation/search but with errors or incorrect results.	**7** Vector embeddings generated and stored; basic similarity search performed with minor issues.	**10** Vector embeddings correctly generated and stored for relevant data (e.g., products, customers, reviews); similarity search queries using MariaDB's vector functions are accurately performed, demonstrating ANN concepts.

Criteria	Beginning	Developing	Proficient	Mastery
Refined AI Data Preparation	**1** No refinement of AI data preparation.	**2** Superficial attempt at feature engineering or normalization; data still unsuitable for AI.	**4** Basic feature engineering and normalization applied; data is somewhat prepared for AI consumption.	**5** Thorough refinement of data preparation techniques for AI, including effective feature engineering and normalization of data for vector generation and other AI models.
PowerPoint Slide Deck Quality				
Content & Organization	**4** Slide deck is absent or poorly organized; content is minimal or unclear.	**8** Slide deck is present but lacks organization or clarity; content is incomplete.	**14** Slide deck is organized and clear; content is mostly complete and accurate.	**20** Slide deck is well-structured, clear, and comprehensive; all required content is accurate and professionally presented.
Video Presentation Quality				
Clarity & Demonstration	**4** Video is absent or very poor quality; demonstrations are unclear or missing.	**8** Video is present but hard to follow; demonstrations are incomplete or confusing.	**14** Video is clear and mostly concise; demonstrations are present but could be more effective.	**20** Video is clear, concise, and engaging; demonstrations are effective and clearly showcase the work.
AI Usage Disclosure & Reflection				

Criteria	Beginning	Developing	Proficient	Mastery
Disclosure & Lessons Learned	**2** No disclosure of AI usage or superficial mention.	**4** Mentions AI usage but lacks detail on application, corrections, or lessons learned.	**7** Adequately describes AI usage and corrections; provides some reflection on lessons learned.	**10** Clearly states AI tools used, details specific corrections/refinements, and provides insightful reflection on benefits, limitations, and impact of AI usage.

Milestone 5: Hybrid Data Architectures & Specialized NoSQL Integration (Weeks 9-10)

Objective

Understand the role of data lakes/warehouses and propose hybrid data architectures, integrating a complementary NoSQL database, to manage AetherMart's growing and diverse multi-channel data.

Business Case Narrative

The data team at AetherMart was facing a new kind of deluge. Beyond the structured product catalog, a tidal wave of diverse data was pouring in from all corners of their multi-channel strategy: granular customer clickstreams from the website and mobile app, sensor telemetry from their smart home devices in the field, unstructured social media interactions, and ever-growing customer support chat logs. In a recent data strategy meeting, Alex Kim pointed to a complex diagram. "David, we're drowning in data from all angles. Our primary MariaDB database, while excellent for transactional product, customer, and order data, simply isn't built for this volume and variety of unstructured or semi-structured information. How do we get a '360-degree view' of our products and customers across all channels without creating massive, inefficient data silos?"

Maria Rodriguez emphasized her team's urgent needs. "Exactly! I want to analyze granular customer journeys, correlate product purchases with real-time social media trends, and understand how our smart devices are being used in the wild. This kind of data isn't fitting neatly into our existing tables, and trying to force it in is both inefficient and costly. We need to unify insights from our online, app, and pop-up store sales data."

David, having already started researching, explained "This is where hybrid architectures come in. We can't put everything in MariaDB. We need to explore concepts like data lakes for raw, unstructured data ingest, and specialized NoSQL databases for specific use cases. For example, a document store could be ideal for flexible customer profiles that capture preferences from different channels, or maybe a graph database to map out complex supplier networks and product interdependencies. My plan is to propose a high-level data flow to connect these disparate systems, potentially on separate EC2 instances for specialized workloads, creating a truly integrated view."

Sarah Chen, always pragmatic, raised a crucial point. "And how does security play into this, David? If our data is spread across different platforms and even different types of EC2 instances, how do we ensure it's all protected consistently? A fragmented security approach is a recipe for disaster for AetherMart, especially with customer PII now residing in multiple places." David acknowledged, "That's a critical point, Sarah. We'll need to discuss comprehensive security implications and best practices for this heterogeneous environment, ensuring data integrity and access control across all platforms and data types."

The path forward was clear: AetherMart needed to embrace a multi-faceted data strategy to truly leverage its growing data assets and maintain its competitive edge.

Key Tasks

- Data Lakes vs. Data Warehouses: Differentiate and discuss their interaction with MariaDB and other sources.
- Hybrid Data Architecture Proposal: Analyze scenarios and propose a data flow strategy for a hybrid architecture.
- Specialized NoSQL Integration (Conceptual/Proof-of-Concept on AWS EC2 & Ubuntu): Choose a NoSQL model, evaluate, and implement a POC.
- Security in Heterogeneous Environments: Discuss security implications and best practices for diverse data platforms.

Deliverables

- A PowerPoint slide deck covering data lakes/warehouses, hybrid architecture, NoSQL analysis, POC documentation, and security considerations, recorded as part of your demo.
- A video presentation (Technical & Feature Walkthrough) demonstrating your progress and explaining your PowerPoint slides.

Rubric

Criteria	Beginning	Developing	Proficient	Mastery
Technical Implementation				
Data Lakes vs. Data Warehouses	**2** No differentiation or incorrect understanding.	**4** Superficial differentiation; interaction with MariaDB is unclear.	**7** Clear differentiation; some discussion of interaction with MariaDB and other data sources.	**10** Clear and accurate differentiation between data lakes and data warehouses; thorough discussion of their roles and interaction with the operational MariaDB cluster and other multi-channel data sources.
Hybrid Data Architecture Proposal	**4** No proposal or a fundamentally flawed diagram/explanation.	**9** Basic diagram/explanation, but lacks detail, justification, or clear data flows for multi-channel data.	**15** Generally sound diagram/explanation, but some components or data flows are unclear or not fully justified for all data types.	**20** Comprehensive and well-justified hybrid data architecture diagram with clear explanations of data flows between MariaDB, conceptual data lake, and specialized NoSQL stores, supporting insights across all sales channels.
Specialized NoSQL Integration (POC)	**3** NoSQL POC is absent or completely non-functional.	**7** NoSQL POC attempted but has major errors or lacks basic CRUD functionality.	**11** NoSQL POC is functional but limited in scope or has minor issues; basic evaluation of its use cases.	**15** Chosen NoSQL DB correctly set up on EC2; basic CRUD operations effectively demonstrated; thorough evaluation of its architectural considerations and ideal use cases for AetherMart's multi-channel data.

Criteria	Beginning	Developing	Proficient	Mastery
Security in Heterogeneous Environments	**1** No discussion of security implications.	**2** Superficial discussion of security, missing key challenges or best practices.	**4** Identifies some security challenges; proposes basic best practices.	**5** Thoughtful and comprehensive discussion of security implications and best practices for data storage across diverse, heterogeneous data platforms, including PII considerations.
PowerPoint Slide Deck Quality				
Content & Organization	**4** Slide deck is absent or poorly organized; content is minimal or unclear.	**8** Slide deck is present but lacks organization or clarity; content is incomplete.	**14** Slide deck is organized and clear; content is mostly complete and accurate.	**20** Slide deck is well-structured, clear, and comprehensive; all required content is accurate and professionally presented.
Video Presentation Quality				
Clarity & Demonstration	**4** Video is absent or very poor quality; demonstrations are unclear or missing.	**8** Video is present but hard to follow; demonstrations are incomplete or confusing.	**14** Video is clear and mostly concise; demonstrations are present but could be more effective.	**20** Video is clear, concise, and engaging; demonstrations are effective and clearly showcase the work.
AI Usage Disclosure & Reflection				
Disclosure & Lessons Learned	**2** No disclosure of AI usage or superficial mention.	**4** Mentions AI usage but lacks detail on application, corrections, or lessons learned.	**7** Adequately describes AI usage and corrections; provides some reflection on lessons learned.	**10** Clearly states AI tools used, details specific corrections/refinements, and provides insightful reflection on benefits, limitations, and impact of AI usage.

Milestone 6: Advanced ETL Orchestration, Data Governance & Project Synthesis (Weeks 11-12)

Objective

Design advanced ETL orchestration, apply data governance principles, assess end-to-end security, and synthesize all learned concepts in a final project presentation for AetherMart's comprehensive multi-channel data strategy.

Business Case Narrative

As the financial quarter drew to a close, the AetherMart executive team gathered for a critical review. The ambition to build a scalable and intelligent data infrastructure had been a significant investment, now encompassing product, customer, order, and review data from all sales channels. Alex Kim, CTO, opened the discussion. "David, Sarah, Maria, we've built a powerful data ecosystem with multiple databases and complex pipelines. But now we need to ensure it's sustainable, trustworthy, and, frankly, defensible to regulators. How do we guarantee data quality across all these systems, from ingestion on our EC2 instances to final consumption? What about data lineage – can we trace where every piece of product, customer, or order data comes from and how it's transformed? And crucially, what's our end-to-end security posture for this entire complex setup, especially with sensitive customer PII?"

Sarah Chen emphasized her ongoing need for reliable data for product strategy across all channels. "I need to be absolutely confident that the data I'm seeing for product performance and customer behavior is accurate and up-to-date, regardless of its source or which EC2 instance it lives on. Inaccurate data leads to bad business decisions, and we cannot afford that, especially with our multi-channel growth."

Maria Rodriguez, mindful of looming privacy regulations and the need for unified customer insights, added, "And from a compliance perspective, especially with the sensitive customer data we're now collecting across different platforms, we need clear governance rules and audit trails. How do we manage all this metadata effectively so we know exactly what data we have and where it is, and how it flows between systems? It's becoming increasingly complex, and the legal team is asking tough questions about data privacy and usage."

David, having coordinated the efforts throughout the project, was ready to present their comprehensive final strategy. "This milestone is all about bringing it together into a cohesive, manageable system. We've designed the advanced orchestration for our complex ETL pipelines, ensuring robust error handling, detailed logging, and proactive monitoring across all our EC2-hosted components, integrating data from all channels. We'll propose a detailed data governance framework for AetherMart, covering quality standards, metadata management, and compliance protocols for all data types, including sensitive customer information. Finally, we'll present a thorough end-to-end security assessment across our entire hybrid data landscape. The final presentation to the board will highlight how this robust data management architecture supports AetherMart's strategic goals, enables future growth, and solidifies our position as a trusted and innovative brand in the tech market."

The team knew this final push would define the success and long-term impact of their ambitious data initiative.

Key Tasks

- Advanced ETL Orchestration: Design an advanced ETL orchestration strategy for complex data pipelines.
- Data Governance Principles: Explain key data governance principles and propose their application.
- End-to-End Security Assessment: Assess the security posture of the entire system and propose mitigations.
- Project Synthesis & Presentation: Synthesize knowledge, propose optimal solutions, and present the overall architecture.

Deliverables

- A PowerPoint slide deck for advanced ETL orchestration.
- A PowerPoint slide deck outlining the application of data governance principles to your project.
- A PowerPoint slide deck for an end-to-end security assessment.
- A comprehensive final project PowerPoint presentation synthesizing all aspects, recorded as part of your demo.
- A video presentation (Technical & Feature Walkthrough) demonstrating your progress and explaining your PowerPoint slides.

Rubric

Criteria	Beginning	Developing	Proficient	Mastery
Technical Implementation (Conceptual/Design Focus)				
Advanced ETL Orchestration Design	**4** No design or a fundamentally flawed design for orchestration.	**9** Basic design for orchestration, but lacks error handling, logging, or monitoring mechanisms for all data types.	**15** Orchestration design is mostly sound but could be more robust or detailed in error handling/monitoring for diverse data.	**20** Robust design for advanced ETL orchestration, including comprehensive mechanisms for error handling, logging, and monitoring across distributed components and all data types (products, customers, orders, reviews).
Data Governance Principles	**3** No explanation of data governance or inaccurate understanding.	**7** Basic explanation of principles, but application to project data is superficial or incorrect for new data types.	**11** Clear explanation of principles; thoughtful application to most aspects of project data.	**15** Clear and accurate explanation of key data governance principles; thoughtful and comprehensive application to the project's product catalog, customer, order, and review data, including PII.
End-to-End Security Assessment	**3** No security assessment or a superficial/inaccurate one.	**7** Basic security assessment, but misses key vulnerabilities or mitigation strategies for new data types.	**11** Comprehensive assessment, but some vulnerabilities or mitigation strategies are overlooked or not fully detailed.	**15** Thoughtful and comprehensive end-to-end security assessment of the complex data management system; identifies potential vulnerabilities (including PII risks) and proposes robust mitigation strategies across all data types and EC2 deployments.
PowerPoint Slide Deck Quality				

Criteria	Beginning	Developing	Proficient	Mastery
Content & Organization	**4** Slide deck is absent or poorly organized; content is minimal or unclear.	**8** Slide deck is present but lacks organization or clarity; content is incomplete.	**14** Slide deck is organized and clear; content is mostly complete and accurate.	**20** Slide deck is well-structured, clear, and comprehensive; all required content is accurate and professionally presented.
Video Presentation Quality				
Clarity & Demonstration	**4** Video is absent or very poor quality; demonstrations are unclear or missing.	**8** Video is present but hard to follow; demonstrations are incomplete or confusing.	**14** Video is clear and mostly concise; demonstrations are present but could be more effective.	**20** Video is clear, concise, and engaging; demonstrations are effective and clearly showcase the work.
AI Usage Disclosure & Reflection				
Disclosure & Lessons Learned	**2** No disclosure of AI usage or superficial mention.	**4** Mentions AI usage but lacks detail on application, corrections, or lessons learned.	**7** Adequately describes AI usage and corrections; provides some reflection on lessons learned.	**10** Clearly states AI tools used, details specific corrections/refinements, and provides insightful reflection on benefits, limitations, and impact of AI usage.