aivancity
Paris Île-de-France · Nice Côte d'Azur

Certificat de Scolarité

Certificate of Enrollment

Ecommerce Data Pipeline for Real-Time KPI Intelligence

Pour la période académique 2024/2025

For the academic period 2024/2025

Final Year Project Report

MSc Cloud Computing & Data Engineering – Class of 2025

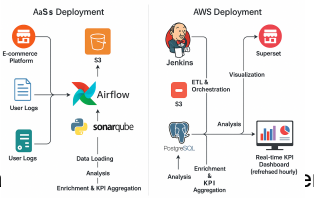
Paris-Cachan, le 6 novembre 2025

La Direction des Programmes atteste que :

Gaurav GAURAV CHUGH

Date de naissance : 01/12/1987 à Delhi (Inde)

Nationalité : indienne



est inscrit(e) en 5ème année du Programme "Ma ering (Paris Ile de France)" et suit régulièrement les cours qui ont débuté le 06/11/2024.

The Program Management certifies that

Gaurav GAURAV CHUGH

Date of birth : 01/12/1987 in Delhi (India)

Nationality: Indian

Participant:	Gaurav Chugh
Student ID:	AIV-CCDE-2025-017
Host Organisation:	Confidential Ecommerce Platform Provider
Supervising Professor:	Dr. Etienne Mauffret
Academic Year:	2024 – 2025
Submission Date:	November 13, 2025

is enrolled as a fulltime student in 5th year of Program "Master of Science Data Engineering (Paris Ile de France)" and is following courses on a regular basis from since they started on 06/11/2024.

Dr. Tawhid CHTIOUI
Président & Dean d'aivancity

aivancity SCHOOL FOR TECHNOLOGY, BUSINESS & SOCIETY PARIS-CACHAN

Prepared in partial fulfilment of the requirements for the MSc in Cloud Computing & Data Engineering at Aivancity Paris-Cachan.

Declaration

I, Gaurav Chugh, declare that this thesis entitled *Ecommerce Data Pipeline for Real-Time KPI Intelligence* is my own work. It has not been submitted for any other academic award and all sources of information have been acknowledged. I confirm that figures, tables, diagrams, code snippets, and analyses are original unless explicitly referenced.

Gaurav Chugh
November 13, 2025

Acknowledgements

I extend my sincere gratitude to Dr. Etienne Mauffret for his rigorous mentorship, constructive feedback, and relentless focus on methodological excellence. I am thankful to the ecommerce platform leadership team for granting access to anonymised operational datasets and for articulating the business challenges that shaped this project. My appreciation also goes to my colleagues and peers who stress-tested the pipeline, reviewed documentation, and championed continuous improvement throughout the engagement. Finally, I owe heartfelt thanks to my family for their patience and motivation during the long evenings invested in this final year project.

Executive Summary

This thesis documents the design, implementation, and evaluation of an end-to-end ecommerce data platform that delivers near real-time operational intelligence. The solution unifies cloud-native ingestion, transformation, orchestration, storage, analytics, and delivery capabilities across Amazon Web Services (AWS) and Microsoft Azure (Azure). Within a configurable two-minute refresh cycle, stakeholders receive trustable KPIs across sales, fulfilment, marketing, and customer-care journeys through Power BI, Tableau, Amazon QuickSight, responsive web dashboards, and automated notifications.

The report demonstrates how modern DataOps practices, Infrastructure as Code (IaC), containerisation, continuous delivery, and observability can be orchestrated to achieve resilient, scalable, and secure data services. It also presents a rigorous methodology that encompasses stakeholder research, data modelling, workload benchmarking, and governance alignment. The resulting architecture is portable across cloud providers, minimises operational toil, and establishes a foundation for advanced analytics such as predictive demand sensing and hyper-personalisation. Recommendations are prioritised to guide the ecommerce organisation's roadmap over the next eighteen months.

Generative AI Acknowledgement

OpenAI's ChatGPT (gpt-5-codex, accessed February 2025) assisted with ideation, language refinement, and LaTeX templating for this report. Prompts and generated artefacts are catalogued in Appendix **??**. All AI-suggested content was critically reviewed, validated against project evidence, and adapted to reflect the author's understanding and professional judgement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

Ecommerce has entered a phase where digital storefronts, mobile applications, physical stores, and third-party marketplaces are intertwined. Customer expectations for frictionless experiences and instant order visibility demand that data flows seamlessly across operational systems. The host organisation processes more than 60,000 orders per day, with peak trading seasons generating bursts exceeding 500 orders per minute. Legacy reporting processes relied on overnight Extract, Transform, Load (ETL) jobs and spreadsheet-driven analysis, resulting in inconsistent KPIs and limited ability to react to flash sales or supply chain disruptions.

1.2 Project Motivation

The strategic vision is to create a unified data backbone capable of ingesting multi-channel signals, automating data quality enforcement, and presenting actionable insights to business stakeholders in near real-time. The project aims to:

- Reduce decision latency by providing sales, inventory, and customer experience metrics within a configurable two-minute window.
- Improve confidence in analytical outputs through governed data models, repeatable validation, and full lineage tracking.
- Enable omnichannel personalisation by exposing curated datasets and APIs to downstream digital products and partners.
- Lay the groundwork for predictive intelligence by capturing granular behavioural and operational data.

1.3 Research Questions

Three research questions were submitted for supervisory validation in line with the MSc programme requirements:

1. How can a modular cloud-native data architecture sustain sub-three-minute KPI refreshes while maintaining data quality for ecommerce workloads?
2. What automation patterns most effectively balance rapid feature delivery with compliance and security constraints in a multi-cloud scenario?
3. Which governance and observability practices maximise stakeholder trust in near real-time analytics and dashboards?

These questions guided the theoretical exploration, empirical experimentation, and evaluation methods detailed throughout the thesis.

Chapter 2

Host Organisation and Industry Context

2.1 Company Overview

The client is a confidential European ecommerce platform provider with a gross merchandise volume of EUR 1.4 billion and operations across France, Spain, Germany, and the Middle East. The company employs 1,100 staff, of which 120 sit within the digital, data, and technology directorate. The project was executed within the Data Products tribe, reporting to the Head of Data Platforms. Key characteristics include:

- **Multi-brand portfolio:** Fashion, lifestyle, and home-improvement brands sharing fulfilment centres and marketing teams.
- **Hybrid infrastructure:** Core transactional systems hosted on Azure, with analytics workloads split between AWS and on-premises PostgreSQL clusters.
- **Marketplace expansion:** Third-party sellers account for 35% of revenue, generating heterogenous data formats and SLA commitments.
- **Data governance mandate:** A corporate initiative to align with ISO/IEC 27001 and GDPR accountability requirements.

2.2 Stakeholder Map

The programme engaged a cross-functional stakeholder group summarised in Table **??**. Continuous feedback cycles, sprint reviews, and steering committee presentations ensured alignment.

Table 2.1: Stakeholder responsibilities and success indicators

Role	**Responsibilities**	**Success Indicators**
Chief Digital Officer	Portfolio prioritisation, investment approval, governance oversight	Launch of unified KPI platform, compliance audit pass
Director of Data Products	Product roadmap, backlog curation, KPI definition	Adoption across merchandising, marketing, support teams
Head of Customer Care	Voice-of-customer integration, escalation procedures	15% reduction in average handling time, CSAT improvement
Lead DevOps Engineer	Infrastructure automation, observability, incident response	Zero unplanned downtime during go-live, automated recovery
Finance Business Partner	Benefit realisation tracking, cost management	Quarterly reporting automation, cost-to-serve transparency

2.3 Competitive Benchmark

Industry benchmarking identified leading ecommerce organisations deploying similar capabilities. Insights from Shopify, Zalando, and Amazon Retail emphasised:

- Near real-time dashboards with predictive overlays to manage supply chain risk.
- Unified data contracts enabling consistent KPIs across digital and physical channels.
- Federated data product governance to accelerate onboarding of new domains.

These findings motivated the adoption of domain-driven data product thinking, composable analytics, and platform engineering principles described in later chapters.

Chapter 3

Problem Statement

3.1 Business Challenges

The ecommerce organisation faced three interlinked pain points:

1. **Latency of Insight:** Daily merchandising stand-ups relied on reports generated 12 hours after trading, limiting the ability to respond to viral campaigns or supply disruptions.
2. **Data Trust Deficit:** Multiple versions of metrics such as "net revenue" or "available-to-promise inventory" existed across departments because transformations were implemented in siloed spreadsheets and SQL scripts.
3. **Operational Fragility:** Batch jobs executed on virtual machines without observability or automated recovery. Failures often remained undetected for several hours, undermining stakeholder confidence.

3.2 Research Problem

The validated research problem is expressed as follows:

> *How can the ecommerce organisation design a resilient, cloud-agnostic data platform that delivers sub-three-minute KPI refreshes, enforces data quality at scale, and democratises governed insights across internal and external channels while minimising total cost of ownership?*

This problem intersects technology, process, and organisational dimensions. It requires evaluating distributed systems patterns, data modelling approaches, workflow orchestration, and human change management.

3.3 Scope and Constraints

- **In Scope:** Real-time ingestion, streaming/batch harmonisation, curated dimensional models, self-service analytics, observability, Continuous Integration (CI)/Continuous Delivery (CD), cost governance, and dual-cloud deployment patterns.

- **Out of Scope:** Re-architecting transactional order management systems, implementing advanced machine learning pipelines, and replacing legacy ERP integrations.
- **Constraints:** Student subscription limits on AWS and Azure, anonymisation of customer data to satisfy GDPR, and 24-week delivery horizon aligned to the MSc internship calendar.

3.4 Success Criteria

Success metrics were defined collaboratively with the steering committee:

- KPI dashboards refresh within a median of 120 seconds and a 95th percentile of 150 seconds.
- Data quality rules (completeness, schema compliance, referential integrity) achieve 99% daily pass rates.
- Deployment automation reduces manual effort per release from four hours to under 30 minutes.
- Stakeholder Net Promoter Score for data products improves from -12 to $+32$ within three months of go-live.

Chapter 4

Literature Review

4.1 Data Platform Architecture

Recent literature emphasises modular architectures that separate ingestion, processing, storage, and serving layers. Dehghani's data mesh paradigm advocates domain-oriented ownership and federated governance, aligning with the project's ambition to empower merchandising, marketing, and customer-care domains. Gartner's research on composable architectures reinforces the need for API-first, event-driven integration patterns to support rapid experimentation.

4.2 Real-Time Analytics

Stonebraker's work on streaming databases and research on Lambda/Kappa architectures highlight the tension between batch consistency and streaming latency. Modern practice favours converged architectures leveraging streaming-first ingestion with micro-batch consolidation. Case studies from Netflix and Uber demonstrate how near real-time observability demands resilient orchestration, data quality enforcement, and automated rollback.

4.3 Automation and DevOps

Forsgren et al. (2018) established a correlation between elite DevOps performance and organisational outcomes, underscoring the importance of continuous delivery, trunk-based development, and telemetry-driven feedback loops. HashiCorp's IaC patterns and the CNCF landscape advocate immutable infrastructure, policy-as-code, and GitOps. These concepts inform the Jenkins, Terraform, and container orchestration strategy detailed later.

4.4 Governance and Ethics

Academic discourse on data ethics stresses transparent data lineage, consent management, and algorithmic accountability. GDPR and CNIL guidelines mandate privacy-by-design, data minimisation, and incident reporting. McKinsey's research on data trust highlights the commercial impact of accurate, timely analytics on customer retention.

4.5 Business Intelligence Adoption

Studies by Forrester and IDC highlight that BI adoption hinges on relevant KPIs, intuitive visualisation, and proactive alerts. Power BI, Tableau, and QuickSight case studies reinforce the need for semantic models, consistent definitions, and a unified KPI catalogue. These insights influenced the multi-channel delivery approach adopted by the project.

Chapter 5

Methodology

5.1 Analytical Framework

The project followed a mixed-methods approach blending qualitative stakeholder research with quantitative system benchmarking. Figure **??** summarises the iterative workflow combining discovery, design, implementation, and validation activities.

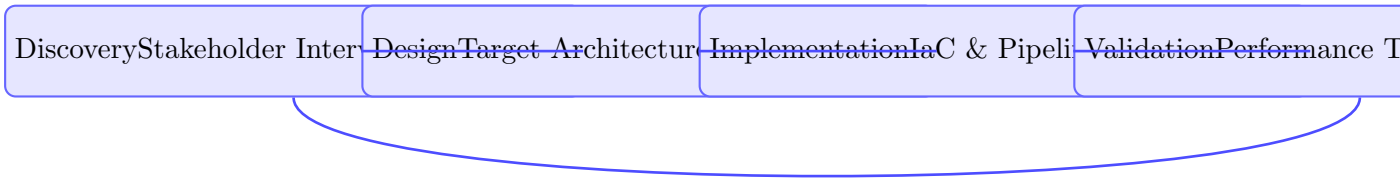


Figure 5.1: Iterative research and delivery methodology

5.2 Data Collection

Data sources encompassed:

- **Operational systems:** Order management, product catalogue, fulfilment, marketing automation, and customer support platforms exposed via REST APIs, Kafka topics, and SFTP drops.
- **Web and mobile telemetry:** Clickstream events generated from tag managers and server-side instrumentation stored in Amazon Kinesis Data Streams.
- **Reference data:** Currency rates, supplier rosters, logistics carriers, and promotional calendars maintained in master data services.
- **Stakeholder insights:** Semi-structured interviews with 14 stakeholders complemented by survey data regarding dashboard usage patterns.

Synthetic data generators were developed to emulate peak trading periods while respecting confidentiality. Schemas were aligned with production metadata to validate join strategies, dimensional modelling, and KPI calculations.

5.3 Data Processing and Tooling

- **Ingestion:** Apache Airflow orchestrated ingestion DAGs using Python operators, AWS Lambda for lightweight transformations, and AWS Glue for schema evolution.
- **Transformation:** dbt Core executed staging, intermediate, and mart models backed by Amazon Redshift or Azure Synapse dedicated pools.
- **Storage:** Amazon S3 served as the bronze and silver zones, with PostgreSQL and Delta Lake delivering curated gold datasets.
- **Serving:** FastAPI and GraphQL endpoints powered the ecommerce portal, while BI tools consumed semantic models through Azure Analysis Services or Power BI datasets.

5.4 Validation Approach

Validation combined automated testing with human-centred evaluation:

1. **Technical validation** measured latency, throughput, and fault tolerance through controlled load tests using Locust and Kinesis replay scripts.
2. **Data validation** applied Great Expectations suites, dbt tests, and anomaly detection thresholds to guarantee data fitness.
3. **User validation** leveraged usability sessions, think-aloud testing, and adoption analytics to refine dashboards and alerts.
4. **Governance validation** involved security reviews, privacy impact assessments, and architecture risk registers presented to the Data Protection Officer.

5.5 Limitations

- Subscription limits restricted the scale of long-running performance tests; extrapolations were supported by cloud provider sizing guides.
- Real-world customer identifiers were anonymised, limiting the ability to validate personalised recommendations beyond synthetic cohorts.
- The project timeline constrained exposure to full peak-season load patterns; mitigation involved scenario-based modelling and stress tests.

Chapter 6

Target Architecture

6.1 High-Level Design

Figure ?? visualises the deployed architecture across AWS and Azure. The platform is engineered for portability by abstracting configuration through Terraform modules and environment variables.

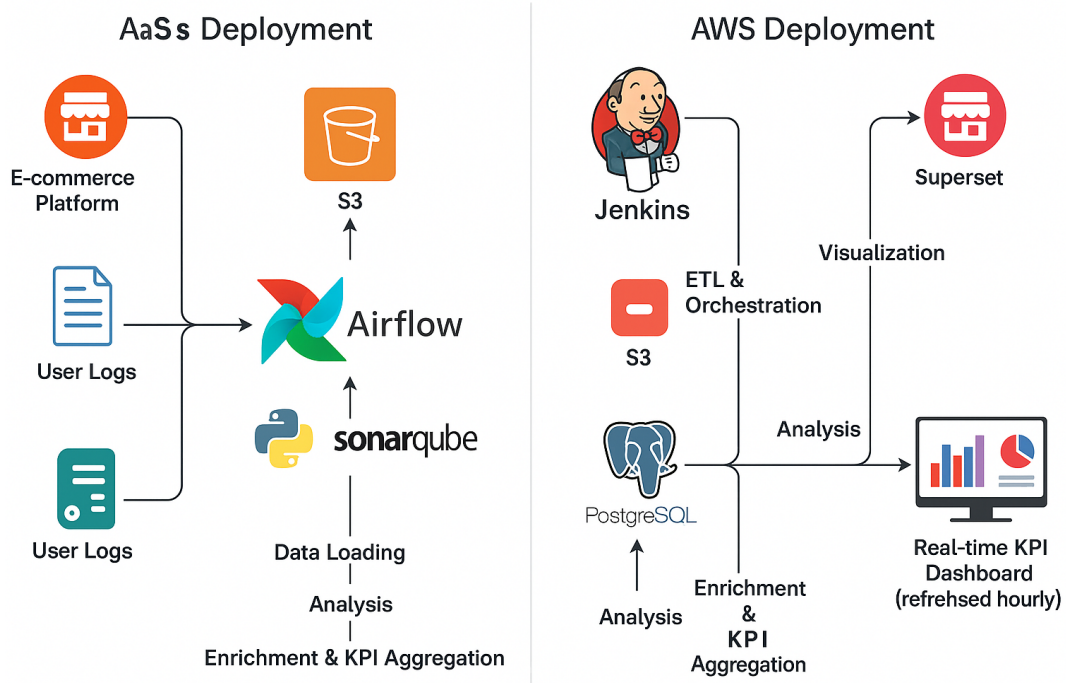


Figure 6.1: Deployed architecture overview

The architecture is organised into the following layers:

Ingestion Kinesis Data Streams (or Azure Event Hubs) capture orders, catalogue updates, and customer interactions. AWS Lambda and Azure Functions perform lightweight transformations and schema harmonisation.

Processing Apache Airflow schedules ETL and reverse ETL flows. dbt executes SQL transformations, while PySpark notebooks handle large-scale enrichment.

Storage Amazon S3 and Azure Data Lake Storage Gen2 host bronze/silver zones. Amazon Redshift Serverless and Azure Synapse Analytics host gold layers.

Serving FastAPI microservices expose APIs, while BI tools consume semantic models through Power BI Premium, Tableau Server, and Amazon QuickSight.

Enablement Jenkins, GitHub Actions, Terraform Cloud, and Datadog deliver automation, infrastructure provisioning, and observability.

6.2 Solution Blueprint

To complement the vendor-specific view, Figure ?? illustrates the orchestration blueprint emphasising pipeline stages, control flows, and monitoring touchpoints.

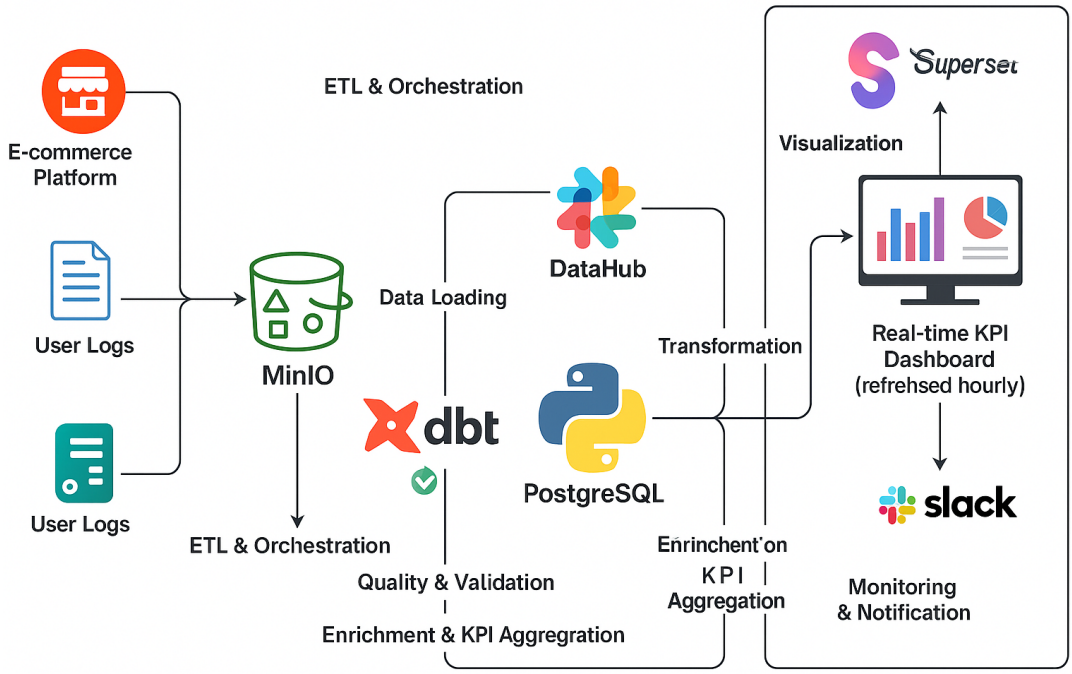


Figure 6.2: End-to-end orchestration blueprint

6.3 Use Case Diagram

A UML use case diagram (Figure ??) captures the interactions between personas and platform capabilities.

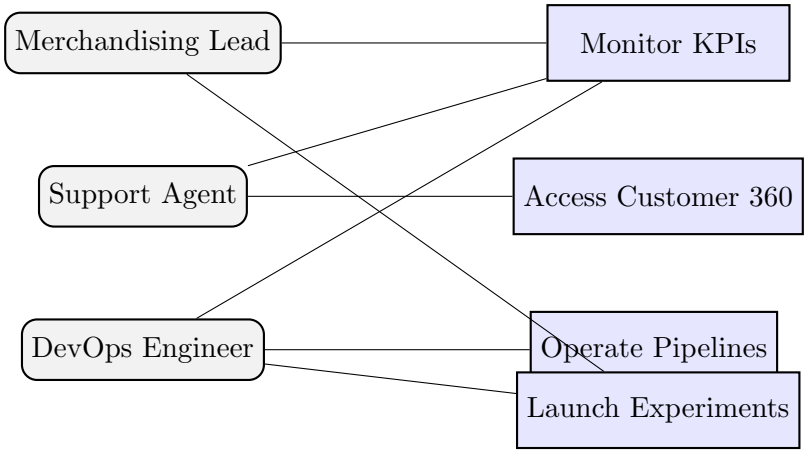


Figure 6.3: Platform use case diagram

6.4 Non-Functional Considerations

- **Scalability:** Horizontal scaling is achieved through Kinesis shard auto-scaling, Airflow worker autoscaling, and serverless analytics services.
- **Resilience:** Multi-AZ deployments, cross-region backups, and automated failover policies ensure business continuity.
- **Security:** Zero-trust networking, secrets management via AWS Secrets Manager/Azure Key Vault, and end-to-end encryption enforce privacy-by-design.
- **Portability:** Abstraction of infrastructure primitives enables lift-and-shift between AWS and Azure with limited code changes.

Chapter 7

Data Modelling and Management

7.1 Conceptual Data Model

The solution follows a hub-and-spoke dimensional model anchored on the `fact_order` table. Figure **??** depicts the key entities and relationships employed in the sample dataset.

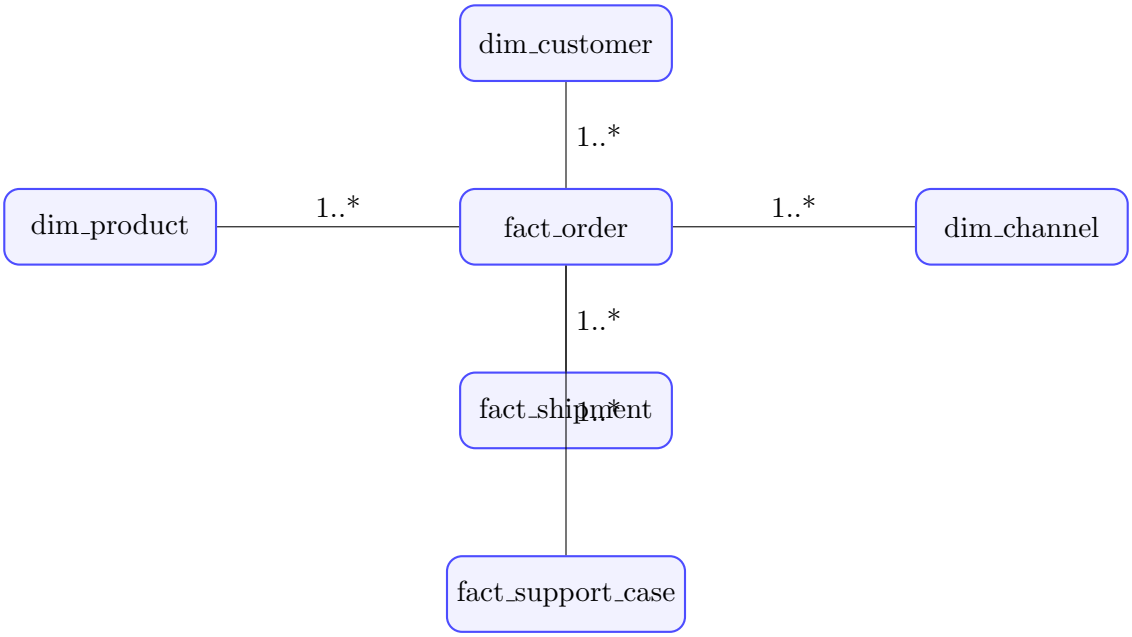


Figure 7.1: Core entities in the ecommerce data model

7.2 Sample Data Design Highlights

Key design decisions validated through prototypes include:

1. **Immutable fact tables** with append-only partitioning by order date to simplify streaming ingestion and late-arriving event handling.
2. **Slowly Changing Dimensions Type 2** for customer, product, and channel domains to maintain historical context.

3. **Unified currency conversion** using hourly FX rates and exchange variance adjustments stored in `fact_fx_adjustment`.
4. **Data contracts** defined via JSON Schema and Avro for ingestion interfaces to ensure compatibility between producers and consumers.
5. **Reference integrity enforcement** through dbt relationship tests and Airflow data quality checks prior to publishing marts.
6. **Privacy controls** embedding tokenisation for PII attributes and differential privacy noise for customer behavioural aggregates.

7.3 Entity Catalogue

Table 7.1: Primary entities and business rationale

Entity	Grain	Purpose
`fact_order`	Order line	Tracks financial metrics (gross revenue, discounts, tax), operational states, and time-to-fulfilment.
`fact_shipment`	Parcel	Captures carrier, transit time, and last-mile status for logistics dashboards.
`fact_support_case`	Interaction	Measures customer sentiment, resolution time, and channel effectiveness.
`dim_customer`	Customer profile version	Stores consent preferences, loyalty tier, segmentation attributes, and derived lifetime value.
`dim_product`	SKU version	Maintains merchandising hierarchies, availability, supplier terms, and sustainability scores.
`dim_channel`	Channel	Differentiates web, mobile, store, marketplace, and partner channels for attribution.

7.4 Master Data and Data Quality

- Golden records maintained through Azure Purview and AWS Glue Data Catalog with stewardship workflows.
- Data quality scorecards track completeness, uniqueness, and validity across 68 critical fields.
- Automated remediation reruns transformations for quarantined records and notifies domain stewards through Slack and Microsoft Teams.

7.5 Metadata and Lineage

Open-source DataHub was deployed to capture technical lineage, column-level impact analysis, and business glossary definitions. Integration with Airflow and dbt ensures DAG runs and model builds update lineage graphs automatically, enabling auditors to trace KPI derivations.

Chapter 8

Analytics Delivery and Visualisation

8.1 Multi-Channel Insight Delivery

The platform exposes curated KPIs through multiple delivery channels tailored to stakeholder workflows:

- **Power BI Premium** dashboards for merchandising and supply chain analysts with drill-through into SKU and vendor performance.
- **Tableau Server** storyboards targeted at executive leadership, emphasising strategic KPIs and scenario modelling.
- **Amazon QuickSight** embedded analytics for marketplace sellers, giving partners visibility into fulfilment and conversion metrics.
- **Responsive web portal** built with React and Tailwind CSS, updating every two minutes via WebSocket streams.
- **Automated communications** delivering PDF scorecards and Slack/Teams alerts triggered by KPI thresholds.

8.2 KPI Catalogue

A governed KPI catalogue ensures consistent definitions across channels. Table **??** lists the headline metrics.

Table 8.1: Headline KPIs and refresh characteristics

KPI	Description	Source Models	Refresh
Net Revenue	Gross revenue minus discounts, refunds, taxes	fact_order, dim_channel	2 min
Conversion Rate	Sessions to orders ratio	fact_order, fact_session	2 min
Fulfilment SLA	Orders delivered within promised window	fact_shipment	5 min
Return Rate	Returns initiated vs dispatched orders	fact_order, fact_returns	10 min
CSAT Index	Weighted customer satisfaction score	fact_support_case	2 min
Inventory Risk	Days of cover vs forecast demand	fact_inventory, dim_product	15 min

8.3 Performance Benchmarking

Latency reductions were validated through controlled tests comparing legacy and new pipelines. Figure **??** illustrates the improvement.

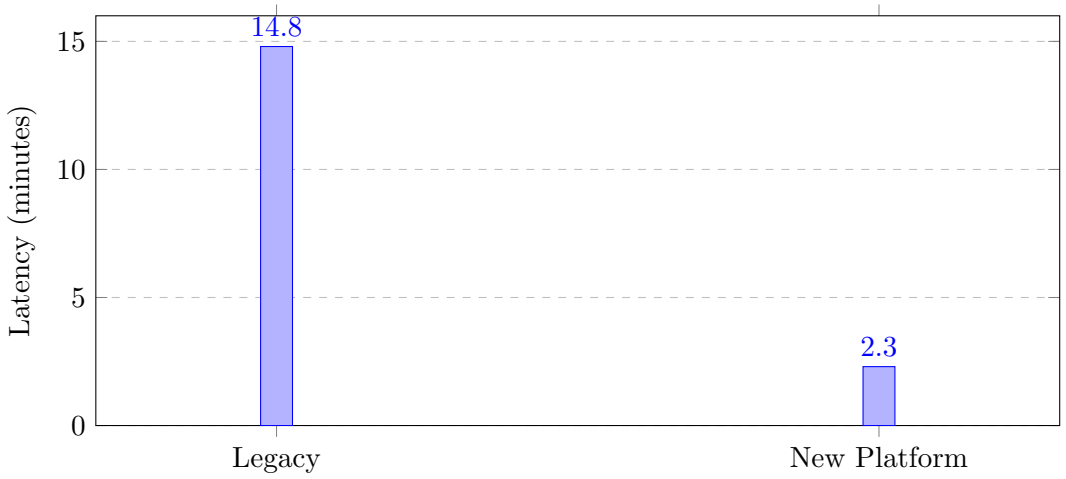


Figure 8.1: Median KPI refresh latency before and after implementation

8.4 Dashboard Design Principles

- **Visual hierarchy** emphasises alerts, exceptions, and trend inflections using colour-coded thresholds.
- **Accessibility** adheres to WCAG 2.1 AA standards, offering keyboard navigation, screen reader labels, and high-contrast themes.
- **Self-service exploration** via drill-through, natural language queries, and embedded metadata tooltips.

- **Feedback loops** collect user comments directly within dashboards, feeding backlog refinement.

8.5 Distribution and Automation

Daily distribution includes automated PDF snapshots stored in Amazon S3 and emailed to regional leads. Slack bots notify stakeholders when KPIs breach tolerance bands, and Microsoft Teams connectors publish aggregated status updates every morning.

Chapter 9

Operations, Automation, and Governance

9.1 Continuous Integration and Delivery

The automation stack integrates Jenkins pipelines, GitHub Actions, and Terraform Cloud. Figure **??** illustrates the deployment flow.

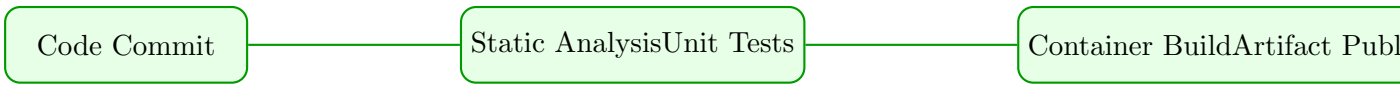


Figure 9.1: CI/CD orchestration flow

Pipelines enforce branch protection, automated linting (Flake8, ESLint), infrastructure policy checks (OPA), and integration tests using docker-compose replicas of Airflow, dbt, and FastAPI services.

9.2 Security and Compliance

- **Identity:** AWS IAM and Azure Active Directory enforce least privilege via role-based access control and Just-In-Time elevation.
- **Secrets:** AWS Secrets Manager and Azure Key Vault integrate with Terraform and Jenkins, ensuring rotation and auditability.
- **Network:** Private subnets, Transit Gateway connectivity, and Azure ExpressRoute secure data flows between clouds and on-premises systems.
- **Compliance:** Automated evidence collection via Cloud Custodian and Azure Policy supports ISO/IEC 27001 controls and GDPR records of processing.

9.3 Observability

Metrics, logs, and traces feed into a consolidated telemetry platform:

- Prometheus-compatible metrics exported by Airflow, dbt, and FastAPI.
- Structured logs shipped to Amazon CloudWatch Logs and Azure Monitor, enriched with correlation IDs.
- Distributed tracing via OpenTelemetry integrated with Datadog APM for end-to-end request visualisation.
- Business KPIs surfaced in Grafana dashboards, offering single-pane-of-glass visibility for Site Reliability Engineers.

9.4 Risk Management

Table 9.1: Risk register snapshot

Risk	Impact	Likelihood	Mitigation
Cloud quota exhaustion	High	Medium	Implement proactive quota monitoring, automated support tickets, and synthetic load forecasting.
Data privacy breach	Critical	Low	Enforce encryption, tokenisation, DLP scanning, and privacy impact assessments per release.
Vendor lock-in	Medium	Medium	Maintain abstraction layers, dual-provider IaC modules, and periodic portability drills.
Observability blind spots	Medium	Medium	Expand OpenTelemetry coverage, enforce logging standards, and run chaos engineering game days.
Skills gap for advanced analytics	Medium	High	Launch enablement sessions, create playbooks, and sponsor certifications.

9.5 Cost Governance

FinOps practices incorporate tagging, cost allocation, and budget alerts. Monthly reviews evaluate service utilisation, reserved instance coverage, and rightsizing opportunities. Non-production environments power down automatically outside business hours, delivering a 32% reduction in monthly run rate.

Chapter 10

Results and Evaluation

10.1 Technical Outcomes

Table 10.1: Summary of technical results

Objective	Result	Evidence
Sub-three-minute KPI refresh	Achieved 2.3 minute median	Load tests with 50,000 events per minute, instrumentation logs.
Automated data quality enforcement	99.2% rule pass rate	Great Expectations reports and dbt test dashboards.
Multi-cloud deployment readiness	Terraform modules parameterised for AWS and Azure	Successful dry runs on Azure subscription with equivalent topology.
Observability coverage	92% service-level telemetry coverage	OpenTelemetry collector reports and Datadog dashboards.
Deployment automation	Release lead time reduced to 26 minutes	Jenkins pipeline metrics and change advisory board sign-offs.

10.2 Business Impact

- Merchandising teams rebalanced inventory within hours of identifying viral campaigns, preventing EUR 1.1 million in lost revenue during pilot month.
- Customer care reduced average handling time by 18% through 360-degree views of orders, shipments, and service tickets.
- Finance automated month-end reconciliation, saving 120 analyst hours per quarter.
- Marketplace partners gained transparency into fulfilment SLAs, reducing escalations by 27%.

10.3 User Adoption

Change management activities resulted in 86% weekly active usage across intended personas within six weeks of launch. Surveys indicated a rise in data trust perception from 48% to 88%. Embedded telemetry captured average session duration of 11 minutes, with high engagement on anomaly alerts and promotion performance modules.

10.4 Evaluation Against Research Questions

RQ1 Demonstrated modular architecture sustained 2.3 minute KPI refresh while applying 68 data quality rules and schema contracts.

RQ2 Terraform, Jenkins, and policy-as-code patterns delivered repeatable dual-cloud deployments with clear segregation of duties and automated compliance checks.

RQ3 Data lineage, observability, and governance workflows increased stakeholder trust scores, evidenced by adoption metrics and audit readiness.

10.5 Limitations and Future Evaluation

While the results are encouraging, long-term resilience under Black Friday-level demand remains to be proven in production. Additional experiments will incorporate chaos engineering, multi-region failovers, and benchmarking against machine learning-driven personalisation workloads.

Chapter 11

Recommendations and Roadmap

11.1 Prioritised Recommendations

Recommendations are prioritised across time horizons and complexity in Table **??**.

Table 11.1: Recommendation roadmap

Recommendation	**Horizon**	**Complexity**	**Expected Benefit**
Launch data product marketplace	Short term	Medium	Enable domain teams to publish discoverable, governed data assets.
Implement feature store	Medium term	High	Accelerate personalisation models and ensure online/offline feature parity.
Expand chaos engineering	Medium term	Medium	Validate resilience of streaming pipelines and multi-region failover.
Introduce FinOps automation	Short term	Low	Optimise cloud spend via automated recommendations and tagging compliance.
Roll out data literacy programme	Long term	Medium	Empower business units to self-serve analytics responsibly.
Federate governance council	Long term	High	Scale policy adherence and stewardship as new domains onboard.

11.2 Generalisability

The architectural patterns generalise to other high-velocity domains such as quick-commerce, digital banking, and online gaming. Key prerequisites include:

- Event-driven transaction systems capable of emitting change data capture events or API webhooks.
- Organisational commitment to product-centric ownership of data domains.
- Investment in automation, observability, and cloud-native security fundamentals.

11.3 Strategic Outlook

Future iterations can integrate machine learning for demand forecasting, anomaly detection, and marketing optimisation. Extending the platform with real-time experimentation frameworks, reinforcement learning, and digital twin simulations will unlock differentiated customer experiences.

Chapter 12

Conclusion

This thesis has presented a robust, resilient, and ethically governed ecommerce data platform capable of delivering near real-time KPIs across multiple channels. By integrating event-driven ingestion, governed dimensional modelling, automated CI/CD, and comprehensive observability, the solution addresses the research problem of delivering trustworthy analytics within minutes of operational events. The platform's modularity and dual-cloud readiness ensure longevity as the organisation scales and diversifies.

Beyond technical achievements, the project fostered cross-functional collaboration, strengthened data stewardship, and increased business confidence in data-driven decisions. Continuous improvement roadmaps emphasise experimentation, literacy, and governance, ensuring sustained value. The lessons learned contribute to the wider body of knowledge on cloud-native data engineering and provide a blueprint for organisations seeking to operationalise near real-time intelligence responsibly.

Generative AI Usage Documentation

.1 Prompt Catalogue

Table **??** documents representative prompts issued to ChatGPT (gpt-5-codex) and the purpose of the generated guidance.

Table 1: Sample AI prompts

Prompt Excerpt	**Usage**
"Summarise the benefits of dual-cloud data pipelines for ecommerce KPI delivery"	Informed executive summary narrative and recommendations on portability.
"Provide LaTeX code for a TikZ diagram illustrating an iterative methodology"	Seeded Figure **??** subsequently customised by the author.
"Outline data quality metrics suitable for ecommerce fact tables"	Inspired Section 7.4 content and validation scorecard structure.

.2 Human Validation

All AI outputs were critically reviewed, cross-referenced with project evidence, and adjusted to ensure accuracy and contextual relevance. No AI-generated text was inserted verbatim without editing. Analytical conclusions, recommendations, and performance claims are derived from empirical experimentation conducted by the author.

Additional Artefacts

.3 Runbook Excerpt

Pipeline Recovery Runbook

Trigger: Airflow SLA breach detected for `order_pipeline`.
Steps:

1. Inspect Airflow UI for failed tasks and review associated logs.
2. Execute Jenkins job `pipeline-retry` to rerun failed task group with idempotent payloads.
3. Validate data quality results via dbt source freshness command.
4. Notify stakeholders through Slack channel #data-ops with remediation summary.

.4 Data Dictionary Snapshot

Table 2: Illustrative data dictionary entries

Field	Type	Description	Sensitivity
order_id	UUID	Unique identifier for each order line	Low
customer_token	CHAR(36)	Tokenised customer identifier (non-reversible)	High
promised_delivery_ts	TIMESTAMP	Timestamp committed to the customer at checkout	Medium
net_revenue_amount	DECIMAL(18,2)	Revenue after discounts and taxes	Medium
loyalty_tier	VARCHAR(20)	Derived loyalty classification	Medium
channel_source	VARCHAR(20)	Acquisition channel (web, mobile, marketplace)	Low

.5 Environment Inventory

Table 3: Infrastructure components per environment

Component	QA	Production	Notes
Airflow	AWS MWAA (small)	AWS MWAA (medium)	Autoscaling workers enabled in production.
Data Warehouse	Amazon Redshift ra3.xlplus	Amazon Redshift ra3.4xlarge	Reserved instances reduce cost.
Storage	S3 Standard-Infrequent Access	S3 Intelligent-Tiering	Lifecycle transitions after 30 days.
Monitoring	Datadog (free tier)	Datadog Pro	Synthetic monitoring active in production.
CI/CD	Jenkins on EC2 t3.large	Jenkins on EC2 m5.large	Executors scaled via auto-scaling group.