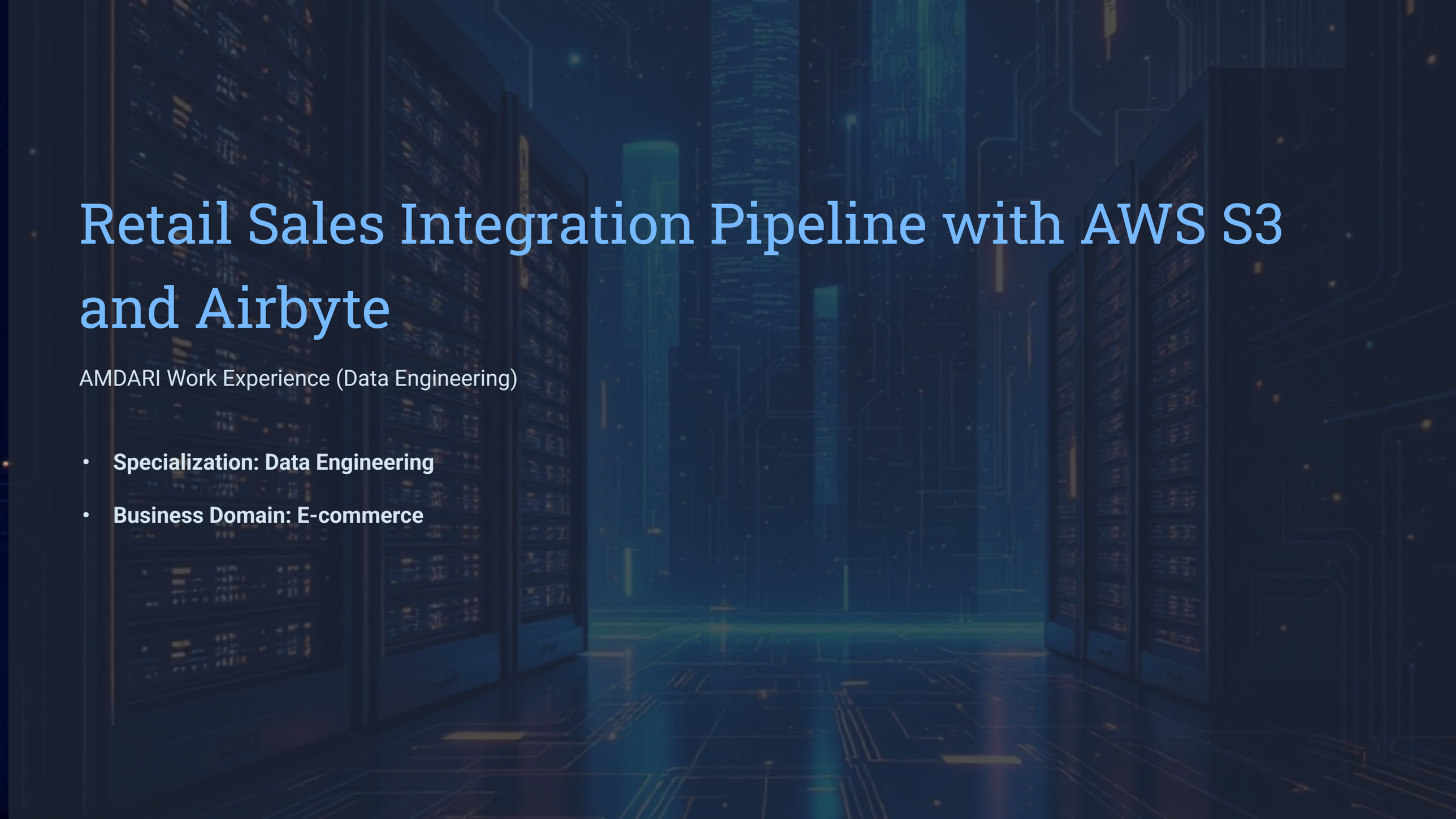


Retail Sales Integration Pipeline with AWS S3 and Airbyte

AMDARI Work Experience (Data Engineering)

- **Specialization: Data Engineering**
- **Business Domain: E-commerce**

Company Overview

Retailio is a rapidly growing mid-sized retail and e-commerce company operating across multiple cities. The company records thousands of daily sales transactions from its branches and online channels. Each regional branch currently exports sales, customer, and inventory data as CSV reports that are manually uploaded into a shared drive by branch managers.

As Retailio scales, management has realised that these manual data flows are inefficient, time-consuming, and unreliable. To improve decision-making, the company has decided to move toward a cloud-based data integration and analytics system where all data is stored centrally, cleaned automatically, and made available for business intelligence teams in near real-time.

AMDARI®

Current Business Challenge

Retailio's data and operations teams face multiple pain points in their current workflow:

Manual Data Uploads

Every branch exports daily reports manually. This process leads to inconsistent file naming conventions, missed uploads, and duplication of data.

Data Silos

Each branch maintains its own records, which makes company-wide performance analysis very difficult.

Limited Visibility

Business analysts rely on stale reports from previous weeks, making it impossible to perform real-time sales analysis or forecast inventory needs.

High Error Rate

Manual copying and merging of spreadsheets have introduced data inaccuracies, leading to poor insights and mistrust in reports.

Lack of Central Repository

Without a unified data lake or warehouse, there's no single source of truth for executive decision-making.

The management team wants a lightweight yet scalable cloud data pipeline that can integrate multiple data sources, store them in a central data lake, and automatically push the data to a warehouse for analytics and visualisation.

Project Objectives

By the end of this project, you will:

1. **Design and Implement a Cloud-Native Data Pipeline**

   Create an end-to-end workflow from raw data generation to automated integration into a cloud data warehouse.

2. **Build a Central Data Lake**

   Configure AWS S3 as the storage layer for all raw and semi-processed data, with versioning and folder hierarchy (bronze layer).

3. **Use Airbyte for Automated Integration**

   Connect the S3 source to a warehouse destination and configure data sync jobs to run on a defined schedule.

4. **Learn ELT Concepts in Practice**

   Experience the Extract-Load-Transform (ELT) workflow, where data is loaded before transformation to optimise performance.

5. **Develop Query Validation Skills**

   Write SQL queries to validate data consistency between the data lake and the warehouse.

Tech Stack

Python & Faker
Used to generate and simulate retail data such as sales, customers, and product records.

AWS S3
Serves as the central data lake where all raw and structured data files are stored.

Airbyte
Handles automated extraction from S3 and loading into the data warehouse (ELT process).

MotherDuck
Acts as the target data warehouse for analytics and reporting.

CSV / JSON
Lightweight file formats used for storing and transferring data between systems.

Airbyte UI & SQL Queries
Tools for monitoring pipeline health, validating data, and ensuring successful ingestion.

Project Scope and Deliverables

Data Simulation and Upload

- Sales, products, and customers dataset will be provided
- Each dataset will contain 10,000+ records.
- Upload datasets to AWS S3 using the boto3 SDK.
- Define structured folder hierarchy
- Implement file versioning and lifecycle policies on the S3 bucket.

Airbyte Configuration

- Using Airbyte Cloud
- Add S3 as the data source connector, defining the bucket name, region, and path prefix.
- Add MotherDuck as the destination connector, authenticating via credentials.
- Configure the pipeline to extract new data from S3 daily.
- Load into staging tables within the warehouse.
- Maintain schema consistency between runs.

AMDARI®

Warehouse Setup and Data Validation

Warehouse Setup

- Configure MotherDuck as the target destination in Airbyte using the authentication token.
- Create a target schema named retail_data to store datasets ingested from S3.
- Allow Airbyte to automatically create tables for each dataset
- Verify that Airbyte successfully writes to the target schema during the first sync.
- Use MotherDuck's SQL interface to explore ingested data and confirm schema structure.
- Ensure all data types and column mappings match the S3 source files.

Data Validation and Testing

- Confirm that the total record count in each MotherDuck table matches the source files in S3.
- Run validation queries to check for missing or null values in key fields.
- Perform schema consistency checks to verify column names, data types, and timestamp accuracy.
- Conduct data profiling queries to ensure values fall within expected ranges (e.g., quantity, price).
- Design analytical queries such as:
  - sales_summary – total sales per day or per region
- Review Airbyte logs after every sync to confirm successful ingestion and identify any errors.

Automation, Monitoring and Documentation

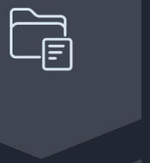
- Airbyte's built-in scheduler can handle daily syncs automatically.
- Logs from Airbyte's UI will serve as monitoring checkpoints for ingestion success/failure.

Documentation and Presentation

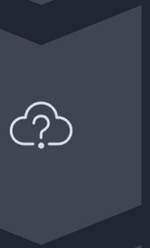
- Document every step: from S3 setup and permissions to Airbyte connection configuration.
- Include screenshots of successful syncs and SQL query outputs.
- Present the final workflow using a Medallion Architecture diagram.

AMDARI®

Data Architecture



Data Retrieval



AWS S3 Data Lake

(Bronze Layer – Raw Data)

MotherDuck Data Warehouse

(Silver/Gold Layer – Clean)

BI / Analytics Tools

AMDARI®

Success Criteria

1

Data Accuracy

Data ingested into the warehouse matches the records generated and uploaded to S3.

2

Automation

Airbyte syncs data automatically on schedule without manual intervention.

3

Scalability

The system can easily handle larger data volumes by adding new S3 files or connectors.

4

Transparency

Airbyte provides clear logs for every run, including success and failure details.

5

Query Readiness

Analysts can query the integrated data using SQL directly from the warehouse.

6

Documentation Quality

The final project deliverable includes a well-structured README or PDF outlining setup, architecture, and validation steps.