Zalando Case Study

Data Modeling & BI Platforms

Helen Karlsson

Table of Contents

About Zalando

Zalando is one of Europe's leading e-commerce platforms for fashion and lifestyle, offering a selection from over 6,000 brands. In 2023, Zalando processed 244 million orders and generated €10.1 billion in revenue. Zalando primarily operates in two business segments: B2C (Business-to-Consumer) and B2B (Business-to-Business).

- B2C means selling directly to consumers, where Zalando offers an online, multi-brand shopping experience to over 50 million active customers across 25 markets.

- B2B is aimed at other businesses and brands—through its partner program, Zalando allows brands to sell products on Zalando's platform. Zalando also provides logistics solutions via Zalando Fulfillment Solutions (ZFS), handling warehousing, packing, and shipping on behalf of its partners.

Zalando operates as a marketplace where customers can purchase products either directly from Zalando's own inventory or from third-party sellers via its partner program. This means the company combines its own retail operations with a marketplace model.

"As an online platform for fashion and lifestyle, we connect customers, brands and partners"
-Zalando

Zalando’s need for structured data modeling

Operating e-commerce at the scale Zalando does demands advanced data infrastructure, technology, and logistics solutions. Since the volume of data Zalando processes fluctuates dramatically over the year, peaking during events like Black Friday and other promotional periods, there is a clear need for a flexible and scalable data infrastructure.

To handle these fluctuations, systems must handle sudden spikes in traffic, transactions, and queries without degrading performance. This demands a dynamic architecture where capacity can scale up or down in real time, ensuring resources are used efficiently and costs stay under control. Additionally, the data model must be responsive, adapting to shifting customer behaviors to deliver fast, relevant recommendations, accurate inventory forecasts, and timely price updates from partners.

Key data flows

Zalando processes massive volumes of data, and their core data flows can be categorized into the following groups:

Customer Data & User Behavior
CRM data, order history, returns and customer support. Web- and app-behavior such as searches, cart activity and abandoned check-outs.

Product & Inventory
PIM data on products (descriptions, colors, sizes) and WMS data on stock levels and forecasts.

Orders & Transactions
Order details (products, payment, delivery) and transaction data (ID, method, amount).

Logistics & Fulfillment
Shipment status, delivery times, return handling and refunds.

Marketing
Insights from analytics and advertising tools, plus CRM- and email-campaign performance.

B2B & Finance
Partner data, contracts, sales volumes, revenue, billing and HR data.

Scope

For this case, I have scoped the data flows and objects to focus on the most central, and for our case-purposes, most “relevant” entities around product and order management. This selection is intended to illustrate an end-to-end view of a data model and pipeline for a company like Zalando.

Data model

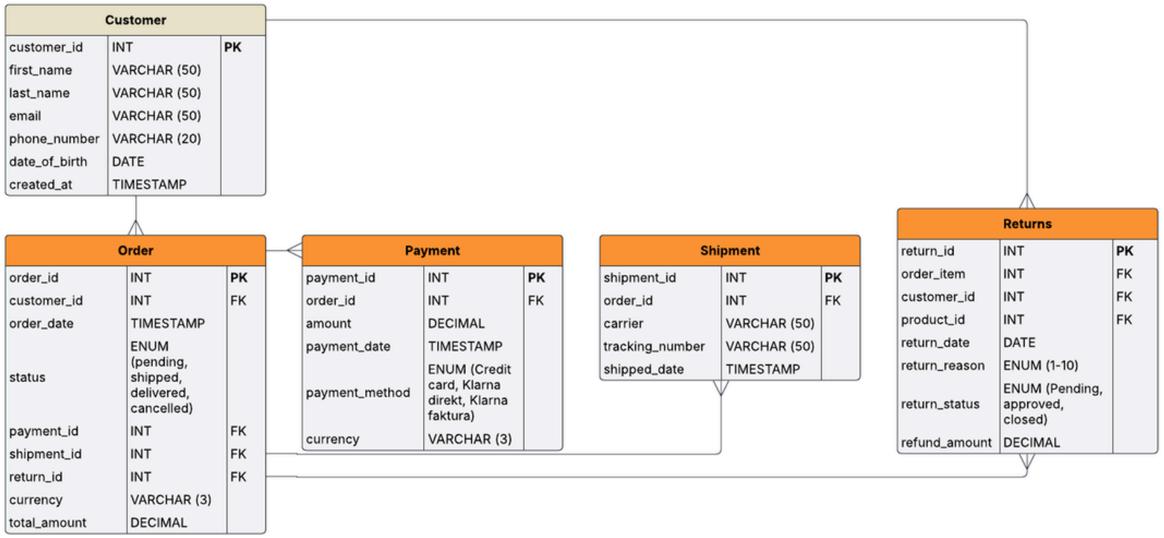
To build a data model for Zalando's order management, we start by defining one table per entity, an entity being any "thing" we want to track in our model. Each table then lists that entity's attributes (properties) along with their data types.

Relationships between tables are enforced via primary keys and foreign keys. By leveraging PK–FK constraints, we eliminate redundancy and store data in the most cost-effective, normalized way.

The diagram to the right illustrates a streamlined version of our upcoming model. Order, Payment, Shipment, and Returns are the main fact tables, while Customer serves as a key dimension table.

The relationships between these entities are:
Customer → Orders & Returns
A single Customer can place many Orders and have many Returns (1 Customer:N Orders/Returns), but each Order or Return belongs to exactly one Customer (N:1).

Order → Payments, Shipments & Returns
A single Order can generate multiple Payments, Shipments and Returns (1 Order:N Payments/Shipments/Returns), but each Payment, Shipment or Return is linked to exactly one Order (N:1).



Link to Lucidcharts 1 of 2: https://lucid.app/lucidchart/7f3e78c9-42b7-4c61-a24a-2817a5a2b65b/edit?invitationId=inv_f10f010e-9e91-4acc-960e-4d28cb50b101

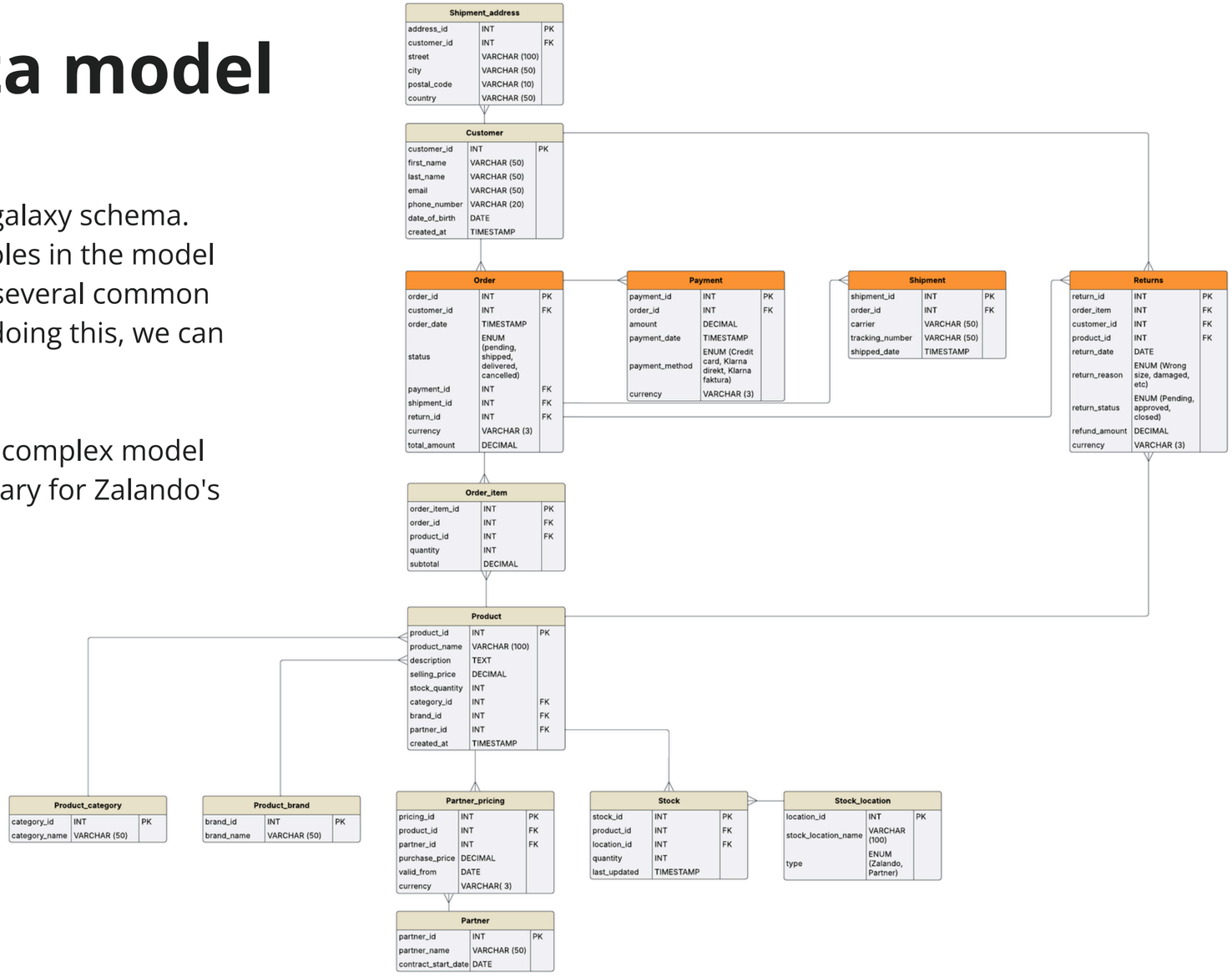
Structure of the data model

Zalando's data model is structured according to a galaxy schema. This is necessary because there are several fact tables in the model (Order, Payment, Shipment, Returns) that refer to several common dimension tables (e.g. Customer and Product). By doing this, we can avoid redundancy.

The galaxy schema allows us to implement a more complex model with multiple layers of dimensions, which is necessary for Zalando's business.

See the extended data model on the right:

Link toLucidcharts 2 of 2: https://lucid.app/lucidchart/7f3e78c9-42b7-4c61-a24a-2817a5a2b65b/edit?invitationId=inv_f10f010e-9e91-4acc-960e-4d28cb50b101



Business value

A data model like the one above enables efficient handling of large amounts of data and enables data-driven insights, which can help Zalando make smart decisions that provide business value. From the above model, it could, for example, be about analyzing how price-sensitive different product categories are or about investigating whether different transport companies are better/worse in different areas. Based on this, Zalando could offer a recommendation on transport choices in their checkout process.

Similarly, Zalando has developed a system to be able to recommend size choices to their customers for certain products. Using data, machine learning and technology, a system has been developed that takes into account several data points, such as the customer's previous size choice, reasons for returns of the current product, the product's dimensions and the model's size in the product images. Based on this, an algorithm has been developed that can notify Zalando’s customers whether the product is, for example, “Large in size”, “Small in size” or “We recommend size Medium, based on your previous purchases”.

By doing this, Zalando has reduced size-related returns by 10%, compared to products that do not have size recommendations. This is something that has enormous business value for companies like Zalando, and perhaps especially companies in the clothing industry, that are under high pressure to demonstrate strategies to reduce their CO2 impact.

Having a well-thought-out and structured data model is a prerequisite for being able to trust your data and enable support for business development and decision-making like the example above. It is important that the data model is flexible and scalable so that it can be adapted and expanded according to Zalando's needs.

Challenges in data modeling

Implementing a data model for Zalando comes with several challenges, especially given the size of the business and the large amount of data that needs to be managed. It requires a detailed data strategy, and an understanding that the scope of this work is more of an organizational/management question, rather than an isolated task for the IT department. It is a resource-intensive task that requires knowledge and understanding of the needs of the entire business.

The model needs to be adapted for extremely large amounts of transactions, handle connections to partner's product and inventory data, and be dynamic and adaptable based on traffic volumes. During periods such as Black Friday, christmas sales or other traffic-driving events, it is important that real-time processing for handling orders and transactions is stable. Here it becomes important that we use referring information that already exists in our data model, instead of collecting the same data in repeated places. An example of this is using Primary Keys and Foreign Keys.

Another challenge is managing data from different partners and third-party sources. There is a high risk that the format will differ from our own data, and we therefore need to have a structured process for transformation using ETL tools.

The data model should be tested at an early stage to detect any errors and areas for improvement early on. It is more cost-effective to invest resources in ongoing testing, than to have to redo an entire model after completion.

Big Data & Modern Data Stack

Big Data is a term that describes how companies handle very large amounts of data. Zalando handles huge amounts of data from customers, orders, inventory data, logistics and marketing on a daily basis. To be scalable and enable real-time analysis, Zalando needs to use modern Big Data technologies.

Modern Data Stack refers to the “toolbox” or set of cloud-based technologies and tools that Zalando uses to collect, store, transform and analyze data.

Big Data Challenges

Volume: Zalando handles huge amounts of data, millions of order lines, product information and customer interactions, which requires scalable storage and processing solutions.

Speed: Real-time updates of inventory status, transactions and customer behavior require fast and efficient technologies to ensure Zalando’s general functionality, but also to ensure a good experience for customers.

Variety: Data comes in both structured (orders, transactions) and unstructured form (click logs, reviews), from multiple data sources. This requires flexible data models, structured transformation/harmonization and advanced analytics.

Value: To create business insights, Zalando must be able to filter, analyze and visualize data in a way that contributes to better customer experiences and business decisions. The right data needs to be available to the right person at the right time.

Reliability: Ensuring data quality, consistency and compliance with privacy directives requires robust validation processes, security measures and data policies.

Big Data & Modern Data Stack

To address these challenges, Zalando can use a Modern Data Stack that handles ***distributed storage, distributed computing, data orchestration, ETL/ELT processes,*** and good conditions for analytics, ***Business Intelligence, and AI***.

Zalando's Modern Data Stack needs to be optimized for:

- Scalability and flexibility at high data volumes (e.g. during Black Friday).
- Ability to handle both structured and unstructured data (transactions, logs, customer behavior).
- Making data available quickly for analysis and personalization. The data needs to be operationalized, i.e. made available, to the right target group (team) at Zalando, at the right time, and in a format that is understandable.

Distributed storage
To handle both real-time and batch data, Zalando can use a combination of operational databases, a Data Lake and a Data Warehouse. Operational database (Database Management System) for OLTP (Online Transaction Processing) to handle transactional data. Using OLTP for this type of data is advantageous to meet requirements for real-time updates, high integrity level and to avoid redundancy as much as possible. For this, Amazon Aurora can be used, which is a relational database, this service is managed through Amazon RDS. For unstructured or semi-structured raw data, a Data Lake, Amazon S3, is used. And for structured processed data, we build a Warehouse using Amazon Redshift.

Distributed data processing
To process and analyze large amounts of data simultaneously, Zalando can use distributed (partitioned) data processing, where multiple servers process the data simultaneously. This enables faster analysis and scalability. Apache Spark can be used for both fast batch processing and stream processing for real-time analysis. AWS Lambda can filter, validate, and enrich data before sending it to the Data Lake.

Big Data & Modern Data Stack

Data Orchestration
To coordinate complex data flows, a tool like Apache Airflow and Amazon Cloudwatch can help schedule and monitor data pipelines, ensuring efficient and automated data processing. AWS Lambda can send alerts in case of failed runs.

ETL/ELT processes
Using tools like Amazon Glue, Zalando can extract, transform, and load data (ETL) or first load raw data and then transform it (ELT), enabling flexible and scalable data management.

Analysis and Business Intelligence
BI platforms like Tableau, Looker, or Power BI are used to create dashboards and reports that provide business insights based on customer behavior, sales trends, and logistics flows. With these tools, we can analyze processed data and visualize it for, for example, decision makers or product owners in different areas at Zalando. Here, adjustments can be made so that the right data is made available to the right person or department, at the right time.

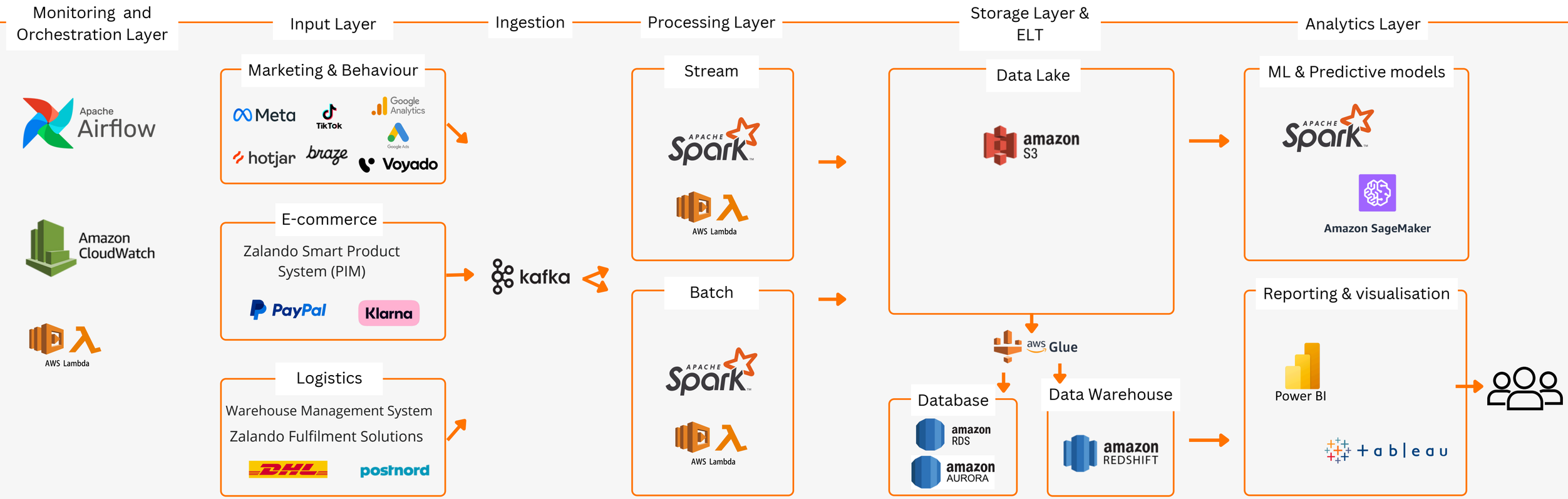
AI and Machine Learning
Using technologies like Amazon SageMaker, Zalando can build ML models for recommendation functions, predictive demand analysis, and personalization of customer experiences.

There is significant business value and potential competitive advantage for Zalando in implementing a tailored data model with the right set of tools:

- **Real-time analysis of customer behavior:** Faster personalization, more profitable marketing, and enabling predictive analytics.
- **Efficient inventory and order management:** Optimized deliveries and sales forecasts.
- **Scalability at peak times:** Robust infrastructure even during high loads.

Data pipeline

Based on the previously mentioned parts that should be included in Zalando's Modern Data Stack, here is a visualization of what their Data Pipeline could look like.



Data pipeline and destinations

These are the steps and destinations in the pipeline:

Step 1. Apache Airflow is used to schedule and manage the data execution in the Kafka, Spark, and ELT process with Amazon Glue. Amazon Cloudwatch is used to set up alerts and notifications, monitor the performance of Amazon tools, and can provide us with insights related to costs. AWS Lambda can send alerts on failed executions.

Step 2. Data sources such as marketing tools, e-commerce platforms, PIM tools, and third-party APIs from partners and vendors. Apache Kafka ingests the data.

Step 3. Apache Kafka distributes both real-time streaming data and batch data to optimize resource allocation.

Step 4. Data is processed through Apache Spark and undergoes minimal processing before being sent to Amazon S3 as a Data Lake. AWS Lambda filters, validates, and enriches the data before sending it to the appropriate Data Lake bucket based on data type. This is where semi-structured and raw data is stored. Machine learning tools and predictive models can retrieve raw data directly from the Data Lake.

Step 5. Amazon Glue is used between Amazon S3, Amazon Aurora and Amazon Redshift to transform, format and structure the raw data. Amazon Aurora for transactional data and Amazon Redshift for structured and processed data. Analysis tools are connected to the Data Warehouse and enable the right data, in the right format, to be available to the right team and people.

The data destinations included in the pipeline are therefore a combination of a Data Lake, a Data Warehouse and an operational OLTP database. This is to suit both transactional data and data that needs to be suitable for large and advanced analyses and queries.

Maintenance and data quality

The data pipeline should be continuously monitored to evaluate performance, processing times, identify any bottlenecks or changes in data format. It is an iterative process to constantly improve and develop the set-up to best suit Zalando's needs. In order to evaluate what is actually good or bad, you should have defined a kind of requirements specification that you follow up on. You can also evaluate which queries are most often run to optimize the set-up after this.

Transformation & ETL/ELT

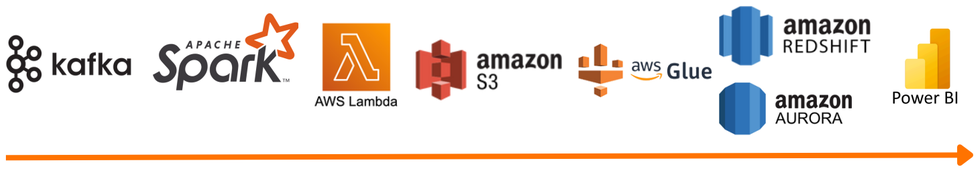
Example of a scenario with data from two data sources being transformed and integrated into the data model. AWS Glue can be used for both batch and stream data sources, enabling integration of data from multiple sources.

Order data from e-commerce platform JSON-format

{"order_id": "ORD12345","customer_id": "CUST001","order_timestamp": "2024-02-20T12:34:56Z","items": [{"product_id": "P001", "quantity": 2, "price": 50.00}, {"product_id": "P002", "quantity": 1, "price": 20.50}],"payment_method": "credit_card","currency": "EUR","campaign_code": "GADS456"}

Data from Google Ads CSV-format

campaign_code,ad_spend,campaign_name,clicks,conversions
GADS456,50.00,Spring Sale,200,5
GADS789,75.00,Summer Deals,300,8



1. The order data and marketing data are loaded as raw data into Amazon S3.
2. Transformation with Amazon Glue and Apache Spark:
   - The list of products needs to be restructured from JSON to table format.
   - The timestamp needs to be converted to the correct format for analysis.
   - The timestamp is missing from the Google Ads data and may need a handling rule.
   - The campaign codes need to be normalized to match the order data.
   - The data type for ad_spend is a string and needs to be converted to a numeric value.
   - The order value is calculated and added to the order data.
3. The Google Ads data and order data are joined through the campaign code.
4. The ROI is calculated.
5. Storage in Amazon Aurora and Amazon Redshift, dashboard with campaign results is created in Power BI.

Based on this, Zalando can gain insights into the profitability of campaigns and optimize its marketing budget by focusing on campaigns with the highest ROI.

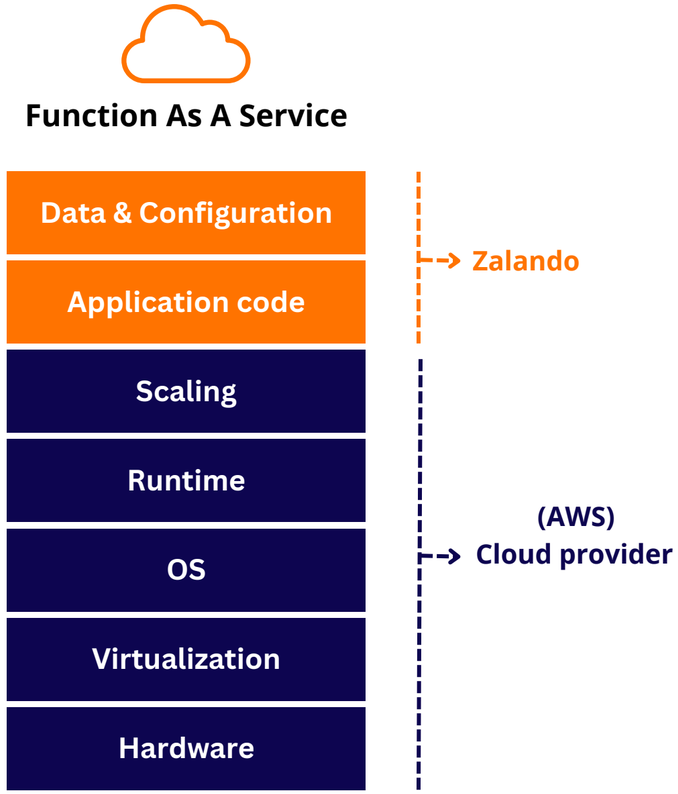
Cloud solution

A cloud solution like AWS Lambda in the form of Function As A Service (FaaS) can provide Zalando with a high-performance, scalable and cost-effective solution for managing its data flows and applications.

It can be strategically advantageous for Zalando to purchase services for managing the majority of the cloud infrastructure. By doing this, Zalando can focus on allocating its resources to developing business-critical applications and functions, and focusing on analytics and AI models instead of maintaining servers.

Zalando will pay for executed code, which means they will not receive any costs for idle capacity. Zalando can also avoid manual infrastructure management during ups and downs of traffic and order flows. AWS also has “disaster recovery” capabilities to be able to return to a functioning state faster after critical events with the infrastructure.

By outsourcing infrastructure and operations but maintaining control over their data and platforms, Zalando gets a scalable, cost-effective and secure cloud architecture.



Cloud solution

Components of the Data Warehouse
In the Data Warehouse, data is divided into different layers or “buckets” based on how much and how the data has been processed.

Staging layer
Acts as an intermediate storage between the Data Lake and the Data Warehouse, enables data reloading in the event of errors and improves performance by separating raw data from processed data.

Cleansed layer
Cleansed, deduplicated and structured data ready for further processing. For example, normalized currencies and date formats, linking order ID to customer ID, handling duplicates and incorrect data. This ensures that only high-quality data is used in the analysis.

Business layer
Data that is structured based on business needs and optimized for analysis. For example, revenue by product category or partner, customer segments by purchasing behavior or return frequencies by geographic region.

Analytics layer
Customized datasets for different departments, such as marketing, sales, business operations, etc.
This enables faster analysis in different areas and reduces the load on the main warehouse.

BI-Implementation

Zalando can implement a cloud-based BI solution where Power BI plays a central role in analyzing and visualizing data. This is especially important for operationalizing the data - making it available for decision-making.

To ensure that information is available in real time, Power BI can retrieve data directly from Zalando's Data Warehouse, where analyzed and structured data is stored in Amazon Redshift. To enable this integration, direct connections (DirectQuery) and import methods are used in Power BI, depending on the needs of the analysis. For real-time analysis during sales follow-up or inventory optimization, DirectQuery can be used, which means that Power BI queries the database directly without storing the data locally. This makes it possible to always work with the latest data without delay.

For more comprehensive reports and historical analysis, data is loaded via Import Mode, where Power BI regularly refreshes its datasets with scheduled queries to Redshift. Data in Redshift is already prepared and transformed through ETL and ELT processes with AWS Glue and dbt. This means that Power BI gets access to processed, aggregated datasets that are optimized for the analysis you want to do in Power BI.

The future of BI platforms and data visualization is moving towards more AI-driven insights, automated analysis and real-time visualization. For Zalando, this means an opportunity to further personalize customer experiences, optimize inventory and predict demand with greater accuracy and precision. By integrating AI and machine learning into Power BI, and by ensuring a flexible and scalable data architecture, Zalando can prepare to jump on future trends in BI.

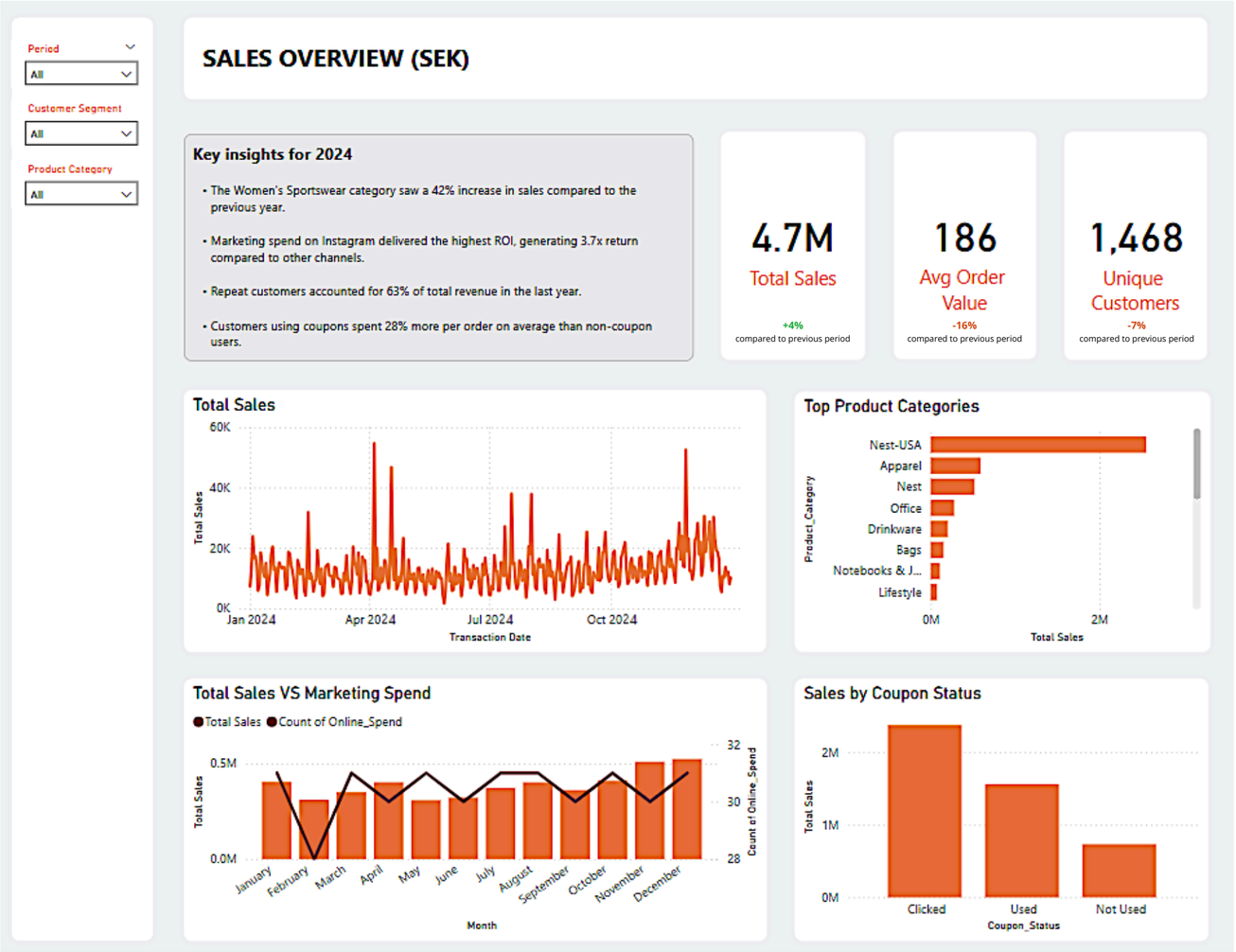
BI-Dashboard

This dashboard was created in Power BI and contains sales data from this Kaggle-dataset: https://www.kaggle.com/code/arpitppatel/marketing-insights-for-e-commerce-company/input

The dashboard can be used to get a very simple overview of the 2024 sales year, and answers questions such as:

- How have sales, in terms of sales value, looked per month?
- What was the average order value for the year?
- How many unique customers was it?
- How did marketing spend impact total sales?
- Which product categories have generated the most revenue?

To add extra value to this overview, I would recommend taking a closer look at comparisons and changes over time to gain insights into what has contributed to this year's results. To do so, further analysis and more advanced reporting than what this overview provides would be required.



My reflections and lessons learned

During this course, Data Modeling and BI Platforms, I have gained a deeper understanding of how data is structured, managed and analyzed on a larger scale than just through the analysis tools that I primarily use in my professional role today: Google Analytics, Piwik Pro and Matomo.

I have learned how data pipelines are built, from data sources to storage in Data Lakes and Data Warehouses, as well as how ETL/ELT processes are used to transform and integrate data from different systems. I have also gained a better understanding of cloud infrastructure, such as AWS, Azure and Google Cloud.

In my role as a digital analyst, this means that I can now better understand how user data can be integrated with other data sources, such as sales and product data, to create more “depth” to my analyses.

In summary, the course has given me a more technical understanding of how data is structured and flows through an organization, which is very important for me who wants to help businesses work more data-driven.

In the future, I want to spend more time exploring Power-BI and Tableau, to become more confident in my use and be able to visualize what I want to convey in a better (and nicer) way.

Sources

This report has been primarily based on course material and lessons learned from the course Data Modeling & BI Platforms, arranged by IHM Business School.

About Zalando
https://corporate.zalando.com/en/investor-relations/zalando-se-exceeds-its-own-profitability-guidance-fy-2024-after-better-expected
https://corporate.zalando.com/en/financials/zalando-full-year-23-results
https://corporate.zalando.com/en/about-us/what-we-do/how-zalando-leverages-technology-help-customers-find-right-size

Dataset to BI-dashboard
https://www.kaggle.com/code/arpitppatel/marketing-insights-for-e-commerce-company

Link to Lucidcharts 1 of 2
https://lucid.app/lucidchart/7f3e78c9-42b7-4c61-a24a-2817a5a2b65b/edit?invitationId=inv_f10f010e-9e91-4acc-960e-4d28cb50b101

Link to Lucidcharts 2 of 2
https://lucid.app/lucidchart/721f86ad-38c5-4a82-9f0e-c164be8feedf/edit?invitationId=inv_4032b474-4c67-42d6-aae2-b9eeedebc0eb