
Katarzyna C.

DATA ENGINEER

Education

Materials Science and Engineering

Language proficiency

English – C2

Polish – C2

Domains

Banking

IoT

Fintech

Certificates

[HashiCorp Certified: Terraform Associate](#)

Data Engineer with 8+ years of experience.

Experienced Data Engineer with a strong background in designing and delivering scalable data platforms across cloud and on-premises environments. Proficient in building distributed data processing pipelines using Apache Spark and PySpark, orchestrating complex workflows with Apache Airflow, and implementing modern Lakehouse architectures with Apache Iceberg and dbt. Hands-on experience with open-source platform engineering, including self-hosted component deployment on Kubernetes, infrastructure provisioning with Terraform, and data governance through OpenMetadata. Skilled in working with regulated and sensitive data domains, collaborating with cross-functional teams, and translating business requirements into production-grade data solutions.

Programming languages

Python, SQL.

Data Engineering

Apache Spark, PySpark, Apache Airflow, dbt, Apache Iceberg, OpenMetadata.

Cloud

GCP (BigQuery, Cloud Build, Cloud Composer, Cloud Functions, Cloud Monitoring, Cloud Storage, Dataflow, IAM, Looker Studio, Pub/Sub, Secret Manager, etc.).

Databases

PostgreSQL, Trino, TimescaleDB.

Data Science

Pandas, NumPy.

Message brokers

Kafka.

BI and Data Visualization

Apache Superset, OpenSearch.

Data Modeling

Data Mesh, Dimensional Modeling (Star Schema, Snowflake Schema), Medallion Architecture, Entity-Relationship (ER) Modeling, Normalization / Denormalization.

DevOps

Docker, Docker Compose, Kubernetes, Terraform, GitLab CI/CD, Bash scripting, Prometheus, Grafana.

AI Tools

Cursor, Claude.

Version control systems

Git, GitHub, GitLab.

Projects

REGULATORY REPORTING PLATFORM

An open-source platform for automated collection, processing, and governance of regulatory reporting data across banking business lines. Responsible for building core ingestion and transformation pipelines, implementing data quality controls, and contributing to platform architecture decisions from early stage.

Project roles

Data Engineer

Period

12.2022 - till now

Responsibilities

- Designed and implemented modular PySpark transformation pipelines processing regulatory financial data with incremental batch ingestion into Apache Iceberg tables;
- Configured Apache Iceberg table formats with hidden partitioning and snapshot-based time travel to support audit and reconciliation requirements for regulated data;
- Built and maintained Apache Airflow DAGs with

custom sensors and retry logic, and idempotent tasks, enabling failed ingestion runs to be safely replayed;

- Developed the dbt transformation layer covering incremental models, generic tests, exposures, and lineage documentation in dbt docs;
- Deployed and configured Trino as a distributed SQL query engine for federated queries across PostgreSQL and Iceberg sources, tuning resource groups for concurrent analytical workloads;
- Designed hypertables and continuous aggregates in TimescaleDB for time-series financial metrics, applying compression and retention policies;
- Containerized platform services and contributed Helm chart and StatefulSet configurations for Kubernetes deployments, supporting persistent storage and release procedures;
- Provisioned on-premises infrastructure using Terraform modules with self-hosted provider configurations and remote state management;
- Automated GitLab CI/CD pipelines for build, test, and deployment of platform components across environments;
- Integrated OpenMetadata for data lineage tracking, access governance, and audit trail coverage over sensitive regulatory datasets;
- Configured OpenSearch indices for log analytics and platform observability, supporting internal search and alerting use cases;
- Set up Prometheus metrics collection and Grafana dashboards to monitor pipeline health, Trino query latency, and Kubernetes resource utilization;
- Developed reusable Python tooling for pipeline automation, data validation, and dependency management across platform services;
- Collaborated with stakeholders to translate regulatory data requirements into architectural proposals and platform design decisions.

ENTERPRISE DATA OCEAN PLATFORM

Enterprise-level data platform for ~250 source systems handling raw and consumption-ready data layers. Part of a platform engineering team responsible for architecture, configuration, and data flow stability across multiple business units.

Environment

Python, SQL, Apache Spark, PySpark, Apache Airflow, dbt, Apache Iceberg, OpenMetadata, Pandas, NumPy, Trino, PostgreSQL, TimescaleDB, OpenSearch, Apache Superset, Prometheus, Grafana, Docker, Docker Compose, Kubernetes, Terraform, Bash scripting, GitLab CI/CD, Kafka, Cursor, Claude, GitLab.

Project roles

Data Engineer

Period

06.2020 - 11.2022

Responsibilities

- Collaborated with stakeholders across business analytics, data engineering, and data architecture teams to gather requirements and translate them into scalable platform solutions;
- Implemented GCP-based Lakehouse architecture using BigQuery and Cloud Storage to support governed, high-performance analytical workloads across enterprise data domains;
- Developed and operated in Cloud Composer, managing dependencies and schedules for hundreds of production workflows;
- Maintained and developed scalable PySpark transformation jobs processing large volumes of structured and semi-structured enterprise data;
- Participated in migrating legacy ETL pipelines to GCP, improving scalability, reliability, and operational efficiency of core data flows;
- Worked closely with technical risk management teams to identify and mitigate risks associated with enterprise data platform components;
- Authored complex SQL queries in BigQuery to support data validation, transformation logic, and sensitive data protection;

IoT DEVELOPMENT AND ANALYTICS PLATFORM

Cloud-based platform enabling communication, monitoring, and analytics for a large network of IoT devices. Part of a development and monitoring team responsible for building

- Implemented fine-grained access control and data lineage visibility using BigQuery Data Catalog and IAM policy management;
- Optimized BigQuery query performance through partitioning strategies, clustering, materialized views, and slot utilization tuning;
- Utilized Looker Studio for analytical serving and business intelligence reporting across downstream consumer teams;
- Provisioned and maintained GCP infrastructure components using Terraform, automating environment configuration across multiple deployment stages;
- Developed and maintained CI/CD pipelines using Cloud Build for automated deployment of data platform components;
- Conducted code reviews, mentored junior engineers, and onboarded new developers into platform standards and development practices;
- Created platform documentation and training materials, serving as L2/L3 support lead for internal engineering teams.

Environment

Python, SQL, Apache Spark, PySpark, Apache Airflow, GCP (BigQuery, Cloud Storage, Cloud Composer, Cloud Build, Dataflow, IAM, Cloud Functions, Pub/Sub, Looker Studio, Secret Manager, etc.), PostgreSQL, Pandas, NumPy, Docker, Docker Compose, Bash scripting, Terraform, GitHub.

Project roles

Data Engineer

Period

07.2018 - 05.2020

Responsibilities

- Built and maintained ETL pipelines for data ingestion, transformation, and analysis from heterogeneous IoT device sources;

data pipelines, designing data models, and supporting technical and business observability.

- Developed and optimized PySpark structured streaming jobs to process high-volume real-time telemetry data with near real-time latency for critical system alerts;
- Implemented complex PySpark transformations for cleansing and normalization of heterogeneous IoT protocol data into unified analytical schemas;
- Contributed to the BigQuery analytical data model and implemented curated tables for structured and semi-structured IoT events.
- Utilized Pub/Sub for streaming and buffering high-volume IoT telemetry events, enabling reliable real-time ingestion and downstream processing;
- Built Python-based Cloud Functions to automate small-scale ETL tasks, trigger downstream workflows, and handle edge-case ingestion logic;
- Integrated external REST APIs into ingestion pipelines, enabling data collection from third-party systems and external sensors;
- Monitored platform performance using Cloud Monitoring, detecting anomalies in HTTP error rates, latency, and system health across markets;
- Queried and analyzed large-scale datasets in BigQuery for ad-hoc diagnostics and reporting;
- Developed documentation to assist non-technical users in interpreting dashboards and platform health metrics;
- Delivered infrastructure changes using Terraform for GCP resource provisioning.

Environment

Python, SQL, Apache Spark, PySpark, Apache Airflow, GCP (BigQuery, Cloud Storage, Pub/Sub, Cloud Functions, Cloud Composer, Cloud Monitoring, IAM, Dataflow, etc.), PostgreSQL, Kafka, Pandas, NumPy, Docker, Docker Compose, Bash scripting, Terraform, GitHub.

Professional skills

SKILLS		EXPERIENCE IN YEARS	LAST USED
Programming languages	Python	8	2026
	SQL	8	2026
Data Engineering	Apache Spark	8	2026
	PySpark	8	2026
	Apache Airflow	5	2026
	dbt	4	2026
	Apache Iceberg	4	2026
	OpenMetadata	3	2026
BI and Data Visualization	Apache Superset	4	2026
	OpenSearch	4	2026
Clouds	GCP	4	2022
Data Science	Pandas	8	2026
	NumPy	8	2026
Message brokers	Kafka	6	2026
Databases	PostgreSQL	8	2026
	Trino	4	2026
	TimescaleDB	4	2026

SKILLS		EXPERIENCE IN YEARS	LAST USED
DevOps	Docker	8	2026
	Docker Compose	8	2026
	Bash scripting	8	2026
	Terraform	8	2026
	Prometheus	4	2026
	Grafana	4	2026
AI Tools	Cursor	2	2026
	Claude	1	2026
Version control systems	Git	8	2026
	GitLab	4	2026
	GitHub	4	2022