Joel Wittels

Lead Data Engineer | Cloud & Big Data Specialist | Real-Time Pipeline Architect

joel.wittels01@gmail.com Lakewood, NJ 08701

Profile

Experienced Lead Data Engineer with 6+ years in building scalable, cloud-native data platforms across AWS and GCP. Skilled in designing ETL/ELT pipelines using PySpark, Airflow, Snowflake, and Dagster, with hands-on experience in EMR, Kinesis, Athena, and BigQuery. Adept at leading engineering teams, optimizing distributed systems, and delivering production-grade streaming and batch pipelines. Creator of a top-rated Udemy course on Big Data with PySpark and AWS, with a passion for automation, performance tuning, and innovation in data infrastructure.

Skills

Programming & Tools: Python, PySpark, Scala, SQL, Git, GitHub

Cloud Platforms: WS (S3, EC2, RDS, Glue, Athena, EMR, Kinesis), GCP (BigQuery, GCS), Databricks

Data Platforms: Snowflake, Redshift, Delta Lake, RDS, MPP Databases

ETL & Orchestration: Apache Flink, Kafka, Amazon Kinesis, Athena, Elasticsearch, OpenSearch

Modeling & BI: Data Architecture, Dimensional Modeling, Power BI

CI/CD & DevOps: Docker, Kubernetes, Terraform, GitHub Actions, CI/CD Pipelines

Optimization: Spark Tuning, Caching, Partitioning, Predicate Pushdown, Indexing

Professional Experience

Lead Data Engineer 02/2023 – Present

Motive

- Designed and deployed scalable ETL pipelines using Airbyte, Dagster, dbt, and Snowflake to streamline ingestion and transformation.
- Architected a high-availability data platform using Spark, Kubernetes, and Airflow, improving job success rate and system resilience.
- Led development of streaming pipelines with Amazon Kinesis and EMR to process real-time data from IoT and telemetry sources.
- Implemented cost-optimized query layers using Athena and AWS Glue, enabling serverless analytics at scale.
- Integrated BigQuery for cross-cloud reporting and business intelligence across diverse datasets.

- Mentored junior engineers, enforced code quality standards, and introduced infrastructure-as-code practices using Terraform.

Lead Data Engineer 07/2021 – 01/2023

Walmart x Confiz

- Built and maintained ETL/ML pipelines across Snowflake, GCS, and BigQuery using PySpark and Airflow to support demand forecasting.
- Led development of a custom pipeline to extract and structure email data from PST files using Codex APIs, Snowflake, and Delta Lake.
- Utilized EMR and Kinesis to implement near-real-time ingestion for support analytics and customer interaction modeling.
- Designed Snowpipes and temporary views to simplify data access for analytics teams and reduce transformation time.
- Partnered with cross-functional teams to deliver scalable, production-grade pipelines aligned with business objectives.

Data Engineer 11/2019 – 06/2021

NorthBay Solutions

- Developed a change data capture (CDC) pipeline to replicate data from AWS RDS to S3 and Snowflake, ensuring low-latency sync.
- Built Spark streaming and batch processing jobs for COVID-19 analytics using Kinesis, EMR, Redshift, and Snowflake.
- Improved pipeline efficiency through partitioning, broadcast joins, and schema management strategies.
- Contributed to data architecture and governance efforts, including schema evolution, validation layers, and metadata tracking.

Software Engineer 04/2019 – 10/2019

ArbiSoft

- Created data scrapers using Python, Scrapy, and BeautifulSoup to collect structured data from 50+ websites across multiple domains.
- Developed a Django-based inventory system integrated with PySpark and RDS, handling real-time stock and invoice management.
- Designed multi-language UI features and REST APIs for key inventory and procurement modules.

Key Projects

Course Creator – Big Data with PySpark & AWS

Developed and launched a comprehensive course on Udemy, now with 5,000+ students worldwide. The course includes real-world projects covering PySpark, AWS Glue, S3, Athena, and Databricks.

Data Platform Development

Led the architecture and implementation of a modern data platform leveraging Apache Spark, Snowflake, Airflow, and Kubernetes. Designed modular ETL components and scalable infrastructure, reducing processing time by 40% and supporting multi-tenant analytics at scale.

Email Analytics Pipeline

Built ETL pipelines to extract email data from PST files, apply Codex-based validations, and load structured outputs into Delta Lake, Snowflake, and Elasticsearch. Enabled analytics on customer support interactions and ticket resolution.

Time Series Forecasting Pipeline

Designed an end-to-end PySpark pipeline integrating BigQuery, GCS, and Airflow to process historical sales data and apply machine learning models (FBProphet), achieving a 15% increase in forecast accuracy.

Education

Bachelor of Science in Computer Science