Pedro Alves Senior Data Engineer

Sunnyvale, CA · pedro.alves@gmail.com · +1 408-555-0172 · linkedin.com/in/pedroalves

PROFILE SUMMARY

- **Senior Data Engineer** with **7 years** on **delivery marketplace data** at **DoorDash and Instacart**, owning **Kafka and Flink streams, CDC ingestion, and Spark batch jobs** behind **Dasher pay, ETAs, and merchant menus**.
- Writes **Python**, **Scala**, and SQL, processes data in **Apache Spark**, **Flink**, and **dbt**, streams with **Kafka** and **Debezium**, schedules in **Airflow**, and lands tables in **Delta Lake** and **Snowflake** on AWS.
- Specializes in **event-time processing** with watermarks, **CDC merges** into Delta tables, and **Spark job tuning**, and checks each change against **freshness SLAs and cost per run** before it ships to production.
- Works with **ETA model engineers**, **logistics analysts**, and the **service teams that own source databases**, shares the **data platform on-call rotation**, and writes a **design doc** before any new pipeline.
- Mentors **3 mid-level engineers** in weekly pairing sessions, sits on the **logistics data hiring loop**, and wrote the **Flink backfill playbook** the team follows after outages.

TECHNICAL SKILLS

Languages:	Python, SQL, Scala, Bash
Processing & Transformation:	Apache Spark, PySpark, dbt, Trino, Polars
Streaming & Messaging:	Apache Kafka, Apache Flink, Debezium, Kafka Connect, Confluent Schema Registry, Protobuf
Orchestration & Ingestion:	Apache Airflow, Fivetran, AWS Step Functions
Warehouses & Lakehouses:	Databricks, Delta Lake, Snowflake, PostgreSQL, Apache Parquet
Cloud & Infrastructure:	AWS (S3, EKS, MSK, IAM), Terraform, Docker, Kubernetes, GitHub Actions
Data Quality & Observability:	Monte Carlo, Great Expectations, dbt tests, Datadog, PagerDuty

EDUCATION

University of Illinois Urbana-Champaign — **B.S. in Computer Science** — Urbana, IL · Aug 2015 to May 2019

WORK EXPERIENCE

DoorDash — **Senior Software Engineer, Data Engineering** — Sunnyvale, CA · Jul 2022 to Present

- Built a **Flink** job on **Kafka** that joins order and Dasher location events into ETA model features with watermarks for late data, cutting feature lag from 15 minutes to **22 seconds**.
- Swapped nightly Postgres dumps of menu and store-hours data for **Debezium** CDC merged into **Delta Lake**, so menu edits reach the search tables in **6 minutes** instead of the next day.
- Tuned the daily Dasher earnings jobs in **Apache Spark** on **Databricks** with Z-ordered Delta tables, broadcast joins, and right-sized job clusters, cutting their compute spend by **$37,400 a month**.
- Added **Monte Carlo** freshness and volume monitors on the merchant payout tables, each paging the owning team, so finance now reports **78% fewer** payout data issues.
- Wrote the Protobuf event contract template that **11 logistics teams** now use for new **Kafka** topics, working out field ownership with ML and analytics leads in a monthly design review.

Instacart — **Data Engineer** — San Francisco, CA · Jul 2019 to Jun 2022

- Split the single SQL script behind the fulfillment dashboards into **dbt** staging, intermediate, and mart models on **Snowflake**, with snapshots for store fee history, cutting dashboard load times by **58%**.
- Rebuilt the shopper supply **Airflow** DAGs with cross-DAG sensors and idempotent, date-partitioned tasks, so only affected partitions rerun after a late load, cutting manual reruns from 30 a month to **4**.
- Moved the team's Python jobs into Docker images on **Kubernetes**, with **Terraform**-built staging and production environments and CI tests on every pull request, cutting deploy rollbacks by **64%**.