Amandeep Singh

San Francisco, CA | amndp64@gmail.com | (916) 296-7074 | linkedin.com/in/bitterfq | github.com/bitterfq

About

Data Engineer with 3+ years of experience building scalable data infrastructure, real-time and batch pipelines, and cloud-native ETL systems. Proven track record driving reliability and efficiency at Meta and The Nature Conservancy across petabyte-scale workflows. Specialized in orchestration, streaming, and analytics-ready modeling using tools like Airflow, Kafka, dbt, and Snowflake on AWS.

Experience

Cloud Data Platform Engineer **San Francisco, CA**
ZTF Real-Time Astronomical Data Pipeline *2025*

- Architected a modular, fault-tolerant ingestion pipeline for **400K+ daily alerts** from ZTF using **Kafka, Airflow, and AWS (S3, EC2)**, supporting low-latency, cloud-native event processing at scale.
- Implemented partitioned alert storage and **hash-based deduplication**, achieving **100% idempotent ingestion** across 30+ days and cutting redundant S3 usage by **30%**.
- Integrated with **Snowflake** using external stages and `COPY INTO`; modeled normalized, query-optimized tables using **dbt** with full backfill support.
- Deployed orchestration stack via **Docker Compose** on **Raspberry Pi 5** to test low-cost edge infrastructure; mirrored workloads on **AWS EC2** to validate deployment portability and performance tradeoffs.

Data Engineer **Remote**
The Nature Conservancy *Oct 2024 - Present*

- Built an end-to-end **Python** system to automate Planet Labs API workflows for acquiring satellite imagery tied to conservation zones, enabling monitoring of **25+ critical ecosystems**.
- Designed **ETL pipelines using Airflow**, processing **3+ TB of imagery weekly** and storing assets in **AWS S3** with versioned metadata for reliable research access.
- Architected infrastructure to improve pipeline runtime by **40%**, enabling faster turnaround on river ecosystem monitoring and reducing resource utilization.
- Developed **automated reporting and alerting systems** to support early detection of ecosystem stress, preventing rivers from going dry without detection in **20+ at-risk watersheds**.
- Created **interactive dashboards** using Streamlit that visualized critical environmental metrics, making complex satellite data accessible to **40+ non-technical stakeholders**.

Data Engineer (Infrastructure) **California**
Meta *Mar 2022 - May 2024*

- Supported and maintained **mission-critical real-time data pipelines** processing **5+ PB daily**, including ODS layers and batch ingestion jobs powering internal analytics systems.
- Diagnosed and resolved **74% of infra-related support tickets** (over 300 annually), reducing average incident resolution time from 24 to 8 hours.
- Optimized high-volume **Python and SQL pipelines** across petabyte-scale warehouses, enhancing query efficiency by **35%** and reducing compute utilization.
- Built internal tools and documentation that **reduced repeat support requests by 40%**, improving engineer self-sufficiency and speeding up onboarding processes.
- Collaborated cross-functionally with data engineers and product teams to ensure end-to-end data reliability, maintaining **99.5% SLAs** for critical data systems.

Education

University of San Francisco **San Francisco, CA**
M.S. in Data Science and Engineering (GPA: 3.91/4.00) *2024 – 2025*
University of California, Davis **Davis, CA**
B.Sc. in Computer Science *2018 - 2022*

Skills

Data Engineering: ETL Pipelines, Data Warehousing, Orchestration, Data Modeling, API Integration, Data Quality, Monitoring
Big Data: Batch/Stream Processing, Distributed Systems, Data Lake Architecture, Scaling
Cloud/Infra: AWS (S3, EC2, Lambda), Docker, Kubernetes, Terraform, CI/CD, Linux
Languages/Tools: Python, SQL, Bash, PySpark, Java, C++, Airflow, Kafka, Snowflake, dbt, Streamlit, Hive, BigQuery