John Doe - Data Engineer

San Francisco, CA | (555) 555-5555 | john.doe@gmail.com | linkedin.com/in/johndoe | github.com/johndoe

Summary

Experienced Data Engineer with 5 years of expertise in designing and optimizing scalable data pipelines, building data lakes, and deploying cloud-based solutions. Adept at leveraging big data technologies and distributed systems to transform raw data into actionable insights. Proven track record of success at LinkedIn and Google, driving data-driven decision-making through efficient systems.

Professional Experience

LinkedIn - Sunnyvale, CA

Data Engineer (January 2021 - Present)

- Designed and maintained scalable data pipelines using Apache Spark, Kafka, and Airflow to process over 5TB of daily user activity data.
- Spearheaded the migration of legacy ETL workflows to a cloud-native solution on AWS, reducing processing time by 40%.
- Partnered with the machine learning team to provide real-time feature engineering pipelines, improving LinkedIn's recommendation systems by 20%.
- Built a data quality framework that detected anomalies, reducing data discrepancies by 30% and ensuring high availability for dashboards.
- Optimized LinkedIn's job recommendation algorithm by implementing Delta Lake for faster query performance, reducing latency by 50%.

Google - Mountain View, CA

Junior Data Engineer (June 2018 - December 2020)

- Developed and maintained scalable data pipelines to ingest and process billions of rows of data daily using BigQuery and Dataflow.
- Automated data validation workflows using Python and Cloud Functions, saving 15+ hours of manual effort weekly.
- Worked closely with product teams to design and implement a Data Warehouse architecture that powered insights for Google Ads campaigns.
- Contributed to the adoption of Data Catalog, enhancing metadata management and improving data discoverability for analysts and engineers.
- Improved query performance on terabyte-scale datasets by tuning SQL queries and optimizing table partitioning, reducing costs by 25%.

Technical Skills

- Big Data Technologies: Apache Spark, Hadoop, Kafka, Airflow, Hive
- Cloud Platforms: AWS (S3, Redshift, Glue), Google Cloud Platform (BigQuery, Dataflow, Cloud Storage)
- Programming: Python, SQL, Java, Scala
- Data Warehousing: Snowflake, Redshift, BigQuery
- Tools: Tableau, Looker, Power BI, Git, Docker, Kubernetes
- Frameworks: Pandas, PySpark, TensorFlow (for feature engineering)

Education

Bachelor of Science in Computer Science

University of California, Berkeley - Berkeley, CA (Graduated: May 2018)

Projects

- Real-Time Analytics Platform: Developed a real-time streaming data pipeline for a retail client using Kafka and Spark Streaming, enabling near-instant analysis of sales and inventory data.

- ETL Pipeline for Social Media Insights: Built an automated ETL pipeline to process social media engagement data using AWS Glue and Redshift, empowering marketers with weekly insights.
- Ad Performance Optimization: Designed and implemented a scalable BigQuery solution that processed 100M+ ad impressions daily, improving reporting performance by 30%.

Certifications

- AWS Certified Data Analytics - Specialty (2022)
- Google Cloud Professional Data Engineer (2021)
- Certified Apache Spark Developer (2020)