Aditya Singh

aditya.dhariwaL4@gmail.com +1(916)743-6606 LinkedIN

Data Engineering & Analytics | Software Engineering in Data | Python | SQL | ETL |
Data Infrastructure | Bigdata | Spark | AWS cloud | Hadoop

Experienced AI focused Data Engineer with 6+ years of experience in designing and optimizing data infrastructure. Skilled in developing and orchestrating data pipelines, ETL processes, and deployment strategies. Adept at transforming structured, semi-structured, and unstructured data into actionable insights, driving efficient data management and analysis.

EDUCATION

California State University, Sacramento — Sacramento, California
Master of Science in Computer Science — *Aug 2016 – Dec 2018*

Dr. A.P.J. Abdul Kalam Technical University — Meerut, India
Bachelor of Technology in Computer Science — *Aug 2011 – May 2015*

CERTIFICATIONS

AWS Certified Data Engineer – Associate

WORK EXPERIENCE

JCPenney — Data Engineer II — San Francisco, California (Nov 2023 – Present)

Tech Stack: **Python, Spark, SQL, Hadoop, Snowflake, Hive, Redshift, S3 Data lake, Jenkins, Airflow, HDFS, Sqoop, Kafka, Datameer, Postgres, Great Expectations, DBT, DataHub, UNIX, GIT**

- Orchestrating the ingestion of data from different sources such as MySQL, Kafka, and edge node servers into Enterprise Data Lake.
- Developing data validations frameworks to ensure 98% data accuracy for downstream users.
- Automating reporting data pipeline using Spark, S3, Jenkins, Redshift that reduced report generation time up to 60%.
- Developing, optimizing and monitoring ETL data pipelines, ensuring seamless and high-performance data integration.
- Eliminating data deduplication, using compressions, performing data archiving with regular auditing to optimize storage cost resulting in $20k monthly savings.
- Executing intricate data transformation and massaging tasks, proficiently loading data into Amazon Redshift and Hive tables.
- Scheduling hourly, daily and weekly jobs using Airflow and Jenkins workflow to load the data into Data warehouse.
- Applying advanced analytical techniques to dissect and comprehend data originating from disparate systems, facilitating the convergence of datasets within the data lake to extract valuable insights and discern trends.
- Achieved 25% reduction in data storage cost by leveraging parquet's compression feature and decreasing the runtime by 40%.
- Processing more than 5 million hourly data using PySpark and ensuring robust infrastructure using partitioning, error handling with low latency and throughput.

WARNER BROS. DISCOVERY — Data Engineer II — San Francisco, California (Apr 2021 - Sep 2023)

Tech Stack: **Python, SQL, Snowflake, Databricks (Workflows, Delta Lake), AWS Lambda, Glue, Athena, Airflow, Terraform, Jenkins CI/CD, Kafka, Docker, Kubernetes EKS, Druid, Grafana, Spark, Django, Prometheus, Cloudwatch, GIT**

- Built and maintained Realtime and batch processing personalize recommendation data pipelines.
- Maximized, migrated content and legacy data pipelines for new streaming platform MAX.
- Performed data validation, sanity checks, data quality checks for the imported data feeds.
- Engineered content engagement metrics, including viewership duration, click-through rates and user engagement from Databricks delta tables.
- Collected and built content engagement (viewership) metrics that were used to generate insights into the performance of platform's content (LATAM/EUROSPORT).
- Designed, created robust and scalable data pipelines to ingest, transform and load data from multiple sources.
- Automated current processes to run effectively through Airflow DAGs and transformed data in Scala/Spark using Databricks workflow. Architected Snowflake tables/views for stakeholder in APAC and LATAM.
- Collaborated with engineers, product managers, and data scientists for their data needs and provided them content viewership related metrics.
- Built data product for performance monitoring and alerting for applications in Kubernetes cluster by using Grafana, Prometheus.
- Built data product to view Kafka cluster metrics such as cluster messages in-out rate, recommendation latency, messages per batch.
- Improved FNK and Magnolia recommendation engine pipeline using Glue and push notifications and developed internal data products using Django, Grafana and Prometheus that power Discovery's product offerings.
- Cost optimization up to 20% by improving data storage and data retrieval processes in delta lake and Snowflake.

- Deployed Discovery+ legacy data pipelines from non-managed Airflow servers to a managed airflow server along with productionizing the code from EC2 jobs to Databricks.

Capgemini Senior Consultant — New York, NY
(Sep 2019 - Apr 2021)

Tech Stack: **Python, SQL, PySpark, AWS CloudFormation, Lambda, DynamoDB, Glue, Tableau, DB2, Attunity, Bamboo, Swagger, Sybase, Airflow, SparkSQL, Redshift, HDFS, Jenkins, Cloudwatch, Shell Scripting**

- Migrated micro services onto cloud platform while extracting, transforming, and loading data into target database with all business requirements.
- Used Kinesis streams to trigger Lambda and automated the workflow to capture everyday transactions from the streams.
- Curated, designed, and cataloged high quality data models, and decision-making insight of business processes to stakeholders.
- Wrote initial data capture glue jobs to fetch and transform data feeds from different data sources.
- Enhanced quality of data by 6% by performing change data capture load with high-quality business requirements.
- Automated the workflow process and performed daily aggregations on the database.
- Ingested, transformed data from middleware and web API into the data pipeline and provided it to the stakeholders for critical business requirements.
- Built an automation framework to test end to end ETL pipeline data flow.
- Automated the workflow to support multiple cross-functional teams and their need for structured data for analytical purpose.
- Improved latency of the pipeline by 10% using data normalization for ingesting data into the core schema.

cherre Data Engineer — New York, NY
(Jul 2019 - Aug 2019)

Tech Stack: **Python, Airflow, BigQuery, Google Analytics, PostgreSQL, NoSQL, GCP, GraphQL, Jenkins**

- Collected**,** integrated public and private data into data warehouse.
- Build custom dashboards using Google analytics to track marketing channels and different types of searches.
- Worked with data scientists on model optimization and provided them clean transformed data for predictive modeling.
- Developed and unit tested assigned features to meet product requirements.
- Design data warehouse solutions to support ETL processes and data analytics applications.

State Compensation Insurance Fund — Pleasanton, California
Data Engineer Intern — (Apr 2018 - Sep 2018)

Tech Stack: **Python, AWS, Oracle SQL, Tableau, SAS, Redshift, Adobe Analytics, Github**

- Worked on predictive analytics to generate 66% attribute matching. Used SAS to generate and record 100k attribute matches every day.
- Wrote python scripts to fetch data from the source and reduced the load time by 3%.
- Used Tableau to generate ad hoc reports.
- Helped in data migration for predictive modeling team.