Jay Pasupuleti

+1 747 -363-7540 | pasupuletij398@gmail.com

Summary

Data Engineer with 5 years of experience specializing in building scalable data pipelines, integrating structured and unstructured data across cloud platforms such as AWS, Azure, and GCP. Proven track record of optimizing data quality, integrity, and security, leading to a 20% improvement in data processing efficiency and enhanced analytics capabilities. Expertise includes data pipeline development, BI concepts, and innovative data solutions that translate business needs into technical outcomes, supporting data-driven decision-making in dynamic healthcare environments.

Technical Skills

- **Programming Languages:** Python, Java
- **Cloud Platforms & Technologies:** AWS Glue, Lambda, Redshift, S3, Step Functions, EMR, Kinesis, IAM, KMS, Lake Formation, CloudWatch, CodePipeline, SageMaker, Glue Catalog, Azure Data Factory (ADF), Synapse Analytics, Azure SQL, Cosmos DB, Databricks, Event Hub, Stream Analytics, DevOps, Purview, AKS (Azure Kubernetes Service), Google BigQuery, Google Dataflow, Pub/Sub, Google Cloud Storage, Google Data Proc, Google Data Catalog
- **Data Integration & ETL Tools:** Apache Airflow, DataStage, Talend, Informatica, SSIS
- **Data Modeling & Warehousing:** Dimensional Modeling, Change Data Capture (CDC), AWS Glue, Azure Data Factory, Azure Synapse Analytics, Snowflake, Google BigQuery, Relational and Multi-dimensional Modeling Concepts
- **Databases & Data Storage:** SQL (SQL Server, PostgreSQL, MySQL, AuroraDB), NoSQL, Cosmos DB
- **Big Data & Data Processing:** Apache Spark, PySpark, Hadoop, AWS EMR, Azure Databricks, Delta Lake, Parquet
- **Version Control & Development Practices:** Git, Agile, Scrum
- **Business Intelligence & Visualization:** PowerBI, Tableau, QuickSight
- **Soft Skills & Methodologies:** Strong problem-solving skills, Attention to detail, Excellent communication skills, Documentation skills, Ability to work independently, Ability to work collaboratively in a fast-paced, Agile environment

Professional Experience

Senior Data Engineer, *Nationwide Insurance* 05/2023 – Present | Columbus, OH

- Designed and developed 20+ ETL pipelines using GCP Dataflow, BigQuery, and Cloud Functions, enabling scalable processing of both structured and unstructured data and accelerating real-time analytics performance improvements by 45%.
- Built and maintained real-time streaming pipelines with GCP Pub/Sub and Cloud Functions, managing ingestion of over 1 million records per hour, supporting high-velocity data environments and reducing data latency by 30%.
- Architected data models in BigQuery and Redshift, leveraging RDBMS best practices to optimize query performance by 60%, while reducing storage costs by 25% through effective partitioning and UDF implementation.
- Automated cross-cloud data flows using AWS Lambda and Step Functions, decreasing manual data movement by 90% and improving data synchronization reliability between AWS and GCP environments.
- Integrated S3, Redshift, and GCP Cloud Storage to unify multi-cloud pipelines, transferring over 10+ TB of data monthly across platforms, enhancing data accessibility and operational efficiency.
- Optimized data models in BigQuery and Redshift by applying advanced modeling and performance tuning techniques, reducing query latency by 60% and storage costs by 25%.
- Automated cross-cloud data workflows using AWS Lambda, Step Functions, and CodePipeline, ensuring seamless data integration for 30+ datasets, which improved data refresh times by 40%.
- Integrated S3, Redshift, and GCP Cloud Storage to streamline data pipelines, enabling Change Data Capture (CDC) for near-real-time synchronization and improving data freshness by 35%.
- Implemented CI/CD workflows using Cloud Build, Jenkins, and Terraform, reducing deployment times for over 50 workflows by 70% and enhancing reproducibility across environments.
- Strengthened security using IAM, Cloud VPC, and AWS VPC, maintaining compliance with SOC 2 and GDPR, safeguarding over 500 data transactions daily with encryption and access controls.
- Designed and maintained real-time dashboards in Looker, delivering actionable insights and supporting 150+ stakeholders, which contributed to a 20% increase in operational efficiency.
- Led migration of 30+ legacy on-premises workloads to GCP, including redesign of Hadoop-to-BigQuery pipelines and Hive-to-SQL conversions, achieving zero data loss and minimal downtime.
- Led orchestration of batch and streaming data flows, utilizing MapReduce jobs for pre-ingestion processing of 10 TB of unstructured data, enhancing data processing speed by 50%.
- Deployed CI/CD workflows with Cloud Build and Jenkins, reducing deployment time by 70% and increasing test coverage of ETL processes to 95%, ensuring production stability.
- Enforced security standards by applying IAM policies and encryption techniques, ensuring compliance with SOC 2 and GDPR for over 500 data transactions daily.
- Developed custom APIs to support data exchange across 6 internal systems, increasing data availability and integration efficiency by 50%.

- Managed datasets exceeding 50 TB in BigQuery and Redshift, applying advanced modeling techniques to scale analytics and reduce processing times by 40%.
- Migrated legacy on-prem workloads to GCP using Migrate for Compute Engine, integrating 30+ data sources into BigQuery with zero disruption to business operations.
- Established end-to-end data lineage tracking with CloudWatch and Cloud Logging, streamlining root cause analysis and increasing troubleshooting efficiency by 80%.
- Utilized Databricks notebooks to perform advanced data transformations and machine learning model training, integrating with AWS S3 and Redshift, boosting forecasting accuracy by 35%.
- Collaborated with ML teams to operationalize models on AWS SageMaker and GCP AI Platform, enhancing predictive analytics capabilities and supporting data-driven decision-making.

Data Engineer, *Cigna Health* 09/2020 – 08/2022 | Chennai, Tamil Nadu

- Led development of 15+ ETL pipelines using AWS Glue, Python, and Snowflake, transforming raw data from 10+ S3 buckets into structured formats for Redshift and Snowflake CDC frameworks, supporting enterprise-wide data analytics.
- Built real-time streaming pipelines with AWS Kinesis, reducing data latency to under 60 seconds, enabling faster decision-making during high transaction volumes, contributing to a 20% operational efficiency gain.
- Optimized Redshift performance by applying RDBMS best practices and advanced modeling techniques, decreasing query times by 50% and lowering storage by 30%, supporting scalable analytics.
- Automated data transformations using AWS Lambda and Step Functions, eliminating over 20 hours of manual effort per month and improving processing reliability.
- Deployed monitoring and logging solutions with CloudWatch and CloudTrail, increasing ETL pipeline uptime to 99.8%, ensuring high availability for critical data processes.
- Leveraged S3 and Redshift Spectrum to process 5+ TB of data weekly, reducing infrastructure costs by 40% and accelerating analysis workflows.
- Automated ETL validation using AWS Glue and Python scripts, achieving 98% accuracy in data quality checks, ensuring high data integrity standards.
- Streamlined analytics by enhancing Redshift and Snowflake data models, improving report generation times by 35% and supporting data-driven insights.
- Resolved pipeline failures in real time using CloudWatch alerts and diagnostics, reducing incident resolution time from hours to minutes, ensuring business continuity.
- Built automated reporting systems and data lineage tracking, improving transparency across 100+ data workflows and facilitating compliance.
- Developed 15+ ETL pipelines with AWS Glue, transforming raw data from over 10 S3 buckets into structured formats, supporting efficient CDC implementations and near-real-time updates.
- Built dynamic dashboards with QuickSight and Tableau, improving data transparency for 20+ internal teams and reducing manual reporting time by 60%.
- Integrated CloudWatch with AWS EMR to pre-process 15+ TB of data, optimizing performance and costs through Hadoop and Spark-based workflows.
- Leveraged Databricks to build, optimize large-scale data pipelines for batch and streaming processing, improving workflow efficiency by 40% and reducing ETL processing times.
- Automated infrastructure provisioning using Terraform and CloudFormation, improving deployment consistency and reducing setup time across 10+ environments.
- Containerized 20+ ETL jobs in Docker, enabling scalable deployments across AWS ECS and on-prem staging setups, facilitating continuous integration.
- Documented and managed data transformation workflows using DBT, maintaining 98% accuracy in Snowflake models and improving change management.
- Executed migration of 25+ legacy data sources into Redshift and Snowflake, ensuring data integrity with zero data loss during transition periods.
- Secured data in transit and at rest through encryption and IAM role enforcement, maintaining full compliance with internal and external security standards.

Data Engineer, *Yes Bank* 04/2019 – 04/2020 | Mumbai, India

- Engineered 10+ ETL pipelines using Azure Data Factory, SSIS, and MS SQL Server, integrating data from over 8 source systems into Azure Synapse Analytics, adhering to RDBMS best practices.
- Architected dimensional and star-schema data models in Azure Synapse, boosting query performance by 40% and reducing storage costs by 25%, supporting scalable BI reporting.
- Automated data extraction and transformation across 12 systems by developing logic with Azure Functions and SQL Server stored procedures, reducing manual effort by 70%.
- Deployed 6 real-time dashboards with Power BI integrated with Azure SQL, enabling 24/7 interactive insights for 100+ users, improving decision response times.
- Implemented CI/CD pipelines using Azure DevOps and Git, decreasing ETL deployment time by 70% and supporting weekly iteration cycles.

- Leveraged Azure Blob Storage to store 10+ TB of structured and unstructured data, ensuring 99.9% data durability and fast retrieval speeds.
- Secured sensitive data by configuring Azure Active Directory and enforcing RBAC, effectively eliminating unauthorized access incidents.
- Directed migration of 30+ legacy SQL Server databases to Azure SQL Database, improving system performance by 50% and minimizing downtime during migration.
- Accelerated query execution by optimizing SQL logic and transformations within Azure Synapse, achieving 60% faster analytics processing.
- Enabled real-time data ingestion by integrating Azure Functions with Data Factory, decreasing data processing latency from 30 minutes to less than 2 minutes.
- Increased ETL pipeline reliability by 95% through automated data quality checks, detecting and preventing anomalies before data loads.
- Enforced data encryption standards across all Azure services, maintaining full compliance with GDPR and internal governance policies.

Education

Union Commonwealth University
Masters in Information Systems Management

Projects

Unstructured Data Ingestion and Processing Pipeline

- Engineered a comprehensive data pipeline to ingest and process unstructured data sources, including log files and multimedia content, utilizing Apache Spark with PySpark, Hadoop, and Databricks notebooks between 2021 and 2023. Developed custom ETL workflows integrating with AWS S3 and GCP Cloud Storage, enabling seamless multi-cloud data transfer and storage. Applied advanced data transformation techniques, improving data extraction efficiency by 50% and reducing processing times from several hours to under 30 minutes. Implemented data quality, security, and governance measures, resulting in a 40% reduction in data anomalies, while supporting scalable analytics platforms. The pipeline supported near-real-time data flow, facilitating more timely business insights and enhanced decision-making across healthcare analytics teams.