Pradeep

SUMMARY:

- Highly skilled and results-driven AWS Data Engineer with extensive experience in designing, developing, and managing scalable data pipelines on AWS cloud platforms.
- Proficient in utilizing a broad range of AWS services, including AWS S3, Redshift, Lambda, Glue, Kinesis, and EMR to build and optimize data architectures.
- Strong expertise in data integration, ETL processes, and real-time data streaming.
- Adept at ensuring data accuracy, integrity, and performance through best practices in data warehousing, big data processing, and automation.
- Experienced in collaborating with cross-functional teams to deliver actionable insights, enhance data workflows, and support data-driven decision-making.
- Excellent problem-solving skills with a focus on delivering high-quality solutions in fast-paced environments.

TECHNICAL SKILLS:

- **Languages:** Python, Pandas, Athena, Pyspark, PostgreSQL
- **Cloud Analytics:** AWS - Glue, Lambda, RDS, Airflow(MWAA), Kafka (MSK),Redshift, Quick sight, S3, SNS/SQS
- **Frameworks/Tools:** Flask, FastAPI, Data Bricks, CI/CD, PowerBI, PyCharm
- **Other Tools:** Git, Docker, Azure(ADF), ETL, ELT, MongoDB, Data Pipelines, Snowflake(Snowpark), Stream lit Dashboard, Web Development

Experience: Senior Data Engineer | Feb 2019 - Present

- **Customer Acquisition and Experience:** Collaborated with the Marketing/Revenue team to process campaign data and improve user targeting, driving customer acquisition and enhancing user experience.
- **ETL Pipeline Development:** Built and optimized ETL pipelines to integrate data from multiple sources into a data warehouse, enabling Data Science teams to create feature stores and perform predictive analysis on customer behavior.
  - Developed and scheduled Airflow DAGs to orchestrate ETL workflows, creating dimension and fact tables in Redshift and Athena.
  - Designed and implemented data pipelines for cleaning, transforming, and processing data across multiple layers (Bronze, Silver, Gold) using AWS Glue, Lambda, Databricks, and S3.
- **Geospatial Data Processing:**
  - Processed 70TB of geospatial data stored in S3 using EMR jobs with threading, improving post-processing concurrency.
  - Preprocessed geospatial data using geo-pandas in AWS Lambda.
- **REST API Development:** Designed and developed REST APIs to expose data insights, enabling real-time data access for applications and external stakeholders. Implemented authentication, rate-limiting, and caching mechanisms to ensure API security and performance. Utilized AWS Lambda for serverless API execution, ensuring scalability and cost efficiency.
- Automated API deployment and versioning using CI/CD pipelines with AWS CodePipeline and Terraform.

- **Workflow Optimization:**
  - Identified and optimized AWS Glue jobs, reducing execution time by 30% and cutting costs by 80% through threading implementation
  - Automated project tasks, significantly reducing workloads across teams.
- **Reporting and Visualization:** Migrated Oracle Discoverer Tableau reports to AWS QuickSight and developed a Streamlined Dashboard App in a Snowflake - SnowPark environment

KEY PROJECTS:

PCMS 2.0

- Enabled real-time monitoring of production data through a web UI for data-driven decision-making and report generation,including insights on energy consumption, alarms, and shift performance.
- Migrated on-premises applications to AWS, leveraging services like MSK, S3, Glue, Redshift, Lambda, SNS, and SQS, while implementing Infrastructure as Code (IaC) using Terraform/CloudFormation for automated resource provisioning and management.
- Developed an ETL pipeline to process Kafka streaming data into PostgreSQL and Redshift using AWS Glue and PySpark.
- Designed and implemented Power BI dashboards, visualizing insights from the Redshift data warehouse.
- Implemented JSON-based data ingestion and transformation processes in AWS Glue, enabling schema evolution and efficient data processing for semi-structured data, with S3 as the primary storage layer.
- Established a CI/CD pipeline using GitHub Actions and AWS CodeBuild to automate deployments of ETL pipelines and infrastructure changes

Theiox

- Led the complete backend development of key modules, including Data Acquisition, User Management, Alarm Configuration, Dashboards, and Reports, using Python-Flask. [Received company-wide recognition for the achievement.]
- Built real-time data ingestion pipelines to Kairos DB and Timescale DB from Kafka.
- Deployed and managed Docker services, troubleshooting and enhancing performance.
- Implemented Modbus, SNMP, MQTT protocols, and WebSocket for seamless IoT device communication.
- Designed and deployed dashboard and widget APIs with JWT-based user authorization for secure API access.
- Automated data processing workflows with AWS Glue ETL jobs, triggered by Amazon S3 events using AWS Lambda, integrating PySpark for large-scale distributed data transformations.
- Utilized AWS Glue Data Catalog to store and maintain a unified view of metadata for all datasets, enabling efficient SQL-based data querying and management.
- Implemented Infrastructure as Code (IaC) using Terraform and AWS CloudFormation to provision and manage cloud resources efficiently, ensuring scalability and consistency across environments.
- Developed and deployed CI/CD pipelines with Jenkins and Docker to automate the release of IoT applications and microservices

EDUCATION:
Bachelor of Computer Applications (BCA) From Bangalore University - 2018