**Pranitha Bollepalli**

**Professional Summary:**

* Data & Cloud Engineering Expert with overall 11 years of IT experience, including 10 years of experience in building cloud-native data platforms across AWS, Azure, and GCP, enabling secure, scalable, and analytics-ready environments.
* Designed and implemented end-to-end data pipelines using AWS Glue, EMR, Redshift, Snowflake, and Kinesis to support batch and real-time processing needs.
* Developed and maintained dbt models integrated with Snowflake to standardize transformations and improve data quality and governance.
* Created real-time ingestion workflows using Kafka, Kinesis, Lambda, and Step Functions for event-driven data processing and alerting.
* Designed cloud-native data lake architectures with schema enforcement and partitioning strategies across S3, ADLS, and GCS to enable scalable analytics and secure storage.
* Led data center-to-cloud migration initiatives, modernizing legacy Hadoop ecosystems and orchestrating seamless cutovers using Data Migration Services and IaC templates.
* Integrated data governance and lineage frameworks using Azure Purview, AWS Glue Catalog, and dbt, enabling traceability, audit readiness, and federated data validation.
* Embedded security, privacy, and compliance controls (HIPAA, PCI-DSS, SOC2) across data pipelines using IAM, encryption, RBAC, and VPC segmentation strategies.
* Automated CI/CD pipelines using Jenkins, Terraform, Argo, and GitHub Actions to standardize testing, validation, and deployment across multi-cloud environments.
* Orchestrated infrastructure deployment using CloudFormation and built containerized applications with AWS EKS to ensure scalability and automation.
* Applied monitoring and observability tools like CloudWatch and Athena to track pipeline performance, throughput, and data health.
* Collaborated with cross-functional teams including BI, data science, and compliance to deliver certified datasets, dashboards, and secure data access.
* Built and optimized data applications using Python, Scala, Spark, and Pandas, with unit and integration testing frameworks like Pytest and Scalatest.
* Collaborated with cross functional teams for deploying LLM-integrated pipelines for real-time insights and chatbot solutions using Hugging Face, Databricks, and SageMaker, enabling scalable semantic search and summarization use cases.

**Technical Skills:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | * **Cloud Platforms:** AWS (S3, Glue, Glue Catalog, Lambda, EMR, Redshift, Athena, Step Functions, Kinesis, MSK, DynamoDB, SNS, SQS, IAM, CloudTrail, CloudWatch, Macie, RDS, CodeBuild, CodePipeline, CodeCommit, EC2, EKS, EMR, QuickSight, Lake Formation, CloudFormation), Azure (Data Factory, Synapse Analytics, ADLS Gen2, Databricks, Azure Monitor, Azure SQL, Event Hubs, Azure Functions, Azure Key Vault, Azure Purview, Azure IoT Hub, AKS, HDInsight, ARM templates), GCP (BigQuery, Dataflow, Pub/Sub, DataProc, Cloud Functions, Vertex AI, Cloud Composer, GCS, GKE, Stackdriver) | | * **Big Data & Streaming:** Apache Spark (Core, SQL, Streaming, Structured Streaming, MLlib), PySpark, Apache Flink, Hadoop, Hive, HDFS, YARN, HBase, Kafka, NiFi, Kinesis, AWS IoT Core, Azure IoT Hub, Pub/Sub, MapReduce | | * **ETL & Orchestration:** Apache Airflow, Cloud Composer, Azure Synapse Pipelines, DBT, Step Functions, metadata-driven pipelines. | | * **Data Warehousing, Data Lake & Storage:** Snowflake, Amazon Redshift, Google BigQuery, Azure Synapse, Hive, Delta Lake, Apache Iceberg, DynamoDB, MongoDB, Cassandra, HBase, Cosmos DB, Bigtable, MySQL, PostgreSQL, Oracle, PL/SQL, Amazon RDS, GCS, S3, ADLS Gen2, Apache Hudi * **Data Modeling & Governance:** Star Schema, Snowflake Schema, Fact/Dimension Modeling, Data Vault, Logical & Physical Data Models, Collibra, Meta Data Management * **Programming Languages & Scripting:** Python, Java, Scala, SQL,R | | * **Frameworks, APIs:** REST API, FastAPI * **AI, Machine Learning & NLP:** Databricks AutoML, SageMaker, Azure ML, Vertex AI, MLflow, Scikit-learn, Pandas, NumPy, Sentiment Analysis, Predictive Modeling * **Data Visualization & Reporting:** Tableau, Power BI, Amazon QuickSight, Looker, Google Data Studio (Looker Studio), Kibana * **Security, Privacy & Compliance:** IAM (AWS, Azure, GCP), Role-Based Access Control (RBAC), Azure Key Vault, AWS KMS, HIPAA, PCI-DSS, SOC2, GDPR | | * **DevOps, CI/CD & Infrastructure Automation:** Jenkins, GitHub, GitLab, Docker, Kubernetes, Terraform, CloudFormation, TDD. * **Monitoring, Logging & Observability:** AWS CloudWatch, Azure Monitor, Log Analytics | | * **Project Delivery, Agile & Documentation:** Agile, Scrum, Sprint Planning, Backlog Grooming, Daily Standups, Stakeholder Collaboration, Requirement Analysis, System Design & Architecture Diagrams, API Documentation, User-Centered Design (UCD), Technical Mentorship, Cross-functional Collaboration, Business Engagement, Regulatory Alignment * **Domain Expertise:** Healthcare (HIPAA), Finance (PCI-DSS, SOX), E-Commerce, Federal/DoD (Security Accreditation), Retail, Marketing, Sales, Inventory | |

**Education**

* **Bachelor’s and Master's in Computer Science**
* **Saint Louis University, St. Louis, Misso**

**Professional Experience**

**Lead Data Engineer – AWS**

**Federal Home Loan Bank of Chicago, Chicago, IL Jan 2023 – Present**

* Directed modernization of existing AWS workloads by replacing legacy infrastructure with scalable, fully managed services like MSK, Glue, and EMR.
* Architected scalable cloud infrastructure using AWS services such as S3, Glue, EMR, Athena, Redshift, EC2, VPC, IAM, and CloudWatch to support high-performance data lakes and analytics pipelines.
* Designed and managed EC2-based compute clusters, customizing instance configurations across multiple Linux environments to meet workload-specific demands.
* Implemented secure and compliant AWS environments, incorporating IAM roles, encryption best practices, and Amazon Macie to protect sensitive financial data (GDPR).
* Refactored legacy SQL pipelines into dbt, significantly improving maintainability and reducing query execution time by 40%.
* Drive best practices in data modeling, version control, and testing with dbt, enhancing pipeline governance and maintainability.
* Designed and developed end-to-end ETL pipelines using AWS Glue, Apache Spark, Kinesis, and Airflow, automating ingestion, cleansing, transformation, and scheduling workflows.
* Integrated structured and semi-structured sources including Hive, MySQL, and Amazon RDS with Apache Spark and Python for unified batch and real-time processing.
* Refined and streamlined Apache NiFi pipelines, optimizing processor configuration, flow file handling, and throughput for high-volume ingestion use cases.
* Utilized Snowflake and Redshift to architect scalable data warehouses, enabling enterprise-wide ad-hoc analytics and BI across complex data domains.
* Implemented dbt (Data Build Tool) for modular data transformation, testing, versioning, and documentation—ensuring clean, analytics-ready data layers.
* Optimized storage and querying strategies through intelligent partitioning, denormalization, and caching in AWS and hybrid environments.
* Developed event-driven architectures leveraging AWS Kinesis, Lambda, and Step Functions to enable real-time streaming ingestion and trigger-based workflows.
* Automated cloud resource provisioning via AWS CloudFormation and Step Functions, ensuring repeatable deployments across development, test, and production environments.
* Deployed and fine-tuned Amazon EMR clusters, enhancing distributed processing performance for large-scale data transformation and analytics workloads.
* Continuously monitored and tuned resource usage with AWS CloudWatch, Auto Scaling Groups (ASG), and job-level optimization to minimize costs and maximize throughput.
* Engineered robust error handling, logging, validation, and data lineage frameworks across all pipelines to ensure quality, auditability, and fault tolerance.
* Collaborated cross-functionally with DevOps, data science, compliance, and security teams to align data architecture with enterprise goals and regulatory needs.
* Supported governance and compliance through role-based access control (RBAC), encryption, and comprehensive lineage tracking in cloud data platforms.
* Streamlined complex data ingestion workflows using Python,. increasing pipeline throughput and reducing operational latency.

**Senior Data Engineer - GCP & AWS**

**Shell, Houston, TX Nov 2020 – Dec 2022**

* Designed and developed scalable ETL/ELT pipelines using Google Dataflow, Dataproc, Apache Spark, and PySpark to process large volumes of structured and unstructured data.
* Architected and implemented centralized data platforms using Snowflake to enable cross-functional data access and governance.
* Integrated BigQuery ML for building and deploying machine learning models within the data warehouse environment.
* Built and managed infrastructure as code (IaC) using Terraform for automating GCP and AWS resource provisioning.
* Deployed scalable ML/AI workloads using Google Kubernetes Engine (GKE) and Docker, optimizing compute resource usage.
* Utilized Google Cloud Functions to enable lightweight, event-driven orchestration across data pipelines and services.
* Enforced security and compliance standards using IAM, VPC Service Controls, and encryption at rest/in transit.
* Monitored and optimized Apache Kafka for real-time streaming, ensuring high availability and secure communication via TLS, Kerberos, and ACLs.
* Collaborated with data scientists and ML engineers to operationalize machine learning models using Databricks and CI/CD pipelines.
* Integrated and synchronized multi-cloud data sources (AWS S3, GCS, RDS, Cloud SQL) for unified analytics and reporting.
* Implemented data validation, lineage, and auditing frameworks to ensure quality, traceability, and compliance across pipelines.
* Used Cloud Composer (Apache Airflow) to schedule and manage complex workflows and data dependencies.
* Optimized storage and compute costs through partitioning, clustering, and resource monitoring via Cloud Monitoring and AWS CloudWatch.
* Led Agile ceremonies and collaborated with product owners, analysts, and engineers to align engineering tasks with strategic business goals.
* Contributed to a regulatory reporting system that required high data accuracy; used dbt for building tested, version-controlled SQL models on top of Snowflake.
* Migrated existing SQL-based transformation scripts into dbt, helping the team streamline workflows and reduce manual errors.
* Integrated dbt models into Airflow DAGs to automate the daily transformation and validation of data stored in Bigquery.

**Tools & Technologies: G**oogle Dataflow, Dataproc, Apache Spark, PySpark, Google BigQuery, Snowflake, BigQuery ML, Terraform, GCP, AWS, dbt, Google Kubernetes Engine (GKE), Docker, Google Cloud Functions, IAM, VPC Service Controls, TLS, Kerberos, ACLs, Apache Kafka, Databricks, CI/CD, AWS S3, Google Cloud Storage (GCS), Amazon RDS, Cloud SQL, Cloud Composer, Apache Airflow, Cloud Monitoring, AWS CloudWatch.

**Senior Cloud Engineer – Azure**

**Walgreens Boots Alliance, Deerfield, Illinois Sep 2018 – Oct 2020**

* Migrated enterprise data assets from Oracle and SQL Server to Azure Blob Storage and HDInsight Hive using Azure Data Factory, ensuring accuracy, scalability, and business continuity.
* Engineered Hive partitioning strategies on Azure HDInsight, significantly reducing query times and enhancing processing throughput.
* Boosted performance of distributed workloads by implementing RDD caching in Azure Databricks for large-scale Spark operations.
* Orchestrated real-time data ingestion using Azure Stream Analytics, enabling live analytics across critical data streams.
* Applied enterprise-grade security protocols using Azure Active Directory and Azure Key Vault for encrypted and controlled data access.
* Architected storage solutions using Azure Data Lake, Blob Storage, and Azure Synapse Analytics tailored to analytic and compliance needs.
* Leveraged Azure Monitor and Logic Apps for real-time tracking, process automation, and system health alerts across the migration lifecycle.
* Validated data accuracy and scalability across Oracle, MS SQL Server, and NoSQL environments through rigorous testing.
* Automated file transfers and job scheduling on Azure Linux VMs using UNIX shell scripts, enhancing operational consistency.
* Refactored legacy MapReduce workloads to PySpark in HDInsight, improving scalability and reducing compute costs.
* Tuned MySQL and NoSQL databases for faster performance, applying insights from continuous monitoring and diagnostics.
* Consolidated multi-source data (RDDs, JSON, databases) into Azure Databricks for unified analytics workflows.
* Developed insightful dashboards using Power BI, delivering clear, actionable intelligence to non-technical stakeholders.
* Wrote modular, optimized Python and Scala code within Databricks, utilizing native libraries for enhanced processing performance.
* Automated complex ETL workflows with shell scripts, ensuring fault-tolerance and efficient data movement across Azure services.

**Tools & Technologies:** Azure, Databricks, Azure Data Factory, Apache Spark, PySpark, Azure Blob Storage, Azure SQL Data Warehouse (Synapse), Azure Data Lake, Azure Key Vault, Python, Scala, Power BI, ETL Automation, Cloud Migration, Real-Time Processing, Data Security.

**Senior Data & Cloud Engineer – AWS**

**Bayer Pharmaceuticals, Whippany, NJ Mar 2017 – Aug 2018**

* Implemented ETL workflows using AWS Glue for data extraction, transformation, and loading into various data stores.
* Configured and managed AWS EMR clusters for distributed data processing and analytics tasks.
* Developed ETL (Extract, Transform, Load) pipelines to ingest and transform data from various sources into Snowflake.
* Utilized AWS S3 for scalable and cost-effective storage of big data sets.
* Developed serverless data processing applications using AWS Lambda and Step Functions for event-driven workflows.
* Set up monitoring and alerting using AWS CloudWatch to proactively detect and address performance issues in data pipelines.
* Implemented data streaming solutions with AWS Kinesis for real-time data ingestion and processing.
* Created continuous integration and delivery pipelines with AWS CodePipeline for automated deployment of data applications.
* Loaded curated data into Snowflake, designing incremental loading workflows with COPY INTO and Snowpipe for patient-level fact tables and clinical dimension tables.
* Automated infrastructure provisioning using AWS CloudFormation for scalable and repeatable deployments.
* Implemented event-driven data processing and notifications using AWS SNS and SQS.
* Developed NoSQL data models and managed data in AWS DynamoDB for flexible and scalable data storage.
* Set up and optimized AWS Redshift clusters for fast and efficient data warehousing and analytics.
* Utilized AWS Redshift Spectrum for querying data directly from S3, enabling cost-effective storage and analysis.
* Performed ad-hoc querying and analysis on large datasets using AWS Athena.
* Provisioned and managed AWS EC2 instances for flexible and scalable computing resources.
* Processed real-time Kafka streams and visualized key metrics on operational dashboards.
* Developed performance and maintainable data applications using both Python and Scala for transformation, analytics, and processing.
* Orchestrated containerized applications using AWS Elastic Kubernetes Service (EKS) for scalable and automated deployment.
* Developed Spark applications for distributed data processing and analysis.
* Leveraged Python and Pandas for data manipulation, cleansing, and preprocessing tasks.
* Developed data processing applications using Scala for enhanced performance and parallel processing.
* Created and optimized SQL queries for data transformation and analysis.
* Implemented Bash scripting for automating data workflows and managing data processing tasks.
* Implemented test cases using Pytest and Scala test to ensure code reliability and functionality.

**Tools & Technologies:** AWS Glue, AWS EMR, PySpark. Amazon EMR, Apache Spark, EC2, Snowflake, dbt, SQL, Python, Amazon S3, AWS Lambda, AWS Step Functions, AWS CloudWatch, SNS, SQS, Amazon Kinesis, GitHub, AWS DynamoDB, Amazon EKS, Pytest

**Hadoop Data Administrator**

**Adobe, San Jose, CA Jan 2015 – Feb 2017**

* Designed and maintained robust ETL pipelines leveraging Python to cleanse, profile, and transform raw datasets into structured formats for downstream analytics.
* Engineered efficient data storage and retrieval strategies by creating and tuning external Hive tables using HQL for high-volume query performance.
* Utilized Sqoop for bulk data movement between Oracle databases and HDFS, and configured Flume agents to stream real-time log data into the Hadoop environment.
* Modernized legacy batch processes by introducing Apache Spark components such as Spark SQL, DataFrames, and Paired RDDs, significantly accelerating compute operations.
* Built scalable data ingestion pipelines using MapReduce, HBase, and Hive to handle structured and semi-structured data from various internal and external sources.
* Developed high-throughput Spark applications in Scala for rapid batch processing, reducing ETL execution time and improving overall testability.
* Conducted performance tuning and capacity planning to ensure cluster reliability and efficiency across multi-node Hadoop environments.
* Ensured structured data was consistently ingested into HDFS in CSV format to support seamless integration with Spark-based analytics workloads.
* Automated job scheduling and monitoring using Oozie and Shell scripts to orchestrate complex workflows and ensure timely data availability across the platform.
* Implemented data quality checks and validation frameworks to detect anomalies **early in the ETL** lifecycle, improving the accuracy and trustworthiness of analytics outputs.

**Tools & Technologies:** Hadoop, HDFS, Hive, HQL, Python, Oracle, Sqoop, Flume, Apache Spark, Spark SQL, Spark Context, DataFrames, Spark YARN, Scala, Paired RDDs, HBase, MapReduce, ETL pipelines, data ingestion, real-time streaming, cluster administration, performance optimization, CSV, relational databases.

**Data Analyst**

**Salesforce, San Francisco, CA Jan 2014 – Dec 2014**

* Gather and integrate data from various internal and external sources into Salesforce’s data infrastructure, ensuring data accuracy, consistency, and alignment with business needs.
* Utilize SQL, Salesforce’s reporting tools, and other analytics platforms to perform detailed analysis of customer, sales, and product data, providing actionable insights to support decision-making across departments.
* Design and develop interactive dashboards and data visualizations using Salesforce’s Analytics Cloud (Tableau CRM), enabling teams to track key performance indicators (KPIs) and trends in real-time.
* Implement data validation rules, cleansing procedures, and establish best practices for maintaining high-quality data within the Salesforce ecosystem, reducing errors and improving data-driven decision-making.
* Collaborate closely with business leaders, product teams, and marketing to understand data requirements, deliver tailored reporting solutions, and assist in defining key metrics that align with organizational goals