HARSHAVARDHAN ANNAPUREDDY

[annapureddy687@gmail.com](mailto:annapureddy687@gmail.com%20)

**Ph:**

**LinkedIn:** [Harshavardhan A | LinkedIn](https://www.linkedin.com/in/vardhan686/)

# PROFESIONAL SUMMARY

Innovative Data Engineer with 6 years of experience in Information Technology. Recognized consistently for performance excellence and strong technical skills. Strengths in Programming, Databases and Cloud Technologies with expertise in Python, AWS, Databricks, SQL Server, and Snowflake.

# TOOLS AND TECHNOLOGIES

* **Programming Languages –** Python, Shell Scripting, PySpark, MATLAB.
* **Query Languages -** SQL, NoSQL.
* **Databases -** Snowflake, MySQL, PostgreSQL, Oracle, Teradata, MongoDB, Amazon Redshift.
* **ETL/BI Tools -** Tableau, Power BI, MS Excel, OBIEE, DataStage.
* **Amazon Web Services -** IAM, EC2, ECS, Lake Formation, Lambda, Step Functions, S3, CloudWatch, API Gateway.
* **Azure –** Azure Data Factory, Azure Synapse Analytics.
* **Hadoop Framework –** HDFS, HBASE, MapReduce, Yarn, HIVE, PIG, Zookeeper, Oozie, Kafka.
* **Operating Systems –** Unix (Linux, macOS), Windows.
* **Other Tools** - Microsoft Office Suite, LabVIEW, GIT, Wireshark, DbVisualizer, FileZilla, Erwin, Docker, Kubernetes.

# WORK EXPERIENCE

**Zions Bancorporation** **-** **Dallas, TX Jun 2021 – Current**

**Big Data Engineer**

**Responsibilities:**

* Daily activities include extracting, transforming, integrating, and loading client data, developing batch, and streaming data pipelines, performing data analysis, supporting production activities, and resolving production issues.
* I have automated the ingestion and storing of data points within the Snowflake Data Platform to provide a consolidated budget report for the business users to extract integration budget metrics within a single report.
* Created automation scripts to streamline ETL processes, Data imports/exports, and API pulls using Python and AWS services such as Lambda, EC2, S3, Step Functions and API Gateway.
* Played a vital role in configuring Amazon Web Services cloud setup and successfully performed complete migration of data, scripts, ETL processes, schedules, applications, tools, SFTP connections, data models, schemas, tables, stored procedures, sequences and BI dashboards from physical databases and Linux servers to AWS and Snowflake.
* I have implemented the adoption of serverless architectures, utilizing AWS Lambda for automated cron job migrations and building Restful Micro Services using Django.
* Designed detail migration plans for workloads from third party data and cloud platforms to AWS and executed them successfully.
* Played a vital role in configuring Amazon Web Services cloud setup and successfully performed complete migration of data, scripts, ETL processes, schedules, applications, tools, SFTP connections, data models, schemas, tables, stored procedures, sequences and BI dashboards from physical databases and Linux servers to AWS and Snowflake.
* Utilized AWS Glue Data Catalog to catalog and discover datasets within Databricks notebooks, streamlining metadata management and promoting data governance best practices.
* Used AWS Key Management Service (KMS) for encryption of data at rest and in transit within Databricks, ensuring compliance with regulatory requirements and data security standards.
* Implemented cost-effective solutions within EventBridge by optimizing resource utilization and leveraging on-demand scaling, resulting in measurable cost savings for the organization.
* Utilized AWS Lambda functionality in conjunction with EventBridge for seamless event processing.
* Designed and implemented micro-services architecture with a focus on decoupling components using Amazon EventBridge, fostering agility and scalability.
* Leveraged AWS Identity and Access Management (IAM) roles to securely authenticate and authorize Databricks clusters, ensuring granular control over access permissions and data confidentiality.
* Developed automated monitoring and alerting system using CloudWatch, SNS and Lambda functions.
* Used Amazon kinesis to capture and analyze log data from applications and servers.
* Designed and developed scalable solutions leveraging big data and container technologies like Spark, Docker, and Kubernetes to ingest, process and analyze large datasets to exceed business requirements.
* Built data infrastructure and automated ETL processes using PySpark and Python and designed semantic layers for business reporting.
* Developed PySpark framework to run data quality analysis on large datasets and visualized scorecards using Tableau.
* Created data ingestion pipelines utilizing PySpark and Scala for processing structured files stored on AWS S3. These pipelines are designed to seamlessly load data into Snowflake, incorporating advanced features like automated backlog file processing, robust schema and data validation, and comprehensive error handling.
* Developed the PySpark code for AWS glue jobs and integrated them with AWS data lake and its services.
* Used Spark's Java API to read data from various sources such as files (CSV, JSON, Parquet), and databases.
* Developed Python scripting to conduct rigorous testing and debugging activities, aligning with design specifications to validate functionality and resolve defects early in the development lifecycle.
* Utilized Databricks for building and optimizing data warehouses on Delta Lake, enabling scalable and reliable analytics while fostering effective collaboration through version control features.
* Created automated workflows on Databricks using GitHub to streamline the Continuous Integration/Continuous Deployment (CI/CD) process and managed code deployment.
* Developed detailed SQL. logic and designed Stored Procedures for different types of tables (Dimension, FACT, Lookup, Aggregate, and Flatten) in Snowflake, catering to various data models. This facilitated the smooth data transfer between different layers like Raw, Staging, Core, and Refined to meet specific business needs.
* Designed and implemented audit control framework for the stored procedures, involving the logging of data in Snowflake tables and returning a JSON payload containing logs to the invoking application.
* Process and load bound and unbound Data from Google pub/subtopic to BigQuery using cloud Dataflow with Python.
* Architected and developed microservices using Go, reducing response times by 40% and improving system scalability.
* Collaborated with cross-functional teams to integrate Go modules into existing projects, enhancing overall performance and maintainability.
* Implemented custom matching rules and thresholds in Apache Flink to enhance data accuracy.
* Used Flink’s probabilistic matching capabilities to cleanse and standardize large datasets, resulting in improved data quality.
* Utilized Apache Flink for large-scale batch data processing, achieving high throughput and fault tolerance.
* Created reports and data visualization dashboards using complex SQL logic and BI tool Tableau.
* Used ETL tools like Informatica to design data pipelines and mappings to efficiently process data.
* Optimized Flink jobs for performance by analyzing execution plans, parallelism, and resource utilization.
* Worked on collecting and analyzing large real-time streams of data using Kafka and Flink.
* Implemented and automated performance tuning and optimization of data processing jobs, employing techniques like query optimization, indexing, and caching to improve query performance and reduce latency.

# Environment: AWS, AWS EventBridge, GCP, Spark Core, Spark SQL, Snowflake, Databricks, BigQuery, PySpark, Kafka, Tableau, MySQL, Linux, PySpark, MapReduce, Apache Flink.

# Aisle Rocket - Chicago, IL Aug 2019 – Jun 2021

# Data Engineer

**Responsibilities:**

* Developed logical and physical data models using Erwin Data Modeler and system standards and architecture.
* Developed and implemented ETL workflows using Java and Spark, ensuring efficient data extraction from various sources, transformation, and loading into target systems.
* Developed optimized Java and Spark code for large-scale data processing systems, ensuring scalability and performance under high-volume data scenarios.
* Created Pipelines in Azure Data Factory using Linked Services, Datasets, and Pipeline to Extract, Transform, and load (ETL) data in Azure Databricks from different sources like Azure SQL, Data Lake storage, Azure SQL Data warehouse, write-back tool, and backward.
* Worked on Azure BLOB and data lake storage and loading data into Azure SQL Synapse analytics.
* Developed ETL workflow which pushes web-server logs to an Azure Blob Storage.
* Developed Azure functions to automate workflows, reducing manual intervention, and improving the efficiency of data processing tasks.
* Worked on the design and development of ADF Pipelines to transform data from heterogeneous sources to Azure SQL DW/Synapse.
* Worked with Azure key vault, Logic apps, and Automation account.
* Worked Extensively with MS SQL Server Integration Services (SSIS) to import, export and transform data between servers.
* Developed stored procedures, triggers, and functions to account for automated insertion of records into the tables.
* Used BCP utility and bulk insert to load data from different data sources to SQL Server and used SQL Profiler to monitor the server performance, debug T-SQL and slow running queries.
* Administered database performing activities like managing users and roles, DDL and DML operations, and database tuning.
* Automated all the jobs for extracting the data from different Data Sources like MySQL to pushing the result set data to Hadoop Distributed File System.
* Developing Scripts and Batch Job to schedule various Hadoop Program.
* Implemented robust quality control measures throughout the data lifecycle, from ingestion to consumption, to uphold data integrity and reliability.
* Optimized data structures to ensure efficient storage, retrieval, and manipulation of large-scale datasets, balancing performance, and scalability requirements.
* Worked on analyzing Hadoop cluster and different bigdata analytic tools including HBase and Sqoop.
* Performed extensive data QA by running complex SQL queries to make sure the data is migrated properly, joins between the tables are working as expected and the sequences are properly set.
* Directed the reporting tools SQL data sources to the migrated server databases and fixed the issues as needed.
* Configured tasks like performance tuning, indexing, stored procedures, checking deadlocks, creating jobs and scheduling.
* Developed data pipeline using Kafka, Spark and Hive to ingest, transform and analyzing data.
* Developed Spark scripts by using Scala as per the requirement.
* Developed custom data processing logic in Scala, including data cleansing, enrichment, aggregation, and statistical analysis, to extract actionable insights and drive data-driven decision-making processes.
* Enforced data governance policies within Kafka and Scala workflows, conducted code reviews, and implemented best practices in Scala.
* Applied functional programming principles in Scala, including immutability, higher-order functions, and pattern matching, to architect and implement concise, resilient, and maintainable data processing logic for diverse data pipelines.
* Utilized Spark's Java API to apply data transformations, filtering, aggregations, joins and other required data manipulations.
* Successfully optimized ETL processes by leveraging Spark's distributed computing capabilities, leading to significant performance improvements and reduced processing times.
* Responsible for writing Hive queries for data analysis to meet the business requirements.
* Responsible for importing and exporting data into HDFS and Hive using Sqoop.
* Apply different HDFS formats and structure like Parquet, Avro, etc. to speed up analytics.
* Worked extensively in creating hive partition and bucket tables on HDFS.
* Involved in creating Hive tables, loading with data, and writing hive queries which will run internally in map reduce.
* Designed and implemented MapReduce based large-scale parallel relation-learning system.
* Developing parser and loader MapReduce application to retrieve data from HDFS and store to HBase and Hive.
* Involved in developing Spark code using Scala and Spark-SQL for faster testing and processing of data.
* Ingested data from Snowflake views, Amazon S3, and other external sources using Power Query and custom connectors.
* Enhanced performance of Power BI reports by optimizing Snowflake SQL queries, utilizing clustering keys, partitioning, and query profiling techniques.
* Created advanced analytical reports with custom Power BI visuals, including R & Python visuals.
* Managed and visualized large datasets by implementing incremental data refreshes, aggregations, and dataflows in Power BI.
* Used OBIEE to monitor account activity in real-time and create alerts when certain thresholds or patterns are exceeded, such as many high-value transactions in a short period of time.
* Utilized Bitbucket as a central repository for version control, branch management, and code collaboration.

# Environment: Spark Core, Spark SQL, Azure Data Factory, Azure Synapse Analytics, OBIEE, Power BI, Java, Scala, Kafka, Hive, Tableau, MySQL, Linux, Bitbucket.

# Smart Search Solutions – Hyderabad, India Dec 2016 – Dec 2017

**Software Developer**

**Responsibilities:**

* Created the data lake in S3 bucket with structure and unstructured data.
* Collaborated with other teams to design and implement Salesforce Integrations with other internal applications using SOAP and REST web services to communicate with other internal systems.
* Extracted the data from various SAP systems (S/4Hana, ECC, BW) and Salesforce and load it into the S3.
* Using CLI commands moved the data from local server to S3 bucket.
* Created the AWS glue jobs to load the data from on-premises databases to S3 bucket.
* Created AWS glue job to move data from S3 bucket to redshift DB.
* Used the Athens to Query the structure files in S3 bucket.
* Developed the PostgreSQL tables to store the metadata information.
* Designed detail migration plans for workloads from third party data and cloud platforms to AWS and executed them successfully.
* Transformed, manipulated, and cleansed the data using Python.
* Setup Spark EMR to process huge data which is stored in Amazon S3
* Worked on AWS environment and technologies such as S3, EC2, EMR, Glue, CFT, Lambda and databases Oracle, SMS, DynamoDB, MongoDB.
* Used version control systems such as Git and collaboration platforms like Jira and Confluence, facilitating effective teamwork and project management within data engineering teams.
* Designed detail migration plans for workloads from third party data and cloud platforms to AWS and executed them successfully.
* Developed Python scripts to parse application log files, later replaced this step with AWS native monitoring tools - CloudWatch, SNS and Lambda.
* Implemented logging framework - ELK stack (Elastic Search, Logstash Kibana) on AWS.
* Setup Spark EMR to process huge data which is stored in Amazon S3.
* Worked on CI/CD deployments using Jenkins.
* Performing Data validation operation between ETL and Apache Hive tables.
* Implementing the Partitioning and bucketing for faster query processing in Hive Query Language.
* Developed the shell scripts for migrating and deploying in Production Servers.
* Used IBM DataStage tool to perform Extract, Transform, and Load (ETL) operations.
* Worked on DataStage client tools like DataStage Designer, DataStage Director and DataStage Administrator.
* Experienced in scheduling sequence, parallel and server jobs using DataStage Director, UNIX scripts and scheduling tools.
* Designed and developed parallel jobs, server and sequence jobs using DataStage Designer.

**Environment:** Python, Salesforce, SAP, EC2, S3, RDS, VPC, Lambda, SNS, CloudWatch, Redshift, Glue, Athena, Oracle database, Data warehousing, ETL, GIT, Jira, PLSQL, Unix shell scripting.

# CERTIFICATIONS AND TRAININGS

* AWS Certified Data Engineer – Associate (Validation number: 2648c06994594e9797250a839c83d901).
* Python and SQL Skills Certified – Hacker Rank.
* Big Data Specialization and Python Specialization – Coursera.

# EDUCATION

**Master of Science, Computer and Information Science - 2019**

Southern Arkansas University, Magnolia, AR

# Bachelor of Technology, Electronics and Communication Engineering - 2017

KL University, Andhra Pradesh, India