**Shiva Chitneni**

**BIG DATA, PYTHON, AWS, GCP CLOUD ENGINEER**

[**shivacharan6991@gmail.com**](mailto:shivacharan6991@gmail.com)

**409-422-9094**

**PROFESSIONAL SUMMARY**

* 8 years of technical expertise in complete software development life cycle (SDLC), which includes 5 years of Hadoop Development and 4 years of Python Development, Design, and Testing.
* Hands-on experience working with Apache Spark and Hadoop ecosystems like MapReduce (MRv1 and YARN), Sqoop, Hive, Oozie, Flume, Kafka, Zookeeper, NoSQL Databases like Cassandra and Orchestration tools like **Airflow**, Data Pipelines, CloudFormation.
* 2+ years of Experience in AWS services such as EC2, ELB, Auto-Scaling, EC2 Container Service, S3, IAM, VPC, RDS, DynamoDB, Cloud Trail, Cloud Watch, Lambda, Elastic Cache, SNS, SQS, Cloud Formation, Cloud Front, EMR, AWS code deploy, Serverless Deployment.

**Airflow:**

* Worked on **Airflow** 1.8(Python2) and **Airflow** 1.9(Python3) for orchestration and familiar with building custom Airflow operators and orchestration of workflows with dependencies involving multi-clouds.
* Orchestration experience using Azure Data Factory, **Airflow** 1.8 and **Airflow** 1.10 on multiple cloud platforms and able to understand the process of leveraging the **Airflow** Operators.
* Built **Airflow** pipelines Migrated Petabytes of data from Oracle, Hadoop, MSSQL, MySQL sources to the AWS cloud.

**Apache Spark:**

* Excellent experience with Spark Core architecture.
* Hands-on expertise in writing different RDD (Resilient Distributed Datasets) transformations and actions using Scala, Python, and Java.
* Created Data Frames and performed analysis using Spark SQL & PySpark Transformations.
* I worked on Spark Streaming and Spark Machine Learning Libraries.

**MapReduce and HDFS:**

* Excellent understanding/knowledge and working experience on HDFS architecture and various components such as HDFS, Name Node, Job Tracker, and Task Tracker.
* Experienced in writing MapReduce programs using Java API.
* Implemented MapReduce programs to perform joins using secondary sorting and distributed cache.
* Implemented Custom Input Format, and Custom Record Reader for MapReduce.

**Apache Sqoop:**

* Used Sqoop to Import data from Relational Database (RDBMS) into HDFS and Hive, storing using different formats like Text, Avro, Parquet, Sequence File, ORC File, and compression codecs like Snappy and Gzip.
* Performed transformations on the imported data and exported it back to RDBMS.

**Apache Hive:**

* Implemented Partitioning and Bucketing on Hive tables for Hive Query Optimization.
* Experience in writing queries in HQL (Hive Query Language), to perform data analysis.
* Created Hive External and Managed Tables.

**Apache Oozie:**

* Experienced in writing Oozie workflows and coordinator jobs to schedule sequential Hadoop jobs.

**Apache Flume and Apache Kafka:**

* Implemented custom interceptors for flume to filter data and defined channel selectors to multiplex the data into different sinks.
* Used Apache Flume to ingest data from different sources to sinks like Avro, HDFS.
* Excellent knowledge and hands-on experience in Fan Out and Multiplexing flow.
* Excellent knowledge of Kafka Architecture.
* Integrated Flume with Kafka, using Flume both as a producer and consumer (concept of FLAFKA).
* Used Kafka for activity tracking and Log aggregation.

**SQL and NoSQL:**

* Worked on Relational Databases like MySQL.
* Ability to write complex SQL queries to analyze structured data.
* Strong understanding of Cassandra architecture and Data Modelling.

**Version Control and Build Tools:**

* Experienced in using GIT and SVN.
* Ability to deal with build tools like Apache Maven and SBT.

**Python:**

* Utilized the concepts of multi-threaded programming in developing applications.
* Extensively involved in developing and consuming web services/API’s/micro-services using requests library in python, implemented security using OAuth2 protocol, etc.
* Experience in developing web-based applications using Python 3.x (3.6/3.7), Django 2.x and Flask.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Languages | Scala, Python, R, Java |
| Big Data Ecosystem | Hadoop, HDFS, Map Reduce, YARN, Sqoop, Flume, Hive, Hue, PIG, HBase, Oozie, Zookeeper, Spark, Scala, Kafka, Spark ML, Stream Sets, Kudu |
| Cloud Technologies | AWS, Terraform, CloudFormation, Serverless Framework, Azure, GCP |
| Scripting Languages | Python, Shell Scripting |
| IDE | Eclipse IDE, IntelliJ, NetBeans, Jupyter, SSMS, Visual Studio |
| Databases | Oracle 11g, MySQL, Teradata, MS SQL Server, Cassandra, Cosmos DB, DB2 |
| BI Tools | SQL Server Reporting and Analysis Services (SSRS & SSAS) |
| Build Tools | ANT, Maven, SBT |
| Application Servers | WebLogic 10.3, Apache Tomcat 6/7, JBoss |
| Web Services | SOAP (JAX-WS), WSDL, SOA, Restful (JAX-RS), JMS |
| Version Control System | GIT, TFS, SVN, CVS |
| Other Tools | Junit, Putty, Jira, Airflow |
| Process methodologies | Agile, Waterfall |
| Operating Systems | Linux, UNIX, Windows |

**PROFESSIONAL EXPERIENCE**

**Client: Disney Streaming April 2023 – Till date**

**Role: Data Engineer \ Data Analyst**

**Responsibilities**:

Our project’s goal is to design a Single Customer View (SCV), we gathered and processed data from customer centric data stores to obtain the data of every customer as a single record. Involved in building an Enterprise data Lake to bring ML ecosystem capabilities to production and make it readily consumable for data scientists, researchers, and business users. Integrated data from business and analytical applications to anticipate customer needs, patterns and provide actionable insights.

* Worked with the Hive for improving the performance and optimization in Hadoop using components.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Designed good understanding of Partitions, bucketing concepts in Hive, and designed both Managed and External tables in Hive to optimize performance.
* Created Hive external tables, views, and scripts for transformations such as filtering, aggregation, and partitioning tables.
* Followed agile methodology and SCRUM meetings to track, optimize and tailored features to customer needs.
* Gained very good business knowledge of a different category of products and designs within.
* Involved in developing Thought spot reports and workflows automated to load data.
* Implement Spark Kafka streaming to pick up the data from Kafka and send it to Spark pipeline.
* Develop python scripts to schedule each dimension process as task and set dependencies for the same.
* Good Understanding of Data ingestion, **Airflow** Operators for Data Orchestration and other related python libraries.
* Develop, tested and deployed python scripts to create **airflow** DAGS, Integrate with Databricks using **airflow** operator to run Notebooks on scheduled basis.
* Built an ETL framework for Data Migration from on premise data sources such as Hadoop, Oracle to AWS using **Apache Airflow**, Apache Sqoop and Apache Spark (PySpark).
* Designed library for emailing executive reports from Tableau REST API using python, Kubernetes, Git, AWS Code Build, and **Airflow**.
* Designed and implemented Confidential Serverless Backend leveraging AWS Amplify REST APIs, GraphQL APIs (DynamoDB) and S3 Storage to streamline development and reduce time to market.
* Managed Serverless functions with the Serverless Framework allowing for cloud provider flexibility.
* Developed Serverless Framework AWS Lambda functions.
* Working experience on serverless deployments through AWS CLI.
* Configured alerting rules and set up PagerDuty alerting for Kafka, Zookeeper, Druid, Cassandra, Spark and different microservices in Grafana.
* Communicate workarounds to be followed to L1 and L2 till the issue/work order gets complete/resolve. Raising issues to development team and working with them closely for permanent fixes through calls.
* Developed Spark SQL to load tables into HDFS to run select queries on top and developed Spark code and Spark-SOL/Streaming for faster testing and processing of data.
* Created HBase tables to load large sets of structured data.

**Environment: Hive, SQL, Python, Java, Scala, Unix, Shell scripting, Bitbucket, spark, Kafka, HBase, HDFS, GIT, Jenkins, MYSQL database (IDE- Data Grip), Airflow, AWS Serverless deployments, Kubernetes Deployments.**

**Client: USAA – San Antonio, Texas May 2022 – April 2023**

**Role: Data Engineer \ Data Analyst**

**Responsibilities**:

* Automated workflows using Python scripts.
* Used Pandas as API to put the data as time series and tabular format for data manipulation and retrieval.
* Developed automation scripts to change the passwords automatically using python scripts.
* Worked on creating a Jenkins ansible code for running our test cases in different instances and hosts.
* Worked on python scripts to check the NPI data in the database schemas.
* Working on enhancing the scripts of automation on databases to increase performance.
* Deep understanding of moving data into **GCP** using **SQOOP** process, custom hooks for MySQL, and cloud data fusion for moving data from Teradata to GCS.
* Build data pipelines in **Airflow**/Composer for orchestrating EL-related jobs using different **airflow** operators.
* Experience in **GCP** Dataproc**, Cloud functions, Cloud SQL & Big Query**.
* Used Cloud shell SDK in GCP to configure data Proc, Storage, and Big Query services.
* Process and load bound and unbound Data from Google pub/subtopic to big query using Cloud Dataflow with Python.
* Streaming data analysis using Dataflow templates by leveraging Cloud Pub/Sub service.
* Monitoring **Big Quer**y, Dataproc, and cloud Data flow jobs via Stack driver for all the different environments.
* Got involved in migrating the on-prem Hadoop system using GCP (Google cloud platform).
* Migrated previously written corn jobs to **airflow**/compose in GCP.
* Worked on Migrating Map Reduce programs into spark transformations using spark and Scala.
* Created SSIS packages for loading different formats of files (Excel, flat file) and databases (SQL server) into the SQL Sen data warehouse using SQL Server Integration Services (SSIS).
* Created SSIS Packages to load data from Oracle to Sq| Server Database
* Experience in ETL Data Warehouse environment. Created Jobs in SIS and was responsible for ETL job scheduled to run daily. Solved the issues associated with ETL Data Warehouse Failure.
* Created SQL Jobs to schedule SSIS Packages Extracted meaningful data from dealer CSV, text, and mainframe files and generated Python panda’s reports for data analysis.
* Technical discussion with business stakeholders and SMEs to provide a feasible solution with Ao to teh high-level business requirement/ Project charter, white boarding sessions with Business stakeholders and SMEs. Supporting internal PRE-SALES for proposals and AoA with technical solution presentations
* Used WEB API to develop REST Services to manipulate data in JSON format for various read and write operations.
* Designed and Coordinated with the Data Science team to implement Advanced Analytical Models in Hadoop Cluster over large Datasets.
* Designed the jobs, workflows, data flows and used various transformations to load the data from Clarity Source into staging area, dimensions and fact tables using SAP Data Services.
* Designed the ETL code using SAP Data Services to implement Type II Slowly Changing Dimensions with Surrogate keys.
* I worked on building our own package of automation and putting it on the Artifactory for another team to utilize it.

**Environment: SAP IQ, DB2, CyberArk, Bitbucket, Dbeaver, SQL, Python, Unix, Shell scripting, GCP, Secret Manager.**

**Client: Vanguard – Charlotte, NC**  **May 2021 – April 2022**

**Role: Data Engineer**

**Responsibilities**:

* Worked as **Data Engineer** to review business requirements and compose sources to target data mapping documents.
* Involved in **Agile** development methodology and an active member in **scrum** meetings.
* Involved in **Data Profiling**and**merging data** from multiple data sources.
* Involved in **Big data** requirement analysis, developing and designing ETL and business intelligence platform solutions.
* Data from HDFS into **Spark RDDs,**for running predictive analytics on data.
* Modeled **Hive** partitions extensively for data separation and faster data processing and followed Hive best practices for tuning.
* Developed Spark scripts by writing custom **RDDs** in **Scala** for data transformations and performing actions on **RDDs.**
* Developed **Spark RDD** transformations, actions, Data frames, case classes, and Datasets for the required input data and performed the data transformations using Spark-Core.
* Created Data pipelines for **Kafka cluster** and processed the data by using spark streaming and worked on streaming data to consume data from KAFKA topics and load the data to the landing area for reporting in near real-time.
* Worked with cloud-based technology like **Redshift, S3, AWS, EC2** Machine, etc., and extracted the data from the Oracle financials and the Redshift database Created Glue jobs in AWS and loaded incremental data to the S3 staging area and persistence area.
* Developed Pyspark code for AWS Glue jobs and for EMR
* Created AWS Lambda function for extracting the data from SAP database and post the data to AWS S3 bucket on scheduled basis using AWS cloud watch event.
* Worked on building a centralized Data Lake on AS Cloud utilizing primary services like s3, EMR, Redshift and Athena.
* Migrate on the in-house database to **AWS Cloud** and design, built, and deployed a multitude of applications utilizing the AWS stack (Including **S3, EC2, and RDS**) by focusing on high availability and auto-scaling.
* Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using **Python and PySpark.**
* Parsed Semi-Structured **JSON** data and converted to Parquet using Data Frames in PySpark and Created Hive DDL on Parquet and **Avro data** files residing in both HDFS and S3 buckets.
* Created **AWS Glue** job for archiving data from Redshift tables to S3 (online to cold storage) as per data retention requirements.
* Updated **Python scripts** to match training data with our database stored in **AWS Cloud** Search, so that we would be able to assign each document a response label for further classification.
* Created monitors, alarms, notifications, and logs for **Lambda functions, Glue Jobs, and EC2** hosts using CloudWatch and used **AWS Glue** for the data transformation, validation, and data cleansing.
* Deployed applications using Jenkins’s framework integrating **Git- version control** with it.
* Worked on Commercial lines Property and Casualty P&C Insurance, including policy, claim processing, and reinsurance.
* Worked on Renaissance P&C Insurance Billing System implementation projects.

**Environment:  Bitbucket, SQL, Python, PySpark, Unix, Shell scripting, AWS Redshift, S3, EC2, Glue, Kafka, Hive**

**Client: Thermo Fisher**  **April 2019 – May 2021**

**Role: Data Engineer/ Data Analyst**

**Responsibilities**:

* Automated workflows using Alteryx to send emails to the stakeholders.
* Created rich dashboards using Tableau Dashboard and prepared user stories to create compelling dashboards to deliver actionable insights.
* Wrote queries to gather data from different tables by using joins, sub-queries, correlated sub-queries, and derived tables on the SQL Server platform.
* Wrote SQL queries that were highly tuned using concepts like Explain, Stats, CAST, and volatile tables and used SAS in importing and exporting data for SQL server tables.
* Developed dashboards in Tableau Desktop and published them in Tableau Server, ETL processes using Informatica.
* PowerCenter, SQL, and shell scripting to extract data from source systems, transform the data into the necessary format, and load it into the data warehouse.
* Created various Documents such as Source-To-Target Data mapping Documents, Backroom Services Documents, Unit Test Cases, and Data Migration Documents.
* Responsible for interaction with business stakeholders, gathering requirements, and managing the delivery.
* Connected Tableau server to publish the dashboard to a central location for portal integration.
* Created visualization for logistics calculation and department spending analysis.
* Generate KPIs for customer satisfaction survey results.
* Support and develop ETL processes to support data warehouse and reporting requirements.
* Integrated data from different sources such as Azure Blob Storage, Azure SQL Database, Azure data lake, SQL server DB, flat files, MS-Excel, etc. in CDM
* Designed and developed ETL packages in SIS and pipelines in Azure Data Factory to transform source data as per transformation logic and ingest that into the CDM platform based on defined architectures of the CDM model.
* Worked extensively on Azure data factory including data transformations and integration runtimes. Azure Key Vaults. Triggers and migrating data factory pipelines to higher environments using ARM Templates
* Engaged with subject matter experts to define and maintain reference data to be consumed during the data ingestion process.
* Automated workflows using Alteryx to send emails to the stakeholders.
* Automated transactional **workflows** using **Talend.**
* Managed **SQL** databases and box folders for the stakeholders to view the metrics.
* Development and enhancements in Tibco Spotfire dashboards as per the client’s requirement.
* Maintain the dashboards on a quarterly and yearly basis.
* Worked on **AWS Redshift** on the ongoing migration project and helped in writing POC on it
* Analyzing the workflows in **Talend** and **Alteryx** and resolving them if any bugs are encountered.

**Environment:** Alteryx, Tableau, SQL, ETL processes, Informatica, PowerCenter, KPIs

**Client: Trianz, Dallas, Offshore, Hyderabad**  **November 2015 – July 2018**

**Role: Data Analyst**

**Responsibilities**:

* Responsible for analyzing functional specifications and preparing technical design specifications.
* Involved in all Software Development Life Cycle (SDLC) phases of the project from domain knowledge sharing, requirement analysis, system design, implementation, and deployment.
* Converted Talend Job lets to support the **snowflake functionality**.
* Used COPY to bulk load the data.
* Created data sharing between two **snowflake** accounts.
* Created internal and external stages and transformed data during load.
* Redesigned the Views in **snowflake** to increase the performance.
* Unit tested the data between Redshift and **Snowflake**.
* Developed data warehouse model in **snowflake** for over 100 datasets using Were scape.
* Creating Reports in Looker based on **Snowflake Connections**.
* Implement One-time Data Migration of Multistate level data from SQL server to **Snowflake** by using Python and Snow SQL.
* Day to day responsibility includes developing ET Pipelines in and out of the data warehouse and developing major regulatory and financial reports using advanced SQL queries in **snowflake**.
* Created SSIS packages for loading different formats of files (Excel, flat file) and databases (SQL server) into the SQL Sen data warehouse using SQL Server Integration Services (SSIS).
* Created SSIS Packages to load data from Oracle to Sq| Server Database
* Experience in ETL Data Warehouse environment. Created Jobs in SIS and was responsible for ETL job scheduled to run daily. Solved the issues associated with ETL Data Warehouse Failure.
* Created SQL Jobs to schedule SSIS Packages Extracted meaningful data from dealer CSV, text, and mainframe files and generated Python panda’s reports for data analysis.
* Accomplished in reproducing the issues reported by azure end customers and helped in debugging the issue.
* Implemented rigorous model testing, validation, and optimization strategies within automated ETL and training pipelines.
* Automated the existing scripts for performance calculations using scheduling tools like **Airflow.**
* Used Git source control to manage simultaneous development.
* Create and maintain detailed, up-to-date technical documentation.
* Experience with continuous integration and automation for CI/CD using Jenkins.

**Environment**: Python, PyQuery, Snowflake, JSON, Databricks, SQL, UNIX, Windows, NoSQL, **Airflow** and python libraries such as NumPy, datetime-time delta, Auto Broadcast,

**Education**:

* Masters in Statistical Analysis Computing and Modeling, Texas A&M University – 2020
* Bachelors in Electronics and Communication Engineer, Amrita University - 2017