**Swapn Shah**

**E-Mail: Sudheer.r@tekprosol.com**

**Contact:609 981 8255**

**Professional Summary:**

I am a data engineer with 10+ years of experience in designing, developing and maintaining data pipelines and architectures using various technologies such as python, SQL, snowflake, AWS. I have successfully delivered data solutions for various domains such as e-commerce, healthcare, finance, and education. I have strong skills in data modeling, data quality, data governance and data visualization. I am proficient in using cloud platforms such as AWS to deploy and scale data applications. I am also well-versed in using python script to automate data processes and tasks. I am passionate about solving complex data problems and delivering value to the business.

* **Data Transformation and Analysis:**
  + Created Data Frames and performed analysis using Spark SQL.
  + Hands-on expertise in writing different RDD (Resilient Distributed Datasets) transformations and actions using Scala, Python, and Java.
  + Performed transformations on imported data and exported back to RDBMS.
  + Developed complex mappings and loaded data from various sources into the Data Warehouse, using different transformations/stages like Joiner, Transformer, Aggregator, Update Strategy, Rank, Lookup, Filter, Sorter, Source Qualifier, Stored Procedure transformation, etc.
  + Implemented POC to migrate MapReduce jobs into Spark transformations using Python.

**CI/CD and Cloud Integration:**

* **Continuous Integration and Deployment:**
  + Worked in agile projects, delivering end-to-end continuous integration/continuous delivery pipelines by integrating tools like Jenkins and AWS for VM provisioning.
  + Implemented continuous integration and deployment (CI/CD) through Jenkins for Hadoop jobs.
* **Cloud Services and Integration:**
  + Good knowledge of Cloudera distributions and Amazon services such as Amazon S3, AWS Redshift, Lambda, Amazon EC2, and Amazon EMR.
  + Worked on Amazon Web Service (AWS) to integrate EMR with Spark 2 and S3 storage and Snowflake.
  + Capable of using AWS utilities such as EMR, S3, and CloudWatch to run and monitor Hadoop and Spark jobs on Amazon Web Services (AWS).
  + Experienced with Dimensional modelling, data migration, data cleansing, data profiling, and ETL processes for data warehouses.

**Big Data and Ecosystem Expertise:**

* **Hadoop Ecosystem:**
  + Excellent understanding of Hadoop Architecture and good exposure to Hadoop components like Hadoop MapReduce, HDFS, HBase, Hive, Sqoop, Cassandra, Kafka, and Amazon Web services (AWS) API testing, documentation, and monitoring with Postman, which easily integrates tests into your build automation.
  + Capable of using AWS utilities such as EMR, S3, and CloudWatch to run and monitor Hadoop and Spark jobs on Amazon Web Services (AWS).
  + Knowledge in installing, configuring, and using Hadoop ecosystem components like Hadoop MapReduce, HDFS, HBase, Oozie, Hive, Sqoop, Zookeeper, and Flume.
* **Data Ingestion and Processing:**
  + Used Apache Flume to ingest data from different sources to sinks like Avro and HDFS.
  + Excellent knowledge of Kafka Architecture.
  + Integrated Flume with Kafka, using Flume as both a producer and consumer (concept of FLAFKA).

**Cloud Computing and Infrastructure:**

* **Cloud Development and Infrastructure:**
  + Software development involving cloud computing platforms like Amazon Web Services (AWS).
  + Capable of using AWS utilities such as EMR, S3, and CloudWatch to run and monitor Hadoop and Spark jobs on Amazon Web Services (AWS).
  + Strong understanding of the entire AWS Product and Service suite, primarily EC2, S3, VPC, Lambda, Redshift, Spectrum, Athena, EMR (Hadoop), and other monitoring service products, their applicable use cases, best practices, and implementation and support considerations.
* **Infrastructure as Code:**
  + Experience in writing Infrastructure as Code (IaC) in Terraform, AWS CloudFormation.
  + Created reusable Terraform modules in AWS cloud environments.
  + Worked on AWS EC2, EMR, and S3 to create clusters and manage data using S3.

**Scripting and Automation:**

* **Scripting and Source Control:**
  + Strong experience in Unix and Shell Scripting. Experience on Source control repositories like SVN, CVS, and GITHUB.
  + Extensive experience in designing and implementing continuous integration, continuous delivery, continuous deployment through Jenkins.
* **Airflow Workflow Management:**
  + Installed and configured Apache Airflow for workflow management and created workflows in Python.
  + Skilled in designing and orchestrating complex data pipelines and workflows for improved data processing efficiency.
  + Experienced in defining task dependencies, handling retries, and managing task execution order in Airflow DAGs.
  + Capable of creating custom operators for tailored workflows, enhancing functionality.
  + Proficient in scheduling intervals, monitoring task execution, and troubleshooting within Airflow.
  + Expertise in developing and maintaining Directed Acyclic Graphs (DAGs) for workflow management.
  + Familiar with integrating Airflow with cloud services for seamless data pipeline orchestration.
  + Skilled in using the Airflow web UI and CLI for workflow management.
  + Proficient in dynamically generating scalable workflows with Airflow's templating and macro features.

**Technical Skills:**

|  |  |
| --- | --- |
| Databases | Oracle 11g, 10g, 9i, SQL Server 2012/2008R2, Hadoop, HIVE, MongoDB, and MySQL, Netezza, SAS, Hadoop, U-SQL, Snowflake |
| Languages | SQL, PL/SQL, TSQL, Python (matplotlib, Seaborn, NumPy, SciPy, SciKit-Learn), Power BI, R-Studio, Alteryx, SSRS (SQL Server Reporting Service), HTML, XML, Nodejs, UNIX, JAVASCRIPT |
| Analytics & Reporting | Tableau, SSRS, Cognos, Power BI, Crystal Reports, Data bricks |
| ETL & Cloud Platform | Informatica 10, SQL Server DTS, Visual Studio 2012/2010, Azure Data Lake & Factory, Pandas, AWS EC2, S3, Lambda, NumPy, Salesforce CRM, Apex Data Loader, SAP MDM |
| BI Tools | SSIS, BI Development Studio, Power BI, Visual Studio 2012, Performance Monitor, Power Pivot, Spark, Scala, Kafka, DevOps |
| Machine Learning | Linear Regression, Logistic Regression, LDA, PCA (Principal Component Analysis), K-Means, Clustering, K-Nearest Neighbors (KNN), Decision Tree, Ada Boosting, Gradient Boosting Trees, Neural Networks. |
| Design Tools | ERWIN, ER Studio, MS VISIO |
| Version Control Systems | SVN, Git, GitHub |

**Professional Experience**

**Client: DTCC Oct 2022 – Present**

**Role: AWS Data Engineer**

**Location: Florida**

**Responsibilities:**

* Orchestrated and managed multi-node clusters on Amazon Web Services (AWS) EC2, ensuring efficient installation and configuration for cloud-based operations.
* Proficiently utilized AWS Management Tools, including CloudWatch and CloudTrail, to monitor and enhance system performance and security.
* Implemented robust data storage solutions by incorporating versioning within AWS S3 buckets for highly sensitive information, enhancing data integrity.
* Leveraged AWS Lambda for seamless integration with DynamoDB, storing item values and facilitating DynamoDB stream backups.
* Demonstrated automation prowess by scripting routine AWS tasks, such as snapshot creation, using Python for increased efficiency.
* Spearheaded the design and deployment of data warehouses on various high-performance platforms, including AWS Redshift.
* Established Apache Airflow for AWS S3 buckets and constructed custom DAGs for streamlined workflow execution.
* Engineered automation scripts in Pyspark and Scala to streamline data ingestion from diverse sources, including APIs, AWS S3, Teradata, and Redshift.
* Effectively harnessed AWS Glue to discover, prepare, and integrate data from multiple sources, supporting analytics, machine learning, and application development.
* Mastered AWS Glue Studio for the visual creation, execution, and monitoring of ETL pipelines, streamlining data lake loading.
* Utilized AWS Glue interactive sessions for data exploration and processing, employing PySpark and Scala notebooks.
* Employed AWS Glue streaming ETL jobs for real-time data processing from sources such as Amazon Kinesis Data Streams and Apache Kafka.
* Crafted multiple custom scripts to automate ETL/ELT processes using PySpark and ensured seamless data transfer across platforms.
* Engineered PySpark scripts, incorporating SQL and RDD, for in-depth data analysis and secure storage in AWS S3.
* Designed and executed Spark SQL code to implement business logic using Python as the programming language.
* Developed, compiled, and executed Apache Spark programs in Scala for ETL operations, effectively managing ingested data.
* Employed Spark Streaming to partition streaming data into batches, optimizing input for batch processing within the Spark engine.
* Maintained Kubernetes clusters with precision, handling patches, and overseeing upgrades in a production environment.
* Actively crafted user-defined functions (UDFs) in Map-Reduce and Python for Pig and Hive to enhance data processing and analysis.
* Ensured data integrity by conducting comprehensive integrity checks using Hive queries, Hadoop, and Spark.
* Leveraged Scala to implement machine learning algorithms within Spark using Python, enhancing data-driven insights.
* Successfully designed and implemented Scala programs using Spark Data Frames and RDDs for transformations and actions on input data.
* Boosted Hive query performance through the implementation of partitioning, clustering, and optimized file formats like ORC.

**Environment**: AWS, JMeter, Kafka, Ansible, Jenkins, Docker, Maven, Linux, Red Hat, GIT, Cloud Watch, Python, Shell Scripting, Golang, Web Sphere, Splunk, Tomcat, Soap UI, Kubernetes, Terraform, PowerShell.

**Client: Spectra Laboratories Jun 2021 – Sep 2023**

**Role: Aws Data Engineer**

**Location: Owings Mills, MD**

**Responsibilities:**

* Developed Python scripts to automate ETL processes and data analysis, integrating SQL for efficient data manipulation.
* Managed MySQL and PostgreSQL databases, ensuring data integrity, security, and optimal performance.
* Leveraged Snowflake for data warehousing, focusing on efficient data loading and querying.
* Built and fine-tuned machine learning models with TensorFlow, Keras, PyTorch, and Scikit-Learn for predictive analytics and pattern recognition.
* Conducted data analysis using Pandas and NumPy, generating visualizations with Seaborn and Matplotlib for informed decision-making.
* Excelled in data analysis, visualization, and reporting using Excel, automating routine tasks with macros.
* Orchestrated ETL processes using AWS Glue, Lambda, and S3, optimizing data integration and storage.
* Designed and developed web applications with Django, implementing RESTful APIs and user-friendly interfaces.
* Implemented PySpark for big data processing, optimizing Spark jobs for enhanced performance and scalability.
* Collaborated with cross-functional teams to gather data and business requirements, ensuring alignment with project goals.
* Designed and implemented data solutions for migrating source data from Data Warehouse to Atlas Data Lake, enhancing data accessibility.
* Conducted data validation using SQL and HIVE scripts to ensure data quality in various applications.
* Employed Cognos report validation to maintain accurate reporting and data consistency.
* Deployed applications on AWS, configuring services like EC2, S3, and RDS for high-availability and fault tolerance.
* Managed AWS S3 buckets for scalable and cost-effective data storage, ensuring data security.
* Developed Lambda functions for data aggregation and storage in DynamoDB and S3.
* Ensured data governance and security on AWS cloud platforms, maintaining compliance with industry standards.
* Contributed to the successful migration of databases from Solaris to OEL, coordinating with multiple teams for seamless transitions.
* Mitigated business risk through documentation, integration, and interdependency tracking, ensuring project success.
* Optimized workload management for load balancing and failover, improving system performance and scalability.
* Provided architectural leadership to support data-driven decision-making, maximize technical deliverables' quality, and enhance data engineering processes.

**Environment**: Aws Ec2, Aws S3, RDS, Snowflake, Dynamo DB, SQL, Aws Glue, Python, Pyscript, Tensor flow,

**Client**: **IM Health Science Feb 2018 – Nov 2020**

**Azure Data Engineer**

**Location: India**

**Responsibilities**

* Spearheaded the process of gathering data and business requirements from both end users and management, and subsequently designed and implemented data solutions to facilitate the migration of existing source data from Data Warehouse to Atlas Data Lake (Big Data).
* Conducted comprehensive analysis of large data volumes, devising a wide range of HIVE and SQL scripts to validate Dataflow within various applications. Furthermore, carried out Cognos report validation and leveraged MHUB for thorough Data Profiling and Data Lineage verification.
* Demonstrated proficiency in designing and deploying multiple applications across the AWS stack, encompassing EC2, S3, AMI, Route53, RDS, SNS, SQS, and IAM. Focused on ensuring high availability, fault tolerance, and Auto-Scaling through AWS CloudFormation.
* Took charge of configuring an AWS Virtual Private Cloud (VPC), NACL, and Database Subnet Group to effectively isolate resources within Confidential RDS and Aurora DB clusters.
* Developed a Lambda function responsible for aggregating data from incoming events, with the subsequent storage of the resulting data in Confidential DynamoDB and S3.
* Oversaw the maintenance and design of data governance and security for data platforms within the AWS Cloud environment.
* Successfully led Project Management for a Datacentre Migration Program involving over 1000 databases, transitioning from Solaris to OEL (Linux). Collaborated with multiple teams to ensure successful Go lives and contributed to the decommissioning of old databases.
* Mitigated business risk through technology documentation, integration, and interdependency tracking, and actively participated in the development and review of solution, integration, workflow, and web service standards.
* Efficiently managed Workload Management to optimize load balancing and failover, enhancing system performance, reliability, and scalability.
* Provided architectural leadership to support informed decision-making, drive development process efficiencies, and maximize the quality of technical deliverables.
* Executed technical feasibility assessments, solution estimations, and proposal development for migrating identified workloads to Windows on AWS.
* Utilized AWS Data Migration Service to create source and destination endpoints for Aurora RDS Migration, including the transfer of data and schema conversion using the AWS schema conversion tool.
* Configured, monitored, and automated Confidential Web Services, actively participating in deploying cloud platform content on Confidential Web Services through ECS, Elastic Beanstalk, and EBS.
* Played a pivotal role in the installation, configuration, and management of RDBMS and NoSQL tools such as DynamoDB.
* Led the implementation of Continuous Integration (CI) and Continuous Delivery (CD) processes using Jenkins in combination with Shell scripts to automate routine tasks.
* Possessed a strong working knowledge of Service Mesh platforms, with hands-on experience with Istio.
* Established and managed a Docker Registry for local Docker image upload and download, as well as for accessing images from Docker Hub.
* Orchestrated Docker container management using Docker Swarm and created a Continuous Delivery pipeline using Docker, Jenkins, and GitHub.
* Designed a private cloud using Kubernetes to support development, testing, and production environments, along with managing Kubernetes charts using Helm for reproducible builds of Kubernetes applications.
* Skilfully configured Docker containers created Docker files for various applications, and managed Docker volumes.
* Crafted Docker images from scratch, customized and modified base images from existing environment configurations, and maintained the image repository for development teams.
* Administered and engineered Jenkins to manage a weekly Build, Test, and Deploy chain, incorporating a GIT workflow with Dev/Test/Prod Branching Model for weekly releases.
* Played a key role in Configuration Automation and Centralized Management with Ansible, writing Ansible Playbooks with Python SSH as the wrapper to manage AWS Node configurations and provision dev servers.
* Undertook the installation, configuration, monitoring, and troubleshooting of Oracle Cluster ware and RAC for databases on Linux and Solaris environments, encompassing Oracle 12c, 11gR2, and 10g.
* Contributed to the migration of a quality monitoring tool from AWS EC2 to AWS Lambda, including the creation of logical datasets to facilitate quality data monitoring on Snowflake warehouses.
* Actively participated in the department's transformation into Squads using Agile methodology for projects, utilizing tools like Jira and cloud platforms such as AWS.

**Environment**: SQL Server 2014/12, Python, Databricks, Spark, Snowflake, Triggers, controllers, Eclipse, Pandas, AWS EC2, S3, Lambda.

**Client: Digital Minds Technologies Sep 2015 to Jan 2018**

**Role: Data Engineer**

**Location: India**

**Responsibilities:**

* Created and presented models for potential holdings to fund managers. Achieved 20% better returns v/s historical performance and Predicted stock price 25% better than traditional figures.
* Facilitated User Acceptance Testing and identified KPIs in the Reference data to predict failure of client's payments.
* Fashioned with machine learning algorithms like linear regression, KNN and decision tree for trading problems, estimated machine learning algorithm's performance for time series data (stock price data).
* Calculated and presented data visualization report for daily returns, cumulative returns, simple moving averages, Sharpe ratio, portfolio value for stock performance optimization.
* Analyzed large data sets using pandas and used regression models using SciPy to predict future data and visualized them and used SQL to manipulate data.
* Proficient in importing/exporting large amounts of data from files to Teradata and Created Teradata specific physical data models which included primary indexes, secondary indexes and joined indexes.
* Implemented Dimensional Data Modeling to deliver Multi-Dimensional STAR schemas and Developed Snowflake Schemas by normalizing the dimension tables as appropriate.
* Worked with data compliance teams, Data governance team to maintain data models, Metadata, Data Dictionaries; define source fields and its definitions.
* Used HIVE to do transformations, event joins and some pre-aggregations before storing the data onto HDFS.
* Used Redshift, S3 within AWS along with Informatica Cloud Services to load data into s3 bucket.
* Worked with HDFS file formats like Avro, Sequence File and various compression formats like Snappy.
* Collaborated with Data engineers and operation team to implement ETL process, Snowflake models, wrote and optimized SQL queries to perform data extraction to fit the analytical requirements.
* Written SQL Scripts and PL/SQL Scripts to extract data from Database and for Testing Purposes.
* Created data pipeline package to move data from amazon S3 bucket to MYSQL database and executed MySQL stored procedure using events to load data into tables.
* Designed and developed Tableau graphical and visualization solutions based on business requirement and expert with the backend data retrieval team, data mart team to guide the proper structuring of data for Tableau reporting.
* Worked on metadata in Tableau desktop to alter data types, assign roles, rename fields, joins, refresh data and change extract data options to prepare data sources to be utilized by users.

**Environment**: Hadoop, ETL operations, Data Warehousing, Data Modelling, Cassandra, AWS Cloud computing architecture, EC2, S3, Advanced SQL methods, NiFi, Python, Linux, Apache Spark, Scala, Spark-SQL, HBase

**Role: Data Engineer Aug 2014 - Aug 2015**

**Quess Corp Limited, India**

**Responsibilities:**

* Experience in using T-SQL for creating stored procedures, indexes, cursors and functions.
* Generated server-side T-SQL scripts for data manipulation and validation and created various snapshots and materialized views for remote instances.
* Creating tables, indexes and designing constraints and wrote T-SQL statements for retrieval of data and involved in performance tuning of T-SQL Queries and Stored Procedures.
* Executed stored proc using BATCH files. Developed stored procedures and functions for designing various reports.
* Extensive hands-on experience with SSIS to build huge packages which involves sourcing the data from a Tab Delimited Text Files, DB2, XML, Excel files and Flat Files and setting up the Variables and Parameters and mapping the obtained result sets to a Map Network Drive which uses the Merge Scripts.
* Worked on integrations involving Microsoft Visual C# to use in the script components of SSIS.
* Enhancing and deploying the SSIS Packages from development server to production server.
* Wrote the ETL scripts to load the data into database from various source files.
* Create, manage SQL Server AZURE Databases.
* Setting up Connection Strings and connecting SQL Server AZURE Databases from locally Installed SQL Server Management Studio (SSMS) for Developers.
* Handling and troubleshooting the scripts in failure to load the data into database.
* Developed DTS packages to copy tables, schemas and views and to extract data from Excel and Oracle using SSIS.
* Created packages in SSIS with custom error handling and worked with different methods of logging in SSIS.
* Expert in SSIS Deploying and Scheduling
* Defined report layouts for formatting the report design as per the need.
* Used TFS for code check-in and deployment.
* Developed Sub Reports, Matrix Reports, Charts, and Drill down reports, using SQL Server Reporting Services (SSRS).
* Generated Reports using Global Variables, Expressions and Functions for the reports.
* Worked on querying data and creating on-demand reports using Report Builder in SSRS reports and sending the reports via email.

**Environment**: SQL Server 2008/2012/2014/2016 Enterprise Edition, Power BI, SSIS, SSRS, power shell, VSO, Power BI, MS PowerPoint, MS Project, C#, VB.Net, XML, Visual Studio 2017/2015/2012/2008, VSTS.

**Role: Data Engineer**

**Anuvega Pvt Ltd , India Jul 2013 to March 2014**

**Responsibilities:**

* Worked on Data Profiling, Data Cleansing and Data Mining.
* Modified the logical and physical data models for the new feeds.
* Developed BTEQ scripts to load the data from the staging tables to the base tables.
* Performance tuning of long running scripts. Stats/ index recommendations.
* Performed developer DBA tasks like creating users/roles/profiles and space allocation tasks.
* Worked on Teradata Manager to monitor and manage resource utilization.
* Capacity planning for new applications.
* Production system monitoring and providing support for any batch failures.
* Environment: Teradata V12 on UNIX MP-RAS, Informatica 7.1, BTEQ, FASTLOAD, MULTILOAD, FASTEXPORT, UNIX Shell Scripting, Teradata Manager/PMON.

**Education:**

* Master of Science – Computer Science MAY 2021 University of South Florida, Tampa, FL
* Bachelor of Engineering – Computer Science 2011- Rajalakshmi Engineering College, Chennai, India