**Name: Hakeel**

**Email: hakeel302@gmail.com**

**CONTACT NO: +1 9809992615**

**SUMMARY:**

* Having 9 plus years of hands-on experience in implementing end-to-end Big Data and Cloud Engineering solutions, encompassing various technologies such as **Snowflake, Data Warehouse, Data Mart, Data Visualization, Reporting, Data Quality, Data virtualization, and Data Science Solutions.**
* Proven expertise in **Data transformation**, **Data mapping** from source to target **database schema,** and executing Data Cleansing procedures to ensure data accuracy and integrity.
* Proficient in programming languages like **R** and **Python**, with a strong command of Big Data technologies like **Hadoop and Hive** for scalable data processing and analysis.
* An excellent understanding of relational database design, **data warehouse/OLAP ideas**, and techniques is required, with a particular emphasis on creating **star and snowflake** schema for **ODS architecture** and **data warehouses**.
* Competent in **relational, dimensional, and multidimensional modeling**; knowledgeable in **OLTP/OLAP System Study**, **Analysis**, and **E-R modeling**; capable of creating database schemas such as **Star schema** and **Snowflake schema.**
* Broad involvement with planning, assembling, and carrying out total Hadoop biological systems including **MapReduce**, **HDFS**, **Hive**, **Impala**, **Sqoop**, **Oozie**, **HBase**, **MongoDB**, **Flash**, and **Kafka**.
* Skilled in utilizing **Erwin**, **ER-Studio**, and **MS Visio** for data design and data architecture; has practical expertise in coding **SQL** for creating **Procedures**, **Triggers**, and **Packages**.
* Experienced in making separate virtual information stockrooms with various size classes in **AWS Snowflake**, and capable of mass stacking and dumping information into Snowflake tables utilizing Duplicate order.
* Talented in information changes using **Snow SQL in Snowflake**, composing flash streamingand flash cluster occupations, and involving **flash Mallik** for examination
* Active involvement with bringing in and sending out information utilizing **Sqoop** from **HDFS** to Social Data set Frameworks (RDBMS) like **Prophet**, **DB2**, **SQL Server**, as well as the other way around, close by broad involvement with Standardization, De-Standardization, and Normalization procedures.
* Solid understanding of **AWS**, **Redshift**, **S3**, **EC2**, and **Apache Spark**, **Scala process**, and concepts, with hands-on experience inmachine learning, big data, data visualization, **R** and **Python** development**, Linux, SQL, GIT/GitHu**b.
* Proficient in data visualization using tools like **plot**, **Matplotlib**, **Seaborn**, and **Tableau**, andutilizing Tableau software to publish and present dashboards, and storylines on web and desktop platforms.
* Experienced in python information control for stacking and extraction, capable in python libraries like **NumPy**, **SciPy**, and **Pandas** for information examination and mathematical calculations.
* Skilled in **RStudio** for data pre-processing and building machine learning algorithms on different datasets, and experienced in working on NoSQL databases like **MongoDB** and **HBase**.
* Extensive experience working with **Python**, including **Scikit-learn**, **SciPy, Pandas, and NumPy,** handling and editing data**,** andcreating machine learning models.
* Proficient in Machine Learning techniques and Statistical Modeling, adept at implementing machine learning algorithms on large datasets to understand hidden patterns and capture insights**.**
* Experienced in building and improving large information pipelines, designs, and datasets utilizing **TensorFlow**, Information Programming interface, **Flash**, and **Hive**.

**TECHNICAL SKILLS**:

|  |  |
| --- | --- |
| **Big Data Technologies** | Hadoop, MapReduce, HDFS, Sqoop, PIG, Hive, HBase, Oozie, Flume, Knife, Kafka, Zookeeper, Yarn, Apache Spark, Mahout, Sparkled |
| **Databases** | Oracle, MySQL, SQL Server, MongoDB, Cassandra, DynamoDB, PostgreSQL, Teradata, Cosmos |
| **Programming** | Python, Spark, Scala, Java, C, C++, Shell script, Perl script, SQL |
| **Cloud Technologies** | AWS |
| **Frameworks** | Django REST framework, MVC, Hortonworks |
| **Tools** | PyCharm, Eclipse, Visual Studio, SQL\*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Eclipse, Postman |
| **Versioning tools** | SVN, Git, GitHub |
| **Operating Systems** | Windows 7/8/XP/2008/2012, Ubuntu Linux, MacOS |
| **Network Security** | Kerberos |
| **Database Modelling** | Dimension Modeling, ER Modeling, Star Schema Modeling, Snowflake Modeling |
| **Monitoring Tool** | Apache Airflow |
| **Visualization/ Reporting** | Tableau, ggplot2, Matplotlib, SSRS and Power BI |
| **Machine Learning Techniques** | Linear & Logistic Regression, Classification and Regression Trees, Random Forest, Associative rules, NLP and Clustering |

**PROFESSIONAL EXPERIENCE:  
S&P Global, New York, NY June 2021 – Present**

**Role: Sr. AWS Data Engineer  
  
Project Summary:** The task included utilizing AWS distributed computing design and different information innovations, for example, Apache Flash, Hadoop, and Cassandra to create effective ETL pipelines, information warehousing arrangements, and AI calculations. Obligations included planning, upgrading, and keeping up with information processes while guaranteeing versatility, unwavering quality, and security**.**

**Responsibilities:**

* Skilled in working with different data formats such as **JSON** and implementing machine learning algorithms in **Python** for advanced data analysis.
* Developed a task scheduling application to run in an **EC2** environment across multiple servers, optimizing resource utilization and improving overall efficiency.
* Strong understanding of various **Data warehousing methodologies** and **Data modeling concepts**, ensuring effective data organization and management.
* Created **Hive partitioned tables** using **Parquet Avro format**, enhancing query performance and optimizing storage space utilization.
* Responsible for **Database Design** and **Creation of User Database**, ensuring data accessibility and security.
* Successfully moved **ETL pipelines** from **SQL server** to **Hadoop Environment**, improving scalability and performance.
* Developed Spark applications using **Spark and Spark-SQL** for **data extraction**, **transformation,** and **aggregation**, facilitating in-depth data analysis and insights generation.
* Implemented **CI/CD pipeline using Jenkins**, **Airflow**, **Docker**, and **Kubernetes**, streamlining development and deployment processes.
* Utilized **SSIS**, **NIFI**, **Python scripts**, and Spark Applications for ETL Operations, creating **data flow pipelines** and **transforming data** for analytics purposes.
* Supported current and new services leveraging **AWS cloud computing architecture**, including EC2, S3, and other managed service offerings.
* Experience in writing SAM template to deploy serverless applications on AWS cloud.
* Proficient in advanced SQL methods for coding, testing, debugging, and documenting complex database queries.
* Develop Informatica SCD Type-2 Mappings based on the requirements.
* Data Extraction, aggregations and consolidation of Adobe data within AWS Glue using Spark.
* Designed relational database models for small and large applications, ensuring efficient data organization and retrieval.
* Create external tables with partitions using Hive, AWS Athena and Redshift.
* Developed Scala workflows for data extraction and transformation from cloud-based systems, enhancing data processing capabilities.
* Skilled in developing reliable, maintainable, and efficient code in SQL, Linux shell, and Python.
* Develop Informatica mappings to be implemented based on client requirements and for the analytics team.
* Utilized Python Libraries like Boto3, NumPy for AWS.
* Implemented Apache Spark code to read and filter real-time records from multiple tables, storing computation results in Cassandra and HBase tables for further analysis.
* Create end-to-end solution for ETL transformation jobs that involve writing Informatica workflows and mappings.
* Proficient in troubleshooting and resolving complex production issues, **providing data analysis** and **validation support.**

**Environment:** Data Warehousing, Data Modeling, Cassandra, AWS Cloud Computing Architecture, EC2, S3, Advanced SQL Methods, Teradata, SQL Server, Hadoop, ETL Operations, Knife, Python, Linux, Apache Spark, Scala, Spark-SQL, HBase

**Rocket Mortgage, Charlotte, NC Jan 2020 - May 2021**

**Role: Data Engineer**

**Project Summary:** The role involved gathering, analyzing, and translating business requirements into technical specifications, acquiring and structuring data from multiple sources, and developing predictive algorithms to extract valuable insights. Responsibilities also included designing and implementing secure data pipelines, conducting exploratory data analysis, and building machine learning models for forecasting and optimization. Additionally, the position entailed collaborating with various departments, utilizing ETL processes and tools like Talend and Snowflake, and creating interactive dashboards for comprehensive reporting.

**Responsibilities:**

* Gathered, analyzed, and translated business requirements into technical specifications, collaborating with various departments to understand client needs and available **data** sources.
* Acquired, cleaned, and structured data from multiple sources, maintaining databases and data systems for efficient storage and retrieval.
* Identified, analyzed, and **interpreted trends** and patterns within **complex datasets** to extract valuable insights.
* Developed, prototyped**,** and **tested predictive algorithms** to **forecast outcomes** and **improve decision-making processes.**
* Filtered and cleaned data, reviewing reports and performance indicators to ensure data accuracy and reliability.
* Designed and implemented data collection systems and strategies to optimize statistical **efficiency and enhance data quality**.
* Created and statistically analyzed large datasets comprising internal and external data sources to derive actionable insights.
* **Collaborated closely** with **marketing teams** to leverage extensive data from various campaigns and customer interactions, such as web portal usage and email responses, to drive business strategies.
* Performed incremental and full data loads, **transferring data from OLTP systems** to Snowflake Data Warehouses using ETL processes.
* **Designed secure data pipelines** to ingest data into Snowflake warehouses from both on-premises and cloud-based sources.
* **Utilized Kafka** as a message broker for **collecting** and **analyzing** large volumes of data within distributed systems.
* **Developed ETL processes using Talend Tool** for seamless data loading and transformation from source systems to Snowflake.
* **Implemented Snowpipes** for continuous data injection using event handlers from AWS S3 buckets.
* Designed and implemented end-to-end ETL processes from source systems to staging areas, Data Marts, and final data loads.
* Loaded data into **Snowflake tables** from internal staging areas using **SnowSQL commands**.
* Prepared data warehouses using **Star/Snowflake schema concepts** in **Snowflake,** leveraging SnowSQL for database operations.
* Conducted data cleaning, feature scaling, and engineering using **NumPy** and **Pandas** in **Python** for improved data quality.
* Conducted Exploratory Data Analysis (EDA) using Python libraries such as **Matplotlib** and **Seaborn** to uncover underlying patterns and correlations between features.
* Worked with **NoSQL databases** like **HBase** to create tables for loading large sets of semi-structured data from source systems.
* Tuned performance of Informatica session for large data files by increasing block size, data cache size, sequence buffer length and target based commit interval.
* Utilized information value, principal components analysis, and Chi-square feature selection techniques for data analysis.
* Implemented machine learning algorithms using Python and R scripts to predict and forecast data outcomes.
* Developed R packages in RStudio with a shiny interface for enhanced user interaction and visualization.
* Used Informatica Power Center to create mappings, sessions and workflows for populating the data into the dimension, facts,
* Evaluated model efficiency and accuracy using Python and R scripts, refining models for improved performance.
* Using InformaticaRepository Manager maintained all the repositories of various applications, created users, user groups, and security access control
* Experimented with various classification algorithms, including Random Forest and Gradient Boosting, using **Python Scikit-Learn** for customer discount optimization.
* Built predictive models using Python and PySpark to forecast attendance probabilities for campaigns and events.
* Used Pmcmdcommands to run the Informatica mappings using Unix Scripts.
* Conducted data visualization and designed interactive dashboards using **Tableau**, generating comprehensive reports for team and stakeholder interpretation.

**Environment:** Kafka, AI, Powershell, SQL, Matplotlib, Seaborn, SciPy, Scikit-Learn, Numpy, Pandas, R/R Studio, Flash, HBase, Hive, Hadoop, AWS S3, SnowSQL, and Snowflake are among the apparatuses and innovations used.

**IU Health, Indianapolis, IN Sep 2017- Dec 2019**

**Role: AWS Data Engineer**

**Project Summary:** The job involved planning and creating ETL processes involving PySpark in AWS Paste to relocate information from S3 for report age. Obligations included arranging fundamental AWS cloud administrations, overseeing foundation with arrangement apparatuses like CloudFormation and Terraform, and laying out cloud information stores in S3. Furthermore, the position included setting up CI/Compact disc pipelines, carrying out information ingestion modules utilizing AWS Paste, utilizing the Kinesis family for streaming information, and computerizing processes with Lambda works and Step Capabilities.

**Responsibilities:**

* Designed and Developed **ETL Processes** with **spark** in **AWS Glue** to migrate data from **S3** to generate Reports.
* Responsible for configuring essential **AWS Cloud services** to ensure **scalability, flexibility, and cost optimization**.
* Implemented **VPCs** and **subnets**, both private and public, across multiple regions and zones to effectively manage global operations.
* Managed **AWS infrastructure** utilizing orchestration tools like **CloudFormation (CFT)**, **Terraform**, and **Jenkins Pipeline**.
* Developed Terraform scripts to automate deployment processes for various **AWS** resources such as **EC2 instances**, **S3 buckets**, **EFS, EBS, IAM Roles, Snapshots**, and **Jenkins Server.**
* Established Cloud data stores in **S3**, structured with logical layers for managing Raw, Curated, and transformed data efficiently.
* Implemented data ingestion modules using **AWS Glue** to load data into different layers within **S3** and facilitate reporting using **Athena and QuickSight**.
* Defined bucket policies and lifecycle configurations for **S3 storage** in compliance with organizational and regulatory guidelines.
* Utilized AWS Systems Manager to create parameters and **SSM documents** for managing configurations and automation tasks.
* Set up continuous integration and continuous deployment (CI/CD) pipelines using **Jenkins** and **Git Bucket** for **version control,** build, and deployment of Python code base.
* Developed Glue Jobs for technical data cleansing tasks such as deduplication, NULL value imputation, and redundant column removal. Also implemented standard and business data transformations using Glue Jobs.
* Leveraged the **Kinesis Family** (including Kinesis Data streams, Kinesis Firehose, and Kinesis Data Analytics) for collecting, processing, and analyzing streaming data.
* Established **Athena data sources on S3 buckets** to enable **adhoc querying** and **business da**shboard creation using Quicksight and Tableau reporting tools.
* Transferred Fact/Dimension data and aggregate outputs from S3 to Redshift for historical data analysis, facilitating visualization using **Tableau and Quicksight**.
* Implemented **Lambda functions** and **Step Functions** to automate the triggering of Glue Jobs and orchestrate the data pipeline.
* Utilized **PyCharm IDE** for **Python/PySpark** development and **Git** for version control and repository management.

**Environment:** AWS - EC2, S3, VPC, EBS, ELB, CloudWatch, CloudFormation, ASG, Lambda, AWS CLI, GIT, Glue, Athena and QuickSight, Python and PySpark, Shell scripting, Jenkins

**Flipkart , Bengaluru, India**     Feb **2015 - Jul 2016**

**Role: Data Engineer**

**Project Summary:** The role involved working extensively with database objects such as tables, procedures, functions, and views, along with ETL process development using AWS Glue to migrate campaign data into Redshift. Responsibilities also included debugging and fixing issues in order management systems, developing data processing solutions with Apache Spark on AWS EMR, and optimizing database performance on Redshift. Additionally, the position entailed implementing data governance frameworks, building data ingestion workflows, collaborating with stakeholders, and ensuring day-to-day production support operations.

**Responsibilities:**

* Worked on tables, packages, procedures, functions, collections, triggers, cursors, ref cursors, exceptions, views, synonyms, sequence, performance tuning, interfaces, **API**, Lookups, processing constrain.
* Create, design, and develop ETL processes in **AWS Glue** to migrate campaign data from external sources like **S3, ORC/Parquet/Text files** into **AWS Redshift**.
* Debugged Order management, Purchase order, and Pricing issues in **IAT, UAT**, and production and fixed the issue.
* Developed DE fix scripts for the hold orders and corrected the process for the old existing orders.
* Prepared **test plans** and test cases for various types of testing like unit, functional, performance, and regression.
* Optimized database performance by tuning **SQL queries**, indexing strategies, and partitioning large datasets on **Amazon Redshift**.
* Implemented data governance frameworks and data quality checks to ensure compliance with regulatory standards.
* Building data ingestion workflows using **AWS Glue, AWS Data Pipelin**e, or custom scripts to extract, transform, and load (ETL) data from various sources.
* Developing data processing solutions using **Apache Spark**, leveraging **AWS EMR** for distributed computing.
* Implementing data lakes and data warehousing solutions on **AWS**, optimizing data storage and retrieval performance.
* Creating, debugging, scheduling, and monitoring jobs using **Airflow** for **ETL batch** processing to load into Snowflake for analytical processes.
* Worked on preparing an estimation plan to implement the change request based on the code freeze dates in different instances.
* Collaborated with business stakeholders to understand requirements and deliver data-driven insights and reports.
* Responsible for day-to-day Production **Support operations**, **Job monitoring, Incident ticket resolution,** timely deliveries, and code deployment**.**

**Environment:**Python,AWS Redshift,PostgreSQL, SQL, TOAD, AWS Glue, Shell Scripts, UNIX, Airflow

**Omega HealthCare**

**Bengaluru, India Aug 2014 - Jan 2015**

**Role: Data Analyst**

**Project Summary:** The role involved designing and delivering business reports and dashboards using Excel and Tableau for senior leaders, analyzing data sets to understand construction delays, and evaluating cost projections using regression analysis. Responsibilities also included collaborating on product design, staying updated on Power BI capabilities, supporting forecasting with R, leading a team in Excel, and managing project stakeholder interactions and dependencies.

**Responsibilities:**

* Design, develop, and deliver business reports and dashboards using **Excel** and **Tableau** to provide senior leaders with insights into data trends and patterns.
* Apply advanced Excel functions, such as **IF, SUM, INDEX, MATCH, VLOOKUP**, and **PIVOT** tables, to validate data.
* Analyze **data sets** to identify key information that can help a team understand trends in construction delays.
* Collaborate with the **Lead Designer** on product design and provide input on architectural issues as necessary.
* Evaluate cost projections using regression analysis in **Excel** to ensure that projects are profitable and meet financial goals.
* Manage the interactions between all project stakeholders, including architects, managers, and contractors.
* Stay up-to-date with the latest **Power BI** capabilities through self-learning activities.
* Support the forecasting process using the **R** statistical language.
* Lead and oversee a team of 10 colleagues in **MS Excel**.
* Identify and manage dependencies and risks related to work assignments.

**Environment:** Microsoft Excel, Tableau, Power BI, MS Office, R.