Data Engineer

Name: Kiran Kumar
Location: Newark, Delaware
Mail Id: kiran.jobs1995@gmail.com
Phone: +1(215)-253-7649

LinkedIn URL: https://www.linkedin.com/in/kiran-kumar-633836173/

Summary:

- With AWS and Power BI Certified Data Engineer specialist and experience in Big Data/Cloud Engineering, Data Warehousing, Data Modelling, Data Visualization, Reporting, Data Quality, and Data Analytics.
- Strong practical experience in cloud data migration utilizing AWS, Azure, and Snowflake.
- Experienced in AWS and Azure deployments, focusing on transferring on-premises servers and data to the cloud.
- Used AWS services like EC2 and S3 for small data processing and extensive experience administering Hadoop clusters running on AWS EMR.
- Detailed exposure to Azure tools such as Azure Data Lake, Azure Databricks, Azure Data Factory, HDInsight, Azure SQL Server, and Azure DevOps.
- Good understanding of the Snowflake Data Platform, including Snowflake Multi-Cluster Warehouses, Importing data from local systems and AWS S3 Buckets, Snowflake Database, Schema and Table structures, Snowflake Clone, and Time Travel.
- Experience with the Big Data ecosystem, including Hadoop MapReduce, NoSQL, Apache Spark, PySpark, Python, Scala, Hive, Impala, Sqoop, AWS, Azure.
- Experience in ETL pipelines in and out of data warehouses using Python and Snowflake's SnowSQL to extract, load, and transform data, and writing SQL queries against Snowflake.
- Proficient in SSAS, SSRS, SSIS, Amazon Redshift, Azure Data Warehouse, and Teradata.
- Skilled in data analysis techniques using Python libraries like NumPy, Pandas, and SciPy, and visualization libraries.

JP Morgan Chase & Co — **Wilmington, Delaware**
Data Engineer — **Oct 2023 – Present**
Responsibilities:

- Developed data pipeline definitions in JSON format for production code.
- Extensively used AWS Athena to import structured data from S3 into various systems, including Redshift, and generate reports.
- Worked on Snowflake modeling, proficient in data warehousing techniques for data cleansing, slowly changing dimensions (SCD), surrogate key assignment, and change data capture (CDC).
- Extracted, transformed, and loaded data into CSV files using Python and SQL queries.
- Used Data Build Tool (DBT) to create SQL queries for data transformations, generating datasets and models in Snowflake.
- Designed, developed, and maintained data integration applications for standard and non-traditional source systems, working with RDBMS and NoSQL data storage in Hadoop and RDBMS contexts.

- Analyzed Hive data using the Spark API and EMR Cluster Hadoop YARN, enhancing existing Hadoop algorithms using Spark Context, Spark SQL, Data Frames, and Pair RDDs.
- Developed AWS Lambda functions to monitor EMR cluster status updates and jobs.
- Designed Jenkins jobs to integrate processes and executed CI/CD pipelines using Jenkins.
- Developed ETL systems for data extraction, transformation, and loading from various sources. Launched and configured Amazon EC2 instances for individual applications using AWS (Linux/Ubuntu).
- Supported continuous storage in AWS using Elastic Block Storage (EBS), S3, and Glacier.
- Created volumes and configured snapshots for EC2 instances.
- Created on-demand tables on S3 files using Lambda Functions and AWS Glue with Python and PySpark.

Elevance Health **Cincinnati, Ohio**
Data Engineer **Jan 2023 – Sep 2023**
Responsibilities:

- Maintain a Python framework for data processing and write quality checks for processed data.
- Write SQL queries to ensure data adherence to the schema and identify discrepancies.
- Troubleshoot and develop production hot fixes in case of failures. Build numerous Lambda functions using Python and automate processes using event creation.
- Analyzed, created, and developed data solutions to enable data visualization using Azure PaaS services.
- Contributed to developing PySpark Data Frames in Azure Databricks, enabling users to read data from Data Lake or Blob Storage and manipulate it using Spark SQL context.
- ETL data from various source systems to Azure Data Lake Storage (ADLS) using a combination of Azure Data Factory (ADF), Spark SQL, and Azure Databricks for data processing.
- Install and configure Cloudera Hadoop Distribution.
- Build ETL pipelines to scale up data processing flow to meet rapid data growth, improving existing algorithms using Spark-Context, Spark-SQL, Data Frame and Spark YARN.
- Implement Teradata-specific features like PI, USI/NUSI, PPI, and compression based on requirements.
- Develop a CI/CD pipeline. Write Python modules to extract/load asset data from the MySQL source database. Perform end-to-end unit testing and document results in unit test plans.
- Involved in data mining solutions and generating visualizations using Tableau, Power BI.

Ciesto Information Technologies **Hyderabad, India**
Jr. Data Engineer/ Data Analyst **Sep 2017 – Jul 2021**
Project: 1
Responsibilities:

- Performed ETL using Python and Redshift to read data from Amazon S3 service. Created scripts in Python to read CSV, JSON, and Parquet files from S3 buckets and load them into Redshift. Used Apache Airflow to orchestrate ETL workflows.
- Designed Fact and Dimension tables using Snowflake methodologies with PostgreSQL.

- Set up S3 buckets and Access Control policies using IAM. Configured IAM roles and attached policies, set up Virtual Private Cloud (VPC) components (subnets, Internet Gateway, Security Groups), and managed EC2 instances for an AWS Redshift cluster using Python AWS SDK.
- Processed files from Data Lake to populate Fact and Dimension tables using Apache Spark and wrote them back to S3 in Parquet format. Worked on PowerBi dashboards and Ad-Hoc DAX quires for PowerBi.
- Created documentation in Markdown language and Jupiter Notebooks.

Project: 2

Responsibilities:

- Developed and optimized complex SQL queries for ETL processes and data analysis, working with databases such as SQL Server 2008, Teradata 13.1, DB2, MS SQL, and Excel.
- Involved in logical modelling using dimensional modelling techniques such as star schema and snowflake schema.
- Created reusable SSIS packages to extract data from multi-formatted flat files, Excel, and XML files into databases and billing systems.
- Developed PL/SQL programming, including stored procedures and triggers, and worked with DataStage and DB2.
- Utilized SDLC and Agile methodologies. Generated periodic reports based on statistical data analysis using SQL Server Reporting Services (SSRS).
- Extracted and loaded data from flat files, SQL Server, Oracle, DB2, and Sybase into flat files, Oracle, and SQL Server using DataStage.

Education:

- Master's in information Technology, Wilmington University Aug 2021- Dec 2022
- Bachelor's in engineering, Acharya Nagarjuna University. Jun 2012- May 2016

Technical Skills:

- AWS: Amazon EC2, Amazon S3, Amazon SimpleDB, Amazon MQ, Amazon ECS, Amazon Lambda, Amazon RDS, Amazon Elastic Load Balancing, Elasticsearch, Amazon SQS, AWS Identity and Access Management, AWS CloudWatch, Amazon EBS, Amazon CloudFormation, AWS SageMaker, AWS Glue, AWS Athena
- MS Azure: Cloud Services (PaaS & IaaS), Active Directory, Application Insights, Azure Monitoring, Azure Search, Data Factory, Key Vault, SQL Azure, Azure DevOps, Azure Analysis Services, Azure Synapse Analytics (DW), Azure Data Lake
- Big Data Technologies: Hadoop, MapReduce, Sqoop, Hive, Spark, Zookeeper, Kafka
- ETL Tools: Snowflake, Data Build Tool (dbt), Informatica
- Reporting Tools: Power BI, Tableau Application Servers: Apache Tomcat, WebSphere
- Hadoop Distributions: Hortonworks, Cloudera
- Programming & Scripting Languages: Python, Scala, SQL, Shell Scripting
- Databases: Oracle, MySQL, Teradata, HBase, Cassandra, DynamoDB
- Version Control: GIT IDE Tools: Eclipse, Jupyter
- Development Methodologies: Agile, Waterfall