**Manjula M S**

**Data Engineering**

**Email: manjumsiva@gmail.com**

**Phone: 669-241-0543**

**PROFESSIONAL SUMMARY**

* Highly dedicated, inspiring, and expert Sr Data Engineer with an overall **10+** years of IT industry experience exploring various technologies, tools, and databases.
* Good working knowledge of Snowflake and Teradata databases.
* Excellent Programming skills at a higher level of abstraction using Scala and Python.
* Hands-on experience in developing **SPARK** applications using Spark tools like RDD transformations, Spark core, Spark ML lib, Spark Streaming, and Spark SQL.
* Strong experience and knowledge of real-time data analytics using Spark Streaming, Kafka, and Flume.
* Working knowledge of Amazon's Elastic Cloud Compute (EC2) infrastructure for computational tasks and Simple Storage Service (S3) as a Storage mechanism.
* Worked on reading multiple data formats on HDFS using Scala.
* Experienced with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDDs, Spark YARN
* Knowledge about unifying data platforms using Kafka producers/ consumers and implementing pre-processing using storm topologies.
* Experience data processing like collecting, aggregating, and moving from various sources using Apache Kafka.
* Developing **ETL pipelines** in and out of the data warehouse using a combination of Python and Snow SQL
* I have expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Good experience in working with cloud environments like Amazon Web Services (AWS**)** EC2andS3.
* Experience in Implementing Continuous Delivery pipelines with Maven, Ant, Jenkins, and AWS.
* Configured, supported, and maintained all networks, firewall, storage, load balancers, operating systems, and software in AWS EC2.
* Experience with using PostgreSQL in a cloud environment, such as AWS, Azure, or GCP.
* Experience in AWS EC2, configuring the servers for Auto scaling and Elastic load balancing.
* Experience with all stages of the SDLC and Agile Development model right from the requirement gathering to Deployment and production support.
* Managing Database, Azure Data Platform services (Azure Data Lake (ADLS), Data Factory (ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes
* Excellent understanding of best practices of Enterprise **Data Warehouse** and involved in Full life cycle development of Data Warehousing.
* Involved in building **Data Models and Dimensional Modeling** with 3NF, Star, and Snowflake schemas for OLAP and Operational data store (ODS) applications.
* Hands-on experience working on NoSQL databases including HBase and its integration with Hadoop cluster.
* Good working experience using Sqoop to import data into HDFS from RDBMS and vice-versa.
* Good knowledge of stored procedures, functions, etc. using SQL and PL/SQL.
* Expertise in using Version Control systems like GIT.
* Knowledge of Amazon Web Services and Microsoft Azure.
* Optimized and tuned ETL processes and SQL Queries for better performance.
* Performed complex data analysis and provided critical reports to support various departments.
* Extensive Python scripting experience for Scheduling and Process Automation.
* Experience with Unix/Linux systems with scripting experience and building data pipelines.
* Responsible for migration of applications running on-premises onto Azure cloud.
* Experience in Cloud Databases and Data warehouses (SQL Azure and Confidential Redshift/RDS).

**TECHNICAL SKILLS**

RDBMS: **Snowflake**, **Microsoft SQL Server 2017, Teradata 15.0, Oracle 12c, MS Access, RDBMS, MySQL, DB2, Hive, Microsoft Azure SQL Database**

BI Tools: **Tableau 10, Tableau server 10, Tableau Reader 10, SAP Business Objects, Crystal Reports**

Methodologies: **Agile, RAD, JAD, RUP, UML, System Development Life Cycle (SDLC), Waterfall Model**.

Programming Languages: **SQL, PL/SQL, UNIX shell Scripting, PERL, AWK, SED, Python**.

Other Tools: **TOAD, BTEQ, MS-Office suite (Word, Excel, Project and Outlook).**

Cloud Management: **AWS (Amazon Web Services), MS Azure**

Hadoop Ecosystem: **MapReduce, Spark 2.3, HBase 1.2, Hive 2.3, Pig 0.17, Solr 7.2, Flume 1.8, Sqoop 1.4,**

**Kafka 1.0.1, Oozie 4.3, ] Neo4j, Hadoop 3.0, Apache Nifi 1.6, Cassandra 3.11**

**PROFESSIONAL EXPERIENCE**

**Client: Door dash, CA Nov 2022 – Till Date**

**Role: Sr. Data Engineer**

**Responsibilities:**

* As a Data Engineer, Maintained data pipelines, data integration, ETL processes, and data management delivery within a global data lake and data warehouse environment.
* Interact with senior management/architects of business & technical teams to develop quarterly
* Objectives and Key Results (OKRs).
* Design, develop and manage the Data Warehouse in Redshift, Snow Flake & Data Lake for
* Analytics and Reporting
* Building Data pipelines & Data integration using Snowflake’s snow pipes.
* Interact with the Business Partners (Marketing, Engineering, Finance & Product team) to
* understand the business of the application, gather requirements to define key metrics and provide
* data for analysis.
* Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) Big Query.
* Utilize Data Frames and Spark SQL for data processing.
* Responsible for Data architecture Design and decision. Design and implement data schemas and models
* for a data warehouse.
* Experience in developing Spark Programs for Batch and Real-Time Processing. Developed Spark Streaming applications for Real-Time Processing.
* Used Spark Streaming to divide streaming data into batches as input to the Spark engine for batch processing.
* Migrated on-premises environment in GCP (Google Cloud Platform)
* Performed Data Ingestion from multiple internal clients using Apache Kafka.
* Worked on integrating Apache **Kafka** with **Spark** Streaming process to consume data from external REST APIs and run custom functions.
* Developed Spark scripts by using Scala Shell commands as per the requirement.

***Environment*:** Spark, Spark-Streaming, Spark SQL, Python, Jira, GitHub, GCP,S3, Redshift, Snowflake, Hadoop, HIVE, Presto, SQL, Spark SQL, Kafka, JSON ,RDS, API’s, Bash, Airflow

**Client: Honeywell, CA June 2019 – Oct 2022**

**Role: Sr. Data Engineer**

**Responsibilities:**

* Involved in implementing and following an Agile development methodology within the cross-functional team.
* Involved in porting the existing on-premises Hive code migration to GCP (Google Cloud Platform) BigQuery.
* Developed comprehensive test suites using Python and PY Spark to validate data transformations and quality checks at each stage of the data processing pipeline.
* Experience in designing and implementing data lake architectures with **Amazon Athena**.
* Conducted database migration to **RDS** from on-premises databases.
* Experience in writing and deploying cloud functions on **AWS Lambda**.
* Proficient in designing and implementing complex **Spark SQL**-based data processing pipelines that involve ETL, and data warehousing using Spark Data Frames and Spark SQL.
* Worked with application development teams to implement serverless architectures and event-driven computing using **AWS Lambda** and **AWS API Gateway**.
* Skilled in creating and managing **AWS SNS** topics and subscriptions, enabling pub/sub messaging for real-time data processing and notifications.
* Experience in setting up and configuring **Kinesis data streams**, enabling ingestion and processing of high-volume and high-velocity data.
* Integrated **AWS ECR** seamlessly with **AWS ECS** and **EKS** (Elastic Kubernetes Service) for seamless deployment and scaling of containerized data services
* Proficient in **Snowflake's** data loading mechanisms, including bulk loading and Snow pipe.
* Skilled in optimizing Spark SQL queries and improving query performance by tuning Spark configurations, leveraging Spark caching, and using advanced SQL optimization techniques.
* Familiarity with Oozie's integration with other AWS services, such as **AWS S3**, **AWS Glue**, and **AWS Lambda**, to build serverless data processing pipelines.
* Strong understanding of data security and access control in **Athena**, including AWS Identity and Access Management (IAM) roles and policies.
* Orchestrated **CI/CD** workflows with **GitHub** Actions, enabling continuous integration and deployment of data pipelines.
* Involved in writing **Python** scripts to extract data from different API’s.
* Extensive experience in building and maintaining data pipelines on AWS Databricks using **Python** and SQL.
* Leveraged **TDD** practices to maintain code quality, reduce bugs, and improve the efficiency of data processing workflows.
* Design and Develop ETL Processes in **AWS Glue** to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into **AWS Redshift**.
* Designed and implemented data warehousing solutions using **AWS Redshift**, including complex data modeling, tuning, and optimization.
* Created and managed **RDS** instances for various database engines such as MySQL, PostgreSQL, Oracle, and SQL Server.
* Experience in designing and implementing **Snowflake** data models, schemas, tables, and views to support efficient data organization and querying.
* Experience in integrating **SSIS** with other Microsoft BI tools like SQL Server Analysis Services (SSAS) and SQL Server Reporting Services (SSRS) to deliver end-to-end BI solutions.
* Designed and implemented scalable, fault-tolerant, and highly available data architectures using AWS services such as Elastic Load Balancing, Auto Scaling, and **CloudFormation**.
* Hands-on experience in building ETL pipelines using **AWS Glue** using AWS SDKs.
* Experience in setting up **AWS CloudWatch** metrics, alarms, and dashboards to monitor and visualize data engineering workflows and infrastructure.
* Skilled in writing and managing DAGs in **Airflow**, representing data workflows as code, and enabling modularity and reusability.
* Experience with real-time streaming data processing using **AWS Databricks** Streaming and integrating with AWS services like Kinesis or **Kafka.**
* Integrated **Terraform** with other DevOps tools, such as Ansible and Jenkins, to automate infrastructure deployment pipelines and streamline continuous integration and continuous deployment (CI/CD) processes.
* Expertise in designing and implementing scalable and fault-tolerant applications using Amazon DynamoDB, a fully managed NoSQL database service in AWS.
* Created ETL Pipeline using Spark and Hive for ingest data from multiple sources.
* Analyzed the **SQL** scripts and designed the solution to implement using **PySpark**.

***Environment:*** Redshift, Glue, Athena, Snowflake DB, PySpark, RDS, Lambda, Databricks, **Python** 3.5, Kafka2.8, Oozie, JIRA, Rest API.GCP

**Client: BMC CA, USA Feb 2016 – May 2019**

**Role: Data Engineer/Data Analyst**

**Responsibilities:**

* Worked with the analysis teams and management teams and supported them based on their requirements.
* Generated PL/SQL scripts for data manipulation, validation, and materialized views for remote instances.
* Created and modified several database objects such as Tables, Views, Indexes, Constraints, Stored procedures, Packages, Functions and Triggers using SQL and PL/SQL.
* Created large datasets by combining individual datasets using various inner and outer joins in SAS/SQL and dataset sorting and merging techniques using SAS/Base.
* Developed live reports in a drill down mode to facilitate usability and enhance user interaction.
* Extensively worked on Shell scripts for running SAS programs in batch mode on UNIX.
* Wrote Python scripts to parse XML documents and load the data in database.
* Used Python to extract weekly information from XML files.
* Developed Python scripts to clean the raw data.
* Developed a python script to transfer data, REST API’s and extract data from on-premises to AWS S3. Implemented **Micro Services** based Cloud Architecture using **Spring Boot**.
* Worked on Ingesting data by going through cleansing and transformations and leveraging AWS Lambda, AWS Glue
* Worked on AWS CLI to aggregate clean files in Amazon S3 and on Amazon EC2 Clusters to deploy files into Buckets.
* Used AWS CLI with IAM roles to load data to Redshift cluster.
* Responsible for in-depth data analysis and creation of data extract queries in both Netezza and Teradata databases.
* Used Scala to write code for all Spark use cases.
* Assigned name to each of the columns using case class option in Scala.
* Developed multiple Spark SQL jobs for data cleaning.
* Developed Spark SQL to load tables into HDFS to run select queries on top.
* Developed analytical component using Scala, Spark, and Spark Stream.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster processing of data.
* Designed custom Spark REPL application to handle similar datasets Creating Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Created functions and assigned roles in AWS Lambda to run python scripts, and AWS Lambda using java to perform event driven processing. Created Lambda jobs and configured Roles using AWS CLI.
* Used Apache Spark Data frames, Spark-SQL, Spark ML Lib extensively and developing and designing POC's using Scala, Spark SQL and ML lib libraries.
* Work on data that was a combination of unstructured and structured data from multiple sources and automate the cleaning using Python scripts.
* Designed reports in SSRS to create, execute, and deliver tabular reports using shared data source and specified data source.
* Used Hive, Impala and Sqoop utilities and Oozie workflows for data extraction and data loading.
* Development of routines to capture and report data quality issues and exceptional scenarios.
* Creation of Data Mapping document and data flow diagrams.
* Developed Linux Shell scripts by using Nzsql/Nzload utilities to load data from flat files to Netezza database.
* Created Dashboards style of reports using QlikView components like List box Slider, Buttons, Charts and Bookmarks.
* Worked on QA the data and adding Data sources, snapshot, caching to the report.
* Involved in troubleshooting at database levels, error handling and performance tuning of queries and procedures.

***Environment***: SAS, SQL, Teradata 13, Oracle 11c, PL/SQL, UNIX, XML, Python, AWS, SSRS, TSQL, Hive 2.1, Sqoop 1.2

**Client: T- Systems, India Oct 2012 – July 2014**

**Role: Data Engineer**

**Responsibilities:**

* Skilled in integrating Oozie with other AWS services such as **AWS CloudFormation**, AWS Step Functions, and **AWS** Data Pipeline, to build scalable and cost-effective data processing pipelines.
* Experienced in using **YAML** to define infrastructure as code (IAC) for cloud-based data solutions, ensuring reproducibility and consistency in data engineering environments.
* Design AWS architecture, Cloud migration, AWS EMR, Dynamo DB, Redshift, and event processing using lambda function.
* Proficient in designing and developing large-scale data solutions using **Teradata**, a leading relational database management system (RDBMS).
* Skilled in implementing Spark SQL-based data pipelines on large-scale, distributed clusters, including **AWS EMR**.
* Created private repositories in **ECR** to securely store **Docker** images for data engineering pipelines and data applications.
* Experience in integrating **SNS** with other AWS services, such as **AWS Lambda**, to trigger serverless functions based on events and notifications.
* Expertise in integrating **Athena** with other AWS services, such as **AWS Glue** for automated data cataloging and **AWS Lambda** for serverless data processing workflows
* Developed **Spark** streaming application to pull data from cloud to hive table and used Spark SQL to process the huge amount of structured data.
* Knowledgeable in working with Spark SQL on real-time data processing and streaming applications, using tools such as Apache **Kafka**.
* Automated cloud deployments using **python**, and **AWS Cloud Formation** Templates.
* Designed ETL process using Talend Tool to load from Sources to Targets through data Transformations.
* Used **Kafka** and **Spark** streaming to ingest real time or near real time data in HDFS.
* Experience in optimizing cost and managing data storage in **Athena**, leveraging partitioning, compression, and lifecycle policies in **Amazon S3**.
* Proficient in using **CRON**, a time-based job scheduler, to schedule and automate the execution of **shell scripts** and data processing tasks at specified intervals.
* Used **Hive** to analyze the partitioned and bucketed data and compute various metrics for reporting on the dashboard.
* Skilled in designing and implementing Oozie-based workflows for data processing and ETL.

***Environment:*** Hadoop, Rest API, Scala, **Python**, MySQL, MongoDB, Apace Airflow, Qu bole, SQL, AWS EMR, Clusters, Amazon S3, AWS Athena, Java, Jupyter notebooks, Jira, Tableau, JSON, XML, Hive, Dynamo DB, Kibana, Big Data.

**Client: Micro land Limited, India May 2010 to Sept 2012**

**Role: Software Engineer**

**Responsibilities:**

* Built a GC analytic tool using **Python,** to help clients analyze data and profile risk against a wide spectrum

of reinsurance options.

* Developed custom reports using **HTML, Python** and **MySQL**.
* Developed monitoring and notification tools using **Python**.
* Documented the revised workflow using confluence wiki.
* Created Data tables utilizing **PyQt** to display customer and policy information and add, delete, update
* customer records.
* Used **Python** based GUI components for the front-end functionality such as selection criteria.
* Key performance index and filtering of displayed columns to help customers understand the analytics data that

is provided by GC.

* Implemented navigation rules for the application and page outcomes, written controllers using annotations.
* Written queries in **MySQL** and Native **SQL**.
* Created **Py Unit** test cases for unit testing.
* Added the navigations and paginations, filtering columns, adding, and removing the desired columns for view.
* Worked in an **agile** development environment.
* Used **GitHub** for version control.
* Used **Py Query** for selecting particular **DOM** elements when parsing **HTML**.

 **Environment:** Python, Py QT, Py Query, MySQL, HTML, CSS, JavaScript, DOM, Ajax, agile, Web Services

 (SOAP), XML, PL/SQL, and Oracle SQL Developer.