MANAV SHARMA

**Data Engineer**

**Phone: 469-656-8349**

**Email: manav208sharma@gmail.com**

**PROFESSIONAL SUMMARY**

* Hands on 10+ years of professional IT experience, Specialized in Data Engineering / Azure / Snowflake.
* Results-driven Data Engineer with expertise in designing and implementing scalable data ingestion pipelines using Azure Data Factory, Oracle, SQL, PL/SQL, Terraform, EC2, Python, AWS, Spark and Snowflake.
* Snowflake and azure certified developer.
* Proficient in leveraging Azure Databricks and Spark for distributed data processing and transformation tasks.
* Proficient in using SnowSQL for complex data manipulation tasks and developing efficient data pipelines.
* Skilled in the art of devising roles, shaping views, and implementing techniques for tuning performance, all geared towards elevating the efficiency of the Snowflake system.
* Exhibit mastery in managing data storage and retrieval through AWS S3, capitalizing on its scalability and cost-effectiveness.
* Skillfully navigate AWS EMR for the processing of large-scale data, encompassing technologies such as Hadoop, Spark, Hive, MapReduce, and PySpark.
* Experienced in crafting partitioning strategies and orchestrating multi-cluster warehouses within Snowflake, ensuring the optimization of query speed and system scalability along with Amazon Event Bridge & AWS Step Functions and suitable Lambda functions.
* Experienced in utilizing AWS Glue for ETL workflows, enabling efficient data extraction, transformation, and loading.
* Strong knowledge of AWS CloudWatch for monitoring and managing AWS resources, setting up alarms, and collecting metrics.
* Possess hands-on experience with relational databases, proficiently using SQL, sqlalchemy, and managing data through AWS DynamoDB.
* Hands on experience in implementing data pipeline solutions using Hadoop, azure, ADF, Synaps, Pyspark, Map-Reduce, Hive, Tez, Python, Scala, Azure functions, Logic apps, stream sets, ADLS Gen2 and snowflake.
* Ability to troubleshoot and optimize Datastage jobs for performance and efficiency.
* Proficient in deploying and managing applications on GCP using tools like Cloud Deployment Manager and Cloud Console.
* Strong knowledge of GCP networking concepts and services like Virtual Private Cloud (VPC) and Cloud Load Balancing.
* Experience in using Snowflake Clone and Time Travel.
* Participates in the development improvement and maintenance of snowflake database applications
* Build the Logical and Physical data model for snowflake as per the changes required
* Implemented data pipelines using snow sql, Snowflake Integrated services and snow pipe.
* Implemented Server-less solutions using azure functions. Experience in using Snowflake Clone and Time Travel.
* Hands on experience in using ORC, Parquet, Avro and delimited file formats.
* Participates in the development improvement and maintenance of snowflake database applications
* Skilled in ensuring data quality and integrity through validation, cleansing, and transformation operations.
* Build the Logical and Physical data model for snowflake as per the changes required
* Implemented query performance in Hive using bucketing and partitioning techniques, and have extensive hands-on experience tuning spark Jobs.
* Experience with Snowflake Multi - Cluster Warehouses.
* Strong expertise in optimizing Spark jobs and leveraging Azure Synapse Analytics for big data processing and analytics.
* Proven track record in performance optimization and capacity planning to ensure scalability and efficiency.
* Experienced in developing CI/CD frameworks for data pipelines and collaborating with DevOps teams for automated pipeline deployment.
* Adept at designing cloud-based data warehouse solutions using Snowflake, optimizing schemas, tables, and views for efficient data storage and retrieval.
* Proficient in scripting languages such as Python and Scala.
* Skilled in working with Hive, SparkSQL, Kafka, and Spark Streaming for ETL tasks and real-time data processing.
* Have a good working experience in Hadoop, HDFS, Map-Reduce, Hive, Teg, Python, PySpark.
* Hands on working experience and developing large scale data pipelines using spark and hive
* Experience in using Apache Sqoop to import and export data from HDFS and Hive.
* Hands-on experience in setting up workflow using Apache Oozie workflow engine for managing and scheduling Hadoop jobs.
* I am experienced in setting up workflow using Apache Oozie workflow engine for managing and scheduling Hadoop jobs, and importing and exporting the data using SQOOP from HDFS to Relational Database systems.
* I have optimized query performance in Hive using bucketing and partitioning techniques, and have extensive hands-on experience tuning spark Jobs.
* Highly proficient in Agile methodologies, including JIRA for project management and reporting.

**Technical Skills:**

|  |  |
| --- | --- |
| Azure Services | Azure data Factory, Azure Data Bricks, snowflake, Logic Apps, GCP, Functional App, Snowflake, Azure DevOps, Purview, DLP |
| Big Data Technologies | MapReduce, Hive, Teg, Python, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop, Zogcpokeeper. |
| Hadoop Distribution | Cloudera, Horton Works |
| Languages | SQL, PL/SQL, Python, HiveQL, Scala. |
| Web Technologies | HTML, CSS, JavaScript, XML, JSP, Restful, SOAP |
| Operating Systems | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| Build Automation tools | Ant, Maven |
| Version Control | GIT, GitHub. |
| IDE &Build Tools, Design | Eclipse, Visual Studio. |
| Databases | MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB |

**Education:**

* Bachelors from Jawaharlal Nehru Technological University Hyderabad, India.

**Professional Experience:**

**Role: Azure Snowflake data engineer. May 2022 – Till Now**

**Client: Cisco, Richardson, Texas**

* Designed and implemented scalable data ingestion pipelines using Azure Data Factory, ingesting data from various sources such as SQL databases, CSV files, and REST APIs.
* Developed data processing workflows using Azure Databricks, leveraging Spark for distributed data processing and transformation tasks.
* Designing and implementing data processing pipelines using GCP technologies like BigQuery, Dataflow, and Pub/Sub.
* Ensured data quality and integrity by performing data validation, cleansing, and transformation operations using Azure Data Factory and Databricks.
* Designed and implemented a cloud-based data warehouse solution using Snowflake on Azure, leveraging its scalability and performance capabilities.
* Created and optimized Snowflake schemas, tables, and views to support efficient data storage and retrieval for analytics and reporting purposes.
* Implemented Informatica for diverse source integration, mapping transformations, and data orchestration
* Collaborated with data analysts and business stakeholders to understand their requirements and implemented appropriate data models and structures in Snowflake.
* Developed and optimized Spark jobs to perform data transformations, aggregations, and machine learning tasks on big data sets.
* Building and managing scalable and reliable architectures on GCP.
* Leveraged Azure Synapse Analytics to integrate big data processing and analytics capabilities, enabling seamless data exploration and insights generation.
* Configured event-based triggers and scheduling mechanisms to automate data pipelines and workflows.
* Implemented data lineage and metadata management solutions to track and monitor data flow and transformations.
* Identified and resolved performance bottlenecks in data processing and storage layers, optimizing query execution and reducing data latency.
* Implemented partitioning, indexing, and caching strategies in Snowflake and Azure services to enhance query performance and reduce processing time.
* Conducted performance tuning and capacity planning exercises to ensure the scalability and efficiency of the data infrastructure.
* Developed CI/CD framework for data pipelines using Jenkins tool.
* Collaborated with DevOps engineers to developed automated CI/CD and test-driven development pipeline using azure as per the client requirement.
* Hands on programming experience in scripting languages like python and Scala
* Involved in running all the Hive scripts through Hive on Spark and some through SparkSQL
* Collaborated on ETL tasks, maintaining data integrity and verifying pipeline stability.
* Worked on combination of structured and unstructured data from multiple sources and automated the cleaning process using python scripts.
* Hands on experience in using Kafka, Spark streaming, to process the streaming data in specific use cases.
* Developed a data pipeline using Kafka, Spark, and Hive to ingest, transform and analyzing data.
* Designed and implemented real-time data processing solutions using Kafka and Spark Streaming, enabling the ingestion, transformation, and analysis of high-volume streaming data.
* Developed Spark core and Spark SQL scripts using Scala for faster data processing.
* Working with JIRA to report on Projects, and creating sub tasks for Development, QA, and Partner validation.
* Experience in full breadth of Agile ceremonies, from daily stand-ups to internationally coordinated PI Planning

**Environment**: Azure Databricks, Data Factory, Snowflake, GCP, Logic Apps, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Purview, DLP, GIT, JIRA, Jenkins, kafka, ADF Pipeline, Power Bi.

**Role: Azure Snowflake data engineer. Nov 2021 – April 2022**

**Client: Colaberry, Arizona**

* Implemented end-to-end data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from diverse sources into Snowflake.
* Designed and implemented data processing workflows using Azure Databricks, leveraging Spark for large-scale data transformations.
* Built scalable and optimized Snowflake schemas, tables, and views to support complex analytics queries and reporting requirements.
* Developed data ingestion pipelines using Azure Event Hubs and Azure Functions to enable real-time data streaming into Snowflake.
* Experience with performance tuning and optimization in Azure Synapse, including optimizing SQL queries and optimizing Spark jobs.
* Used Informatica Power Center for extraction, transformation, and loading (ETL) of data in the data warehouse.
* Building and managing scalable and reliable architectures on GCP.
* Leveraged Azure Data Lake Storage as a data lake for storing raw and processed data, implementing data partitioning and data retention strategies.
* Utilized Azure Blob Storage for efficient storage and retrieval of data files, implementing compression and encryption techniques to optimize storage costs and data security.
* Integrated Azure Data Factory with Azure Logic Apps for orchestrating complex data workflows and triggering actions based on specific events.
* Generated report on predictive analytics using Python and Tableau including visualizing model performance and prediction results.
* Implemented data governance practices and data quality checks using Azure Data Factory and Snowflake, ensuring data accuracy and consistency.
* Implemented data replication and synchronization strategies between Snowflake and other data platforms using Azure Data Factory and Change Data Capture techniques.
* Responsible for manipulating data set and building models for document classification and OCR using Python and SQL
* Developed and deployed Azure Functions for data preprocessing, data enrichment, and data validation tasks in data pipelines.
* Implemented advanced analytics and machine learning workflows using Azure Machine Learning and Snowflake, enabling predictive analytics and data-driven insights.
* Designed and implemented data archiving and data retention strategies using Azure Blob Storage and Snowflake's Time Travel feature.
* Developed custom monitoring and alerting solutions using Azure Monitor and Snowflake Query Performance Monitoring (QPM) for proactive identification and resolution of performance issues.
* Integrated Snowflake with Power BI and Azure Analysis Services for creating interactive dashboards and reports, enabling self-service analytics for business users.
* Optimized data pipelines and Spark jobs in Azure Databricks for improved performance, including tuning of Spark configurations, caching, and leveraging data partitioning techniques.
* Implemented data cataloging and data lineage solutions using tools like Azure Purview and Apache Atlas to provide a comprehensive understanding of data assets and their relationships.
* Collaborated with cross-functional teams including data scientists, data analysts, and business stakeholders to understand data requirements and deliver scalable and reliable data solutions.

**Environment**: Azure Databricks, Data Factory, Logic Apps, GCP, Snowflake, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, Purview, DLP, GIT, JIRA, Jenkins, kafka, ADF Pipeline, Power Bi.

**Role: AWS Snowflake Data Engineer. Aug 2017 – Oct 2021**

**Client: American Express, Phoenix, AZ**

* Designed and implemented Snowflake stages to efficiently load data from various sources into Snowflake tables.
* Created and managed different types of tables in Snowflake, such as transient, temporary, and persistent tables and optimized Snowflake warehouses by selecting appropriate sizes and configurations to achieve optimal performance and cost efficiency.
* Architect and implement end-to-end data pipelines on AWS, integrating disparate data sources and ensuring data quality and consistency.
* Crafted intricate SnowSQL queries, orchestrating the extraction, transformation, and loading of data from diverse sources into the Snowflake platform.
* Applied advanced partitioning techniques within Snowflake, elevating both query performance and the efficiency of data retrieval.
* Effectively configured and oversaw multi-cluster warehouses in Snowflake, ensuring their capacity to adeptly manage high-concurrency workloads.
* Established access roles and privileges within Snowflake, meticulously crafting a framework for data security and governance.
* Deployed innovative Snowflake caching mechanisms, resulting in optimized query performance and reduced costs associated with data transfer.
* Worked with Snowpipe to seamlessly ingest real-time data into the Snowflake environment, guaranteeing a steady flow of data and automating the loading process.
* Harnessed the power of Snowflake's time travel functionalities, enabling the tracking and restoration of historical data to facilitate both auditing and analytical pursuits.
* Carried out regular expressions in Snowflake, employing them for tasks involving pattern recognition and the extraction of specific data.
* Integrated AWS SNS and SQS for real-time event processing and messaging.
* Implemented AWS Athena for ad-hoc data analysis and querying on S3 data.
* Utilized AWS CloudWatch for monitoring, setting up alarms, and collecting metrics along with managing resources
* Created intricate Snowflake scripting solutions, effectively automating the execution of data pipelines, ETL processes, and transformative operations. Developed Snowflake scripting solutions to automate data pipelines, ETL processes, and data transformations.
* Developed ETL workflows using AWS Glue to extract, transform, and load data from various sources into Redshift.
* Designed and implemented data streaming solutions using AWS Kinesis for real-time data processing.
* Managed DNS configurations and routing using AWS Route53 for efficient application and service deployment.
* Developed data processing pipelines using Hadoop, including HDFS, Sqoop, Hive, MapReduce, and Spark.
* Implemented Spark Streaming for real-time data processing and analytics.
* Build scheduling and job automation using IBM Tivoli, Control-M, Oozie, and Airflow.
* Designed and developed database solutions using Teradata, Oracle, and SQL Server.
* Proficient work in version control systems such as Git, GitLab, and VS for efficient code repository management and collaborative development processes.

**Environment**: AWS, AWS S3, redshift, EMR, SNS, SQS, AWS Athena, glue, Cloudwatch, Kinesis, Route53, IAM, Sqoop, MYSQL, HDFS, Apache Spark, Hive, Cloudera, Kafka, Zookeeper, Oozie, PySpark, Ambari, JIRA, IBM Tivoli, control-m, OOZIE, airflow, Teradata, oracle, SQL

**Role: Snowflake data Developer. May 2014 – Jul 2017**

**Client: Taylor Technology, Dallas Texas**

* Prepared an ETL framework using Sqoop, Pig, and Hive to bring in data from various sources and make it available for consumption.
* Processed HDFS data and created external tables using Hive, along with developing scripts for table ingestion and repair for reuse across the project.
* Developed ETL jobs using Spark and Scala to migrate data from Oracle to new MySQL tables.
* Utilized Spark (RDDs, DataFrames, Spark SQL) and Spark-Cassandra Connector APIs for various tasks, including data migration and business report generation.
* Developed a Spark Streaming application for real-time sales analytics.
* Analyzed source data, efficiently handled data type modifications, and used Excel sheets, flat files, and CSV files to generate Power BI ad-hoc reports.
* Analyzed SQL scripts and designed solutions using PySpark.
* Extracted data from various data sources into HDFS using Sqoop.
* Handled data import from various sources, performed transformations using Hive and MapReduce, and loaded data into HDFS.
* Extracted data from MySQL into HDFS using Sqoop.
* Implemented automation for deployments using YAML scripts for streamlined builds and releases.
* Worked with Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka, and Sqoop.
* Implemented data classification algorithms using MapReduce design patterns.
* Extensively worked on creating combiners, partitioning, and distributed cache to enhance the performance of MapReduce jobs.
* Utilized Git and GitHub repositories to maintain the source code and enable version control.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Shallscript, Cassandra, YAML, ETL.

**Role: Data Engineering Apr 2013 – Apr 2014**

**Client: TMW Systems, Cleveland Ohio**

* Imported data from MySQL to HDFS on a regular basis using Sqoop for efficient data loading.
* Performed aggregations on large volumes of data using Apache Spark and Scala, and stored the results in the Hive data warehouse for further analysis.
* Worked extensively with Data Lakes and big data ecosystems, including Hadoop, Spark, Hortonworks, and Cloudera.
* Loaded and transformed structured, semi-structured, and unstructured data sets efficiently.
* Developed Hive queries to analyze data and meet specific business requirements.
* Leveraged HBASE integration with Hive to build HBASE tables in the Analytics Zone.
* Demonstrated ability to design and implement data integration strategies between Snowflake and external systems, leveraging technologies such as Apache Airflow or custom-built orchestration frameworks to ensure seamless data movement and synchronization.
* Utilized Kafka and Spark Streaming to process streaming data for specific use cases.
* Developed data pipelines using Flume and Sqoop to ingest customer behavioral data into HDFS for analysis.
* Utilized various big data analytic tools, such as Hive and MapReduce, to analyze Hadoop clusters.
* Implemented a data pipeline using Kafka, Spark, and Hive for ingestion, transformation, and analysis of data.
* Wrote Hive queries and used Hive QL to simulate MapReduce functionalities for data analysis and processing.
* Migrated data from RDBMS (Oracle) to Hadoop using Sqoop for efficient data processing.
* Developed custom scripts and tools using Oracle's PL/SQL language to automate data validation, cleansing, and transformation processes.
* Implemented CI/CD pipelines for building and deploying projects in the Hadoop environment.
* Utilized JIRA for issue and project workflow management.
* Used Spark Streaming to process streaming data in batches for efficient batch processing.
* Leveraged Zookeeper to coordinate, synchronize, and serialize servers within clusters.
* Utilized the Oozie workflow engine for job scheduling in Hadoop.
* Used Git as a version control tool to maintain the code repository

**Environment**: Sqoop, MYSQL, HDFS, Apache Spark Scala, Hive Hadoop, Cloudera, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, Python, PySpark, Ambari, JIRA.