**Gopi Chand**

**Sr Data Engineer**

**+1 646-969-0565**

**PROFESSIONAL SUMMARY**

* Over 10 years of extensive hands-on experience in Big Data with expertise in Hadoop Eco System and cloud-based platforms.
* Deep knowledge and experience in Cloud Computing and Hadoop architecture, including components like HDFS, MapReduce, Spark, Name node, Data Node, Job Tracker, Task Tracker, and Secondary Name Node.
* Proven track record of migrating SQL databases to Azure Data Lake, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse, while effectively managing database access and security using Azure Data Factory.
* Strong experience in setting up AWS Data Platform, leveraging AWS CloudFormation, AWS Glue, EMR, Jupyter/SageMaker Notebooks, Redshift, S3, and EC2 instances.
* Extensive experience working with different Hadoop Distributions (Cloudera, Hortonworks, MapR, Amazon EMR) to implement and leverage new Hadoop features.
* Proficient in Google Cloud Platform Technologies like Big Query, Dataflow, Dataproc, Pub sub, and Airflow.
* Coded in Python, shell scripting and PL/SQL queries in my day-to-day activities
* Deep knowledge of troubleshooting and optimizing Spark applications and Hive scripts to achieve optimal performance.
* Expertise in developing Spark applications using Spark RDD, Spark-SQL, and Data frame APIs.
* Worked on the Tableau security setting up users using windows run as user. Setting up SSL /SMPT mail server configuration for alters on report failures.
* Developed complex workflows using Alteryx.
* Responsible for designing, building, and testing workflows in Alteryx .
* Improved Reports performance in tableau using extracts and user filters.
* Implemented security at the Data (Data security) and object (Roles and Permissions) level.
* Experience in Publishing Tableau Workbooks to Tableau Server and assiging User Access.
* Skilled in real-time data processing and streaming techniques using Spark Streaming and Kafka.
* Proficient in leveraging Alteryx for seamless data blending, cleansing, and transformation, optimizing ETL processes for efficient data integration.
* Demonstrated ability to harness Alteryx's powerful features to streamline workflows and enhance overall data quality, contributing to successful data-driven decision-making.
* Strong experience in data integration using Apache Sqoop to move data between HDFS and Relational Database Systems (RDBMS).
* Extensive experience in HIVE data warehouse infrastructure, including table creation, partitioning, bucketing, and optimizing HQL queries.
* Proficient in replacing existing MR jobs and Hive scripts with Spark SQL and Spark data transformations for efficient data processing.
* Proven experience in designing, developing, documenting, and testing ETL jobs in distributed environments.
* Skilled in developing Kafka producers and consumers for streaming high-volume event data.
* Strong database design, modeling, migration, and development skills with MySQL, MS SQL Server, DB2, and Oracle.
* Experience working with NoSQL databases including MongoDB, Cassandra, and HBase.
* Proficient in software development tools such as JIRA, Play, and GIT.
* Good understanding of data modeling concepts such as Dimensional & Relational modeling, Star-Schema Modeling, Snowflake schema modeling, Fact and Dimension tables.
* Expertise in manipulating and analyzing large datasets to identify patterns and insights within structured and unstructured data.
* Experience with Snowflake data warehouse architecture, including roles, warehouses, databases, schemas, tables, and views for data validation and complex SQL queries.
* Proficient in creating notebooks in Databricks platform using Python, Spark, PySpark, Scala, and SQL for various data transformations, joins, and development logics.
* Proven track record of designing and implementing Alteryx workflows to expedite ETL tasks, resulting in improved operational efficiency and data reliability within complex data environments.
* Strong experience with ETL and/or orchestration tools such as Talend, Oozie, and Airflow.
* Experienced in Agile methodologies including extreme programming, SCRUM, and Test-Driven Development (TDD).
* Proficient in using Informatica Power Center for ETL processes, extracting, transforming, and loading data from heterogeneous source systems into target databases.
* Experience in SQL Server Database, developing T-SQL queries, and stored procedures.
* Expertise in enhancing database performance through the creation of indexes, triggers, and constraints.
* Experience with utilizing source code control systems such as TFS (Team Foundation Server) and GIT.

Experience in working with project management and bug-tracking tools like JIRA.

**TECHNICAL SKILLS:**

❖ **Programming Languages and Frameworks:** Scala, PySpark, PL/SQL, Python, Bash scripting

❖ **Big Data Technologies**: Hadoop, HDFS, MapReduce, YARN, Spark, Hive, HBase, Kafka, Impala, Hue

❖ **Database Systems:** SQL Server 11.0, MySQL

❖ **SQL and Database Tools:** PL/SQL, Oracle SQL Developer, Oracle Enterprise Manager

❖ **ETL and Data Integration:** Sqoop, Alteryx, Talend, Informatica, Fivetran

❖ **Cloud Computing:** Azure, AWS

❖ **Cloud Data Platforms:** Azure Data Factory, Databricks, Snowflake

❖ **Workflow and Orchestration**: Airflow

❖ **Data Visualization:** looker, Tableau, power BI, Unix and Shell Scripting

❖ **Other Tools and Technologies**: Elastic search, Erwin (Data Modeling Tool), Scheduling Tool, Shell Scripting, Docker, Jenkins, Terraform, Maven, Kubernetes, NiFi, JSON, Kibana, Grafana

❖ **NoSQL Databases:** Cassandra, MongoDB

**Education:**

**Degree:** Bachelor’s in Computer Science, JNTUK, India, 2013

**Professional Experience**

**JP Morgan Chase & Co -Columbus, OH**.

**Oct 2021 - Till Date**

**Sr. Data Engineer**

**Responsibilities:**

* Migrate legacy ETLs to Spark-based processing framework.
* Improve performance and scalability of ETL processes.
* Reduce costs associated with ETL processing.
* Developed and implemented ETL pipelines using AWS Glue on S3 parquet files in a data lake.
* Leveraged Snowflake to evaluate design considerations, build logical and physical data models, and optimize views for improved performance.
* Authored Python (PySpark) scripts in Databricks for custom UDFs, including row/column manipulations, merges, aggregations, and data cleaning.
* Utilized AWS services like EC2 and S3 for processing and storage of small data sets, maintaining Hadoop cluster on AWS EMR.
* Developed and executed ETL pipelines using Informatica, Spark, and Hadoop to migrate legacy data to a modern data warehouse.
* Developed complex data mappings and transformations using PL/SQL and Python.
* Experience with Unix shell scripting, Tableau, Ab Initio, and Control-M.
* Experience in issue analysis and resolution, including usage of issue resolution processes for application production problems.
* Led the modernization initiative by designing and implementing a scalable and robust data processing architecture using Java and Spark.
* Collaborated closely with stakeholders to understand business requirements and translate them into technical designs and data mappings.
* Involved in Exploring, Designing, Building, and Deployment of the high-level and rich dashboards for decision making of critical metrics using Tableau.
* Worked with team to design and develop Alteryx workflows and Tableau dashboard.
* Translate business logic from various systems, such as Microsoft Excel, on-premise SQL servers, and cloud (Microsoft Azure) SQL Database into Alteryx in a quick and accurate manner.
* Developed complex technical designs that included data mappings, ensuring efficient and accurate data transformations.
* Leveraged cloud technologies, primarily AWS, to establish a scalable and cost-effective infrastructure for data storage and processing.
* Implemented data pipelines using Spark, enabling real-time data ingestion, processing, and analytics for immediate insights and decision-making.
* Conducted issue analysis and resolution, utilizing issue resolution processes to address application production problems promptly.
* Supported the migration of the data center, ensuring smooth data transfer and minimal downtime during the transition.
* Provided mentorship and guidance to junior team members, fostering knowledge sharing and skill development within the team.

**Environment:** Databricks, Java, Apache Spark, PL/SQL, Python, Hadoop, HDFS, Cloudera, AWS, lambda, EMR, Glue, S3, Snowflake, Hadoop, Spark, AWS, Kafka, Unix, Python, Java, Scala, Teradata, Oracle, Alteryx, Tableau, and Unix shell scripting

**Credit Suisse -Morrisville, NC**

**Feb 2019 - Sep 2021**

**Sr Data Engineer**

**Responsibilities:**

* Developed and implemented ETL pipelines using AWS Glue on S3 parquet files in a data lake.
* Tasked with converting Alteryx workflow to PL/SQL in GCP BigQuery called score integration internally.
* Leveraged Snowflake to evaluate design considerations, build logical and physical data models, and optimize views for improved performance.
* Authored Python (PySpark) scripts in Databricks for custom UDFs, including row/column manipulations, merges, aggregations, and data cleaning.
* Utilized AWS services like EC2 and S3 for processing and storage of small data sets, maintaining Hadoop cluster on AWS EMR.
* Proved expertise in Oracle database Oracle PL/SQL, SQL tuning and partitioning to help the supporting team.
* Transformed and moved large data sets using AWS EMR, integrating with Amazon S3, Amazon DynamoDB, and other data stores.
* Developed Spark applications with PySpark and Spark-SQL for data extraction, transformation, and aggregation from various file formats.
* Provided BI Analysis using various BI tools (such as Alteryx, Tableau, etc.) to review key metrics
* Created report schedules, data connections.
* Designed and executed data modeling for Snowflake, incorporating dimensional and relational schemas (Star and Snowflake).
* Collaborated with stakeholders and conducted sessions to finalize user requirement documentation.
* Scheduled and managed job workflows using Airflow, Python, and Lambda, ensuring task dependencies and smooth execution.
* Conducted regression testing using Python scripts to validate ETL processes across multiple databases, including AWS and SQL Server.
* Employed Unix shell scripting for automating data pipeline workflows, enhancing operational efficiency.
* Utilized relational databases such as Teradata and Oracle to manage and store structured data, ensuring data integrity and availability.
* Collaborated with BI Reporting teams, integrating Tableau and other reporting tools to enable comprehensive data visualization and reporting capabilities.
* Worked in an Agile setting, participating in sprint planning, daily stand-ups, and retrospectives to ensure timely project delivery and alignment with business objectives.
* Leveraged ETL tools and implemented process improvements, utilizing scheduling tools and Python for automation.
* Supported production environments, maintaining mappings, sessions, and workflows for data loading into the data warehouse.

**Environment**: AWS, EC2, S3, ETL, Java, PySpark, PL/SQL, Snowflake, Databricks, Airflow, Kafka, Spark, Lambda, Hadoop, Hadoop, Spark, AWS, Kafka, Unix, Python, Java, Scala, Teradata, Alteryx, Oracle, Tableau, and Unix shell scripting

**Group Health Cooperative of - South Central WI , Atlanta, GA.**

**Mar 2018 - Jan 2019**

**Data Engineer**

**Responsibilities:**

* Designed the business requirement collection approach based on the project scope and SDLC methodology.
* Developed JSON scripts for deploying the pipeline in Azure Data Factory and computing large volumes of data.
* Created, scheduled, and monitored Azure Data Factory pipelines and Spark jobs on Azure SQL.
* Designed and implemented database solutions in Azure SQL Data Warehouse and Azure SQL.
* Authored Python (PySpark) scripts for custom UDFs, performing row/column manipulations, merges, aggregations, and cleaning tasks.
* Developed Tableau data visualization using Cross tabs, Heat Maps, Box and Whisker charts, Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
* Evaluated Snowflake design considerations and built logical and physical data models accordingly.
* Redesigned views in Snowflake to enhance performance and conducted data unit testing.
* Developed pipelines in Azure Data Factory to extract, transform, and load data from various sources.
* Designed and developed ETL workflows and datasets in Alteryx.
* Used Alteryx for Data Preparation and then Tableau for Visualization and Reporting.
* Participated in Projects from requirements gathering through design, development using Alteryx and deployment in Tableau and Alteryx Server and publish online.
* Used excel sheet, flat files, CSV files to generated Tableau adhoc report
* Built a learner data model for real-time data ingestion from Kafka to Cassandra.
* Developed Kafka consumer API in Scala for consuming data from Kafka topics.
* Implemented real-time processing and core jobs using Spark Streaming with Kafka.
* Employed developmental toolsets such as Terraform to automate infrastructure provisioning and configuration management, ensuring reliable and reproducible deployments.
* Leveraged ETL tools like Informatica and Python for process improvements.
* Used Spark Streaming with Python to receive real-time data from Kafka and store it in HDFS and NoSQL databases like HBase and Cassandra.
* Utilized Spark for interactive queries, processing of streaming data, and integration with popular NoSQL databases.

**Environment**: Azure, Hadoop, MapReduce, Cloudera, ETL, Azure Data Factory, Spark, terraform, Java, Snowflake, Alteryx, Kafka, HBase, Hive, Informatica, Python, PL/SQL, Oracle, UNIX, Shell Scripting, Lambda, Tableau

**Maisa Solutions Private Limited -Hyderabad, India.**

**Jul 2016 - Sep 2017**

**Hadoop & Spark Developer**

**Responsibilities:**

* Worked on a data science team uncovering web behavior and performing analytics.
* Utilized Sqoop, Kafka, Flume, and Hadoop File System API's for implementing data ingestion pipelines.
* Created data pipelines for ingestion, aggregation, and loading consumer response data into Hive external tables in AWS S3.
* Implemented real-time streaming and performed transformations using Kafka and Spark Streaming.
* Created Hive tables for data loading and analysis, including partitions and buckets.
* Implemented schema extraction for Parquet and Avro file formats in Hive.
* Used Spark API over Cloudera Hadoop YARN for analytics on Hive data.
* Handled large datasets using Spark's in-memory capabilities, efficient joins, and transformations during ingestion.
* Employed developmental toolsets such as Terraform to automate infrastructure provisioning and configuration management, ensuring reliable and reproducible deployments.
* Loaded data into Spark Data Frames, performed advanced procedures like text analytics using Spark's in-memory computation.
* Integrated with Cassandra using DataStax Spark-Cassandra connector, leveraging CQL for data analysis.

**Environment**: Hadoop YARN, Scala, Python, Spark Core, Spark SQL, Spark Streaming, Spark MLlib, Kafka, Hive, Sqoop, Elastic Search, Impala, Cassandra, Zookeeper, Talend, Cloudera, MySQL, Linux, Tableau

**JAF IT Solutions-Hyderabad, IND.**

**Jan 2013 - Jun 2016**

**Hadoop Developer**

**Responsibilities:**

* Utilized Sqoop, Kafka, and Hadoop File System APIs for implementing data ingestion pipelines.
* Worked on real-time streaming, performing transformations on the data using Kafka and Spark Streaming.
* Collaborated with business users and analysts to understand ETL and reporting requirements.
* Worked with various data formats and sources, including relational tables, flat files, and CSV, and created mappings.
* Developed Hive tables, loaded data, and wrote queries for data processing.
* Implemented partitions and bucketing on Hive tables, optimizing performance, and developing UDFs as per business use-cases.
* Automated data loading into HDFS and Hive for pre-processing using Oozie.
* Collaborated in data modeling, data mining, machine learning methodologies, and ETL optimization.
* Experienced in working with Cloudera Data Platform and Cloudera Manager for Hadoop administration.
* Analyzed application and cluster logs for Hadoop production support tasks.

**Environment**: Hadoop, HDFS, MapReduce, YARN, Spark, Hive, HBase, Kafka, Scala, Impala, Hue, Oracle 10g, PL/SQL, Oracle SQL Developer, Oracle Enterprise Manager, SQL Server 11.0