MANIDEEP CHARUGUNDLA
DATA ENGINEER

Location: FL | Email: charugundlamanideep42@gmail.com | **Phone:** (512) 865-1298

SUMMARY

- Around **5 years** of experience as Data Engineer working with **Spark, AWS, Python, R programming, Scala, Hive, HBase, Pig, Sqoop** and **Java for complex business problems**, which involves real time effective analysis and processing of semi-structured and unstructured data.
- Have **3+ years** of industrial experience in Big Data analytics, Data manipulation using Hadoop Eco system tools Map-Reduce, **HDFS**, Yarn/MRv2, Pig, **Hive**, HBase, **Spark**, **Kafka**, **Flume**, Sqoop, Oozie, Avro, AWS, Spark integration with Cassandra, Zookeeper.
- Demonstrated strong proficiency in implementing and optimizing **SPARK** core components, including **SPARK SQL**, **MLlib**, **GraphX**, and Spark Streaming, for efficient data analysis and machine learning tasks.
- Applied advanced **data analytics** techniques to manipulate and analyze large datasets, enabling real-time insights and decision-making for complex business challenges.
- Sound knowledge in developing highly scalable and resilient **Restful APIs, ETL** solutions, and third-party integrations as part of Enterprise Site platform using **Informatica**.
- Utilized **Amazon Elastic Cloud Compute (EC2)** infrastructure for computational tasks and **Simple Storage Service (S3)** for scalable and reliable data storage, ensuring high availability and durability of data.
- Managed and optimized databases, including **Oracle**, **MySQL**, **Teradata**, and **Netezza**, writing complex SQL queries and developing custom User Defined Functions **(UDFs)** for enhancing data processing capabilities in Hive and Pig.
- Proficient in working with **NoSQL** databases like **HBase**, **Cassandra**, and **MongoDB**, designing schemas and optimizing data retrieval for efficient data processing.
- Designed and implemented end-to-end data pipelines to extract, transform, and load **(ETL)** large volumes of data, ensuring data quality and reliability throughout the process.
- Involved in migration of the legacy applications to cloud platform using DevOps tools like **GitHub**, **Jenkins, JIRA, Docker**, and **Slack**.
- Experienced in leveraging various AWS services such as **EMR**, **Redshift**, **S3**, **Athena**, and **Glue** for building scalable and cost-effective big data solutions in the cloud environment.

TECHNICAL SKILLS:

Programming Languages:	Shell scripting, SQL, PL/SQL, Python, R, Pig, Hive QL, Scala
Big Data Ecosystem:	HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala
Hadoop Distribution:	Cloudera CDH, Horton Works HDP, Apache, AWS
DevOps Tools:	Jenkins, Git, and Maven
Database:	Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB)
Cloud Technologies:	MS Azure, Amazon Web Services (AWS), GCP
Web Development:	HTML, DHTML, XHTML, Java Script
Operating Systems:	Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS
Others:	Machine learning, NLP, Stream Sets, Spring Boot, Jupyter Notebook, Docker, Kubernetes, Jenkins, Jira

WORK EXPERIENCE:

Client: CITIBANK, TAMPA, FL **May 2022 - Present**
Data Engineer

- Created and maintained a data pipeline to ingest and process customer behavioral data and financial histories into a **Hadoop** cluster for analysis, ensuring data quality and reliability throughout the process.
- Collaborated in an **Agile** team to deliver and support business objectives. Utilized **Java**, **Python**, and shell scripting, along with other related technologies, to acquire, ingest, transform, and publish data in the Hadoop Ecosystem.
- Leveraged **AWS Redshift**, **S3**, **Spectrum**, and Athena services to query large datasets stored on S3, creating a Virtual Data Lake without the need for traditional **ETL** processes, improving data accessibility and agility.
- Evaluated and compared different tools for test data management with Hadoop in **Azure**, identifying the most suitable tools for efficient data management and testing processes.
- Experience in building and architecting multiple Data pipelines, **end to end ETL** and ELT process for **Data ingestion** and transformation in **GCP** and coordinate task among the team.
- Worked on **POC** to check various cloud offerings including Google Cloud Platform (GCP).
- Designed sequence diagrams for Hadoop data flow, ensuring clear documentation of data processing workflows.
- Imported/exported data between **HDFS** and relational systems (**Oracle, MySQL, DB2**) using **Sqoop** for seamless data integration.
- Implemented data cleansing, event enrichment, data aggregation, denormalization, and data preparation processes for downstream model learning and reporting, enhancing data quality and usability.

- Assisted Application and Operations teams in **troubleshooting** performance issues, ensuring smooth operation of Hadoop applications and jobs.
- Implemented Partitioning, Dynamic Partitions, and bucketing in **Hive** for efficient data access and query performance, improving overall system performance.
- Developed **Spark** scripts using **Python** on **AWS EMR** for **Data Aggregation**, Validation, and Adhoc querying, leveraging the scalability and performance benefits of Spark for large-scale data processing.
- Worked with **EMR** cluster in the **AWS** cloud and managed data storage and processing using S3, ensuring data availability and durability. Explored the usage of Spark for improving the performance and optimization of existing algorithms in Hadoop, using Spark Context, **Spark SQL**, and Spark Yarn, enhancing algorithm efficiency and speed.
- Worked on the Spark **SQL** and Spark Streaming modules of Spark, using **Scala** and Python to write code for various Spark use cases, enabling real-time data processing and analysis.
- Migrated historical data to **S3** and developed a reliable mechanism for processing incremental updates, ensuring data consistency and integrity.
- Involved in building own **PaaS** with **Docker**, deployed various applications through Dockers containers
- Used **Oozie** workflow engine to manage independent Hadoop jobs and automate various types of Hadoop jobs, such as Java **MapReduce**, **Hive**, and **Sqoop**, improving operational efficiency and job management.
- Monitored and debugged Hadoop jobs/applications running in production, ensuring timely resolution of issues and smooth operation of data processing workflows.

Client: Tolaram Group, Lagos, Nigeria **May 2018-Nov 2020**
Data Engineer

- Provided user support and application support on Hadoop infrastructure within **Azure** cloud environment, ensuring smooth operation and resolving issues in a timely manner.
- Automated the creation and termination of **Azure** HDInsight clusters, streamlining the process and ensuring resource optimization.
- Developed batch scripts to fetch data from Azure Blob Storage and perform required transformations in Scala using the Spark framework, improving data processing efficiency and accuracy.
- Collected all logs from source systems into HDFS using **Kafka** and performed analytics on them within the Azure ecosystem, extracting valuable insights and monitoring system performance.
- Ingested streaming data with Kafka in Azure, ensuring real-time data processing and analysis capabilities.
- Worked on fine-tuning **Spark** applications in Azure Databricks to improve overall processing time for pipelines, optimizing resource utilization and enhancing performance.
- Designed and documented operational problems by following standards and procedures using **JIRA**, ensuring clear communication and effective problem resolution within the Azure environment.
- Developed dynamic content of the presentation layer using JSP and Servlets, ensuring a user-friendly and interactive interface for data visualization and analysis within Azure web applications.
- Implemented Partitioning, Dynamic Partitions, and bucketing in **HIVE** for efficient data access in Azure Synapse Analytics, improving query performance and overall system efficiency.
- Experience in working with **EC2** Container Service plugin in **JENKINS** which automates the **Jenkins** master- slave configuration by creating temporary slaves.
- Implemented data ingestion and processing solutions using Big Data technologies in Azure, handling over 2 million data records monthly with technologies like Hadoop, MapReduce, HBase, Hive, and Sqoop.
- Transformed over 100 complex Hive/SQL queries into Spark transformations monthly using Spark RDD, Scala, and Python in Azure Databricks, achieving a 25% increase in data processing efficiency.
- Enhanced data infrastructure and pipelines within Azure, focusing on optimizing data collection, storage, and processing for datasets using Azure Data Factory, Azure Data Lake Storage, and Azure Blob Storage.
- Managed and processed data from Azure Data Lake Storage into Spark Data Frames monthly in Azure Databricks, performing vital data transformations and actions.
- Demonstrated proficiency in Microsoft Azure Services, managing data across services like Azure VM, Azure Blob Storage, Azure SQL Database, Azure Synapse Analytics, Azure Data Factory, **IAM**, and **Power BI**.
- Authored **PowerShell** scripts to monitor and manage Azure cloud resources, opt mizing infrastructure handling for substantial data volumes.

EDUCATIONAL

Masters in Data Engineering
University of North Texas Denton, TX, May 2022

Bachelors of Technology in Mechanical Engineering
Vellore Institute of Technology Vellore, TN, Apr 2018