K. CHARAN

 ☏ +1 (469)-277-8789 | www.LinkedIn.com | ✉ Charank4190@gmail.com

Data Engineer with 5 years of experience specializing in building, optimizing, and maintaining scalable data pipelines and ETL processes. Proven expertise in managing large-scale data infrastructures, integrating diverse data sources, and enabling real-time analytics. Proficient in Python, SQL, Apache Spark, and cloud platforms like AWS, Azure and Google Cloud. Adept at collaborating with cross-functional teams to design data models, improve data quality, and drive business intelligence solutions.

**Sr. Data Engineer** Jul 2023 – Present

# [EPSILON, USA](https://www.epsilon.com/us)

* Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks, and analysis. Developed a data pipeline using Kafka and Storm to store data in HDFS.
* Analyzed data using Python, SQL, Hive, PySpark, and Spark SQL for Data Mining, and Data Cleansing.
* Implemented and managed ETL solutions and automated operational processes.
* Optimized ETL processes using Python, SQL, and Airflow, reducing data processing time by 30% and improving the reliability of financial data ingestion.
* Worked on Databricks to write Scripts in PySpark, Python, and SQL and integrate Databricks with AWS.
* Orchestrated and automated end-to-end data ingestion pipelines using Azure Data Factory (ADF), successfully integrating data from diverse sources (SQL Server, flat files, APIs, etc.) into Azure Synapse and Azure Data Lake Storage.
* Implemented data governance and security measures using AWS IAM, KMS, and Redshift Spectrum to ensure secure access and encryption of sensitive financial data.
* Worked on AWS Data Pipeline to configure data loads from S3 to Redshift and have used AWS components Downloading and uploading data files (with ETL) to AWS system using S3 components and Used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket.
* Designed and executed on-prem to AWS cloud migration projects for state agencies.
* Leveraged AWS services (EC2, ECS, Lambda, Glue) for managing data pipelines and handling large-scale data ingestion, transformation, and storage across distributed systems.
* Developed a deep understanding of company intellectual property to proactively identify and mitigate potential threats.
* Conducted risk assessments to safeguard sensitive corporate information and supported the development of risk mitigation strategies.
* Advanced proficiency in SQL and Excel for financial data manipulation, model validation, and performance tracking.
* Automated data validation and error handling processes, reducing the manual intervention required for pipeline failures by 40%.
* Delivered complex findings in a clear and accessible manner through written reports, visual summaries, and engaging presentations tailored for diverse stakeholders.

**Sr. Data Analyst** May 2018 – Nov 2021

# [PER SFT,](https://persft.com/about.html) India

* Analyzed data using Python, SQL, Hive, PySpark, and Spark SQL for Data Mining, and Data Cleansing.
* Implemented and managed ETL solutions and automated operational processes.
* Applied object-oriented design patterns in Python to develop modular, reusable code components.

Implemented unit tests to maintain code quality and reliability.

* Used Pandas, NumPy, Seaborne, SciPy, Matplotlib, Scikit-learn, and NLTK in Python for developing various machine learning algorithms. Implemented using Machine Learning techniques with Python Pandas and Scikit-learn packages
* Assisted with tasks such as data pipeline engineering, data analytics, data scrapping and mining, data visualization, data dashboarding, machine learning (ML), and Cloud Computing.
* Built Data Models and Dimensional Modelling with Star and Snowflake schemas for OLAP and Operational data store (ODS) applications.
* Responsible for performing Machine-learning techniques regression/classification to predict the outcomes.
* Involved in data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling and data mining, machine learning, and advanced data processing.
* Extensively chipped on Spark Context, Spark-SQL, RDD's Transformation, Actions, and Data Frames.
* Built and maintained ETL pipelines using PySpark to process and analyze large datasets.
* Developed high-performance data ingestion pipelines from multiple sources using Azure Data Factory and Azure Databricks.
* Integrated Scala with Spark SQL and Data Frames to process and analyse large datasets, contributing to the development of real-time analytics solutions that provided actionable insights for business operations.
* Implemented medium to large-scale BI solutions on Azure using Azure Data Platform services.

## Technical Skills:

**Programming/Scripting:** Python, Scala, R, Shell Scripting, SQL, PySpark.

**Big Data Ecosystem:** Hadoop, MapReduce, Pig, Hive, HBase, YARN, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, Spark, Ambari, Elastic Search, Parquet, Snappy, Airflow, NiFi.

**Hadoop Distributions:** Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapReduce, Apache EMR.

**Cloud Platforms**: Amazon Web Services (AWS), MS Azure, Google Cloud Platform (GCP).

**Cloud Platforms**: Azure (ADLF, ADL), AWS ((EC2, ECS, S3, ELB, Auto Scaling, KMS, Elastic Beanstalk, CloudFront, CloudFormation, Elastic File System, RDS, DMS, VPC, Direct Connect, Route 53, CloudTrail, IAM, SNS, SQS, Lambda). **Machine Learning Algorithms**: Linear Regression, Logistic Regression, Linear Discrimination Analysis (LDA), Decision Trees, Random Forests with Adaboost and Gradient Descent Boosting.

**CI/ CD**: Azure DevOps, Jenkins.

**Ticketing Tools**: JIRA.

**Operating Systems:** Windows, Linux, Unix.

**Databases (RDBMS/ NoSQL)**: Oracle, SQL Server, Cassandra, Teradata, PostgreSQL, HBase, MongoDB.

**ETL Tools**: Informatica.

**Python Libraries**: Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, SciPy.

**DWH Schemas**: Star Schema, Snowflake Schema.

**Data Modelling Tools:** Erwin, MS VISO

**Web/ Application Server:** Apache Tomcat, WebLogic, WebSphere

**Version Control**: Git, GitHub, Bitbucket, Subversion

**Reporting Tools**: PowerBI, Tableau, R, SAS Visual Analytics, SQL Server Reporting Services (SSRS)

## Certification

[Certified Business Analysis professional (CBAP)](https://www.linkedin.com/learning/certificates/47e23115895b40f797e235181bea7c87ec540e0106e66ac2cc154bca2d26b72f?lipi=urn%3Ali%3Apage%3Ad_flagship3_profile_view_base%3B0LPJIuStS4KutVL9JdWXrw%3D%3D) [Google Project Management Certification.](https://coursera.org/verify/AA32YYL54MBZ)

**Achievements**: Recognized for leading a critical data integration project that improved data accessibility for analytics purposes by 50%, enhancing data-driven decision-making across the business

## Education

University of North Texas, Denton, USA January 2022 – May 2023 Master of Science in Business Analytics, CGPA 3.66