**Sheshank J**
Data Engineer
Sheshankj6@gmail.com +1 7793792186


## SUMMARY

Around four years of experience as a Data Engineer in different Technology, performing Statistical Modelling, Data cleaning, Data Exploration, and Data Visualization of structured and unstructured datasets as well as implementing large-scale Machine Learning algorithms to deliver resourceful insights, and inferences. Proficient in leveraging a diverse set of technologies and tools to build scalable and efficient data pipelines, enabling data-driven decision-making and business insights. Skilled in managing and processing large volumes of structured and unstructured data, with a strong focus on data quality, reliability, and performance. Demonstrated expertise in cloud platforms such as Azure, AWS, GCP and Snowflake, as well as proficiency in technologies like Python (NumPy, Pandas, Matplotlib, SciPy, Scikit-learn, Seaborn), SQL, Apache Airflow, Bigquery, Azure Synapse and Scala. Proven ability to collaborate effectively with cross-functional teams to deliver impactful data solutions and drive organizational success.

## SKILLS

**Programming Languages:** Python, Scala, SQL, C, C++, Java, Shell Scripting.
**Big Data Stack:** PySpark, Spark, Hadoop, Sqoop, Pig, HDFS, Hive, Kafka, Spark SQL, Apache Sqoop
**Databases:** MySQL, SQL Server, PostgreSQL, Snowflake, MongoDB, HiveQL, Spark SQL, HBase.
**Orchestration:** Airflow.
**Dev/Ops:** Docker, Git, Jenkins.
**Cloud:**

- **AWS**: Lambada, AWS Glue, Athena, Redshift, EMR, EC2, S3.
- **Azure**: Azure Data Factory, Databricks, Blob storage, Azure Data Lake
- **GCP**: Composer, Big Query, Cloud Storage (GCS), Spanner, Google Kubernetes Engine.

**IDE:** DataBricks, Visual Studio Code, PyCharm, Anaconda, IntelliJ.
**Visualization:** Tableau, Power BI.
**Project Management:** Agile, Waterfall

## PROFESSIONAL EXPERIENCE

**CVS Health, CT** **July 2023 – Current**
**Data Engineer**

- Interacted and gathered requirements for Designing and developing common architecture for storing Claims data within Enterprise and building Data Lake in Azure cloud.
- Developed Spark applications for data extraction, transformation and aggregation from multiple systems and stored on Azure Data Lake Storage using Azure Databricks notebooks.
- Developed applications using PySpark to integrate data coming from other sources like ftp, csv files processed using Azure Databricks and written into Snowflake.
- Written Unzip and decode functions using Spark with Scala and parsing the xml files into Azure blog storage.
- Designed and implemented end-to-end data pipelines using Azure Data Factory to orchestrate data workflows across multiple systems.
- Developed and managed ETL processes in Azure Data Factory to extract, transform, and load data from diverse sources into Snowflake.
- Utilized DBT (Data Build Tool) for transforming raw data into curated data models in Snowflake, ensuring data quality and consistency.
- Developed PySpark scripts from source system like Azure Event Hub to ingest data in reload, append, and merge mode into Delta tables in Databricks.
- Design and Develop ETL Processes in Databricks to migrate Campaign data from external sources like Azure Data Lake, and gen2 in ORC/Parquet/Text Files.

- Implemented data ingestion solutions using Azure Data Factory to efficiently collect and ingest data from various sources, including relational databases, files, APIs, and streaming data sources.
- Optimized PySpark applications on Databricks, which yielded a significant amount of cost reduction.
- Created Pipelines in Azure Data Factory to copy parquet files from ADLS Gen2 location to Azure Synapse Analytics Data Warehouse.
- Worked on replacing existing Hive scripts with Spark Data-Frame transformation and actions for faster analysis of the data.
- Developed PySpark scripts to Reduce costs of organization by 30% by migrating customers data in SQL Server to Hadoop.
- Experience in handling JSON datasets and writing custom Python functions to parse through JSON data using Spark.
- Used Spark for interactive queries, processing of streaming data, and integration with popular NoSQL databases for huge volumes of data.
- Responsible for loading Data pipelines from web servers using Kafka and Spark Streaming API
- Designed and implemented end-to-end data pipelines using Azure Data Factory to orchestrate data movement and transformation processes across on-premises and cloud data sources.
- Generate weekly based reports and ops reports, customer goals reports, mobile scan and pay goals and usage in data by using power BI.

**Vanguard, PA** **Aug 2022 – Jun 2023**
**Data Engineer**

- Using an Agile software Jira to report progress on software projects.
- Designed, implemented, and maintained data pipelines using AWS services such as Amazon S3, Glue, and Redshift, integrating data from various sources into a unified data lake.
- Worked as a Data Engineer with Big Data and Hadoop ecosystem components.
- Involved in converting Hive/SQL queries into Spark transformations using Scala and Python.
- Created Spark data frames using Spark SQL and prepared data for data analytics by storing it in AWS S3.
- Used Spark Data frame and Spark API to implement batch processing of Jobs.
- Improving Efficiency by modifying existing Data pipelines on AWS Glue to load the data into AWS Redshift.
- Used Airflow for orchestration and scheduling of the ingestion scripts.
- Identify source systems, their connectivity, related tables, and fields and ensure data suitability for mapping, preparing unit test cases, and providing support to the testing team to fix defects.
- Worked in real-time data streaming data using AWS Kinesis, EMR, and AWS Glue.
- Streamlined financial reporting processes with SQL queries in a relational database, leading to a 27% decrease in report generation time and improved data accuracy for a hedge fund.
- Applied complex transformations using SQL and Python within Databricks, aggregating and cleansing data to meet specific business requirements and enhance data quality.
- Developed Extract, Transform, Load (ETL) processes using AWS Glue, PySpark, or other AWS-native tools to prepare and transform data for analytics and reporting.
- Enhanced financial modeling proficiency in Microsoft Excel and Python to create accurate financial forecasts, resulting in a 12% reduction in budget variances for a FinTech startup.
- Developed interactive and visually appealing Tableau dashboards to present data insights and analysis to stakeholders, improving data visualization and decision-making processes.

**Accion Labs India PVT LTD, India** **Jan 2019 - Dec 2019**
**Data Engineer**

- Developed business process models in Agile to document existing and future business processes.
- Designed, developed, and maintained ETL workflows using SQL Server Integration Services (SSIS) to extract, transform, and load data from various sources into target databases and data warehouses.
- Managed and administered SQL Server databases using SQL Server Management Studio (SSMS), ensuring data integrity, availability, and optimal performance.
- Designed and implemented database schemas, tables, views, and indexes to support efficient data storage and retrieval.
- Published Power BI Reports in the required originations and Made Power BI Dashboards available in Web clients and mobile apps.

- Developed the necessary Stored Procedures and created Complex Views using Joins for robust and fast retrieval of data in SQL Server.

---

**EDUCATION**

**Bachelor of Technology in Computer Science** **GPA 3.53**
**University of Central Missouri, 2018 - 2022**
**Academic honor:** Spring semester (Dean's list)