**Professional Summary:**

* Total 9+ years of experience in IT industry working on multiple domain and technologies.
* 4+ yrs of experience in Big Data Engineering in AWS & Azure Cloud technologies.
* Having experience in designing and implementing complex data solutions using AWS services like EC2, EMR, S3, Airflow and Azure Technologies like Databricks, DataFactory, DataLake etc.
* Experience in Developing ETL(Extract, Transform, Load) solutions using SQL and PySpark in EMR, Azure Databricks for data extraction, transformation and aggregation.
* Experienced in optimizing Spark DataFrame and RDD performance by tuning various configuration settings, such as memory allocation, caching, and serialization.
* Expertise in using Spark RDD and DataFrame transformations (filtering, mapping, reducing,joins, aggregating) to handle both structured and unstructured data and along with Optimization techniques to improve processing time and data accuracy.
* Strong understanding of integrating Spark RDD with other big data technologies like Hadoop, Hive optimizing data processing workflows and enhancing performance across distributed systems.
* Worked on Delta tables and its features like Timetravelling, Versioning, Schema Evolution, handling streaming data, Autoloader features etc.
* Have good knowledge on Spark architecture and data processing.
* Proficient in designing efficient data models with Hive tables and partitions, optimizing query performance and resource utilization.

**Technical Skills:**

**LANGUAGES:** PySpark, SQL, Python

**CLOUD PLATFORM:** Azure and AWS.

**TOOLS:** EMR, Airflow, Azure Databricks, ADF, Azure Data Lake & Key Vault.

**STORAGE:** S3,DataLake

**FRAMEWORK:** Hadoop and Spark.

**Education:**

* B.TECH (Electronics & Telecommunication) from BAMU, Aurangabad (Maharashtra).

**Professional Experience:**

**CAPGEMINI May 2025 – Present**

**DIENSTEN TECH LIMITED Mar 2025 - May 2025**

**FLIPKART PVT LTD Dec 2023 - Mar 2025**

**STERLITE TECH Apr 2017 - Dec 2023**

**Role: BIG DATA ENGINEER**

**Responsibilities:**

* data processing, real-time analytics, and batch jobs.
* Managed Spark clusters, optimized resource utilization through partitioning, caching, and tuning configurations, and monitored jobs using YARN or Mesos.
* Familiarity with Hive performance tuning tools, such as Hive Query Profiler, Hive Query Plan Visualization, and Hive Load Testing Tools, and their features and limitations.
* Proficient in optimizing Hive query performance by tuning various configuration settings, such as memory allocation, parallelism, and compression.
* Experienced in identifying and resolving performance bottlenecks in Pipelines, such as data skew, inefficient joins, and excessive shuffling.
* Used AWS Step Functions to manage and monitor PySpark job execution workflows.
* Debugged and fixed performance issues in PySpark jobs on AWS EMR.
* Configured AWS EMR clusters for running optimized PySpark applications.
* Built end-to-end PySpark pipelines on AWS EMR, reading data from AWS S3.
* Worked on ingesting data from Azure Data lake into Databricks Delta Lake using Workflows created through PySpark Notebooks and scheduled to run daily.
* Transformed ingested data to create insightful tables/views with advanced logic such as joins and aggregations.Having experience on performing complex transformation logic like aggregations, joins etc to process the data.
* Manage Lakehouse storage for Delta tables across both raw and transformed dataset the customers using Azure Databricks as an integration between multiple platforms.
* Contributed to designing the overall architecture and implemented data transfer pipelines using PySpark. Orchestrated workflow execution with Azure Data Factory for streamlined automation.

**Role: Talend ETL Developer**

**Responsibilities:**

* Proficient in using ETL testing tools and frameworks, such as QuerySurge, Talend Data Quality, or Informatica Data Validation Option.
* Skilled in validating and verifying data transformation rules and business logic applied during ETL processes.
* Strong understanding of data governance and regulatory compliance requirements in ETL testing.
* Proficient in testing data integration and synchronization between different systems using ETL processes.
* Expertise in testing ETL transformations, such as data aggregation, filtering, sorting, and joining.
* Knowledgeable about metadata management and testing metadata-driven ETL processes.
* Experienced in identifying and reporting data quality issues and data anomalies during ETL testing.
* Strong understanding of data warehouse concepts and testing data loading into data warehouse systems.
* Skilled in performance testing of ETL processes to ensure scalability and efficiency.
* Knowledgeable about testing data extraction and loading from different database systems, such as Oracle, SQL Server, or MySQL.