Nikhil Sai Srinivas Pokuri
Phone: (+91)-7032565184
saisrinivaspokuri@gmail.com

PROFESSIONAL SUMMARY

- Around 2+ years of experience as a Big Data Developer with expertise in Hadoop Ecosystem technologies (HDFS, Hive, Sqoop, Apache Spark and AWS).
- Developed Spark applications for distributed data processing.
- Performed data cleansing and preprocessing using Spark transformations.
- Worked with Spark's data serialization formats (Avro, Parquet, JSON, etc.).
- Expertise in using Spark RDD transformations and actions to process large-scale structured and unstructured data sets, including filtering, mapping, reducing, grouping, and aggregating data.
- Strong understanding of Spark RDD integration with other big data technologies, such as Hadoop, Hive, and Kafka, and their impact on data processing workflows and performance.
- Proficient in developing and implementing Spark DataFrame-based data processing workflows using Python programming language.
- Skilled in using Spark DataFrame persistency and caching mechanisms to reduce data processing overhead and improve query performance.
- Proficient in processing serialized data in Spark using various formats, such as Avro, Parquet, ORC and their features and limitations.
- Optimized Spark jobs and data processing workflows for scalability, performance, and cost efficiency using techniques such as partitioning, compression, and caching
- Proficient in handling hive partitions and buckets with respect to the business requirement.
- Experience in handling hive schema evolution with avro file format
- Skilled in handling semi structured/serialised data processing using hive (AVRO,PAQUET,ORC)
- Experienced in efficiently using Hive managed and external table with respect to the business requirement including partitioned tables.
- Knowledge of Hive query tuning best practices, such as minimizing data transfers, avoiding unnecessary data conversions, and using appropriate data formats.
- Performed Hive integration with other big data technologies, such as Hadoop, Spark, and their impact on query performance.
- Knowledge of Hive Avro schema evolution best practices, such as versioning schema files, using schema registries for centralized schema management, and testing schema changes in a staging environment before deployment.
- Experienced in importing and exporting large datasets between Hadoop and relational databases using Sqoop.
- Adept in scheduling and automating Sqoop jobs for incremental runs.
- Proficient in using Sqoop to import and export data in various file formats such as CSV, Avro, and Parquet.
- Experienced in using Sqoop to import and export data from and to cloud-based data storage services such as Amazon S3
- Skilled in optimizing Sqoop jobs for high throughput and low latency using tuning parameters such as batch size and number of mappers.
- Developed large-scale distributed data pipelines using PySpark on AWS EMR.
- Configured Spark jobs on AWS EMR to efficiently read and write data from AWS S3.
- Used AWS Step Functions to orchestrate PySpark workflows and automate data pipelines.
- Built end-to-end PySpark pipelines on AWS EMR, reading data from AWS S3.
- Used AWS Hive to perform SQL queries on datasets processed by Spark on AWS EMR.
- Integrated AWS EC2 instances for managing and deploying AWS EMR clusters.
- Utilized AWS S3 for storing intermediate and final datasets processed by PySpark.
- Integrated PySpark job outputs with AWS S3 for downstream reporting and analytics.

TECHNICAL SKILLS

Data Eco System	:	Hadoop, Sqoop, Hive, Apache Spark
Cloud Skills	:	AWS
Databases	:	MySQL
Languages	:	Python, SQL
Operating Systems	:	Linux and Windows

PROFESSIONAL EXPERIENCE

Company Name: Gigabyte Infocomm Pvt Ltd *Jan 2024– Present*

Client: Ixigo

Responsibilities

- Worked with Spark's data serialization formats (Avro, Parquet, JSON, etc.).
- Experienced in integrating Sqoop with other Hadoop ecosystem components such as Hive, and Spark.
- Integrated Spark with data lakes such as AWS S3, HDFS, EMR, EC2.
- Designed and implemented ETL processes using Spark.
- Implemented data partitioning and shuffling strategies for optimization.
- Worked with Spark DataFrame APIs for structured data analysis.
- Designed and optimized Spark jobs for join operations.
- Maintained and monitored Spark clusters on AWS EMR, ensuring high availability and fault tolerance.
- Developed Spark applications for distributed data processing.
- Created and managed RDDs (Resilient Distributed Datasets) for data transformations.
- Utilized DataFrames for structured data manipulation and analysis.
- Designed and implemented Spark jobs using Pyspark.
- Performed data cleansing and preprocessing using Spark transformations.
- Optimized Spark jobs for performance and resource utilization.
- Implemented Spark SQL queries for data querying and aggregation.
- Created and managed Spark clusters for distributed computing.
- Implemented Spark partitioning and caching strategies.
- Conducted Spark job scheduling and orchestration.
- Monitored Spark jobs using cluster management tools like YARN.
- Conducted Spark job scheduling and orchestration.
- Developed Some Kafka Jobs to stream the data to s3 continuously

Technologies: Spark, Hive, Sqoop, Python, HDFS, Hive, AWS

Company Name: Lumen Technologies *Sep 2023– Jan 2024*

Client: Cisco

Responsibilities

- Performed data cleansing and preprocessing using Spark transformations.
- Optimized Spark jobs for performance and resource utilization.
- Worked with Spark's data serialization formats (Avro, Parquet, JSON, etc.).
- Designed and implemented ETL processes using Spark.
- Automated the sqoop jobs for the incremental data.
- Implemented data partitioning and shuffling strategies for optimization.
- Certified in Azure-900 Fundamentals

Technologies: Spark, Hive, Sqoop, Python, HDFS, Hive

Company Name: VRJ Technologies Pvt Ltd *Aug 2022– Sep 2023*

Client: Rupeek

Responsibilities

- Worked with Spark's data serialization formats (Avro, Parquet, JSON, etc.).
- Conducted Spark job scheduling and orchestration.
- Integrated Spark with data lakes such as AWS S3, HDFS, EMR, EC2.
- Designed and implemented ETL processes using Spark.
- Implemented data partitioning and shuffling strategies for optimization.
- Tuned Spark configurations for resource utilization.
- Worked with Spark DataFrame APIs for structured data analysis.
- Designed and optimized Spark jobs for join operations.
- Maintained and monitored Spark clusters on AWS EMR, ensuring high availability and fault tolerance.
- Developed Spark applications for distributed data processing.
- Created and managed RDDs (Resilient Distributed Datasets) for data transformations.
- Utilized DataFrames for structured data manipulation and analysis.
- Designed and implemented Spark jobs using Python.
- Performed data cleansing and preprocessing using Spark transformations.
- Optimized Spark jobs for performance and resource utilization.
- Implemented Spark SQL queries for data querying and aggregation.
- Created and managed Spark clusters for distributed computing.
- Implemented Spark partitioning and caching strategies.
- Monitored Spark jobs using cluster management tools like YARN .
- Orchestrated the entire pipeline with StepFunction.

Technologies: Spark, Hive, Sqoop, Python, HDFS, Hive, AWS

WORK EXPERIENCE

Gigabyte Infocomm Pvt Ltd – Jan 2024 – Present.
Lumen Technologies – Sep 2023 – Jan 2024.
VRJ Technologies Pvt Ltd – Aug 2022 – Sep 2023.

EDUCATION

Institute/College	Duration	CGPA Obtained
QIS Institute Of Technology	2019-2023	6.55
Future Focus Jr College	2017-2019	8.98
Gitanjali Em High School	2016-2017	9.8