Divya Biradar
Phone: (+1)-6823748912
divyabiradar8570@gmail.com

PROFESSIONAL SUMMARY

- 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.
- Having a total of 2 experience in big data design and development.
- Skilled in using Spark RDD persistency and caching mechanisms to reduce data processing overhead and improve query performance.
- Familiarity with Spark RDD lineage and fault tolerance mechanisms and their impact on data processing reliability and performance.
- Knowledge of Spark RDD optimization techniques, such as data partitioning, shuffle tuning, and pipelining, and their impact on query performance and resource utilization.
- 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.
- Ability to troubleshoot common issues with Spark RDD, such as data processing errors, performance bottlenecks, and scalability limitations.
- Experience working with Spark RDD in production environments and implementing performance monitoring and alerting systems to detect and resolve performance issues proactively.
- Familiarity with Spark RDD-based data processing libraries and frameworks, such as Apache Spark SQL, MLlib, and GraphX, and their features and limitations.
- Knowledge of Spark RDD best practices in data engineering and data science domains, such as data preprocessing, feature engineering, model training, and inference.
- Proficient in developing and implementing Spark DataFrame-based data processing workflows using Scala, Java, or Python programming languages.
- Experienced in optimizing Spark DataFrame performance by tuning various configuration settings, such as memory allocation, caching, and serialization.
- Expertise in using Spark DataFrame transformations and actions to process large-scale structured and semi-structured data sets, including filtering, mapping, reducing, grouping, and aggregating data.

TECHNICAL SKILLS

- **Technical Skills:** Python, HTML, C, C++, SQL, Java
- **Operating system:** Windows, Linux, IOS, Android
- **Database:** MySQL, PostgreSQL, Oracle, NOSQL.
- **IDE:** Eclipse, PyCharm, Jupyter notebook, Git
- **Cloud:** AWS, Azure, Databricks

- **Framework / Methodology:** Agile, Scrum, Jira, Apache Spark, Apache Kafka, Hadoop, Hive, Spark Scala, Confluence, Apache Airflow
- **CI/CD pipeline:** Git, Maven
- **Distribution:** Cloudera 5.12
- **Data visualization:** Tableau, Microsoft PowerBI
- **Microsoft Tools:** Excel, PowerPoint, Word, Microsoft Teams.

PROFESSIONAL EXPERIENCE

Project Name: Freshworks *August 2022 – August 2023*

Project Role: Data Engineer

Responsibilities:

- Experienced in optimizing Spark RDD performance by tuning various configuration settings, such as memory allocation, caching, and serialization.
- 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.
- Integrated AWS EC2 instances for managing and deploying AWS EMR clusters.
- Utilized AWS S3 for storing intermediate and final datasets processed by PySpark.
- Implemented fault-tolerant PySpark jobs on AWS EMR with data storage in AWS S3.
- Orchestrated multi-step data processing workflows using AWS Step Functions.
- Experienced in ETL (Extract, Transform, Load) testing methodologies and processes.
- Proficient in testing data extraction processes from various sources, including databases, files, and APIs.
- Skilled in validating and verifying data transformation rules and business logic applied during ETL processes.
- Strong understanding of data warehouse concepts and testing data loading into data warehouse systems.
- Experience in handling hive schema evolution with avro file format
- Skilled in handling semi-structured/serialized data processing using hive (AVRO, PAQUET, ORC)
- Experienced in efficiently using Hive-managed and external tables for business requirements.
- Experienced in importing and exporting large datasets between Hadoop and relational databases using Sqoop.
- Proficient in performing data validation and cleansing during data transfer using Sqoop's validation and cleansing options.

Technologies: Spark, AWS, Hive, Sqoop, Confluence, Outlook, Cloudera

Project Name: HealthifyMe *June 2021 – June 2022*

Project Role: Data Engineer intern

Responsibilities:

- Proficient in developing and implementing Spark DataFrame-based data processing workflows using Python programming languages.
- Developed Spark applications for sentiment analysis.

- Conducted performance testing and profiling of Spark applications.
- Utilized Spark for log parsing and parsing unstructured data.
- Integrated Spark with messaging systems like Kafka.
- Proficient in managing and optimizing data storage solutions using Azure Data Factory, ensuring efficient data organization, access control, and security.
- Experienced in deploying and managing data processing clusters with Azure HDInsight, leveraging its scalability and automation features for large-scale data analysis.
- Knowledgeable about testing data extraction and loading from different database systems, such as Oracle, SQL Server, or MySQL.

Technologies: PySpark, Azure, Jira, Spark SQL, Kafka, Python

PROJECT

Event Detection in Twitter Streams: Python, AWS Kinesis, AWS Lambda, Apache Airflow

- Developed a real-time system leveraging Google Cloud Platform services, including Big Query and Dataflow, to continuously monitor and analyze Twitter data. Employed advanced event detection algorithms to identify specific events or trends from tweet content in real-time.
- Implemented automated actions triggered by significant events, facilitating applications such as emergency response, real-time tracking of breaking news, and monitoring social media trends.
- Orchestrated workflows using Apache Airflow, ensuring smooth coordination between data extraction, transformation, and event detection processes.
- Enhanced system efficiency by optimizing stream processing pipelines and reducing data processing time using parallelization techniques in Google Dataflow.

Classification of the category using machine learning on BBC news dataset: NLTK, Python, sci-kit-learn, TensorFlow

- Involved in cleaning and transforming the raw data into a format suitable for machine learning algorithms. It may include removing stop words, tokenization, stemming, and vectorization.
- Load the BBC news dataset, which consists of articles from different categories like business, entertainment, politics, sport, and tech.
- Preprocess the text data by removing stop words, tokenizing the text, and converting it into numerical representations using techniques like TF-IDF or word embeddings.

PUBLICATIONS

- Leveraging big data for disease surveillance and public health interventions (IJGIS JULY 2024) HTTPS://DOI.ORG/10.21428/E90189C8.8F592E36
- beyond words: exploring emotion detection in speech using sound (INCET 2024) HTTPS://IEEEXPLORE.IEEE.ORG/DOCUMENT/10593428
- A study on integrating machine learning techniques for waste management (ICCPCT 2024) https://ieeexplore.ieee.org/document/10672957

CERTIFICATES

- Snowflake - Hands-on Essentials - Data Warehouse

- Microsoft - Introduction to Microsoft Azure Cloud Services
- Microsoft - Harnessing the Power of Data with Power BI
- Forage - Accenture Data Analytics and Visualization Job Simulation certificate on the Forage platform.
- Forage - BCG Data Science Job Simulation on Forage

EDUCATION

Institute/College	Duration	Percentage Obtained
University of Texas at Arlington	2023-24	3.4
Bharat Institute of Engineering and Technology	2018-22	3.0