ASHISH PRADIPRAO AAGE

74, Sadguru nagar,Manewada,Besa Road, Nagpur,440027 · +91-8087135160
ashish26aage@gmail.com www.linkedin.com/in/ashish-p-aage-85682b235

Objective

To be associated with a progressive organization that provides an opportunity for a challenging and rewarding career by applying my knowledge, skills, and potential in the profession. I would also like to make a positive contribution to your organization by promoting team spirit.

Professional Summary

Overall, 3.2 years of extensive experience in different domains in the IT industry and **mainly 2.4 years** in Hadoop Ecosystem Development and PySpark API.

- Experience with Big Data core components and Ecosystems including Data Ingestion and Data Processing (pySpark, Hive, HDFS, HBase).
- Injected and processed a large amount of data from various structured and semi-structured sources into HDFS (AWS Cloud).
- Having experience in using optimization techniques in Hive and Spark if required.
- Having hands-on experience on Amazon web services mainly S3, EC2, IAM Roles, AWS EMR.
- Also having knowledge of AWS Glue and AWS Lambda,AWS Redshift.
- Knowledge of Git and Github.
- Experience in data ingestion using spark, transformation, spark-SQL, and performance tuning Experience in transferring data from RDBMS to HDFS and Hive tables using PySpark and Spark API.
- Have experience in writing data extraction logic in SQL.
- Knowledge of writing python scripts.
- Involved in working on Spark SQL Code as an alternative approach for Faster Data Processing and better Performance.
- Developed PySpark programs and created the data frames and worked on transformations.
- Experience in creating tables, partitioning, bucketing, and loading in Hive.

Technology Proficiency

Platforms:	Windows, Linux.
Hadoop Ecosystem:	Hadoop, Hive,Hbase
Dev. IDE:	Pycharm, Anaconda, Jupyter notebook.
Cloud platforms:	AWS (EC2, RDS, EMR, S3, IAM)
Programming languages:	Python
Relational database:	Oracle, Postgres.
Version control tool:	Git, Github
Ticketing Tool:	Jira
Distribution Tool:	CDH
SDLC Methodology:	Agile

Work Experience

Organization: Zielotech Software Pvt.Ltd.

Duration: 6/09/2019 – Till now

Designation: Data Engineer

Project1: **Financial Intelligence Unit** 19/04/2021 – Till now

Description & Responsibilities:

- Import Data into the Hbase from Relational Database (Postgre) for Historical data and HDFS for delta data using Spark.
- Write Phoenix DDL to create Hbase Table.
- Write Hive DDL to Create Hive Table for Optimize Query Performance.
- CSV Files like Delimited, Fixed Length, etc. into Hive Warehouse.
- Implement Data Validation, Quality Checks.
- Involved in importing data from Hbase into Hive Managed Tables using Spark which includes incremental Load and some transformations.
- Involved in importing data from Hive Managed Table into Hive External Tables which includes Queries using Spark.

- Designed both Managed and External Hive Tables and Defined static and dynamic partitions as per requirement for optimized performance on production datasets.
- Written Hive Queries for Data Analysis to meet the business requirements.
- Create PySpark jobs for importing data into Hbase.
- Create PySpark jobs for data transformation and aggregation.
- Produce unit tests for Spark transformations and helper methods.

Environment: Hadoop, Hive, Hbase, Apache Spark, AWS RDS, AWS S3, PySpark

Project2: **Retail store big data pipeline** 25/03/2020 -16/02/2021

V Mart Basically Its chain of retail stores, serving customers world-wide with a wide range of products like groceries, home decor, furniture, electronics has selling products world-wide and generating the sales data at their POS (point of sale) Systems. The daily data from POS (point of sale) Systems world-wide, get pushed into hdfs and then from hdfs it is pushed to big data pipeline for cleaning, processing and Business analysis.

Description & Responsibilities:

- Firstly, we have to build a Data Ingestion Framework. To migrate the data from the AWS S3.
- At the one end of this project Architecture, there is an Oracle RDS instance and another end was our AWS S3.
- On the EMR cluster we have Hadoop and Spark.
- The cluster was designed in such a manner so that AWS S3 data could be transferred into Oracle RDS.
- Importing from AWS S3 to HDFS stored as CSV via PySpark.
- Transformation as per business logic was done over these CSV files using PySpark and later ingested into Oracle RDS.
- And after separating table column having null values, we perform join operation and store it into Oracle RDS then that data will be used for visualization.

Environment: Oracle, AWS S3, PySpark, HDFS, Power BI.

Education

2010-2013: Bachelor of Engineering from Nagpur University.