

Vaishnavi Boje



+918600442876



vaishnaviboje2699@gmail.com



<https://www.linkedin.com/in/vaishnavi-boje/>



Pune, Maharashtra

Hardworking and reliable Data Engineer with strong ability in designing pipeline and analytics with 2.5 years of experience.



Skills

Big Data Technologies: Hadoop, HBase, Pig, Hive, Kafka, SQL, Pyspark, Snowflake, Airflow.

Data Processing: Python (NumPy, Pandas). Basic - Machine Learning

Cloud Platforms: Google Cloud Platform (GCP).

GCP Services: Cloud Storage, Compute Engine, BigQuery, Dataflow, Cloud Functions, Pub/Sub, Cloud Composer, ETL.

Programming Languages: Basic- Scala | Intermediate - C, C++, Java | Advanced - Python.

Databases: MySQL, MongoDB, Cassandra.

Others: Git, CI/CD, Bitbucket, Basic- GenAI.



Work History

2023-01 - Current

Data Engineering Analyst | [Accenture](#)

Postal service pipeline implementation and enhancements:

Designed and implemented a robust data pipeline utilizing Apache Airflow and Google Cloud Dataflow, seamlessly ingesting Pub/Sub messages into BigQuery tables, enhancing real-time data processing and storage efficiency.

Engineered a comprehensive solution for transforming BigQuery tables as per ML model results and publishing details to Pub/Sub topics, demonstrating proficiency in end-to-end data processing, orchestration, and message communication.

Demonstrated expertise in GCP's data ecosystem by orchestrating the entire workflow, ensuring smooth integration of Pub/Sub, BigQuery, and Dataflow. Sound knowledge on implementing pipeline with Apache beam Python SDK.

Data Engineering Associate | [Accenture](#)

Collaborated with the development team to write Hive queries and its enhancement on Hadoop with YARN cluster manager. worked on on-premises to GCP migration and conducted root cause analysis for failing queries, ensuring seamless data processing and analysis.

Contributed Individually on Python & Pyspark Automation:

1. Utilized Pyspark scripting to automate the creation of 100+ source tables, increasing 50% speed of QA and DEV initial efforts.
2. Implemented a Pyspark script to automate HDFS Commands and Hive query to perform comprehensive verification of 1:1 and transformation logic with report generation saved time required for quality check.
3. Developed a Python Machine Learning - Time Series model (ARIMA) to accurately predict the duration of job execution, resulting in improved efficiency and optimized resource allocation.
4. Implemented a robust script that parses JIRA exports and calculates defect counts for each entity, reducing manual effort and ensuring quick and accurate defect analysis.



Education

2018-08 - 2021-07

Bachelor of Engineering: Information Technology

Government College of Engineering Aurangabad - Aurangabad

2015-07 - 2018-06

Diploma: Information Technology

Puranmal Lahoti Government Polytechnic Latur - Latur



Certifications

2023-05

Google Cloud Certified Professional Data Engineer

2023-01

Google Cloud Certified Associate Cloud Engineer