Bhavya Joshi

Data Engineer

Dublin, Ireland | +353 894913787 | joshibhavya2000@gmail.com

linkedin.com/in/joshibhavya2

Professional Summary

Data Engineer with 4+ years of experience developing and optimizing production data pipelines and large-scale data processing solutions. Experienced in using Python and PySpark to transform complex datasets, validate data quality, improve processing performance, and deliver reliable data for analysis. Skilled in Spark DataFrames, SQL, HiveQL, batch and incremental processing, and workflow orchestration with Apache Oozie. Works directly with business and technology stakeholders to understand data requirements, resolve production issues, and improve delivery processes. Experienced with Linux-based Cloudera environments, including HDFS, YARN, Hive, Spark, and platform security.

Technical Skills

Data Engineering: PySpark, Apache Spark, Spark DataFrames, ETL/ELT, Batch Processing, Incremental Processing, Data Transformation, Data Validation, Data Quality, Data Ingestion
SQL & Data Processing: SQL, HiveQL, Query Optimization, PostgreSQL, SQL Server
Big Data Technologies: Cloudera Data Platform (CDP), HDFS, YARN, Apache Oozie, Parquet, ORC
Orchestration & Automation: Apache Airflow, Apache Oozie, DAGs, Shell Scripting, Job Scheduling, Monitoring, Alerting
Programming: Python, PySpark, SQL, HiveQL, Shell Script
Cloud & Platforms: Databricks, AWS S3, AWS Glue, AWS Lambda, Amazon Athena
Data Platform Security: Kerberos, Apache Ranger, TLS, Active Directory
Tools & Technologies: REST APIs, JDBC, ODBC

Professional Experience

Deloitte Dublin, Ireland
Senior Data Consultant – Data Engineering September 2022 – Present

- Designed, developed, and maintained production ETL pipelines using **PySpark, HiveQL, and Apache Oozie** to ingest, clean, transform, pseudonymize, and deliver structured datasets for downstream analytics.
- Developed **PySpark transformations using Spark DataFrames** to process production datasets from multiple source systems and formats, including CSV, JSON, and relational databases, applying business rules and standardized transformations to produce analytics-ready datasets.
- Built and maintained **historical and incremental data processing pipelines**, enabling reliable ingestion and transformation of new and updated source data while maintaining data integrity across processing stages.
- Implemented **data-quality validation frameworks in PySpark** to perform record-level validation, schema checks, null checks, reconciliation, and data-integrity checks between source and target datasets.
- Orchestrated complex production workflows using **Apache Oozie**, managing dependencies, scheduling, error handling, and automated execution of data-processing jobs; reduced data ingestion time by **83% through automation**.
- Wrote and optimized **HiveQL** queries for large production datasets, including files exceeding 29 GB, and improved storage and processing efficiency using columnar formats such as **Parquet and ORC**.
- Optimized **Spark jobs and YARN resource configurations** to improve distributed processing performance, resource allocation, job scheduling, and reliability across large-scale workloads.
- Mentored new team members on **ETL pipelines, Hive, Spark, and Cloudera CDP**, providing technical guidance and reviewing their work to help them contribute to production delivery.
- Prepared, cleaned, transformed, and delivered high-quality transport datasets through robust ETL pipelines for a **CSEIA award-winning client analytics dashboard**.
- Worked across the complete distributed data platform stack, including **YARN, HDFS, Hive, Spark, Kerberos, Apache Ranger, and TLS**, troubleshooting pipeline, performance, connectivity, authentication, and data-access issues affecting production workloads.
- Led a major **RHEL operating system upgrade across 22 servers** (7.9 Maipo to 8.10 Ootpa), including 15 production servers supporting the enterprise data platform, coordinating upgrade activities and post-upgrade validation to ensure continued availability and stability of **Spark, Hive, HDFS, YARN, and production data pipelines**.
- Performed multiple **Cloudera Manager and Cloudera Runtime upgrades**, coordinating Python and Java version upgrades to maintain compatibility with the **Cloudera support matrix** and validating platform services and production pipelines after each upgrade.
- Designed and implemented a secure **Disaster Recovery (DR) data platform from the ground up** for Cloudera CDP, configuring inter-cluster data replication and recovery mechanisms to support data availability

and business continuity.

- Led **client knowledge-transfer sessions** on data pipelines and platform usage, explaining technical processes to business and technology stakeholders and translating their requirements into engineering solutions.

Projects

Automated ETL Pipeline with Apache Airflow and NASA APOD API
Apache Airflow, Python, PostgreSQL, Docker, REST API

- Designed and implemented an end-to-end ETL pipeline using **Apache Airflow**, extracting JSON data from NASA's APOD REST API, transforming API responses, and loading structured data into PostgreSQL.
- Developed scheduled Airflow DAGs using the **TaskFlow API, Operators, and Hooks** to manage task dependencies, database operations, automated ingestion, monitoring, and error handling.
- Containerized Airflow and PostgreSQL using **Docker**, providing isolated and reproducible development environments for the complete data pipeline.

E-Commerce Data Engineering Project
Databricks, PySpark, SQL, Delta Lake

- Built an end-to-end e-commerce data pipeline in Databricks using Bronze, Silver and Gold layers to transform raw data into analytics-ready datasets.
- Developed PySpark and SQL transformations to clean, join and aggregate data for business analysis and reporting.
- Implemented Delta Lake tables and organised data assets using Unity Catalog within a lakehouse architecture.
- Created a business insights dashboard and explored Databricks Genie for natural-language querying of curated datasets.

Certifications

- Databricks Certified Data Engineer Associate
- Databricks Certified Databricks Fundamentals
- Databricks Certified Generative AI Fundamentals
- AWS Certified Cloud Practitioner
- Cloudera Education – Preparing with Cloudera Data Engineering
- DANA-161 Cloudera Education – Introducing Modern Data Architecture

Education

MSc, Data Science and Statistical Learning September 2021 – August 2022
University of Limerick, Ireland

Bachelor of Engineering, Information Technology July 2017 – June 2021
University of Mumbai, India

Volunteering

Volunteer Educator - Junior Achievement Ireland
Volunteered with the **It's My Business** programme at Presentation Primary School, teaching fourth-class students about entrepreneurship and guiding them through creative product/service brainstorming.

Volunteer Mentor - Time to Talk
Supported individuals experiencing long-term unemployment through the **Time to Talk programme**, offering conversation and encouragement to help build their confidence.