Shiv Kumar

☎ +91 8696147894 ✉ shivkumarkr21@gmail.com in /shivkumar31 /shivkumar31

Experience

Software Development Intern **Oct 2025 – Dec 2025**
Imperative Code *Jaipur, India*

- Orchestrated automated data ingestion pipelines using Python and Pandas to parse and validate raw client data, reducing manual intervention and data entry time by **40%**.
- Refactored legacy SQL queries to optimize data retrieval speeds for internal dashboards, achieving a **25% reduction** in query execution time through indexing and join optimization.
- Designed and maintained normalized database schemas for **5+ client websites**, ensuring strict referential integrity and consistency across user activity logs.

Education

VIT Bhopal University **2022 – 2027**
Integrated M.Tech in Computer Science and Engineering (Computational and Data Science) *Madhya Pradesh*

Projects

Scalable Hybrid Data Lakehouse | *PySpark, AWS Kinesis, S3* **2025**

- Designed a hybrid ingestion layer using **AWS Kinesis Firehose** to stream real-time logs into **S3**, reducing data availability latency from hours to minutes compared to batch loads.
- Integrated **AWS CloudWatch** for log monitoring and configured SNS alerts to notify stakeholders immediately in case of ETL job failures or data quality anomalies.
- Optimized storage footprint by **60%** by enforcing Columnar Parquet format with Snappy compression and implementing Partitioning based on event timestamps.
- Developed robust exception handling logic to manage schema evolution (Schema-on-Read), ensuring zero data loss during high-volume ingestion batches.

Retail Data Warehousing & Modeling | *PySpark, Spark SQL, AWS* **2025**

- Engineered a Dimensional Data Model (Star Schema) using PySpark, transforming raw transactional streams into optimized Fact and Dimension tables for BI reporting.
- Leveraged **Broadcast Joins** to optimize heavy join operations between large Fact tables and small Dimension tables, reducing network shuffle and execution time by **30%**.
- Utilized **Spark SQL Window Functions** to compute complex aggregations (Rolling Averages, Cumulative Sums) across millions of retail records efficiently.
- Automated the loading of processed data into partitioned S3 buckets, enabling seamless integration with external analytics tools like AWS Athena.

Automated Data Quality & Skew Optimization | *PySpark, AWS EMR, Python* **2025**

- Architected a data validation framework using PySpark to audit massive datasets for nulls and anomalies, ensuring **99.9% data reliability** prior to warehouse loading.
- Resolved critical **Data Skew** bottlenecks by implementing Salting techniques and custom repartitioning, preventing Executor OOM (Out of Memory) failures.
- Implemented automated referential integrity checks across distributed nodes to validate relationships between datasets without compromising performance.

Technical Skills

Languages & Core: Python, SQL, Java, Data Structures & Algorithms
Big Data & Frameworks: Apache Spark, PySpark, Spark SQL, AWS Kinesis, Delta Lake
Cloud & Storage: AWS (S3, EMR, EC2, IAM), HDFS, Parquet
Data Engineering: ETL/ELT Architecture, Dimensional Modeling, Query Optimization, Skew Handling
DevOps & Tools: Git, GitHub

Certifications

Introduction to Machine Learning | *NPTEL* **Nov 2024**