Punitkumar More

Data Engineer | Cloud ETL & Big Data Specialist (PySpark Focus)

(1) **Contact:** punitmore31@gmail.com | **LinkedIn:** https://www.linkedin.com/in/punit-more/ | **Mobile:** +919284026736 | **Location:** India

🔑 Technical Skills

Category	Skills	Proficiency/Tools
Big Data Frameworks	**Apache Spark (PySpark), Spark SQL**, Pandas UDFs, Broadcast Join, Partitioning, Caching	**(Self-Trained)**
Cloud Platform (GCP)	**BigQuery (Advanced),** Google Cloud Composer (Airflow), GCS, **Dataproc (Applied),** Dataflow, Cloud Shell	Associate Google Cloud Engineer Certified
Data Languages	**Python (Advanced), SQL (Advanced)**, Shell Script, Unix	
Data Lake/DWH	**Delta Lake** (Self-Project), Parquet, SCD Type 1 & 2, Data Modeling	
Orchestration/CI/CD	**Apache Airflow (DAG Development, Maintenance),** ETL/ELT	
Legacy ETL	IBM InfoSphere Datastage, Teradata BTEQ	(Historical Experience)

💼 Professional Experience (Total: 3.10 Years)

Datametica Birds, Data Engineer

Dec 2021 – Present

Project Phase II: DataStage Migration to Google Cloud Platform

Client: IBC, Health Insurance Domain

- **Cloud-Native Data Pipeline:** Led the transition from DataStage to a **scalable, cloud-native ETL framework** utilizing Airflow and BigQuery, successfully reducing infrastructure costs and processing time by **30-40%.**
- **Data Modeling & MERGE Logic:** Designed and implemented robust **Slowly Changing Dimension (SCD Type 1 & 2)** logic for critical datasets, utilizing highly efficient **MERGE** statements to enable incremental load capabilities. (Directly transferable to Delta Lake operations).

- **Performance Optimization:** Optimized BigQuery ETL processes using advanced SQL features (**Window Functions, Clustering, and Partitioning**) to ensure performance parity with legacy systems and achieve a **30% reduction** in system runtime.
- **Orchestration & Automation:** Developed and maintained custom **Apache Airflow DAGs** for end-to-end data workflows, improving system efficiency by **40%** and minimizing manual intervention.
- Collaborated with testing teams (SIT/UAT) to validate data accuracy and ensure readiness for production deployment across migration phases.

Project Phase I: Teradata EDW Migration to Google Cloud Platform

Client: IBC, Health Insurance Domain

- **Legacy Analysis:** Analyzed complex Teradata BTEQ and KSH scripts to accurately replicate core business logic within the BigQuery cloud environment.
- **Migration Success:** Spearheaded the successful migration of Teradata jobs to BigQuery, completing the project within the designated timeline with **zero data loss** to production systems.
- Collaborated with Development and Field teams to identify and resolve bugs, resulting in a **30% improvement** in cloud-based ETL performance and laying the groundwork for enhanced future migration efficiency.

💡 Applied Data Engineering Projects (PySpark & Distributed Computing)

This section demonstrates hands-on experience with the PySpark framework and distributed processing techniques.

Project 1: Large-Scale Transaction Processing Engine (PySpark/Dataproc/Delta Lake)

- **Goal:** Engineered an end-to-end data pipeline to ingest 10GB of simulated financial transaction data, focusing on high-speed join and aggregation operations.
- **Key Achievement:** Developed PySpark transformations that utilized **Broadcast Joins** to efficiently link customer reference data (small dimension table) to transaction streams (large fact table), achieving a **5x speedup** in the join phase.
- **Data Lake Target:** Wrote optimized output to **Delta Lake** tables on GCS, enabling efficient upserts and time-travel functionality.
- **Tools:** PySpark, Spark SQL, Dataproc, GCS, Delta Lake.

Project 2: Legacy Data Standardizer (Airflow & PySpark)

- **Goal:** Built an Airflow DAG to orchestrate a PySpark job that cleans and standardizes messy text files into a quality-controlled Parquet format, preparing data for analysis.
- **Key Achievement:** Implemented data quality checks using PySpark and leveraged **Pandas UDFs (Vectorized)** for efficient, parallel string operations (e.g., standardizing address formats), showcasing best practices for custom logic execution on Spark clusters.

- **Tools:** PySpark, Apache Airflow, Parquet, Python Pandas.

🏆 Education & Certifications

Education:

- **BE Computer** – SSBT's College Of Engineering and Technology, Jalgaon (Aug 2018 – Sep 2021)
- **Diploma In Computer Engineering** – Government Polytechnic, Jalgaon (Aug 2015 – May 2018)

Certifications & Awards:

- **Google Cloud Certification:** Associate Google Cloud Engineer
- **HULK Award & SPOT Award** – Datametica Birds (Recognized for efficiency and technical expertise in resolving critical production issues).