

Utkarsh Singh

utkarsh.singhqwe@gmail.com | +91 88515 52639 | [LinkedIn](#) | [GitHub](#)

Experience

IBM

Dec 2025 – Present

Data Engineer | Data Platform Practice | (PySpark, AWS Glue, Neo4j, ANTLR4, sqlglot, Databricks, Airflow)

- Led R&D and primary development of an enterprise data-lineage platform for a UK banking estate, building multi-source parsers (Tableau, QlikView, TWS, Spark) that emit column- and job-level lineage into a unified Neo4j knowledge graph with deterministic SHA-256 node IDs and APOC-free pure-Cypher MERGE writes.
- Built an internal Ab Initio-to-PySpark conversion platform, developing a two-layer wrapper architecture, DML-to-PySpark schema conversion, and an IR/Neo4j graph store, backed by comprehensive automated test suites (DML test cases and a pytest-based PySpark fixture harness).
- Engineered and hardened PySpark/AWS Glue pipelines and parser layers with ANTLR4 grammars, lxml, and sqlglot; diagnosed and resolved lineage-accuracy defects (silent zero counts, datasource inflation, edge/resource undercounting) and stood up cross-platform pytest collection and module-mocking for reliable CI validation.

NextBigGames

Jun 2024 – Nov 2025

Data Engineer | (Azure Data Factory, Azure Data Lake, PySpark, Delta Lake, Azure SQL)

- Built an end-to-end gaming market intelligence pipeline using Medallion architecture (Bronze-Silver-Gold), ingesting multi-source data from AppMagic, Play Store, and App Store across Idle, Action, and RPG genres.
- Developed automated ADF ingestion workflows into the Bronze Lakehouse, and Silver-layer logic for cleaning, filtering, outlier detection, and cross-source normalization.
- Engineered Gold-layer datasets surfacing market trends, failure analysis, and competitive benchmarking; automated a narrative report generator converting datasets into stakeholder-ready documents.
- Integrated Gold-layer datasets with internal dashboarding systems, improving analytics accessibility and usability across business and product teams.

Projects

Ariadne — Enterprise Multi-Parser Data Lineage Platform

[GitHub](#)

Python, ANTLR4, lxml, sqlglot, libcst, Neo4j (Cypher), PostgreSQL, SHA-256

- Architected a multi-parser lineage platform extracting column- and job-level lineage from Tableau, QlikView, TWS, and Spark sources into a unified Neo4j knowledge graph using deterministic SHA-256 node IDs and APOC-free pure-Cypher MERGE writes.
- Built parsers with ANTLR4 grammars, lxml, sqlglot, and Python AST/libcst; diagnosed and fixed lineage-accuracy bugs (silent zero counts, datasource inflation, edge/resource undercounting) and authored realistic fixture suites for end-to-end parser validation.

NextBigIdeas — Multi-Source Gaming Market Intelligence Platform

[GitHub](#)

Azure Data Factory, Azure Data Lake, PySpark, Delta Lake, Azure SQL, Python, SQL

- Designed and built an end-to-end market intelligence platform on Medallion architecture (Bronze-Silver-Gold), ingesting competitive data from AppMagic, Play Store, and App Store across the Idle, Action, and RPG genres.
- Developed automated Azure Data Factory ingestion workflows into the Bronze layer and Silver-layer PySpark processing for cleaning, filtering, outlier detection, and cross-source normalization, reconciling schema and metric differences between providers.
- Engineered Gold-layer datasets surfacing market trends, failure analysis, and competitive benchmarking; automated a narrative report generator and integrated outputs with internal dashboarding systems.

Skills

- Languages: Python, SQL, Java, JavaScript.
- Big Data & Streaming: Apache Spark (PySpark), Databricks, Hadoop, Kafka, Airflow.
- Cloud & Platforms: AWS (S3, Glue, Athena, EMR, Redshift, Lambda, Kinesis, EventBridge, EC2, VPC), Azure (ADLS, Synapse, ADF), Microsoft Fabric (OneLake, Lakehouse, Warehouse, Pipelines, Notebooks).
- Databases: PostgreSQL, MySQL, Snowflake, Redshift, DynamoDB.
- Tools & DevOps: Docker, Git, Jupyter.

Certifications

- AWS Certified Data Engineer – Associate (DEA-C01); AWS Certified Machine Learning – Associate (MLA-C01); AWS Certified AI Practitioner (AIF-C01).
- Microsoft Certified: Fabric Data Engineer Associate (DP-700); Microsoft Certified: Azure Fundamentals (AZ-900); Microsoft Certified: Azure Data Fundamentals (DP-900).
- Claude Certified Associate – Foundations (CCAO-F); Claude Certified Developer – Foundations (CCDV-F).

Education

Amity University, Noida — Master of Computer Applications (MCA), Computer Science

2023 – 2025

Kristu Jayanti University — Bachelor of Computer Applications (BCA), Computer Science

2019 – 2022