Aditya Agarwal

+1 (930) 333 2884 | www.linkedin.com/in/adityaagarwal5 | agarwaladitya1202@gmail.com

TECHNICAL SKILLS

- **Programming & Scripting:** Python, SQL, Scala, Java, Go, C, C++, C#, HiveQL, Unix Shell Script
- **Big Data & ETL Tools:** Apache Spark, Flink, PySpark, Hadoop, HDFS, Presto, SSIS, SSRS, Hive, Sqoop, Kafka, Apache Airflow, DBT
- **Databases:** Snowflake, Databricks, Amazon Redshift, Google BigQuery, MS SQL Server, PostgreSQL, MySQL, MongoDB, DynamoDB
- **Cloud Platforms:** AWS (S3, EC2, EMR, Glue, Athena, Lambda, Kinesis, MSK, IAM), Azure (ADF, Synapse, ADLS), Terraform, Docker
- **Data Engineering:** Batch & Real-Time Pipelines, CDC, Data Modelling, Data Warehousing, ETL/ELT, Data Quality, Data Governance, Privacy
- **Visualization & BI Tools:** Power BI, Looker, Tableau, Qlik
- **Data Formats:** Parquet, Avro, JSON, XML
- **Version Control & CI/CD:** Git, Bitbucket, Jenkins, Azure DevOps
- **Methodologies:** Agile (Scrum, Kanban), JIRA, Confluence

WORK EXPERIENCE

MyEdMaster LLC — **Leesburg, VA (Remote)**
Software Developer — *Jun 2025 - Present*

- Designed **AWS-hosted ETL/ELT** pipelines integrating AI product data using **Python** and **MySQL,** reducing API latency by **40%** and boosted user engagement by **15%**.
- Developed automated data ingestion scripts for cross-platform AI product integration, improving data availability and reducing manual processing by **50%**.

Indiana University — **Bloomington, IN (Remote)**
Research Assistant — *Jun 2025 - Present*

- Architected multi-platform graph data ingestion pipelines for **Neo4j** and **PuppyGraph**, improving large scale query performance by **25%**.
- Built benchmarking scripts and performance dashboards using Python and graph query languages, providing actionable insights that optimized research workflow.

HealthEdge Software — **Bangalore, India**
Data Engineer II — *Apr 2022 – Jul 2023*

- Engineered modern data pipelines **(ETL, ELT, CDC)** using **Airflow, AWS Glue, ADF, Databricks**, migrating **70TB** of **US Health Claims** data to **AWS S3** and **Azure Delta Lake**, which elevated data availability by 80% and improved data governance crucial for product reliability**.**
- Orchestrated real-time streaming pipelines with **Kafka (AWS MSK)** and **Airflow**, processing **500GB/day** into cloud data lakes to power analytics, reporting, and machine learning initiatives that directly informed product development.
- Integrated **OLTP systems** with **OLAP platforms** (**Snowflake**, **Databricks**) to accelerate reporting and support critical business decisions.
- Devised a **Python**-based data validation framework that reduced manual testing by **30%** while improving data quality and pipeline reliability.
- Spearheaded 3+ **proof-of-concept (POC)** initiatives for cloud migration and streaming architecture, demonstrating **2x scalability** and presenting outcomes to leadership to influence future product infrastructure.

HealthEdge Software — **Bangalore, India**
Data Engineer I — *Jan 2020 – Mar 2022*

- Constructed analytics layers using **dbt** on **Snowflake** and **Databricks SQL** to standardize data models, improving analyst efficiency by **40%**.
- Automated **Snowflake** schema deployments via **CI/CD** pipelines using **Bitbucket**, ensuring version control and reliable data integration.
- Authored **20+ SQL procedures** for **reporting**, **auditing**, and **business intelligence**, delivering high-impact insights to stakeholders.
- Produced **15+** dashboards in **Tableau** and **Power BI**, reducing manual reporting by **40%** and enabling real-time KPI monitoring.
- Contributed to **Agile Scrum** ceremonies, helping the team achieve **90%** sprint goal completion and ensuring alignment on priorities.
- Mentored and trained **5+ interns**, improving team productivity by **25%** through knowledge sharing and collaboration.

AWARDS & ACHIEVEMENTS

- Awarded **Best Performing Employee** at **HealthEdge** for **2021, 2022, 2023**, recognizing exceptional contributions to data engineering and process automation.
- **HIPAA Certified**, ensuring compliance with data security and privacy regulations for handling **sensitive healthcare data**.

PROJECTS

Real-Time Change Data Capture (CDC): MySQL, Apache Kafka, Debezium, Spark, Airflow, Google BigQuery — *Mar 2025 – Apr 2025*

- Implemented a real-time CDC pipeline using Debezium & Kafka to stream MySQL data changes, automating ETL with Airflow orchestration, reducing data latency from hours to seconds and enabling instant reporting.

ETL Pipeline for JSON Data Processing: Python, PySpark, Apache Spark, Hive, MySQL — *Jan 2024 – May 2024*

- Built a **PySpark** data pipeline to process large **JSON** files, flattening nested fields, creating normalized tables, and optimizing **Spark** jobs to reduce execution time by **60%**.

EDUCATION

Indiana University — **Bloomington, IN**
Master of Science (MS) in Computer Science *(GPA: 3.83/4.0)* — *Aug 2023 –May 2025*

REVA University — **Bangalore, India**
Bachelor of Technology (B.Tech) in Computer Science and Engineering *(GPA: 9.29/10)* — *Aug 2016 – Jun 2020*