Aditya Agarwal

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

TECHNICAL SKILLS

- **Programming & Scripting:** Python, SQL, Scala, Java, 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, Redshift, MS SQL Server, PostgreSQL, MySQL, MongoDB, DynamoDB
- **Cloud:** AWS (S3, EC2, EMR, Glue, Athena, Lambda, Kinesis, Firehose, IAM), Azure (ADF, Databricks, Synapse, ADLS), Terraform, Docker
- **Data Engineering:** ETL/ELT Pipelines, Data Warehousing, Real-time Streaming, Data Modeling, Data Governance, Data Security
- **Data Visualization & Reporting:** Power BI, Looker, Tableau, Qlik
- **Data Formats:** Parquet, Avro, JSON
- **Version Control & CI/CD:** Bitbucket, Git, Jenkins
- **Collaboration & Methodologies:** Agile (Scrum, Kanban), JIRA, Azure DevOps, Confluence

WORK EXPERIENCE

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

- Configuring **AWS** cloud platforms to ensure scalable and reliable hosting of multiple AI-based health and educational products.
- Designing robust **data pipelines** to facilitate seamless **data integration** and efficient data flow across various product ecosystems.
- Developing critical backend functionalities and high-performance **API services** using **Python** and **MySQL**, directly contributing to the core features and operational excellence of **AI-driven** applications.

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

- Leading data preparation and loading for a graph database benchmarking project, integrating datasets like the Stack Overflow Graph into **4 different graph database systems** (**Neo4j**, **GraphFrames**, **TigerGraph**, **PuppyGraph**).
- Architecting and executing comprehensive benchmarking queries and analytics tasks, systematically collecting and analyzing performance data to identify optimization opportunities across multiple evaluated systems.
- Assisting in the integration of new graph mining algorithms into **two key graph database libraries** (**Neo4j** and **PuppyGraph**).

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%**.
- Tuned **SQL queries** in **MS SQL Server**, **Snowflake**, and **Oracle**, achieving up to **80%** improvement in query performance.
- 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*

- Developed a real-time CDC pipeline using Debezium & Kafka to stream MySQL data changes, automating ETL with Airflow and resolving schema/authentication challenges.

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*