HARSHDEEP HARIKESH MISHRA

harshdeepmishra82@gmail.com | +1 716 426-8682 | linkedin.com/in/harshdeep-mishra/ | github.com/harshdM99 | Portfolio

PROFESSIONAL EXPERIENCE

Haver AI Inc - Boston, MA, USA | *Data Engineer Intern*   Jun 2024 – Aug 2024

Led the development of an open-source patient healthcare product, building a high-quality data lake, and data pipelines to automate AI-driven predictions based on user-defined prediction parameters.

- Built a **data lake** by automating ingestion of 5+ GB of monthly data using AWS Lambda and Eventbridge, fetching FHIR claims data from BCDA API and ensuring real-time alerts via AWS SNS for reliability.
- Achieved a 50% increase in data processing speed with AWS Glue, integrating vendor data into Redshift for scalable analytics.
- Developed a serverless REST API backend with AWS API Gateway and Lambda, enabling 100% on-demand pipeline execution.
- Engineered a **scalable MLOps pipeline** on AWS SageMaker with AutoML, automating ML workflows and predictions.

Skuad - India | *Data Engineer*   Dec 2022 - May 2023

- Executed client **business KPI** calculations on user activity and network performance with PySpark on GCP, accelerating analytics.
- Gathered client requirements and developed data processing pipelines using AWS Glue and GCP Dataproc ephemeral clusters, reducing cloud costs by 30%.
- Improved SLA resolution time by 50% by developing Grafana dashboards for GCP service monitoring.

Quantiphi - India | *Data Engineer*   Jul 2021 - Nov 2022

Devised cloud-based solutions and orchestrated scalable workflows on GCP, collaborated with Google teams on POCs, and executed data and workflow migrations for North and South America-based clients.

- Spearheaded creation of a cloud solution on GCP with Apache Airflow and Cloud Vision API for parallel processing of **1M+ PDF documents** and reducing processing time from 28 days to 10 days.
- Enhanced the observability platform by implementing structured logging in GCP cloud logging, reducing debugging time by 70%.
- **Migrated 20+ GB** of parquet data from on-prem HDFS to BigQuery using Azure Blob Storage and Google Transfer Service.
- Led end-to-end migration of oozie projects from Hadoop to GCP Composer (Airflow), improving resource utilization by 40%.

EDUCATION

University at Buffalo, The State University of New York   Aug 2023 - Dec 2024

Master of Science, Computer Science and Engineering   **GPA: 3.7/4.0**

Relevant Coursework: Algorithms, Database Systems, ML/AI, Data Computing, Data Warehousing & Mining

University of Mumbai   Aug 2017 - May 2021

Bachelor of Engineering, Computer Engineering   **GPA: 9.15/10**

PROJECTS

Real-Time Data Streaming and Analytics for Log Data | *Kafka, Spark Streaming, Cassandra, AWS, Plotly, Dash*   **| Demo | GitHub**

- Streamed and processed **100+ logs per minute** using Kafka and PySpark, achieving less than 10s latency for real-time insights.
- Stored and indexed 200K+ processed logs in Cassandra and S3, enabling real-time trends visualization with Plotly Dash.

Data Pipeline with Snowflake and dbt | *Snowflake, dbt, Dagster, SQL, Streamlit*   **| Demo | GitHub**

- Designed an **ELT pipeline** with Snowflake and dbt, optimizing data transformations and increasing developer efficiency by 20%.
- Implemented dbt incremental models, reducing query costs by 80% by processing only delta data.
- Automated data workflows with Dagster, streamlining transformations and visualizations, reducing manual effort by 60%.

Data Mining and Blockchain Analysis | *Tor, Python, JavaScript, Selenium, Neo4j, GraphQL, SQS*   **| Demo | GitHub**

- Developed a dark web crawler in Python with Tor and Selenium, extracting Bitcoin addresses from 100K+ onion sites.
- Designed blockchain transaction scraper, tracing fund movements and **mapping 15K+ relationships** in Neo4j for pattern analysis.

TECHNICAL SKILLS

- **Programming**   : Python, JavaScript, SQL, C, C++, Java
- **Web**   : Node.js, React.js, REST APIs, HTML, CSS, Express.js
- **Data**   : Spark, Beam, Kafka, MySQL, Hadoop, Hive, Snowflake, Apache oozie, Data Modeling, ETL, ELT
- **Cloud platforms** : Amazon Web Services (AWS) – EC2, S3, Glue, lambda, Redshift, API Gateway
                          Google Cloud Platform (GCP) – Dataproc, BigQuery, Composer, Pub/Sub
- **Additional Skills** : Git, Linux, Docker, terraform, Shell Scripting, Machine Learning

PROFESSIONAL CERTIFICATIONS & ACHIEVEMENTS

- Google Cloud Platform Associate Cloud Engineer (link) (*January 2022*) – Certified by Google.
- SnowPro Core Certification (link) (*March 2022*) – Certified by Snowflake.
- **Winner**, Smart India Hackathon (SIH) 2019 – A national-level hackathon held at NIT Trichy, competing against **50+** teams.