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 development for Open-Source Version of the patient healthcare product to deliver a high-quality data lake, and data pipelines to be used by AutoML to drive innovation and predict user customized features.

- Built a data lake by automating **monthly data ingestion jobs** using AWS Lambda and AWS Eventbridge to fetch FHIR claims data from BCDA API, ensuring reliability with real-time alerts via AWS SNS.
- Achieved a 50% increase in data processing speed utilizing AWS Glue and Spark and integrated vendor data into Redshift.
- Designed and developed a REST API backend service using AWS API Gateway and Lambda.
- Engineered a **scalable MLOps pipeline** using AWS SageMaker and AutoML to execute ML workflows and generate predictions.

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

- Implemented business KPIs leveraging PySpark on GCP with Customer 360 data, **accelerating customer analytics** by 10%.
- Gathered client requirements and developed data processing pipelines on AWS using Glue, S3, Athena, and on GCP using Dataproc ephemeral clusters and BigQuery, reducing cloud costs by 30%.
- Built Grafana dashboards for GCP service monitoring, reducing SLA ticket resolution time by 50%.

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

Developed 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.
- Modernized the observability platform by leveraging structured log formats, reducing overall debugging time by 70%.
- **Migrated 20+ GB** client 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**

TECHNICAL SKILLS

- **Programming**　: Python, JavaScript, SQL, C, C++, Java
- **Web**　　　　　: Node.js, React.js, REST APIs, HTML, CSS
- **Data**　　　　 : Spark, Beam, Kafka, MySQL, Hadoop, Hive, Snowflake, Apache oozie, Data Modeling
- **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

PROJECTS

Real-Time Data Streaming and Analytics for Log Data | *Kafka, Spark Streaming, Cassandra, AWS, Plotly, Dash*　　| Demo | GitHub
- Streamed **100+ logs per minute** into Kafka and processed them with <10 seconds latency using PySpark structured streaming.
- Stored processed data in Amazon Keyspaces/Cassandra and raw logs in S3.
- Visualized real-time and hourly data on an interactive Dash dashboard with Plotly charts.

Snowflake End-to-End Food Aggregator Data Pipeline | *Snowflake, ELT, SQL, Streamlit*　　　　　　| Demo | GitHub
- Architected a **3-layer data warehouse** in Snowflake, developing staging, intermediate, and analytics layers for data processing.
- Implemented ELT process using Snowflake's COPY command and Streams to efficiently load and transform data.

Data Mining and Blockchain Analysis | *Tor, Python, JavaScript, Selenium, MongoDB*　　　　　　| Demo | GitHub
- Developed a dark web crawler in Python using Selenium on Tor network to track illicit cryptocurrency transactions.
- Extracted Bitcoin addresses by **crawling 100K+ onion sites**, storing the data in MongoDB.

PROFESSIONAL CERTIFICATIONS & ACHIEVEMENTS

- Google Cloud Platform Associate Cloud Engineer (link) (January 2022), offered by Google.
- SnowPro Core Certification (link) (March 2022), offered by Snowflake.
- Secured 1st place at the Smart India Hackathon (SIH) 2019 amongst **50+** teams held by the Government of India.