# Roshan Sah

Email: roshan.unt@gmail.com

Linkedin Github

Mobile: 781-778-8910


## PROFESSIONAL SUMMARY

- Data Engineer with around 4+ years of Experience in Data Engineering, Machine Learning, Computer Vision, and Data Science including Data Mining, Statistical Analysis with domain knowledge in the Healthcare and Banking industries.
- Proficient in designing, implementing, and maintaining data warehousing solutions using tools such as Amazon Redshift, and Apache Hive.
- Skilled in building efficient ETL pipelines using Apache Airflow.
- Strong knowledge of SQL and extensive experience working with relational databases like MySQL, PostgreSQL, and Oracle.
- Proficient in database modeling, schema design, and query optimization.
- Expertise with distributed systems design and parallel processing, comprehensive knowledge of the Spark and MapReduce execution frameworks.
- Experience with tools such as Hadoop MapReduce, Apache Hive, and Apache Pig.
- Familiarity with data integration platforms like Apache Kafka and Apache NiFi.
- Implemented data quality checks and data governance processes to ensure the accuracy, consistency, and integrity of data.
- Experience in implementing data analysis with various analytic tools, such as Anaconda 4.0 Jupiter Notebook 4.X, and Excel.
- Experienced the full software lifecycle in SDLC, Agile, DevOps, and Scrum methodologies including creating requirements, and test plans.
- Proficient in Python for data manipulation, automation, and workflow orchestration.
- Skilled in scripting languages such as Bash and PowerShell.
- Strong experience in working with UNIX/LINUX environments and writing shell scripts.
- Experience working with cloud-based data technologies such as Amazon Web Services (AWS S3).
- Experience with Machine Learning algorithms such as logistic regression, KNN, SVM, random forest, neural network, linear regression, lasso regression, and k-means
- Good Knowledge and experience in deep learning algorithms such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), LSTM, and RNN-based speech recognition using TensorFlow.
- Proficient in data visualization tools like Tableau, and Power BI for creating dashboards and reports.
- Good working experience in Real-time streaming technologies Spark and Kafka.
- Excellent organizational, communication, and time management skills.

## Technical Skills:

| | |
|---|---|
| **Big Data:** | Apache Spark, Hadoop, HDFS, Map Reduce, Pig, Hive, HBase, Cassandra, Airflow, Snowflake |
| **Programming & Scripting Languages:** | Python, C, C++, HTML, JavaScript, XML, Git |
| **Database:** | Oracle 10g/11g, PostgreSQL, DB2, SQL Server, MySQL, Redshift |
| **NoSQL Database:** | HBase, Cassandra, MongoDB |
| **IDE:** | Eclipse, Net Beans, Jupyter Notebook, VS Code |

| | |
|---|---|
| **ETL Tools:** | Mage AI |
| **Reporting Tool:** | Tableau, Power BI |
| **Operating Systems:** | Windows, UNIX, Linux |
| **AWS:** | Dynamo DB, Redshift, RDS, Data Pipelines, Lake formation, S3, CloudFormation. |
| **Machine Learning Libraries** | PyTorch, TensorFlow, Keras, Scikit-Learn, Matplotlib, Seaborn, Ploty |

## WORK EXPERIENCE

**Graduate Teaching Assistant** **Aug 2022 – May 2023**

University of North Texas Denton, Texas

- Offer one-on-one or group assistance to students, helping them understand data engineering concepts, interpreting data, and programming languages such as Python.
- Lead workshops on data engineering techniques, statistical concepts, and programming languages (such as Python, and SQL) to enhance student's analytical skills.
- Guided in setting up lab infrastructure, configuring software tools, and troubleshooting technical issues.
- Mentored students on data engineering concepts, technologies (Hadoop, Spark), and career pathways.
- Designed and delivered hands-on sessions to help students develop practical skills in data engineering.

**Environment:** Python, Machine Learning Libraries and Algorithms, SQL, Hadoop, MapReduce, Spark.

**Junior Data Engineering** **Jan 2022 – Jul 2022**

Ava Tech Systems LLC Eagan, Minnesota

- Ensured data integrity and optimized performance through proper data validation and error-handling techniques.
- The use of Spark in Python improved ingestion and performance by 67% while processing massive streams of data.
- Worked with clients to comprehend business requirements and translate those requirements into reports that could be acted on in Tableau, saving 17 hours of manual labor per week.
- Utilize Hive scripts to create tables in Hive and load, process, and analyze data. Dynamic Partitions, Buckets, and Partitioning have been implemented in Hive.
- Data was ingested from various sources and created data views for use in business intelligence (BI) applications like Tableau using a combination of Python, Google Analytics API, and SQL.
- Built and constructed a real-time data pipeline to analyze semi-structured data by combining 150 million raw records from 30+ data sources with Kafka and PySpark.

**Environment**: Tableau, Python, Hadoop, Hive, ETL, SQL, AWS, Kafka, Pyspark

**Software Engineer** **Mar 2018 – Jul 2021**

RS Technology Pvt. Ltd. Lalitpur, Nepal

- Develop data security protocols with the help of cross-functional teams to make sure that sensitive data is always secured.
- Used Power BI to develop a data visualization solution that would aid business users in comprehending and analyzing data.
- Utilized Spark, Redshift, S3, and Python to maintain a data pipeline uptime of 99.8% while consuming streaming data from 3 separate primary data sources.

- Created a model for predictive analytics using machine learning techniques, increasing customer satisfaction by 11%
- Created Kafka producer API to send live-stream data into various Kafka topics.
- Developed MapReduce jobs in both PIG and Hive for data cleaning and pre-processing.
- Used Hive queries to analyze huge data sets of structured, unstructured, and semi-structured data.

**Environment:** Python, Pandas, Scikit-learn, NumPy, Power BI, Hive, S3, Redshift, Spark, Kafka

**Big Data Intern** **Sept 2017 – Nov 2017**
ERA-InfoTech Ltd Dhaka, Bangladesh

- Maintained a large database and used various statistical techniques to collect, analyze and interpret data from customers.
- Maintain existing data management, data query, and ETL procedures.
- Advance demonstrable experience in building, validating, and leveraging machine learning models.

## ACADEMIC PROJECTS

- **Uber Data Analytics:**
  The project focused on leveraging advanced data analytics techniques to build a robust and scalable data infrastructure. I ensured seamless data ingestion, processing, storage, and retrieval using Google Cloud services. I utilized the powerful Mage AI ETL tool to efficiently load the raw data, perform essential transformations, and export it to the Google Big Query data warehouse. By analyzing the ride data, I uncovered valuable insights into ride pickup and drop-off patterns, customer preferences, and fare statistics. To enhance data visualization and accessibility, I implemented Google Looker, creating an interactive dashboard that provided stakeholders with real-time insights.
  **Technologies Used:** Google Cloud Storage, Python, Compute Instance, Mage Data Pipeline Tool, Big Query, Looker Studio, XML, and SQL.

- **Twitter Hate Speech Detection:**
  The Project was to detect hateful tweets and send a reminder to users if a tweet is hateful and what part of the text is hateful. This is to reduce hate and abusive tweets.
  **Technologies Used:** Natural language Processing, Python, AWS

## EDUCATION

- **University of North Texas** **Denton, Texas**
  Master of Science in Artificial Intelligence; GPA: 3.88 Aug 2021 – May 2023
  **Courses:** Big Data & Data Science, Data Visualization, Data Mining, Machine Learning, Deep Learning, Feature Engineering, Natural Language Processing, Methods in Empirical Analysis, Software Development of AI

- **Hajee Mohammad Danesh Science & Technology University** **Dinajpur, Bangladesh**
  Bachelor of Science in Computer Science and Engineering; GPA: 3.42 Feb 2013 – Dec 2017