SOVIT NAYAK

Newark, NJ | 716-295-3415 | sovit.nayak03@gmail.com | linkedin.com/in/sovitnayak/ | github.com/sovitnayak123

EDUCATION

New Jersey Institute of Technology, Newark, NJ *Dec 2024*
Master of Science in Data Science (Statistics), GPA: 3.75

- Relevant Coursework: Data Engineering, Machine Learning, Big Data Analytics, Data Visualization
- Key Projects: Developed ETL pipelines using AWS Glue and Redshift, optimized data retrieval time by 35%, and analyzed large datasets using Python.

SKILLS

- **Programming & Query Languages**: Python, Java, R, SAS, SQL, NoSQL (DynamoDB, MongoDB), HiveQL, SparkSQL
- **Big Data Technologies**: Hadoop, Hive, Apache Spark, AWS EMR
- **ETL Tools & Processes**: Informatica, Apache Airflow, AWS Glue, SSIS
- **Data Modeling & Warehousing**: Star Schema, Snowflake Schema, Redshift, Snowflake
- **Data Visualization**: Tableau, Quicksight, Power BI
- **Cloud Technologies**: AWS (Redshift, EC2, S3, Glue, Athena)
- **Statistical Analysis & Machine Learning**: Scikit-learn, Pandas, NumPy, Matplotlib, Regression, Classification, Time Series Analysis, NLP, TensorFlow, PyTorch, Git, Docker
- **Project Management**: Agile Methodology, Jira, Cross-Functional Team Collaboration

PROFESSIONAL EXPERIENCE

Vigil AI, Atlanta, GA *Aug 2024 - Present*
Data Engineer

- Engineered **15+ scalable ETL pipelines** using **AWS Glue** and **Informatica**, reducing data integration time by 50% and processing 2TB+ of data weekly for enterprise reporting.
- Designed and implemented data models using **Redshift** and **Snowflake**, boosting query performance by 40% and reducing data retrieval times from 60 seconds to under 30 seconds.
- Utilized **HiveQL** and **SparkSQL** for large-scale data querying, cutting down processing time by 35% across 1B+ records.
- Built a distributed streaming data pipeline with **Apache Kafka** and **Spark**, increasing data processing speed by 60%, handling over 10M events per day.
- Enhanced existing **ETL pipelines using NoSQL databases (DynamoDB)** for optimized data retrieval, reducing query latency by 30% and handling over 1 million read/write operations per day.
- Automated **data validation** scripts in Python, reducing data quality issues by 40% and enhancing data accuracy for analytics.

Career Labs, Bengaluru, India *Sept 2021 - July 2023*
Data Analytics Engineer

- **Orchestrated the development of 20+ scalable ETL pipelines** using **Python**, **SQL**, and **Apache Airflow**, improving data integration efficiency by 50% and supporting weekly ingestion of 1.5B records.
- Designed **data models for financial analytics** using **Star Schema**, decreasing query execution times by 30%, from 45 seconds to 30 seconds, and supporting 15+ interactive dashboards.
- Developed **Tableau** and **QuickSight** dashboards for 25+ KPIs, increasing data visibility and enabling data-driven decisions, leading to a 20% increase in stakeholder engagement.
- Deployed **AWS EMR** and **Hadoop** for distributed data processing, reducing analysis time by 40% for datasets exceeding 500GB.
- Implemented data transformation projects using **Informatica**, automating 95% of manual processes and increasing pipeline throughput by 30%.
- Leveraged **AWS (S3, Redshift, Glue)** to automate data integration across 10+ data sources, reducing data retrieval time by 40% and increasing reporting efficiency, supporting 50+ daily business intelligence queries.
- Integrated **NoSQL** solutions **(DynamoDB, MongoDB)** for data storage and retrieval, improving data processing efficiency by 25% across 500 million records.
-

PROJECTS

Marketing Attribution Insights for Amazon Music | SQL, Tableau, AWS Redshift

- Developed a comprehensive dashboard using **Tableau**, analyzing over 50M+ marketing records, enabling data-driven decisions and reducing time to insight by 45%.
- **Optimized AWS Redshift** data queries and developed **advanced data models**, reducing query latency by 50% and improving reporting efficiency by 35%, supporting real-time analytics across 1M+ records daily for 10+ business intelligence use cases.

Real-Time ETL Pipeline for E-Commerce Analytics | AWS EMR, Spark, Hive, Python

- Built a real-time ETL pipeline using **AWS EMR** and **Apache Spark**, processing 1B+ records weekly, increasing data processing capacity by 50%.
- Designed a **star-schema data model**, reducing query times by 40% and supporting 100+ daily business intelligence queries.
- Automated **data ingestion with AWS Glue**, boosting ETL efficiency by 30% and minimizing manual intervention