Shouvik Sharma

2829 S Wells Street, Chicago, IL-60616|Phone: 3124592008 | shouvik19@gmail.com | Linkedin | GitHub | Medium

SUMMARY

Results-driven Senior Data Engineer with 5+ years of experience in developing resilient services and client APIs using Python. Extensive experience in building, scaling, and optimizing data pipelines and production data models using Databricks, Snowflake, dbt, and Python. Proficient in managing large-scale data processing, data quality management, and real-time data processing. Strong background in ETL/ELT processes, analytics engineering, and data visualization with tools like Sigma. Experienced in working within a quantitative/systematic trading environment and processing security reference data. Passionate about developing scalable data systems and advocating for best practices in data product usage. Seeking to leverage my skills in developing resilient services and client APIs for high-impact projects.

WORK EXPERIENCE

Data Engineer at Avant LLC, Chicago: **(Aug 2021 –Present)**
- Implemented robust data quality checks using SODA for real-time validation, developed custom Python scripts to rectify anomalies, and automated anomaly detection, resulting in a 30% reduction in data discrepancies and enhanced overall data reliability.
- Developed and optimized complex SQL queries for ETL pipelines, improving query performance by 25%
- Implemented Apache Airflow for workflow orchestration, resulting in a 30% increase in pipeline reliability
- Designed and maintained data pipelines using Python and dbt, achieving a 20% improvement in analytics process efficiency
- Collaborated with cross-functional teams to implement data quality checks using SODA, reducing data discrepancies by 30%, further utilized AWS services including Redshift and S3 for large-scale data processing and storage
- Developed Tableau dashboards for marketing attribution, contributing to a 4% increase in application rates. Extracted insights from marketing data to create source attribution funnel dashboard in Looker, this helped business stakeholders to improve customers' application experience, and increase application rate by 4%.
- Built a refinance product data pipeline for existing loan customers based on their TransUnion credit scores and existing features, contributing to a 10% increase in the overall refinance book.
- Understand and write complex sql queries as a source of ETL pipelines, providing required recommendations to the DBA team such as adding an index on the frequently used tables, ultimately to improve query optimization.
- Utilized Azure Databricks extensively for data processing and analysis tasks, including building and optimizing Spark-based data pipelines for large-scale data processing, and implementing machine learning models for predictive analytics.
- Develop and implement collaborative data strategies to enhance analytics workflows, ensuring data integrity and accessibility, resulting in measurable improvements in cross-functional collaboration and data-driven decision-making within set timeframes.
- Leveraged AWS services including AWS Databricks and S3 for large-scale data processing and storage and utilized Snowflake for efficient data management.

Data Engineer Intern at CNH Industrial Inc., Racine: **(Mar 2021 – Aug 2021)**
- Contributed to the platform's growth by participating in team code reviews and system reviews, standardizing methodologies, and improving CI/CD tooling.
- Architected and managed production data models in Snowflake, including documentation and data tests, improving data processing efficiency by 20%.

Data Engineer at Daten Solutions Inc., Chicago: **(May 2020 – Mar 2021)**
- Developed and automated data migration pipeline from SQL Server to Snowflake using SnowSQL and SnowPipe, and further enhanced data quality by performing dimensional modeling on the migrated data.
- Automated ETL processes using Prefect (Python), enhancing data wrangling capabilities and achieving a 40% reduction in time through large-scale data conversions. Facilitated the seamless transfer of BAAN data into standardized formats for integration into Snowflake.
- Developed client APIs using Python, facilitating seamless integration with research and trading systems.
- Automated reporting process using Excel VBA (Macros) and MySQL maintaining accuracy and saving ~ 75% of time, maintained version control Git, Mercurial, SVN.

Data Engineer – Practicum Student at Labelmaster, Chicago: **(May 2020 – Dec 2020)**
- Minimize technical debt in managing new data requirements, ensuring scalability and sustainability of data solutions, resulting in a 20% reduction in maintenance time within six months.
- Automated hourly status report saving 10 man-hours/week, thus decreasing response time for fixes and campaign failures.

Big Data Developer at Cartesian Consulting, Mumbai: **(Apr 2018- Jul 2019)**
- Designed and implemented ETL processes using DynamoDB, resulting in a 50% reduction in processing time.
- Developed dimensional data models and a MariaDB-powered data warehouse, strictly adhering to integrity and normalization rules. This infrastructure supported the creation of a campaign data-mart and a comprehensive customer one-view for marketing campaigns.

EDUCATION

- **MS in Computer Science and Mathematics**, Illinois Institute of Technology, **GPA: 3.8**   **(Aug 2019 - May2021)**

Related Courses: Big Data Technologies, Applied Statistics, Database Management, Data Preparation and Analysis.

- **MS in Statistics**, NMIMS University, **GPA: 3.35**   **(Jul 2016 - Apr 2018)**
- Certifications: Snowflake Pro Certification, Databricks Certified Associate Data Engineer

SKILLS

- **Programming:** Python, SQL, Scala, Java, HTML, Excel VBA (Macros).
- **Big Data Ecosystem**: Spark, Hadoop, MapReduce, Hive, Pig, Kafka, Flume, Hbase, Microsoft Azure.
- **Distributed Data/Computing Tools**: MapReduce, Hive, Spark
- **Cloud Technologies**: Azure Data Factory, Azure Blob, NoSQL, Cassandra, MongoDB, Kubernetes, Snowflake, CircleCI, Airflow, Prefect, Google Data Studio, Azure Synapse Analytics, DynamoDB.
- **Tools:** Tableau, Power BI, Azure ML, RStudio, Jupyter Notebook, DBT, Databricks, IBM-Unica, SSIS, MS Office, JIRA, Looker.
- **Libraries:** Numpy, Pandas, Matplotlib, Seaborn, Scikit-Learn, Keras, Nltk, Gensim, Scipy, Beautiful Soup.