Shouvik Sharma

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

SUMMARY

A results-driven Data Engineer with a proven track record of designing and implementing scalable data solutions. Proficient in building infrastructure for data extraction, transformation, and loading using Python, SQL and GCP technologies, as well as developing analytics tools to provide actionable insights into key business metrics.

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.
- Designed and implemented optimal data pipeline architecture on GCP cloud services, assembling large, complex datasets to meet functional and non-functional business requirements. Resulted in a 20% improvement in data processing efficiency.
- 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 GCP Databricks extensively for data processing and analysis tasks, including building and optimizing Pyspark-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.

Data Engineer Intern at CNH Industrial Inc., Racine: **(Mar 2021 – Aug 2021)**

- Designed and implemented efficient ETL processes using Dataflow, optimizing SQL queries, collaborating with cross- functional teams, automating reporting systems, and conducting performance tuning, resulting in enhanced data management and accessibility.
- Collaborated with data scientists to define data requirements, implemented ETL processes for seamless data integration, optimized SQL queries, and facilitated the development of predictive models, fostering a synergistic environment for data-driven insights.
- Implement and utilize engineering best practices and methods to deploy and maintain quality, curated datasets using Airflow, enabling efficient data processing and management.

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.
- Developed and maintained data pipelines using Azure services resulting in a 40% increase in data processing speed.
- 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.
- Created Tableau dashboards to explain variation in success Metrics and Time Series Analysis to higher management using Big Query.
- Automated reporting process using Dataprep 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)**

- Involved in designing databases, data marts, E-R model for OLTP and multi-dimensional model for OLAP using SnowSQL.
- 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.
- Implemented data governance policies resulting in a 30% reduction in data quality issues.
- Formulated and implemented data governance policies, resulting in a substantial 30% reduction in data quality issues within the MariaDB environment.
- 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, Big Query.
- **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.