Subhranil Das

Chicago, IL | +1 (812)-671-5394 | dassubhranil1998@gmail.com | LinkedIn

Education

Indiana University, Bloomington **08/2022 - 05/2024**
Master of Science in Data Science
Coursework: *Machine Learning, Data Mining, Advanced Database Concepts, Data Visualization, Applied Algorithms, Deep Learning Systems, Applied Database Technologies, Statistics & Information Retrieval.*

Skills

Programming/Scripting: Python, SQL, R, Java, C, C++
Databases: MySQL, PostgreSQL, MongoDB, Cassandra, Snowflake, NoSQL, PL/SQL, AWS Redshift, Spark/SQL, SSMS
Frameworks: Spark, Hadoop, PyTorch, TensorFlow, Keras, Django, Flask - RestAPI
Data Management and Analytics tools: dBT, Apache Airflow, Kafka, Power BI, Tableau, Excel, Google Analytics
CI/CD and DevOps: Kubernetes, Docker, Linux, JIRA, Git, GitLab, Jenkins, Bitbucket
Cloud Platforms: AWS - S3, Glue, Sagemaker Kinesis, EMR, Athena, Lambda, Google Cloud Platform - Bucket, BigQuery

Experiences

Data Analyst **03/2024 - Present**
School of Public Health, Indiana University *Remote, USA*

- Developed efficient methods to extract and move large data files from **Excel** and **R** from various on premise systems to cloud storage, enabling Data Scientists to easily access, collaborate, and utilize the data.
- Collaborated with cross-functional teams to implement data quality checks and cleansing procedures using **AWS Glue** and **EMR**, integrated with **CI/CD** pipelines, resulting in a **15%** reduction in data-related errors and ensuring smooth workflow updates.
- Designed and implemented efficient data pipelines using **Apache Airflow** and **AWS Lambda**, reducing data processing time by **15%** and increasing processing throughput by **30%**.
- Developed a scalable data infrastructure using **Amazon RDS** and **S3**, accommodating a **50%** increase in data volume while maintaining consistent performance and reducing storage costs by **20%**.
- Integrated **AWS Kinesis** for real-time data streaming and analytics, improving the timeliness of decision-making by **30%** and enabling faster identification of trends and anomalies.
- Established data governance policies to ensure data integrity, security, and compliance with industry regulations such as **GDPR**, **HIPAA**, and **AES-256** achieving data robustness and reducing the risk of data breaches by **25%**.
- Collaborated with stakeholders to gather requirements and address ad-hoc requests, developing real-time analytics solutions and interactive **Tableau** dashboards, resulting in a **40%** increase in data-driven decision-making.
- Implemented automated data validation and error-handling mechanisms using **AWS Glue** and **Lambda functions**, improving data pipeline reliability by **25%** and reducing manual intervention by **30%**.

Solution Success Engineer **06/2021 - 07/2022**
eGain Corporation *Pune, India*

- Worked on data from various industries, including healthcare, finance, and retail, applying domain-specific insights and leveraging **SQL** and **Python** to design efficient data pipelines and ensure industry-compliant data processing and analysis.
- Collaborated with the engineering team to identify and fix bugs in **ETL** processes, while also analyzing pipeline failures and implementing enhancements that reduced **DAG** errors by 30%.
- Automated data monitoring and alerting using **Azure Monitor**, integrated with custom SQL queries and **stored procedures**, reducing issue resolution time by **25%** through faster failure detection, identification of bottlenecks, and immediate notifying relevant teams.
- Collaborated with cross-functional teams on **20+ projects**, gaining insights into the product cycle and support processes, while managing data infrastructure on **Azure** and **J2EE** platforms, effectively overseeing servers and applications.
- Developed and optimized SQL queries for data extraction and reporting, improving performance by 40%, while also **creating comprehensive documentation** and **training articles** for pipeline processes, which enhanced team knowledge and efficiency.
- Created interactive **Power BI** dashboards using **DAX**, integrating data from **Azure SQL Database** and **Azure Data Lake** with **Power Query** for transformation, enhancing data accessibility and decision-making.
- Utilized **JIRA** for task tracking and backlogs, adhering to the **Agile/Scrum** methodology.

Projects

Dynamic-Commentary | *Python, YouTube API, Pydub, ChatGPT API, NLTK, Data Visualization* **Link**

- Directed a project on tonal shifts and narrative strategies in esports commentary, enhancing stakeholder insights, and improving sentiment analysis accuracy by **20%** using **ChatGPT API** and **NLTK's VADER**.
- Created interactive **Python** visualizations to map sentiment plots, identifying key tonal shift moments of the commentators accompanied with that of the live audience chat with **90%** precision.

Turbocharge Retail Insights | *Python, Apache Airflow,dBT, Soda, Docker, GCP, Big Query, Metabase* **Link**

- Led the development of a comprehensive **Apache Airflow** ETL pipeline integrating **GCP Bucket**, **BigQuery**, **Soda**, and **dbt**, achieving **25%** enhancement in process efficiency.
- Optimized over **20** SQL scripts and dbt models, and implemented automated testing procedures (**DAGs**) for financial data ingestion, quality checks, and transformations, significantly improving data quality.

Generative AI for Pathology Datasets | *Python, Tensorflow, Scikit-learn, Keras, OpenCV* **Link**

- Directed a project on **Generative AI** for pathology datasets, utilizing **GANs** to synthesize nuclear detection datasets in medical imaging while addressing **HIPAA**-related data access restrictions to create synthetic data for model training.
- Utilized advanced GAN architectures such as **DCGAN**, **Variational Autoencoders**, and **StyleGAN3**, enhancing model performance by **40%**, and trained a **YOLOv8** architecture to achieve an impressive accuracy of **85%** in nuclei detection models.