Shubham Idekar

Boston, MA | idekar.s@northeastern.edu | linkedin.com/in/shubham-idekar-8231b0208/ | github.com/shubhamidekar?tab=repositories

TECHNICAL SKILLS

Programming Language: Python, SQL (MySQL, PostgreSQL, T-SQL)
Database: *RDBMS* – SQL Server, MySQL, Postgres, Azure Data Studio0; *NoSQL* - Cassandra, Azure Cosmos DB, ArangoDB
Cloud: AWS S3, Lambda, Azure Data Factory, Azure Synapse, Azure BLOB-
Big Data Tools: Apache Spark, Apache Airflow, Apache Kafka, Alteryx, Talend
Python libraries: Pandas, NumPy, Matplotlib, sklearn, TensorFlow, Keras, NLTK, selenium, BeautifulSoup, scipy, seaborn, statsmodel
Tools: Docker, Terraform (IaaS), Power Automate, Git, Agile Scrum

WORK EXPERIENCE

Data Engineer | Niraj Cement Structurals Limited | Mumbai, India **May 2021- July 2022**

- Designed and developed batch ingestion ETL pipelines using SSIS, Spark, Python scripts for data warehousing, reducing execution time by 24%
- Spearheaded initiatives to elevate data analytics for 5000 users, employing SQL query optimization and performance tuning techniques
- Implemented Python wrapper functions to streamline REST API interactions, enhancing data retrieval and integration from web services by 25%
- Monitored and optimized existing CI/CD pipelines in Agile Scrum setup, achieving 28% decrease in rollback incidents
- Crafted Tableau dashboards to visualize, boost KPIs tracking by 12%, implementing RLS, LoD for comparing metrics to foster decision-making
- Orchestrated transition from on-prem to Google cloud, employing Terraform, Kubernetes Cluster (GKE), reducing system downtime by 34%
- Collaborated with data science team to develop ML models leveraging TensorFlow to predict cement quality variations, leading to 17% improvement in product consistency

Business Analyst | Netscribes India Pvt Limited | Mumbai, India **Jan 2021- Apr 2021**

- Optimize data pipelines using ADF and Power Automate, streamlining reporting by 32% to enable real-time decision-making for stakeholders
- Led migration to centralized Azure Synapse data warehouse, integrate 15+ disparate sources using Spark enable enterprise-wide analytics
- Partnered with QA to develop 10+ comprehensive use case scenarios, guide UAT process and ensuring 96% alignment to functional requirements
- Conducted process mapping and gap analyses using Power BI to improve efficiency, implement KPI to boost customer satisfaction by 18%

Data Analyst | Tekman India Pvt Ltd | Thane, India **May 2019 - Sep 2019**

- Employed Excel and Power BI to create 12 dashboards and 25 ad hoc reports for periodic reviews, enabling managers to monitor deliverables and resulting in 23% increase in on-time project completion
- Analyzed multi-year data trends and identified cost-saving opportunities that contributed to 10% improvement in long-term profitability
- Executed Pl/SQL to implement CDC in Oracle, delivering solution to meet business requirements for 21% improvement in data processing
- Communicated data insights into clear and concise manner to technical and non-technical stakeholders for business initiatives

EDUCATION

Master of Science in Information Systems | Northeastern University, *Boston, MA* **May 2024**
Coursework: Data Science/Architecture, Designing Data Architecture for Business Intelligence, Data Management & Database Design **GPA : 3.8**
Bachelor of Engineering in Computer Engineering | Vidyalankar Institute of Technology, *Mumbai, India* **May 2021**
Coursework: Data Structures and Algorithms, Database Management Systems, Artificial Intelligence and Machine Learning **GPA: 3.5**

PROJECTS

Real-time Data Streaming *(Docker, Cassandra, PostgreSQL, Apache - Airflow, Kafka, Spark)* **Dec 2023 – Feb 2024**

- Leveraged Apache Airflow to orchestrate seamless data extraction of user data in JSON format employing API requests from web, resulting in 40% increase in data ingestion efficiency and ensured 99% uptime with distributed synchronization through Apache Zookeeper
- Organized Docker containerized setup and achieved 20% increase in data throughput efficiency using Apache Kafka for data streaming
- Leveraged Apache Spark within multi-tech stack including Cassandra and PostgreSQL, utilizing downstream Cassandra analytics

NYC Motor Vehicle Collision ETL *(Talend Studio, Alteryx, Azure SQL Database, Google Big Query)* *LINK* **May 2023 - Aug 2023**

- Fetched data from Google BigQuery, ensuring quality, integrity, and performed in-depth data profiling using Alteryx reducing anomalies by 30%
- Engineered dimensional data model through meticulous data cleaning, integration, and transformation with Talend Studio for ETL, loading dimension and fact tables holding 10 million rows of data into Azure SQL Database, visualized using Power BI and Tableau

E-commerce Data Engineering and Visualization *(Azure Data Factory, Power Automate, Cosmos DB, Power BI)* *LINK* **Jan 2023 - Apr 2023**

- Automated data fetching and refreshing using Power Automate flow to move over 2 million records from OneDrive to Azure Blob Storage
- Architected and executed 2 data transformation pipelines to migrate data to Cosmos Document DB using NoSQL API for JSON format through Azure Data Factory and using python scripts for Cosmos Graph DB using Gremlin API
- Crafted visually engaging and interactive dashboard for top 4 product categories as KPIs through e-commerce data in Power BI

Ecommerce Text Classification *(Python- NLTK, TensorFlow, sklearn, Jupyter Notebook)* *LINK* **Sep 2023 – Nov 2023**

- Implemented NLP-based data preprocessing with tokenization using Porter Stemmer library resulting in 20% reduction of irrelevant information
- Enhanced the Multinomial Naive Bayes baseline model by using TfIDF for advanced text vectorization, achieving 94% accuracy
- Applied CNN model with 5 layers, optimizing intricacies using ReLU for pattern recognition and softmax functions for enhanced performance, achieving accuracy 97% ensuring performance against overfitting risks