Sofia Dutta

Technical Leader, Software Engineer

(443) 554-4170 | sofia.dutta17@gmail.com | https://sofiadutta.github.io

WORK EXPERIENCE

Technical Leader @ NewWave Telecom & Technologies, Inc., Windsor Mill, MD, USA   **Jan 2021 – Present**

- Designed the complete system architecture, database schema, and data workflow for the Imersis data quality analytics platform. Built the project's cloud infrastructure setup from scratch.
- Built scripts to process hundreds of millions of healthcare records from the Centers for Medicare & Medicaid. Created pre-processing scripts for consuming large batches of unstructured customer data. Developed Apache PySpark code to compute data quality metrics for customer data.
- Created Looker dashboard visualizations with drill-down options that "explains" why the data quality came out to a certain value.
- Due to the lack of large quantities of data, re-designed the machine learning goal of the project into a data quality "Explainable AI" system.
- Built a data quality system for State governments that allows them to understand where to improve their upstream data ingestion processes and helped them observe how that improves their data quality over time.
- Reduced costs at four levels of the project:
  - Pre-processing: Built scripts that brought down the data pre-processing time from ten days to a couple of hours.
  - Uptime of cluster: Analyzed causes of high cloud expenditure and deployed Apache Airflow workflow management platform scripts to automate resource uptime only during hours of usage. Reduced cost from thousands of dollars to a couple hundred.
  - Storage v/s data transfer: Performed the cost advantage analysis of using a Google compute engine with large storage versus using more network data transfer.
  - Partially working in local servers: Built a system that handled pre-processing in our local data farm to reduce project costs from over ten thousand dollars to a few hundred in a month.
- Quickly learned new technologies like Apache Airflow, Databricks, and Google Cloud Platform and guided team members in their technology ramp-up for the past 2 years and helped them in setting up data workflows in the cloud. Provided educational expertise and mentoring to junior team members. Built product feature lists with stakeholders, conducted system design sessions with other architects on the team and led code review meetings.
- Investigated root causes for customer-found defects. Carried out several customer demonstrations to help sell the product and handled rapid prototyping and solution building for ad-hoc requirements and last-minute feature requests from the customer.
- Advised management, business, and technical staff on the usage of specific technologies like Apache Airflow and Google Cloud Platform.

Data Scientist Intern @ NewWave Telecom & Technologies, Inc., Windsor Mill, MD, USA   **May 2020 – Dec 2020**

- Successfully improved computation speed by 10-fold by deploying data analysis workflow in Google Cloud Platform (GCP) clusters and using Apache Spark for quality metrics computations.
- Collaborated with the Product Owner and other engineers in creating mechanisms for generating fake training data using Python programming to test out the efficacy of machine learning algorithms used in the project.
- Carried out necessary DevOps tasks for setting up Big Data Analytics environment by configuring GCP environment to execute Python programs and connected the cloud infrastructure with Looker's dashboards for delivering computed results to be presented to customers.

Software Engineer, Technical Leader @ Tata Consultancy Services, Kolkata, India   **Nov 2010 – Feb 2018**

- Led the design, development, and delivery of API interfaces using PL/SQL stored procedures for several projects of TCS.
- Carried out change based regression impact analysis, created software functional specifications, prepared test plans for several projects of TCS.
- Performed system integration, user-acceptance and performance testing and ensured client systems had very high uptime even when carrying out data migration activities. Saved millions of dollars in potential revenue lost to the client and was awarded for said effort by clients.

SKILLS

Coding languages	Python, Java, SQL, PL/SQL, T-SQL
Data Science tools	PyTorch, Sci-kit Learn, Apache Spark, Keras, Tensorflow, Hive, Hadoop, Looker, LookML, OpenCV
Tools	Databricks, Google Cloud Platform, Apache Airflow, Google Dataproc, Google Compute Engine, Google Cloud Storage, Google Cloud SQL, Google Big Query, AWS S3, Jupyter Notebook, Google Colab, Git
Back-end	Google BigQuery Table, Oracle Databases, PostgreSQL, Microsoft SQL Server, MongoDB, JSON
Domain knowledge	Big Data Analytics, Deep Learning, Machine Learning

EDUCATION

Udacity **Nanodegree**: Deep Learning		October 2021
Master's in Data Science @ University of Maryland, Baltimore County, Baltimore, MD, USA	GPA: 4.0	2019 – 2020
Bachelor's in Computer Science @ West Bengal University of Technology, Kolkata, India	GPA: 3.5	2006 – 2010