Manikaran Kathuria

+918285437110| kathuriamanikaran@gmail.com | linkedIn/manikaran-kathuria | github/Manikaran1996

WORK EXPERIENCE

SUPER HIGHWAY LABS (SHUTTL) | Data Engineer — Gurugram, IN | Feb 2021 – Present

- **Designing Data Lake**: The whole infrastructure of the organization is hosted on AWS. The goal of this project is to move all the teams to use Athena for querying the data which would be single source of truth and the company wide data present in AWS S3.
  - Data Ingestion
    - * **Scraping data from third party** Set up pipeline to download multiple dataset using third party API and store the data to data lake on AWS S3.
    - * **Ingesting business data stored in google sheets** Set up pipeline to download the data stored in google sheets, dump it to s3 and create Athena tables.
    - * **Migrating data from legacy MySQL instance to S3** Moved the data from MySQL database hosted on EC2 instance to S3 using AWS DMS.
  - File compaction
    - * Slowness in query result was observed while querying the data from Athena due to the presence of thousands of very small files in data lake.
    - * Wrote a generic spark job to compact the files stored in s3. Repartitioned the older data to lesser number of partitions to optimize the query performance
    - * Query time improved by more than **10x**
- **Reporting tool**: designed a generic configuration based reporting tool which queries the data via Athena and deliver the result to the concerned parties via email or slack notification.

ALPHONSO LABS | Technologist — Bengaluru, IN | Jul'19-Jan'21

- **Data Quality Project**: I have been a key player in improving the Data Quality in the Organization.
  - **Outlier Detection and Removal**: Aim was to find out the outliers in the data and remove/flag the data points.
    - * Performed several analysis to find out the potential issues and came up with a plan to flag the data in the production dataset.
    - * Leveraged time series decomposition algorithm (STL decomposition) to detect the anomaly in the data and notify the concerned team via slack notification.
    - * Presented the effect of removing/flagging the outliers and its consequence at different levels backed by the different statistics.
    - * Made use of databrick's delta format to flag the already existing data.
    - * Solved multiple challenges like the underlying data getting updated by multiple processes at the same time, firing jobs dependent on the outliers flag etc.
  - **Data Skewness** dealing with location and viewership based skewness in the data
    - * Derived threshold values to trim the data from some areas resulting in more balanced dataset.
    - * Flagged the production data not to be used when more balance data is required for analysis/reporting.
- **Data Insight Project**
  - **Data Monitoring** Daily monitoring of the majority datasets used in the Organization.
    - * Set up daily jobs using Airflow to generate the stats of several datasets used in the organization.
    - * Generated stats are stored in different datasets like InfluxDB, Elastic Search or HDFS.
    - * Generalized the code to generate stats of any dataset based on the configuration file.
  - **Google sheet based dashboard**: leveraged google sheets to share the state of the data to all the concerned teams
    - * Used google API to publish the stats on google sheets
    - * To detect the anomaly, came up with global thresholds for different metrics monitored.

        - Highlighted the row in the sheet having anomalous metric with different colors depending on the criticality of the metric.
    - Created a dashboard on Kibana for data visualization.
- I had managed multiple ETL jobs to process the incoming data from our clients and generate aggregated datasets which was used by multiple downstream processes.
- Was part of the critical discussions to make improvements in data processing, disk space usage, solving small file problems, releasing new version of data etc.

EDUCATION

M.Tech in Computer Science (GPA: 8.98) Delhi, IN | July 2019
Indian Institute of Technology, Delhi

B.Tech in Computer Science (90.78%) Delhi, IN | July 2017
Hansraj College, Delhi University

Class XII (95.6%) Delhi, IN | May 2013
Kendriya Vidyalaya

PROJECTS

GOVERNMENT AND CORPORATE INTERLOCK (THESIS PROJECT) Python, Pandas, Statsmodels, Web Scraping, MatplotLib

- Study of the CSR amount spent by the companies in India, performed regression analysis on the amount spent by the companies in different Lok Sabha constituencies
- Analysis of the interlock between the Government (Politicians/Bureaucrats) and the Corporate Sector
- Created a Neo4j Graph database containing information of Politicians, Bureaucrats, Board of Directors and Companies.

OBJECT DETECTION
Implemented Convolutional Neural Network using PyTorch Framework to predict the objects in the image

MNIST HANDWRITTEN DIGIT RECOGNITION
Implemented Neural Networks to recognize Handwritten digits from MNIST dataset

PUBLICATIONS

What Drives Location Preference for Corporate Social Responsibility (CSR) Investments in India? In Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '20). Association for Computing Machinery, New York, NY, USA, 296–300.

SKILLS

Languages: Python, Scala, Java, C++, Bash, C, SQL, Terraform
Technology: Git, AWS, Spark, Hadoop, Python Flask, InfluxDB, Android SDK, Elastic Search, Neo4j
Python Packages Numpy, Pandas, PyArrow, Matplotlib, ML Library Sklearn, boto3, Seaborn and statstools