NAISARGI DAVE

7743128988 • naisargidave29@gmail.com • www.linkedin.com/in/naisargidave • http://naisargidave.github.io

SUMMARY

Data Engineer with experience in developing, optimizing complex data pipelines and looking for a role with data engineering/data science background.

EDUCATION

Worcester Polytechnic Institute (WPI), Worcester, MA
MS in Data Science, GPA 3.90/4.0 — Aug 2019 - May 2021

Mukesh Patel School of Technology Management and Engineering, NMIMS University, Mumbai, India
B.Tech in Electronics and Telecommunication Engineering, GPA 3.13/4 — Aug 2013 - May 2017

SKILLS

Programming Languages: SQL, Python, R, HIVEQL, Scala, Java, HTML, JavaScript
Frameworks: AWS, PyTorch, TensorFlow, Keras, MySQL, SQLite, SAP-HANA, HADOOP, MapReduce, HDFS, Hive, PIG, Spark, MongoDB, Scikitlearn, ReactJS.
Areas: Data Science, Data Analytics, Business Analytics, Data Visualization, Machine Learning, Statistics, Artificial Intelligence, Big Data, Database Management, Data Mining
Tools: DataNet, DataCraft, Cradle, Redshift, Andes, S3, Glue, Athena, Qlik Sense, Tableau, Anaconda, SharePoint Online, IBM Modeler, Git, JIRA, MATLAB, VISUAL STUDIO, MS O365 Apps (Planner, Forms, PowerApps, Teams, Excel, Word, Powerpoint)

INDUSTRIAL EXPERIENCE

Data Engineer I, **Amazon**, Seattle, WA, USA — Jun 2021 – Present

- **Inflation Pipeline:** Worked with economists to build a cascading multi-stage Fisher Index pipeline based on **Cradle, SQL** to analyze the month over month glance view weighted price inflation trends at Amazon. Created **Redshift** tables to store the staging data and the final result. Received the Pathfinder award for finding solutions that impact and change business processes.
- **Net Price Competitiveness(NPC) pipeline:** Built a pipeline that tracks Amazon's price competitiveness by incorporating newly launched promotions using **DataNet** Extract and Load Jobs. Created **Andes** and **Redshift** tables to store the resultant data and backfilled it. Generated automated weekly reports using **SQL Metric Jobs** to highlight the issues in our pricing systems and aid decision making.
- **Price Consistency Metric:** To ensure similar products (e.g. two same tshirts, differing only in color) are priced the same, built a **DataNet** and **Redshift** based metric pipeline, to report the price consistency of such products at Amazon.
- **Data Storage Cost Optimization:** Analyzing data stored in **Andes** tables and **EDX files** to identify redundant, extraneous data to be deprecated to reduce storage costs. Projected expense saved - 30%.
- **Pricing Evaluations Data Pipeline:** Read nested **JSON** data from **DynamoDB stream** using **DataCraft** and converted it to TSV format by defining complex **SDL schema** in Cradle to **replace unscalable and expensive** DynamoDB scans.
- **List Price Update:** Created a pipeline to nudge vendors to update the price of their products on Amazon marketplace. Implemented the logic to identify the products that need price update using **SQL** in **Cradle** with output data written to **S3 buckets**. Created a **Glue** database to store and query the output data and attached schema using **Crawler**. Performed validations and data analysis using **Athena**.

Data Visualization Co-Op, **AbbVie**, Worcester, MA, USA — May 2020 – Nov 2020

- **Change Over Application:** Developed an application using **SharePoint** and **PowerApps** with the Quality Assurance team for capturing the changeover details of equipment used in drug manufacturing processes.
- Dashboards:
  - **Talent Dashboard:** Represented AbbVie employee details and the **acquisition, attrition rates** using visualizations in a **Qlik Sense** Dashboard for the Strategic Operations team to monitor. Implemented data masking script using **Python** to handle sensitive data.
  - **Training Dashboard:** Created a dashboard for supervisors and employees across multiple departments to **track and highlight upcoming and past due requirements** to ensure compliance.

Developer, Reliance Industries Ltd., Mumbai, India Aug 2017 – Jun 2019

CIO Dashboard:

- Created a dashboard for the **CIO of Reliance** to monitor performance of the teams.
- Integrated data from **flat files, SQL tables and SAP-HANA**, modeled it and used **Tableau** to visualize the Key Performance Indicators (KPIs).

Text Classification:

- **Automated the classification of user queries** entered in the Grievance Redressal Portal of Reliance by clustering and classifying them using natural language processing techniques, **Naive Bayes and Support Vector Machine classifiers**.
- **Improved the efficiency** of the team by approximately **30%** by automating the old manual process.

Team Ranking Scorecard:

- Developed a scorecard to track the performance of several teams and rank them using 17 KPIs.
- Performed **statistical modeling in Python** and developed **visualizations using Tableau**.

Procurement Spend Analysis:

- Performed **complex data modeling** for procurement spend analysis of the organization using **Hive on Hadoop - MapReduce** framework. The results of the analysis were visualized using **Zoomdata**.

ACADEMIC PROJECTS

Space Missions Visualization, Worcester Polytechnic Institute**,** USA Aug 2020 – Dec 2020

- Created visualizations using d3.js and react to analyze the details of space missions launched since 1957.
- Implemented techniques such as highlighting, brushing, and filtering to make the visualizations interactive.

Melanoma Classification, Kaggle Jun 2020 – Aug 2020

- **Detected melanoma** among images of benign and malignant skin lesions from the SIIM ISIC Melanoma Challenge Dataset.
- Used **EfficientNet** for feature extraction, incorporated metadata, performed **data augmentation**, and image preprocessing to **crop out regions of interest** from the images. Achieved **80%** accuracy.

Human Protein Classification, Kaggle Jun 2020 – Aug 2020

- Performed **multi-class classification** to identify all the types of proteins present in the cell images from Human Protein Classification dataset.
- Performed data pre-processing and data augmentation, used **Transfer Learning** with **pre-trained ResNet50** model to make predictions.

Car Review Analysis, Worcester Polytechnic Institute**,** USA Jan 2020 – May 2020

- Created a **search engine** to fetch the most relevant reviews to the given user query using **BM25**.
- Performed **topic modeling using LDA**.
- Generated more accurate user ratings based on **Vader sentiment analysis** of the user reviews.

Twitter Data Analysis, Worcester Polytechnic Institute**,** USA Jan 2020 – May 2020

- **Performed data mining of Twitter data** using twitter API and tweepy library for tweets containing the keywords “Donald Trump” and “Joe Biden”.
- Performed **exploratory and sentiment analysis** to figure out the general sentiment for each candidate and **predict the winner** of the 2020 presidential elections.

Travel Itinerary Application, Worcester Polytechnic Institute**,** USA Jan 2020 – May 2020

- Developed an **Entity Relationship Model** and a **Database** using **SQLite** to capture the customer and travel details such as preferred transport, hotel, restaurant and tourist attraction.
- Created an **Android application** to allow users to **plan the trip and view their itinerary**.

PageRank, Worcester Polytechnic Institute**,** USA Jan 2020 – Mar 2020

- **Crawled web pages** using BeautifulSoup, performed **text preprocessing using natural language processing** (NLP) techniques.
- Implemented **PageRank** algorithm to rank the webpages.

Skewed Join Optimization, Worcester Polytechnic Institute**,** USA Aug 2019 – Dec 2019

- Used **Apache Spark and Scala** to optimize the join operation between two **large data sets** (13M and 0.1M records) with one of them **skewed**.

Human physical activity recognition model, Worcester Polytechnic Institute**,** USA Aug 2019 – Dec 2019

- Performed **feature extraction** and developed a **human physical activity recognition model** for predicting the activity performed based on the person’s movement data
- Evaluated and compared several statistical learning methods including **Logistic Regression, Random Forest and Support Vector Machine**.