JHANVI ARORA

+1 (514) 553-9507 | jhanviarora509@gmail.com

SUMMARY

I am a dedicated and knowledgeable professional in the field of Big Data and Data Engineering, boasting 3 years of industry experience. I am actively seeking a challenging role in a dynamic organization where I can apply my acquired skills and expertise to contribute to the growth and success of the company. My background includes proficiency in data science, big data technologies, and workflow automation. I have cross-functional expertise in ETL processes, database management, data warehousing, and BI reporting, which allows me to develop impactful business solutions. With strong attention to detail, a drive for innovation, and a commitment to delivering exceptional results, I approach every use-case with unwavering passion and dedication.

TECHNICAL SKILLS

Data Technologies	Hadoop HDFS, MapReduce, Spark-Streaming, Kafka, Hive, Sqoop, Apache Flume, Spark Scala, Airflow
Programming Languages	Python, Shell, Java, SQL, Spark, C, C++, MATLAB (basic), JavaScript (basic), Bash
ML Libraries	PyTorch, TensorFlow, Keras, Scikit-Learn, Pandas, NumPy
Micro Services	Docker, Kubernetes
Databases (SQL & NoSQL Systems)	Oracle, MongoDB, Dynamo DB, Couchbase
Operating Systems	Linux, Mac, Windows.
Visualization and Computational Tools	Microsoft Power BI, Tableau
Cloud Technologies	Microsoft Azure, GCP
IDE	PyCharm, Github, SQL Developer, Bitbucket, IntelliJ, Jupyter, Colab

RELEVANT WORK EXPERIENCE

Data Scientist, Duca Financial   **Aug 2021 – Present**

- Managed large financial datasets and perform tasks to store, analyze, and process data in Hadoop using big data technologies.
- Provided insights and KPIs using Business Intelligence (BI) reporting tools like Power BI, complying with internal needs and stakeholder requirements.
- Developed ETL data pipelines using Azure Data Factory (ADF) and integration runtime to connect with different Azure resources like storage accounts, Azure Data Lake (ADL), etc.
- Restructured Azure Data Flows for transformation and created Databricks notebook(s) using Python, and PySpark, and assisted in creating supervised and unsupervised Machine Learning pipelines.
- Worked on data visualization libraries such as ggplot, Plotly, matplotlib, and Seaborn to visualize complex data and analytical results.
- Handled large-scale datasets in Azure data warehouse, SQL database performing DDL, DML, and DCL operations.
- Managed to save about 10% full-time effort because of process improvement activity.
- Implemented Spark using Scala and Spark SQL for faster testing and processing of data.
- Worked on Spark code in Scala language. Such as Streaming data from the cluster using Spark streaming.
- Implemented NumPy, Scikit-Learn and TensorFlow library's functions for training, testing, and validation of data and increase the significant percentage of accuracy in projects.

- Familiar with agile methodology working in sprints for business stakeholders from different time zones.
- Created and updated technical documentation of flow and working process in the project.

Data Analyst, CSI Consulting   **July 2020 - Jul 2021**

Projects/Roles:
1. **Data Science Developer:**
   - Worked in a start-up environment, wrote SQL queries to create custom views for key business metrics, Build and publish Tableau dashboards with relevant business metrics, integrated with Apache Airflow for automated updates and data-driven decision-making.
   - Designed and led the development of a Health Assistant Bot using RASA stack and trained BERT- that takes in health history and advises users on preliminary procedures of treatment.
   - Identified new problem areas in the bot, integrated voice support via Google Cloud API, and deployed it on the GCP cloud platform.
   - Applied statistical methods and data visualization techniques to present findings and trends, aiding in informed decision-making processes.
   - Conducted thorough data validation and quality checks, ensuring accuracy and reliability in analytical outputs.
   - Collaborated with cross-functional teams to understand business requirements and translate them into actionable data analysis tasks.
   - Actively participated in data governance initiatives, ensuring data integrity, security, and compliance with regulations.
2. **AI Research & Development Engineer:**
   - Assisted the Company research in Legal Automation using Natural Language Processing devised recommendation systems using topic embedding and vector space approaches.
   - The proposed approach for finding precedents achieved a high Precision @ 10 score of 0.1 and the statute mapping algorithm a BPREF score of 0.2812.
3. **Chatbot Developer:**
   - Researched deployment feasible technology and use case suiting implementation practices- amongst existing tech stack of bots – RASA, Botpress, Dialogflow.
   - Developed Initial Prototype for rigorous testing alongside setting up a continuous integration and development pipeline.
   - Accumulated bulk conversational data through end-user testing of the prototype.

EDUCATION

Master Of Applied Computer Science
CONCORDIA UNIVERSITY

Bachelor of Engineering in Information and Communication Technology
PANDIT DEENDAYAL PETROLEUM UNIVERSITY

PROJECTS

CUSTOMER RECORD HANDLING OF E-COMMERCE WEBSITE
- Pre-processed the incoming data, such that it preserved and extracted only meaningful information.
- Designed a storage Back-end containerized in Docker, running on Amazon EC2 Server utilizing the Couchbase no-schema technology to store the huge amount of batch data.
- Tested and reviewed if the system retains the properties of a distributed system such as Fault Tolerance, Concurrency, etc.
- Designed a driver that interacts with the database and executes queries in the form of various transactions.

PERSON RETRIEVAL BASED ON COGNIZABLE HUMAN DESCRIPTIONS

- Built using an amalgamation of Computer Vision and NLP-based approaches- the intended system takes in verbal user-described attributes to identify a person in given video surveillance footage.
- Constructed the pipeline of converting speech to text, identifying CRFs from text, and mapping the extracted attributes to video frames with an accuracy of 94.8%.
- Utilized Mask RCNN for semantic segmentation and identification of subjects in training data frames of CCTV footage, fine-tuned Efficient Net – B5 for each soft biometric targeted, e.g., torso color, gender.
- The achieved results were 10% higher than the reviewed method results.

PNEUMONIA CLASSIFICATION
- Set on the task of pneumonia classification- achieved an F1-Score of 0.92 using a training set of only 10 images.
- Researched and reported performance comparisons of various suiting deep learning techniques such as– meta learning and ensemble model approaches.

PUBLICATIONS

- Arora, J., Pandya, U., Shah, S., & Doshi, N. (2019). Survey-Pollution monitoring using IoT. Procedia Computer Science, 155, 710-715. Orally presented research at FNC 2019 conference proceedings held in Halifax, Canada
- Arora, J., & Bharti, S. K. (2021). Rhetorical Analysis and Classification of Poem Text. International Journal of Semiotics and Visual Rhetoric (IJSVR), 5(1), 57-71. doi:10.4018/IJSVR.2021010105
- Arora, J., Patankar, T., Shah, A., & Joshi, S. (2020). Artificial Intelligence as Legal Research Assistant. In FIRE (Working Notes) (pp. 60-65). Orally presented research at Forum of Information Retrieval – 2020.