Dharmendra Kanjaria

Linkedin: https://www.linkedin.com/in/djkanjariya//
Github: https://github.com/idjey

Address : 4820 Carmella Dr, Arbutus, MD 21227
Email : djkanjaria@gmail.com
Mobile : 727-688-8450

TECHNICAL SKILLS

Programming Languages	Python, Scala, Java, R
Python ML Libraries	Numpy, Pandas, Scikit-Learn, Apache Spark MLlib, SparkML
Software / Cloud	PySpark, Hadoop, (MapReduce), Airflow, Kafka, AWS, Azure, Linux, REHL, Shell
Database	SQL, PostgreSQL, Redshift MySQL, Cassandra
Machine Learning	Decision Tree, KMeans Clustering, Classification, Self Organizing Maps
Data Visualization	Tableau, Power BI, Matplotlib, Seaborn

WORK EXPERIENCE (5 YEARS)

1. **Data Analyst - Genius Infotech** June 2011 - February 2016
   - Developed machine learning algorithm for diamond shape suggestion using KMeans Clustering and Classification.
   - Closely worked and collaborated with the team of machine learning engineers and implemented Auto shape suggestion rough diamond cutting tool, used **python**, **Apache Spark**
   - Achieved Cutting accuracy up-to **56%** which helped our client to reduce wastage during cutting process
   - Worked on **SQL query optimization** using various methods to **improve database** performance.
   - Worked on **ETL data pipeline** using **Informatica ETL** tool to load large volume of image data from multiple sources to data warehouse

EDUCATION

M.S in Information Systems, University of Maryland Baltimore County (UMBC), USA 2017 - 2019

B.S in Electrical Engineering, LukhdhirJi Engineering College (LEC), INDIA 2007 - 2011

PROJECTS

1. **Big data on Apache Spark, Machine learning, Python and MLlib (Complete)**
   - Worked on large scale dataset of energy company to detect anomalies by using clustering algorithm using PySpark, Kafka, MLlib, Scikit-learn, Python, pandas and Jupyter environment
   - Used **Hortonworks HDP** to gain hands-on experience on various big data components like Hadoop,Spark, Hive, HBase, Hue, used Parquet, Avro and CSV file format
2. **Udacity Data Engineering project**
   - Worked on Data engineering project using music streaming Dataset (**JSON**)1 million entries, used PostgreSQL and Apache Cassandra NoSQL database
   - Built out an **ETL pipeline** to **optimize queries** in order to understand what songs users listens most
   - Created a **NoSQL** database using **Apache Cassandra** (both locally and with docker container)
   - Developed denormalized tables and optimized for a specific set queries and business needs
   - Created a **data warehouse** using **Amazon Redshift**, Develop an ETL Pipeline that copies data from **S3 buckets** into staging tables to be processed into a star schema
   - Developed a star schema with optimization to specific queries required by the data analytic team
   - Scaled up the current **ETL pipeline** by moving the data warehouse to a data lake
   - Created an EMR Hadoop Cluster, Further develop the ETL Pipeline copying datasets from S3 buckets
   - Data processing using Spark and writing to S3 buckets using efficient partitioning and parquet format
   - Writing custom operators in Airflow to perform tasks such as staging data, filling the data warehouse, and validation through data quality checks