RAKESH DASH

Data Engineer

+917008339812 | rakeshdash6180@gmail.com | 48 Sashank Nivas, Vagadevi Layout 2nd Cross,Munekolala, BENGALURU, 560037, India

http://www.linkdin.com/in/rakeshdash007

ABOUT ME

Professional Summary Experienced **Python Developer** with **7 years in IT**, specializing in **automation, database engineering, and data pipeline optimization**.

Strong expertise in Python **development, data transformation, and workflow automation**.

- **5+ years of experience in Automating pipeline development & optimization.**
- **3+ years in PySpark**, handling **large-scale data transformations**
- **1+ year in Scala**, applied in **data processing & transformation.**
- **Proficient in SQL, schema modeling, and query optimization.**
- **Strong focus on performance tuning, data integrity, and scalability**

Passionate about **building efficient, automated, and scalable data solutions** that enhance business insights and streamline data workflows.

WORK EXPERIENCE

Data Engineer | Hsbc Electronic Data Processing India | Bangalore | Jul 2023 - Present

Skills Acquired:

- Python, PySpark, Hadoop, Git, GCP
- Data Transformation, ETL Logic Building
- Beginner-level Scala Development

Projects:

1. Data Quality Check Package

- **Technologies Used:** OOP Python, PySpark
- **Description:** Developed a comprehensive data quality check package to ensure data integrity across assets.
- **Key Features:**
- **Data Completeness Check** – Ensures that all required fields contain valid data without missing values.
- **Coverage Check** – Verifies the extent of data availability across different sources.
- **Derived Field Validation** – Assesses the accuracy of computed fields, including GAR field ratio checks, by ensuring that values remain within acceptable decimal limits.
- **Null Values Check** – Identifies and flags missing or incomplete data points.
- **Inconsistency Identification** – Detects discrepancies in data formatting, relationships, and expected values.
- **Results Output:**
- Generates structured reports in both CSV format and an Excel workbook.
- The Excel workbook includes a **summary results sheet** containing:
- test_case_id, expected_output, actual_output, exception_found, status (Fail/Pass), and exception_id (if applicable).
- If exceptions are found, **additional sheets** are generated to document exception details for different test cases.

2. Streamlined Data Processing Pipeline

- **Technologies Used:** Python, Pandas, openpyxl
- **Description:** Developed a comprehensive master source output for HSBC vendor data.
- **Key Contributions:**
- Consolidated and transformed vendor data to align with business requirements.
- Worked closely with the vendor management team to ensure data accuracy and completeness.
- Optimized data processing workflows for enhanced efficiency.

3. ESG Data Pipeline Development

- **Technologies Used:** PySpark, Python, Spark SQL
- **Description:** Integrated vendor data from MSCI and SBTi while ensuring data integrity and adherence to industry standards.
- **Key Features:**
- Implemented a **scalable workflow** using a DAG scheduler.
- Conducted rigorous **testing** to validate data transformations.
- Collaborated with cross-functional teams and documented the end-to-end process.

4. Automated Completeness Check

- **Technologies Used:** Python (OOP), CI/CD (Jenkins, Airflow DAGs)
- **Description:** Automated the completeness check process for vendor data across **27 vendors** by developing a unified ETL pipeline.
- **Key Features:**
- Standardized the ETL process to ensure seamless **integration and execution**.
- Automated **data validation and reporting**, reducing manual intervention.
- Implemented robust **error-handling mechanisms** to identify and log discrepancies.

Data Analyst | ZETA GLOBAL | Hyderabad | Jul 2021 - May 2023

Client: Christmas Tree Shop

Fuzzy Matching Algorithm & ETL Pipeline Development

- Designed a customer transaction data model, improving entity resolution. - Developed Python-based ETL workflows to extract, transform, and load structured and unstructured data into Snowflake. - Built a fuzzy matching algorithmic package leveraging NLP and probabilistic matching techniques in PySpark. - Achieved a 20% revenue increase in Q3 2022 by enhancing data accuracy and reducing duplicate records.

Client: Charter
Skills Used: Python, Snowflake Connection, SQL

- Developed Snowflake database schema modifications to improve storage and query efficiency. - Automated data migration scripts for handling millions of records while maintaining referential integrity. - Optimized indexing strategies and query execution plans, improving report generation speed by 50%.

Client: Ralph Lauren
Skills Used: SQL, Python

- Developed customized data models to meet client's unique requirements, enabling efficient data retrieval and analysis.
- Analyzed complex queries and performed data modeling for Ralph Lauren, catering to ad-hoc requests from APAC, Europe, and North America customers.
- Leveraged Oracle SQL for efficient data extraction, manipulation, and aggregation.

Decile Column Transformation:

- Leveraged Python and Snowflake connection to update and alter column values in the database based on the analysis findings.
- Developed efficient scripts to modify and update column values for millions of records, ensuring data integrity and consistency.
- Collaborated closely with the database team to optimize the performance of data transformation operations and ensure minimal impact on the database infrastructure.
- Implemented data quality checks and validations during the column value changes to maintain data accuracy and reliability.
- Documented the changes made to the database schema and provided clear documentation for future reference and auditing.

Client: AMCN, CNN
Skills Used: Tableau

- Developed and maintained Tableau dashboards for AMCN client, incorporating data extraction from SnowSQL server, Google Analytics, and Google Ads.
- Utilized Tableau's data visualization capabilities to create interactive and dynamic dashboards, showcasing weekly performance metrics and insights from multiple data sources.
- Conducted regular data extraction, transformation, and loading (ETL) tasks from SnowSQL server, Google Analytics, and Google Ads to ensure up-to-date and accurate data visualization in Tableau dashboards, enabling data-driven decision-making for the client.

Analyst | Cyient LTD | Hyderabad | Apr 2018 - Jul 2021

Project-Scheduled Interruption:

Skills used: Python, SQL complex queries, database management (T-SQL)

- Automated the process flow of data extraction from XML and HTML files into CSV format using Python scripts.
- Utilized Python's built-in libraries such as xml.etree.ElementTree and BeautifulSoup to extract data from XML and HTML files, and pandas library for data manipulation and conversion to CSV format.
- Developed efficient and scalable Python scripts to automate the extraction process, ensuring accurate and reliable data extraction from XML and HTML files into CSV format, saving time and effort in manual data extraction tasks.

Logbooks and Removals:
Skills used: Python, SQL complex queries, Power BI

- Extracted data from Teradata database using Power BI, leveraging Power BI's native connectors and query capabilities to retrieve relevant data for weekly, monthly, and yearly trend analysis.
- Designed and implemented Key Performance Indicators (KPIs) in Power BI to monitor and analyze weekly, monthly, and yearly trends, using measures and calculated fields to derive meaningful insights from the extracted data.
- Developed visually appealing and interactive Power BI reports and dashboards to present the KPIs and trends in a clear and concise manner, enabling data-driven decision-making and performance tracking at different time intervals.

Aircraft Identification and Statistics:
Skills used: Python, SQL complex queries, Power BI

- Extracted data from Teradata, manipulated and modeled it using Power BI's data modeling capabilities, and visualized the insights using interactive reports and dashboards.
- Leveraged Python for additional data manipulation, advanced analytics, and machine learning modeling to enhance the data analysis and visualization capabilities in Power BI.
- Combined Teradata, Power BI, and Python to create a comprehensive end-to-end data analysis and visualization solution, driving data-driven decision-making with meaningful insights and actionable recommendations.

EDUCATION

Bachelor, Institution of engineers (India), Bhubaneswar | 2016

Professional Accreditation

SKILLS

Python SQL Data Transformation ETL PySpark Data Warehouse Data Mining DataBricks Cloud (GCP,Microsoft Azure) Git

LANGUAGES

English Hindi

COURSES

Analysing Data with Python | EDX | Nov 2019

Structuring Machine Learning Projects | deeplearning.ai | Feb 2020

Design Databases With PostgreSQL | Code Academy | Jan 2021

How to Analyze Business Metrics with SQL Course | Code Academy | Oct 2020

INTERESTS

Data Architect

Python Data Engineer/Automation Specialist

PySpark Data Engineer

Python Developer