AISHWARYA THM
Data Engineer

Bengaluru, India • 8762397506 • aishuthm95@gmail.com • linkedin.com/in/Aishwarya-thm/

"Data-driven data engineer with 7 years of experience working with Python, Apache Spark, Databricks, Apache NiFi, ETL, Azure cloud and AWS. My focus on data integrity and cross-functional integration across all company departments has consistently boosted efficiency by >75%, optimized data pipelines and data migration to Databricks. Looking forward to bringing my skills in management and data analytics to develop and deliver scalable solutions to meet growing data handling requirements."

PROFESSIONAL EXPERIENCE

Cargill CFB, Camsdata Technologies Dec'23 – Present
Technical Lead Data Engineer in Master Data Governance

- Working solely on the entire master data governance in identifying and addressing any data inconsistencies to ensure accuracy and reliability in the data extracted from multiple ERP systems like SAP TC1, SAP TCC, etc.
- Identifying and removing duplicate SKUs within the data to prevent errors in analysis and reporting and also to maintain a unique product code throughout the system.
- Mapping the product hierarchy to ensure consistency and alignment across systems. Aligning the product hierarchy used in the ERP system with the global hierarchy used in TC2 for compatibility and consistency.
- Loading cleaned and mapped data into LEAP and CDP to maintain a central repository for all the ERP systems for the reference and avoiding the ambiguous hierarchy mapping.
- Developed a commercial Power BI dashboard by connecting it to database and projected the data analysis of FY24.
- Enforcing the data governance rules for the new SKUs and creating new product lines to map the main products of Cargill Food and Beverages across APAC.
- Established standardized data definitions and formats for the master data table to have a single, reliable source of data across APAC.

HP Research and Development Center Bengaluru, HCL Technologies Oct'2021 – Dec'23
Technical Lead - Data Engineer (Business Project)

- Developed 25+ key **data pipelines** on AWS for the production and staging EC2 instances to process over400 TB of data by consolidating data from multiple disparate sources into a single destination, enabling quick data analysis for reliable business insights using **AWS EMR, Data Pipelines, Apache NIFI and Amazon Redshift.**
- Evaluated current architecture and designed the data flow that allowed horizontal unlimited scaling of Hadoop clusters on EMR.
- Designed and implemented a **REST API** External Calls Handler to facilitate the generation of monthly and quarterly retailer and distributor compensation payment summary reports. This integration seamlessly connected the Java frontend code, enabling real-time report refreshment on the portal.
- Experience in building **ETL pipelines using Databricks, Azure Data factory.**
- Reduced processing time by **63**% by developing a **Pyspark** script in Phase2 privacy implementation of project that efficiently compares the previous and current **snapshots** of the **delta table** and updates only changed or appended records in **Databricks**.
- Improved data refresh efficiency in Phase1 implementation of project by automating the comparison and update process of data post refresh in **Pyspark** and **SQL** and unloads only modified records in **Amazon S3** thereby reducing memory savings by **30%,** manual effort by 50%.
- Solely Designed and implemented a new ETL flow and database tables for the launchof toner products to small and medium businesses through the commercial channel. This provided asolution for remote home office workers to manage printing supplies for lots of individuals in separate locations versus a shared office space.
- Cut projected time for data generation into backend tables by one week by developing reusable ETL components and building multiple data flows.
- Knowledge of Azure SQL Database, Data Factory, Databricks, PowerBI.

- Worked on various Python libraries like Cryptography, Pandas, Numpy, Boto3, relativedelta, datetime, etc.
- Developed and operated a POC project in analyzing the production logs using the Azure Data Factory pipeline.
- Migrated the **HP Ink's Instant Paper and SMB Business Compensation Model** projects to Data bricks connecting to AWS S3, delta tables and Redshift.
- Significantly contributed to the seamless migration of tables from **Redshift Spectrum to a Redshift schema**, resulting in substantial improvements in query execution times and memory efficiency.
- Successfully led the migration of processes from Apache NiFi to Azure Databricks as part of the comprehensive Retailer Compensation project in a duration of 45 days managing a team of 13 people.
- Learnt and developed a **Microsoft Power BI dashboard** for the Retailer Compensation Project within the span of 6 days. The dashboard facilitated the approval process of reports by Data Analysts. Improved data analysis post-refresh and resulted in significant time savings in monthly report delivery.
- Developed 5 automation scripts using **Python and Sql** to halt the data refresh of already approved reports, effectively resolving major escalations from country managers.
- Implemented customized **Power BI reports** for clients, resulting in a 40% increase in data analysis efficiency.
- Developed bash scripts for executing the pyspark scripts automatically reducing the manual effort by 16 hours and 4 resources.

- Successfully identified and resolved a **critical escalation issue** at Dixons GB, mitigating a potential $7.5 million losses. This issue was attributed to exceptions in a native PySpark script that had persisted for fiveyears.
- Mentored two interns and three resources from client side technically for the design and implementation of a project to execute a country-wide experience that promotes an Instant Ink first approach across the entire shopper journey lifecycle which drove the adoption rate to 65%. Acquired proficiency in web scraping techniques using **BeautifulSoup (BS4)** to extract data for subsequent analysis and data-driven insights.

OTE, Accenture,Jul'2019 –Oct' 2021

Python Developer (Operations Team)

- Accurately wrote more than 100 Python and bash scripts to automate the ETL scripts runs every hour.
- Developed ETL scripts in Python to get data from one database table and update the resultant to another.
- Automated themanual data entry into the CRM portals from the input csv files using Python.
- Acted as Scrum Master with the experience of over 6 years for issue tracking systems, preferably Jira.
- Excellent problem-solving skills, in particular a methodical approach to dealing with problems across distributed systems.

Telia, Accenture Jun'2017 – Jul'2019

Python Developer (Oracle Product Hub Team)

- Designed and implemented a centralized website for the project using Java, Web Development and Oracle SQL to integrate the design, development, and testing teams thus successfully reducing the amount of search time of data over multiple sources.
- Automated the data migration between Oracle Product Hub, CRM and Billing systems using python and selenium increasing the average sales rate to 53%.
- Designed merchandising strategies for telecom retailers identifying key store attributes to increase revenue and conversion, resulting in a 28% sales lift.
- Designed, executed, and implemented the data flow process using python, SQL, Selenium, and jQuery for automated data entry into the CRM portal from the stakeholder requirements, resulting in tremendous timesaving from 3 days to 2 hours 54 minutes.
- Developed a Microsoft Access Tool for the data validation and export of xml files from the CRM system using java, Apache Tomcat server, SQL, and Eclipse reducing manual efforts from 4 resources and 4 hours to 16 minutes.

Personal Projects:

- **Azure End-To-End Data Engineering Project - Olympic Data Analytics**: Analyzed Tokyo 2021 Olympic data utilizing a suite of tools and technologies, including **Azure Data Factory, Data Lake Gen 2, Synapse Analytics, and Azure Databricks.**

- **Web-scraping using Python**: Implemented automated data extraction of gold, silver, and platinum rates from various jewelry websites, streamlining the process of sending WhatsApp messages daily using **python libraries and regex**. Developed a system to send messages and save data in excel format for further analysis using **Power BI**.

- **Twitter Data Pipeline using Airflow:** Created End-To-End Data Engineering Project using **Airflow** and **Python**. In this project, we will extract data using Twitter API, use **python** to transform data, deploy the code on **Airflow/EC2** and save the result on **Amazon S3**.

EDUCATION

- **10th - 10 CGPA**, DAV Public School, Central Board of Secondary Education, 2011.
- **12th – 91%**, National Pre-University College, 2013
- **Graduation – 8.67 CGPA,** B.E, B.V.B College of Engineering & Technology, 2017.

SKILLS & OTHER

Database: Oracle SQL Developer, AWS Redshift,Azure SQL Database.
Programming Languages: Python, PySpark, Java, SQL, Responsive Web Development including HTML, CSS, JavaScript, jQuery.
Shell Scripting: Linux.
Version Control: GitHub, Bitbucket
Integrated Platforms: Apache Nifi, Mobaxterm, Postman API, Azure Databricks, SSMS.
AWS: Amazon S3, EC2 Instances, Amazon Elastic MapReduce (EMR), Data Pipelines
Azure: Azure Data Factory, Azure Databricks, Data Lake Storage, Azure SQL Database, Power BI, Synapse Analytics.
Cloud: AWS, Microsoft Azure
Certifications completed: Amazon Web Services Professional, Microsoft Certified Azure Data Engineer DP-203 course by Udemy, Microsoft AZ-900 Fundamentals, Responsive WebDesign, JNCIA - 2016, Oracle Cloud Infrastructure Foundations Associate, Python Master Class.

Total Experience: 7 Years.