HARSHITA PATHAK

Data Engineer | Genpact Headstrong Capital Markets
Noida – UP

☎ **091- 7827745827** ✉ **harshitapathak25@gmail.com**

8.5 years of comprehensive experience in the Information Technology (IT) industry with experience in data warehousing, ETL, Big Data, Data engineering and Cloud management.

EMPLOYEMENT SUMMARY

- **Data Warehousing** and **Data Integration** projects in Big Data - Design, development, implementation, maintenance with meeting and exceeding the business requirements in **AGILE** methodology.
- **ETL solution** - Big Data in **AWS** and **On-Premise** server in **Python** and **Spark**
- **Spark Developer** in **Python** – Creating **complex scripts** to transform and load data from **HDFS Data Lake** into **Hive Tables**.
- **Data Analysis** on data present in **Redshift, Hive** and **Athena** tables**.**
- **PySpark** RDD Actions & Transformations and **Dataframe API**.
- **AWS solution Designer** – Data extraction, Transformation from AWS **S3** Buckets, AWS **ATHENA**, AWS **Glue** tables running **spark** application in AWS **EMR** cluster and using other AWS services.
- **CI/CD** pipeline implementation with **Jenkin**, **docker** & **AWS ECS container services**.
- **Data validation** and **Data Quality (DQ)** - Ensure data quality along with end-to-end testing of deliverables.
- **AWS cloud services**, **IAM** roles, **policies**, troubleshooting and management.
- Job scheduling **Airflow** – Responsible for automation by scheduling the jobs.
- Experience in setting up several **DevOps** tools like **Jira**, **Jenkins**, **GitHub** etc.
- Involved in Training and Mentoring new team members.
- Team player with excellent communication, problem solving skills and willing to learn new technologies.
- Received various recognitions and awards.

ACADEMIC QUALIFICATION

MBA – Information Technology

IMT Ghaziabad
2015 - 2017

B. Tech - Electronics & Communication

IIMT College of Engineering, Gr. Noida | UPTU
2007 – 2011

TECHNICAL PROFICIENCY

- Python, PySpark APIs
- Data Analysis using Spark Core & Spark SQL.
- AWS EMR, AWS Glue CatLog, AWS Athena, AWS Glue Crawler, AWS S3, AWS Cloud watch, AWS Glue, AWS Redshift, AWS IAM roles, Policies
- GIT, Jira, Confluence, Docker & Container Services
- SQL, Oracle, Unix shell scripting
- Python, Pandas, NumPy

WORK EXPERIENCE

Genpact Headstrong Capital Markets

Data Engineer April 2019 – Present

- Worked towards design & creation of Finance data mart for a travel client to cater the reporting needs and to support the BI team.
- Integrating data from various data sources - MySQL and Kinesis streams into S3 buckets.
- Loading the data into Redshift stage layer and the performing data validation and transformation using spark to finally load the data into DataMart in Redshift.
- Performing Data Quality checks on the input data.
- Help the BI team with adhoc queries and performance tuning of existing tables.
- AWS Infra (EC2, S3, VPC, EMR, Redshift) Provisioning and Management.
- Setting up IAM roles, user, groups, policies for different AWS services for the team.
- Troubleshooting connectivity, access, performance issues etc.
- Project and Collaboration platforms – Jira, Confluence and GitHub.
- Participating in the daily stand-up and running Jira scrum board.
- Performing Team handling activities.

DXC Technologies

Technical Lead May 2014 – April 2019

- Worked as part of Global Development platform team to create a Datamart solution in Hive by pulling data from HDFS in Hadoop by using Python and Spark. in AWS using AWS EMR, Athena, Glue to process the data.
- Part of Hadoop Data Lake implementation to create a one stop solution for all the raw data in parquet format.
- Migration of on-premise solution to AWS cloud using EMR, ATHENA and S3 buckets and AWS CLI.
- Reading the data from AWS **S3** via **spark**, performing data **transformation**, **joining**, data **validation** and **loading** the data to AWS **Athena** tables which have data catalog in AWS **Glue**.
- ETL creation for pulling data from Hadoop Data Lake, mapping data to correct columns, putting filter validation, missing validation and QA validations in **python** and **PySpark**.
- Worked towards creating **ETL** and transformation logic in **python** scripts.
- Worked towards creating **Python** Wrappers Classes to encapsulate the whole ETL logic into one entity and writing data to **Hive** and **HDFS**
- Analyzed the **Pyspark** script and work on **optimization** of the execution of the script.
- Involved in **creating Hive tables**, and **loading** and **analyzing** data using **hive queries**
- Implemented **Partitioning**, **Dynamic Partitions**, **Buckets** in **HIVE**.
- Used Reporting tools like Quick sight to connect with Hive for generating daily reports of data.
- Troubleshooting connectivity, access, performance issues etc.
- Reviewing the peer codes and merging these into GIT.
- Technically Helping people in python, spark and guiding them in different teams.
- Responsible for Shortlisting candidates and interviewing people in recruitment drive.
- Visiting the UK Team to understand the requirements and help to speed up projects.

Sapient Corporations (Infinite Computer Solutions)

Associate July 2013 – May 2014

- Working as a part of Replatforming team on a retail project to test the ETL solution in Hive using Hive queries.
- Working as a QA to manually test the ETL solution which extract the data from HDFS path from an on-premise Hadoop cluster and loads the data into Hive tables.

- Two set of Hive tables - source data and transformed data was in hive tables.
- Responsible for understanding the requirements thoroughly and converting them into Test Scenarios.
- Working towards creating negative and positive cases in Test scenarios.
- Responsible for creation and execution of Test cases.
- Involved in Defect logging and tracking.
- Involved in giving Internal Demo
- Responsible for functional and UI testing

Wipro Ltd

Associate May 2012 – Mar 2013

- Resolved the queries related to network issues for Verizon U.S. customers.
- Performed the L1 troubleshooting.
- Provided support during weekends.
- Worked 24X7 for support.