Shubham Dubey

dubeys7151@gmail.com | Personal Website | LinkedIn | github

SKILLS

PROGRAMMING
Python •Shell • Java

DATA ENGINEERING
Apache Hadoop • Spark • Hive • SQL (MYSQL)

DEVOPS
Git• Jenkins

FAMILIAR
Machine Learning • Web Development

LINKS

Github:// **shubham7151**
LinkedIn:// **dubeys7151**
LeetCode:// **dubeys7151**

EDUCATION

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
B.Tech in Computer Science and Engineering
2020

CERTIFICATE

BIG DATA HADOOP AND SPARK DEVELOPER
SimpliLearn | View

LANGUAGE

English
Hindi

EXPERIENCE

INFOSYS LIMITED | Senior System Engineer
April 2022 - Present

- Build ETL pipeline to ingest historical data and transform it based on business requirement.
- Understanding business requirement and interaction with stack holders.
- Worked on 2 GB out of total 25 GB of data.
- Ingested data from S3 to Databricks and created notebook for transformation.
- Used spark dataframe to apply business logic on data and write transformed data to downstream storage (S3).
  Tech Stack : AWS S3, Databricks, Python, Pyspark, Spark Dataframe, Spark sql.

INFOSYS LIMITED | System Engineer
01/2021 - 03/2022

- Worked as Test Engineer for Client Apple Inc.
- Product owner for corp features like group, person service, SCIM.
- Regression analysis was part of the day to day task.
- Worked efficiently under an agile environment with devops tools.
  Tech Stack : Java, TestNg, Git, Jenkins.

PROJECT

LOG ANALYSIS USING SPARK | View

- Project focuses on analysis of log using Spark Dataframe on Databrick. Source : EDGAR log file data set.
- The data have over 1 lakhs records, loading the data to Databrick and using the platform to analyze the data.
- Used Spark UDF to convert data to different formats.
  Tech Stack : Python, Apache Spark, Databricks, Spark Sql, Spark Dataframe.

MOVIE DATA REVIEW USING APACHE HIVE | View

- Project focuses on loading data into Apache hive and analyzing the same. Source: https://simplilearn.com
- The data is collected from IMDB, with 49000+ records. The records are analyzed by querying using Hive query Language.
- Used Hive UDF to convert data from one form to other to facilitate better analysis.
  Tech Stack: HDFS, Hive, Python, HQL, Matplotlib.