

Shubham Dubey

+919205497291 | dubey7151@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](#)
Hacker-Rank:// [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](#)
AWS - Services overview | Infosys Internal
AWS-Compute and Network Services | Infosys Internal
AWS - Storage, Databases and Migration Services | Infosys Internal

LANGUAGE

English
Hindi

EXPERIENCE

INFOSYS LIMITED | SENIOR SYSTEM ENGINEER | DATA ENGINEER

12/2021 - 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 | AUTOMATION TEST ENGINEER

01/2021 - 11/2021

- 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.