Shubham Jain
Mobile (+91) 7828733244
Email ID- shubhamakachamp@gmail.com
LinkedIn URL- www.linkedin.com/in/shubham-jain-59326680

Work Experience and Notable Highlights

Quantiphi Analytics | Bigdata Engineer | May-19 till date

- Design and Implementation of data lake on AWS using S3, Redshift, Glue and DMS.
- Implementation of document style Data Hub on AWS using DynamoDB, Glue, Step functions, Athena, S3, EC2 and Infoworks.
- Expertise in creating ETL pipelines with the help of Apache Spark over Glue and EMR.
- Optimizing Apache Spark jobs by optimizing joins and reducing the data shuffle over the network.
- Optimization of Redshift cluster performance, using optimum dist keys and sort keys and tuning sql queries for maximum performance.
- Optimization of DynamoDB queries by implementing appropriate GSI and choosing the best partition and sort key.
- Implementation of Flask based API to fetch data from dynamodb and serve customer and deploying it on AWS Elastic Container Service using Docker images.
- Worked on creating Docker images for enabling Spark history server, glue local dev endpoints and enabling data lineage for glue jobs and deploying them on ECS.
- Implemented a data model on Azure using Datafactory, Databricks and Blob storage.

Tata Consultancy Services | System Engineer | Dec-16 to May-19

- Experience in Data Analytics with Cloudera CDH 5.13 and services include Spark, Python, HDFS, Hive, Impala, Flume, Sqoop and Oozie.
- Design and implementation of python based framework to import data from various relational databases(Teradata, Oracle, SQL Server, DB2) to hadoop HDFS using Sqoop, hive and Oozie.
- Implementing spark jdbc based data ingestion framework from Informix database.
- Developed spark based application to parse complex xml data received as file.
- Developed spark based reconciliation processes to maintain the data integrity in hadoop.
- Imlementation of Sqoop export framework for exporting the data from hadoop to Teradata.
- Excellent understanding of Hadoop Architecture and underlying Hadoop framework including Storage Management
- Analyzed and processed complex data sets(structured and semi-structured) using advanced querying and analytics tools

Achievements

- Amongst Top 5% pyspark contributor on stackoverflow.
- Published well received technical blogs on Medium

Tech Stack

Cloud : AWS (S3, EC2, Glue, EMR, ECR, ECS, Redshift, DynamoDB, Step function), Azure (Blob storage, Datafactory, Databricks), GCP (GCS, Dataproc, Bigquery, Airflow)
Programming Language: Python (Preferred), Shell scripting, Java
OpenSource Stack: Apache Spark, Hive, Impala, Presto, Oozie, HDFS, Livy, Spline(Spark Lineage), Docker
Operating System: Linux (preferred), windows
API Development: Flask

Personal Dossier

Qualification	: Bachelor of Engineering (Computer Science)
Languages known	: English & Hindi
Date of Birth	: 3rd July 1994
Current Address	: Jain Mechanical Works, pali road, sheopur, MP Pin: 476337
Portfolio	: http://jnshubham.github.io
Github	: http://www.github.com/jnshubham