GAURAB KUMAR JHA

Staff Engineer | Data Analytics Engineer

☎ +91-9923832293
✉ gaurabkjha@gmail.com
in linkedin.com/in/gaurabjha/

OVERVIEW

Get a challenging position where I can put all my experience and knowledge to give a significant growth to the organization.

PROFESSIONAL SUMMARY

- Almost **9 years of experience** as JAVA, Python & Big Data Developer
- 6+ years of experience in Big Data & Cloud Development – Batch and NRT Processing of the clinical data, SSD/HDD Test Data
- Fair Knowledge on Java, Python, C#, ruby scripting and relational databases, SQL (Structured Query Language) etc.
- Designed and Contributed to the design and implantation of various applications and tools.
- Interviewed, Mentored, evaluated, and graduated new hires and interns in the team.
- Good Knowledge on Dependency management, Source Control, and comprehension tool Such as Apache Maven, Git etc.
- Participated in peer code review to maintain high coding standards.
- Good collaborator with excellent communication skills, interpersonal skills with problem solving, trouble-shooting capabilities.
- Managed responsibility of Release and Deployment manager for pushing the code into production environment
- Worked on Cloud Storage & AWS – EMR, EC2, SaaS (MSK, EKS , MWAA etc), GCP – GKE

TECHNICAL SKILLS

- **Languages**: JAVA, Python, C#
- **Front End**: HTML, JavaScript
- **Framework**: Apache (Crunch, Spark, Kafka, Hive), Spring Boot
- **Hadoop Ecosystem**: MapReduce, HDFS, YARN, Oozie, HBase
- **IDE's**: JetBrains IntelliJ, Eclipse, PyCharm, SQL Developer, Visual Studio Code
- **Serialization Format**: AVRO, JSON, CSV, Parquet
- **Database**: MySQL, Oracle11g, Oacle12c, HBase
- **Build Tools & Source Control**: Apache Maven, GitHub, BitBucket, Github
- **CI/CD & Code Review Tools**: Jenkins, Crucible, Nexus, Spinnaker, Splunk
- **Cloud Infrastructure Provider**: GCP, AWS (S3), Cloudera
- **Other Tools**: Node Management (CHEF/Ansible), Kubernetes, Mesosphere, Cloudera Manager CDH6, Confluent, Putty, Hue, Airflow

WORK HISTORY

SanDisk(Wester Digital)	May'21 - Present	Bangalore, Karnataka
Cerner Healthcare	Aug'18 - May'21	Bangalore, Karnataka
NTT Data GDS	Feb'17 - Aug'18	Chennai, Tamil Nadu
Tech Mahindra Ltd	Oct'14 - Jan'17	Pune, Maharashtra

PROJECTS

#1. Western Digital Corporation | Sandisk India Device Design Centre Pvt Ltd

May 2021 - Present
Environment: Java, JDK 11, Python, Spark, Confluent Kafka, Kubernetes, Apache Airflow, GKE/GCP etc.
Role: Staff Engineer | Data Analytics Engineer

Responsibilities:

- I am part of DAO responsible for development and maintenance of Western Digital's **Big Data platform applications**.
- An essential part of this Big Data system is writing, extracting, and transforming (**ETL**) raw data into its final materialized format suitable for machine learning analytics.
- Work in a strong teamwork environment with Agile Scrum methodology to manage tasks assignments.
- I am responsible to design and **build big data infrastructure and ETL processes** in production.
- Developed pipeline using Kafka streaming where above materialized data will get populated to Kafka Topic and it will get consumed by different Team members.
- Write Airflow DAGs for various use cases, like running scoring models, report generation, batch decoding etc.
- Maintain and Support **Airflow Instances and the Kafka Instances**
- **Mesosphere to GKE Migration**

#2. Oracle Cerner Corporation (Healthcare Solutions, Services)

August 2018 – May 2021
Environment: Java, JDK 1.8, Crunch, Spark, Kafka, HDFS, HBase, Oozie, CHEF etc.
Role: Java | Big Data Developer

Responsibilities:

- Batch ETL (Extract Transform Load) of client's clinical data by deploying multiple data pipeline implemented in Spark (PySpark) and Apache Crunch (MapReduce Jobs).
- Perform Near Real Time (NRT) Processing and Transformation of clinical data using Apache Kafka and Apache Storm – deploying storm Topologies for each client.
- Perform data cleansing, transform and normalization of clinical data ingested into HDFS from various data sources – serializing it using AVRO serialization format and storing the data it into HBase Tables.
- Configure & deploy Oozie workflow using CHEF to Schedule MapR/Crunch Jobs.
- Wrote Unit and Integration Tests using Mockito, JUnit, Scala testing frameworks.
- Worked on various language e.g., Ruby, Shell to write script for functional testing of the functionality. Also, Setup different pipeline in Jenkins using Jenkins file and groovy script.
- Generate smoke test data and run an integration test to check – if the data is processed correctly.

#3. NTT Data GDC - Client: Owens & Minor (healthcare logistics)

February 2017 – August 2018
Environment: Hive, Hbase, Storm, Maven, Kafka, ZooKeeper, Linux, Splunk, Apache Crunch and Oracle12c
Role: Application Development Consultant

Responsibilities:

- As part of the Data Engineering Team, managed the batch and reporting solution.
- Created Shell Scripts to implement the business logic and automate the data loading on the Hive tables in Big-data Ecosystem.
- Wrote Java Program to transform the data which were used as library in the Apache Crunch programs to run Business logic to load the data into the system.

- Scheduling done by Rundeck and Oozie

#4. Tech Mahindra Ltd - Client: AT&T Inc. USA (Telecom)

October 2014 – January 2017
Environment: Java 1.7, SQL, PL/SQL, Servlets, JSP, Spring, Hibernate, Eclipse, Putty, SOAP UI, SQL Developer, Toad and Oracle Database (10g and 11g)
Role: Software Engineer

Responsibilities:

- Create/modify/update Web modules of the myAT&T web application using the spring and Hibernate framework.
- Managed code by addressing the defect and feature request.
- Create/update XMLs and XSDs for proper DOM management.
- Updating and maintaining the database of client customers and application.

ACHIVEMENTS & AWARDS

- *Received **Award** from **Analytics Magazine India** for a paper on Data Engineering | Handbook for Data Pipeline Orchestration Solution. - (Over 140 Companies participated)*
- *"1st Place in Q4 2020 **Hackathon** at Cerner for Automations of the deployment and running the workflow on AWS EMR."*
- *"**NOTT** Award for Extra Ordinary Performance in Q2 2019 and Q2 2020", and few organizational awards for helping to set up new team within the organization, setting up an on-boarding project for new hires in the team, which gives them the exposure to technical stack which the team uses.*
- *"**Bravo**" for commendable contribution to the myATT project.*
- *Awarded with **Golden Certificate**, Technical Topper of the Badge and Technical Topper of the ELITE (Training Program of TechM) 2014.*

This is to certify that all the information given above is true to the best of my *knowledge.*