Gunjan Kumar **+918106660562**

Senior Data Engineer (**AWS Certified**) **Kumar.gunjan@outlook.in**

Experience Summary:

Total industry experience: 9 years 8 months

- **Samsung R&D Institute, Bangalore (May 2022 - till date**)
  - Highest Designation: Senior Data Engineer
  - Responsibilities: Data Pipeline Implementation end to end in GCP/AWS
  - Database: Bigquery, Hive
  - Technologies: AWS/GCP, Airflow, Terraform, Cloud Formation, Pyspark, Dataproc, EMR, Superset, Hive, CICD(git actions),Sagemaker Feature store
  - Language: Python
- **Happiest Minds Technologies, Bangalore (Aug 2020 - till date)**
  - Highest designation: Senior Data Engineer
  - Responsibilities: Individual contributor, developing features using Pyspark, python. Job scheduling and orchestration using Databricks Notebook and Airflow. Making sure Jira story is moved to code review within due date. Writing unit test and Pytest for testing the newly developed feature
  - Database: Snowflake, Delta Lake, AWS RDS for metadata
  - Technologies: Pyspark, Databricks, DBFS, Delta Lake, Snowflake, Sftp, s3, Jenkins, Jira, Gitlab
  - Languages: python

- **Wipro Technologies – Pune (Nov 2018 to Aug 2020)**
  - Highest designation: Senior Project Engineer
  - Responsibilities: Data Pipeline development and RCA for high production Issues. Belonged to Big Data and Artificial Intelligence Team(DAII)
  - Databases: **Azure Cosmos Db**, Azure SQL database, Oracle, MySQL, **HBase**
  - Technologies: Kafka, Azure Databricks, Azure SQL Data warehouse, Hadoop stack
  - Languages: **Scala, Java**

- **Infosys Ltd. – Pune (Dec 2014 to Oct 2018)**
  - Highest designation: Senior System Engineer
  - Responsibilities: Development of **HQL** Script, **Aws Resources Management**
  - Technologies: Hive, Pyspark SQL, SQOOP, Flume, Azure EMR, EC2, RDS, SQS, Kinesis
  - Languages: Python, Scala

Technical Skills:

- Databases Systems
  - **BigQuery, Hive, HBase, Neo4j Graph** Database, **Oracle** 11g, MySQL,

Azure database for **PostgreSQL**

- Subject Area
  - Building data pipelines with TDD framework for both Batch and streaming Application on AWS and GCP. Driving POC in New Technologies. Cloud deployment and release Automation. Cloud provision automation. Providing RCA to high priority Production Issues. Competitive programming with good knowledge of Data structure and algorithm
- Big Data platforms
  - Google cloud and **AWS** cloud services for data/analytics, **Kafka**, **Azure Databricks**, Azure Stream Analytics, Azure Data Lake Store, Azure Data Factory, Azure Blob, **Amazon S3**, **Amazon RDS,** AWS Glue, EMR, Azure HDInsight, on-premises Cloudera cluster Kinesis, Hive, **Oozie, Airflow**, Hue, Yarn
- Languages/Platforms
  - Scala, Java, Python PL/SQL, Cypher for graph, PL-SQL, Unix shell scripting, Linux platforms, **CI/CD – Jenkins, Jira, Nexus, GIT, Atlassian Stack**

Certification & Training:

- Neo4j Certified Professional by Neo Technology
- **AWS Certified Data Engineer**
- **Az-900 Microsoft** Azure Fundamentals by Microsoft
- Java Sun Certified in JAVA SE6 by NIIT and Oracle
- PL/SQL certified by Infosys
- Training for GCP certified Data Engineer Professional
- Hands On Snowflake -WebUI Essentials by snowflake
- Dockers Essentials a Developer Introduction by IBM
- Build and Monitor Spark Application by MAPR
- **Best Rank under 10k** globally in Python Programming in Hacker Rank.

Details of the projects worked on:

Name of the Company	Samsung R&D Institute Bangalore
Project Name	Automated Personalized Engine
Client Name	Samsung USA
Description of the Project: Targeting Samsung customers by creating different ML models. Data pipeline for Ingesting the 2000+ features into Sagemaker Feature Store. 7 Data Pipelines were developed using spark and dataproc. Job was scheduled in airflow	
Duration	May 2022 to current
Role	Senior Data Engineer

Responsibility • Migration Of all the data pipeline from AWS to GCP • Writing unit and integration test cases for newly developed feature (TDD development) • Automation, scheduling Dataproc/EMR jobs using Airflow or inbuilt scheduler • Implementing CI/CD pipeline using Jenkins/Git Actions • Monitoring Junior data Engineer for the assigned jira task • Developing Data Quality Dashboard for monitoring Data pipeline Jobs • writing Bigquery SQL for data validation and EDA • Monitoring and alerting tool development in superset/Grafans • Making plan for Deployement of DE jobs in GCP/AWS • Followed best practices for git versioning and branching • Optimizing spark jobs for bigquery • Optimize the Bigquery implementation for cost optimization • Best practices for fast debugging and testing like using packaged wheel instead of scripts • Unit testing of Airflow DAGS even before deployment	
Technologies	Pyspark, Composer, aws, GCP,EMR,DataProc

Name of the Company	Happiest Minds Technologies Pvt. Ltd
Project Name	Audience Insights and campaign Insights
Client Name	Healthgrades
Description of the Project: Healthgrades, is a US company that provides information about physicians, hospitals and health care providers.	
Duration	Aug 2020 to current
Role	Senior Data Engineer
Responsibility • Developing new features in pySpark for Databricks • Pyspark package development and deployment into Databricks • Airflow Dag Development for scheduling pipeline job • Alerting and monitoring Dashboard Development in superset and Prometheus/Grafana • Development of very complex CI CD pipeline and automated code review in jenkins • Writing unit and integration test cases for newly developed feature • Automation, scheduling Databricks jobs using Airflow or inbuilt scheduler • Working on snowflake for making sure clean data is stored in the respective dB	

Technologies	Pyspark, Databricks, aws, git, Jenkins, jira

Name of the Company	Wipro Technologies Pvt. Ltd.
Project Name	Enterprise Customer Risk Rating and Behavior
Client Name	TD Bank
Description of the Project: An Anti-Money Laundering Project. The scope of the project was to build stream and Batch pipeline using lambda Architecture to find High Risk customers and their behavior.	
Duration	Nov 2018 to July 2020
Role	Big Data Solution Developer
Responsibility • Developing custom Kafka Producer for 29 sources of data in java • Developing Stream Data Transformation using Azure Databricks and **spark streaming**. • collaboration with the business and onsite teams to Provide **RCA** for legacy application of Map Reduce. • Developing Hive Query script for Data Extraction to Downstream • Developed shell and python script for Job scheduling and management. • showcasing POC in GCP cloud Dataflow using Python API.Scheduling using Airflow • Enhanced Legacy Map Reduce Application to tolerate Job failure and Restarting • Hands on with **BITBUCKET** as code base and version management • Azure Infrastructure Automation using Terraform, ARM template, Azure Devops • Metadata Management for failed and success jobs in MySQL • Maven for Code Build and Nexus as Repository for Code Release	
Tools used	Kafka, Blobstorage, **AzureDatabricks**, Spark stream, Hive
Platform	MySQL, Python, Linux

Name of the Company	Infosys Technologies Pvt. Ltd
Project Name	Customer Behavior
Client Name	NordStrom
Description of the Project: Building Streaming application using **spark** streaming and use of other data Acquisition tools like Flume and Kafka. Streaming data was visualized on a web application	
Duration	March 2016 to Nov 2018

Role	Senior Big Data Engineer
Responsibility • Developing Spark Application for structured and unstructured Dataset Reading Data from server logs to Kafka Topics using custom Producer. • Creating Direct Stream from Kafka Topic to spark streaming. • Data processing by creating RDD and data-frame in spark core. • window operation on streaming data. • Creating temp table or Data Frame in Spark-SQL to be used by other users. • Row level Transformation and Actions on RDD. • Writing RDD to HDFS using different compression method. • Developed application using SQS for Job Failure Email Notification	
Provide Support to Production Team for deployment and Job Scheduling	
Technologies	Sqoop, Flume, Spark, Awscloud, EMR, **AWS Glue**, SQS, Kafka

Name of the Company	Infosys Technologies Pvt. Ltd
Project Name	Data Ingestion/Extraction
Client Name	NordStrom
Description of the Project: Client Historical dataset of about 13 TB was mostly logs which didn't conform to any specific schema. Client took close to half a day to move the data into their BI systems weekly. They wanted to reduce this time. Queries over this data set took hours.	
Duration	June 2015 to Feb 2016
Role	Big Data Engineer
Responsibility: • Extracting Data from different web server sources in form of logs to HDFS. • Creating Hive schema and External tables to be used by other visualization tools. • Incremental Data load using SQOOP. • Using Oozie to automate data loading into the Hadoop Distributed File System and PIG to pre-process the data. • Created table namespace for different Data Formats	
Technologies	Sqoop, Flume, Spark, Awscloud, EMR, AWS Glue, SQS, Kafka

Name of the Company	Infosys Technologies Pvt. Ltd
Project Name	Training
Client Name	NA

Description of the Project: Training On different Technologies Stack by Infosys 1.Java 2.PL/SQL 3. Advance SQL 4. Hadoop Stack	
Duration	Dec 2014 to March 2015
Role	Trainee
Responsibility: Training in different technologies. Was amongst top 20 percentile of all Trainees	
Technologies	Java, SQL, PL/SQL, Hadoop Stack

Personal Details:

Education: B.tech (EEE) from KIIT University, Bhubaneswar(2014)
Current Job Location: -Bengaluru
Dob: -18 sep 1992
Passport: Yes
LinkedIn: https://www.linkedin.com/in/gunjan075/