Pooja Dangi

LinkedIn: https://www.linkedin.com/in/pooja-dangi/

Email: nitt.pooja281991dangi@gmail.com **| M:** 9986961537 **|** Bangalore, India

Professional Experiences

❖ **Envestnet Yodlee**
Member Technical Staff **|** Bangalore
Date: - 06/2020 to till now.

❖ **PwC India**
Senior Associate **|** Bangalore
Date: - 08/2019 to 06/2020

❖ **Tavant Technologies**
Senior Software Engineer **|** Bangalore
Date: - 08/2015 to 08/2019

Education Qualification

❖ Master of Computer Applications (**MCA**) from **National Institute of Technology (NIT) Tiruchirappalli** in 2015 with 8.73 pointer

Certification

❖ **Databricks Certified Apache Spark Developer -2016**

Profile Summary

- Over **6.5** Years of IT experiences in **Big Data and Cloud Technologies**.
- Having **5+** years of extensive hands-on experience in **Big Data** technologies **Spark v-3.0, Scala, Python**, **Hive**, **HDFS**, Amazon EMR, Amazon Step Functions, Airflow, Amazon S3, Azure Data Factory, Azure HDInsight, Azure Databricks, Azure SQL database, Azure Data Lake.
- Having **3+** years of experience in **Microsoft Azure Cloud technology**
- Having **2+** years of experience in **AWS Cloud technology**
- Extensive experience with design, development, optimization and performance tuning of big data pipelines with Spark SQL using Scala on cloud environment.
- Experienced with processing different file formats like, Parquet**,** ORC, JSON and Sequence file formats using Spark.
- Experience in implementing ETL/ELT workflow in Azure Data Factory, AWS Step functions and Airflow.
- Experience in designing table partitioning, bucketing and optimizing hive scripts using different performance utilities and techniques.
- Good domain experience on Manufacturing and Finance Industry.

Awards & Achievement

- Received "**Spot Award"** for successfully implementation of PoC for reports ETL redesign using Spark and Scala on AWS.
- Received "**Tavant Excellence Award**" for successfully delivery of IR-TMAP project.
- Received "**Tavant Quarterly Award**" for successfully building TMAP platform

Tools & Technologies

❖ **Bigdata Technologies** : Spark, Scala, Python, HDFS
❖ **Cloud Technologies** : Microsoft Azure , AWS
❖ **ETL & Orchestration Service** : Azure Data Factory, Airflow, AWS Step Function
❖ **Databases & Wahehouses** : Oracle , Postgres, SQL Server, Hive, Kudu, Snowflake
❖ **NoSQL** : Neo4j Graph DB
❖ **Versioning Control** : GitHub ,Perforce
❖ **BI & Other tools** : Power BI, Neo4j Bloom, Putty, WinSCP, Unix shell script, Jira ,Confluence

Work Experience

Member Technical Staff – (Envestnet Yodlee)
Jun 2020 to **Present**
Skills: - Spark, Scala, Python, HDFS, Hive, Airflow, Oracle, Snowflake, Kudu, Jira, Perforce, Cloudera CDH, AWS EMR, AWS Step Functions, AWS Lambda Function, AWS S3, AWS CloudWatch, AWS SNS, AWS RDS, AWS Data pipeline, Putty.

Responsibilities:

- Hands on experience with scalable solutions on big data ecosystem using Spark, Scala, Python, Hadoop, HDFS, Hive, Airflow, Postgres RDS, Oracle, Jira, GitHub, and AWS.
- Responsible for Data Ingestion, Data Cleansing, Data Standardization and Data Transformation.
- Optimized Spark jobs while submitting to EMR cluster.
- Developed Apache spark jobs to ingest data from various source systems to cloud storage.
- Read data from Postgres hosted on AWS into Spark Dataframe and Implemented Spark Dataframe transformations to map business analysis and load the data into Postgres, Hive and Kudu.
- Write transformed data by spark jobs to different data sources like Snowflake, Hive, Kudu, Oracle.
- Performed Data enrichments such as Filtering, Grouping, Sorting, Aggregation etc. in spark dataframe.
- Incremental loading of records implements using Apache spark jobs.
- Worked on spark code performance optimization to handle data skew and parallel read and write.
- Used Spark-SQL to process data and to run complex query on Spark SQL.
- Explored and implemented various spark optimization technique to improve spark job performance.
- Worked on AWS step functions to develop workflow to automate data pipeline.
- Worked on Airflow to develop DAG to automate data pipeline.
- Configured Airflow data pipeline to run multiple spark jobs in parallel
- Collaborating with business users/product owners/developers to contribute to the analysis of functional requirements.
- Supported code/design analysis, strategy development and project planning.
- Involvement in design, development, and testing phases of Software Development Life Cycle
- Define Data Architecture standards, best practices and participate in governance activities during Big Data projects.
- Define and implement standards and best practices for data ingestion, analytics, and integration in Data Lake.

Senior Associate (PwC India)
Aug 2019 to **Jun 2020**
Skills: Scala, Apache Spark, Hive, HDFS, Neo4j Database, Neo4j Bloom, Linkurious, Hortornworks Data Platform IaaS on Azure, Bash script, Linux, Eclipse, Putty.

Responsibilities:

- Collaborated with team for better understanding of the project functional and technical aspects.
- Analyzed OLAM data sources and process flow of OLAM Peru Coffee.
- Involved in creation of DDL hive scripts for three different data sources CTRM, DO and SAP systems and target tables in hive as per data model.
- Generated dynamic DML hive scripts for inserting data into target table using bash scripts.
- Created Spark-Scala jobs for 10 different module of OLAM Peru coffee supply chain process.
- Built and deployed Spark Scala job on cluster.
- Involved in creation of bash scripts to automate spark jobs on cluster.
- Involved in Unit testing.
- Created Spark jobs to load the data into Neo4j Database.
- Involved in Graph Data modelling.
- Created cypher queries to create nodes and relationships in Neo4j Database as per data model.
- Visualized graph data using Neo4j Bloom & Linkurious.
- Applied optimization techniques in ETL job level and query level.
- Proven team player, good communication skills and quick learner.

Senior Software Engineer (Tavant Technologies)
Aug 2015 to **Aug 2019**
Skills: - Cloud Technology, Azure Datalake, Azure HDInsight, Azure Datafactory, Azure SQL Database, Azure Blob, Hive

Responsibilities:

- Part of Data Engineering Team responsible for complete ETL flow.
- Involved in setting up and deployment of necessary resources in Azure subscription for the project.
- Involved in Data Modeling for IR-TMAP
- Implemented data factory pipelines for loading data from on-premises Oracle Database to Azure Cloud.
- Implemented full load and incremental load strategy for data load.
- Implemented data factory pipelines for creation of dimensions and facts as per data model of the project.
- Used Copy activity to move data to Azure, Hive activity for transformation in Hive and Stored procedure activity for SQL Database in data factory pipelines.
- Utilized Hive tables and HQL queries for weekly and monthly reports.
- Worked on creating Hive managed and external tables based on the requirement.
- Implemented Partitioning, Dynamic Partitions and bucketing in HIVE table for efficient data access.
- Implemented full automation and workflow of the data factory pipelines.
- Deployed and managed HDInsight cluster in Azure.
- Part of automation scripts for the cluster creation and deletion using Power Shell.
- Created compact and reusable hive scripts for data transformation.
- Created all the SQL Database tables for facts and dimensions and queries for intermediate tables required for the Power BI Dashboards.
- Involved in data cleaning and testing of data counts and data integrity in the ETL flow.
- Done testing for incremental data load on the daily basis.
- Successfully delivered IR-TMAP Project which compact and reusable code in both UAT and Production environment.
- Successfully given 3-month support for IR-TMAP.