

Name: Venkatesh Pasupuleti
Number: 9989797232
Email: pasupuletiVenkatesh80@gmail.com

Professional Summary:

- IT professional with around **4.3** years of experience in Software Development and implementation of Big Data and Big Data related technologies.
- Efficient on working with **Big Data and Hadoop Distributed File System (HDFS)**.
- Good working knowledge with **Apache Airflow, GCP & AWS** as Support role
- Expert in data ingestion, data transformations, data preprocessing, data exploration and data integration using Sqoop, **Hive** Scripts.
- Experienced with **Apache Spark** by improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
- Performed Hive operations on large datasets with proficiency in writing HiveQL queries using transactional and performance efficient concepts: Partitioning, Bucketing, efficient and effective Join operations.
- Experience and strong knowledge on implementation of **SPARK core, PYSPARK, Spark SQL**.
- Written **Sqoop scripts** In Order to Integrate with **Hive** and **MYSQL** Database.
- In-depth understanding of Apache-Spark Architecture and performed querying using Spark(Core, SQL), RDDs, Data frames, Datasets.
- Experience in creating Hadoop environment in GCP with the help of Google cloud Data proc and its components like **HDFS, Hive and SPARK**.
- Implemented **Spark** using **Scala** and **Spark SQL** for faster testing and processing of data
- Experience in **Scala** programming.
- Experience in working with Echo systems like **Sqoop, Hive, Hbase**.
- Worked with **Sqoop** to Import and Export data from a relational database into Hadoop.
- Exposure to open-source scheduling to **Airflow**.
- Good exposure to **ETL frameworks and fundamentals**.
- Good understanding/knowledge of **GCP Architecture and Hadoop Architecture** and its various components.
- Experience in using different file formats like **Parquet,Text ,AVRO,ORC and CSV**.
- Worked in **Scrum/Agile methodologies**.

Carrier Profile:

- Currently Working as a BigData Engineer, in **Wipro** From December 2021 to Till Date.
- Worked as a Software Developer, in **Virtusa** From July 2018 to November 2021.

Academic Qualifications & Certification:

- **Bachelor of Technology (B.Tech) in Electronics and Communication Engineering** from JNT UNIVERSITY, Hyderabad

Professional Experience:

Technical skills:

BIGDATA Technologies	Hadoop, Hive, AWS, Google cloud Data proc
Languages	PYSPARK, SPARK CORE, SPARK SQL
Cloud Platforms	AWS, EC2, EMR, DataBricks, Airflow
Build Tools	DataBricks, Jupyter Notebook
Data Formats	CSV, JSON, XML, and Parquet, ORC
Project Execution Tools	Agile–Sprint, Scrum, JIRA, Confluence, GitHub
Operating Systems	Windows, Unix and Linux.

PROJECT 1 :

Title : GSK - Cleo
Client : GSK
Role : Hadoop Developer
Organization : GSK Consumer Healthcare

Description:

GSK Consumer Healthcare India is a leading fast-moving consumer healthcare company with over 500 employees in India and category-leading brands such as Sensodyne, Eno, Crocin and Otrivin. Its Mission Health program touches the lives of approximately 90,000 people by focusing on health and well being with projects dedicated to neglected tropical diseases, dengue management, cleft and palate surgeries and plastic waste management.

Responsibilities:

- Developed pre-processing scripts.
- Able to do perform ingestion from source to Raw layer of s3 bucket using databricks.
- Developed ETL framework which can handle all file formats and write to S3 buckets using JSON parameters.
- Developed complex pyspark scripts and scheduled
- Checked code to Git and used Jenkins as CICD
- Worked on snowflake, Sagemaker for analytics workloads.
- Worked on aws Glue to ingest data for China region

Environment: Agile, Hadoop, Hive, ETL, AWS, Py spark, Spark, Python.

- Scheduling jobs in Apache Airflow

PROJECT 2 :

Client: Medieasy

Project: Sterling Logix LLC,USA

Role: Hadoop Developer

Description :

Medi Easy is a cross platform patient medical records suite for capturing monitoring information such as measurements for hypertension or blood sugar, patient reminders, symptoms reporting, issues reporting. We will process the metadata datasets coming thousands of records of patients data from various hospitals. The purpose of publishing this dataset is to figure out the patients depending upon various constraints like type of disease, based on gender, age, seasonal diseases, on location base.

Responsibilities:

- Written the Hive scripts to process the HDFS data.
- Created Hive tables to store the processed results in a tabular format.
- Developed Sqoop Scripts in order to make the interaction between Hive and mysql databases.
- Implemented **Spark** using Scala and Spark SQL for faster testing and processing of data.
- Involved in gathering the requirements, designing, development and testing.
- Writing the script files for processing data and loading in HDFS
- Writing CLI commands using HDFS
- Involved in requirement analysis phase.

PROJECT 3 :

Client Name: Credit-Suisse

Role: Data Engineer

Skills: Bigdata, Pyspark, Spark-sql, Hive, Python, Git, AWS,

Description:

Credit-Suisse is the leading financial in the EU region growth market. Their major focus is to improve couple of their business areas which include credit cards, loans, customer financial activities/behavior. Here we are building the unified storage (Data warehouse) for the financial client leveraging bigdata technologies that to boost the decision taking dashboards.

Responsibilities:

- Performed Data preparation
- Developed ETL pipelines using Pyspark.
- Developed all ETL transformation and job dependencies via pyspark.
- Took the ownership of bitbucket-dev branch.
- Performed HiveQL for data analysis.
- Also analyzed couple of complex shell scripts and leveraged them.
- Developed reconciliation script in order to perform unit testing.

Declaration: I hear by, confirming that all the above information provided by me is true

Venkatesh

Signature:

Place:

Date: