Yashwanth P

**Sr. Data Engineer**

yashwanth2191@gmail.com **Contact Number +1 (732) 313 2677**

****

**PROFESSIONAL SUMMARY:**

* Seasoned Data Engineer with a 9-year career, proficient in a wide range of data-related technologies and tools.
* Expertise spans Big Data, Data Warehousing, Programming, Databases, and Web Technologies, ETL, and Cloud platforms.
* Deep understanding of **Hadoop/Big Data, including HDFS, MapReduce, Hive, Pig, HBase, Sqoop, Impala, Oozie, Kafka, Spark, Zookeeper, Yarn.**
* Proficient in cloud-based Big Data solutions like **AWS, AWS S3, AWS EMR, and Redshift**.
* Strong experience in Big Data processing using **PySpark** and **NiFi**, along with **ETL** using **AWS** Glue.
* Skilled in data modelling, encompassing **OLTP, OLAP, Hive, Snowflake,** and **Teradata**.
* Extensive experience in designing, implementing, and optimizing data warehouses for efficient data retrieval and analysis.
* Proficient in multiple programming languages, including **Scala, Python, Unix, Linux shell scripts,** and **PL/SQL**.
* Competent in working with databases such as **Oracle, PL/SQL Server, Cassandra, MongoDB, Teradata,** and **DB2**.
* Proficient in managing and optimizing database systems for data storage, retrieval, and analysis
* Expertise in cloud platforms, including AWS and Azure, and leveraging cloud services like **AWS S3, Redshift, Glue, Lambda, Athena.**
* Skilled in optimizing data pipelines and automating **ETL** processes to deliver actionable insights from complex datasets.
* Dedicated to enhancing operational efficiency and enabling data-driven decision-making within organizations through technology and data utilization.

**TECHNICAL SKILLS:**

* **Hadoop/Big Data:** HDFS, MapReduce, Hive, Pig, HBase, Sqoop, Pig, Impala, Oozie, Kafka, Spark, Zookeeper, Yarn, AWS, AWS S3, AWS EMR, Redshift, PySpark, NiFi, AWS Glue.
* **Data Warehouse:** Erwin, OLTP, OLAP, Hive, **Snowflake** and Teradata
* **Programming languages:** Scala, Python, Unix & Linux shell scripts and PL/SQL
* **Databases:** Oracle, PL/SQL Server, Cassandra, MongoDB, Teradata and DB2.
* **ETL Tools:** Informatica, Talend, AWS Glue and Databricks.
* **Cloud Technology:** AWS S3, Redshift, Glue, Lambda, Athena, RDS, Azure.

**PROFESSIONAL EXPERIENCE:**

**Publicis Sapient, Hyderabad India Dec 2022 – Aug 2023**

**Role: Senior Associate, Data Engineering L1 Responsibilities:**

* + Enhanced data processing efficiency by 30% by performing comprehensive data wrangling using **Python, NumPy,**

and **Pandas**, ensuring clean and structured datasets for thorough and accurate analysis.

* + Conducted advanced data analysis and visualization using **Matplotlib**, creating detailed plots such as **histograms, scatter plots, and line charts** that provided actionable insights, improving decision-making processes by 25%.
  + Developed interactive dashboards and comprehensive reports in Power BI, effectively visualizing key performance metrics and trends for stakeholders, increasing engagement and understanding by 35%.
  + Orchestrated the utilization of **Apache Spark** to conduct in-depth big data processing and analytics, enabling real- time identification and mitigation of potential fraudulent patterns.
  + Implemented **Apache Airflow** workflows to automate data pipelines for supply chain data processing, resulting in 55% reduction in manual effort and 79% increase in data accuracy.
  + Leveraged **SQL and NoSQL** Databases **PostgreSQL, MongoDB** to structure and manage diverse datasets, ensuring seamless storage and accessibility of both structured and unstructured data for fraud analysis.
  + Used Spark SQL/Dataframes API to load, query, and store data from and to a wide variety of data sources into Hive.
  + Tuned Spark jobs for improving performance and load balancing.
  + Worked with XML transformer and other complex transformation to load data and create JSON API endpoints.

**Environment:** Cloudera Hadoop, MapReduce, Informatica, PySpark, HDFS, Nifi, Hive, Pig, Sqoop, Oozie, Zookeeper, Cassandra, HBase, Erwin, Spark, Spark-PL/SQL, Scala, AWS EMR, S3, AWS Glue, Redshift, Mongo DB, AWS Lambda, Snowflake, Databircks, Kafka, PL/SQL, Data Warehousing, PL/PL/SQL, RDBMS (Oracle, Teradata).

**Mind tree Ltd, Bangalore India Sep 2021 – Nov 2022**

**Role: Data Engineer**

**Responsibilities:**

* + Introduced a proprietary identification system with unique property codes, streamlining customer activities and enhancing tracking capabilities, resulting in a 90% reduction in data retrieval time. Utilized property codes as a sophisticated identification system, enabling the efficient retrieval and analysis of customer details, contributing to an 85% improvement in customer service’s response time.
  + Implemented a dynamic property codes generation system, resulting in 10% fewer errors and enhancing the overall accuracy of customer data. Implemented interactive dashboards for different departments, providing accurate insights and facilitating data-driven decision-making, leading to an 85% improvement in operational effectiveness.
  + Managed data from various sources and maintained **HDFS**. Visualized HDFS data for customers using **Hive ODBC**

Driver and BI tools.

* + Created a data pipeline for event ingestion, aggregation, and loading consumer response data into AWS S3. Utilized Lambda Functions and AWS Glue to create on-demand tables using Python and PySpark.
  + Analyzed and optimized PL/SQL scripts using **PySpark PL/SQL** for improved performance. Encoded and decoded JSON objects in PySpark for data frame manipulation.
  + Developed Big Data Analytics and Machine Learning applications with **Apache Spark** using **Python**. Executed machine learning use cases under **Spark ML and Mllib**.
  + Employed Scala API for programming in **Apache Spark**, imported data with Sqoop from Teradata, and developed POCs with Scala and **PySpark** on Yarn cluster.
  + Loaded **JSON data** using **PySpark-PL/SQL**, created schema **RDDs** and data frames, and loaded them into Hive Tables. Managed structured data with **Spark-PL/SQL**.
  + Enhanced existing Hadoop algorithms using **Spark Context, Spark-PL/SQL, Data Frames, and Pair RDDs**.
  + Developed a data pipeline with Amazon AWS to extract weblogs data and store it in HDFS.
  + Created Hive Generic UDFs for policy-based business logic. Imported data from Relational Databases into Hive Dynamic partition tables using Sqoop.
  + Customized Pig Loaders and storage classes to work with various data formats like JSON and XML.
  + Using **machine learning** to predict patient outcomes, disease progression, or treatment responses.
  + Building, training, and fine-tuning machine-learning models using frameworks like **TensorFlow or PyTorch**.
  + Utilized Spark for **ETL** tasks like removing duplicates, joins, and aggregation before storing in a Blob.

**Environment:** Hadoop, HDFS, HBase, Spark, MapReduce, Teradata, Informatica, MyPL/SQL, Java, Hive, Pig, Data Warehousing, Sqoop, Flume, Oozie, PL/SQL, Cloudera Manager, Cassandra, Scala, Python, AWS (EMR, S3, EC2, Athena, Glue, Redshift), PL/SQL, Elastic Search, Kafka, Tableau, ETL.

**Tech Mahindra, Delhi India Aug 2019 – Aug 2021**

**Big Data Analytics Engineer**

**Responsibilities:**

* + Overcame challenges associated with storing and processing large volumes of **structured/semi-structured data**, handling an average of 136.5 Terabytes of data via the Hadoop Framework.
  + Successfully transferred and analysed customer journey data related to Adobe products, managing and processing approximately 172.6 Petabytes of data in the **Hadoop Distributed File System (HDFS) using Hive**.
  + Conducted RCA for over 100 issues related to Hadoop services by demonstrating a 70% reduction in recurring issues.
  + Effectively processed 2-3 billion events data per day to filter out user activities to render insights for the client.
  + Analyzed Log files for Hadoop ecosystem services & conducted **Root Cause Analysis** to diagnose & resolve 100+ issues.
  + Transferred data into HDFS & analyzed Customer Journey about Adobe products via Hive.
  + Getting the files from different sources like SAP Hana via Snaplogic on daily basis and getting the data dumped into the HDFS.
  + Managed end-to-end Hadoop jobs utilizing technologies like **Sqoop, PIG, Hive, MapReduce, Spark, and Shell scripts** for data extraction and loading into the **Data Lake (Amazon S3).**
  + Oversaw data from various sources, maintained **HDFS**, and visualized HDFS data using a **BI** tool with the **Hive ODBC**

Driver.

* + Utilized **Apache Spark** with Python for **Big Data Analytics** and **Machine learning** applications, including **Spark ML and Mllib** use cases.
  + Optimized existing algorithms in **Hadoop using Spark Context, Spark-PL/SQL, Data Frames**, and **Pair RDDs**. Improved performance and optimization using **Spark Context, Spark-PL/SQL, Data Frame, Pair RDDs, and Spark YARN.**
  + Developed an Amazon AWS data pipeline for data extraction from weblogs, storage in HDFS, and processing with Partitions and Buckets based on State for Hive joins.
  + Leveraged AWS services to implement scalable data storage and processing solutions, enhancing data accessibility and security, and reducing data retrieval times by 40%.
  + Presented analytical findings to stakeholders using advanced presentation skills, clearly communicating data insights and recommendations, which informed strategic business decisions and improved project outcomes by 20%.
  + Enhanced data quality and reliability by 30% through comprehensive data cleaning and wrangling, utilizing Python and NumPy to address missing values, outliers, and inconsistencies.
  + Conducted advanced statistical analysis using SciPy and developed detailed visualizations in Power BI, delivering actionable insights that facilitated data-driven decision-making for clients.
  + Enhanced data accuracy and integrity by designing and implementing stored procedures, triggers, and user- defined functions in SQL Server, automating data processing tasks and reducing manual errors.

Environment: Hadoop, HDFS, HBase, Spark, Map Reduce, Teradata, Informatica, Python, Hive, Pig, Data Warehousing, Sqoop, Flume, Oozie, PL/SQL, Cloudera Manager, Cassandra, Scala, Python, AWS (EMR, S3, EC2, Athena, Glue, Redshift), PL/SQL, Elastic Search, Kafka, Tableau, ETL.

**Client scape Services India Pvt Ltd, Hyderabad India June 2015 - August 2019 Role: Software Engineer**

**Responsibilities:**

* + Promoting the updated scripts to STAGE, PROD environments using CICD. Schedule and Monitoring the Tidal jobs daily for automation.
  + Involved in creating the Control - M jobs creation (changing from Tidal-job-scheduler to Control - M)
  + Analyzing End-to-End scripts for causes of mismatches in final view.
  + Created interactive dashboards and reports in **Power BI/Tableau** to present actionable insights to stakeholders.
  + Conducted data preprocessing and transformation using **Python**, ensuring data quality and consistency.
  + Utilized Teradata for managing and querying large datasets, optimizing data retrieval processes.
  + Performed statistical analysis, hypothesis testing, and regression analysis to extract valuable insights from data.
  + Proficient in writing **PL/SQL** queries for data extraction, aggregation, and analysis across various database systems, including Teradata.
  + Developed and implemented machine learning models for predictive analytics and pattern recognition.
  + Integrated data from multiple sources, ensuring accurate and cohesive datasets for analysis.
  + Effectively communicated complex data findings and trends to technical and non-technical stakeholders.
  + Identified business challenges, formulated data-driven solutions, and contributed to process improvements.
  + Managed **end-to-end data analysis projects**, meeting deadlines and delivering actionable recommendations.
  + Ensured data governance and compliance with data privacy regulations, maintaining data security.
  + Stayed updated with the latest data analysis tools, techniques, and trends, continually enhancing analytical skills.

**Environment:** Teradata, HDFS, Map Reduce, Pig, Tableau, Rest API, Maven, Strom, ETL, PySpark, Shell Scripting.

**EDUCATION:**

**Saint Peter’s University,** Frank J. Guarini School of Business **Sep 2023 – Nov 2024**

Master of Science, Data Science

**Jawaharlal Nehru Technological University**

Bachelors in Electronics and Communication engineering, Hyderabad India **Aug 2011-May 2015**