PALLAVI MANUPATI

Chicago, IL | (913) 405-1488 | pallavi.mpati@gmail.com

SUMMARY

- Over 4+ years of experience as a Data Engineer, including profound expertise and experience in **statistical data analysis** such as **transforming business requirements** into **analytical models**, designing **Strategic solutions**, and **strategic solutions** that scale across massive volumes of data.
- Hands-on experience in developing **SPARK applications** using **Spark tools** like **RDD transformations**, **Spark core**, **Spark MLlib**, **Spark Streaming**, and **Spark SQL**.
- Experience in writing **Spark transformations** and actions using **Spark SQL** in **Scala**.
- Experience in layers of **Hadoop Framework - Storage (HDFS)**, Analysis (**Pig** and **Hive**), Engineering (Jobs and Workflows), extending the functionality by writing **custom UDFs**.
- Experience with **PySpark** for using Spark libraries by using **Python** scripting for data analysis.
- Experience in **Extraction, Transformation, and Loading (ETL)** of data from multiple sources into **Data Warehouse** and **Data Mart**.
- Experience in using **Tableau Desktop** to extract data for analysis using filters based on the business use case.
- Experience in working with **Docker** and managing **Docker** images on clusters.
- Experience in **Hive Query language (HQL)** and experience in **hive performance optimization** using Static Partitioning, Dynamic-Partitioning, Bucketing, and Parallel Execution concepts.
- Experience in **Dimensional Data Modeling Star Schema**, **Snow-Flake Schema**, Fact, and Dimensional Tables, concepts like **Lambda Architecture**, and **Batch processing**.
- Experience in using **Apache Kafka** for collecting, aggregating, and moving large amounts of data.
- Experience in **Amazon Web Service (AWS)** concepts like **EMR** and **EC2** web services which provide fast and efficient processing of **Teradata Big Data Analytics**.
- Experienced in creating **shell scripts** to push data loads from various sources from the edge nodes onto the HDFS.
- Utilize **algorithmic** problem-solving expertise to optimize data processing workflows.
- Experience in using **SQL** and **PL/SQL** for the development of Procedures, Functions, Packages, and Triggers.
- Experience with **AGILE Methodologies**.
- Demonstrated proficiency in troubleshooting and analysis, leveraging strong analytical skills to identify and resolve complex data-related issues efficiently.
- Implement and manage CI/CD pipelines using Jira, Bitbucket, and DevOps tools.
- Self-motivated exuberant learner and adequate with challenging projects and work in ambiguity to solve complex problems independently or in a collaborative team.
- Strong skills in analytical, presentation, communication, and problem-solving with the ability to work independently as well as in a team and follow the best practices and principles defined for the team.

SKILLS

- **Programming Languages**: Java, Python, SQL, Scala, MATLAB, HTML, PHP
- **Big Data Tools**: Hadoop, Hive, Apache Spark, Pig, MapReduce, HDFS, HBase, Sqoop, Kafka, Oozie, Airflow
- **Cloud Technologies**: AWS (S3, EC2, DMS, Redshift, Lambda, Glue, Snowflake, Kinesis), Azure
- **Database**: Oracle, MySQL, Microsoft SQL, PostgreSQL, Teradata
- **NoSQL Databases**: MongoDB, Apache Cassandra, Amazon DynamoDB, Redis, Elasticsearch
- **Data Formats**: CSV, JSON
- **Version Control**: Git, GitHub
- **Integration Tools**: Jenkins
- **Operating Systems**: Red Hat Linux, Unix, Windows, macOS
- **Reporting & Visualization**: Tableau, Power BI
- **Python Modules**: NumPy, Pandas, TensorFlow
- **IDE Tools**: Eclipse, Jupyter, Anaconda, PyCharm, VS Code

WORK EXPERIENCE

Data Engineer **01/2023 – Current**

AutomotiveMastermind Inc. **New York, NY**

- Interacted with clients to gather business and system requirements which involved documentation of processes based on the user requirements.
- Developed **Spark scripts** by using **Scala** as per the requirement.
- Developed **Spark programs** using **Scala**, involved in creating **Spark SQL queries** for faster data processing than the standard Map Reduce programs.
- Developed **Spark code** using **Scala** and Spark-SQL/Streaming for faster processing of data.

- Developed and implemented core API services using **Scala** and Spark.
- Responsible for designing and managing the **Sqoop** jobs that migrate the data from DB2 (Mainframe) to HDFS.
- Developed **Pig scripts** to parse the raw data, populate staging tables, and store the refined data in partitioned DB2 tables for Business Analysis.
- Involved in developing applications in **Python** language for multiple platforms and good experience in handling data manipulation using **Python** scripts.
- Developed ETL jobs using **PySpark**, and DataLiniage in which the data has been transformed in multiple stages and actions like aggregations are performed.
- Developed Tableau **data visualization** using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic maps, Pie Charts Bar Charts, and Density charts.
- Involved in Designing **Star Schema** (identification of facts, measures, and dimensions), **Snowflake Schema** for **Data Warehouse**, and ODS architecture by using tools like Data Model, and Erwin.
- Designed and implemented data pipelines utilizing **Amazon Kinesis Firehose** to ingest, transform, and load large volumes of streaming data into **AWS data lakes** and **warehouses**.
- Written multiple **MapReduce programs** for data extraction, transformation, and aggregation from multiple file-formats including XML, JSON, CSV, and other compressed file formats.
- Used **Kafka producer** to ingest the raw data into Kafka topics and run the **Spark Streaming** app to process click stream events.
- Collaborated with cross-functional teams to integrate **IAM** solutions into data pipelines, ensuring seamless access management within data workflows.
- Created **Hive queries** and tables that helped the line of business identify trends by applying strategies on historical data before promoting them to production.
- Worked on **MongoDB** by using CRUD (Create, Read, Update, and Delete), Indexing, Replication, and Shading features.
- Designed and implemented data pipelines using **Cradle**, ensuring efficient **extraction, transformation, and loading (ETL)** processes for large-scale datasets.
- Implemented monitoring and alerting systems on **Quick Site** to proactively identify and resolve **data pipeline** issues, ensuring continuous data availability and reliability.
- Worked on **SQL queries** in dimensional **data warehouses** and relational **data warehouses**.
- Performed Data Analysis and Data Profiling using Complex **SQL queries** on various systems.
- Implement **Data Modeling** techniques, **Data Warehousing** standard methodologies, and **object-oriented** concepts including **Test-Driven Development (TDD)**.
- Develop and apply comprehensive **data modeling** strategies, integrating **Data Warehousing** standard methodologies and practices, to effectively support both **OLAP** and **OLTP** databases.
- Utilized **Matillion** for data integration and transformation tasks to streamline data processes and enhance efficiency.
- Implemented CI/CD pipelines and **Release Lifecycle Management (RLM)** process using **Jira**, **Bitbucket**, and other DevOps tools.
- Utilized cloud solutions such as **AWS** (e.g., EC2, S3, Lambda, API Gateway, GLUE) and **SaaS** solutions to optimize data processes and storage.

Data Engineer **05/2018 – 07/2021**
Deloitte **Hyderabad, India**

- Involved in analyzing business requirements and prepared detailed specifications that follow project guidelines required for project development.
- Involved in the development of Spark jobs in **PySpark** and SparkSQL to run on top of **hive** tables and create transformed datasets for downstream consumption.
- Developed **Spark scripts** by writing custom RDDs in **Scala** for data transformations and performing actions on RDDs.
- Wrote Spark applications for Data validation, cleansing, transformations, and custom aggregations.
- Developed **Scala** scripts and UDFs using Data Frames and RDD in Spark for data aggregation, queries, and writing data back into the OLTP system through **Sqoop**.
- Developed the Map Reduce programs to parse the raw data and store the pre-aggregated data in the portioned tables.
- Developed ETL (Extraction, Transformation, and Loading) procedures and Data Conversion Scripts using Pre-Stage, Stage, Pre-Target, and Target tables.
- Optimized **SQL queries** and database performance, leveraging **PHP**-based tools and techniques to enhance data retrieval and analysis speed.
- Create comprehensive **reports** and **dashboards** using **PowerBI** to visualize data insights.
- Used various sources to pull data into **Power BI** such as SQL Server, Excel, Oracle, etc.

- Involved in designing and deploying rich Graphic visualizations with drill-down and Drop-down menu options and Parameters using Tableau.
- Write **Python** scripts to parse JSON documents and load the data into a database.
- Performed transformations like event joins, filter bot traffic, and some pre-aggregations using **Pig**.
- Implemented data ingestion and cluster handling in real-time processing using Kafka.
- Worked with building data warehouse structures, and creating facts, dimensions, and aggregate tables, by dimensional modeling, Star and **Snowflake schemas**.
- Migrated an existing on-premises application to AWS.
- Used AWS services like **EC2** and **S3** for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS **EMR**.
- Implemented AWS Elastic Container Service (ECS) scheduler to automate application deployment in the cloud using **Docker** Automation techniques.
- Used **hive** to do transformations, event joins and pre-aggregations before storing the data to HDFS.
- Created **HBase** tables to store various data formats coming from different applications.
- Analyzed the SQL scripts and designed the solution to implement using **PySpark**.
- Extracted files from **MongoDB** through **Sqoop** and placed in HDFS and processed.
- Use **SQL queries** and other tools to perform data analysis and profiling.
- Followed **agile methodology** and was involved in daily SCRUM meetings, sprint planning, showcases, and retrospectives.
- Participated in the status meetings and status updating to the management team.

EDUCATION

Master of Science: Computer Science
University of North Texas - Denton, TX