**Deepthi N**

**Sr. Data Engineer/Python Developer**

**Professional Summary:**

* Highly skilled Data Engineer/Python Developer with **9+ years** of experience in the IT Industry in areas of Data Analysis, Statistical Analysis, Machine Learning, Deep Learning, Data mining with large data sets of structured and unstructured data source and Big Data.
* Results-driven Python Data Engineer with extensive experience in designing, developing, and optimizing data processing solutions.
* Proficient in leveraging Python and associated libraries to build efficient **ETL** pipelines, conduct data analysis, and develop insightful data visualizations. Adept at collaborating with cross-functional teams to integrate data-driven solutions, ensuring accuracy and effectiveness in decision-making processes.
* Strong background in performance optimization, code efficiency, and adherence to best practices in software development.
* Experience in complete **Software Development Life Cycle** including Analysis, Design, Development, Testing, and Implementation using Python, Django, and Flask technologies.
* Expertise in managing **full life cycle of Data Engineer** project includes transforming business requirements into Data Collection, Data Cleaning, Data Preparation, Data Validation, Data Mining, and Data Visualization
* Experienced in using various Python libraries (**Beautiful Soup, NumPy, SciPy, matplotlib, python-twitter, Pandas**).
* Good experience in **RDBMS** concepts, **Data warehousing** concepts, OLTP & OLAP.
* Experienced in building/maintaining **Docker** container clusters managed by **Kubernetes**, **Linux**, **Bash**, **GIT** and Docker on **Google Cloud Platform** (GCP)
* Experience in developing pipelines in **spark** using **Scala** and **PySpark**.
* Experienced in data cleaning, handling missing values, reshaping and implementing exploratory data analysis with advanced data mining and Natural Language Processing techniques; Accessed data through **API/RESTful services**
* Experienced in **Amazon Web Services** (AWS) and **Microsoft Azure**, such as **AWS EC2, S3, RD3, Azure HDInsight, Azure Storage, and Azure Data Lake, Azure Data Factory (ADF)** etc.
* Solid understanding of big data technologies like **Hadoop**, **Spark**, **HDFS**, **MapReduce**, **Pig** and **Hive**.
* Experienced in **Agile Methodologies**, Scrum stories and sprints experience in a Python based environment, along with data analytics, data wrangling and Excel data extracts.
* Experience in writing **UNIX Shell Scripts** and automation of the ETL processes using UNIX Shell Scripting and SQL.
* Experienced in integration of various relational and non-relational sources such as **DB2**, **Oracle**, **SQL** **Server**, **NoSQL-MongoDB, XML** and **Flat Files**, to **Netezza** database.

**Technical Summary:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, SQL, Java, JavaScript, PHP, C, MATLAB, C++, Swift, Mathematica |
| **Machine Learning** | Classification, Regression, Clustering, NLP, Deep Learning |
| **Statistical Methods** | Regression Models, Hypothesis Testing, PCA, |
| **Python Packages** | Scikit-learn, NumPy, Pandas, Django, Flask, Matplotlib, PySpark, NLTK, TensorFlow, PyTorch, MATLAB |
| **Big Data Technologies** | Hadoop, PySpark, MapReduce, Apache Kafka, Spark, Pig, Hive, YARN, Sqoop |
| **Databases** | MySQL, SQL Server, MongoDB, HBase, Teradata, Redis |
| **Cloud technologies** | Azure, AWS (EC2, Redshift, RDS, S3), Google Cloud Platform EMR, Step Functions, Big Query, Redshift, CloudWatch, S3, DynamoDB, EC2, S3, Glue, Lambda, Dataproc, AWS Batch, AWS Quick sight. |
| **Data Visualization & BI Tools** | Tableau, Power BI, SSIS, SSRS, ELT, ETL |
| **Robotic Control****Package /Simulation** | ROS, Multisim, Gazebo |
| **Framework** | Spring Boot, Restful API, Microservices |
| **Version Control** | Git, Internal CICD, Airflow, GitHub, Gitlab, SVN, Bitbucket. |
| **Streaming & Message broker** | Kafka, Flink, Kinesis, Cloud Pub/Sub. |
| **ETL Tools** | AWS Glue, Apache MiFi, Dataflow, Data stage. |

**Educational Qualification:**

* **Masters in Computer Science.**

**Professional Experience:**

**Humana - Louisville, KY Dec 2022 – Till Date**

**Sr. Data Engineer/Python Developer**

**Responsibilities:**

* Design and develop ETL processes in AWS Glue to migrate payments data from external sources like **S3, Parquet/Text files** into **AWS Redshift**.
* Implemented **DBT (Data Build Tool**) in an AWS environment to streamline and automate data transformation processes, enabling efficient and scalable data pipeline management for analytics and reporting.
* Created End-End Automation with CI Procedures using Jenkins & automated Maven builds by integrating them with Continuous Integration tools Jenkins
* Extract Transform and Load data from Sources Systems to AWS Storage services using a combination of AWS Glue, **Apache MiFi**.
* Implemented end-to-end data ingestion, enrichment, and accessibility solution on AWS, utilizing services such as AWS Glue, **AWS Lambda**, Amazon Redshift
* Designed and developed data ingestion pipelines using Snowflake's features, including bulk loading and streaming, to seamlessly integrate data from various sources into the Snowflake platform.
* Integrated **Kafka & Flink** with other tools like **Hadoop, Spark**, and Elastic search to enable real-time data analytics and insights.
* Designed infrastructure for AWS application and workflow using Terraform and had done implementation and continuous delivery of AWS infrastructure using Terraform.
* Developed Python and PySpark code for ETL and used them in **AWS** services like AWS **Lambda, Glue, Ec2, EMR**, etc.
* Developed a shell script to run the ETL jobs from Control-M job scheduler
* Hands on experience in migrating on premise data to Google Cloud Platform (GCP) using cloud native tools such as **Big Query, Cloud Data Proc, Google Cloud Storage, Composer**, Pub/Sub & Cloud **Dataflow**.
* Implemented real time data streaming pipeline using Athena, **AWS Kinesis, Airflow,** Dynamo DB and deployed AWS Lambda code from Amazon S3 buckets.
* Extensive experience in designing and implementing NoSQL databases like **MongoDB** and Cassandra for high volume data storage and retrieval.
* Developed Tableau data visualization using **Cross tabs, Heat maps, Bar Charts, Gantt charts, Waterfall charts, Scatter Plots, Geographic Maps, Pie Charts** and **Donut Charts**.
* Used Git version control to manage the source code and integrating Git with Jenkins to support build automation and integrated with Jira to monitor the commits.
* Integrated Pig with Hadoop ecosystem tools like **Hive, HBase, Presto** and Spark to enable advanced analytics and data modeling.
* Worked in container-based technologies like **Docker, Kubernetes** and **OpenShift**.
* Developed RESTful API'S using **Python** with **flask and Django** framework and done the integration of various data sources including Java, JDBC, RDBMS, Unix & Shell Scripting, Spreadsheets, and Text files.
* Loaded data into **Spark** Data Frames and used **Spark-SQL** to explore data insights.

**Environment:** AWS,ETL, YARN, HDFS, Hive, Eclipse, Sqoop, Spark, Spark-SQL, Spark Streaming, Scala, Linux file system, Python, Django, flask, Linux Shell Scripting, Oozie, Domo, Tableau, NoSQL, MongoDB, Cassandra

**Volkswagen Credit - Libertyville, IL Sep 2020 - Nov 2021**

**Sr. Data Engineer/ Django**

**Responsibilities:**

* Involved in the project life cycle including design, development and implementation and verification and validation
* Extensively utilized Python frameworks like **Django, Flask, PyUnit**, and libraries like matplotlib.
* Gathered design requirements from the client, analyzed, and provided solutions and met design requirements.
* Updated and manipulated content and files by using **python scripts.**
* Created logic using **Python** programming language to node, functions and pipelines to retrieve and manipulate data.
* Implemented **Pandas, Pandas Data Frame query**, **NumPy** Python libraries, performed complex table joins, extractions and transformation of financial data retrieved from NetSuite and migrated to SAP.
* Implemented **ETL** pipeline to source data files and cross reference tables from **AWS S3** and perform data conversion using logic implemented in **Python**.
* Copied data from **AWS S3** into **Redshift** using **SQL** and performed text mining and analysis using **DBeaver**.
* Fetched twitter feeds for certain important keyword using the python-twitter library (Tweepy), Used JSON to store the twitter data which was further represented using matplotlib visualization and generated graphical reports for business decision using matplotlib library.
* Wrote queries using **SQL** in **DBeaver** to load data from **AWS** **S3**, transfer into **AWS Redshift** and perform transformation**.**
* Used **Spark** **(PySpark SQL)** in **Spark Standalone** mode and managed constant data changes to data analysis and migration.
* Participated in requirement gathering and analysis phase of the project in documenting the business requirements by conducting workshops/meetings with various business users.
* Maintain source file, cross reference table and output file location and parameters in **YAML file** for **Kedro** pipeline to retrieve and upload data files.
* Use **Jupyter Notebook** and **I Python** for testing design concepts and solutions.
* Use **GitHub** for version control and creation of repositories.
* Responsible for updating project status on **JIRA**.
* Scheduled and ran **Airflow DAG** jobs to sync and update cross reference tables in Google Drive to AWS S3.
* Update stakeholders on the results of the conversion program in daily standup meetings and log status in **JIRA.**

**Environment:** Python, PyCharm, Hadoop, Django, Flask, AWS S3, AWS Redshift, Pandas, NumPy, Spark, GitHub, SQL, Jira, Agile and MacOS.

**KeyBank - Cincinnati, OH Sep 2019 - Aug 2020**

**Data Engineer/ Flask**

**Responsibilities:**

* Used **Pandas API** to put the data as time series and tabular format for east timestamp data manipulation and retrieval.
* Worked on Data mapping, logical data modeling, created class diagrams and **ER** diagrams and used **SQL** queries to filter data within the Oracle database.
* Created a user review data classification pipeline using Apache Spark and Apache Airflow.
* Created a data extraction pipeline for migrating user review data from PostgreSQL dB to **AWS Redshift**.
* Deployed **AWS Lambda** functions to ingest user purchase data on a daily basis and stage it in S3.
* Designed a REST API backend using **Django** to provide access to marketplace trends dataset, this was used by product management team.
* Wrote and executed various **MYSQL** database queries from Python using Python-MySQL connector and MySQL dB package
* Used AWS lambda to run servers without managing them and to trigger to run code by S3 and **SNS.**
* Setup storage and data analysis tools in Amazon Web Services (**AWS**) cloud computing infrastructure.
* Updated Python scripts to match training data with our database stored in AWS Cloud Search, so that we would be able to assign each document a response label for further classification.
* Performed troubleshooting, fixed and deployed many **Python** bug fixes of the two main applications that were a main source of data for both customers and internal customer service team.
* Used Python scripts to update content in the database and manipulate files.
* Generated Python Django Forms to record data of online users
* Used Python and Django creating graphics, **XML** processing, data exchange and business logic implementation
* Wrote Python scripts that ran on AWS Lambda to transform and move large amounts of data in and out of AWS S3 & **DynamoDB**.
* Developed a Tableau dashboard on top of the metric dataset to visualize mobile ads performance and report to leadership team.
* Updated and manipulated content and files by using python scripts.
* Migrated key data pipeline components from **MATLAB** to Python, deployed using Lambda and used **CloudWatch** for monitoring, reducing firmware maintenance costs for product.
* Implemented web applications in Flask and **spring** frameworks following **MVC** architecture.
* Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
* Designed and developed the UI of the website using **HTML**, **XHTML**, **AJAX**, **CSS**, and **JavaScript**.
* Used Django configuration to manage URLs and application parameters.
* Built various graphs for business decision making using Python matplotlib library.

**Environment:** AWS, Python, API, PySpark, Django, Java, Oracle, Snowflake, Teradata, Tableau, Unix/Linux, Oracle/SQL & DB2, Agile, Apache Airflow, EMR, JavaScript.

**KinderCare – Portland, OR May 2016 – Aug 2019**

**Data Engineer/Python**

**Responsibilities:**

* Analyzed pre-existing predictive model developed by advanced analytics team and factors considered during model development.
* Analyzed metadata and processed data to get better insights of the data.
* Analyze and prepare data, identify the patterns on dataset by applying historical models.
* Designed and developed data management system using **MySQL**.
* Built application logic using **Python**.
* Angular.js is used to build efficient backend for client web application.
* Used Python to extract information from **XML** files.
* Worked on development of **SQL** and stored procedures on MYSQL.
* Designed and developed a horizontally scalable **API**s using **Python Flask**.
* Used supervised machine learning techniques such as Logistics Regression and Decision Tree Classification.
* Worked on datasets containing large values which are structured and unstructured data.
* Performed Data cleaning process- Forward filling methods on dataset for handling missing values.
* Created initial data visualizations in tableau to provide basic insights of data to the project stakeholders.
* Performed extensive exploratory data analysis using Teradata to improve the quality of the datasets
* Experienced in various Python libraries like **Pandas, One dimensional NumPy** and Two dimensional NumPy.
* Worked for the Analytics team, to update the regular reports and providing solutions.
* Creating visualizations for the data extracted with the help of **Tableau**.
* Identifying patterns and meaningful insights from data by analyzing it.
* Twisting SQL queries for improving performances.
* Performing Data Modeling in Tableau Desktop.

**Environment**: MS Access, SQL Server, Data modeling, Python (Pandas, NumPy, Sci-kit learn), PyTorch, Flask, Tableau, Excel

**Harvard Pilgrim Healthcare - Boston, MA Sep 2014 – Apr 2016**

**Python Developer**

**Responsibilities:**

* Writing **Python scripts** to parse XML documents as well as **JSON** based REST Web services and load the data in database.
* data and for building the pipeline.
* Experienced in the conversion of unstructured data in the format of **CSV,** and **JSON** into structured data in parquet form and saved in AWS s3 buckets.
* Experienced in the creation of Data zones that included AWS Glue databases and AWS Glue catalog that help in giving easy access to business and **Tableau** connection setup.
* Performed SQL optimization for the creation of AWS Views in Glue DB.
* Writing ORM’s for generating complex SQL queries and building reusable code and libraries in **Python** for future use.
* Working closely with software developers and debug software and system problems.
* Profiling **Python** code for optimization and memory management and implementing multithreading functionality.
* Involved in creating stored procedures that get the **data** and help analysts to spot the trends.
* Designed and developed the server module and resolved issues and responsible for its enhancements.
* Worked on **Django** ORM module for signing complex queries.
* Used WPF to create user interfaces for Windows operating system.
* Implemented business logic in **Python** to prevent, detect and claim duplicate payments Rewrite existing Python/**Django** modules to deliver certain format of data.
* Used Django Database **APIs** to access database objects.
* Developed tools using Python, Shell scripting, **XML** to automate some of the menial tasks.
* Developed scalable and effective applications to mane technology **tradeoffs.**
* Maintained web servers and platforms in cloud with collaborations with outside vendors. Used GitHub for version control.
* Performed **troubleshooting**, fixed, and deployed many **Python** bug fixes of the two main applications that were a main source of data for both customers and internal customer service team.
* Implemented automation using its **Python** API for test case scenario.

**Environment:** Python, Oracle, JSON, XML, Django, API, SQL, REST, AWS.