|  |  |
| --- | --- |
| **Amruta Raikwar** |  |
| **Role**: Data Engineer |  |



**SKILLS SUMMARY**:

Experience professional having 8 + years of collective experience in Data Engineering and Big Data Technology with proven ability to quickly learn evolving technologies, self-motivated and adaptable. Good verbal and written communications, proactive and a team player.

|  |  |
| --- | --- |
| Hadoop/Big Data Technologies | HDFS, Sqoop, Spark, Scala, Hive, Oozie, Apache Nifi, Apache Airflow |
| AWS | AWS EMR, S3, CloudWatch, Athena, Glue |
| Cloud Solutions | AWS |
| Python Libraries | Pandas, NumPy, Matplotlib, Scikit-learn, Pyathena |
| Data Visualization | Tableau, pyplot, seaborn, ggplot, Excel, Unidash |
| Programming & Scripting | Python, Java, R, shell-scripting, Advanced SQL. |
| Databases | Oracle, MY SQL, MongoDB, PrestoDB, Terraform |
| Version Control | SVN, GIT, Github, Teamcity, Mercurial |
| Other Tools: | PyCharm, Jupyter Notebook, Eclipse, Microsoft Word, PowerPoint, JIRA, ServiceNow, Tableau. |

**EDUCATION:**

Bachelor of Engineering | Electronics and Telecommunication (2010 - 2014)

Ramdeobaba college of Engineering Nagpur RTMN University, Nagpur, India

**CERTIFICATIONS:**

Post Graduate Program (PGP) | Big Data Analytics and Optimization (2018 – 2019)

International School of Engineering Bangalore, India (Certified by Language Technologies Institute (LTI) of Carnegie Mellon University (CMU), USA)

**WORK EXPERIENCE:**

|  |  |
| --- | --- |
| Client: Airbnb, CA |  |
| Role: Data Engineer | Jan 2024 – present |

Responsibilities:

* Working as a member of the Finance Data Engineering team.
* Designing, developing, and maintaining data pipelines and ETL processes for payments datasets.
* Collaborating with cross-functional teams, including Finance, Engineering, and Data Science, to ensure accurate and timely data ingestion, transformation, and loading.
* Performing data analysis and data profiling to identify data quality issues, anomalies, and trends within payments datasets.
* Optimizing data pipelines for performance and scalability to handle large volumes of payments data.
* Asisting in data modeling and schema design for payments datasets to facilitate efficient data access and analysis.
* Building and maintaining data pipelines and ETL (Extract, Transform, Load) processes to ensure the availability and accuracy of financial data.
* Participating in cross-team initiatives and projects to improve data engineering processes, scalability, and developer productivity.
* Executing data migration tasks, ensuring seamless and accurate transfer of payments data from legacy systems to new platforms or databases.
* Documenting data processes, pipelines, and data dictionaries to facilitate knowledge sharing and maintain data lineage.

Environment: AWS, Hive, SQL, Scala, Spark, Spark SQL, Python, Airflow

|  |  |
| --- | --- |
| Client: Meta, Fremont CA |  |
| Role: Data Engineer | April 2022 – Dec 2023 |

Responsibilities:

* Working as a member of the PPAO (Privacy Program Audit and Oversight) and storage hardware (RTP) teams.
* Collaborated with clients and business teams to understand their needs and create dashboards that tell a story with data.
* Responsible for creating and maintaining various data pipelines.
* Develop reports and dashboards for leaders and business users.
* Implement and fine-tune data pipelines.
* Improve data quality and reliability of data pipelines through monitoring, validations, and failure detections.
* Crafting and executing complex SQL scripts and building pipelines leveraging Spark, Python, and Airflow like in-house tools like Dataswarm.
* Analyze data and present findings in an understandable format to business users.
* Data warehousing and advanced SQL techniques.
* Worked on developing metrics to improve the efficiency of power utilization and network utilization within the fleet while also addressing spare planning and capacity.
* Built ODS/Rapido queries to extract the fleet storage capacity, power, and throughput data.
* Deployed scheduled Dataswarm pipelines to import the quarterly extracted data into Hive, for further processing and analysis.
* Developed consolidated reports and presented the dashboards on top of hive using presto queries in various time series charts and graphs using the processed data to provide comprehensive insights into fleet Storage capacity, power utilization, and throughput.
* Worked on building/finding Platform Hardware Security Metrics such as Attestation Coverage, Re-attestation required assets, Median Attestation Lifetime, Attestation Failures / Fraction of Attestation Failures Automatically Remediated / Fraction of Attestation Failures that are Undiagnosed.
* Used data from attestation history, repair insights and resort tables to formulate Daiquery queries to extract the Attestation and repair insights data.
* Build Dataswarm pipelines to import the extracted data into new Hive tables, for further processing and analysis. Developed consolidated reports and presented the dashboards on top of hive using presto queries in various time series charts and graphs.

Environment: Hive, MySQL, Presto SQL, Dataswarm Pipelines, Daiquery, Unidash, Python, Spark, Mercurial.

|  |  |
| --- | --- |
| Client: 84.51°, Cincinnati OH |  |
| Role: Data Engineer | Aug 2020 – Mar 2022 |

Responsibilities:

* Working as a member of the data Ingestion team, responsible for data ingestion and enterprise data movement for big data systems including Hadoop and cloud storage leveraging Apache NiFi, Spark, and python.
* Responsible for developing new ETL processes and maintaining current processes in Nifi.
* Developed end to end automation by automating data movement between servers and cloud solutions (GCP and Azure).
* Experience in using message brokers RabbitMQ. developed python scripts to interact.
* Worked on development of product called data Shuttle using Nifi to automate the data movement between different Hadoop systems and Cloud Solutions (GCP and Azure).
* Created Nifi flow files to save and pull metadata from MongoDB and use it for other Nifi data pipelines.
* Developed Spark code for faster testing and processing of data.
* Optimized spark internals and improved performance of spark jobs by 10%.
* Created Airflow DAGs as a scheduling script to run Python, Pyspark, shell commands.
* Assisted in the orchestration process to automate workflow and scheduling of the ingestion scripts using Airflow.
* Build data pipelines (ETL scripts) that would transfer data from clients' sources to on premises HDFS locations and vice-versa using MQFTE, Seeburger etc.
* Managed and analyzed the logs registered during the data transfers.
* Hands-on experience in using message brokers RabbitMQ. Created Airflow DAGs as a scheduling script to run Python, Pyspark, shell commands. For software development and collaboration and testing, worked with tools such as Git, optimized spark internals and improved performance of spark jobs by 20%.
* Used RabbitMQ messages as a trigger for initiating the data transfers between multiple platforms.
* Created, configured, and set up daily running Airflow DAGs as per the requirements to run spark command both in parallel and sequential.
* Performed tuning techniques to optimize the spark jobs performance.
* Built orchestration process to automate workflow and scheduling of the ingestion scripts using Airflow.
* Developed scripts using Python and Pyspark, managed CI/CD processes using Github and TeamCity.
* Implemented and maintained multiple Data pipeline DAG's and worked on Airflow performance tuning of the DAG.
* Maintained various versions of code repositories using Github, used Teamcity to deploy scripts. ● Got some experience on Terraform.

Environment: Hadoop, HDFS, SQL, Spark 2.4, Apache Nifi, Apache Airflow, Pyspark, Python, Rabbit MQ, MongoDB, git, TeamCity, linux.

|  |  |
| --- | --- |
| Client: Groundtruth, Mountain View CA |  |
| Role: Data Engineer | Oct 2019 – Mar 2020 |

Responsibilities:

* Worked on POC’s with Apache Spark using Pyspark to implement spark in the project.
* Pulling the data from data lake (HDFS) and massaging the data with various RDD transformations.
* Worked on generating business insights for Small Medium Businesses through digital marketing.
* Worked on a large set of location data on HDFS to discover patterns, meaningful relationships and created base tables using AWS Athena for deriving insights.
* Created SQL scripts for daily extracts, ad-hoc requests, reporting and analyzing large data sets from S3 using AWS Athena, Hive and Spark SQL.
* Creation of ETL, built data pipelines using spark SQL, Pyspark, AWS Athena and AWS Glue.
* Developed spark scripts, UDFs using both Data frames/SQL in Pyspark and RDD in spark for data aggregation, queries.
* Optimizing existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and RDDs.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Joins, Transformations.
* Analyzed the SQL scripts and designed the solution to implement using Pyspark.
* Collaboration with peer data scientists to understand and infer new business insights.
* Performed statistical data analysis of the user’s visitation data to different geo locations to understand users’ journey and their behavioral patterns.
* Built home geo- location data pipeline (in Pyspark) to map user’s residential locations.
* Built Machine Learning prediction model using Random Forest to predict number of user’s visits to a geolocation to improve sales using digital campaigns.
* Used Reporting tool Tableau for generating daily reports of data.
* Actively involved in migrating spark jobs to AWS EMR spark jobs.
* Developed an automated pipeline which is highly scalable using AWS Glue, Pyspark and s3.
* Data Extraction, aggregations, and consolidation of data within AWS Glue using Pyspark. ● Optimized spark jobs on AWS EMR saving costs.

Environment: Hadoop, S3, Hive, HDFS, SQL, Spark 2.4, AWS EMR, AWS Athena, AWS Glue.

|  |  |
| --- | --- |
| Company: Accenture, Bangalore India |  |
| Client: Schaeffler**,** Velux | Dec 2016 -- Feb 2019 |
| Role: Software Product & Platform Engineering Analyst |  |

Responsibilities:

* Installed, Configured and Maintained Hadoop clusters for application development and Hadoop tools like Hive, Spark, Zookeeper and Sqoop.
* Worked in a Hadoop ecosystem implementation/administration, installing software patches along with system upgrades and configuration.
* Created SQL-scripts which pulled all the product lifecycle data from the production server and moved it to all the testing and development servers.
* This reduced rehosting time of development and testing servers by 40%.
* Designed complex SQL queries to extract and transform data for further analysis and to detect anomalies.
* Developed reports in excel using data from SQL server, helping clients in analyzing daily and monthly active users’ trends.
* Developed spark scripts using Spark-SQL, Data Frames and RDDs, UDFs using both Data frames/SQL in Pyspark and RDD in Spark for data aggregation, queries.
* Worked on different file formats like Parquet, Avro and ORC file formats to store on HDFS.
* Rendered and delivered reports in desired formats by using reporting tools such as Tableau.
* Executed Hive queries on Parquet tables stored in Hive to perform data analysis to meet the business requirements.
* Analyzed multiple products raw data using excel, Python libraries, Pyspark based on location, time and quantity to estimate production cost.
* Involved in modifying various existing packages, procedures, functions, triggers according to the new business needs.
* Worked with Sqoop in Importing and exporting data from different databases like MySQL, Oracle into HDFS and Hive.
* Performed data cleansing, de-duplicating the data and has a good knowledge on best practices.
* Connected to various sources of data via tableau to validate and build dashboards and created story lines and made presentations on the findings.
* Involved in data validations of the results in Tableau by validating the numbers against the data in the database
* Performed analysis and presented results using SQL, SSIS, Excel.
* Wrote SQL queries to process the data using Pyspark and Spark SQL.
* Wrote Shell scripts, automating using crontab.
* Participated in trouble shooting the production issues and coordinated with the team members for the defect resolution under the tight timelines.

Environment: Hadoop, Agile, Oracle 11g, Apache Hive, HDFS, SQL, SPARK SQL, PL/SQL, ETL, Sqoop, Tableau 10.5, PLM

|  |  |
| --- | --- |
| Company: ITC Infotech, Pune, India | June 2014 - Nov 2016 |
| Role: Associate IT Consultant |  |

Responsibilities:

* Worked on Project and Report Management in Product Lifecycle Management.
* Build complex SQL queries using Query builder using custom Java Methods to fetch Project dashboard data in Reports.
* Created Query Builder Jobs which provided reports of users’ status from Active Directory of multiple clients used In Windchill PLM tool.
* Requirement gathering and analyzing data sources. Assisted in design and implementation of Database.
* Created SQL Server Reports and designed stylish report layout parameterized, performance, Ad-hoc reports.
* Extensively used Joins and subqueries to simplify complex queries involving multiple tables.
* Worked with Stored Procedure, Queries, Triggers, Functions, Indexes, User-defined functions
* Used SSIS and T-SQL stored procedures to transfer data from source databases to the staging area and finally transfer into the data warehouse.
* Extract, Transform and Load (ETL) data into the Data warehouse from databases such as Oracle, MySQL.
* Scheduled reports for daily, weekly, monthly reports for executives, business analyst and customer representatives for various categories and regions based on business needs using SQL Server Reporting Services (SSRS)
* Creation of objects like stored procedures, triggers, tables, views and analyzing tables and indexes for performance tuning.
* Worked with business users to understand metric definitions, presentation, and user needs.

Environment: Oracle, SSRS, SQL, SQL SERVER, MS Excel, PLM.