**Prajwal Rajeev**

**EMAIL : prajwalbhatia1327@gmail.com**

**Mobile: 201 – 680 – 1630**

**Sr. DATA ENGINEER**

**Summary:**

* Sr. Data Engineer with over 10 + years of experience in Data warehousing, Data engineering, Feature engineering, big data, **ETL/ELT**, and **Business Intelligence.**
* Experienced as a big data architect and engineer, specializing in **AWS** and **Azure frameworks**, Cloudera, Hadoop Ecosystem, **Spark/PySpark/Scala, Data bricks, Hive, Redshift, Snowflake**, relational databases, tools like **Tableau, Airflow, DBT, Presto/Athena**, and Data **DevOps Frameworks/Pipelines** with strong Programming/Scripting skills in **Python**, Expertise on designing and developing the big data Analytics platforms for Retail, Logistics, **Healthcare** and **Banking** Industries using **Big Data, Spark, Real-time streaming**, **Kafka**, Data Science, Machine Learning**, NLP** and **Cloud.**
* Worked in Data Engineering, Data Pipeline Design, Development, and Implementation as a Sr. Data Engineer and **Data Modeler/Data Analyst** using AWS, **Azure Cloud.**
* Experienced in **AWS**, **Azure DevOps**, Continuous Integration, Continuous Deployment, and Cloud Implementations.
* I have extensive experience in Text Analytics, generating Data Visualization using **Python** and creating dashboards using tools like **Tableau.**
* Developed Consumer-based custom features and applications using **Python, Django, HTML, and CSS**.
* Experienced with Software Development Life Cycle, Database designs, **agile methodologies**, coding, and testing of enterprise applications and IDEs such **as Jupiter Notebook**, **PyCharm, Emacs, Spyder** and **Visual Studio.**
* Proficient in managing the entire data science project life cycle and actively involved in all the phases of the project life cycle including Data acquisition, Data cleaning, Feature scaling, Dimension reduction techniques, Feature engineering, Statistical modeling, and Ensemble learning.
* Good understanding on **Apache Zookeeper** and **Kafka** for monitoring and managing Hadoop jobs and using **Cloudera CDH4**, and **CDH5** for monitoring and managing Hadoop clusters.
* Good working experience on **Spark** (Spark Core Component, Spark SQL, Spark Streaming, Spark MLlib, Spark GraphX, SparkR) with **Scala** and **Kafka.**
* Hands-on experience on Google Cloud Platform **(GCP)** in all the big data products **Big Query**, **Cloud Data Proc**, **Google Cloud Storage**, and Composer (Air Flow as a service).
* Used Explain to optimize **Teradata SQL** queries for better performance.
* Used the Teradata tools like **Teradata SQL Assistant**, Administrator and PMON extensively.
* Used DBT to test the data and ensure data quality.
* In-depth knowledge of **SSIS**, **Power BI, Informatica**, **T-SQL**, and reporting and analytics.
* Experience in building and architecting multiple Data pipelines, **end-to-end ETL**, and ELT processes for Data ingestion and transformation in **GCP** and coordinating tasks among the team.
* Understanding structured datasets, data pipelines, **ETL tools**, and extensive knowledge on tools such as **DBT**, data stage.
* Understanding of **Spark Architecture** including **Spark SQL, Data Frames, Spark Streaming**.
* Strong experience in analyzing Data with **Spark** while using **Scala.**
* Hands-on experience in using other **AWS** (Amazon Web Services) like **S3, VPC, EC2, Auto scaling, RedShift**, **Dynamo DB, Route53, RDS, Glacier**, **EMR.**
* Experience in Analysis, Design, Development, and **Big Data in Scala, Spark, Hadoop, Pig, and HDFS** environments.
* Used Data Stage stages namely Sequential file, Transformer, Aggregate, Sort, Datasets, Join, Funnel, Row Generator, Remove Duplicates, Teradata Extender, and Copy stages extensively.
* Built machine learning solutions using **PySpark** for large sets of data on the **Hadoop ecosystem**.
* Expertise in statistical programming languages like **Python** and R including **Big-Data** technologies like **Hadoop, HDFS, Spark, and Hive.**
* Extensive knowledge of designing Reports, Scorecards, and Dashboards using **Power BI.**
* Worked on technical stack like **Snowflake, SSIS**, **SSAS**, and S**SRS** to design warehousing applications.
* Integrated **DBT** with data warehouse technologies such as **Snowflake**, **Big Query**, or Redshift to leverage their capabilities for efficient data processing and storage.
* Experience in data mining, including predictive behavior analysis, Optimization and Customer Segmentation analysis using **SAS** and **SQL**.
* Experience in Applied Statistics, Exploratory Data Analysis, and Visualization using **Matplotlib**, **Tableau**, **Power BI**, and **Google Analytics**.

**Technical Skills:**

|  |  |
| --- | --- |
| **Hadoop Distributions** | Cloudera, AWS EMR, and Azure Data Factory. |
| **Languages** | Scala, Python, SQL, Python, Hive QL, KSQL. |
| **IDE Tools** | Eclipse, IntelliJ, PyCharm. |
| **Cloud platform** | AWS, Azure, GCP. |
| **AWS Services** | VPC, IAM, S3, Elastic Beanstalk, Cloud Front, Redshift, Lambda, Kinesis,  Dynamo DB, Direct Connect, Storage Gateway, EKS, DMS, SMS, SNS, and SWF |
| **Reporting and ETL Tools** | Tableau, Power BI, Talend, AWS GLUE. |
| **Databases** | Oracle, SQL Server, MySQL, MS Access,  NoSQL Database (HBase, Cassandra, Mongo DB) |
| **Big Data Technologies** | Hadoop, HDFS, Hive, Pig, Oozie, Sqoop, Spark, Machine Learning, Pandas,  NumPy, Seaborn, Impala, Zookeeper, Flume, Airflow, Informatica,  Snowflake, Data bricks, Kafka, Cloudera |
| **Machine Learning and**  **Statistics** | Regression, Random Forest, Clustering, Time-Series Forecasting, Hypothesis,  Explanatory Data Analysis |
| **Containerization** | Docker, Kubernetes |
| **CI/CD Tools** | Jenkins, Bamboo, GitLab CI, UDeploy, Travis CI, Octopus |
| **Operating Systems** | UNIX, LINUX, Ubuntu, CentOS. |
| **Other Software** | Control M, Eclipse, PyCharm, Jupyter, Apache, Restful API,  Jira, Putty, Advanced Excel, TOAD, Oracle SQL developer,  MS Office, FTP, Control-M, SQL Assistant, Rally, JIRA, GitHub, JSON |
| **Frameworks** | Django, Flask, WebApp2 |

**Education Details:**

* Bachelor’s in computer science engineering from Terna Engineering College, Maharashtra, India - May 2013
* Master’s in computer science from Pace University, Seidenberg School of Computer Science, and Information Systems, New York - May 2020

**PROFESSIONAL EXPERIENCE:**

**Client: Albertsons, Pleasanton, CA Feb 2023 – Till Now**

**Role: - Sr.Data Engineer**

**Responsibilities:**

* Transformed business problems into **Big Data** solutions and defined **Big Data strategy** and Roadmap.
* Installed, configured, and maintained Data Pipelines.
* Conducted thorough testing and quality assurance of **SnapLogic** pipelines, ensuring high data integrity and supporting the organization's data-driven decision-making process.
* Develop and deploy the outcome using **spark** **and Scala** code in **Hadoop** cluster running on **GCP.**
* Designed the business requirement collection approach based on the project scope and **SDLC methodology.**
* Created Pipelines in **Azure Data Factory** using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like **Azure SQL, Blob storage, Azure SQL Data warehouse**, write-back tool, and backward.
* Worked with Data governance and Data quality to design various models and processes.
* Involved in all the steps and scope of the project reference data approach to **MDM**, have created a Data Dictionary and Mapping from Sources to the Target in **MDM Data Model.**
* Chipped away at **Python** content to extricate information from **Netezza data** sets and move it to **AWS S3**.
* Implemented **restful API** to transfer data between systems.
* Experience in analyzing and writing **SQL** queries to extract data in **JSON** format through rest API calls with API keys.
* POC evaluation of **IBM cloud** for Data with NexGen and **Snaplogic** against each use case and bench marking to arrive at most suitable Enterprise Architecture **ETL** tool.
* Designed and coded different **SQL** statements in Teradata for generating reports.
* Involved in query translation, optimization, and execution.
* Automated the data processing with Oozie to automate data loading into the **Hadoop** Distributed File System.
* Collaborated with cross-functional teams to understand data requirements, identify data sources, and create scalable and maintainable **DBT** pipelines for data extraction, transformation, and loading.
* Designed and Developed and **Oracle PL/SQL** Shell Scripts, Data Import/Export, Data Conversions, and Data Cleansing.
* Engineered robust data pipelines using **SnapLogic** to automate data flow, ensuring timely and accurate data availability for analytics and reporting.
* Performed data analysis, statistical analysis, generated reports, listings, and graphs using **SAS tools**, **SAS/Graph, SAS/SQL, SAS/Connect, and SAS/Access**.
* Developed Spark applications using **Scala** and **Spark-SQL** for data extraction, transformation, and aggregation from multiple file formats.
* Created on-demand tables on S3 files using **Lambda Functions** and **AWS Glue** using **Python** and **PySpark.**
* Used Athena to transform and clean the data before it was loaded into data warehouses.
* Used Data Build Tool **(DBT)** to debug complex chains of queries. They can be split into multiple models and macros that can be tested separately.
* Authored **Python** **(PySpark)** Scripts for custom UDF’s for Row/ Column manipulations, merges, aggregations, stacking, data labeling, and for all Cleaning and conforming tasks.
* Designed and maintained **DBT** models, including defining the schemas, managing incremental data loads, and handling data lineage for efficient data processing.
* Collaborated with business analysts and data scientists to understand data needs, translating complex requirements into scalable **SnapLogic** pipeline solutions.
* Used ORC, Parquet file formats on **HDInsight**, **Azure Blobs**, and **Azure** tables to store raw data.
* Involved in writing **T-SQL** working **on SSIS, SSAS**, Data Cleansing, **Data Scrubbing**, and **Data Migration**.
* Worked on Dimensional and Relational Data Modeling using Star and **Snowflake** Schemas, **OLTP**/**OLAP** system, Conceptual, Logical and Physical data modeling using Erwin.
* Performed **PoC** for **Big data** solution using Hadoop for data loading and data querying.
* Wrote Pig Scripts to generate Map Reduce jobs and performed ETL procedures on the data in HDFS.
* Used Sqoop to channel data from different sources of HDFS and RDBMS.
* Involved in Normalization and De-Normalization of existing tables for faster query retrieval.
* Developed and maintained data dictionary to create metadata reports for technical and business purposes.
* Developed Json Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Cosmos Activity.

**Environment:** Python, Azure HD Insights, Hadoop (HDFS, MapReduce), YARN, Spark, Spark Context, AWS, Azure, GCP, Spark-SQL, PySpark, DBT, Pair RDDs, Spark Data Frames, Spark YARN, Hive, Pig, HBase, Oozie, Hue, Sqoop, Restful API, Flume, Oracle, NIFI, Kafka, Erwin9.8, BigData3.0, Hadoop3.0, Oracle12c, Pig0.17, Sqoop1.4, Oozie4.3.

**Client: PNC Bank, Pittsburgh, PA July 2020 - Jan 2023**

**Role: AWS Data Engineer**

**Responsibilities:**

* Developed and deployed data ingestion and storage solutions utilizing AWS S3, Redshift, and Glue, enhancing data management and storage capabilities.
* Successfully designed and implemented a scalable and cost-effective AWS-based Data Lake architecture for a large telecommunication company, handling terabytes of data.
* orchestrated data pipelines using AWS Glue for data extraction, transformation, and loading (ETL) into Redshift, resulting in a 30% reduction in data processing time.
* Architected and maintained Delta Lake using Databricks, ensuring data consistency, transactional capabilities, and improved data quality for a telecom analytics system.
* Extract, Transform, and load data from source systems to AWS data storage services using a combination of AWS Data Pipeline, AWS Glue, Apache Spark, and Amazon Redshift Spectrum.
* Developed and implemented optimized Snowflake stages to streamline data loading from a variety of sources into Snowflake tables.
* Designed and implemented ETL processes within SnapLogic, enabling the extraction, transformation, and loading of large datasets from various sources into a centralized data warehouse.
* Managed continuous replication of data using AWS DMS to ensure data consistency between source and target.
* Developed ETL (Extract, Transform, Load) processes to cleanse, enrich, and transform raw data in AWS S3 before loading it into AWS Aurora Postgres DB for analysis and reporting.
* Implemented data pipeline orchestration using Boto3 to schedule and manage data workflows.
* Designed and developed streaming data pipelines with Delta Lake integration, enabling real-time data ingestion, processing, and analytics.
* Utilized Matillion ETL solution to develop a pipeline that extracts and transforms data from multiple sources and loads to Snowflake.
* Utilized partitioning features in Complex SQL to segment large datasets for easier management and quicker data retrieval.
* Utilized Terraform to establish ECS clusters for containerized data processing, enhancing scalability and isolation of analytical workloads.
* Developed ETL pipelines using AWS Glue and Lambda, ingesting data from DynamoDB and MongoDB into Snowflake.
* Optimized Informatica ETL workflows and AWS Redshift query performances and improvement in data processing rates.
* Ensured data security by integrating Informatica's data masking and encryption features with AWS Key Management Service (KMS) and IAM roles.
* restored critical data from Aurora snapshots during a data corruption incident, with minimal downtime.
* Data intake is performed using Sqoop, and data ingestion is performed using MapReduce and HBase.
* Extensively worked on Spark Streaming and Apache Kafka to fetch live stream data.
* Developed and implemented a Snowpipe pipeline to ingest real-time data from Amazon S3 into Snowflake.
* Contributed to data processing by utilizing an ETL pipeline orchestrated by AWS Data Pipeline with Hive.
* Implemented Kafka manager to monitor Kafka Metrics and track consumer lags. Also responsible for managing topics and partitions.
* Proficient in creating configuration files for deploying SSIS packages across various environments.
* Skilled in writing SQL and R queries for extracting, transforming, and loading (ETL) data from large datasets using Data Staging.
* Utilized Data Pipeline and Glue extensively to orchestrate ETL transformations for slowly changing dimensions, enabling data updates, and ensuring accurate analysis.
* Profiled and tuned MongoDB queries using indexes, leading to improvement in application response times.
* Used Snowflake's time travel feature to analyze past customer data to give the best offers.
* Secured DynamoDB data using IAM, encryption, and VPCs.
* Created Databricks notebooks utilizing SQL and Python, and automated notebooks using jobs.
* Interacting with other data scientists and architected custom solutions for data visualization using tools like Tableau, packages in R.
* Proficient in integrating star schema models with the AWS data ecosystem, utilizing services such as Glue for data cataloging and ETL, and Lambda for serverless data processing.
* Developed predictive models using Python and R to predict customer churn and classify customers.
* Familiar with data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling and data mining, machine learning, and advanced data processing.
* Designed and implemented by configuring Topics in the new Kafka cluster in all environments.

**Environment:** AWS, AWS S3, redshift, EMR, SNS, SQS, Athena, glue, CloudWatch, kinesis, route53, IAM, Sqoop, MYSQL, HDFS, Apache Spark, Hive, Cloudera, Kafka, Dynamo DB, PySpark, Ambari, control-m, airflow, Oracle, SQL.

**Client: Tire Hub, Dunwoody, GA Sept 2019 – June 2020**

**Role: Big Data Engineer**

**Responsibilities:**

* Expertise in designing and deployment of Hadoop clusters and different Big Data analytic tools including Pig, Hive, HBase, Oozie, Sqoop, Flume, Spark, and Impala.
* Migrated data from On-Prem to cloud databases, using Teradata Utilities & Informatica for better loading, and created data pipelines using Azure data factory & Azure data bricks.
* Developed and deployed the outcome using Spark and Scala code in Hadoop cluster running on GCP.
* Used Snaplogic for child pipelines.
* Implemented advanced procedures like text analytics and processing using in-memory computing capabilities like Apache Spark written in Python.
* Build data pipelines in airflow in GCP for ETL-related jobs using different airflow operators, both old and newer operators.
* Implemented Spark using Python and Spark SQL for faster testing and processing of data.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, and Scala.
* Worked with Spark to create structured data from the pool of unstructured data received.
* Implemented intermediate functionalities like events or records count from the flume sinks or Kafka topics by writing Spark programs in Java and Python.
* Implemented information into CSV records and stored them in AWS S3 by using AWS EC2 and loading them into AWS Redshift.
* Developed Glue Jobs to read data in CSV format in the raw layer and write data to parquet format in publish layer.
* Developed and implemented data engineering solutions using DBT to transform raw data into clean, structured, and analytics-ready formats.
* Developed a Glue job that does deletes, updates, and incremental loads from source to target.
* Extracted Real-time feed using Kafka and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Experienced in transferring Streaming data, data from different data sources into HDFS, No SQL databases.
* Created ETL Mapping with Talend Integration Suite to pull data from Source, apply transformations, and load data into the target database.
* Created Data bricks notebooks using SQL, Python, and automated notebooks using jobs.
* Used DBT to test the data and ensure data quality.
* Used PySpark and Pandas to calculate the moving average and RSI score of the stocks and generated them into the data warehouse.
* Got involved in migrating the on-prem Hadoop system to using GCP (Google Cloud Platform).
* Developed multiple Kafka Producers and Consumers from scratch to as per the software requirement specifications.
* Optimized existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames, and Pair RDDs.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the AWS EMR Spark cluster.
* Extensive experience in building ETL jobs using Jupyter Notebooks with Apache Spark.
* Running analytics on power plant data using PySpark API with Jupyter notebooks in an on-premises cluster for certain transforming needs.
* Designed and developed data loading strategies and transformations for business to analyze the datasets.
* Experienced in writing Spark Applications in Scala and Python (PySpark).
* Implemented design patterns in Scala for the application.

**Environment:** Hadoop, Hive, Flume, DBT, Map Reduce, Sqoop, Kafka, Spark, Yarn, Cassandra, Oozie, shell Scripting, Scala, Maven, MySQL AWS (Lambda, Glue, EMR), GCP, NoSQL, Python, HDFS, Amazon Elastic Compute Cloud, Amazon Simple Storage Service(S3), Cloud Watch Triggers (SQS, Event Bridge, SNS), REST, ETL, Dynamo DB, JSON, Tableau.

**Client: Marsh McLennan, LA, California June 2018 – Aug 2019**

**Role: AWS Data Engineer**

**Responsibilities:**

* Worked on code migration of quality monitoring tool from Amazon EC2 to AWS Lambda and built logical datasets to administer quality monitoring on snowflake warehouses.
* Used Explains to optimize Teradata SQL queries for better performance.
* Used the Teradata tools Teradata SQL Assistant, Administrator and PMON extensively.
* Automation of cloud infrastructure using Terraform, and application configuration and deployment.
* Creating and managing access to AWS services for IAM user accounts and for role-based users.
* Used Tableau to design a dashboard to show operational metrics.
* Hands-on experience integrating AWS services: EC2, S3, Network Protocol, Transit VPC, VPC Peering, VPC Endpoints, VPC Private Link.
* Created Dax Queries to generate computed columns in Power BI.
* Used Power BI, and Power Pivot to develop data analysis prototype, and used Power View and Power Map to visualize reports.
* Developed multiple POC’s using Spark, and Scala and deployed them on the Yarn Cluster, compared the performance of Spark, with Hive and SQL.
* Published Power BI Reports in the required originations and created Power BI Dashboards available in Web clients and mobile apps.
* Partnered with DBT on the delivery of data definitions and aligning with the instance data conversion team.
* Improved the performance of dashboards and visualizing customer data using Amazon Quick Sight.
* Created monitors, alarms, notifications, and logs for Lambda functions, Glue Jobs, and EC2 hosts using CloudWatch.
* Used Amazon Elastic Cloud Compute (EC2) infrastructure for computational tasks and Simple Storage Service (S3) as a storage mechanism.
* Evaluated Snowflake design considerations for any change in the application.
* Worked on Spark data bricks cluster for estimating the cluster’s size, monitoring, and troubleshooting on the AWS cloud.
* Used Splunk to create dashboards, search queries, and reports for multiple applications.
* Implemented Installation and configuration of the multi-node cluster on the Cloud using AWS.
* Involved in PL/SQL query optimization to reduce the overall run time of stored procedures.
* Worked on Kibana dashboards based on log stash data and integrated several source and target systems into Elastic search for near real-time log analysis of end-to-end transaction monitoring.
* Integrated Apache Airflow with AWS to monitor multi-stage machine learning processes with Amazon Sage Maker jobs.

**Environment:** Hadoop, DBT, Map Reduce, HDFS, Python, Horton Works Data Platform, Power BI, CDH4, Spark, AWS, Apache NiFi, Oozie, HBase, JSON, CSV, XML, Hive, Sqoop, Pig, Splunk, MySQL, Jira.

**Client: TatvaSoft, India July 2016 – Feb 2018**

**Role: Sr Data Analyst / Python Developer**

**Responsibilities:**

* Performed Column Mapping, Data Mapping, and Maintained Data Models and Data Dictionaries.
* Collaborated with cross-functional team in support of business case development and identifying modeling method (s) to provide business solutions.
* Determines the appropriate statistical and analytical methodologies to solve business problems within specific areas of expertise.
* Integrated Teradata with R for BI platform and implemented corporate business rules.
* Participated in Business meetings to understand the business needs & requirements.
* Arrange and chair Data Workshops with SME’s and related stakeholders for requirement data catalogue understanding.
* Designed a Logical Data Model which will fit and adopt the Teradata Financial Logical Data Model (FSLDM11) using the Erwin data modeler tool.
* Generating Data Models using Erwin9.6 and developing a relational database system.
* Logical modeling using the Dimensional Modeling techniques such as Star Schema and Snowflake Schema.
* Guide the full life cycle of a Hadoop solution, including requirements analysis, platform selection, technical architecture design, application design and development, testing, and deployment.
* Consult on broad areas including data science, spatial econometrics, machine learning, information technology and systems and economic policy with R.
* Performed Data mapping between source systems to Target systems, logical data modeling, created class diagrams and ER diagrams and used SQL queries to filter data.
* Enabled speedy reviews and first-mover advantages by using Oozie to automate data loading into the Hadoop Distributed File System and PIG to pre-process the data.
* Used various techniques using R data structures to get the data in the right format to be analyzed which is later used by other internal applications to calculate the thresholds.
* Maintaining conceptual, logical, and physical data models along with corresponding metadata.
* Done data migration from an RDBMS to a NoSQL database and gives the whole picture for data deployed in various data systems.
* Developed triggers, stored procedures, functions, and packages using cursors and ref cursor concepts associated with the project using PL SQL.
* Used Meta data tool for importing metadata from repository, new job categories and creating new data elements.

**Environment:** R, Oracle 12c, MS-SQL Server, Hive, NoSQL, PL/SQL, MS- Visio, Informatica, T-SQL, SQL, Crystal Reports 2008, Java, SPSS, SAS, Tableau, Excel, HDFS, PIG, SSRS, SSIS, Metadata.

**Zoof Software Solutions – India Nov 2013 – June 2016**

**Role: ETL Developer**

**Responsibilities:**

* Analysed key business requirements via process flows, data analysis, workflow analysis, requirement workshops, business process storyboarding, helped translate customer need into specific requirements.
* Revitalized key metrics/KPIs to measure performance of business objectives, resulted in critical adjustments to achieve strategic goals.
* Built technology solutions to identify trends using data analytics, data visualization and data modelling, helped other departments, managers and executives make business decisions to improve and streamline operations.
* Devised risk-treatment plan (including Risk Analysis, Identification, Mitigation, and contingency plan) lead to on time project delivery with improved efficiency.
* Extensively participated in translating business needs into Business Intelligence reporting solutions by ensuring the correct selection of toolset available across the Tableau BI suite.
* Conduct code review to ensure the work delivered by the team is of high-quality standards.
* Maintain relationships with assigned customers post integration support their needs and build the relationship to encourage future growth of business with the customer.
* Used shell scripts and PMCMD commands to conduct basic ETL functionalities.

**Environment:** Informatics, Oracle 9i, Toad, Unix KSH, Tortoise SV.