***Candidate Details:***

**Name: Pranay Chaitanya**

**Email:** [**Pranaychaitanyareddy@gmail.com**](mailto:Pranaychaitanyareddy@gmail.com)

**Mobile: (737) 702-2051**

***Eployer Details:***

**Name: Rama Krishna Mettapalli**

**Email:** [**Ramakrishna@Cognierinc.com**](mailto:Ramakrishna@Cognierinc.com)

**Mobile: (404)-445-26*98***

**Professional Experience:**

* Around 8+ Years of experience working with Big Data technologies and hands on experience with both **AWS** and **Azure cloud** and hands on experience working with services such as **AWS EC2**, **EMR**, **S3** and **Azure Data Factory**, **Azure SQL Server**, **Blob Storage**, **Databricks** and **Snowflake**.
* Experienced in using distributed computing architectures such as **AWS** products (e.g., **EC2**, **Redshift**, and **EMR**, **Elasticsearch**), **Hadoop**, **Python**, **Spark**, and effective use of map-reduce, SQL and Cassandra to solve big data type problems.
* **Big Data - Hadoop** (**Map Reduce** & **Hive**), **Spark** (**SQL**, **Streaming**), **SQL Data warehouse**, **AWS** **Redshift, Athena, Lambda, Step Function** and **SQL**.
* Strong knowledge in **Spark ecosystems** such as **Spark** **core**, **Spark SQL**, **Spark Streaming** libraries.
* Expertise in **Business Intelligence**, **Data warehousing** technologies, **ETL** and **Big Data** technologies.
* Experience in writing **PL/SQL** statements - Stored Procedures, Functions, Triggers, and packages.
* Good knowledge in **Apache Hadoop ecosystem** components **Spark**, **Cassandra**, **HDFS**, **Hive**, **SQOOP**.
* Experience in building data pipelines using Azure Data factory, Azure data bricks and loading data to **Azure Data Lake**, **Azure SQL Database**, **Azure SQL Data warehouse** and controlling and granting database access.
* Proficiency in **SQL** across several dialects (we commonly write **MySQL, PostgreSQL**, **Redshift**, **SQL** **Server**, and **Oracle**)
* Experience in Database Design and development with Business Intelligence using **SQL Server** **2014/2016**, **Integration Services (SSIS**), DTS Packages, **SQL Server Analysis Services (SSAS)**, DAX, **OLAP** Cubes, Star Schema and **Snowflake** Schema.
* Strong skills in visualization tools **Power BI**, Microsoft Excel - formulas, **Pivot Tables**, Charts and DAX Commands.
* Proficient in **Data Analysis**, Cleansing, Transformation, **Data Migration**, Data Integration, Data Import, and **Data Export** through use of **ETL** tools such as **Informatica**, **SSIS**.
* Well experience in **Normalization** and **De-Normalization** techniques for optimum performance in relational and dimensional database environments.
* Skilled in **Tableau** **Desktop** **versions 10x** for data visualization, Reporting and Analysis.
* Developed reports, dashboards using **Tableau** for quick reviews to be presented to Business and IT users.
* Extensive knowledge in various reporting objects like Facts, Attributes, Hierarchies, Transformations, filters, prompts, Calculated fields, Sets, Groups, Parameters etc., in Tableau.
* Hands on learning with different **ETL** tools to get data in shape where it could be connected to **Tableau** through **Tableau** **Data Extract**.
* Analyzed data and provided insights with **R Programming** and **Python Pandas**
* Expertise in writing complex **SQL** queries, made use of Indexing, Aggregation, and materialized views to optimize query performance.
* Segmentation Analysis, Regression Models, and Clustering.
* Migration of on-premises data (**Oracle/ SQL Server/ DB2/ MongoDB**) to **Azure Data Lake Store** (**ADLS**) using **Azure Data Factory (ADF V1/V2).**
* Experience in designing star schema, Snowflake schema for Data Warehouse, ODS architecture.
* Good knowledge of Data Marts, OLAP, Dimensional Data Modeling with **Ralph Kimball** Methodology (Star Schema Modeling, Snow-Flake Modeling for FACT and Dimensions Tables) using Analysis Services.
* Strong analytical and problem-solving skills and the ability to follow through with projects from inception to completion.
* Ability to work effectively in cross-functional team environments, excellent communication, and interpersonal skills.

**Technical Experience:**

**Big Data:** HDFS, Map Reduce, Spark, Airflow, Yarn, NiFi, HBase, Hive, Pig, Flume, Sqoop, Kafka, Oozie, Hadoop, Zookeeper, Spark SQL.

**Languages:** SQL, PL/SQL, PYTHON, Java, Scala, C, HTML, Unix, Linux.

**Cloud Platform:** AWS (Amazon Web Services), Microsoft Azure.

**RDBMS:** Oracle 10g/11g/12c, Teradata, My SQL, MS SQL.

**NO SQL:** HBase, DynamoDB.

**Data Modeling Tools:** ERwin, Power Designer, Embarcadero ER Studio, IBM Rational Software Architect, MS Visio, ER Studio, Star Schema Modeling, Snowflake Schema Modeling, Fact and Dimension tables

**ETL Tools:** AWS Redshift, Informatica PowerCenter, SSIS, Azure Data Factory.

**Reporting Tools:** Power BI, SSRS and Tableau.

**Other Tools:** Azure Databricks, Azure Data Explores, Azure HDInsight.

**Project Experience:**

**Client: Liberty Mutual Group, Bloomington, IL Nov 2020 - Nov 2023**

**Title: Senior AWS Data Engineer**

**Responsibilities:**

* Involved in designing and deploying multi-tier applications using all the **AWS services** like (**EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM**) focusing on high-availability, fault tolerance, and auto-scaling in **AWS Cloud Formation**.
* Strong experience in working with **ELASTIC MAPREDUCE (EMR)** and setting up environments on **Amazon** **AWS EC2** instances.
* Developed **Spark Scala** and **python scripts** using both **Data frames/SQL/Data sets** and **RDD/MapReduce** in Spark for **Data Aggregation**, queries and writing data back into **OLTP** system through Sqoop.
* Developed **Hive** queries to pre-process the data required for running the business process.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of **Spark** using **Scala**.
* Extract 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 writing real-time processing and core jobs using Spark Streaming with **Kafka** as a data pipe-line system.
* Experience in building Real-time Data Pipelines with **Kafka** Connect and **Spark Streaming**.
* Developed multiple **Kafka** Producers and Consumers as per the software requirement specifications.
* Configured **Spark** streaming to get ongoing information from the **Kafka** and store the stream information to **HDFS**.
* **Hive** tables are created as per requirement were Internal or External tables defined with appropriate static, dynamic partitions and bucketing, intended for efficiency.
* Developed **Hive queries** for the analysts by loading and transforming large sets of structured, semi structured data using **hive**.
* Used **Spark** and **Spark-SQL** to read the parquet data and create the tables in **hive** using the **Scala API**.
* Experienced in using the **spark** application master to monitor the **spark** jobs and capture the logs for the **spark** jobs.
* Supporting Continuous storage in **AWS** using **Elastic Block Storage**, **S3, Glacier**. Created Volumes and configured Snapshots for **EC2** instances.
* Used **Data Frame API** in **Scala** for converting the distributed collection of data organized into named columns, developing predictive analytic using **Apache Spark Scala APIs**.
* Developed efficient **MapReduce** programs for filtering out the unstructured data and developed multiple **MapReduce** jobs to perform data cleaning and pre-processing on Hortonworks.
* Used Impala and Presto for querying the datasets.
* Collected data using **Spark Streaming** from **AWS S3 bucket** in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in **HDFS**.
* Used **impala** to query the data into the publish layers where all the other teams or business users can access for faster processing.
* Hands on experience working with **Databricks** runtime and performance tuning of **spark** jobs.
* Worked with delta tables on **Databricks**.
* Used Snowflake as **Datawarehouse database**.
* Exported data from **S3** to **Snowflake** and vice versa.
* Hands on experience working on performance tuning of **snowflake** jobs.
* Hands on experience working with airflow and used it for scheduling and orchestration of workloads.
* Worked with Docker and Kubernetes for running data pipelines in a runtime agnostic environment.

**Environment:** Hadoop, AWS, EC2, S3 Bucket, Redshift, Hive, HDFS, Spark, Spark-SQL, Spark-Streaming, Scala, KAFKA, Map Reduce, Impala, Zookeeper, ELK, Sqoop, Oracle 12c, Sql Server, T-SQL, Python and Agile Methodologies.

**Client: Tailored Brands-Fremont, California March 2018 - Oct 2020**

**Title: Senior Big Data Developer**

**Responsibilities:**

* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run **MapReduce** jobs in the **backend**.
* Involved in creating the notebooks for moving data from raw to stage and then to curated zones using **Azure data bricks**.
* Used **SQL Azure** extensively for database needs in various applications.
* Created multiple dashboards in tableau for multiple business needs.
* Installed and configured **Hive** and written **Hive UDFs** and used piggy bank a repository of **UDF’s** for **Pig Latin**.
* Participated in various upgradations and troubleshooting activities across enterprise.
* Knowledge in performance troubleshooting and tuning **Hadoop clusters**.
* Experience in building data pipelines using **Azure Data factory**, **Azure Databricks** and loading data to **Azure Data Lake**, **Azure SQL Database**, **Azure SQL Data warehouse** and Controlling and granting database access.
* Migrated complex **MapReduce** programs into in memory Spark processing using Transformations and actions.
* Used **Avro** serializer and **Avro De serializer** for developing the **Kafka** clients.
* Developed **Kafka** Consumer job to consume data from **Kafka** topic and perform validations on the data before pushing data into **Hive** and **Cassandra** databases.
* Applied **spark streaming** for real time data transforming.
* Create and maintain optimal data pipeline architecture in cloud **Microsoft Azure** using **Data Factory** and **Azure Databricks**.
* Hands on experience working with Delta tables.
* Building the pipelines to copy the data from source to destination in **Azure Data Factory**.
* Queried and analysed data from **Cassandra** for quick searching, sorting and grouping through **CQL**.
* Implemented various **Data Modelling** techniques when building **data pipelines**.
* Worked on data pre-processing and cleaning the data to perform feature engineering and performed data imputation techniques for the missing values in the dataset using Python.
* Created Data Quality Scripts using **SQL** and **Hive** to validate successful das ta load and quality of the data. Created various types of data visualizations using Python and Tableau.
* Create and deploy **Azure** command line scripts to automate tasks.
* Worked with **Spark** using **Scala** and **Spark SQL** for faster testing and processing of data.
* Applied Spark advanced procedures like text analytics and processing using the in-memory processing.
* Experienced in using **Tidal** enterprise scheduler and **Oozie Operational Services** for coordinating the cluster and scheduling workflows.
* Implemented Partitioning, Dynamic Partitions and Buckets in **HIVE** for efficient data access.
* Devised and lead the implementation of next generation architecture for more efficient data ingestion and processing.
* Created and implemented various shell scripts for automating the jobs.
* Implemented **Apache** Sentry to restrict the access on the hive tables on a group level.
* Worked in **Agile environment**, and used rally tool to maintain the user stories and tasks.
* Exported the analysed data to the relational databases using Sqoop for visualization and to generate reports for the **BI** **team** Using **Tableau**.
* Worked on data cleaning and reshaping, generated segmented subsets using Numpy and Pandas in Python.
* Worked with **Machine Learning** engineering teams to help them with **EDA** and feature extraction and also to provide them with the infrastructure to run the **ML models**.
* Implemented test scripts to support **test-driven development** and continuous integration.

**Environment**: Azure SQL, Azure Data Factory, Azure Databricks, Azure Data Lake, HDFS, Hive, pig, Impala, MapReduce, Spark, Python, Kafka, Tableau, Teradata, Pentaho, PIG, Sqoop.

**Client: Ford - Dearborn, MI Aug 2016 - Feb 2018**

**Title: Senior Data Engineer**

**Responsibilities:**

* Worked on **Big data** on **AWS** cloud services i.e. **EC2, S3, EMR** and **DynamoDB**.
* Used **Hive** **SQL**, **Presto SQL** and **Spark SQL** for **ETL** jobs and using the right technology for the job to get done.
* Compiled data from various sources to perform complex analysis for actionable results.
* Created **Entity Relationship Diagrams (ERD)**, Functional diagrams, Data flow diagrams and enforced referential integrity constraints and created logical and physical models using **Erwin**.
* Created ad hoc queries and reports to support business decisions **SQL Server Reporting Services** (**SSRS**).
* Analyse the existing application programs and tune **SQL** queries using execution plan, query analyser, **SQL** Profiler and database engine tuning advisor to enhance performance.
* Involved in the Forward Engineering of the logical models to generate the physical model using Erwin and generate **Data Models** using Erwin and subsequent deployment to Enterprise **Data Warehouse**.
* Wrote various data normalization jobs for new data ingested into **Redshift**.
* Created various complex **SSIS/ETL** packages to **Extract, Transform and Load data**.
* Advanced knowledge on Confidential **Redshift** and **MPP database** concepts.
* Migrated on premise database structure to Confidential Redshift data warehouse.
* Was responsible for **ETL** and **data** validation using **SQL Server Integration Services**.
* Defined and deployed monitoring, metrics, and logging systems on **AWS**.
* Connected to **Amazon Redshift** through **Tableau** to extract live data for real time analysis.
* Implemented **Work Load Management (WML)** in **Redshift** to prioritize basic dashboard queries over more complex longer-running adhoc queries. This allowed for a more reliable and faster reporting interface, giving sub-second query response for basic queries.
* Worked publishing interactive data visualizations dashboards, reports /workbooks on Tableau and **SAS Visual Analytics**.
* Developed **SSRS reports**, **SSIS packages to Extract**, Transform and Load data from various source systems
* Implementing and Managing **ETL** solutions and automating operational processes.
* Optimizing and tuning the **Redshift environment**, enabling queries to perform up to 100x faster for **Tableau** and **SAS Visual Analytics**.
* Defined facts, dimensions and designed the data marts using the **Ralph Kimball's** Dimensional Data Mart modelling methodology using Erwin.
* Analysed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing **ETL** changes.
* Implemented a Continuous Delivery pipeline with **Docker**, and **Git Hub** and **AWS**.
* Built performant, scalable **ETL** processes to load, cleanse and validate data.
* Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other **Agile methodologies**.
* Collaborate with team members and stakeholders in design and development of data environment.
* Preparing associated documentation for specifications, requirements, and testing.

**Environment:** AWS, EC2, S3, SQL Server, Erwin, Oracle, Redshift, Informatica, RDS, NOSQL, Snowflake Schema, MySQL, Dynamo DB, Docker, PostgreSQL, Tableau, Git Hub.

**Client: ALLIANCE Technosoft Private Limited-INDIA May 2013 - Jan 2015**

**Title: Data Engineer**

**Responsibilities:**

* Experience in building and architecting multiple Data pipelines, end to end **ETL** and **ELT** process for Data ingestion and transformation in **AWS**.
* Strong understanding of **AWS components** such as **EC2** and **S3**.
* Created **Hive, Phoenix, HBase** tables and **HBase** integrated **Hive** tables as per the design using **ORC** file format and Snappy compression.
* Devised simple and complex **SQL** scripts to check and validate **Dataflow** in various applications.
* Used Hive to implement data warehouse and stored data into **HDFS**. Stored data into **Hadoop** clusters which are set up in **AWS EMR**.
* Responsible for coding **Java Batch**, **Restful Service**, **MapReduce program**, **Hive query's**, testing, debugging, **Peer code review**, troubleshooting and maintain status report.
* Performed Data Preparation by using Pig Latin to get the right data format needed.
* Used Kafka and Kafka brokers, initiated the spark context and processed live streaming information with **RDD** and Used **Kafka** to load data into **HDFS** and **NoSQL** databases.
* Implemented a Continuous Delivery pipeline with **Docker** and **Git Hub**.
* Experience in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and **SCD (Slowly changing dimension).**
* Devised **PL/SQL** Stored Procedures, Functions, Triggers, Views, and packages. Made use of Indexing, Aggregation and Materialized views to optimize query performance.
* Created Session Beans and controller Servlets for handling **HTTP** requests from **Talend**.
* Used Git for version control with **Data Engineer team** and **Data Scientists** colleagues.
* Performed data engineering functions: **data extract, transformation**, **loading**, and **integration** in support of enterprise **data infrastructures** – **data warehouse**, operational data stores and master data management
* Implemented **Partitioning**, **Dynamic Partitions**, **Buckets** in **HIVE**.
* Develop database management systems for easy access, storage, and retrieval of data.
* Perform DB activities such as indexing, performance tuning, and backup and restore.
* Expertise in writing Hadoop Jobs for analysing data using **Hive QL (Queries), Pig Latin (Data flow language),** and **custom Map Reduce** programs in **Java**.
* Proficient in **Machine Learning** techniques (Decision Trees, Linear/Logistic Regressors) and **Statistical Modeling.**
* Used **PCA** to reduce dimension and compute eigenvalue and eigenvector and used OpenCV to analysis the **CT scan** pictures to figure out the disease in **CT scan**.
* Skilled in data visualization like **Matplotlib** and **seaborn** library.
* Experience implementing **machine learning** **back-end pipeline** with **Pandas**, **NumPy**.

**Environment**: HBase, Cloud Shell, AWS, Apache Airflow, Python, Pandas, Matplotlib, seaborn library, NumPy, ETL workflows, Python, Scala, Spark.