**Sravya**

**Senior Data Engineer**

**Ph No:** **(512) 221-2402**

**SUMMARY**

* Having **Around 9 years** of proven track record in the fields of **Business Intelligence** **Reporting,** Google cloud services, **Bigdata/Hadoop ETL** and supply chain product development.
* Hands on experience on Google Cloud Platform **(GCP)** in all the bigdata products **Big Query**, **Cloud Data Proc**, **Google Cloud Storage**, Composer (Air Flow as a service).
* Strong experience working with **HDFS,** **Map Reduce, Spark, Sqoop, Flume, Kafka, Oozie, Pig** and **HBase.**
* Experience with building data pipelines in **Python/PySpark/HiveQL/Presto/Big Query** and building python **DAG** in **Apache Airflow**.
* Expertise in **Creating, Debugging**, Scheduling and Monitoring jobs using Composer **Airflow**.
* Strong knowledge in data preparation, data modelling and data visualization using **Power BI** and had experience in developing various analysis services using **DAX** queries.
* Hands on experience with different programming languages such as **Python, R, SAS.**
* Expertise in transforming business requirements into analytical models, designing **algorithms,** building models, developing **Data Mining**, **Data Acquisition**, **Data Preparation,** **Data Manipulation**, Feature Engineering, **Machine Learning Algorithm**, Validation and Visualization and reporting solutions that scales across massive volume of structured and unstructured Data.
* Experience in using **SQOOP** for importing and exporting data from **RDBMS** to **HDFS** and Hive.
* Experience with **Unix/Linux** systems with bash scripting experience and building data pipelines.
* Excellent knowledge of studying the data dependencies using metadata stored in the repository and prepared batches for the existing sessions to facilitate scheduling of multiple sessions.
* Strong **SQL** development skills including writing Stored **Procedures, Triggers, Views**, and User Defined functions.
* Knowledge in various file formats in **hdfs** like **Avro, orc, parquet**.
* Extensive experience in writing **python** functions and advanced.
* Experience in building and architecting multiple **Data pipelines**, end to end **ETL** and **ELT** process for Data ingestion and transformation in **GCP** and coordinate task among the team.
* Have good Programming experience with **Python** and **Scala**.
* Hands in experience on **No SQL** database like **HBase, Cassandra**.
* Experience in using stack driver service/ dataproc clusters in **GCP** for accessing logs for debugging.
* Experience with scripting languages like **PowerShell, Perl, Shell,** etc.
* Orchestrated **Azure** Databricks applications using **ADF** pipelines and built data pipelines using **ADF** pipelines and Databricks.
* Experience in **GCP** **Datapost, GCS**, Cloud functions, **Big Query**, Azure Data Factory Data Bricks.
* Experience in building efficient pipelines for moving data between **GCP** and **Azure** using **Azure Data Factory.**
* Good understand and hands on experience with **AWS S3** and **EC2.**
* Strong written and oral communication skills for giving presentations to non-technical stakeholders.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Databases:** | Oracle, MySQL, SQLite, NO SQL, RDBMS, SQL Server, HBase, MongoDB, Teradata, Netezza, Cassandra. Alation, Data Governance. |
| **Database Tools:** | PL/SQL Developer, Toad, SQL Loader, Erwin. |
| **Web Programming:** | Html, CSS, Xml, JavaScript. |
| **Programming Languages:** | R, Python, SQL, Scala, UNIX, C, JAVA |
| **DWH BI Tools:** | Data Stage, Power BI |
| **Google Cloud Platform:** | GCP Cloud Storage, Big Query, Composer, Cloud Data proc, Cloud SQL, Cloud Functions, Cloud Pub/Sub. |
| **Data Visualization:** | Dataiku, Tableau. |
| **Big data Frameworks:** | HDFS, Map Reduce, Pig, Hive, Sqoop, Oozie, Zookeeper, Flume and HBase, Amazon EC2, S3 and Red Shift, Spark, Storm, Impala, Kafka. |
| **Scheduling Tools:** | Autosys, Control-M. |
| **Operating Systems:** | LINUX, UNIX. |
| **Environment:** | AWS, AZURE, Databricks. |
| **Reporting Tools:** | SSIS/SSRS/SSAS. |

**PROFESSIONAL EXPERIENCE:**

**Transunion CH, IL**

**Sr Data Engineer Jan 2023- Present**

**RESPOSBILITIES:**

* Developing **Spark** programs with **Python,** and applied principles of functional programming to process the complex structured data sets.
* Work in a fast-paced agile development environment to quickly analyze, develop, and test potential use cases for the business.
* Used **Python Boto 3** to configure the services **AWS glue, EC2, S3**.
* Got involved in migrating on prem **Hadoop system** to using **GCP** (Google Cloud Platform).
* Wrote scripts in **Hive SQL/Presto SQL**, using python plugin for both spark and presto for creating complex tables with high performance metrics like partitioning, clustering and skewing.
* Leveraged cloud and **GPU** computing technologies for automated machine learning and analytics **pipeline**s, such as **AWS, GCP.**
* Was involved in setting up of **Apache** airflow service in **GCP.**
* Worked with Hadoop infrastructure to storage data in **HDFS** storage and use **Spark / HIVE** **SQL** to migrate underlying **SQL** codebase in **AWS.**
* Converting **Hive/SQL** queries into Spark transformations using **Spark RDDs** and **Pyspark**
* Build data pipelines in airflow in **GCP** for **ETL** related jobs using different airflow operators both old and newer operators.
* Analyzing **SQL** scripts and designed the solution to implement using **PySpark.**
* Export tables from Teradata to **HDFS** using **Sqoop** and build tables in **Hive.**
* Loaded and transformed large sets of structured, semi structured and unstructured data using **Hadoop/Big Data concepts.**
* Worked with **Hadoop** ecosystem and Implemented **Spark** using **Scala** and utilized Data Frames and **Spark SQL API** for faster processing of data.
* Used **AWS** Glue for the data transformation, validate and data cleansing.
* Prepared **Data Mapping** Documents and Design the **ETL** jobs based on the **DMD** with required Tables in the Dev Environment.
* Was responsible for creating on-demand tables on **S3** files using Lambda Functions and **AWS Glue** using **Python** and **PySpark.**
* Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services **(AWS)** on **EC2.**
* Designed and developed architecture for data services ecosystem spanning Relational, **NoSQL**, and Bigdata technologies. Extracted Mega Data from Amazon **Redshift, AWS**, and **Elastic Search** engine using **SQL** Queries to create reports.
* Active participation in decision making and **QA** meetings and regularly interacted with the Business Analysts & development team to gain a better understanding of the Business Process, Requirements & Design.
* Designed and Developed Real time Stream processing Application using **Spark, Kafka**, **Scala,** and **Hive** to perform Streaming **ETL** and apply **Machine Learning**.
* Applying the Data modelling and Data Designing in-between staging and target for creating the views.
* Responsible for gathering requirements, system analysis, design, development, testing and deployment and Responsible to manipulate **HTML5, CSS3** in **jQuery** and provided dynamic functionality using **AJAX, XML and JSON.**
* Extensive knowledge on applying **Data Governance** controls.
* Worked on **SQL** Server concepts **SSIS** (SQL Server Integration Services), **SSAS** (Analysis Services) and **SSRS** (Reporting Services). Using Informatica & **SSIS, SPSS, SAS** to extract transform & load source data from transaction systems.
* Evaluated big data technologies and prototype solutions to improve our data processing architecture. Data modeling, development, and administration of relational and **NoSQL** databases (**Big Query, Elastic Search**)
* Utilized **Spark, Scala, Hadoop, HBase**, **Cassandra, MongoDB, Kafka, Spark** Streaming, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Experience with **Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales**, **PivotTables** and **OLTP** reporting.
* Extensive knowledge on **AQT tool.** Classifying the domains in Enterprise Data Warehouse.

**ENVIRONMENT:** IBM Info sphere DataStage, GCP, Oracle, Flat files, Autosys, UNIX, Erwin, TOAD, MS SQL Server database, XML files, AWS, MS Access database, AQT Tool.

**State of Hartford, Hartford CT**

**Sr Data Engineer Apr 2020- Dec 2022**

**RESPONSIBILITIES:**

* Installed **Hadoop, Map Reduce, HDFS, AWS S3** and developed multiple Map **Reduce** jobs in **PIG** and **Hive** for data cleaning and pre-processing.
* Architected, Designed and Developed Business applications and Data marts for reporting. Involved in different phases of Development life including Analysis, **Design**, **Coding, Unit** **Testing**, Integration Testing, Review and Release as per the business requirements.
* Created monitors, alarms, notifications and logs for **Lambda functions**, **Glue Jobs, EC2** hosts using **CloudWatch.**
* Migrating an entire oracle database to **Big Query** and using of power bi for reporting.
* Build data pipelines in airflow in **GCP** for **ETL** related jobs using different airflow operators.
* Experience in **GCP** Data proc, **GCS**, Cloud functions, **Big Query.**
* Experience in moving data between **GCP** and **Azure** using **Azure** Data Factory.
* Experience in building **Power BI** reports on **Azure** Analysis services for better performance.
* Used cloud **shell SDK** in **GCP** to configure the services **Data Proc, Storage, Big Query**
* Worked on batch processing of data sources using **Apache Spark,** Elastic search.
* Developed Big Data solutions focused on pattern matching and predictive modeling.
* Develop an Estimation model for various product & services bundled offering to optimize and predict the gross margin.
* Used **AWS** glue catalog with crawler to get the data from **S3** and perform **SQL** query operations.
* Designed **OLTP** system environment and maintained documentation of Metadata. Used forward engineering approach for designing and creating databases for **OLAP** model.
* Exploring with the **Spark** improving the performance and optimization of the existing algorithms in Hadoop using **Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.**
* A highly immersive Data Science program involving **Data Manipulation & Visualization**, **Web Scraping, Machine Learning, Python programming, SQL, GIT, Unix Commands, NoSQL, MongoDB, Hadoop.**
* Worked on migrating **PIG** scripts and **Map** Reduce programs to Spark Data frames **API** and Spark **SQL** to improve performance.
* Involved in creating **UNIX shell** scripts for database connectivity and executing queries in parallel job execution.
* Used the **ETL** Data Stage Director to schedule and running the **jobs**, testing and debugging its components & monitoring performance statistics.
* Worked closely with the **ETL Developers** in designing and planning the **ETL** requirements for reporting, as well as with business and **IT** management in the dissemination of project progress updates, risks, and issues.
* Handled importing data from various data sources, performed transformations using **Hive, MapReduce**, and **loaded data** into **HDFS.**
* Developed predictive causal model using annual failure rate and standard cost basis for the new bundled services.
* Design and develop analytics, machine learning models, and visualizations that drive performance and provide insights, from prototyping to production deployment and product recommendation and allocation planning.
* Worked with sales and marketing team for Partner and collaborate with a cross-functional team to frame and answer important data questions.
* Prototyping and experimenting **ML algorithms** and integrating into production system for different business needs.
* Design built and deployed a set of **python** modeling **APIs** for customer analytics, which integrate multiple machine learning techniques for various user behavior prediction and support multiple marketing segmentation programs.
* Designed, developed, and maintained data integration programs in **Hadoop** and **RDBMS** environment with both **RDBMS** and **NoSQL** data stores for data access and analysis.
* Used all major **ETL** transformations to load the tables through Informatica **mappings**.
* Created Hive queries and tables that helped line of business identify trends by applying strategies on historical data before promoting them to production.
* Worked on Data modeling, **Advanced SQL** with Columnar Databases using **AWS.**
* Extensively used **Apache Sqoop** for efficiently transferring bulk data between **Apache** Hadoop and relational databases **(Oracle)** for product level forecast. Extracted the data from Teradata into **HDFS** using **Sqoop.**
* Used classification techniques including Random Forest and Logistic Regression to quantify the likelihood of each user referring.
* Designed and implemented end-to-end systems for **Data Analytics** and Automation, integrating custom visualization tools using **R** and **Power BI**.

**ENVIRONMENT:**  Python, Spark framework, AWS, Redshift, MS Excel, NoSQL, Power BI, T-SQL, ETL, XML, MS SQL Server

**E TRADE, Menlo Park CA**

**Sr Data Engineer Dec 2018 -Apr 2020**

**RESPONSEBILITIES:**

* Developed multi cloud strategies in better using **GCP** (for its **PAAS**) and **Azure** (for its **SAAS)**.
* Involved in loading and transforming large sets of the structured, semi-structured dataset and analyzed them by running **Hive queries**.
* Developed custom python program including **CI/CD** rules for google cloud data Catalog for metadata management.
* Design and develop **Spark** job with **Scala** to implement end to end data pipeline for batch processing.
* Do fact dimensional modelling and proposed solution to load it.
* Processing data with **Scala, spark, spark SQL** and load in **hive** partition tables in parquet file format
* Develop **spark** job with partitioned **RDD** (like hash, range, custom) for faster processing.
* Develop and deploy the outcome using **spark** and **Scala** code in **Hadoop** cluster running on **GCP.**
* Develop near real time data pipeline using **flume, Kafka**, and spark stream to ingest client data from their **web log server** and apply transformation.
* Performs data analysis and design, and creates and maintains large, complex logical and physical data models, and metadata repositories using **ERWIN** and **MB MDR**
* Assist service developers in finding relevant content in the existing reference models.
* Utilized **Spark SQL API** in **Pyspark** to extract and load data and perform **SQL** queries.
* Worked on developing **Pyspark** script to encrypting the raw data by using hashing algorithms concepts on client specified columns.
* Responsible for Design, Development, and testing of the database and Developed Stored Procedures, Views, and **Triggers**
* Developed **Python**-based **API** (**RESTful Web Service**) to track revenue and perform revenue analysis.
* Compiling and validating data from all departments and Presenting to Director Operation.
* Develop **SQOOP** script and **SQOOP** job to ingest data from client provided database in batch fashion on incremental basis.
* Use **DISTCP** to load files from **S3** to **HDFS** and Processing, cleansing, and filtering data using **Scala, Spark, Spark SQL, HIVE**, Impala **Query** and Load in Hive tables for data scientists to apply their **ML** algorithms and generate recommendations as part of data lake processing layer.
* Build data pipelines in airflow in **GCP** for **ETL** related jobs using different airflow operators both old and newer operators.
* Created **Big Query** authorized views for row level security or exposing the data to other teams.
* Good knowledge in using cloud shell for various tasks and deploying services.
* Designed, developed, and maintained end-to-end **ETL** processes using **IBM** DataStage for extracting, transforming, and loading data from various sources into the data warehouse.
* Collaborated with data analysts and business stakeholders to understand data requirements and translated them into effective **ETL** workflows.
* Implemented complex data transformations and mappings, optimizing for performance and accuracy.
* Developed **PL/SQL** scripts and stored procedures for data manipulation, validation, and transformation.
* Played a key role in migrating on-premises **ETL** processes to Google Cloud Platform, leveraging **GCP's** data processing and storage services.
* Utilized Google Cloud Dataflow for building data pipelines, enabling real-time data processing and streaming.

**ENVIRONMENT:** AWS-S3, PL/SQL, ETL pipelines, Airflow, Python, Spark, HIVE, Spark SQL, Scala.

**Bayer, St. Louis MO**

**Data Engineer May 2016-Nov 2018**

**RESPONSEBILITIES:**

* Used panda’s library for data manipulation and **NumPy** for numerical analysis and managed large datasets using **panda’s** data frames and **MySQL**.
* Helping the client partially migrate from **MySQL** to Hadoop ecosystem.
* Define the data pipeline for various clients.
* Loading data in **No SQL** database **(HBase, Cassandra)**
* Using **YARN** in **CLOUDERA** manager to monitor job processing.
* Do participate in architecture council for database architecture recommendation.
* Deep analysis on **SQL** execution plan and recommend hints or restructure or introduce index or materialized view for better performance.
* Deploy **EC2** instances for oracle database.
* Worked briefly in writing **Pyspark** core methods for speeding up Hive **SQL** queries like non-equip joins.
* Used **Apache Sqoop** import and export and handled datatypes after moving.
* Did **POC's** by doing bucketing and partitioning on a hive table to understand performance.
* Wrote and executed various **MYSQL** database queries from **Python** using **Python-MySQL** connector and **MySQL DB** package.
* Used Collections in **Python** for manipulating and looping through different user defined objects.
* Developing **ETL** processes using Datastage to move data between different systems.
* Using **UNIX shell** scripting to automate data processing tasks and system management.
* Optimizing **SQL** queries for efficient data retrieval and analysis.
* Collaborating with other teams to ensure data **quality, integrity**, and **security.**

**ENIVIRONMENT:** EC2, Python, Spark, CI/CD, Jenkin, SQL, ETL.

**Ainslee Software, Hyderabad, India**

**Data Engineer July 2014 - Dec2015**

**RESPONSBILITIES:**

* Carried out data transformation and cleansing using **SQL** **queries, Python** and **Pyspark.**
* Was responsible for **ETL** and data validation using **SQL** Server Integration Services
* Built **SQL** queries to build the reports for presales and secondary sales estimations.
* Hands-on experience with building data pipelines in **python/Pyspark/hive SQL/Presto**.
* Have used **SAS** for data analysis as well as **Python** for building **ETL** pipelines using **panda’s** **framework.**
* Converted previously written **SAS** programs into **python** for one the **ETL** project.
* Carried out data transformation and cleansing using **SQL queries**, **Python** and **Pyspark.**
* Worked on the backend using **Python** and Spark to perform several aggregation logics.
* Created **Oracle** Stored Procedure to implement complex business logic for better performance.
* Developing **Python** programs that can run the **end-to-end** data migration and as well as transformation and load data into sinks such as oracle and **MySQL.**
* Developed **Python** scripts to create data files from database and post them to **FTP** server on daily bases using windows task scheduler.

**ENVIRONEMNT:** MySQL, Python, ETL, SAS, hive SQL.