**Revanth Ch**

**Senior DATA ENGINEER.**

**Email:** revanthc12@gmail.com

**Phone:** 9402778227

 

**Professional summary:**

* About 10 years of experience as Data Engineer in Data Systems installation and maintenance with large data sets of Structured and Unstructured data.
* Excellent Experience in Designing, Developing, Documenting, Testing of ETL jobs and mappings in Server and Parallel jobs using Datastage to populate tables in Data Warehouse and Data marts.
* Experience in improving Spark performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark Core, Spark-SQL, and Data Frames.
* Extensively used Python Libraries Pyspark, Pytest, Pymongo, cxOracle, PyExcel, Boto3, Psycopg, embedPy, NumPy and Beautiful Soup.
* Experience in AWS cloud Services (EMR, EC2, VPC, RDS, EBS, S3, Kinesis, Lambda, Glue, Athena, Elasticsearch, SQS, IAM, CloudFront, Cloud Watch, Autoscaling, DynamoDB, Redshift, ECS).
* Strong experience in Software Development Life Cycle (SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Strong experience in writing scripts using Python API, PySpark API and Spark API for analyzing the data.
* Hands-on use of Spark and ScalaAPIs to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala.
* Expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Experience in developing Map Reduce Programs using Apache Hadoop for analyzing the big data as per the requirement.
* Experience in usage of Hadoop distribution like Cloudera and Hortonworks.
* Deep understanding of MapReduce with Hadoop and Spark. Good working knowledge of Big Data ecosystems like Hadoop (HDFS, Hive, Pig, Impala), Spark (SparkSQL, Spark MLlib, Spark Streaming).
* Establishes and executes the Data Quality Governance Framework, which includes end - to-end process and data quality framework for assessing decisions that ensure the suitability of data for its intended purpose.
* Integrated Kafka with Spark Streaming for real time data processing.
* Skilled in performing data parsing, data manipulation and data preparation with methods including describing data contents.
* Extensive experience in Text Analytics, generating data visualizations using R, Python and creating dashboards using tools like Tableau.
* Experienced on Hadoop Ecosystem and Big Data components including Apache Spark, Scala, Python, HDFS, Map Reduce, KAFKA.
* Good Exposure on Apache Hadoop Map Reduce programming PIG Scripting and Distribute Application and HDFS.
* Good Knowledge on Hadoop Cluster architecture and monitoring the cluster.
* Have good knowledge in Job Orchestration tools like Oozie, Zookeeper and Airflow.
* Experience in working with Business Intelligence including Amazon Redshift and Azure Data Warehouse.
* Managing Databricks Notebooks, Delta Lake with Python, Delta Lake.
* Have extensive working knowledge of Databricks notebooks for manipulation and aggregation tasks.
* Strong Experience in GCP Kubernetes Engine and orchestrating Data Pipelines.
* Strong Hands-on experience in using GCP dataflow to create data pipelines in gcp environment.
* Experience in GCP cloud SQL utilization and GCP BigQuery.
* Experience in GCP DataProc for execution of Apache Spark and Hadoop Clusters.
* Extensive experience with databases like Oracle, Teradata, SQL Server, DB2, and MS Access.
* Experience in execution of Batch jobs through the data streams to Spark Streaming.
* Experience in creating and documenting Metadata for OLTP and OLAP when designing a system.
* Solid knowledge of Power BI and Tableau Desktop report performance optimization.
* Strong experience in writing SQL and Transact SQL programs for Stored Procedures, Triggers and functions.
* Excellent performance in building, publishing customized interactive reports and dashboards with customized parameters including producing tables, graphs, listings using various procedures and tools such as Tableau and user-filters using Tableau.
* Experience with Unix/Linux systems with scripting experience and building data pipelines.
* Experience with Data Analytics, Data Reporting, Ad-hoc Reporting, Graphs, Scales, PivotTables and OLAP reporting.
* Deep experience with the design and development of Tableau visualization solutions.

**Technical Skills:**

|  |  |
| --- | --- |
| **Category** | **Skills** |
| **Big Data** | Cloudera Distribution, HDFS, Yarn, Data Node, Name Node, Resource Manager, Node Manager, MapReduce, PIG, SQOOP, Kafka, Hbase, Hive, Flume, Cassandra, Spark, Storm, Scala, Impala, GCP BigQuery, Snowflake. |
| **Programming** | Python, PySpark, Scala, Java, C, C++, Shell script, Perl script, SQL, PL/SQL |
| **Databases** | Snowflake(cloud), Teradata, IBM DB2, Oracle, SQL Server, MySQL, NoSQL, GCP Cloud Sql. |
| **Cloud Technologies** | AWS(EMR, EC2, RDS, Athena, IAM,S3, Glue, Redshift, Cloud watch, Cloud formation, Lambda, Sagemaker ), Google Cloud Platform (GCP AutoML, GCP Cloud Dataflow, GCP Cloud SQL, GCP DataPrep, GCP Dataflow, GCP DataProc, GCP IAM, GCP Kubernetes Engine, GCP Pub/SubGCP Secret Manager, GCP Stackdriver). |
| **Frameworks** | Django REST framework, MVC, Hortonworks |
| **ETL/Reporting** | Ab Initio, Informatica, Tableau, GCP Dataflow. |
| **Tools** | PyCharm, Eclipse, Visual Studio, SQL\*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Server Management Studio, SQL Assistance, Eclipse, Postman |
| **Machine Learning Techniques**:  | Linear Regression, Logistic Regression, Classification and Regression Trees, Random Forest, Associative rules, NLP and Clustering. |
| **Database Modelling** | Dimension Modeling, ER Modeling, Star Schema Modeling, Snowflake Modeling |
| **Visualization/ Reporting** | Tableau, ggplot2, matplotlib, SSRS and Power BI |
| **Web/App Server** | UNIX server, Apache Tomcat |
| **Operating System** | UNIX, Windows, Linux, Sun Solaris |

**Professional Experience:**

 **Fiserv, Frisco June 2021- present.**

 **Data Engineer**

 **Responsibilities:**

* Development of Pyspark applications for data extraction, transformation, and aggregation from multiple sources.
* Utilized Agile and Scrum methodology for team and project management.
* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, Lambda, Cloud Watch, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
* Worked on Kafka REST API to collect and load the data on Hadoop file system and also used sqoop to load the data from relational databases.
* Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
* Developed a NiFi Workflow to pick up the data from Data Lake as well as from server and send that to Kafka broker.
* Built various jobs using AWS Glue’s crawler capabilities to perform data cataloging and building the ETL pipelines for the target data mart.
* Responsible for building the data ingestion pipelines using AWS EMR and Spark Scala as the Data Processing Engine and AWS Athena as the Consumption Layer.
* Used PySpark and Pandas to calculate the moving average and RSI score of the stocks and generated them into a data warehouse.
* Exploring with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, PostgreSQL, Data Frame, OpenShift, pair RDD.
* Performed data analysis on the raw data residing on HDFS and local storage using Hive queries and Sqoop to import data between Hadoop and RDBMS.
* Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline which can be written to Glue Catalog and can be queried from Athena.
* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Integrated Apache Airflow with AWS to monitor multi-stage ML workflows with the tasks running on Amazon SageMaker.
* Used Jenkins for CI/CD, Docker as a container tool and Git as a version control tool.
* Worked on Dimensional and Relational Data Modelling using Star and Snowflake Schemas, `OLTP/OLAP system, Conceptual, Logical and Physical data modeling using Erwin.
* Generated report on predictive analytics using Python and Tableau including visualizing model performance and prediction results.
* Performance tuning of Spark application for setting right batch interval time, correct level of parallelism and memory tuning.
* Application design for integration with REST API’s, Merchant UI, and custom python libraries.
* Visualization of store layout with adjacency, left/right, opposite &amp, perpendicular mapping.
* Automation of ETL processes to wrangle the data, posting recommendation data store by store.
* Automated store data analysis by department and validation, using Snowflake as the underlying data repository.

**Environment:** Spark (PySpark, SparkSQL, SparkMLIib), Python 3.x (Scikit-learn, Numpy, Pandas), Snowflake, Tableau 10.1, GitHub, AWS EMR/EC2/S3/Redshift, and Pig.

 **AT&T - Texas May 2019 – May 2021**

 **Data Engineer**

 **Responsibilities:**

* Worked with Data ingestion, querying, processing, and analysis of big data.
* Collaborated with a development team following Agile methodologies, employing Git for version control and code collaboration.
* Utilized Apache Airflow to create DAGs (Directed Acyclic Graphs) in Python, allowing the management and scheduling of data workflows in a scalable and reliable manner.
* Utilized Spark and GCP DataProc, a managed Spark and Hadoop service on GCP, to perform data transformations, processing, and analysis efficiently.
* Employed Apache Nifi, an open-source data integration tool, to facilitate the movement of raw data between various systems within the GCP environment, ensuring smooth and controlled data flow.
* Utilized GCP IAM to design roles and groups, assigning appropriate permissions and access controls to users and resources, ensuring secure data handling and management.
* Leveraged GCP Stackdriver to monitor system resources such as compute instances, memory usage, database services, and set up alerts and notifications for automated actions.
* Implemented PySpark to perform fast and scalable data processing, enabling efficient analysis and transformation of large datasets.
* Engaged in the maintenance and optimization of the big data pipeline, ensuring its reliability, performance, and scalability.
* Utilized GCP Dataflow for executing batch and streaming data processing pipelines, to load and transform large sets of data.
* Connected data sources to Tableau, enabling the creation of interactive and visually appealing reports and dashboards.
* Employed GCP Cloud SQL for executing SQL scripts, creating stored procedures, views, tables, and managing database objects efficiently.
* Utilized Python Pandas library for data preprocessing and feature engineering, preparing the data for further analysis and machine learning models.
* Implemented ETL migration workflows using GCP Cloud Dataflow to transform and move data between different data storage systems, such as Cloud Storage, BigQuery, and Cloud SQL.
* Utilized Git for version control and managed source code and tracked changes to ensure efficient collaboration among team members.
* Implemented Docker for containerization, created Docker images that encapsulated the application and its dependencies, ensuring consistent deployment across different environments.
* Configured Jenkins jobs to automate build, test, and deployment processes, triggered by code commits or scheduled intervals.
* Utilized GCP Secret Manager to securely store and manage sensitive data and credentials used in the project.
* Implemented real-time data processing pipelines using GCP Pub/Sub, a messaging service, and GCP Dataflow to process and analyze streaming data.
* Leveraged GCP Kubernetes Engine to optimize and run PySpark jobs on a Kubernetes cluster, enabling faster data processing and analysis.
* Utilized GCP DataPrep for data cleaning and GCP AutoML for data modeling, enabling efficient data preparation and machine learning model development.
* Implemented streaming data processing using GCP Dataflow and Apache Kafka, enabling real-time data ingestion, processing, and analysis.
* Imported data from file-based systems and relational databases into the data lake storage in standard file formats such as Apache Parquet using GCP Stackdriver and Databricks.
* Deployed the trained and optimized model using Databricks' deployment capabilities.
* Used Pyspark to perform EDA and data manipulation tasks in Databricks notebooks.

 **Environment:** Apache Airflow, Apache Kafka, Apache Nifi, Docker, GCP AutoML, GCP Cloud Dataflow, GCP Cloud SQL,

 GCP DataPrep, GCP Dataflow, GCP DataProc, GCP IAM, GCP Kubernetes Engine, GCP Pub/Sub

 GCP Secret Manager, GCP Stackdriver, Git, Jenkins, PySpark, Python, Tableau, Databricks.

 **Equitable, New York May 2017 – Apr 2019**

 **Data Engineer**

 **Responsibilities:**

* Extensively used Agile methodology as the Organization Standard to implement the data Models.
* Developed Airflow dags and plugins to support requirements.
* Loading data into the Hadoop distributed file system (HDFS) with the help of Kafka and REST API
* Worked on AWS and BIG Data Technologies like HDFS, HIVE, SQOOP, EMR, SPARK AWS, REDSHIFT, EMR, EC2, RDS, Athena, IAM, DATA PIPELINE.
* Worked on the Spark SQL and Spark Streaming modules of Spark and used Scala to write code for all Spark use cases.
* Worked on Apache NiFi like executing Spark script, Sqoop scripts through NiFi, worked on creating scatter and gather pattern in NiFi.
* Supporting Continuous storage in AWS using Elastic Block Storage, S3, Glacier, Created Volumes and configured Snapshots for EC2 instances.
* Used PySpark and Pandas to calculate the moving average and RSI score of the stocks and generated them into a data warehouse.
* Generates ETL scripts to transform, flatten, and enrich the data from source to target using AWS Glue and created event- driven ETL pipelines with AWS Glue.
* Developed data pipeline using big data/Hadoop tools Spark, Pig, Impala, and HBase to ingest customers.
* Performed CI/CD operations with Gitlab pipelines, Jenkins and Docker.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python.
* Configure Flume to ingest log file data into HDFS.
* Created an internal tool for comparing the RDBMS and Hadoop such that all the data is located in source.
* Used Java, Python and OOPS concepts to write PySpark codes.
* Created tables, views, sequences, triggers, table spaces, constraints and generated DDL scripts for physical implementation.
* Normalized the incoming files of various sources to 3NF at the staging area before loading into facts and dimension tables.
* Involved in importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
* Wrote generic SQL Procedures and Complex T-SQL statements to achieve the reports generation.
* Involved in the development of PL/SQL stored procedures, functions, and packages to process business data in OLTP system.
* Involved in the creation, maintenance of Data Warehouse and repositories containing Metadata.
* Analyzed the data using Statistical features in Tableau to develop trend analysis.
* Assisted in Datastage Server jobs to load data from sequential files, flat files, and MS Access.

**Environment:** AWS EMR, S3, RDS, Redshift, Lambda, DynamoDB, Amazon SageMaker, Apache Spark, HBase, Apache Kafka, HIVE, SQOOP, Map Reduce, Apache Pig, Python, Tableau.

 **M&T Bank, New York May 2015 – April 2017**

 **Data Engineer**

 **Responsibilities:**

* Spearheaded a complex ETL extraction project from an on-premises Oracle database to the cloud, leveraging Google Cloud Platform (GCP) services and open-source technologies.
* Designed and implemented a robust data extraction strategy using Apache NiFi, extracting data from the on-premises Oracle database and converting it to Parquet format for optimization.
* Orchestrated seamless data transfer from on-premises to Google Cloud Storage, ensuring data security with encryption at rest using Cloud Storage's default encryption feature.
* Employed Google Cloud Dataflow (Apache Beam) and PySpark to perform extensive data transformations, including data cleansing, aggregations, and enrichment, handling complex data structures and nested JSON data effectively.
* Developed and maintained a scalable and fault-tolerant Dataflow pipeline in Python, effectively utilizing the dynamic nature of the Dataflow model for adaptive processing.
* Leveraged the power of Apache Spark on Google Kubernetes Engine (GKE) to handle large-scale data processing tasks, taking advantage of Kubernetes' auto-scaling capabilities to optimize resource utilization.
* Utilized Cloud Dataproc for big data processing jobs, enhancing data processing performance by leveraging Spark's distributed computing capabilities on a managed cluster.
* Designed and implemented a schema-on-read approach using BigQuery to efficiently load and analyze transformed data, enabling fast and flexible querying for analytical purposes.
* Implemented partitioning and clustering strategies in BigQuery to optimize query performance and reduce costs for analytical workloads.
* Ensured data accuracy and consistency by implementing data quality checks and monitoring mechanisms throughout the ETL pipeline.
* Incorporated Airflow as the ETL orchestration tool, scheduling and managing the entire data workflow while facilitating error handling and reruns when necessary.
* Collaborated with cross-functional teams, including data engineers, data analysts, and business stakeholders, to gather requirements and provide actionable insights based on transformed data stored in BigQuery.
* Documented the entire ETL process, data pipeline architecture, and technical specifications, enabling seamless knowledge transfer to new team members and facilitating future enhancements.
* Proactively optimized and fine-tuned the ETL pipeline, reducing processing times by 30% and significantly lowering overall operational costs.
* Conducted knowledge-sharing sessions to enhance team members' understanding of GCP services and best practices in cloud-based ETL and big data processing.

**Environment:** Google Kubernetes Engine (GKE), Google Cloud Storage, Google Cloud Dataflow, BigQuery, Cloud Dataproc,

and Cloud Composer (Airflow), Apache Spark, PySpark,SparkSQL and Apache NiFi.

 **Quiddity Infotech, Hyderabad, India June 2013 to December 2014**

 **Data Engineer/Analyst**

 **Responsibilities:**

* Involved in logical and Physical Database design &amp, development, Normalization and Data

 modeling using Erwin and SQL Server Enterprise manager.

* Followed agile methodologies and implemented them on various projects by setting up Sprint for every two weeks and daily stand-up meetings.
* Installed Oozie workflow engine to run multiple Hive and Pig Jobs.
* Implemented real-time data driven secured REST APIs for data consumption using AWS Cloud and API.
* Worked on creating data ingestion pipelines to ingest huge amount of Stream and customer application data into Hadoop in various file formats like raw text files, CSV, and ORC.
* Developed data pipeline using big data/hadoop tools Sqoop, pig and MapReduce to ingest customer behavioral.
* Created a task scheduling application to run in an EC2 environment on multiple servers.
* Developed Talend Big Data jobs to load heavy volumes of data into S3 data lake.
* Creating ETL mappings and enhancing existing mappings to facilitate the data load in the system.
* Involved in working with NoSQL databases like MongoDB, HBase and Cassandra.
* Implemented Sqoop jobs for large data exchanges between RDBMS (Mysql, Oracle) and Hive clusters.
* Worked with continuous Integration of applications using Jenkins.
* Worked on creating various repositories and version control using GIT.
* Worked extensively with Dimensional modeling, Data migration, Data cleansing, ETL Processes for data warehouses.
* Created action filters, parameters and calculated sets for preparing dashboards and worksheets using PowerBI.
* Loaded data into Spark RDD and in memory data Computation to generate the Output response.
* Developed Impala queries to pre-process the data required for running the business process.
* Actively involved in design analysis, coding, and strategy development.
* Developed Hive scripts for implementing dynamic partitions and buckets for history data.
* Developed Spark scripts by using Scala per the requirement to read/write JSON files.
* Involve in converting SQL queries into Spark transformations using Spark RDDs and Scala.
* Analyzed the SQL scripts and designed the solution to implement using Scala.
* Applied various machine learning algorithms like decision trees, regression models, neural networks, SVM, clustering to identify fraudulent profiles using scikit-learn package in Python.
* Used clustering technique K-Means to identify outliers and to classify unlabeled data.
* Used Spark API using Scala over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Created a Hadoop design which replicates the Current system design.
* Developed Scala scripts using both Data frames/SQL and RDD/MapReduce in Spark for Data aggregation and queries.

**Environment:** Hadoop, Python, Scala, SQL, Maven, AWS (S3, EC2), MongodB, MYSQL, Agile.