**Data Engineer**

**Name: Adithya**

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

**Contact:** (402) 909-1720

## Professional Summary:

* 6+ years of IT experience in a range of businesses working on Big Data technology utilizing Cloudera and AWS and GCP Technologies. Hadoop, Spark, MapReduce, Kafka, Hive, Sqoop are all part of a large working environment.
* Worked in End-to-End Software Development Life Cycle for around 6 years, which includes Requirement Gathering, Analysis, Design, Development, Test planning, Test Case Development, Test Execution, and Test cycle closure.
* Good knowledge of Hadoop architecture and **Hadoop** components such as **HDFS** and **MapReduce**.
* Strong understanding of Spark architecture and various Spark components such as **Spark Core, SparkSQL and Structured Streaming**.
* Hands-on experience developing **Apache Spark** jobs using **Spark SQL** and **Hive** for large-scale data processing and analytics.
* Hands-on experience in setting up workflows using **Apache Airflow**.
* Hands-on experience with cloud services like Amazon Web Services (AWS) which includes **Elastic MapReduce** for compute and **S3** for storage.
* Hands-on experience with cloud services like **GCP** Cloud Storage, Pub/Sub, or IoT Core to ingest data from various sources into the **GCP** environment.
* Experience using **Pub/Sub** as a messaging service to handle real-time data streams.
* Experience working on creating and running **Docker**images with multiple **micro services**.
* Experienced in data manipulation using **python** for loading and extraction as well as with **python libraries** such as **NumPy, SciPy**  for data analysis and numerical computations.
* Good understanding of **SCD Type - 1** and **Type - 2** operations and their implementation using **Delta Lake** API.
* Experienced in handling different file formats like JSON file, CSV file, Avro file, parquet file formats.
* Experienced in using join’s like left outer join, inner join, Broadcast join in SPARK/ HIVE and Dynamic Partitioning and Bucketing in hive.
* Experience in working with window analytical functions like RANK, DENSE RANK, Row Number, Lead, lag, etc.
* Implemented Data Ingestion using Sqoop into AWS S3 from RDBMS and vice-versa.
* Involved in loading the data from S3 buckets and processing the data using AWS (EMR) and analytics using Spark and, hive.
* Hands-on experience in **SQL and NOSQL** databases such as **Snowflake, HBase, Cassandra and MongoDB.**
* Hands-on experience in DevOps tools like **Git/GitHub** **and Jenkins**.
* Have good interpersonal, communication skills, strong problem-solving skills, explore/adapt to new technologies with ease and be a good team member.

**Education:**

Masters in computer and information Systems, Bellevue University.

## TECHNICAL KNOWLEDGE:

| **Databases** | Snowflake, AWS RDS, Teradata, Oracle, MySQL, Microsoft SQL, PostgreSQL. |
| --- | --- |
| **NoSQL Databases** | Hadoop HBase and Apache Cassandra. |
| **Programming Languages** | Python, SQL, Scala , R,Shell Scripting |
| **Cloud Technologies** | AWS S3,EMR,SNS,SQS , Docker,GCP Pub/Sub,Dataproc,Bigquery,Dataflow |
| **Data Formats** | CSV, JSON,Parquet,RC,ORC |
| **Querying Languages** | SQL, NO SQL, PostgreSQL, MySQL, Microsoft SQL |
| **Integration Tools** | Jenkins |
| **Scalable Data Tools** | Hadoop, Hive, Apache Spark, Map Reduce, Sqoop, Power Query,Databricks |
| **Workflows** | Airflow,Oozie |
| **Operating Systems** | Red Hat Linux, Unix, Windows, macOS. |
| **Reporting & Visualization** | Tableau, Power BI, Matplotlib |

**Nike Inc, Beaverton, OR Feb 2023 -Till date**

**Data Engineer/Analyst**

**Responsibilities:**

* Worked on cost-optimization tasks like reducing the **EMR** cluster size based on its utilization in different environments.
* Worked on writing generic **Spark** job which converts data from **json/parquet** format to **delta** formats.
* Worked on writing validation scripts for data converted from **json /parquet** to **delta.**
* Worked on migrating spark jobs from **EMR** to **Databricks**.
* Worked on converting **hive** scripts running on **EMR** to **spark** in **databricks.**
* Worked on creating **kafka** topics that read data from **kafka** topics and write it to **s3** buckets.
* Worked on creating and scheduling DAGs in **airflow** to run **spark** jobs on **Databricks**.
* Hands on experience in scheduling **CRON** jobs on EMR clusters using **Apache Airflow**.
* Hands on experience on monitoring the **Spark** applications through the **Spark UI** and identifying executor failures, data skew and other runtime issues.
* Experience in handling **JSON** datasets and writing custom Python functions to parse through **JSON** data using **Spark**.
* Used **Spark Streaming** to divide streaming data into batches as an input to **Spark** engine for batch processing.
* Involved in writing **Pyspark** code using higher order functions for iterative algorithms in **Spark** for Performance considerations.
* Worked on **PySpark APIs** for data transformations.
* Created **HIVE Queries** to process large sets of structured, semi-structured and unstructured data and store in Managed and External tables.
* Implemented the project under **Agile** Project Management Environment and followed SCRUM iterative incremental model & configured various sprints to execute.
* Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

**Environment:** Spark, AWS, ETL (DataStage, Talend, AWS Glue) Hadoop, Python, Snowflake,HDFS, Hive, MapReduce, PySpark, Docker, GitHub, Apache Spark, Databricks, JSON, Parquet, PostgreSQL, SQL, Jenkins, GitLab, Agile and macOS.

**Prudential, Newark, NJ Aug 2021 -Feb 2023**

**Data Engineer**

**Responsibilities:**

* Worked in a data ingestion team that was responsible for ingesting a wide variety of datasets like **CSV**, **JSON** and **Parquet**.
* Hands-on experience processing deeply **nested** **JSON** datasets using **SparkSQL** and flattening complex JSON structures to parquet format to query them through **Hive** tables.
* Worked on designing and implementing **staging, raw and consumption** layers of part of the data pipeline.
* Hands-on experience in monitoring the Spark applications through the **Spark UI**.
* Worked on implementing **UDF’s** using **PySpark**.
* Worked with Spark to import and export data from **S3** to Relational Database systems like Oracle and vice-versa.
* Developed data processing jobs using **Spark** and **Spark SQL** to populate staging tables and store the refined data in partitioned tables in the Enterprise Data Warehouse.
* Created **Hive** tables and implemented **dynamic and static partitioning** strategies and other performance tunings for faster data access.
* Created **Spark SQL** queries that helped analysts spot emerging trends by comparing new data with reference tables and historical metrics.
* Working experience in tuning Spark jobs and having a good understanding of Spark **Performance optimization** techniques.
* Hands-on experience in setting up workflows using **Apache Airflow** to Databricks.
* Used **Apache Kafka** to aggregate web log data from multiple servers and make them available in downstream systems for analysis.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through **Python**.
* Optimized Hadoop jobs to improve performance and reduce data processing time by 25% using techniques such as data partitioning, compression, and bucketing.
* Designed and implemented a data warehouse solution using **Snowflake,** reducing storage cost by 30% with improved data access and analysis.
* Implemented a Continuous Delivery pipeline with **Docker**, and **GitHub** and **AWS.**
* Implemented automated CI/CD pipelines using GitLab reducing deployment by 50% and collaborated closely with DevOps team to ensure data accuracy and completeness.
* Followed **AGILE** (SCRUM) methodologies, had sprint planning every two weeks and set up daily meetings to monitor the status.
* Participated in the status meetings and status updating to the management team.

**Environment:** Spark, AWS, ETL Hadoop, Python, Snowflake,HDFS, Hive, MapReduce, PySpark, JSON, XML, Apache Kafka, PostgreSQL, SQL, PL/SQL, GitLab, Jenkins, Agile and Windows

**TCS, India Sep 2018 - Aug 2021**

## Data Engineer

## Responsibilities:

* Responsible for designing and implementing ETL process to load data from different sources, perform data mining and analyze data using visualization/reporting tools to leverage the performance of OpenStack.
* Developed **ETL** framework using **Spark and Hive** (including daily runs, error handling, and logging) to useful data.
* Good understanding of python scripts and developed the spark scripts using PySpark.
* Worked as part of an agile team serving as a developer to customize, maintain and enhance a variety of applications for Hadoop.
* Used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
* Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting.
* Worked on Sqoop for importing and exporting the data from HDFS to Relational Database systems like Oracle and vice-versa.
* Developed business specific custom UDF in Hive.
* Worked with cloud services like Amazon EC2, S3 and Elastic MapReduce (EMR).
* Hands-on experience in creating RDD’s using Apache Spark, Spark Streaming and applying operations like Transformations and Actions on them.
* Developed Spark jobs written in Scala & Python to perform operations like data aggregation, data processing and data analysis.
* Experience in handling JSON datasets and writing custom Python functions to parse through JSON data using Spark.
* Worked extensively on performance tuning of Spark jobs and Hive/Impala queries.
* Hands on experience on monitoring the Spark applications through the Spark UI and identifying executor failures, data skew and other runtime issues.
* Responsible for loading and transforming large sets of structured, semi structured, and unstructured data.
* Actively participated and provided feedback in a constructive and insightful manner during weekly Iterative review meetings to track the progress for each iterative cycle and figure out the issues.

**Environment:** Spark, Scala, ETL, Python, AWS(EC2,EMR,S3), HDFS, Hive, Kafka, CSV, JSON, PySpark, SQL, Power Query, Power BI, Jira and Windows.

**SysIntelli,Hyderabad,India August 2017- Sep 2018**

**Java Developer**

**Responsibilities:**

* Created design documents and reviewed with team in addition to assisting the business analyst project manager in explanations to line of business.
* Involved in Different Phases To Gather Requirements,document the functional specifications, design, data modeling, development, testing of the applications.
* Designed and developed the UI using JSP, HTML, CSS.
* Used Tomcat web server for deployment purpose.
* Involved in creation of Test Cases for JUnit Testing.
* Used Oracle as Database and used Toad for queries execution and involved in writing SQL scripts, PL/SQL code for procedures and functions.
* Developed an application using Eclipse and used a build and deploy tool as Maven.
* Responsible for Testing and moving the application into Staging and Production environments.
* Responsible for Project Documentation, Status Reporting and Presentation.
* Used CVS version control to maintain the Source Code.

**Environment:**J2EE/Java, JSP, Spring MVC, Hibernate, Eclipse 3.0, MySQL, JDBC, JBoss, ANT, JavaScript, XML, SOAP, WSDL