**Senior Data Engineer**

**Abhinav Reddy**

**Phone:** +1 (269)-220-0418

**Email:** abhinavreddy.kumbham@gmail.com

**Professional Summary:**

* 8+ years of IT experience as Data Engineer with strong expertise in Hadoop ecosystem components such as HDFS, MapReduce, Yarn, HBase, Pig, ETL Data Stage, and Cloud technologies like AWS, Azure, Snowflake, Sqoop, Spark SQL, Spark Streaming, and Hive for scalability, distributed computing, and high-performance computing.
* Experience in configuring Spark Streaming to receive real-time data from Apache Kafka and storing the stream data to HDFS and expertise in using spark-SQL with various data sources like JSON, Parquet, and Hive.
* Extensively used Spark Data Frames API over the Cloudera platform to perform analytics on Hive data and used Spark Data Frame Operations to perform required data validation.
* Expertise in documenting sources to target data mappings and Business rules associated with the ETL processes.
* Experience in deploying agile development methodology and actively participated in daily scrum meetings.
* Experience with Spark Core, Spark SQL, and its core abstraction layer API like RDD and Data frame.
* Experience in gathering customer requirements, writing test cases, and partnering with developers to ensure a complete understanding of internal/external customer needs.
* Experience in building data engineering pipelines, automating and fine-tuning for batch and real-time data pipelines.
* Good experience of streaming processes like Kafka and Spark
* Experience working with Amazon Web Services (EC2, S3, RDS, Elastic Load Balancing, Elastic Search, Lambdas, SQS, Cloud Watch, and Glue).
* Experience working on AWS Databases like Elastic Cache (Memcached& Redis) and NoSQL databases HBase, Cassandra & MongoDB for database performance tuning & data modelling.
* Experience in designing detailed migration plans for workloads from any platform to AWS.
* Good experience in developing AWS centralized logging solutions for security teams to consolidate AWS logs and analyze them to detect incidents, using elastic search, EC2 server logs.
* Experience in Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault, and Azure Data Lake.
* Experience with Azure transformation projects and implementing ETL and data solutions using Azure Data Factory (ADF) and SSIS.
* Expert in recreating existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database, and SQL Data warehouse environment.
* Expertise in HQL queries to do data analytics on top of Bigdata.
* Good experience with Python libraries such as Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn to analyse the insights of data and perform data cleaning, data visualization and build models.
* Hands on experience in creating a private cloud using Kubernetes that supports DEV, TEST, and PROD environments.
* Experience in developing complex shell/Python scripts in Linux and DevOps tools like Jenkins, Maven, Terraform, Ansible, Docker, and Kubernetes.
* Expertise in writing Sqoop jobs to migrate vast amounts of data from Relational Database Systems to HDFS/Hive Tables and vice-versa according to client’s requirement.
* Experience on NoSQL databases and their integration with Hadoop clusters to store and retrieve vast amounts of data.

**Technical Skills:**

| **Hadoop Distribution** | Cloudera distribution and Hortonworks |
| --- | --- |
| **Hadoop/Spark Eco-system** | Hadoop, MapReduce, Pig, Hive/impala, YARN, Kafka, Flume, Oozie, Zookeeper, Spark, Airflow |
| **Cloud Platforms** | AWS (Amazon EC2, S3, RDS, IAM, Auto Scaling, CloudWatch, SNS, Athena, Glue, Kinesis, Lambda, EMR, Redshift, DynamoDB)  Azure (Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake, Azure HDInsight) |
| **BI Tools** | Tableau, PowerBI, SSIS |
| **CI/CD** | Jenkins, Ant, Maven, Gradle |
| **Programming Languages** | Scala, Hibernate, PL/SQL, R |
| **Scripting** | Python, Shell Scripting, JavaScript, jQuery, HTML, JSON, XML |
| **Web/Application Server** | Apache Tomcat, WebLogic, WebSphere |
| **Version Control** | Git, Subversion, Bitbucket, TFS |
| **Operating Systems** | Linux, Windows, Ubuntu, Unix |
| **Databases** | Oracle, SQL Server, Cassandra, Teradata, PostgreSQL, Snowflake, HBase, MongoDB , CosmosDB |

**Professional Experience:**

**Spencer Health Solutions - Aerial Center Pkwy, Morrisville Jan 2023 - Present**

**Senior Data Engineer**

**Responsibilities:**

* Developed complex data pipelines using Azure Databricks and Azure Data Factory (ADF) to create a consolidated and connected data lake environment.
* Wrote scripts in Hive SQL for creating complex tables with high performance metrics like partitioning, clustering, and skewing.
* Loaded the data through HBase into Spark RDD and implement in memory data computation to generate the output response. Continuously tuned Hive UDF for faster queries by employing partitioning and bucketing.
* Developed PySpark script to encrypting the raw data by using Hashing algorithms concepts on client specified columns.
* Implemented medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HD Insight/Databricks, and NoSQL DB).
* Experienced in configuring, tuning, and scaling NoSQL databases to handle large-scale data and high-throughput applications.
* Knowledgeable in integrating NoSQL databases with distributed data processing frameworks like Apache Spark, Kafka, and Hadoop.
* Configured Spark streaming to receive real time data from the Kafka and store the stream data to HDFS.
* Responsible for development and maintenance of data pipeline on Azure Analytics platform using Azure Databricks.
* Developed purging scripts and routines to purge data on Azure SQL Server and Azure Blob storage.
* Developed Azure Databricks notebooks to apply the business transformations and perform data cleansing operations.
* Implemented Apache Airflow for authoring, scheduling, and monitoring Data Pipelines.
* Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target snowflake database.
* Worked on migration of on-premises data (Oracle, SQL Server, DB2, MongoDB) to Azure Data Lake and Stored (ADLS) using Azure Data Factory (ADF V1/V2).
* Implemented Copy activity, Custom Azure Data Factory Pipeline activities.
* Implemented ad-hoc analysis solutions using Azure Data Lake Analytics/Store, HD Insights.
* Moved data to Azure Data Lake to Azure Data Warehouse using PolyBase. Created external tables in ADW with 4 compute nodes and scheduled.
* Worked with data ingestions from multiple sources into the Azure SQL data warehouse
* Transformed and loading data into Azure SQL Database. Maintained data storage in Azure Data Lake.
* Worked on Data Migration using SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Ingested data into Hadoop from various data sources like Oracle, MySQL using Sqoop tool.
* Created Sqoop job with incremental load to populate Hive External tables.
* Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines.
* Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
* Worked on a direct query using Power BI to compare legacy data with the current data and generated reports and stored dashboards.

**Environment:** Python, Hadoop, Spark, Spark SQL, Spark Streaming, PySpark, Hive, Scala, MapReduce, HDFS, Kafka, Sqoop, HBase, MS Azure, Blob Storage, Data Factory, Data Bricks, SQL Data Warehouse, Apache Airflow, Snowflake, Oracle, MySQL, UNIX Shell Script, Perl, PowerShell, SSIS, Power BI

**Homesite Insurance - Boston, MA Mar 2021 - Dec 2022**

**Senior Data Engineer**

**Responsibilities:**

* Developed Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into customer usage patterns.
* Implemented ETL and data movement solutions using Azure Data Factory, SSIS
* Developed dashboards and visualizations to help business users analyze data and provide insight to upper management with a focus on Microsoft products like SQL Server Reporting Services (SSRS) and Power BI.
* Migrated data from traditional database systems to Azure SQL databases.
* Implemented ad-hoc analysis solutions using Azure Data Lake Analytics/Store, HDInsight
* Designed and implemented streaming solutions using Kafka or Azure Stream Analytics
* Managed Azure Data Lakes (ADLS) and Data Lake Analytics and understanding how to integrate with other Azure Services. Used U-SQL for data transformation as part of a cloud data integration strategy.
* Worked with similar Microsoft on-prem data platforms, specifically SQL Server and related technologies such as SSIS, SSRS, and SSAS.
* Designed and implemented data pipelines using **Azure Data Factory (ADF)** to ingest, transform, and load data into **CosmosDB** for scalable data storage solutions.
* Integrated **JSON**-formatted data from multiple sources into **CosmosDB**, ensuring seamless data flow for business reporting needs.
* Developed ETL pipelines for handling **semi-structured data** in **MongoDB** and **CosmosDB** environments.
* Utilized **Python** and **PySpark** for data transformation, cleansing, and performance optimization in Azure-based data pipelines.
* Designed and implemented end-to-end data solutions (storage, integration, processing, and visualization) in Azure.
* Extracted, Transformed and Loaded data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, and U-SQL Azure Data Lake Analytics.
* Collaborated with data scientists, software engineers, and business analysts to understand business needs, design solutions, and deploy NLP models and pipelines
* Performed data Ingestion to one or more Azure Services (Azure Data Lake, Azure Storage, Azure SQL, and Azure DW) and processed the data in Azure Databricks.
* Integrated Kubernetes with the network, storage of security to provide a comprehensive infrastructure, and orchestrating the Kubernetes containers across multiple hosts
* Recreated existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database, and SQL Data warehouse environment. Experience in DWH/BI project implementation using Azure DF
* Involved in designing Logical and Physical Data Models for Staging, DWH, and Data Mart layer.
* Created POWER BI Visualizations and Dashboards as per the requirements
* Used various sources to pull data into Power BI, such as SQL Server, Excel, Oracle, SQL Azure, etc.
* Developed dashboards and visualizations to help business users analyze data and provide insight to upper management with a focus on Microsoft products like SQL Server Reporting Services (SSRS) and Power BI.

**Environment:** Azure, Hadoop, Scala, Hive, HDFS, Apache Spark, Kubernetes, Oozie, Sqoop, Cassandra, Shell Scripting, Power BI, Mongo DB, Jenkins, UNIX, JIRA, Git

**Lending Tree - Charlotte, North Carolina Feb 2019 - Mar 2021**

**Data Engineer**

**Responsibilities:**

* Optimized the PySpark jobs to run on Kubernetes Cluster for faster data processing
* Build ETL/ELT pipeline in data technologies like PySpark, Hive, Presto, and Databricks.
* Worked in data architecture best practices, integration, and data governance solutions (Data Catalog, Data Governance frameworks, Metadata and Data Quality)
* Implemented a 'server less' architecture using API Gateway, Lambda, and Dynamo DB and deployed AWS Lambda code from Amazon S3 buckets
* Worked datasets stored in AWS S3 buckets, used Spark data frames to perform pre-processing in Glue.
* Designed and developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Worked extensively with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).
* Developed a Data pipeline using Spark, Hive, Impala, and HBase to analyze customer behavioural data and financial histories in the Hadoop cluster.
* Developed and maintained data pipelines using Snowflake, leveraging the platform's built-in support for data integration and data warehousing.
* Skilled in designing flexible, schema-less, and high-performance data models tailored for NoSQL database use cases.
* Proficient in working with NoSQL databases like MongoDB, Cassandra, DynamoDB, and Couchbase for unstructured and semi-structured data.
* Contributed to the development of a business intelligence platform using Snowflake, used SQL to extract and transform data and created visualizations using Tableau
* Developed Python scripts to manage AWS resources from API calls using BOTO3 SDK and worked with AWS CLI.
* Set up the CI/CD pipelines using Maven, GitHub, and AWS.
* Proficient in leveraging NoSQL databases in event-driven architectures using message brokers like Apache Kafka and RabbitMQ.
* Knowledgeable in implementing eventual consistency, ACID compliance (where applicable), and transaction support in NoSQL environments.
* Expertise in working with multiple database types (e.g., combining NoSQL with SQL or graph databases) to meet diverse application needs.
* Worked extensively with importing metadata into Hive using Python and migrated existing tables and applications to work on AWS cloud (S3).
* Wrote real-time processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
* Created UNIX shell scripts for parameterizing the Sqoop and Hive jobs.
* Designed and implemented NLP data pipelines using tools such as Apache Kafka, Apache Spark, and AWS Glue.
* Created an end-to-end Machine learning pipeline in PySpark and Python.
* Implemented ETL solutions between an OLTP and OLAP database in support of Decision Support Systems with expertise in all phases of SDLC.

**Environment:** Hadoop, Python, PySpark, Spark, Hive, AWS, MapReduce, Sqoop, Spark-SQL, Kafka, Oozie, Impala, Nifi, Airflow, Snowflake, Oracle, Maven, Kubernetes, GitHub, Boto3, Unix Shell Scripting, Tableau, MySQL, SQL Server

**IBM - Hyd, India May 2016 - Oct 2018**

**Data Engineer**

**Responsibilities:**

* Loaded structured, unstructured, and semi-structured data into Hadoop by creating static and dynamic partitions.
* Implemented data ingestion and handling clusters in real time processing using Kafka.
* Imported real time weblogs using Kafka as a messaging system and ingested the data to Spark Streaming.
* Built scalable distributed Hadoop cluster running Hortonworks Data Platform.
* Serialized JSON data and storing the data into tables using Spark SQL.
* Worked on Spark framework on both batch and real-time data processing.
* Integrated Oozie with the rest of the Hadoop stack, supporting several types of Hadoop jobs Map Reduce, Hive, and Sqoop and system-specific jobs such as Java programs and Shell scripts.
* Developed Oozie Workflows for daily incremental loads, which get data from Teradata and then imported into hive tables. Involved in data ingestion of log files from various servers using NIFI.
* Designed and built solutions for data ingestion in both real-time& batch using Sqoop, Impala, Kafka, and Spark.
* Developed programs for Spark streaming which takes the data from Kafka and pushes into different sources.
* Developed bash scripts to bring log files from FTP server and then processing it to load into Hive tables.
* Developed MapReduce programs for applying business rules on the data.
* Developed a NiFi Workflow to pick up the data from Data Lake as well as from server and send that to Kafka broker.
* Loaded and transformed large sets of structured data from router location to EDW using an Apache NiFi data pipeline flow.
* Used Tableau to generate reports, graphs, and charts that offer an overview of the presented data.
* Created Tables, Stored Procedures, and extracted data using T-SQL for business users whenever required.

**Environment:** Hadoop HDP, MapReduce, HBase, HDFS, Hive, Pig, Nifi, Oozie, Shell Scripting, Sqoop, Apache Spark, Git, SQL, Tableau, Linux