**Name : Aditya**

**Email : adityamr9601@gmail.com**

**Number: 832-410-3718**

**Senior Data Engineer**

**PROFESSIONAL SUMMARY**

* IT professional with around 8years of experience, specialized in **Big Data ecosystem, Data Acquisition, Ingestion, Modeling, Storage Analysis, Integration, Data Processing**, and **Database Management**.
* Experience in designing interactive dashboards, reports, performing ad-hoc analysis and visualizations using **Tableau, Power BI, Arcadia,** and **Matplotlib**.
* Experience in application development, implementation, deployment, and maintenance using **Hadoop** and **Spark**-based technologies like **Cloudera, Hortonworks**, **Amazon EMR**, **Azure HDInsight**.
* A **Data Science enthusiast** with strong Problem solving, Debugging, and Analytical capabilities, who actively engage in understanding and delivering to business requirements.
* Ample work experience in **Big-Data ecosystem** - Hadoop (HDFS, MapReduce, Yarn), Spark, Kafka, Hive, Impala, HBase, Sqoop, Pig, Airflow, Oozie, Zookeeper, Ambari, Flume.
* Good knowledge of **Hadoop** cluster architectureand its key concepts - Distributed file systems, Parallel processing, High availability, fault tolerance, and Scalability.
* Complete knowledge of **Hadoop architecture** and **Daemons of Hadoop clusters,** which include Name node, Data node, Resource manager, Node Manager, and Job history server**.**
* Expertise in developing Spark applications for interactive analysis, batch processing and stream processing, using programming languages like PySpark, Scala.
* Advanced knowledge in Hadoop based Data Warehouse (**HIVE**) and database connectivity (**SQOOP**).
* Ample experience using **Sqoop** to ingest data from **RDBMS** - Oracle, MS SQL Server, Teradata, PostgreSQL, and MySQL.
* Experience in working with various streaming **ingest services** with Batch and Real-time processing using **Spark Streaming, Kafka, Confluent, Storm, Flume, and Sqoop.**
* Proficient in using **Spark API** for streaming real-time data, staging, cleaning, applying transformations, and preparing data for machine learning needs.
* Experience in developing end-to-end ETL pipelines using **Snowflake**, **Alteryx,** and **Apache NiFi** for both relational and non-relational databases (**SQL** and **NoSQL**).
* Strong working experience on **NoSQL** databases and their integration with the Hadoop cluster- **HBase**, **Cassandra**, **MongoDB, DynamoDB**, and **Cosmos DB**.
* Experience with **AWS** cloud servicesto develop cloud-based pipelines and Spark applications using **EMR**, **LAMBDA** and **Redshift**.
* Extensive knowledge in working with **Amazon EC2** to provide a solution for computing, query processing, and storage across a wide range of applications.
* Expertise in using **AWS S3** to stage data and to support data transfer and data archival. Experience in using **AWS** **Redshit** for large scale data migrations using **AWS DMS** and implementing **CDC** (change data capture).
* Strong experience in developing **LAMBDA** functions using **Python** to automate data ingestion and tasks.
* Working knowledge of **Azure** cloud components (HDInsight, Databricks, Data Lake, Blob Storage, Data Factory, Storage Explorer, SQL DB, SQL DWH, Cosmos DB).
* Experienced 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 database access.
* Extensive experience with **Azure** services like **HDInsight**, **Stream Analytics**, **Active Directory**, **Blob Storage**, **Cosmos DB**, and **Storage Explorer**.
* Good knowledge in understanding the security requirements and implementation using **Azure Active Directory, Sentry, Ranger**, and **Kerberos** for authentication and authorizing resources.
* Experience in all phases of **Data Warehouse** development like requirements gathering, design, development, implementation, testing, and documentation.
* Extensive knowledge of Dimensional **Data Modeling** with Star Schema and Snowflake for FACT and Dimensions Tables using Analysis Services.
* Sound knowledge in developing highly scalable and resilient **Restful** **API**s, **ETL** solutions, and third-party platform integrations as part of Enterprise Site platform.
* Experience in implementing pipelines using **ELK** **(Elasticsearch, logstash, kibana)** and developing stream processes using **Apache Kafka**.
* Sound knowledge and experience in programming languages like **Python**, **Scala**.
* Experience in using various IDEs like **Eclipse**, **IntelliJ**, and repositories **SVN** and **Git** version control systems.
* A team player with strong communication, interpersonal, problem-solving, and debugging skills. Ability to quickly adapt to new environments and technologies.
* Successfully working in a fast-paced environment, both independently and in a collaborative way. Expertise in complex troubleshooting, root-cause analysis, and solution development.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, Yarn, MapReduce, Spark, Kafka, Kafka Connect, Hive, Airflow, StreamSets, Sqoop, HBase, Flume, Pig, Ambari, Oozie, Zookeeper, Nifi, Sentry |
| **Hadoop Distributions** | Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP |
| **Cloud Environment** | Amazon Web Services (AWS), Microsoft Azure |
| **Databases** | MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2 |
| **NoSQL Database** | DynamoDB, HBase |
| **AWS** | EC2, EMR, S3, Redshift, EMR, Lambda, Kinesis Glue, Data Pipeline |
| **Microsoft Azure** | Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Azure Active Directory |
| **Operating systems** | Linux, Unix, Windows 10, Windows 8, Windows 7, Windows Server 2008/2003, Mac OS |
| **Software’s/Tools** | Microsoft Excel, Statgraphics, Eclipse, Shell Scripting, ArcGIS, Linux, Jupyter Notebook, PyCharm, Vi / Vim, Sublime Text, Visual Studio, Postman |
| **Reporting Tools/ETL Tools** | Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, Power BI, Arcadia, Data stage, Pentaho |
| **Programming Languages** | Python (Pandas, Scipy, NumPy, Scikit-Learn, Stats Models, Matplotlib, Plotly, Seaborn, Keras, TensorFlow, PyTorch), PySpark, T-SQL/SQL, PL/SQL, HiveQL, Scala, UNIX Shell Scripting |
| **Version Control** | Git, SVN, Bitbucket |
| **Development Tools** | Eclipse, NetBeans, IntelliJ, Hue, Microsoft Office |

**PROFESSIONAL EXPERIENCE**

**GE HealthCare (Wisconsin) Jan 2023 – Till Date**

**Senior Data Engineer**

**Responsibilities:**

* Worked on **Apache Spark** data processing project to process data from **RDBMS** and several data streaming sources and developed **Spark** applications using **Python** on **AWS EMR.**
* Performed reporting analytics on data from AWS stack by connecting it to BI tools (**Tableau**, **Power Bi**).
* Migrated an entire oracle database to BigQuery and build Data pipelines in airflow in **GCP** for ETL related jobs using different airflow operators.
* Designed and deployed multi-tier applications leveraging **AWS** services like (EC2, Route 53, S3, RDS, DynamoDB) focusing on high availability, fault tolerance, and **auto-scaling** in AWS **Cloud Formation**.
* Utilizing G-cloud function with python to load the data into **BigQuery** for the on-arrival files in **GCS** Bucket. Using Apache Drill beam to build data flow pipelines using which we converted the CSV files, and JSON files to NDJSON files.
* Involved in **Data Mapping** Specifications to create and execute detailed system test plans. The data mapping specifies what data will be extracted from the data warehouse
* Configured and launched **AWS EC2** instances to execute Spark jobs on **AWS Elastic Map Reduce (EMR).**
* Performed data transformations using Spark **Data Frames**, **Spark SQL**, **Spark File formats**, **Spark RDDs**.
* Transformed data from different files (Text, CSV, JSON) using **Python** scripts in **Spark**.
* Loaded data from various sources like **RDBMS** (MySQL, Teradata) using **Sqoop** jobs**.**
* Handled **JSON** datasets by writing custom **Python** functions to parse through JSON data using **Spark.**
* Developed a preprocessing job using **Spark Data Frames** to flatten **JSON** documents to **flat files**.
* Utilized REST API’s with python to ingest the data into big query. Computed **PySpark** Jobs using gsutil and got that executed In Data proc Cluster.
* Improved performance of cluster by optimizing existing algorithms using Spark.
* Performed wide, narrow transformations, actions like filter, Lookup, Join, count, etc. on **Spark Data** **Frames**.
* Worked with Parquet files and Impala using **PySpark,** and Spark Streaming with RDDs and **Data Frames**.
* Aggregated logs data from various servers and made them available in downstream systems for analytics by using **Apache Drill**
* Developed **batch** and **streaming** processing apps using **Spark API**s for functional pipeline requirements.
* Automated data storage from streaming sources to AWS data lakes like **S3**, **Redshift** and **RDS** by configuring **AWS Kinesis (Data Firehose)**.
* Performed analytics using real time integration capabilities of **AWS Kinesis (Data Streams)** on streamed data
* Cleaned and handled missing values in data using **Python** by backward-forward filling methods and applied Feature engineering, normalize and label encoding techniques using **Python** **Scikit-learn** preprocessing.
* Stored data into various tiers of **AWS** **S3** based on business requirements and frequency of data access.
* Imported data from **AWS S3** into **Spark** **RDD** performed transformations and actions on RDD's.
* Worked with database administrating team on **SQL optimization** for databases like **Oracle**, **MySQL**, **MS SQL**.
* Assisted in configuring and implemented **MongoDB** cluster nodes on **AWS EC2** instances.
* Identified executor failures, data skewness, and runtime issues by monitoring **Spark** apps through **Spark UI**.
* Ensured database performance in production by stress testing **AWS EC2** of **DynamoDB** environments.
* Automated deployments and routine tasks using **UNIX Shell Scripting**.
* Collaborated with the Data Science team building machine learning models on **Spark EMR** cluster to deliver the data needs under business requirements.
* Worked in an agile environment to implement projects and enhancements with weekly SCRUMs.

**Environment:** Spark v2.0.2, Hive, Power BI, Tableau, , AWS (EC2, S3, EMR, RDS, Lambda, Kinesis, Redshift, Cloud Formation), Sqoop, Kafka, Spark streaming, ETL, Python (Pandas, NumPy), PySpark, GIT (version control), MySQL, MongoDB

**Spencer Stuart (Chicago)**   **Sep 2020 – Dec 2022**

**Senior Data Engineer**

**Responsibilities:**

* Designed and deployed data pipelines using Azure cloud platform **(HDInsight, DataLake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH, and Data Storage Explorer)**.
* Involved in Data mapping specifications to create and execute detailed system test plans. The data mapping specifies what data will be extracted from an internal data warehouse, transformed, and sent to an external entity
* Responsible for analyzing various data sources such as flat files, ASCII Data, EBCDIC Data, Relational Data (Oracle, DB2 UDB, MS SQL Server) from various heterogeneous data sources.
* Developed custom-built **ETL** solution, batch processing, and real-time data ingestion pipeline to move data in and out of the Hadoop cluster using **PySpark** and **Shell Scripting**.
* Integrated on-premises data **(MySQL, Hbase)** with cloud **(Blob Storage, Azure SQL DB)** and applied transformations to load back to **Azure Synapse** using **Azure Data Factory**.
* Built and published Docker container images using **Azure Container Registry** and deployed them into **Azure Kubernetes Service (AKS)**.
* Imported metadata into **Hive** and migrated existing tables and applications to work on **Hive** and **Azure.**
* Created complex data transformations and manipulations using **ADF** and **Scala**.
* Configured **Azure Data Factory (ADF)** to ingest data from different sources like relational and non-relational databases to meet business functional requirements.
* Optimized workflows by building DAGs in **Apache Airflow** to schedule the ETL jobs and implemented additional components in Apache **Airflow** like Pool, Executors, and multi-node functionality.
* Improved performance of **Airflow** by exploring and implementing the most suitable configurations.
* Configured **Spark streaming** to receive real-time data from **Apache Flume** and store the stream data using Scala to Azure Table and **DataLake** is used to store and do all types of processing and analytics. Created data frames using **Spark Dataframes**.
* Designed cloud architecture and implementation plans for hosting complex app workloads on **MS Azure**.
* Performed operations on the transformation layer using **Apache Drill, Spark RDD**, **Data frame APIs**, and **Spark SQL** and applied various aggregations provided by Spark framework.
* Provided real-time insights and reports by mining data using **Spark Scala functions**. Optimized existing Scala code and improved the cluster performance.
* Processed huge datasets by leveraging **Spark Context**, **SparkSQL,** and **Spark Streaming**.
* Enhanced reliability of Spark cluster by continuous monitoring using **Log Analytics** and **Ambari** WEB UI.
* Improved the query performance by transitioning log storage from **Cassandra** to **Azure SQL** Datawarehouse.
* Implemented custom-built input adapters using **Spark, Hive, and Sqoop** to ingest data for analytics from various sources **(Snowflake, MS SQL, MongoDB)** into **HDFS**. Imported data from web servers and Teradata using **Sqoop, Flume, and Spark Streaming API**.
* Improved efficiency of large datasets processing using **Scala** for concurrency support and parallel processing.
* Developed map-reduce jobs using **Scala** for compiling program code into bytecode for the **JVM** for data processing. Ensured faster data processing by developing Spark jobs using **Scala** in a test environment and used Spark SQL for querying.
* Improved processing time and efficiency by using Spark applications like batch interval time, level of parallelism, memory tuning. Monitored workflows for daily incremental loads from **RDBMS**s **(MongoDB, MS SQL, MySQL)**.
* Implemented indexing to data ingestion using **Flume** sink to write directly to indexers deployed on a cluster.
* Delivered data for analytics and Business intelligence needs by managing workloads using **Azure Synapse**.
* Improved security by using **Azure DevOps** and **VSTS (Visual Studio Team Services)** for **CI/CD**, Active Directory, and Apache Ranger for authentication. Managed resources and scheduling across the cluster using **Azure Kubernetes Service.**

**Environment:** Hadoop, Spark, Hive, Sqoop, HBase, Flume, Ambari, Scala, MS SQL, MySQL, Snowflake, MongoDB, Git, Data Storage Explorer, Python, Azure (Data Storage Explorer, ADF, AKS, Blob Storage)

**Charter Communication (Colorado) Aug 2018 – Aug 2020**

**Data Engineer**

**Responsibilities:**

* Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Drive insight from the data perform impact analysis, suggest, and implement solutions to maintain the quality of the data ensuring timely generation and retrieval of quality client deliverables.
* Onboard new clients into existing studies and assist in launching new studies for the client.
* Revamp existing transaction data model to meet growing needs of the clients and the organization that continues to help Argus establish new revenue generating engagements.
* Extensively worked on Performance Tuning of complex scripts and redesigned the tables to avoid bottlenecks in the system.
* Evaluate correlations among statistical data, identify trends and summarize findings across issuers.
* Responsibilities include gathering business requirements, developing strategy for data cleansing and data migration, writing functional and technical specifications, creating source to target mapping, designing data profiling and data validation jobs in Informatica, and creating ETL jobs in Informatica.
* Ran HDD Explore and HDD Iatric Inbound batches from OC and verified the data is published from stage to hold tables using the workflows ran in the Informatica workflow monitor.
* Developed SQL scripts to validate the data loaded into warehouse and Data Mart tables using Informatica.
* Extensively used Informatica power center for extraction, transformation and loading process.
* Built APIs that will allow customer service representatives to access the data and answer queries.
* Designed changes to transform current Hadoop jobs to HBase.
* Handled fixing of defects efficiently and worked with the QA and BA team for clarifications.
* Responsible for Cluster maintenance, Monitoring, commissioning and decommissioning Data nodes, Troubleshooting, Manage and review data backups, Manage & review log files.
* Extending the functionality of Hive with custom UDF s and UDAF's.
* Implemented Bucketing and Partitioning using hive to assist the users with data analysis.
* Used Oozie scripts for deployment of the application and perforce as the secure versioning software.
* Develop database management systems for easy access, storage, and retrieval of data.
* Responsible for loading the data from BDW Oracle database, Teradata into HDFS using Sqoop.
* Wrote data ingestion systems to pull data from traditional RDBMS platforms such as Oracle and Teradata and store it in NoSQL databases such as MongoDB. Involved in loading and transforming large sets of Structured, Semi-Structured and Unstructured data and analyzed them by running Hive queries. Processed the image data through the Hadoop distributed system by using Map and Reduce then stored into HDFS.
* Performed Data Visualization and Designed Dashboards with Tableau and generated complex reports including chars, summaries, and graphs to interpret the findings to the team and stakeholders.
* Used Git for version control with Data Engineer team and Data Scientists colleagues. Involved in creating Created Tableau dashboards using stack bars, bar graphs, scattered plots, geographical maps, Gantt charts etc. using show me functionality. Dashboards and stories as needed using Tableau Desktop and Tableau Serve
* Performed statistical analysis using SQL, Python, R Programming and Excel.
* Worked extensively with Excel VBA Macros, Microsoft Access Forms
* Used Python& SAS to extract, transform & load source data from transaction systems, generated reports, insights, and key conclusions.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Build an ETL which utilizes spark jar inside which executes the business analytical model.
* Developed story telling dashboards in Tableau Desktop and published them on to Tableau Server which allowed end users to understand the data on the fly with the usage of quick filters for on demand needed information.
* Effectively Communicated plans, project status, project risks and project metrics to the project team planned test strategies in accordance with project scope.

**Environment:** Cloudera CDH4.3, Hadoop, Pig, Hive, Informatica, HBase, Map Reduce, HDFS, Python.

**LOWE’S (NC)**   **Aug 2016 – July 2018**

**Data Engineer**

**Responsibilities:**

* Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data.
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Design and development of ETL processes using Informatica ETL tool for dimension and fact file creation 
* Performed wide, narrow transformations, actions like filter, Lookup, Join, count, etc. on **Spark Data** **Frames**.
* Worked with Parquet files and Impala using **PySpark,** and Spark Streaming with RDDs and **Data Frames**.
* Involved in Uploading Master and Transactional data from flat files and preparation of Test cases, Sub System Testing
* Aggregated logs data from various servers and made them available in downstream systems for analytics by using **Apache Kafka**.
* Improved **Kafka** performance and implemented security.
* Developed **batch** and **streaming** processing apps using **Spark API**s for functional pipeline requirements.
* Worked with Spark to create structured data from the pool of unstructured data received.
* Implemented intermediate functionalities like events or records count from the flume sinks or Kafka topics by writing Spark programs in java and python.
* Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS.
* Experienced in transferring Streaming data, data from different data sources into HDFS, No SQL databases
* Created ETL Mapping with Talend Integration Suite to pull data from Source, apply transformations, and load data into target database.
* Transformed data from different files (Text, CSV, JSON) using **Python** scripts in **Spark**.
* Loaded data from various sources like **RDBMS** (MySQL, Teradata) using **Sqoop** jobs.
* Well versed with the Database and Data Warehouse concepts like OLTP, OLAP, Star Schema
* AWS provides a secure global infrastructure, plus a range of features that use to secure the data in the cloud
* Worked and learned a great deal from Amazon Webservices (AWS) Cloud services like EC2, S3, EBS, RDS and VPC.
* Developed multiple Kafka Producers and Consumers from scratch to as per the software requirement specifications.
* Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of Spark using python.
* Worked with Apache Drill which provides fast and general engine for large data processing integrated with functional programming language python.

**Environment:** Apache Hadoop, HDFS, MapReduce, Sqoop, Flume, Pig, Hive, HBASE, Oozie, Scala, Spark, Kafka, Linux.

**GGK Tech (India) June 2015-June 2016**

**Data Engineer**

**Responsibilities:**

* Participated in the analysis, design, and development phase of the **Software Development Lifecycle (SDLC)**.
* Developed test-driven web applications using **Java J2EE**, **Struts 2.0 framework**, **Spring MVC**, Hibernate framework, **JavaScript,** and **SQL Server** database with deployments on **IBM WebSphere**.
* Designed & developed a web Portal using Struts Framework, J2EE. Developed newsletter as part of process improvement tasks using HTML and CSS to report the weekly activities.
* Developed front-end, User Interface using **HTML**, **CSS**, **JSP**, **Struts, Angular,** and **NodeJS,** and session validation using **Spring** **AOP**.
* Extensively used Java multi-threading to implement batch Jobs with **JDK** 1.5 features and deployed it on the **JBoss** server.
* Ensured High availability and load balancing by configuring and Implementing clustering of Oracle on **WebLogic Server 10.3**.
* Improved productivity by developing an automated system health check tool using **UNIX** shell scripts.

**Environment**: Java/J2EE, Spring, Oracle, Linux, JDBC, Git, HTML, CSS, Angular, NodeJS, Postman, Servlets, Struts, JSP, WebLogic, PL/SQL, Eclipse.