**Name: Abhilash Reddy N**

**Email ID:** **abhilashreddy1305@gmail.com**

**Contact information/Phone: +1 587-487-4187**

**Data Engineer / ML OPS / ETL Engineer / Data Analyst**

**About me**

Data Engineer with 6 years of experience in data warehousing, data engineering, data visualization, feature engineering, big data, ETL/ELT, and business intelligence. As a data engineer, specialize in Azure, AWS frameworks, Cloudera, Hadoop Ecosystem, Python Spark/PySpark/Scala, ML Ops, Data bricks, Redshift, Snowflake, relational databases, Visualizations with Tableau, PowerBI, Data DevOps Frameworks/Pipelines with strong Programming/Scripting Skills

**Professional Summary**

**Azure**

Hands on experience in setting up the **Azure Data factory** and creating the ingestion Pipelines to pull data to **Azure Data Lake Store** and **Azure Blob Storage**. Migration of on-premises databases to Microsoft Azure environment (**Blobs, Azure Data Warehouse, Azure SQL Server, Azure ML studio, PowerShell Azure components, SSIS**). Skilled in Azure, Azure Data Factory (ADF), **Databricks**, ETL, SQL Warehousing, **Power BI**, Azure Data Lakes (ADLS), Data Lake Analytics and integrated with other Azure Services.

**AWS**

* Expertise in **Amazon Web Services (AWS)** Cloud Platform which includes services like **VPC, DynamoDB, Route 53, Elastic Container Services (ECS),** Security Groups, CloudWatch, EC2, S3, Security Groups, Kinesis, Red shift, IAM, CloudFormation, ELB, Cloud Front, Elastic Beanstalk (EBS).
* Experience in using Python included **Boto3** to supplement automation provided by **Ansible** and **Terraform** for tasks such as encrypting Elastic Beanstalk volumes and scheduling Lambda functions for routine AWS tasks.
* Hands on experience in using other Amazon Web Services like **S3, VPC, EC2, Autoscaling, RedShift, DynamoDB, Route53, RDS, Glacier, EMR**.
* Implemented monitoring and established best practices around using **Elasticsearch** and used **AWS Lambda** to run code without managing servers.
* Experience in deploying the AWS infrastructure with **AWS CDK** using **Typescript**.
* Experienced in **predictive analytics** and **Machine learning** models.
* **Implemented a ‘server less’ architecture using API Gateway, Lambda, and DynamoDB and deployed AWS**
* **Good experience** with **Informatica**, **AWS Glue** for designing **ETL** Jobs for Processing of data.

**Data Visualization**

* Proficiency in data visualization tools (Tableau, PowerBI).
* Ability to implement automated processes for data visualization.
* Well Versed with principles of Data warehousing, Fact Tables, Dimension Tables, Star Snowflake schema modeling.

**Python**

* REST API Framework using Django, Experience in implementing Model View Control (MVC) architecture using server-side applications like Django, Flask and Pyramid for developing web applications.
* Extended Testing framework by adding some helper classes and methods.
* Implementation of Python best Practices (PEP-8).

**Big Data & Hadoop**

* Extensive experience in implementing the best practices and design patterns for the Data Lake, enterprise data warehouse, and domain-specific data marts.
* Expertise in Big Data architecture like Hadoop (MS Azure, Hortonworks, and Cloudera) distributed systems.

**Education:**

**Masters –** Concordia University of Edmonton (2019-2021) **Major –** MISSM

**Bachelors –** Holy Mary Institute of Technology and Science (2013-2017) **Major-** Computer Science

**Technical Skills:**

|  |  |
| --- | --- |
| **Operating Systems** | Linux, UNIX, Windows. |
| **Development Methodologies** | Agile/ Scrum, Waterfall. |
| **IDEs** | Eclipse, Net Beans, IntelliJ |
| **Big Data Platforms** | Hortonworks, Cloudera CDH4, CDH5 |
| **Programming Languages** | Python, Scala, SQL,Java, .Net, XML, JSON |
| **Python** | NumPy, Scipy, Pandas, nltk, Matplotlib, BeautifulSoup, TextBlob |
| **Hadoop Components** | HDFS, Sqoop, Hive, Pig, MapReduce, YARN, Impala, Hue, Zookeeper |
| **Spark Modules** | Spark Core Component, Spark SQL, Spark Streaming, Spark MLlib, Spark GraphX, SparkR |
| **RDBMS/NOSQL Databases** | Oracle, MYSQL, Microsoft SQL Server, Hbase, MongoDB, Cassandra |
| **Cloud Technologies** | Amazon Web Services (AWS), Azure |
| **Azure services** | ADF, Databriks, Azure data lake, Azure SQL data warehouse, Logic apps, key vault |
| **AWS Services** | EC2, IAM, S3, Autoscaling, CloudWatch, Route53, EMR, RedShift, Glue |
| **DevOps Tools** | Ant, Maven, Jenkins, Git, GitHub |
| **Visualization and Reporting** | Power BI, Tableau |
| **Big Data** | HDFS, Hue, MapReduce, PIG, Hive, HCatalog, Hbase, Sqoop, Impala, Zookeeper, Flume, Yarn, Cloudera Manager, Kerberos, pysparkAirflow, Kafka Snowflake. |

**Professional Experience**

**Client: LTI - Larsen & Toubro Infotech- Toronto, Canada**  **Jan 2023 – Present**

**Role: Data Engineer (AWS,Python)**

**Project Description:** This project majorly focuses on expanding and optimizing data and data pipeline architecture, as well as building and maintaining the data workflow, designing optimal data pipeline and infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources (SAP, Oralce ERP, Salesforce).

 Development of tasks of designing and managing data pipelines, by making use of ETL, Hadoop ecosystem, designing, developing, maintaining, deploying data and ETL pipeline solutions using Apache Airflow, Kubernetes, Python (RESTful server-side APIs) stacks, and cloud services.

**Key Contributions:**

* AWS Cloud Formation templates were designed to create VPCs, subnets, and NAT to ensure the successful deployment of Web applications and database templates.
* Creating S3 buckets and managing S3 bucket policies, as well as using S3 buckets and Glacier for storage and backup on AWS.
* Migrated Hive and MapReduce jobs from on-premises MapR to AWS cloud using EMR and Quble

**Responsibilities:**

* Worked on Amazon AWS tools, such as Lambda, S3, SNS/SQS, and their python interfaces.
* Architected the AWS pre-prod environment, which includes setting up the **VPC, Subnets, S3, EC2, AWS EMR, Redshift, Web Servers, IAM, Security Groups, Loan Balancer & Lambda to support data warehousing solutions.**
* Built and implemented performant data pipelines using **Apache** Spark on **AWS EMR**. Performed maintenance of data integration programs into Hadoop RDBMS environments from both structured semi- structured data source systems.
* Involved in S3 event notifications, an SNS topic, SQS queue, Lambda function sending a message to Slack channel.
* Transformed Teradata scripts and stored procedures to SQL and **Python running on Snowflake's** cloud.
* Worked on implementing Data warehouse solutions in AWS Redshift, worked on various projects to migrate data from one database to **AWS Redshift, RDS, ELB, EMR, Dynamo DB** and **S3.**
* Worked with Informatica Cloud **Data Integration** to deliver accessible, trusted, and secured data to facilitate more valuable business decisions to identify competitive advantages to better service customers, and build an empowered workforce.
* Implemented a Python/Java based distributed random forest via PySpark and **MLlib**.
* Built and deployed **ML models** using **AWS Sagemaker** to predict customer purchase behavior and improving marketing strategies.
* Created file objects from the web service data (JSON/XML) using HDFS.
* Monitored and managed AWS services using **AWS CloudWatch**.
* Developed Spark Python models for **machine learning** and **predictive analytics** in Hadoop.
* Developed algorithms to support complex data visualizations.
* Developed AWS infrastructure with **AWS CDK** using **constructs** and **typescript**.
* Written scripts from scratch to create AWS infrastructure using languages such as BASH and Python, created Lambda functions to upload code and to check changes in S3, DynamoDB table.

**Environment: AWS**, Hadoop, Python, Pyspark, SQL, Snowflake, Data bricks/Delta Lake, S3, Athena EMR, ML Ops, CDK, HDFS, Json, XML.

**Client: Manulife - Ontario June 2021 – Jan 2023**

**Role: Azure Data Engineer**

Main aim is to implement a single analytics environment for delivering world-class business intelligence capability for all health group business segments. This project will stage core healthcare data for primary services offered by business segments such as Medical Claim, Pharmacy Claim, Membership, Customer, Claim Financial, and Provider. Design and implementation of data pipelines consisting of launching several Spark clusters equipped with Azure services that read the datasets from various data sources and perform transformations, analytics and finally store results to application.

**Key Contributions**

* Implement ad-hoc analysis solutions using Azure Data Lake Analytics/ Store, HDInsight/ Databricks.
* Improved data processing speed by developing and maintaining Apache Spark workflows, and Spark Streaming data processing.
* Used Control-M for scheduling DataStage jobs and used Logic Apps for scheduling ADF pipelines.

**Responsibilities:**

* Worked on Azure Cloud Services (PaaS & IaaS), **Azure Synapse Analytics, Azure Data Lake, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault,** and **Azure Data Lake**.
* Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
* Extract, transform, and load data from source systems to Azure Storage services using a combination of Azure Data Factory, **T-SQL, Spark SQL**, and **U-SQL** of Azure Data Lake Analytics.
* Developed **HIVE UDFs** to incorporate external business logic into Hive script and developed join data set scripts using HIVE join operations.
* Implement one time Data migration of Multistate level data from SQL server to **Snowfalke** by using Python and **SnowSQL**.
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data. Worked with NoSQL databases like HBase in creating HBase tables to load large sets of semi-structured data coming from various sources.
* Extracted and updated the data into HDFS using Sqoop import and export.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL DB, Azure SQL DW), and processing of the data in Azure Databricks. Implement ad-hoc analysis solutions using Azure Data Lake Analytics/ Store, HDInsight
* Analyze and define updates to **ML pipelines**, deployment, and monitoring scripts to accommodate any changes in data models.
* Bulit **machine learning models** identify user behavior using real-time data analysis using **Azure ML studio**.
* Worked with Spark for optimization of the existing algorithms in Hadoop using Java, Spark Context, Spark-SQL, **PySpark**, Pair RDD's, and **Spark YARN.**
* Completed online data transfer from AWS S3 to Azure Blob by using Azure Data Factory (ADF).
* Used Azure Migrate to get started migrating AWS EC2 instances over to MS Azure.
* Designed rich data visualizations to model data into human-readable form with **Tableau** and **Matplotlib**.
* Developed custom Kafka producer and consumer for different publishing and subscribing to Kafka topics.

**Environment**: Python, Hadoop (HDFS, MapReduce), SnowSQL, Snowflake, Spark Context, Spark-SQL, PySpark, Pair RDD's, Spark Data Frames, Spark YARN, Hive, Pig, HBase, Oozie, Hue, Sqoop, Flume, NIFI, Azure cloud, ML Ops,Tableau, Matplotlib.

**Client: EXL - India**  **Aug 2016 – July 2019**

**Data Modeller/Engineer**

**Project Description:** This Project implements Data Analytics, Application Development, Robotic Process Automation, AI, DevOps, and Test Automation Services. Project's goal is to establish a database for a new program and formulate schemas for the data warehouse, encompassing comprehensive technology and operational support for individual consumers, small businesses, middle-market businesses, and large corporations.

**Responsibilities:** Developed business logic using Kafka Direct Stream in Spark Streaming and implemented business transformations. Worked on Cloudera distribution and deployed on **AWS EC2** Instances.

* Worked on connecting Cassandra database to Amazon EMR File System for storing the database in S3.
* Implemented usage of Amazon EMR for processing Big Data across a Hadoop Cluster of virtual servers on **Amazon Elastic Compute Cloud (EC2)** and **Amazon Simple Storage Service (S3)**. Deployed the project on **Amazon EMR**
* Evaluated Hadoop and its ecosystem's suitability for aforementioned project, validated with proof of concept (POC) applications in order to ultimately adopt them to benefit from Big Data Hadoop initiative.
* Worked on complex SQL Queries, PL/SQL procedures and convert them to ETL tasks.
* Worked with **PowerShell** and **UNIX scripts** for file transfer, emailing and other file related tasks.
* Created a risk-based machine learning model (logistic regress, random forest, SVM, etc.) to predict which customers are more likely to be delinquent based on historical performance data and rank order them.

**Environment:** Hadoop, Map Reduce, Hive, Apache Spark, Oracle, GitHub, Tableau, UNIX, Cloudera, Kafka, Sqoop, Scala, NIFI, HBase, Amazon EC2, S3, Kubernetes, Oracle, SQL Server, MySQL, HBase, MongoDB, RedShift, DynamoDB