**Nikhila**

**Data Engineer**

**SUMMARY:**

* **8+** years of experience in Analysis, Design, Development and Implementation as a **Data Engineer**.
* Worked in multiple **BigData** distributions like **HortonWorks, Cloudera** and **MapR**.
* Experienced in **MRV1** and **Hadoop 2.0 (MRV2) YARN** architecture.
* Hands on experience with **Big Data Hadoop** Ecosystem tools like **Map Reduce, YARN, HDFS, HBase, Impala, Tez, Hive, Airflow and Apache Spark for ingestion, storage, querying, processing, and analysis of data**.
* Experience working on **Integrated Developer Environment Tools** like **PyCharm** and **Visual Studio**.
* Vast knowledge in Database Management Systems and associated tools **Oracle, Cassandra, Erwin Data Modeler, Business Intelligence** and **Oracle BI-Publisher**.

# Hands on experience in Data Modelling, Data Mapping, Data Profiling, Data Analytics, Data Mining, Data Integration, Data Cleansing, Data Loading, Data Transformation, Data Migration and Data Validation.

* Experience on Migrating **SQL database** to **Azure data Lake, Azure data lake Analytics, Azure SQL**

**Database, Data Bricks** and **Azure SQL Data warehouse** and controlling and granting database access and Migrating **On premise databases** to **Azure Data Lake** store using **Azure Data factory**.

* Design and implement database solutions in **Azure SQL Data Warehouse, Azure SQL.**
* Strong experience on **Azure Data Factory (ADF), Integration Run Time (IR), File System Data**

# Ingestion, Relational Data Ingestion.

* Created **Azure SQL database**, performed monitoring and restoring of **Azure SQL database**. Performed migration of Microsoft SQL server to **Azure SQL database**.
* Profound knowledge and experience working on **AWS services** using **S3, EC2, EMR, Glue, Athena, Aurora, SNS, Crawler**.
* Proficient in using RDBMS concepts and worked with **MySQL, Oracle, Aurora** databases.
* Strong experience working on **Bigdata** using **Spark, Hive, HDFS, Hadoop** and **Sqoop**.
* Experience in software configuration management using **GIT**.
* Excellent knowledge on using **Jenkins** for continuous integration and continuous deployment.
* Proficient in extracting, transforming, and loading of data using SQL Server along with writing complex

**SQL Queries** and Stored procedures.

* Experience using **Jira** for ticketing issues and tracking the progress of the issues.
* Proficient in SDLC methodologies such as **agile**, **scrum** and **waterfall** models.

**PROFESSIONAL CERTIFICATIONS:**

* **AWS Certified Cloud Practitioner** from **Amazon Web Services Training and Certification**.
* **Databricks Certified Data Engineer Associate (Version 3)** from **Databricks. TECHNICAL SKILLS:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Cloud Technologies | Azure, AWS |  |
| Azure Skills | Azure Data Lake, Data Factory, Databricks, Azure SQL Databases, Azure SQL Datawarehouse |
| AWS Skills | S3, EC2, EMR, Glue, Athena, Aurora, SNS, Crawler |
| Big data Skills | Spark, Hive, HDFS, Hadoop, Sqoop |
| Programming Languages | Python, Spark |
| Databases | Aurora DB, Oracle, MySQL |
| Job Scheduler Tools | Airflow |
| CICD | Bit Bucket, Jenkins |
| IDE | PyCharm, Visual Studio |
| **PROFESSIONAL EXPERIENCE:**  ***Role: Sr Data Engineer September***  ***2022 – Present***  ***Client: XTO Energy, Spring, TX***  **Responsibilities:**   * Created Pipelines in **Azure Data Factory** using Linked Services/Datasets/Pipeline/ to extract, transform   and load data from different sources like **Azure SQL, Azure SQL Data warehouse**.   * Implemented **Lambda architectures** using Azure Data platform capabilities like **Azure Data Lake, Azure Data Factory, Azure Data Catalog, Azure SQL Server**. * Designed end to end scalable architecture to solve business problems using various **Azure Components** like **Data Factory, Data Lake, Storage**. * Prepare Conceptual, Logical & Physical Data models covering all the entities using **Erwin Data Modeler** 9.64. * Expertise in **Erwin Data Modeler** tool for ER, Dimensional Data Modelling. * Developed **JSON Scripts** for deploying the Pipeline in **Azure Data Factory (ADF)** that process the data   using the **SQL Activity**.   * Developed Spark applications using Scala and **Spark-SQL** for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns. * Used **Azure Event Grid** for managing event service that enables you to easily manage events across   many different **Azure services and applications**.   * Used **Azure Databricks** for fast, easy and collaborative spark-based platform on **Azure**. * Used **Databricks** to integrate easily with the whole Microsoft stack. * Data Ingestion to one or more **Azure Services** - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW)   and processing the data in **Azure Databricks**.   * Used **Azure Data Catalog** which helps in organizing and to get more value from their existing investments. * Created Build and Release for multiple projects (modules) in production environment using **Visual Studio Team Services (VSTS)**. | | | |

* Designed and Developed Real time Stream processing Application using **Spark, Kafka, Scala,** and **Hive**

to perform **Streaming ETL** and apply Machine Learning.

* Collected the **JSON** data from **HTTP** Source and developed **Spark APIs** that helps to do inserts and updates in **Hive tables.**
* Used **Azure Data Factory, SQL API and MongoDB API** and integrated data from **MongoDB, MS SQL**,

and **Azure SQL DB.**

* Responsible for resolving the issues and troubleshooting related to performance of **Hadoop** cluster.
* Involved in designing and developing tables in **HBase** and storing aggregated data from **Hive Table**.
* Used **Jira** for bug tracking and **Bit Bucket** to check-in and checkout code changes.
* Analyzed the **SQL** scripts and designed it by using **PySpark SQL** for faster performance.
* Developed spark applications in **python(PySpark)** on distributed environment to load huge number of

**CSV** files with different schema in to Hive ORC tables.

* Worked on reading and writing multiple data formats like **JSON** on **HDFS** using **PySpark**.
* Provide guidance to development team working on **PySpark** as **ETL platform**
* Performed all necessary day-to-day **GIT** support for different projects, responsible for maintenance of the **GIT** Repositories, and the access control strategies.

***Role: Sr Data Engineer October 2021 –***

***August 2022***

***Client: Deutsche Bank, New York, NY***

**Responsibilities:**

* Designed and setup Enterprise **Data Lake** to provide support for various uses cases including Analytics, processing, storing, and Reporting of voluminous, rapidly changing data.
* Designed and developed Security Framework to provide fine grained access to objects in **AWS S3 using AWS Lambda, DynamoDB.**
* Worked on authentication principals to establish secure network communication on cluster and testing

of **HDFS, Hive, Pig** and **MapReduce** to access cluster for new users.

* Performed end- to-end Architecture & implementation assessment of various **AWS** services like

# Amazon EMR, Redshift, S3.

* Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using **S3 data lake.**
* Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (**Amazon S3**) and **Amazon DynamoDB**.
* Used **Spark SQL** for Scala & amp, **Python** interface that automatically converts **RDD** case classes to

schema RDD.

* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Creating **Lambda** functions with Boto3 to deregister unused AMIs in all application regions to reduce the cost for **EC2 resources**.
* Developed reusable framework to be leveraged for future migrations that automates **ETL** from RDBMS systems to the Data Lake utilizing **Spark Data Sources** and **Hive** data objects.
* Conducted Data blending, Data preparation using Alteryx and SQL for **Tableau** consumption and publishing data sources to Tableau server.
* Developed Kibana Dashboards based on the Log stash data and Integrated different source and target systems into Elasticsearch for near real time log analysis of monitoring End to End transactions.
* Implemented **AWS Step** Functions to automate and orchestrate the **Amazon SageMaker** related tasks such as publishing data to **S3**, training **ML** model and deploying it for prediction.
* Integrated **Apache Airflow** with **AWS** to monitor multi-stage ML workflows with the tasks running on

# Amazon SageMaker.

***Role: Sr Data Engineer September 2019 - September***

***2021***

***Client: AT&T, Plano, TX***

**Responsibilities:**

* Experience in migrating existing legacy applications into optimized data pipelines using **Spark** with

**Scala** and Python, supporting testability and observability.

* Experience in developing scalable real-time applications for ingesting clickstream data using **Kafka**

Streams and Spark Streaming.

* Developed optimized and tuned **ETL** operations in Hive and Spark scripts using techniques such as partitioning, bucketing, vectorization, serialization, configuring memory and number of executors.
* Worked on Talend integrations to ingest data from multiple sources into **Data Lake**.
* Developed an **MVP** on exporting data to Snowflake to understand usages and benefits for migration.
* Experience in developing 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 the

customer usage patterns.

* Implemented cloud integrations to **GCP** and **Azure** for bi-directional flow setups for data migrations.
* Developed various scripting functionality using Shell Script and Python.
* Built **Jupyter notebooks** using **PySpark** for extensive data analysis and exploration.
* Implemented code coverage and integrations using Sonar for improving code testability.
* Pushed application logs and data streams logs to **Kibana** server for monitoring and alerting purpose.
* Worked on migrating data from HDFS to Azure HD Insights and **Azure Databricks**.
* Experience designing solutions in Azure tools like **Azure Data Factory**, **Azure Data Lake**, **SQL DWH**, **Azure SQL** & **Azure SQL Data Warehouse**, **Azure Functions**.
* Migrated existing processes and data from our on-premises **SQL Server** and other environments to

# Azure Data Lake.

* Implemented multiple modules in microservices to expose data through **Restful API’s**.
* Developed **Jenkins pipelines** for continuous integration and deployment purpose.

***Role: Data Engineer August 2017 - July***

***2019***

***Client: Amigos Software Solutions, Hyderabad, India***

**Responsibilities:**

* Developed **scripts**, **UDF**s using Data frames and **RDD** in Spark for **Data Ingestion** and **aggregation** from scratch.
* Involved in developing following frameworks using **Python** and **spark**.
  1. Comet Framework:
     1. To filter and validate the data in **AWS S3** using **Pyspark** which is replicated by **Attunity**

replication tool.

* + 1. To Process the **CDC** data in different mode like **NRT** (Near Real Time), **BATCH, ACTIVE** and finally store in **hive**.
  1. File Framework:
     1. For uncompressing the different compressed files (zip, tar, tar.gz).
     2. To process delimited files like **csv, tsv, dat**, fixed width and pipe delimited files using python and spark data frames and finally saving into hive table.
  2. Sqoop Framework:
     1. To import data from **Oracle, DB2** and **SQL server** source to s3 using **Sqoop** import and then processing the data from s3 and creating table on it.
  3. Airflow Framework:
     1. Creating **dags** for the source system to be ingested in production which is python code.
     2. Scheduling and submitting the spark command from **AWS EC2** to **AWS EMR**.
* Replicated data from Oracle databases to S3 using **Attunity** replication tool by creating tasks and providing source and destination connections.
* Tracking the corrupted records for each table and updating those corrupted records separately in audit table.
* Experience in dealing **JSON, CSV, TSV, DAT** and pipe delimiter files.
* Created **ETL** Jobs for converting **CSV** files to Parquet with the help of crawlers.
* Developed scripts using python boto3 to uncompressed zipped files in S3.
* Closely worked with **QA** team to fix all the bugs and made the code more stabilized.
* Experience in bit bucket code repository.
* Hands on deploying code using **Jenkins**.
* Developed **SQL** scripts for extracting metadata from on premise **Oracle DB** to **Cloud Aurora DB**.
* By Using Object Tagging concept in s3 categorized the objects in s3 i.e. tagged the objects after processing using Python scripts.
* Created databases in **AWS Glue** catalogue using **Python script**.
* Developed **Python script** for converting fixed width delimited files to **csv** files.
* Done Validations checkouts for active and history tables.

***Role: Software Developer January 2015 - July***

***2017***

***Client: Hudda Infotech Private Limited, Hyderabad, India***

**Responsibilities:**

* Design and implement database solutions in **Azure SQL Data Warehouse, Azure SQL**.
* Architect & implement 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, HDInsight/Databricks, NoSQL DB).

* Design & implement migration strategies for traditional systems on **Azure** (Lift and shift/**Azure**

**Migrate**, other third-party tools.

* Engage with business users to gather requirements, design visualizations, and provide training to use self-service **BI tools**.
* Used various sources to pull data into **Power BI** such as **SQL Server, Excel, Oracle, SQL Azure** etc.
* Propose architectures considering cost/spend in **Azure** and develop recommendations to right-size data infrastructure.
* Develop conceptual solutions & create proof-of-concepts to demonstrate viability of solutions.
* Technically guide projects through to completion within target timeframes.
* Collaborate with application architects and **DevOps.**
* Identify and implement best practices, tools and standards.
* Design Setup maintain Administrator **the Azure SQL Database, Azure Analysis Service, Azure SQL Data warehouse, Azure Data Factory, Azure SQL Data warehouse**.
* Build Complex distributed systems involving huge amount data handling, collecting metrics building

**data pipeline**, and Analytics.