Naveen
Azure Data Engineer
937-759-3993
naveenpasup12@gmail.com

Professional Summary:

- 4 years of experience as a Azure Data Engineer/SME with expertise in designing data intensive applications using Hadoop Ecosystem, Big Data Analytics, Cloud Data engineering, Data Warehouse/ Data Mart, Data Visualization, Reporting.
- As a data engineer, specialize in AWS and Azure frameworks, Cloudera, Hadoop Ecosystem, and Snowflake, relational databases tools like Tableau, Airflow, DBT, Presto/Athena, and Data DevOps Frameworks/Pipelines with strong Programming/Scripting skills in Python.
- Well versed with big data on AWS cloud services i.e. EC2, S3, Glue, Anthena, DynamoDB and RedShift
- Well versed with HADOOP framework and analysis, design, development, documentation, deployment and integration using SQL and big data technologies.
- Expertise in using major components of Hadoop ecosystem components like HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, Python, HBase, Spark, Ariflow, Spark SQL, C#, Kafka, Spark Streaming, Flume, Oozie, Zookeeper, Hue.
- Hands on experience in setting up workflow using Apache Airflow and Oozie workflow engine for
- Managing and scheduling Hadoop jobs.
- Strong development skills with Azure Data Lake, Azure Data Factory, SQL Data Warehouse Azure Blob, Azure Storage Explorer.
- Recreating existing application logic and functionality in the Azure Data Lake, Data Factory, SQL Database, and SQL Data warehouse environment.
- Architect and implement ETL and data movement solutions using Azure Data Factory (ADF), SSIS.
- Experience with Data warehousing and Data Mining using one or more NoSQL databases like HBase, Cassandra and Mango DB.
- Knowledge in integration of data from various sources like RDBMS, Spreadsheets, text files, JSON files, Delimited files.
- Docker container orchestration using ECS, ALB and lambda.
- Good understanding of Amazon Web Services (AWS) like EC2 for computing and S3 as storage mechanism and EMR, RedShift, DynamoDB.
- Experience in handling services on AWS, Glue, Azure cloud like S3 for storage management, creating/configuring/integrating IAM roles, EC2 etc, knowledge on handling streaming data.
- Well versed in Normalization / De-normalization techniques for optimum
- Performance in relational and dimensional database environments and implemented various data warehouse projects in Agile Scrum/Waterfall methodologies.
- Excellent Software Development Life Cycle (SDLC) with good working knowledge of testing methodologies, disciplines, tasks, resources and scheduling.
- Worked in large and small teams for systems requirement, design & development.
- Key participant in all phases of software development life cycle with Analysis, Design, Development, Integration, Implementation, Debugging, and Testing of Software Applications in client server environment, Object Oriented Experience in using various IDEs Eclipse, IntelliJ, and repositories SVN and Git.
- Used Azure DevOps and VSTS (Visual Studio Team Services) for CI/CD, Active Directory for authentication and Apache Ranger for authorization.

Technical Skills:

Data Processing and Storage	Azure Synapse, Apache Spark, Python, SQL, Databricks, Azure Data Factory, ADLS

Programming Languages	Python, Scala, Unix Shell Scripting
Databases	SQL Server, MySQL, Oracle
Methodologies	Agile methodology, Waterfall model
Azure Cloud Platform	Azure Data Factory v2, Azure Blob Storage, Azure Data Lake Gen 1 & Gen 2, Azure SQL DB, SQL server, Logic Apps, Azure Synapse, Azure Analytic Services, Data bricks, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Key Vault, Azure App Services, Logic Apps, Event Grid, Service Bus, Azure DevOps, ARM Templates.

Professional Experience:

Client:CareSource,Dayton,OH
Duration: Nov 2022 To Present
Role: Azure Data Engineer

Responsibilities:

- Analysed, designed and built Modern data solutions using Azure PaaS service to support visualization of data. Understand current Production state of application and determine the impact of new implementation on existing business processes.
- Extracted 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.
- Developing ETL pipelines in and out of data warehouse using combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.
- Implemented Proof of concepts for SOAP & REST APIs.
- Used Spark SQL for Scala & amp, Python interface that automatically converts RDD case classes to schema RDD.
- Responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
- Used AWS EMR clusters for creating Hadoop and spark clusters. These clusters are used for submitting and executing Scala and python applications in production.
- Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in HDFS.
- Created Airflow Scheduling scripts in Python.
- REST APIs to retrieve analytics data from different data feeds Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
- Extensively worked on SSIS script task with C# and Vb.net scripting.
- Developed Spark applications using Python, Pyspark 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.
- Responsible for estimating the cluster size, monitoring and troubleshooting of the Spark data bricks cluster.
- Working on fetching data from various source systems such as Hive, Amazon S3, and AWS Kinesis.
- Spark Streaming gathers this data from AWS Kinesis in near real-time, performs the necessary transformations and aggregation on the fly, and persists the data in a NoSQL store to build HBase.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
- 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 kinesis firehose and S3 Data Lake.
- Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the AQL Activity.

Environment: SQL Server Management, VSTS, Azure SQL, Azure Storage Explorer, Azure Blob, Power BI, PowerShell, C# .Net, SSIS, DataGrid, ETL Extract Transformation and Load, Business Intelligence (BI), Python, Azure Storage, Azure Data Factory, Azure SQL Analytics, Azure Blob Storage, Azure Backup, Azure Files, Azure Data Lake Storage, Azure App Services, Azure Web Apps, Azure Logic Apps, Azure Virtual Machine (VM), Windows Server, Unix, LINUX.

Client: Xpheno, India
Role: Azure Data Engineer
Duration: July 2019 To Aug 2021

Responsibilities:

- Worked on Azure Data Factory to integrate data of both on-prem (MY SQL, Cassandra) and cloud (Blob storage, Azure SQL DB) and applied transformations to load back to Azure Synapse.
- Evolved in Spark Scale functions for mining data to provide real time insights and reports. Configured spark streaming to receive real time data from the Apache Flume and store the stream data using Scala to Azure Table.
- Data Lake is used to store and do all types of processing and analytics. Ingested data into AzureBlob storage and processed the data using Data bricks.
- Creating pipelines using Azure Data factory and Azure apps services to pull data from Webservices and APIs in Azure SQL.
- Transforming data in Azure Data Factory with ADF Transformations.
- Worked on different files like CSV, JSON, Flat, fixed width to load the data from source to raw tables
- Implemented Triggers to schedule pipelines.
- Use various types of activities: data movement activities, transformations, and control activities; Copy data, Data flow, Get Metadata, Lookup, Stored procedure, Execute Pipeline
- Used Flume sink to write directly to indexers deployed on cluster, allowing indexing during ingestion. Migrated from Oozie to Apache Airflow. Involved in developing Oozie and Airflow. Workflows for daily incremental loads, getting data from RDBMS (MongoDB, MS SQL).
- Managed resources and scheduling across the cluster using AzureKubernetes Service. AKS can be used to create, configure and manage a cluster of Virtual machines.
- Using DAX queries, I created Calculated Columns and Measures in Power BI and Excel according to the needs.
- Extensively used Kubernetes which is possible to handle all the online and batch workloads required to feed, analytics and machine learning applications.
- Azure SQL Data Modelling Azure Data Factory Azure Devops Azure Storage Azure Synapse
- Involved in processing and flattening JSON data using Mapping Data Flows, and Pyspark.
- Actively involved in monitoring and support of Data Factory pipelines and Databricks notebooks.
- Designed and developed SSIS Packages to extract data from various data sources such as Access database, Excel spreadsheet and flat files into SQL server for further Data Analysis and Reporting by using multiple transformations provided by SSIS such as Data Conversion, Conditional Split, merge, union all and lookup transformation and send mail task.
- Worked as a BI developer with following responsibilities
- Maintained and optimized existing databases, monitored database performances and growth, performance tuning,
- Involved in using Spark Data Frames to create Various Datasets and applied business transformations and data cleansing operations using Data Bricks Notebooks. Data Modelling Efficient in writing Python scripts to build ETL pipeline and Directed Acyclic Graph (DAG) workflows using Airflow, Apache NiFi.
- Tasks are distribution on celery workers to manage communication between multiple services. Monitored Spark cluster using Log Analytics and Ambari Web UI. Transitioned log storage from Cassandra to AZURE SQL Data warehouse and improved the query performance.

- Involved in developing data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL. Also Worked with Cosmos DB (SQL API and Mongo API).
- Designed custom-built input adapters using Spark, Hive, and Sqoop to ingest and analyse data (Snowflake, MS SQL, and MongoDB) into Data Modelling HDFS.
- Responsible for creating complete test cases, test plans, test data, and reporting status ensuring accurate coverage of requirements and business processes.
- Experience in SDLC and Agile methodologies such as SCRUM.
- Analyzing requirements and creating and executing ETL test cases.
- ETL Test Case Execution and Adhoc testing
- Performed Integration, End-to-End, system ETL testing.
- Create external tables and views in Azure synapse to read data ADLS Gen2 and integrate this with tableau for reporting.
- Proficient in writing DAX with optimal performance. Tuning SSAS cubes to improve the performance of Power BI reports.
- Tuned the performance of SQL queries and Stored procedures using SQL Profiler. Involved in database and log backups & restoration, backup strategies, scheduling.
- Experience in working with Spark applications like batch interval time, level of parallelism, memory tuning to improve the processing time and efficiency.

Environment:PL/SQL, Python, Azure-Data factory, Data Modelling Azure Blob storage, Azure table storage, Azure SQL server,Apache Hive, Apache Spark, MDM, Typescript Fargate Aurora Container Tech IaC Netezza, Teradata, Oracle 12c, SQL Server,SSAS Teradata SQL Assistant, Teradata Vantage, Microsoft Word/Excel, Flask, Snowflake, DynamoDB, Athena, Lambda,MongoDB,Pig, Sqoop, Tableau, PowerBI

Education:

- ❖ Bachelors in computer science from SRM UNIVERSITY.
- ❖ Masters in computer science from University of Dayton.