Yasaswini Annem
annemyasaswini99@gmail.com | **512-545-2908**

Professional Summary:

- 4+ years of IT experience in a variety of industries working on Big Data technology using technologies such as Cloudera. Hadoop working environment includes Hadoop, Spark, MapReduce, Kafka, Hive, HBase, and Impala.
- Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like MapReduce, YARN, Hive, HBase, Flume, Spark MLlib, Spark GraphX, Spark SQL, Kafka.
- Adept at configuring and installing Hadoop/Spark Ecosystem Components.
- Knowledge of HDFS File system and Hadoop Demons such as Resource Manager, Node Manager, Name Node, Data Node, Secondary Name Node, Containers, Map Reduce programming paradigm, and good hands-on experience in PySpark and SQL queries.
- Developed a data pipeline using Kafka and Spark Streaming to store data into HDFS and performed real-time analytics on the incoming data.
- Proficient with Spark Core, Spark SQL, Spark MLlib, Spark GraphX and Spark Streaming for processing and transforming complex data using in-memory computing capabilities written in Scala.
- Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into a data warehouse.
- Extensive experience in developing applications that perform Data Processing tasks using Teradata, Oracle, SQL Server and MySQL database.
- Extensively worked on Spark using Scala on cluster for computational (analytics), installed it on top of Hadoop performed advanced analytical application by making use of Spark with Hive and SQL/Oracle.
- Able to use Sqoop to migrate data between RDBMS, NoSQL databases and HDFS.
- Has experience in working on Python libraries Pandas, NumPy.
- Managing Database, Azure Data Platform services (Azure Data Lake(ADLS), Data Factory(ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes
- Implemented a 'server less' architecture using API Gateway, Lambda, and Dynamo DB and deployed AWS Lambda code from Amazon S3 buckets. Created a Lambda Deployment function, and configured it to receive events from your S3 bucket.
- Expertise and Vast knowledge of Enterprise Data Warehousing including Data Modeling, Data Architecture, Data Integration (ETL/ELT), and Business Intelligence.
- Experience in developing Map Reduce Programs using Apache Hadoop for analyzing the big data as per the requirement.
- Extensive experience using MAVEN as a Build Tool for the building of deployable artifacts from source code.
- Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
- Designed the data models to be used in data intensive AWS Lambda applications which are aimed to do complex analysis creating analytical reports for end-to-end traceability, lineage, and definition of Key Business elements from Aurora.
- Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats.
- Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python and build models using deep learning frameworks.
- Extensive experience in writing UNIX shell scripts and automation of the ETL processes using UNIX shell scripting.

- Data warehouse solutions using polybase/external table on Azure Synapse/Azure SQL Data warehouse (Azure DW), Using Azure Data Lake as source. Rewriting exiting SSAS cubes to Azure Synapse/Azure SQL Data warehouse (Azure DW).
- Developed automated process for code builds and deployments using Jenkins, Ant, Maven, Sonar type, Shell Script.
- Experience in Converting existing AWS Infrastructure to Server less architecture (AWS Lambda), deploying via Terraform and AWS Cloud Formation templates.
- Strong SQL development skills including writing Stored Procedures, Triggers, Views, and User Defined functions.
- Developed Spark Applications that can handle data from various RDBMS (MySQL, Oracle Database) and Streaming sources.
- Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting Up AWS and Microsoft Azure with Databricks, Databricks Workspace for Business Analytics, Manage Clusters In Databricks.
- Experience in analyzing data using HiveQL, Pig, HBase and custom MapReduce programs.

Technical Skills:

TECHNICAL SKILLS
Big Data Tools: Hadoop, Map Reduce, HDFS, Spark, Airflow, HBase, Hive, Kafka
BI Tools: Tableau, Power BI, SSIS
Programming Languages: SQL, Python, Java
Methodologies: System Development Life Cycle (SDLC), Agile
Cloud Management: Azure, AWS
Databases: MySQL, PostgreSQL, Dbeaver, Oracle, SSMS, DB2, Teradata, MongoDB, Cassandra, HBase
ETL/Data warehouse Tools: Informatica power center, IICS, Informatica MDM, and Informatica IDQ
Version Control Systems: Git, GitHub, Bitbucket
Agile Tools & Scheduling Tools: Jira, Confluence, Service Now, Control-M, Autosys
Software: Postman, Insomnia, REST APIs
Operating Systems: Unix, Linux, Windows

Professional Experience:

Graduate Assistant Data Engineer- Career Services, TXST, Tx (Oct 2021 – Aug 2023)

Responsibilities:

- Developed batch processing solutions by using Data Factory and Azure Data bricks
- Implemented Azure Data bricks clusters, notebooks, jobs and auto scaling.
- Designed relational and non-relational data stores on Azure
- Worked on ETL tool Informatica, Oracle Database and PL/SQL, Python and Shell Scripts.
- Created pipelines in ADF using linked services to extract, transform and load data from multiple sources like Azure SQL, Blob storage and Azure SQL Data warehouse.
- Used Python to perform data cleaning and make data structured and easily converted into DF.
- Ample knowledge of data architecture including data ingestion pipeline design, Hadoop/Spark architecture, data modeling, data mining, machine learning and advanced data processing.
- Designed and created Data Architect Specifications for various ETL projects.
- Successfully implemented POC (Proof of Concept) in a Development Databases to validate the requirements and benchmarking the ETL loads.

- Worked on data cleaning and reshaping, generated segmented subsets using NumPy and Pandas in Python.
- Created stage execute script in python which do the ddl and dml operation on hive landing tables and load this data to hive stage tables. This script handles the SCD type 2 and simple pass through load.
- Experienced in implementing Azure data solutions, provisioning storage account, Azure Data Factory, SQL server, SQL Databases, SQL Data warehouse, Azure Data Bricks and Azure Cosmos DB
- Orchestrated hundreds of Sqoop scripts, python scripts, Hive queries using Oozie workflows and sub- workflows
- Designed and implemented by configuring Topics in new Kafka cluster in all environment.
- Extracted large datasets from Teradata using utilities like fastexport, bteq and Load into Verticas.
- Wrote PL/SQL script, shell script to support ETL dataflow.
- Implemented to reprocess the failure messages in Kafka using offset id.
- Implemented Kafka producer and consumer applications on Kafka cluster setup with help of Zookeeper.
- Used Spring Kafka API calls to process the messages smoothly on Kafka Cluster setup.
- Responsible for data services and data movement infrastructures good experience with ETL concepts, building ETL solutions and Data modeling
- Used ETL to implement the Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
- Designed Data Marts by following Star Schema and Snowflake Schema Methodology, using industry leading Data modeling tools.
- Managed Confidential Redshift clusters such as launching the cluster and specifying the node type.
- Have good experience working with Azure BLOB and Data lake storage and loading data into Azure SQL Synapse analytics (DW)
- Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
- Implemented data ingestion and handling clusters in real time processing using Apache Storm and Kafka.
- Developed Python-based API (RESTful Web Service) to track revenue and perform revenue analysis.

Application Development Associate - Accenture PLC, Hyderabad, India (Feb 2021– July 2021)

Responsibilities:

- Used ETL to implement the Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
- Designed Data Marts by following Star Schema and Snowflake Schema Methodology, using industry leading Data modeling tools.
- Developed Spark Programs for Batch and Real-Time Processing to process incoming streams of data from Kafka sources and transform them into Data frames and load those data frames into Hive and HDFS.
- Utilized Spark, Scala, Hadoop, HBase, Cassandra, MongoDB, Kafka, Spark Streaming, MLlib, and Python and utilized the engine to increase user lifetime by 45% and triple user conversations for target categories.
- Developed Spark Applications by using Scala, and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
- Worked with the Spark for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD's, Spark YARN.
- Creating 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.
- Monitoring end to end integration using Azure monitor.
- Built messaging system to get data from different sources and produce it using Kafka.
- Created Python script to push data to HDFS directory. Created hive landing tables on top of these hdfs data files.
- Performing ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.

- 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.
- Worked on various Spark optimizations techniques pocs for memory management, garbage collection, Serialization, and custom partitioning.
- Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python.
- Experience managing Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services.
- Developed Spark programs to parse the raw data, populate staging tables, and store the refined data in partitioned tables in the EDW.
- Created Stored Procedures, Database Triggers, Functions and Packages to manipulate the database and to apply the business logic according to the user's specifications.
- Creating new workflows and maintaining Data access existing ETL workflows, data management, and data query components.
- Design, develop and orchestrate data pipelines for real-time and batch data processing using AWS Redshift
- Performed Exploratory Data Analysis and Data visualizations using Python and Tableau.
- Worked in writing SPARK SQL query scripts for optimizing the query performance.
- Implemented Spark Scripts using Spark Session, Python, Spark SQL to access hive tables data flow into spark for faster processing of data.
- Stored the data in MongoDB NoSQL and performed different transactions.
- Collaborated with DBAs on performance, backup strategies, security, and standard methodologies.
- Supported the design, integration, and testing of automated data pipelines.

ETL Developer & Tester - Khemas Engineers, Hyderabad, India (Aug 2017 – Jan 2021)

Responsibilities:

- Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python.
- Experience in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
- Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
- 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 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.
- Developed and implemented ETL pipelines using Python, SQL, Spark and PySpark to ingest data and updates to relevant databases.
- Using python, the ETL pipeline was developed and programmed to collect data from Redshift data warehouse.
- Worked on fine-tuning spark applications to improve the overall processing time for the pipelines.
- Used Airflow for scheduling and orchestration of the data pipelines.
- Extensively worked with pyspark / Spark SQL for data cleansing and generating Data Frames and RDDs
- 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
- Designed and developed Informatica workflows to exchange data with Oracle databases, Salesforce, Data Lake, and other operational and warehouse data stores, ensuring seamless data integration.
- Created ETL solutions from a variety of data sources using SQL and big data technologies.
- Successfully managed the conversion of ETL mappings from Oracle to PostgreSQL, enhancing data processing efficiency.
- Collaborated with infrastructure teams to optimize Informatica objects, enhance throughput, and plan for capacity, resulting in improved system performance.

- Configured and tuned complex Informatica maps, workflows, and session logs, meeting performance and recovery objectives.
- Coordinated with development and infrastructure groups to prepare and implement planned maintenance outages, minimizing service disruptions.
- Strong understanding of ETL, data analysis, metadata, data quality, audit, design, version control, and CI/CD within the technology stack.
- Design, Creation, Execution, Review of Unit Test case , SIT(System Integration Testing) Test Cases , Functional Test Cases based on Client Requirements.
- Create Spark code to process streaming data from Kafka cluster and load the data to staging area for processing.
- Involved in Functional Testing, Integration testing, Regression Testing, Smoke testing and performance Testing. Tested Hadoop Map Reduce developed in python, pig, Hive
- Implemented CI/CD pipelines using Git and Jenkins for efficient code integration and deployment.
- Involved in developing Spark SQL queries, Data frames, import data from Data sources, perform transformations, perform read/write operations, save the results to output directory into HDFS.
- Designed, built, and operated data tools, services, and workflows using modern data engineering tools and orchestration tools and Used copy command script to load data into staging tables.
- Developed and optimized ETL processes to leverage cloud and migrated existing data pipelines as needed.
- Loading data from azure data lake storage into enterprise data warehouse.

Education:

- MS in Computer Science - Texas State University, San Marcos, TX, USA | GPA: 3.75 (Aug 2021 - Aug 2023)
- BTech in Information Technology - BVRITH - JNTUH, Hyderabad, TS, IND | GPA: 7.08 (Aug 2016 - Sept 2020)

Achievements:

- One of the youngest Area Directors appointed, Advanced Leadership Bronze & Competent Communicator Award: D98 (TI)
- Winning team of ENLITE 2018, a ten – day boot camp on Innovation & Entrepreneurship by JHUB, JNTUH.

Certifications:

- Diversity, Equity, and Inclusion in the Workplace certification from University of South Florida
- Project Management Essentials certification from Management and Strategic Institute
- LinkedIn certifications from ‘Become a Data Analyst’ path: The Non-Technical Skills of Effective Data Scientists, Learning Excel: Data Analytics, Learning Data Analytics: 1 Foundations, Learning Data Analytics Part 2: Extending and Applying Core Knowledge, Data Fluency: Exploring and Describing Data, Excel Statistics Essential Training: 1
- Ongoing: Google Data Analytics Professional Certification, Google Business Intelligence Professional Certification