Advait Narvekar

Lead AWS Data Engineer

advaitnarvekar03@gmail.com
607 713-8155

Certified Advanced
PYTHON

SUMMARY

Results-oriented IT professional with over 12 years of experience in software and big data applications within AWS and Azure ecosystems. Specialized in tools such as Cloudera, Hadoop, Spark, and Snowflake, with a strong focus on Data Warehousing and ETL processes using Informatica, AWS Glue, and Spark SQL. Proven ability to design efficient ETL architectures, optimize SQL queries, and implement data integration solutions, ensuring data security and governance. Experienced in developing Spark applications using Databricks for data extraction, transformation, and aggregation, as well as real-time data analytics with Spark Streaming and Kafka. Proficient in Python for workflow management using Apache Airflow and for building data pipelines. Skilled in monitoring and maintaining cloud architecture with AWS services and CI/CD practices using Jenkins and Terraform. Strong analytical skills in extracting insights from complex datasets and contributing to organizational data strategy initiatives. Having good experience over AI/ML.

TECHNICAL SKILLS:

Big Data Ecosystems	Hadoop, Map Reduce, HDFS, HBase, Hive, Pig, Sqoop, Kafka, Cassandra, Flume,
Cloud	AWS Snowflake, AWS RDS, AWS Aurora, Redshift, EC2, EMR, S3, Lambda, Glue, Data Pipeline, Athena, Data Migration Services, SQS, SNS, ELB, VPC, EBS, RDS, Route53, Cloud Watch, AWS Auto Scaling, Git, AWS CLI, Jenkins, Microsoft Azure.
Programming Languages	Java Scala, Python, HTML,
Scripting Languages	JavaScript, XML, HTML, Python and shell
Databases	NoSQL, Oracle, Microsoft SQL Server.
UNIX Tools	Apache, Yum, RPM
Tools	Eclipse, JDeveloper, JProbe, CVS, Ant, MS Visual Studio
Application Servers	Apache Tomcat 5.x 6.0, Jboss 4.0
Testing Tools	Net Beans, Eclipse, WSAD, RAD
Methodologies	Agile, UML, Design Patterns

CERTIFICATIONS:

Advance Python Certified
Databricks Certified
AWS Data Engineer Certified

EDUCATIONAL DETAILS:

University of Mumbai | India | Bachelor of Computer Eng. (B.E.) |2012

PROFESSIONAL EXPERIENCE:

Prudential Financial, New York, USA, (Remote) **Nov 2023 – Present**
Lead Data Engineer
Responsibilities

- Performing assessment of migration viability, assess total number of programmable objects and gather statistics related to migration schemas.

- Currently leading the program for onboarding data pipelines for US Regulatory reports like CCAR, FRY9C and FFIEC-031 call report on AWS and Databricks platform handling complex financial data and addressing user challenges around data latency/SLA, lineage and drill down for granular analysis.
- Led a team of 6 data engineers in the migration and transformation efforts, overseeing ETL development and ensuring best practices in data quality and performance optimization.
- Designed complex data transformation workflows with DBT's ability to handle cross-model dependencies, allowing for logical chaining of transformations which resulted in simplified management of complex data relationships.
- Refactored ORCALE 230+ packages, 150+ procedures,30+functions to be compatible with Postgres.
- Migrated 4 application schemas approximately of size 800Gigs, containing 800+ tables.
- Developed data pipelines from on premises to cloud using AWS services like DMS, GLUE, S3, EC2, EMR and Redshift.
- AWS environment coordination with cloud platform team - RDS, policies, IAM roles, S3 buckets, Secrets Manager, CloudFormation.
- Refactored code is maintained and managed in bitbucket and pipeline is deployed via Jenkins with flyway configuration.
- Hands on experience in application development using Java, python, RDBMS, and Linux shell scripting and python scripting.
- Designed and developed high-performance ETL workflows to load large datasets into Netezza data warehouse for enterprise reporting.
- Implemented testing frameworks in DBT to validate data quality by writing custom tests for non-null constraints, uniqueness, referential integrity, and data type checks.
- Streamlined microservices development by leveraging Spring Boot's auto-configuration and embedded servers for rapid prototyping of data tools.
- Managed DB Ops using Flyway and Bitbucket, deploying all database changes to production through automated CI/CD pipelines.
- Developed a generative AI system that creates automated narrative reports for CCAR and FFIEC regulatory filings. Leveraged OpenAI's GPT-3 model to generate comprehensive insights based on structured data inputs, streamlining the reporting process.
- Worked with on-prem platform team in establishing data/file transfer between on-prem and cloud platform.
- Migrated historical data using AWS Database Migration Service (DMS).
- Designed and implemented machine learning models in Databricks using MLlib and TensorFlow, leveraging large datasets for predictive analytics and classification tasks relevant to regulatory reporting.
- Designed, developed, and managed multiple AWS Glue ETL jobs to process, transform, and load data into Amazon S3, Redshift, and other data warehouses.
- Leveraged AWS Glue to integrate data from various sources like S3, Redshift, and Athena, enabling seamless data flow and query execution.
- Optimize long running jobs using explain plan and create index, executing vacuum analyze, create intermediate table loads etc.
- Developed batch and streaming processing apps using Spark APIs 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 wide, narrow transformations, actions like filter, Lookup, Join, count, etc. on Spark Data Frames.
- Writing Python Applications which runs EMR cluster that fetches data from the S3/one lake location and queue it in the Amazon SQS (simple Queue Services) queue.
- Stored data into various tiers of AWS S3 based on business requirements and frequency of data access.
- Data validation through informatica dvo for flat files and oracle vs Postgres tables (includes history and incremental data loads)
- Automation of inbound/outbound file transfers through cyber fusion and email using python shell scripts
- Automation of glue job triggers on schedule using Autosys.
- Worked with database administrating team on SQL optimization for databases like Oracle.
- Ensured database performance in production by stress testing AWS EC2 of PostgreSQL environments.
- Automated deployments and routine tasks using UNIX Shell Scripting.
- Worked in an agile environment to implement projects and enhancements with weekly SCRUMs.

Tools and Technologies Used: AWS (S3, Glue, Redshift, Lambda, Databricks, Step Functions, Kinesis, DynamoDB, MSK, IAM, Lake Formation), PySpark, Terraform, GitLab CI/CD, SQL, Spark, Python, CloudWatch, Salesforce, API Gateway.

Kaiser Permanente, Oakland, CA **Jun 2020 – Oct 2023**
SR.AWS Data engineer
Responsibilities

- Led and executed a comprehensive data migration project, seamlessly transitioning critical Healthcare data assets from on-premise source systems (Teradata and Hadoop) to Snowflake.
- Developing AWS Lambda functions which creates the EMR Cluster and auto terminates the cluster after job is done.
- Developed data transition programs from DynamoDB to AWS Redshift (ETL Process) using AWS Lambda by creating functions in Python for the certain events based on use cases.
- Involved in building a data pipeline and performed analytics using AWS stack (EMR, EC2, EBS, Elastic search, Kinesis, SQS, DynamoDB, S3, RDS, Lambda, Glue, SQS, and Redshift).
- Utilized AWS S3 as a secure and efficient bridge for data transfer, adhering to stringent healthcare data privacy and HIPPA security regulations.
- Developed complex DBT models to transform raw data from Snowflake and prepare it for analytical consumption, utilizing Jinja templating for reusable SQL code components.
- Writing Pyspark Applications which runs on Amazon EMR cluster that fetches data from the Amazon S3/one lake location and queue it in the Amazon SQS (simple Queue Services) queue.
- Built Kinesis dashboards and applications that respond to incoming data using AWS SDKs, exported data from kinesis to other AWS services, such as EMR for analytics, S3 for storage, Redshift for extensive data, and Lambda for event-driven actions.
- Utilized SageMaker Experiments for hyperparameter optimization, tracking experiments, and managing model versions, which facilitated collaboration with the data science team.
- Working on the Spark SQL for analyzing and applying the transformations on data frames created from the SQS queue and loads them into Database tables and querying.
- Working on Amazon S3 for persisting the transformed Spark Data Frames in S3 buckets and using Amazon S3 as a Data-lake to the data pipeline running on spark.
- Developing logging functions which stores logs of the pipeline in Amazon S3 buckets.
- Developing Email reconciliation reports for ETL load in Spark framework.
- Building PySpark applications for interactive analysis, batch processing, stream processing.
- Configuring Spark executor memory to speed-up spark jobs, developed unit tests for PySpark jobs, and perform tuning by analyzing Spark logs and job metrics.
- Worked with Data Science team running Machine Learning models on Spark EMR cluster and delivered the data needs as per business requirements.
- Utilized Spark's in memory capabilities to handle large datasets on S3 Data Lake. Loaded data into S3 buckets then filtered and loaded into Hive external tables.
- Strong Hands-on experience in creating and modifying SQL stored procedures, functions, views, indexes, and triggers.
- Developing End to End ETL Data pipeline that take the data from surge and loading it into the RDBMS using the Spark.
- Worked on 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.
- Developing Data load functions, which reads the schema of the input data and load the data into a table.
- Involved in extracting and enriching multiple DynamoDB table transformations.
- Automated the process of transforming and ingesting terabytes of monthly data using AWS Kinesis, S3, Lambda and Oozie.
- Created a semantic layer in DBT to standardize metrics and dimensions, allowing for consistent reporting across various business units within Kaiser Permanente.
- Performed ETL operations using Python, Spark SQL, S3 and Redshift on terabytes of data to obtain customer insights.
- Involved in writing Python scripts to automate the process of extracting weblogs using Airflow DAGs.
- Involved in writing unit tests, worked along with DevOps team in Installing libraries, Jenkins agents and ETL jobs.
- Used Ansible to provision the environment and deployed applications in a CI/CD process using a Jenkins pipeline. Also managed and deployed configurations using Terraform.
- Experience with analytical reporting and facilitating data for Tableau dashboards.
- Used Git for version control and Jira for project management, tracking issues and bugs.

Tools and Technologies Used: Hadoop 2.x, Hive v2.3.1, Databricks,Spark v2.1.3, AWS, EC2, S3, Lambda, Glue, Elasticsearch, RDS, DynamoDB, Redshift, ECS, Python, SQL, Sqoop v1.4.6, AWS Kinesis, Airflow v1.9.0, Oracle, Teradata, Tableau, Git, Jira.

Erie Insurance, Erie, PA **Mar 2017 – May 2020**
Data engineer

Responsibilities

- Played a lead role in gathering requirements, analysis of entire system and providing estimation on development, testing efforts.
- Designed and implemented a comprehensive data pipeline for Erie Insurance to streamline the migration of legacy data from on-premises systems to a cloud-based architecture on AWS.
- Involved in designing different components of system like Sqoop, Hadoop process involves map reduce & hive, Spark, FTP integration to down systems.
- Have written hive and spark queries using optimized ways like using window functions, customizing Hadoop shuffle & sort parameter
- Developed ETL's using PySpark.
- Used both Dataframe API and Spark SQL API.
- Using Spark, performed various transformations and actions and the final result data is saved back to HDFS from there to target database Snowflake
- Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS EMR
- Implemented Spark Streaming integrated with Kafka to provide Erie Insurance with live data processing capabilities.
- Configured Spark streaming to get ongoing information from the Kafka and store the stream information to HDFS
- Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift
- Used various spark Transformations and Actions for cleansing the input data
- Used Jira for ticketing and tracking issues and Jenkins for continuous integration and continuous deployment.
- Enforced standards and best practices around data catalog, data governance efforts
- Created DataStage jobs using different stages like Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel, Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, Row Generator, Etc.
- Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
- Building ETL pipeline for data ingestion, data transformation, data validation on cloud service AWS, working along with data steward under data compliance.
- Worked on scheduling all jobs using Airflow scripts using python added different tasks to DAG, LAMBDA.
- Used Pyspark for extract, filtering and transforming the Data in data pipelines.
- Monitoring servers using Nagios, Cloud watch and using ELK Stack Elasticsearch Kibana
- Used Data Build Tool (DBT) for transformations in ETL process, AWS lambda, AWS SQS
- Worked on scheduling all jobs using Airflow scripts using python. Adding different tasks to DAG's and dependencies between the tasks.
- 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.
- Responsible for estimating the cluster size, monitoring and troubleshooting of the Spark Databricks cluster.
- Created Unix Shell scripts to automate the data load processes to the target Data Warehouse.
- Responsible for implementing monitoring solutions in Ansible, Terraform, Docker, and Jenkins.

Tools and Technologies Used: PySpark, Spark Streaming, Databricks, EMR, S3, RDS, Redshift, Lambda, Boto3, DynamoDB, Apache Spark, HBase, Apache, HIVE, Map Reduce, Snowflake, Pig, Python, NumPy, Pandas, SSRS, Tableau.

Nike, Beaverton, OR **Oct 2015 – Feb 2017**
Big Data engineer
Responsibilities

- Designed and developed the real-time matching solution for customer data ingestion
- Worked on converting the multiple SQL Server and Oracle stored procedures into Hadoop using Spark SQL, Hive, Scala, and Java.
- Created production Data-lake that can handle transactional processing operations using Hadoop Eco-System.
- Developed PySpark and SparkSQL code to process the data in Apache Spark on Amazon EMR to perform the necessary transformations.
- Involved in validating and cleansing the data using Pig statements and hands-on experience in developing Pig MACROS.
- Analyzed dataset of 14M record count and reduced it to 1.3M by filtering out rows with duplicate customer IDs and

removed outliers using boxplots and univariate algorithms.

- Worked with Hadoop Big Data Integration with ETL on performing data extract, loading, and transformation process for ERP data.
- Performed extensive exploratory data analysis using Teradata to improve the quality of the dataset and created Data Visualizations using Tableau.
- Experienced in various Python libraries like Pandas, One dimensional NumPy, and Two dimensional NumPy.
- Experienced in using PyTorch library and implementing natural language processing.
- Developed data visualizations in Tableau to display day to day accuracy of the model with newly incoming Data.
- Worked with R for statistical modeling like Bayesian and hypothesis test with dplyr and BAS packages, and visualized testing results in R to delivery business insight
- Model validation by Confusion Matrix, ROC, AUC, and developed diagnostic tables and graphs that demonstrated how a model can be used to improve the efficiency of the selection process
- Presented and reported business insights by SSRS and Tableau dashboard combined with different diagrams
- Utilized Jira as project management methodology and Git for version control to build the program
- Reported and displayed the analysis result in the web browser with HTML and JavaScript
- Involved constructively with project teams, supported the project's goal through principle and delivered the insights for team and client

Tools and Technologies Used: Hadoop, Spark SQL, Hive, Scala, Java, MS Access, SQL Server, Pig, PySpark, Tableau, Excel

Mastek Ltd, Addison, TX **Apr 2013 – Oct 2015**
Data Analyst
Responsibilities

- Worked with Data Analyst for requirements gathering, business analysis and project coordination.
- Performed migration of Reports (Crystal Reports, and Excel) from one domain to another domain using Import/Export Wizard.
- Wrote a complex SQL, PL/SQL, Procedures, Functions, and Packages to validate data and testing process.
- Used advanced Excel formulas and functions like Pivot Tables, Lookup, If with and/index, match for data cleaning.
- Redesigned some of the previous models by adding some new entities and attributes as per the business requirements.
- Reviewed Stored Procedures for reports and wrote test queries against the source system (SQL Server) to match the results with the actual report against the Data mart (Oracle).
- Involved with data profiling for multiple sources and answered complex business questions by providing data to business users.
- Performed SQL validation to verify the data extracts integrity and record counts in the database tables
- Created Schema objects like Indexes, Views, and Sequences, triggers, grants, roles, Snapshots.
- Effectively used data blending feature in Tableau to connect different databases like Oracle, MS SQL Server.
- Transferred data with SAS/Access from the databases MS Access, Oracle into SAS data sets on Windows and UNIX.
- Provided guidance and insight on data visualization and dashboard design best practices in Tableau
- Performed Verification, Validation and Transformations on the Input data (Text files) before loading into target database.
- Executed data extraction programs/data profiling and analyzing data for accuracy and quality.
- Wrote complex SQL queries for validating the data against different kinds of reports generated by Business Objects.
- Documented designs and Transformation Rules engine for use of all the designers across the project.
- Designed and implemented basic SQL queries for testing and report/data validation
- Used ad hoc queries for querying and analyzing the data.
- Performed Gap Analysis to check the compatibility of the existing system infrastructure with the new business requirements.

Tools and Technologies Used: SQL, PL/SQL, Oracle9i, SAS, Business Objects, Tableau, Crystal Reports, T-SQL, SAS, UNIX, MS Access 2010.