**Name : Lekhasree Narayanagari**

**Cell: (780) 278-8320**

**Email:** **lsree1792@gmail.com**

**PROFESSIONAL SUMMARY**

* 6+ years of IT Experience in Diverse Domains: Extensive background in end-to-end data analytics solutions encompassing Big Data, Hadoop, Informatica, Data Modeling, and System Analysis.
* Azure DevOps Expertise: Proficiently build reusable YAML pipelines in Azure DevOps, create CI/CD pipelines using cloud-native architectures on Azure Cloud, and implement Git flow branching strategy.
* GCP Proficiency and Cloud Native Technologies: Mastery over GCP services including Big Query, Google Cloud Storage (GCS) buckets, Google Cloud Functions, and Dataflow.
* AWS and Hadoop Expertise: Proficient in AWS cloud services like EC2, S3, Glue, Athena, DynamoDB, and RedShift. Hands-on experience with the Hadoop ecosystem - HDFS, MapReduce, Pig, Hive, Sqoop, Flume, and Spark.
* Legacy Data Migration: Led successful migration projects from Teradata to AWS Redshift and on-premises to AWS Cloud.
* AWS Cloud-Based Pipelines: Utilized AWS services like EMR, Lambda, and Redshift to develop cloud-based pipelines and Spark applications.
* DevOps and Scripting Proficiency: Skilled in PowerShell scripting, Bash, YAML, JSON, GIT, Rest API, and Azure Resource Management (ARM) templates. Implement CI/CD standards, integrate security scanning tools, and manage pipelines effectively.
* Windows Scripting and Cloud Containerization: Proficient in scripting and debugging within Windows environments. Familiarity with container orchestration, Kubernetes, Docker, and AKS.
* Efficient Data Integration: Expertise in designing and deploying SSIS packages for data extraction, transformation, and loading into Azure SQL Database and Data Lake Storage. Configure SSIS Integration Runtime for Azure execution and optimize package performance.
* Data Visualization and Analysis: Create data visualizations using Python, Scala, and Tableau. Develop Spark scripts with custom RDDs in Scala for data transformation and actions. Conduct statistical analysis on healthcare data using Python and various tools.
* Big Data Ecosystem: Extensive experience with Amazon EC2 for computing, query processing, and storage. Proficiently set up Pipelines in Azure Data Factory using Linked Services, Datasets, and Pipelines for ETL tasks.
* Azure Data Services: ETL expertise using Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Ingest data to Azure Services and process within Azure Databricks.
* Hadoop Proficiency: Strong support experience across major Hadoop distributions - Cloudera, Amazon EMR, Azure HDInsight, Hortonworks. Proficient with Hadoop tools - HDFS, MapReduce, Yarn, Spark, Kafka, Hive, Impala, HBase, Sqoop, Airflow, and more.
* Azure Cloud and Big Data Tools: Working knowledge of Azure components - HDInsight, Databricks, Data Lake, Blob Storage, Data Factory, SQL DB, SQL DWH, Cosmos DB. Hands-on experience with Spark using Scala and PySpark.
* Database Migration: Expertise in migrating SQL databases to Azure Data Lake, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse. Proficient in access control and migration using Azure Data Factory.
* Cloud Computing and Big Data Tools: Proficient in Azure Cloud and Big Data tools - Hadoop, HDFS, MapReduce, Hive, HBase, Spark, Azure Cloud, Amazon EC2, DynamoDB, S3, Kafka, Flume, Avro, Sqoop, PySpark.
* Real-time Data Solutions: Build real-time data pipelines and analytics using Azure components like Data Factory, HDInsight, Azure ML Studio, Stream Analytics, Azure Blob Storage, and Microsoft SQL DB.
* Database Expertise: Work with SQL Server and MySQL databases. Skilled in working with Parquet files, parsing, and validating JSON formats. Hands-on experience in setting up workflows with Apache Airflow and Oozie.
* API Development and Integration: Develop highly scalable and resilient RESTful APIs, ETL solutions, and third-party platform integrations as part of an Enterprise Site platform.
* IDE and Version Control: Proficient use of IDEs like PyCharm, IntelliJ, and version control systems SVN and Git.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Technologies** | Kafka, Cassandra, Apache Spark, HBase, Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper |
| **Hadoop Distribution** | Cloudera CDH, Horton Works HDP |
| **Programming Languages** | SQL, Python, PySpark, Scala, Shell Scripting, Regular Expressions |
| **Spark components** | RDD, Spark SQL (Data Frames and Datasets), Spark Streaming |
| **Cloud Infrastructure** | Azure, AWS, GCP |
| **Databases** | Oracle, Teradata, MySQL, SQL Server, NoSQL Databases (HBase, MongoDB) |
| **Version Control** | Git |
| **Build Tools** | Maven, SBT |
| **Containerization Tools** | Kubernetes, Docker |
| **Reporting Tools** | Power BI, Tableau |

**PROFESSIONAL EXPERIENCE**

**Client:** **BMO- ON**

**Role: Data Enginner Dec 2022- Till**

**Responsibilities:**

**Initiation Phase:**

* Contributed to the analysis, design, and development phases of the Software Development Lifecycle (SDLC). Proficient in Agile practices, participated in sprint planning, scrum calls, and retrospectives. Managed projects through JIRA and version control with GitHub.

**Data Ingestion:**

* Orchestrated messaging queues using RabbitMQ for seamless data flow from HDFS.
* Harnessed Kafka and RabbitMQ to capture data streams within Docker virtualized test and dev environments.
* Designed Docker Containers for data ingestion, leveraging Docker Swarm, Mesos, and Kubernetes.
* Developed robust Databricks solutions for data extraction and transformation.

**Data Processing:**

* Leveraged PySpark distributed computing capabilities for large-scale data processing.
* Engineered custom ETL solutions for batch processing and real-time data ingestion.
* Crafted Azure Databricks (Spark) notebooks for efficient data extraction and loading.
* Conducted comprehensive statistical analysis using SQL, Python, Scala, R Programming, and Excel.

**Data Transformation:**

* Engineered custom ETL solutions for data transformation using PySpark and Shell Scripting.
* Crafted Azure Databricks (Spark) notebooks for data movement between storage accounts.
* Transitioned log storage from Cassandra to Azure SQL Data Warehouse for enhanced query performance.
* Developed and optimized Spark jobs using Python and Spark SQL for data transformation.
* Developed and maintained data processing pipelines using Python, automating the extraction, transformation, and loading (ETL) of large and complex datasets.
* Created custom Python scripts to perform data manipulation, cleansing, and enrichment, ensuring data accuracy and consistency for analytical purposes.
* Implemented modular and reusable Python codebase, promoting code efficiency, maintainability, and collaboration across the data engineering team.

**Data Storage:**

* Designed and implemented ETL solutions in Databricks, adhering to bronze, silver, and gold layer architecture.
* Seamlessly integrated on-premises data sources with cloud platforms using Azure Data Factory.
* Transferred metadata into Hive for migration and complex transformations.
* Accelerated data processing using Spark, Hive, and Sqoop for data storage.

**Data Analysis and Reporting:**

* Derived real-time insights and reports using Spark Scala functions.
* Enhanced query performance and data processing efficiency.
* Managed and delivered data for analytics and Business Intelligence needs using Azure Synapse.
* Conducted comprehensive statistical analysis for data-driven insights.

**Data Integration:**

* Seamlessly integrated on-premises data sources with cloud platforms through Azure Data Factory.
* Streamlined data ingestion from varied sources through Azure Data Factory configurations.
* Designed and implemented DAGs within Apache Airflow to schedule ETL jobs.
* Configured Spark streaming for real-time data reception from Apache Flume.

**Data Security and Management:**

* Bolstered security by integrating Azure DevOps, VSTS, Active Directory, and Apache Ranger for CI/CD and authentication mechanisms.
* Ensured optimal performance and scalability in cloud implementation strategies.
* Managed and delivered data for analytics and Business Intelligence needs using Azure Synapse.
* Set and enforced CI/CD standards and best practices for data processing pipelines.

**CI/CD and DevOps:**

* Led Agile delivery and integrated SAFe and DevOps frameworks.
* Architected, developed, and maintained CI/CD pipelines within Azure DevOps.
* Utilized PowerShell scripting, Bash, YAML, JSON, GIT, Rest API, and Azure Resource Management (ARM) templates for CI/CD.
* Automated Azure Databricks jobs and constructed SSIS packages for smooth data transfer.

**Collaboration and Mentorship:**

* Collaborated closely with development teams to diagnose issues and debug code.
* Mentored junior engineers on CI/CD best practices and cloud-native architectures.
* Collaborated with cross-functional teams for project success.
* Collaborated with development teams to diagnose and resolve issues.

**Environment:** Hadoop, Scala, Spark, Hive, Sqoop, HBase, Flume, Ambari, Scala, MS SQL, MySQL, SSIS, Snowflake, MongoDB, Git, Data Storage Explorer, Python, Azure (Data Storage Explorer, ADF, AKS, Blob Storage), RabbitMQ, Docker.

**Client: Loyalty One - ON**

**Role: ETL Developer Jan 2021 – Nov 2022**

**Responsibilities:**

**Initiation Phase:**

* Contributed to the analysis, design, and development phases of the Software Development Lifecycle (SDLC). Proficient in Agile practices, participated in sprint planning, scrum calls, and retrospectives. Managed projects through JIRA and version control with GitHub.

**Data Collection and ETL:**

* Created and managed Amazon EC2 instances, diagnosing common issues and maintaining the health of EC2 instances and other AWS services.
* Engineered RESTful APIs using Python with Flask and Django frameworks, orchestrating integration across diverse data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.

**Data Warehousing:**

* Designed, deployed, and managed high-availability and scalable database solutions using AWS RDS, ensuring optimal performance and reliability for critical applications.
* Developed and optimized complex stored procedures in AWS Redshift to streamline data transformations, enhance query performance, and enable efficient data extraction, transformation, and loading (ETL) processes.
* Designed and implemented data pipelines utilizing AWS Redshift stored procedures to facilitate seamless data movement and transformation between various source and target systems, ensuring accurate and timely delivery of insights to stakeholders.
* Collaborated with cross-functional teams to gather business requirements and translate them into effective stored procedures within AWS Redshift, enabling efficient data processing and analysis for informed decision-making.
* Led the migration of on-premises databases to AWS RDS, minimizing downtime and disruptions while optimizing database performance in the cloud environment.
* Developed and maintained data processing pipelines using Python, automating the extraction, transformation, and loading (ETL) of large and complex datasets.
* Created custom Python scripts to perform data manipulation, cleansing, and enrichment, ensuring data accuracy and consistency for analytical purposes.
* Implemented modular and reusable Python codebase, promoting code efficiency, maintainability, and collaboration across the data engineering team.
* Designed and optimized complex SQL queries for data extraction, aggregation, and reporting, ensuring efficient data retrieval from various database systems.
* Collaborated with cross-functional teams to define data access patterns, data transformation rules, and business logic through SQL stored procedures and functions.
* Engineered and managed data pipelines with AWS Glue to extract, transform, and load data into Snowflake.
* Designed and optimized intricate SQL queries and stored procedures to support data-intensive applications with large databases.
* Crafted and sustained robust data models and schemas, leveraging database management systems such as Postgres, MySQL, and Oracle.
* Implemented Python API for Apache Spark to facilitate large-scale distributed data processing.
* Utilized PySpark for data manipulation, executing operations such as filtering, aggregations, and joins on distributed datasets.
* Developed and optimized PySpark jobs to effectively handle big data workloads.
* Integrated PySpark with complementary big data technologies like Hadoop and Hive to construct comprehensive end-to-end data processing pipelines.

**Cloud Infrastructure and Deployment:**

* Designed and deployed multi-tier applications, harnessing the full spectrum of AWS services (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) with a focus on high availability, fault tolerance, and auto-scaling within AWS Cloud Formation.

**Application Development:**

* Developed and maintained web applications utilizing Python frameworks, including Django and Flask, following Model-View-Controller (MVC) architecture for scalability and maintainability.
* Designed and implemented RESTful APIs using Django Rest Framework (DRF) and Flask-RESTful, ensuring seamless integration with external systems.

**Cloud Management and Support:**

* Provided support for cloud instances on AWS, adeptly managing Linux and Windows instances, Elastic IP, Security Groups, and Virtual Private Cloud.
* Orchestrated data pipeline development utilizing Spark, Hive, Pig, Python, Impala, and HBase to enable efficient customer data ingestion.

**Data Quality and Automation:**

* Proficiently profiled structured, unstructured, and semi-structured data across diverse sources, implementing data quality metrics and pattern identification through SQL queries and Python scripts.
* Automated backups of ephemeral data stores to S3 buckets and EBS, generating nightly AMIs for mission-critical production server backups with AWS CLI.
* Installed and configured automated tools like Puppet and participated in deployment processes on multiple platforms using Chef and Puppet.
* Contributed to OpenShift Pass product architecture, focusing on creating OpenShift namespaces for seamless migration of on-premises applications to the cloud.
* Utilized Docker for server virtualization in testing and development environments, automating configuration tasks through Docker containers.

**DevOps and CI/CD:**

* Worked with Amazon AWS/EC2 and Google's Kubernetes-based Docker cluster management environment.
* Orchestrated Jenkins job creation, load distribution on Jenkins servers, and parallel build execution through Jenkins node configuration.
* Extensively managed Jenkins CI/CD pipelines, overseeing end-to-end automation for artifact build, test, and delivery, and effectively troubleshooting issues during the build process.
* Managed Jenkins artifacts within the Nexus repository, incorporating versioning with timestamps, and deployed artifacts to AWS servers with Ansible and Jenkins.
* Established continuous integration systems with Ant, Jenkins, and Puppet, driving comprehensive automation and swift, error-free deployments.
* Installed and administered the GIT source code tool, designed effective branching strategies, and ensured application reliability.
* Managed Nexus Repository Manager, overseeing repository management and artifact release processes.
* Proficient in on-premises and AWS-based network, application, and server monitoring tools such as Nagios, Splunk, AppDynamics, and CloudWatch.
* Configured and customized JIRA as a defect-tracking system, implementing workflows and plugins to optimize bug and issue tracking.

**Environment**: AWS, Ansible, ANT, MAVEN, Jenkins, Bamboo, Splunk, Confluence, Bitbucket, GIT, JIRA, Python, SSH, Shell Scripting, Docker, JSON, JAVA/J2EE, Kubernetes, Nagios, Red Hat Enterprise Linux, Terraform, Kibana, Fargate.

**Client: Flexiti Financials - Toronto**

**Role: MSBI Developer Sep 2019 – Dec 2020**

**Responsibilities:**

* **Initiation Phase:**
* Developed RESTful APIs using Python with Flask and Django frameworks, integrating diverse data sources.
* Leveraged Apache Spark with Python for Big Data Analytics and Machine Learning applications.
* Designed and deployed SSIS packages for data loading and transformation within Azure databases.
* Configured and managed SSIS Integration Runtime for executing packages in Azure.
* Proficient in version control systems like Git for managing Data Lakehouse pipelines.
* **Data Collection and ETL:**
* Automated monitoring tasks using AWS CloudWatch and Redshift Query Performance Insights.
* Conducted performance tuning and optimization of existing AWS Redshift stored procedures.
* Implemented automated backups and snapshots for AWS RDS instances.
* Leveraged Python libraries for advanced data analysis and integration tasks.
* Designed and developed RESTful APIs for data integration.
* Collaborated with data scientists and analysts to deploy machine learning models using Python.
* Proficient in data validation and cleansing procedures using SQL.
* Proficiently profiled structured, unstructured, and semi-structured data.
* **Data Warehousing:**
* Experience working with Microsoft Azure Cloud services. Executed ETL operations using Azure Data Factory, Spark SQL, and T-SQL. Expertise in data migration to various Azure services.
* Orchestrated data extraction, transformation, and loading across Azure services.
* Automated script execution through Apache Airflow and shell scripting. Constructed pipelines in Azure Data Factory.
* Led Data Migration initiatives employing SQL, SQL Azure, and Azure Data Factory.
* Proficiently profiled structured, unstructured, and semi-structured data from various sources.
* Employed PowerShell and UNIX scripts for various tasks. Leveraged Sqoop for data transfer between RDBMS and HDFS.
* Installed and configured Apache Airflow for data workflows. Employed MongoDB for data storage.
* Developed RESTful APIs, ETL solutions, and platform integrations.
* Proficiently used IDEs and version control systems.

**Environment**: Hadoop, Scala, Spark, Hive, Sqoop, HBase, Flume, Ambari, Scala, MS SQL, MySQL, SSIS, Snowflake, MongoDB, Git, Data Storage Explorer, Python, Azure (Data Storage Explorer, ADF, AKS, Blob Storage), RabbitMQ, Docker.

**Client: HDFC Bank - India**

**Role: Data Engineer May 2017 – July 2019**

**Responsibilities:**

* Contributed to the analysis, design, and development phases of the Software Development Lifecycle (SDLC) in an agile environment.
* Leveraged PySpark extensively for transformations and data processing on Azure HDInsight.
* Developed Azure ML Studio pipelines, integrating Python for machine learning algorithms.
* Orchestrated data loading from diverse sources to Azure Data Lake using Azure Data Factory.
* Implemented business rules for contact deduplication using Spark transformations with PySpark.
* Designed and deployed multi-tier applications on AWS services, focusing on high availability and fault tolerance.
* Developed Graph Database nodes and relations using Cypher language.
* Built microservices with AWS Lambda for third-party vendor API calls.
* Provided support for cloud instances on AWS, managing resources and security.
* Engineered data pipelines using Spark, Hive, Pig, Python, Impala, and HBase for customer data.
* Configured AWS services including EC2, S3, Elastic Load Balancing, and security measures.
* Automated backups and data management tasks using AWS CLI.
* Created and managed Docker containers for application deployment.
* Proficient in container clustering with Docker, Swan, Mesos, and Kubernetes.
* Established and managed Jenkins CI/CD pipelines for automation.
* Managed artifacts within Nexus repository and deployed them using Ansible and Jenkins.
* Utilized monitoring tools like Nagios, Splunk, AppDynamics, and CloudWatch.
* Set up JIRA as a defect tracking system for bug and issue tracking.

**Environment**: Spark, Spark-Streaming, Spark SQL, HDFS, Hive, Apache Kafka, Sqoop, Java, Scala, Linux, Azure SQL Database, Azure ML Studio, Jenkins, Flask Framework, IntelliJ PyCharm, Eclipse, Git, Azure Data Factory, Tableau, MySQL, Postman, Agile Methodologies, AWS Lambda, Azure Cloud, Docker.