NITHYA SANTHOSHINI CHANDA

+1 (832)-510-9992 • chandanithya@gmail.com

Summary

- Over 7 years of hands-on experience in computer science, specializing in Big Data Hadoop development and Java/J2EE technologies.
- Proficient in deploying instances, handling EC2 provisioning, setting up security groups, and establishing Hadoop ecosystems seamlessly within the AWS framework.
- Experienced in distributed computing architectures, migrating raw data to Amazon Cloud (S3), and executing refined data processing.
- Skilled in seamlessly integrating data from various sources, utilizing Snowflake Cloud Data Warehouse and AWS S3 buckets, with expertise in managing nested JSON formatted data.
- Demonstrated proficiency in AWS Lambda and EC2 instances provisioning, with a focus on establishing robust security groups and overseeing the administration of Amazon VPCs.
- Crafted application environments on AWS using EC2 instances, employing Docker, Bash, and Terraform for efficient deployment and system management.
- Utilized tools such as Jupyter Notebooks, Apache Spark, Python, R, SAS, and Snowflake Data Warehouse on AWS for comprehensive data analysis, supporting academic research data management processes.
- Extensive experience in designing and maintaining data pipelines on the Azure Analytics platform using Azure Databricks, PySpark, Python, Pandas, and NumPy libraries.
- Proficient in designing and developing visually compelling dashboards and reports using Tableau Visualizations, addressing diverse end-user requirements.
- Accomplished in establishing Enterprise Data Lakes that facilitate analytics, storage, data processing, and reporting for large and dynamically evolving datasets.
- Proficient in researching and analyzing data sources, working with OLAP cubes, designing data models and schemas, and developing SQL solutions.
- Demonstrated ability to create databases on RDS, load data from AWS S3 to RDS SQL Server, and develop APIs with associated security groups for external access.
- Effectively transformed code from Scala to PySpark within the Data Harmonization Framework (DHF) and oversaw the transition from DHF 1.0 to DHF 2.1.
- Led collaborative project management efforts, overseeing the deployment of instances, provisioning EC2, and configuring Hadoop ecosystems on Cloudera in AWS.
- Executed seamless data integration from multiple sources using Snowflake Cloud Data Warehouse and AWS S3, ensuring data accuracy and consistency.
- Administered Amazon VPCs, implemented security groups, and provisioned EC2 instances on AWS environments to ensure secure and efficient data solutions.
- Collaborated with data scientists to integrate machine learning models into production systems, enhancing the intelligence and capabilities of data solutions.
- Proactively seek opportunities for continuous improvement, staying updated on industry trends, and incorporating relevant advancements into data engineering practices.

- Applied strong problem-solving skills to address data-related challenges, ensuring the timely resolution of issues and the continuity of data processes.
- Certified Cloudera Spark and Hadoop Developer & Oracle Java SE 8 Programmer I & II, with extensive experience in Unit Testing using JUnit, MRUnit, and pytest.
- In-depth understanding of Hadoop Architecture, workload management, schedulers, scalability, and various components, such as HDFS, MapReduce, and YARN.

Technical Skills

Programming Languages
Java 6+, Scala 2.10+, Python, C, R, PHP

Relational & NoSQL Databases
MySQL, Oracle, PostgreSQL, MongoDB

Web Technologies
SOAP, REST, JSP 2.0, JavaScript, Servlet PHP, HTML5, CSS

Big Data Technologies
Apache Spark, Apache Hadoop, Hue, Map Reduce, Apache Hive, Apache Sqoop, Apache Kafka, Apache Flume, Apache Airflow, Apache Zookeeper, HDFS, Cassandra.

Operating Systems
Windows (98/XP/NT/2000/2003/2008), UNIX, LINUX, Ubuntu (12.x, 13.x, 14.x, 15.x, 16.x), RHEL (4.x, 5.x, 6.x, 7.x), SOLARIS, Centos.

AWS Services
EC-2, ELB, VPC, RDS, IAM, Cloud Formation, S3, CloudWatch, CloudTrail, SNS, SQS, SWF, EBS.

CLOUD and CLOUD Services
AWS, Azure, GCP, Splunk, EC2, Load Balancer, EC2, S3, Cloud Formation

Application web server
Apache Tomcat, JBOSS, IBM Web sphere

CI Tools, CM Tools, Monitoring Tools
Tableau 9.x/10.x, Python (Matplotlib, Pandas, and NumPy), Power BI, MicroStrategy, CHEF, PUPPET, Graphana, Kibana, Splunk, Service Now, OPAL, ACSS

Testing Tools
Selenium, Silk International, Automatics (Developed on Selenium Framework), QTP, Soap UI, QA Load

Work Experience

Data Engineer/AWS Developer, 12/2021 to Current
Client: Amazon.com Services – Seattle, WA
Responsibilities:

- Designed and implemented Enterprise Data Lake supporting various use cases, including storage, processing, analytics, and reporting of large, rapidly changing data, utilizing a variety of AWS services.
- Established Kinesis Data streams, Kinesis Data Firehose, and Kinesis Data Analytics to capture and process streaming data, subsequently outputting it into S3, Dynamo DB, and Redshift for storage and analysis.
- Developed ETL integration patterns using Python on Spark for efficient data processing.
- Created a framework for converting existing PowerCenter mappings to PySpark (Python and Spark) Jobs.
- Optimized volumes and EC2 instances, creating multiple Virtual Private (VP) instances, and managed IAM to establish new accounts, roles, and groups.
- Implemented SparkRDD transformations to map business analysis and apply actions on top of transformations.
- Built S3 buckets, managed policies for S3 buckets, and utilized S3 bucket and Glacier for storage and backup on AWS.
- Enhanced and tuned Redshift environment, achieving query performance improvements of up to 100x for Tableau and SAS Visual Analytics.

- Integrated services such as GitHub, AWS Code Pipeline, Jenkins, and AWS Elastic Beanstalk to create a streamlined deployment pipeline.
- Established monitors, alarms, and notifications for EC2 hosts using CloudWatch and implemented new project builds framework using Jenkins as a build tool.
- Designed and implemented ETL jobs for data processing using various AWS services, including Glue Crawlers for automatic schema discovery, S3 buckets for storage, and Athena for data transformation, resulting in a 90% reduction in processing time.
- Extracted data from multiple source systems (S3, Redshift, RDS) and created tables/databases in Glue Catalog by utilizing Glue Crawlers.
- Wrote code to optimize the performance of AWS services used by application teams and provided code-level application security for clients (IAM roles, credentials, encryption, etc.).
- Designed and created a serverless AWS Glue ETL pipeline using Scala to transform and load data from S3 buckets into Amazon Redshift, utilizing CloudWatch events to trigger scheduled Lambda functions for automated ETL processes.
- Utilized Athena for data analysis and QuickSight to generate business intelligence reports from processed data obtained through Glue ETL Jobs.
- Converted and delivered 300M+ CSV records into 3.5M Parquet files using AWS Glue Jobs and Python scripts.
- Redesigned architecture and changed the triggering of long-running ETL jobs from Lambda functions to Workflows.
- Designed dashboards for the Code Quality team to gain insights into developer productivity using Metabase and Python.
- Developed a generic framework using Scala and bash scripts for data ingestion, processing different data formats (CSV, JSON, PARQUET), and feeding the data into Hive tables.
- Implemented Spark and Python scripts to fetch build data and test results from Jenkins, pushing them into AWS S3 and Redshift for faster reporting.
- Developed and maintained Directed Acyclic Graphs (DAGs) in Apache Airflow, defining the sequence and dependencies of tasks for efficient data pipeline execution.
- Created custom operators in Python for specialized tasks, extending the functionality of Apache Airflow to seamlessly integrate with unique data processing requirements.
- Designed and implemented an Enterprise Data Lake supporting various use cases, including the storage, processing, analytics, and reporting of large, rapidly changing data, utilizing a variety of AWS services.
- Established Kinesis Data streams, Kinesis Data Firehose, and Kinesis Data Analytics to capture and process streaming data, subsequently outputting it into S3, Dynamo DB, and Redshift for storage and analysis.
- Developed ETL integration patterns using Python on Spark for efficient data processing.
- Created a framework for converting existing PowerCenter mappings to PySpark (Python and Spark) Jobs.
- Optimized volumes and EC2 instances, creating multiple Virtual Private (VP) instances, and managed IAM to establish new accounts, roles, and groups.
- Implemented SparkRDD transformations to map business analysis and apply actions on top of transformations.
- Built S3 buckets, managed policies for S3 buckets, and utilized S3 bucket and Glacier for storage and backup on AWS.
- Enhanced and tuned the Redshift environment, achieving query performance improvements of up to 100x for Tableau and SAS Visual Analytics.
- Integrated services such as GitHub, AWS Code Pipeline, Jenkins, and AWS Elastic Beanstalk to create a streamlined deployment pipeline.
- Established monitors, alarms, and notifications for EC2 hosts using CloudWatch and implemented a new project builds framework using Jenkins as a build tool.

- Designed and implemented ETL jobs for data processing using various AWS services, including Glue Crawlers for automatic schema discovery, S3 buckets for storage, and Athena for data transformation, resulting in a 90% reduction in processing time.
- Extracted data from multiple source systems (S3, Redshift, RDS) and created tables/databases in Glue Catalog by utilizing Glue Crawlers.
- Wrote code to optimize the performance of AWS services used by application teams and provided code-level application security for clients (IAM roles, credentials, encryption, etc.).
- Designed and created a serverless AWS Glue ETL pipeline using Scala to transform and load data from S3 buckets into Amazon Redshift, utilizing CloudWatch events to trigger scheduled Lambda functions for automated ETL processes.
- Utilized Athena for data analysis and QuickSight to generate business intelligence reports from processed data obtained through Glue ETL Jobs.
- Converted and delivered 300M+ CSV records into 3.5M Parquet files using AWS Glue Jobs and Python scripts.
- Redesigned architecture and changed the triggering of long-running ETL jobs from Lambda functions to Workflows.
- Designed dashboards for the Code Quality team to gain insights into developer productivity using Metabase and Python.
- Developed a generic framework using Scala and bash scripts for data ingestion, processing different data formats (CSV, JSON, PARQUET), and feeding the data into Hive tables.
- Implemented Spark and Python scripts to fetch build data and test results from Jenkins, pushing them into AWS S3 and Redshift for faster reporting.
- Developed and maintained Directed Acyclic Graphs (DAGs) in Apache Airflow, defining the sequence and dependencies of tasks for efficient data pipeline execution.
- Created custom operators in Python for specialized tasks, extending the functionality of Apache Airflow to seamlessly integrate with unique data processing requirements.

Environment:

Python, PySpark, Spark, Scala, AWS Cloud Data, Amazon EMR, Amazon S3 ETL(Informatica), Lambda, Glue Crawlers, CloudWatch, Workflows, Kafka, Hadoop (HDP 3.0), Tableau, Cloudera, Postgres, Teradata, Terraform, Splunk, AWS Redshift, Airflow

Data Engineer, 01/2020 to 11/2021

Client: World Fuel Services – Houston, TX

Responsibilities:

- Analyzed results using Spark SQL queries in conjunction with Hive queries to derive meaningful insights from large datasets.
- Implemented Spark with Scala and Spark SQL to optimize data processing performance, enabling faster testing and processing of data.
- Utilized Spark with Scala, leveraging DataFrames and Spark SQL API to expedite the processing of data.
- Designed and created HBase tables, employing column families to efficiently store and manage project-related data.
- Employed Maven for deployments and managed structured, semi-structured (e.g., XML), and unstructured data in the data lake with a streamlined data flow.
- Established connectors to facilitate data loading to and from Snowflake, utilizing analysis to discover and align with business goals.
- Diagnosed issues and provided business intelligence for Hadoop through the effective utilization of Splunk.
- Visualized data collected from diverse sources using Splunk, enhancing the understanding of data patterns and trends.
- Demonstrated experience in developing/consuming Web Services (REST, SOAP, JSON) and APIs within Service-oriented architectures.

- Constructed Chef-based CI/CD solutions, improving developer productivity and enabling rapid deployments.
- Troubleshooted Linux network and security-related issues, utilizing tools like IP tables and firewall, and captured packets for analysis.
- Extensively worked on importing metadata into Hive, migrating existing tables and applications onto Hive for enhanced data management.
- Leveraged Tableau to create compelling stories and interactive dashboards, offering detailed insights from the analyzed data.
- Wrote various Spark transformations in Scala to cleanse, validate, and summarize user behavioral data.
- Developed a production planning dashboard in Spotfire by creating information links to seamlessly integrate data sources.
- Identified anomalies and new data sources for analysis, enhancing data collection and analysis flow into the data lake.
- Conducted data wrangling using the Pandas library for cleaning, transforming, and reshaping data.
- Utilized R and Python for Exploratory Data Analysis, comparing and assessing the effectiveness of data in the data lake.
- Generated statistics from analyzed data, producing insightful reports for informed decision-making.
- Parsed unstructured data into a semi-structured format by implementing complex algorithms in Spark.
- Loaded transformed data into Hive tables, enabling analytics and insights using Hive queries.
- Implemented partitioning on Hive data to enhance performance in data processing and flow.
- Analyzed data by executing Hive queries (Hive QL) to gain insights into customer behavior.
- Worked on various performance optimizations in Hive, including the use of distributed cache for small datasets and partitioning and bucketing strategies.

Environment:

AWS (IAM, EC2, S3, EBS, Glacier, ELB, Cfn, CloudWatch, Cloud Trail, Hadoop (HDP), Informatica, HDFS, Spark SQL, Git, Kafka, Hive, Java, Scala, HBase, Maven, UNIX Shell Scripting, AWS Cloud Data, Terraform, Python, SQL, Tableau, Pig, Teradata, Splunk

Data Analyst, 04/2017 to 12/2019

Client: Syncor Solutions – Hyderabad, India

Responsibilities:

- Developed and maintained SQL scripts for data extraction, transformation, and loading (ETL) processes.
- Automated repetitive tasks using Python scripts, improving overall efficiency in daily workflows.
- Established and documented coding standards to ensure consistency and code quality within the development team.
- Conducted code reviews and provided constructive feedback to enhance the overall codebase.
- Implemented version control using Git, ensuring collaborative and streamlined development processes.
- Integrated unit testing into the development workflow, enhancing code reliability and reducing bugs.
- Collaborated with cross-functional teams to gather and analyze requirements, ensuring alignment with project goals.
- Contributed to the creation of RESTful APIs, facilitating seamless communication between different components of the application.
- Utilized Docker for containerization, enabling consistent deployment across various environments.
- Implemented monitoring and logging solutions to track system performance and troubleshoot issues effectively.
- Collaborated with DevOps teams to optimize continuous integration and continuous deployment (CI/CD) pipelines.
- Actively participated in Agile development methodologies, including Scrum meetings and sprint planning.
- Developed and maintained technical documentation, ensuring comprehensive knowledge transfer within the team.
- Engaged in ongoing professional development, staying updated on industry best practices and emerging technologies.
- Conducted training sessions for team members on best coding practices and the use of new tools and technologies.
- Established error-handling mechanisms in scripts and applications, enhancing data processing reliability.

- Assisted in the design and implementation of a robust security framework for web applications.
- Implemented caching strategies to optimize data retrieval and processing times.
- Collaborated with UX/UI designers to implement visually appealing and user-friendly interfaces.
- Contributed to the creation of data models, ensuring database structures aligned with application requirements.
- Conducted performance testing to identify and address bottlenecks in applications.
- Collaborated with the Quality Assurance (QA) team to address and resolve reported issues promptly.
- Collaborated with the user support team to address and resolve customer-reported issues.
- Implemented A/B testing strategies to optimize user experience and feature effectiveness.
- Actively participated in technology meetups and conferences to stay updated on industry trends.
- Engaged in cross-functional retrospectives to identify areas for process improvement.
- Implemented data anonymization techniques for compliance with privacy regulations.

Environment:

Python, Django, Mongo DB, Selenium, Pandas, Java, J Query, Zookeeper, bootstrap, My SQL, Linux, Ajax, JavaScript, Apache, JIRA, Cassandra, HTML5 and CSS, Angular JS, Backbone JS

Education

Master of Science**:** Computer Science

Bachelor of Technology: Electronics and communication Engineering