|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | |  | |
|  | **Phanindra S**  **DATA ENGINEER/PYTHON ENGINEER** | | | |  |
| **+1 848-777-6269** | | | **Phanindras007@gmail.com** | | |

**PROFESSIONAL SUMMARY:**

* Over 8+ years of IT experience in Analysis, Design, Development, Implementation, Testing and Production Support (SDLC) of Data Warehousing projects.
* Strong background in diverse programming languages, including Python, Scala, and SQL, and extensive expertise in PL/SQL. Proficient in various Big Data technologies such as Apache Spark, PySpark, Hadoop, Hive, and HBase, focusing on optimizing data processing and analysis. Skilled in utilizing Apache Kafka, Apache Airflow, and Spark SQL for efficient data streaming and workflow management.
* Adept at working with database management systems, including MySQL, Amazon DynamoDB, and PostgreSQL, and experienced in data warehouse solutions like Snowflake and RedShift.
* Extensive cloud platform experience with AWS, including EC2, S3, EMR, and RDS, and expertise in serverless computing using AWS Lambda. Hands-on experience setting up and managing cloud-based infrastructure, utilizing services like CloudWatch, SQS, SNS, and API Gateway.
* Proficient in developing robust web applications using Django and Flask frameworks. Highly skilled in version control and collaboration tools like Git/GitHub and Bitbucket, ensuring seamless team collaboration and code management.
* Experienced in containerization and orchestration technologies like Docker and Kubernetes, enabling efficient deployment and scaling of applications. Proficient in ETL processes using AWS Glue, Apache Nifi, and Talend, ensuring smooth data extraction, transformation, and loading.
* Strong background in automation and CI/CD pipelines, utilizing tools like Jenkins, Ansible, and Puppet to enhance development and deployment workflows. Familiarity with infrastructure management tools like Terraform and configuration management using Apache Ambari.
* Having experience encompassing in Machine Learning, Data Mining with large datasets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Data Visualization
* Proficient in API documentation using Swagger and visualization tools like Tableau and Power BI. Experienced handling diverse data formats and APIs, including JSON and RESTful APIs, ensuring seamless data integration across systems.
* Skilled in operating systems including Linux (Ubuntu, CentOS) and Windows Server, ensuring cross-platform compatibility and smooth system operations. A detail-oriented and dedicated professional with a proven track record of delivering high-quality data engineering solutions.
* Ability to work through a solution end-to-end, meet project deadlines and deliver solutions in a supportable manner.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Programming & Scripting** | Python, Scala, SAS, R, SQL, PL/SQL, MATLAB, HiveQL, PowerShell and BASH and Shell Scripting. |
| **Cloud Technologies and Services** | Amazon AWS - EMR, EC2, EBS, RDS, S3, Athena, Glue, Elasticsearch, Lambda, SQS, SNS, CloudWatch, API Gateway, DynamoDB, Redshift. |
| **Big Data/ Hadoop Ecosystem** | HDFS, YARN, MapReduce, Sqoop, Hive, Oozie, Pig, Spark, Zookeeper, Cloudera Manager, Kafka, Flume, Nifi, Connect, Airflow, Stream Sets, Kafka connect |
| **Python Libraries** | Scikit-learn, SciPy, Pandas, AWS S3, AWS SFTP and NumPy.. |
| **Frameworks** | Django, Flask, NodeJS, ReactJS, Angular. |
| **IDE Tools** | Eclipse, PyCharm, Visual Studio Code, Sublime Text, IntelliJ, Jupyter Notebook. |
| **CI/CD Tools & Automation** | Jenkins, Terraform, Puppet, Ansible. |
| **Version Control & Collaboration** | Git/GitHub, Bitbucket, SVN. |
| **Containerization & Orchestration** | Docker, Kubernetes. |
| **ETL Tools** | AWS Glue, Apache Nifi, Talend. |
| **Databases and ORM** | Oracle, MySQL, Teradata, Postgres, Django ORM, SQL Alchemy and AWS Redshift. |
| **NoSQL Database** | HBase, Cassandra, Dynamo DB, Cosmos, Mongo DB. |
| **Data warehouse** | Snowflake, RedShift. |
| **Data Formats & APIs** | JSON, RESTful APIs |
| **Other Tools** | Maven, Ant, REST/SOAP Webservices, PyUnit. |
| **Operating Systems** | Ubuntu, Mac OS-X, CentOS, Windows 10/8/ 7, Linux. |

**PROFESSIONAL EXPERIENCE:**

**CSAA- Remote**

**Python/Data Engineer Dec 2022 - Present**

**Responsibilities:**

* Developed robust ETL processes using PySpark and Apache Spark to handle large-scale data processing tasks efficiently.
* Designed and implemented scalable data pipelines in Python, leveraging technologies like Apache Kafka and Apache Spark for real-time data ingestion and processing.
* Utilized TensorFlow and NumPy for advanced machine learning tasks, enhancing data-driven decision-making processes within the organization.
* Optimized and fine-tuned SQL queries for complex data manipulations on MySQL and Amazon DynamoDB, ensuring high performance and reliability.
* Implemented data warehousing solutions using Amazon Redshift, optimizing queries for large-scale analytics and reporting purposes.
* Architected and maintained distributed systems on AWS, including EC2, S3, EMR, and RDS, ensuring seamless integration and data flow across various components.
* Utilized Docker for containerization, enabling consistent deployment and scaling of data processing applications across different environments.
* Skilled in monitoring servers using Nagios, Data dog, Cloud watch and using EFK Stack Elasticsearch Fluentd Kibana
* Designed and implemented RESTful APIs using Django and Next.js, enabling smooth communication between diverse microservices within the architecture.
* Developed and maintained data pipelines using Apache Airflow's DAGs (Directed Acyclic Graphs)
* Managed and monitered Airflow components like the web server, scheduler, and workers.
* Implemented real-time data streaming solutions using Apache Kafka, ensuring timely processing and analysis of streaming data for actionable insights.
* Automated deployment workflows using Jenkins and Ansible, streamlining the continuous integration and delivery processes for data-related applications.
* Implemented security protocols and access controls for sensitive data on Hadoop, HDFS, and HBase, ensuring compliance with industry standards and regulations.
* Designed and optimized data models using Pandas and NumPy, enabling efficient data analysis and visualization for stakeholders and decision-makers.
* Implemented data partitioning and sharding strategies on Amazon S3, enhancing data retrieval performance and optimizing storage costs.
* Implemented security best practices in Terraform configurations, manage secrets securely, and ensure that infrastructure adheres to compliance standards and policies.
* Automated routine infrastructure tasks and workflows, such as server provisioning, scaling, and resource cleanup. Implement CI/CD pipelines to automate Terraform deployments.
* Utilized Apache Airflow for workflow orchestration, scheduling complex data processing tasks and ensuring data pipelines were executed reliably.
* Implemented CI/CD pipelines using Git/GitHub, allowing collaborative development and version control for data engineering projects.
* Used AWS glue catalog with crawler to get the data from S3 and perform SQL query operations using AWS Athena.
* Written PySpark job in AWS Glue to merge data from multiple tables and in Utilizing Crawler to populate AWS Glue data.
* Used AWS Glue for transformations and AWS Lambda to automate the process.
* Created Lambda functions to run the AWS Glue job based on the AWS S3 events.
* Utilized Terraform for infrastructure as code, enabling the provisioning and management of cloud resources in a version-controlled and automated manner.
* Implemented event-driven architecture using AWS Lambda and SQS, ensuring seamless integration between various services and components within the ecosystem.
* Designed and implemented data caching mechanisms using Redis, improving query performance and reducing latency for frequently accessed datasets.
* Implemented monitoring and alerting systems using AWS CloudWatch and SNS, proactively identifying and addressing potential issues in data pipelines.
* Collaborated with cross-functional teams using Agile methodology and JIRA, ensuring effective communication and alignment of data engineering tasks with business objectives.

**Environment:** Python, JavaScript, Django, Next.js, NumPy, Pandas, PySpark, Tensor Flow, MySQL, Amazon DynamoDB, Hadoop, HDFS, Hive, MapReduce, HBase, Amazon S3, Ansible, AWS, Ec2, S3, EMR, SQS, SNS, RDS, AWS Glue,Lambda, API Gateway, YARN, Sqoop, Pig, Ambari, Nifi, Apache Kafka, Apache Airflow, Apache Spark, Docker, Redis, Git/GitHub, Swagger, Terraform, Jenkins, Redshift, REST APIs, Athena, Agile methodology, JIRA.

**Walmart- Bentonville, AR**

**Data Engineer June 2021 – Dec 2022**

**Responsibilities:**

* Developed and maintained Python scripts for data extraction, transformation, and loading (ETL) processes.
* Managed SQL and PL/SQL queries for efficient data retrieval and manipulation in various databases.
* Designed and implemented RESTful APIs for seamless data communication between systems.
* Utilized Flask framework to create scalable web applications for data visualization.
* Configured and optimized AWS services for large-scale data storage and processing, including EC2, EBS, and S3.
* Implemented message queues using SQS and SNS for real-time data streaming and event-driven architecture.
* Automated tasks using AWS Lambda, enhancing system efficiency and reducing manual intervention.
* Orchestrated complex data workflows using Apache Airflow, ensuring timely execution and monitoring.
* Setting up and maintaining the Airflow metadata database (e.g., PostgreSQL).
* Managed and maintained stateful data processing components within Kubernetes, ensuring data integrity and durability.
* Set up, maintained, and optimizeed CI/CD pipelines using tools like Jenkins, Travis CI, or AWS CodePipeline for automated build, test, and deployment of applications.
* Architect and configured AWS environments, including Virtual Private Clouds (VPCs), EC2 instances, auto-scaling groups, databases (RDS), and more.
* Continuously assess and optimize the costs associated with running real-time data processing workloads in Kubernetes.
* Written code or defined tasks using Airflow operators and Python scripts to perform data transformations.
* Managed and optimized CloudWatch alarms for system performance and resource utilization tracking.
* Implemented infrastructure as code using Cloud Formation templates for AWS resource provisioning.
* Worked with Snowflake and DynamoDB for NoSQL and cloud-based data warehousing solutions.
* Developed and maintained PostgreSQL databases, ensuring data integrity and high availability.
* Implemented version control using Bitbucket, ensuring collaborative development and code review.
* Automated deployment pipelines using Jenkins for continuous integration and continuous deployment (CI/CD).
* Containerized applications using Docker, enabling seamless deployment across various environments.
* Managed infrastructure configurations using Puppet, ensuring consistency and security across servers.
* Utilized Apache Spark and Hadoop for processing large datasets in distributed computing environments.
* Ingested and organized data from various sources into Amazon S3, ensuring data is structured and stored in an optimal format for querying with Athena.
* Managed AWS EC2 instances utilizing Auto Scaling, Elastic Load Balancing and Glacier for our QA and UAT environments as well as infrastructure servers for GIT and Chef.
* Developed ETL (Extract, Transform, Load) processes using tools like AWS Glue or custom scripts to transform and prepare data for analysis in Athena.
* Developed shell scripts for automating routine tasks and system maintenance on Linux servers.
* Managed Terraform state files to track the current state of infrastructure resources, allowing for updates and modifications while maintaining consistency.
* Developed and maintained data ingestion pipelines to bring data from various sources into Elasticsearch, using tools like Logstash, Beats, or custom scripts.
* Collaborated with other teams, such as Elasticsearch administrators, developers, and data analysts, to ensure data infrastructure aligns with overall objectives.
* Created reusable Terraform modules and templates to standardize and simplify infrastructure provisioning. These modules can be used across various projects and teams.
* Created multiple Glue ETL jobs in Glue Studio and then processed the data by using different transformations and then loaded into S3,Redshift and RDS.
* Created multiple Recipes in Glue Data Brew and then used in various Glue ETL Jobs.
* Design and Develop ETL Processes in AWS Glue to migrate data from external sources like S3, Parquet/Text Files into AWS Redshift.
* Used AWS EMR to transform and move large amounts of data into and out of AWS S3.
* Implemented and optimized data processing workflows using Apache Spark on AWS EMR clusters.
* Collaborated with cross-functional teams using Jira, facilitating efficient project management and issue tracking.

**Environment:** Python, SQL, PL/SQL, Flask, AWS, EC2, EBS, S3, SQS, SNS, Lambda, EMR, CloudWatch, Cloud Formation, Terraform, Snowflake, DynamoDB, PostgreSQL, Bitbucket, Jenkins, Docker, Puppet, Apache Spark, Athena, Hadoop, Apache Airflow, AWS EMR, Glue, Shell Scripting, RESTful APIs, JSON, Docker, Kubernetes, Jira, Linux.

**FM Global - Norwood, MA**

**Data Engineer / DevOps Engineer Sep 2020 - May 2021**

**Responsibilities:**

* Worked in all the phases of SDLC and agile development methodologies including SCRUM, involved in daily SCRUM meetings to keep track of the on-going project status and issues using JIRA.
* Developed and maintained data pipelines using Python, Scala, and Apache Spark Core for efficient data processing.
* Managed AWS resources, including EC2 instances, S3 buckets, RDS databases, and EBS volumes to ensure data availability and reliability.
* Implemented Elastic Load Balancers (ELB) to distribute traffic for high availability of applications.
* Designed and orchestrated ETL workflows using AWS Glue Pipelines and automated data discovery with Glue Crawlers.
* Developed and maintained automation scripts and processes to streamline infrastructure provisioning, scaling, and maintenance.
* Utilized configuration management tools like Ansible, Puppet, or Chef to maintain server configurations and ensure consistency across environments.
* Managed real-time data storage solutions, such as key-value stores or in-memory databases, within Kubernetes for low-latency access to processed data.
* Created AWS Glue crawlers for crawling the source data in S3 and RDS.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs using CloudWatch.
* Monitored system performance and logs using AWS CloudWatch, taking proactive actions to optimize data pipelines.
* Configured Elastic Lod Balancers with EC2 Auto scaling groups.
* Implemented AWS solutions using E2C, S3, RDS, EBS, Elastic Load Balancer, Lambda, Route53, CloudFormation, Auto Scaling
* Utilized Spark SQL for complex data transformations and querying large datasets in distributed environments.
* Administered Hadoop cluster, including Pig, Hive, HBase, and Oozie, to support big data processing and analytics.
* Facilitated data ingestion from various sources by configuring Sqoop jobs and ensured data consistency.
* Managed real-time data streams using Apache Kafka, facilitating seamless data integration and processing.
* Enhanced data querying and analytics capabilities by implementing Impala for fast SQL queries on Hadoop.
* Developed and maintained Snowflake data models for efficient data warehousing and reporting.
* Created and managed ERwin data models to document data structures and relationships.
* Administered Oracle databases, optimizing performance and ensuring data integrity.
* Maintained clear and up-to-date documentation for Terraform configurations, workflows, and infrastructure designs.
* Collaborated with cross-functional teams using Git and GitHub for version control and code collaboration.
* Proficiently managed Linux servers, ensuring the stability and security of data infrastructure.
* Exposed RESTful APIs and processed JSON data for seamless integration with external systems.

**Environment:** Python, Scala, Apache Spark, AWS, EC2, S3, RDS, EBS, ELB, ETL, AWS Glue, Terraform, CloudWatch, Spark SQL, Hadoop, Pig, Hive, HBase, Oozie, Sqoop, Kafka, Hadoop, Snowflake, Erwin, Oracle, Git, GitHub, RESTful API, JSON, Linux.

**Josh Software, Pune, India**

**DevOps Engineer/Data Engineer Dec 2018 - Aug 2020**

**Responsibilities:**

* Collaborated with cross-functional teams to design and implement scalable ETL pipelines using Python and Django, ensuring efficient data processing and transformation.
* Utilized SSIS (SQL Server Integration Services) for extracting, transforming, and loading data from diverse sources into data warehouses.
* Managed and optimized data storage on AWS, including EC2 instances, S3 buckets, and RDS databases, ensuring data availability and reliability.
* Implemented big data solutions using Hadoop ecosystem tools such as Hive, MapReduce, and HDFS, handling large volumes of structured and unstructured data.
* Configured and maintained real-time data streaming pipelines with Kafka, ensuring high-throughput data ingestion and processing.
* Developed automation scripts and workflows to streamline data pipeline processes, making it easier to ingest, transform, and analyze data using Athena
* Maintained the AWS Glue Data Catalog to track metadata and facilitate discovery of data sources and tables for use in Athena.
* Developed and consumed REST APIs to facilitate seamless communication between different data systems and applications.
* Leveraged PySpark and Spark for large-scale data processing, analysis, and machine learning tasks, optimizing data workflows for performance.
* Collaborated with Agile development teams, participating in sprint planning, daily stand-ups, and sprint reviews to ensure alignment between data engineering and software development efforts.
* Managed version control and collaborated on coding projects using GitHub, facilitating collaborative development and code review processes.
* Automated build and deployment processes using Jenkins and Maven, ensuring continuous integration and delivery of data pipelines and applications.
* Tracked project progress, managed tasks, and reported issues using JIRA, ensuring efficient project management and timely issue resolution.
* Designed, implemented, and maintained scalable and secure REST APIs for data access and integration with external applications.
* Set up and configured monitoring and logging tools like AWS CloudWatch, ELK Stack (Elasticsearch, Logstash, Kibana), or Prometheus to track system performance and troubleshoot issues
* Collaborated closely with development and operations teams to bridge the gap between software development and IT operations, promoting a culture of collaboration and shared responsibility (DevOps).
* Implemented data quality checks and validation processes to ensure accuracy, completeness, and consistency of data in data pipelines.
* Assisted in optimizing SQL queries and database performance for efficient data retrieval and processing.
* Participated in code and design reviews, providing constructive feedback and suggesting improvements to ensure high-quality codebase.
* Documented technical specifications, data architecture, and ETL processes to facilitate knowledge transfer and ensure maintainability of data solutions.

**Environment:** Python, Django, ETL, SSIS, AWS, EC2, S3, RDS, Hadoop, Hive, MapReduce, HDFS, Kafka, REST API, GitHub, PySpark, Spark, Agile, Jenkins, Athena, Maven, JIRA.

**CMC Hyderabad, India**

**Python Developer July 2015 - Dec 2018**

**Responsibilities:**

* Collaborated with senior developers to build and maintain Python-based web applications using Django, ensuring code quality and adherence to best practices.
* Developed responsive and user-friendly front-end interfaces by utilizing HTML, CSS, Bootstrap, and JQuery.
* Worked with XML and JSON data formats to facilitate data exchange between different systems and components.
* Participated in the development of RESTful web services, implementing CRUD operations and enhancing system interoperability.
* Leveraged JavaScript to create interactive features and improve the overall user experience.
* Utilized SQLAlchemy for database interaction, including creating models, querying, and database schema management.
* Assisted in the extraction, transformation, and loading (ETL) process of data using Talend, ensuring data accuracy and integrity.
* Extracted data from multiple source systems S3, Redshift, RDS and Created multiple tables/databases in Glue Catalog by creating Glue Crawlers.
* Collaborated with the database team to optimize MySQL databases, improving query performance and efficiency.
* Contributed to version control and code management using Git, ensuring smooth collaboration within the development team.
* Gained experience in configuring and managing web servers like Nginx to deploy web applications.
* Worked within the LAMP (Linux, Apache, MySQL, PHP) stack environment, gaining exposure to server-side scripting and database management.
* Participated in the development of CherryPy-based web applications, gaining proficiency in this Python web framework. Assisted in troubleshooting and resolving technical issues, enhancing debugging skills.
* Conducted code reviews and collaborated with team members to implement code improvements.

**Environment:** Python, Django, HTML, CSS, XML, Bootstrap, JQuery, JSON, JavaScript, SQLAlchemy, ETL, Talend, MySQL, REST, Git, CherryPy, Nginx, LAMP.