**Email:** **raj@abntechcorp.com**

**Phone Number: +1 (469) 722-7347**

**Anil Kumar**

Bitbucket /Git Repo: [BITBUCKET](https://bitbucket.org/abn_tech_corp/datalake/src/main/generic_script/python/) AWS: [aws\_solution\_architect](https://www.credly.com/badges/6f0d13f6-4546-4774-a6ee-cbe4d2414463/linked_in?t=sco272)

LinkedIn: <https://www.linkedin.com/in/anil16/>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**PROFESSIONAL SUMMARY**

* 9+ years of experience across all phases of SDLC, specializing in requirements analysis, application design, development, integration, maintenance, installation, implementation, and testing of client-server and web applications on the Big Data Eco-System.
* Extensive working knowledge of **Amazon Web Services** (**AWS**), including **EC2**, **S3**, **VPC**, **ELB**, **IAM**, **DynamoDB**, **CloudFront**, **CloudWatch**, **Route** **53**, **Elastic** **Beanstalk** (**EBS**), **Auto** **Scaling**, **Security** **Groups**, **Redshift**, **CloudFormation**, **CloudTrail**, **Kinesis**, **IAM,** **SQS**, **SNS**, and **SES**.
* Expertise in **Big** **Data** **processing** using **Hadoop** and its ecosystem (**MapReduce**, **Pig**, **Spark**, **Scala**, **Hive**, **Sqoop**, **Flume**, **HBase**, **Cassandra**, **MongoDB**, **Kafka**), implementation, maintenance, **ETL**, and **Big** **Data** analysis operations.
* Proficient in scripting with **Python** and **PySpark** **APIs** for data analysis.
* Expertise in various Python libraries including **PySpark**, **Pytest**, **Pymongo**, **PyExcel**, **Boto3**, **NumPy**, and **Beautiful** **Soup**.
* Experienced in designing, developing, and maintaining data pipelines on **AWS** **Cloud** using services like **AWS** **Glue**, **AWS** **Lambda**, and **AWS** **EMR**.
* Skilled in optimizing and tuning AWS services, such as **AWS** **Redshift**, for improved query performance and cost efficiency.
* Experienced in building and managing data lakes on **AWS** **S3,** including data ingestion, storage, and access control mechanisms.
* Designed and developed core data pipeline code, using Java and **Python**.
* Hands-on experience with **AWS** **DynamoDB** for **NoSQL** database solutions and designing scalable data models.
* Expertise in working with AWS **Athena** for serverless querying of data stored in **S3** buckets.
* Familiar with AWS **Data** **Pipeline** for orchestrating complex data workflows and **ETL** processes.
* Implemented complex data processing tasks using Scala and **Apache** **Spark**, optimizing performance for large-scale data transformations.
* Conducted performance tuning and optimization of **Scala** **code**, improving application efficiency and reducing processing times.
* Designed and implemented scalable **data** **pipelines** on **Databricks**, leveraging **Apache** **Spark** for efficient data processing and analytics.
* Implemented **Delta** **Lake** on **Databricks** to ensure data reliability, consistency, and efficient handling of batch and streaming data.
* Designed and implemented data warehousing solutions using **GCP** **BigQuery**, enabling fast and scalable **SQL** **queries** on large datasets.
* Developed ETL processes to ingest data into **BigQuery** from various sources, ensuring efficient data loading and transformation.
* Proficient in implementing real-time data streaming solutions using **AWS** **Kinesis** and **AWS** **Lambda**.
* Knowledgeable in setting up and managing **AWS** **EC2** instances for data processing and analytics workloads.
* Skilled in data integration across heterogeneous sources using AWS **Glue** **Data** **Catalog**.
* Experienced in implementing data governance policies and practices using **AWS** **Lake** **Formation**.
* Proficient in monitoring and managing AWS resources using **AWS** **CloudWatch** and **AWS** **CloudTrail**.
* Hands-on experience with AWS **CLI** and **SDKs** for automation, scripting, and managing AWS resources.
* Familiarity with AWS **IAM** for fine-grained access control and security management.
* Proficient in working with **SQL**, **Python**, **Scala**, and other programming languages for data manipulation and analysis.
* Experience with version control systems like **Git** for managing codebase and collaborating with teams.
* Experienced Data Engineer proficient in **AWS** **EMR** and **Hadoop** ecosystems.
* Skilled in leveraging **AWS** **EMR** for big data processing, **data** **lakes**, and analytics.
* Strong knowledge of Sqoop for efficient data ingestion from relational **databases** to **Hadoop**.
* Proficient in managing **Hive** **tables**, including **external** **tables**, for **structured** **data** storage and querying.
* Proficient in implementing and managing **Snowflake** **Data** **Warehouse** on **AWS** **cloud** **infrastructure**.
* Experience with setting up **Snowflake** **accounts**, **warehouses**, **databases**, and **schemas** for data storage and processing.
* Experienced working with **JIRA** for project management, **GIT** for source code management, and JENKINS for continuous integration and code reviews.

**Certifications:**

* AWS SAA-C03
* AZURE AZ-900
* AZURE DP-900

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Cloud Technologies** | Amazon Web Services (IAM, S3, EC2, VPC, ELB, Route53, RDS, Auto Scaling, Cloud Front), Jenkins, GIT, CHEF, CONSUL, Docker, and Rack Space. |
| **Big Data Ecosystem** | HDFS, Yarn, Map Reduce, Spark, Kafka, Kafka Connect, Hive, Airflow, StreamSets, Impala, Sqoop, HBase, Flume, Pig, Ambari, Oozie, Zookeeper, Nifi, Sentry, Ranger |
| **Hadoop Distributions** | Apache Hadoop 3.3/2.x, Cloudera CDP, Hortonworks HDP, Amazon EMR (EMR, EC2, EBS, RDS, S3, Athena, Glue, Elasticsearch, Lambda, SQS, DynamoDB, Redshift, ECS, Quick sight) Azure HDInsight (Databricks, Data Lake, Blob Storage, Data Factory ADF, SQL DB, SQL DWH, Cosmos DB, Azure AD). |
| **Integration Tools** |  Informatica, Autosys |
| **Visualization/ Reporting** | Tableau, Power BI & matplotlib |
| **Data Warehouse** | AWS RedShift, Google Cloud Storage, Snowflake, Teradata, |
| **Operating Systems** | Linux, Windows |
| **Devops Tools** | Urban Code Deploy, Jenkins (CI), Puppet, Chef, Ansible, AWS. |
| **Orchestration Tool** | Apache Airflow, Astronomer, Oozie |
| **Languages** | C, SQL, Languages Shell, and Python scripting. |
| **Databases** | Oracle, SQL Server, MySQL, HBase, MongoDB, RedShift, DynamoDB and Elastic Cache, Cassandra, Oracle toad/11g |
| **Web/App Server** | Apache, IIS, HIS, Tomcat, Application Server. |
| **CI Tools** | Hudson, Jenkins, Bamboo, Cruise Control. |
| **Devops / Build & Release Engineering** | Jenkins, Perforce, Docker, AWS, Chef, puppet, Ant, Atlassian-Jira, GitHub, Ansible, Splunk. |

**PROFESSIONAL EXPERIENCE**

**Client:** **Wynn Las Vegas, Dallas, Texas March 2024– Till date**

**Senior Data Engineer**

**Responsibilities:**

* Designed and implemented data warehousing solutions using **Snowflake**, optimizing data storage and query performance for large-scale datasets.
* Developed and maintained **ETL** **pipelines** to ingest data into **Snowflake**, ensuring efficient **data** **loading**, **transformation**, and **storage**.
* Utilized **Snowflake's** capabilities for data sharing, cloning, and **time** **travel** to enhance data accessibility and backup processes.
* Implemented security best practices in **Snowflake**, including role-based access control, data encryption, and secure data sharing.
* Developed and **orchestrated** complex ETL workflows using **Apache** **Airflow**, automating data processing tasks and improving workflow efficiency.
* Successfully implemented and managed **Infrastructure** as Code (**IaC**) using **Terraform** to provision and manage **AWS** **resources**, resulting in streamlined infrastructure deployment processes and reduced manual intervention.
* Managed configurations of **AWS** **environments** using **Terraform**, enabling consistent and repeatable infrastructure deployments across development, **staging**, and **production** environments.
* Configured **Airflow** **DAGs** to schedule and **monitor** data **pipelines**, ensuring timely and accurate data processing.
* Implemented custom Airflow operators and **hooks** to extend functionality and integrate with various data sources and services.
* Monitored and troubleshooted Airflow tasks, optimizing performance and ensuring reliability of data workflows.
* Containerized applications using **Docker**, improving development, testing, and deployment processes by ensuring consistent environments.
* Developed **Docker** **Compose** files to define multi-container applications, streamlining the deployment and orchestration of complex systems.
* Implemented **CI**/**CD** **pipelines** for **Dockized** applications, utilizing tools such as **Jenkins**, **GitLab** CI, and **AWS** **Code Pipeline**.
* Optimized **Docker** **images** and container configurations to improve resource utilization and application performance.
* Designed and implemented data lake solutions using **AWS** **S3**, enabling scalable and durable storage for structured and unstructured data.
* Integrated **S3** with various **AWS** **services**, such as **Glue**, **Redshift**, and **EMR**, for seamless data processing and analysis.
* Orchestrated complex workflows using **AWS** **Step** **Functions**, improving visibility and reliability of data processing pipelines.
* Integrated **Step** **Functions** with other AWS services, such as **Lambda**, **SNS**, and **SQS**, to build scalable and event-driven architectures.
* Developed Python scripts using **AWS** **Lambda** to gather data from various sources to the **Snowflake** data **warehouse**.
* Working on incremental data load from Snowflake sharable object/secured view to BOD **Snowflake** **tables** using **DBT ELT** tool.
* Worked on Snowflake's storage optimization features, such as micro-partitioning and automatic clustering to minimize data scan and improve query performance.
* Used PySpark to load data from AWS S3 to Snowflake by setting up the **AWS** **credentials**, **installing** required **libraries** and used **snowflake**-**connector**-**python**.
* Integrated Amazon Lex with other AWS services such as **AWS** **Lambda**, **DynamoDB**, and **S3** to enable seamless backend processing, data storage, and content retrieval.
* Involved in building the data pipelines for data ingestion from SQL Server to S3 bucket.
* Implemented a **CI/ CD** pipeline with **Docker**, **Jenkins** and **GitHub** by virtualizing the servers using **Docker** for the Dev and Test environments by achieving needs through configuring automation using Containerization.

**Environment:** Amazon Web Services, IAM, S3, RDS, EC2, VPC, Redshift, Glue, Java, Hudi, Cloud watch, Bit Bucket, Dockers, Python,Kinesis, Firehose Kinesis,Snowflake**,** Airflow.

**Client: Adobe,** **Dallas, Texas June 2022 - Feb 2024**

**Senior Data Engineer**

**Responsibilities:**

* Designed and implemented data warehousing solutions using **AWS services**, optimizing data storage and query performance for large-scale datasets.
* Involved in all phases of SDLC including Requirement Gathering, Design, Analysis, Testing of customer specifications, Development, and Deployment of the Application.
* Collaborated with development teams to build custom **API** **endpoints** and extensions, extending platform capabilities and supporting business-specific use cases.
* Architected and delivered a **REST** **API** leveraging **JWT** authentication, successfully **integrating** **APIs**.
* Tested the Service **API** using **Postman** and then implemented it using **Python** **requests**.
* Led the strategic effort to transition authentication protocols within **JWT** to **OAuth2**, meticulously orchestrating a phased roadmap. Ensured seamless service delivery, prioritizing data security and enhancing the customer experience.
* Involved in designing and deploying a large application utilizing almost the entire AWS stack (**including EC2, S3, Glue, Redshift, RDS, DynamoDB, SNS, SQS, IAM**) focusing on high availability, fault tolerance, and auto-scaling in AWS CloudFormation.
* Designed and implemented scalable data pipelines using **AWS** **services** such as **S3**, **EMR**, and **Glue**, ensuring efficient data ingestion, transformation, and storage for real-time and batch processing needs.
* Orchestrated **ETL** **workflows** using **AWS** **Glue** to automate data processing tasks, reducing manual effort by 50% and improving data accuracy and consistency.
* Leveraged **AWS** **Redshift** for data warehousing, **optimizing** **query** **performance** and improving data accessibility for business intelligence and reporting purposes.
* Designed and implemented event-driven architectures using **AWS** **Event Bridge** to facilitate seamless integration and communication between various services and applications, improving system reliability and scalability.
* Scheduled and orchestrated complex data workflows using **AWS** **Step** **Functions**, improving workflow visibility and reducing processing time by 40%.
* Developed and deployed serverless applications using **AWS** **Lambda**, **optimizing** **resource** utilization and reducing infrastructure costs by 20%.
* Implemented data lake solutions using **AWS** **S3**, **establishing** efficient data storage, organization, and access patterns for structured and unstructured data, improving data discovery and analytics capabilities.
* Utilized **AWS** **EMR** (**Elastic** **MapReduce**) for big data processing tasks, optimizing cluster configurations and job scheduling to handle large-scale data processing workloads effectively.
* Conducted performance tuning and optimization of **SQL** **queries** and data processing workflows on **AWS** **Redshift**, improving query response times and overall system efficiency.
* Developed and optimized **Hive** **queries** to process large-scale datasets efficiently.
* Implemented data transformations and aggregations using **HiveQL**.
* Managed **Hive** **Metastore** and performed schema evolution tasks.
* Utilized **Hive** **partitions** and **bucketing** for data organization and optimization.
* Leveraged EMR for parallel processing of data using **Hadoop** **MapReduce** and **Spark**.
* Proficient in using **Terraform** for infrastructure as code (**IaC**) to provision and manage AWS resources.
* Experienced in writing **Terraform** **scripts** to automate the creation and management of AWS infrastructure, including EC2 instances, S3 buckets, VPCs, and IAM roles.
* Designed and implemented real-time data streaming pipelines using **AWS** **Kinesis** **Streams**, allowing for efficient **ingestion**, **processing**, and analysis of large-scale data streams.
* Configured Kinesis Firehose delivery streams to seamlessly load streaming data into **AWS** data stores (**Amazon** **S3**, **Amazon** **Redshift**), enabling near real-time analytics and insights generation.
* Developed various Spark applications using **PySpark** to perform various enrichments of user behavioral data (click stream data) merged with user profile data.
* Utilized **CI/CD** pipelines with **AWS** **Code Pipeline**, **AWS** **Code Build**, and **AWS** **Code Deploy** for automated build, test, and deployment of containerized applications.
* Involved in data cleansing, event enrichment, **data aggregation**, **de**-**normalization**, and **data preparation** needed for downstream model learning and reporting.
* Utilized **Spark API** to implement batch processing of jobs.
* Used JavaScript and **JSON** to update a portion of a webpage.
* Troubleshot **Spark** applications for improved **error** tolerance.

**Environment:** PySpark, AWS, Amazon Connect, Git, Jenkins, Snowflake, Neptune, spark, Java, IBM DataStage tool, Airflow, DBT tool, SQL, Python, Netezza DB AWS Lambda, S3, EC2, EMR, Redshift, AWS Glue

**EMRaaga Informatics LLP, India. Nov2018 – June 2022**

**Data Engineer**

**Responsibilities:**

* Involved in all phases of SDLC including Requirement Gathering, Design, Analysis and Testing of customer specifications, Development, and Deployment of the Application.
* Worked on data that was a combination of **unstructured** and **structured** data from multiple sources and automate the data cleaning using **Python** **scripts**. Performed data analysis by using **Hive** to retrieve the data from **Hadoop** **cluster**, SQL to retrieve data from Oracle database.
* Loaded data into **S3** **buckets** using **AWS** **Glue** and PySpark. Involved in **filtering** **data** stored in **S3** **buckets** using Elasticsearch and loaded data into **Hive** **external** **tables**.
* Experience implementing Cloud based Linux OS in AWS to Develop Scalable Applications with Python.
* Created Data Quality Scripts using **SQL** and **Hive** to validate successful data load and quality of the data and extracted data from **HDFS**.
* Developed data analysis by using Hive to retrieve the data from **Hadoop** **cluster**, SQL to retrieve data from Oracle database and used ETL for data transformation. Worked on machine learning on large size data using **Spark** and Hive**,** Map Reduce.
* Worked on **AWS** **Redshift** and **RDS** for implementing models and data on **RDS** and **Redshift**.
* Experience in creating **Kafka** **producer** and **Kafka** **consumer** for Spark streaming which gets the data from different learning systems of the patients.
* Automated the existing scripts for performance calculations using scheduling tools like airflow.
* Used **AWS** **EMR** to transform and move large amounts of data into and out of other **AWS** data stores and databases, such as Amazon Simple Storage Service (**Amazon** **S3**) and Amazon DynamoDB.
* Spark Streaming collects this data from **Kafka** in near-real-time and performs necessary transformations and aggregation on the fly to build the common learner data model.
* Implemented and maintained real-time data processing pipelines using **Apache** **Spark** Streaming on Databricks.
* Worked on Snowflake Schemas and **Data** **Warehousing** and processed batch and streaming data load pipeline using Snow Pipe and from data lake Confidential AWS S3 bucket
* Implemented **Spark**/**PySpark**, **Python**, R for potential healthcare project in the **Hadoop**/**Hive** environment with Linux/Windows for big data resources
* Rewrote SSIS packages in Spark using **Spark** **SQL**, **PySpark** and executed in **AWS** **EMR**.
* Created reports and produced rich data visualizations to model data into human-readable form with the Tableau, Power BI, Matplotlib, and Seaborn to show client how prediction can help the business decisions.

**Environment:** AWS, Hadoop, Map Reduce, Snowflake, Cloudera, Airflow v1.9.0, PySpark, Kafka, HDFS, Hive, Pig, Oozie, Scala, Eclipse, TFS, Flume, Kinesis, Oracle, Unix Shell Scripting, Python.

**Wipro, India. June 2017– Nov 2018**

**Data Engineer**

**Responsibilities:**

* Ingested user behavioral data from external servers such as **FTP** **server** and **S3** buckets on daily basis using custom Input Adapters.
* Created Sqoop scripts to **import**/**export** user profile data from **RDBMS** to **S3** **Data** **Lake**.
* Developed various spark applications using Scala to perform various enrichments of user behavioral data (click stream data) merged with user profile data
* Involved in data cleansing, event **enrichment**, **data** **aggregation**, **de**-**normalization** and data preparation needed for downstream model learning and reporting.
* Utilized Spark Scala API to implement batch processing of jobs.
* Troubleshooting Spark applications for improved error tolerance.
* Install, configure, test, monitor, upgrade, and tune new and existing **PostgreSQL**/Snowflake databases. Designed number of partitions and replication factor for **Kafka** **topics** based on business requirements and worked on migrating Map Reduce programs into Spark transformations using Spark and Scala, initially done using python (**PySpark**).
* Implemented Spark using **Python** and **Spark** **SQL** for faster processing of data and worked on migrating Map Reduce programs into Spark transformations using Spark and Scala, initially done using python (**PySpark**)
* Fine-tuning spark applications/jobs to improve the efficiency and overall processing time for the pipelines
* Created **Kafka** **producer** **API** to send live-stream data into various **Kafka** **topics**.
* Developed Spark-Streaming applications to consume the data from **Kafka** **topics** and to insert the processed streams to HBase.
* Utilized Spark in Memory capabilities, to handle large datasets.
* Used broadcast variables in spark, effective & efficient Joins, transformations, and other capabilities for data processing and experienced in working with **EMR** **cluster** and **S3** in **AWS** cloud.
* Creating Hive tables, loading and analyzing data using hive scripts. Implemented Partitioning, Dynamic Partitions, Buckets in Hive.
* Involved in continuous Integration of application using Jenkins.

**Environment**: AWS EMR, Spark, Hive, HDFS, Big Query, Hadoop, Sqoop, Spark, Kafka, Oozie, HBase, Scala, Map Reduce.

**Ivy Technologies India. Aug 2015 to June 2017**

**ETL Data Engineer**

**Responsibilities:**

* Conducted analysis on significant observations to identify insights, potential root causes, and necessary corrective actions.
* Successfully implemented Informatica BDM (Big Data Management) and Hadoop technologies to achieve project objectives and deliver desired outcomes.
* Constructed and maintained end-to-end data pipelines using Sqoop and Spark, enabling seamless data transfer from MySQL and Oracle to HDFS for precise mortgage risk assessments.
* Developed shell scripts to optimize ETL flows within Informatica workflows.
* Utilized Informatica file watch events for monitoring and retrieving SFTP CSV files.
* Created parameterized mapping templates for cloud integration, covering stage, dimension (SCD Type 1, SCD Type 2, and Incremental Load), and fact load processes. Implemented parameterization for DB and table objects to enhance flexibility and operational efficiency.
* Architected data lakes infrastructure, enabling seamless processing of structured/unstructured data; leveraged Hadoop and Cloudera to analyze loan performance data, leading to a 25% improvement in decision-making accuracy and a 30% reduction in processing time.
* Developed and implemented customized Hive queries, enhancing efficiency in data analysis and decision-making processes, resulting in a 30% increase in revenue through targeted marketing strategies.
* Built Hive integration for efficient storage and retrieval, facilitating quicker access to critical market data.
* Managed data importing from various sources, performed transformations using Hive, and loaded data into HDFS, supporting a large-scale marketing campaign analysis project.
* Deployed key Hadoop ecosystem components such as Apache Hive, Apache PySpark, and Sqoop for a big data migration project.
* Leveraged PySpark for rapid data testing and processing, expediting data validation for more accurate analytics.
* Implemented CI/CD pipelines in the Hadoop environment, enhancing development and deployment efficiency for various data projects.
* Enhanced Map Reduce job performance by creating combiners, partitioning, and implementing distributed cache for a project focused on optimizing the organization's data processing pipeline.
* Utilized JIRA for effective project management, improving collaboration and efficiency within cross-functional teams.
* Collaboratively resolved JVM-related issues, ensuring smooth system performance for data processing tasks.
* Provided weekly written and verbal communications to the team.
* Maintained code repositories with Git, version tracking, and efficient team collaboration.

**Environment:** : Informatica Power Center, Informatica BDM, csv files, Sqoop, SQL, HDFS, ETL, Apache Pyspark, Hadoop Hive, Cloudera, Map Reduce, Python, PySpark, JIRA