**** 

**Anil**

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

**PROFESSIONAL SUMMARY**

* 8+years professional experience in all phases of SDLC including requirements analysis, applications design, development, Integration, maintenance, Installation, Implementation and testing of various client server and web applications on Big Data Eco-System.
* Strong experience in writing scripts using **Python API**, **PySpark** **API** for analyzing the data.
* Extensively used Python Libraries **PySpark**, **Pytest**, Pymongo, PyExcel, Boto3, Psycopg, embedPy, **NumPy** and Beautiful Soup.
* Experience implementing Cloud based Linux OS in **AWS** to Develop Scalable Applications with **Python**.
* Experience in Google Cloud components, Google container builders and **GCP** client libraries and cloud SDK's
* Hands-on use of **Spark** and **Scala** APIs to compare the performance of Spark with Hive and SQL, and Spark SQL to manipulate Data Frames in Scala.
* Used SQL, **Numpy,** **Pandas**, Scikit-learn, Spark, **Snowflake**, Hive for Data Analysis & Modeling**.**
* Implemented solutions for ingesting data from various sources and processing the Data-at-Rest utilizing Big Data through Map Reduce Frameworks, **HBase**, and apache **hive**.
* Experience in developing Map Reduce Programs using **Apache Hadoop** for analyzing the big data as per the requirement.
* Hands-on experience on **Spark MLlib** utilities including classification, regression, clustering, collaborative filtering, dimensionality reduction.
* Involved in Data Integration and Data warehouse techniques, using ETL tools like Informatica Intelligent Cloud Services (IICS), Oracle Data Integrator, MS SQL Server and cloud-based database solutions including Azure Synapse, Snowflakes.
* Experience with continuous integration and automation using Jenkins.
* Manage the Space and Storage of the Databases through various Shell Scripts scheduled in the corncob

periodically in regular intervals. Write Complex Shell Scripts that automate the Import of Reports to PostgreSQL and MySQL Databases

* Expertise in Big Data processing using Hadoop, Hadoop Ecosystem (Map Reduce, Pig, Spark, Scala, Hive, Sqoop, Flume and HBase, Cassandra, Mongo DB, Kafka Framework) implementation, maintenance, ETL and Big Data analysis operations Adapt at implementing E2E solutions on Big Data using Hadoop framework, executed, and designed big data solutions on multiple distribution systems like Cloudera (CDH3 & CDH4), Hortonworks.
* Expert in developing SSIS/DTS Packages to extract, transform and load (**ETL)** data into data warehouse/ data marts from heterogeneous sources.
* Worked on implementing advanced machine learning models for image processing. Working along with data science team to implement machine learning and artificial intelligence techniques to improve the process and automating IOL cosmetic inspection system.
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
* Good working knowledge and experience in **Amazon Web Services** (AWS) Cloud Platform which includes services like EC2, S3, VPC, ELB, IAM, DynamoDB, Cloud Front, Cloud Watch, Route 53, Elastic Beanstalk (EBS), Auto Scaling, Security Groups, EC2 Container Service (ECS), Code Commit, Code Pipeline, Code Build, Code Deploy, Dynamo DB, Auto Scaling, Security Groups, Red shift, CloudWatch, CloudFormation, CloudTrail, Ops Works, Kinesis, IAM, SQS, SNS, SES.
* Experienced working with **JIRA** for project management, **GIT** for source code management, **JENKINS** for continuous integration and **Crucible** for code reviews
* Hands-on experience working on **NoSQL databases** including **HBase**, **Cassandra**, **MongoDB**, and its integration with Hadoop cluster, and Kubernetes cluster.
* Experience in Data Analysis, Data Profiling, **Data Integration**, Migration, Data governance and Metadata Management, Master Data Management and Configuration Management.
* Hands-on experience in working with **Cloudera** and **Azure HDInsight**
* Configured **AWS CLI** and performed necessary actions on the AWS services using shell scripting.
* Implemented **ELK** (Elastic Search, Log stash, Kibana) stack to collect and analyze the logs produced by the spark cluster.
* Enabled AWS Managed Streaming for **Kafka** (MSK) for developers using Cloud Development Kit (CDK) in Python
* Extensively worked with **Teradata** utilities Fast export, and Multi Load to export and load data to/from different source systems including flat files.
* Experienced in building Automation Regressing Scripts for validation of **ETL** process between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using Python.
* Knowledge in Informatica Intelligent Cloud Services (IICS) Data Synchronization, Data Replication, Contact Validation and Data Masking.
* Designed, developed, and maintained ETL pipelines using Databricks, ensuring efficient data processing and transformation
* Proficiency in SQL across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle).

**Education:**

• Bachelor of Technology, India. 2017.

**Certifications:**

* AWS CLF-C02
* 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, MapReduce, 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, Quicksight) Azure HDInsight (Databricks, DataLake, Blob Storage, Data Factory ADF, SQL DB, SQL DWH, CosmosDB, Azure AD). |
| **Integration Tools** |  Informatica, Autosys |
| **Operating Systems** | Linux, Windows |
| **Devops Tools** | Urban Code Deploy, Jenkins (CI), Puppet, Chef, Ansible, AWS. |
| **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, WebSphere Application Server, JBoss. |
| **CI Tools** | Hudson, Jenkins, Bamboo, Cruise Control. |
| **Devops/Build & Release Engineering** | Jenkins, Perforce, Docker, Udeploy AWS, Chef, puppet, Ant, Atlassian-Jira, GITHub, Ansible, Open Stack and Salt Stack, Splunk. |

**PROFESSIONAL EXPERIENCE**

**Adobe Express Platform Customer| TCS |Dallas, Texas June 2022 - Present**

**Sr. AWS 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.
* Involved in designing and deploying a large application utilizing almost the entire **AWS stack** (Including **EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM**) focusing on high availability, fault tolerance, and auto-scaling in **AWS Cloud Formation**.
* Architected and delivered a **RESTful** **API** leveraging **JWT** authentication, successfully integrating Adobe Experience Platform APIs (Query Service, Catalog, Data Access); streamlined data flow, improving system efficiency by 50% and elevating user experience.
* Tested the **Adobe Experience Platform Service API using Postman** and then implemented it using **Python requests.**
* Working on **migration** project of moving current applications in traditional data center to **AWS** by using **AWS** services.
* Implemented a proof of concept deploying this product in **AWS S3** bucket.
* Implemented usage of **Amazon EMR** for processing Big Data across a Hadoop Cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3) .
* Spearheaded the strategic initiative to shift authentication protocols in Adobe Experience Platform from **JWT** to **OAuth2**, meticulously planning and executing a phased roadmap; ensured uninterrupted service delivery, safeguarding data security and customer experience.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark databricks cluster
* 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.
* Launching **AmazonEC2** Cloud Instances using **Amazon Web Services** (**Linux/ Ubuntu/RHEL**) and configuring launched instances with respect to specific applications.
* Installed application on **AWS EC2** instances and configured the storage on **S3 buckets**. Assisted the team experienced in deploying **AWS** and **Cloud Platform**
* Managed **IAM policies**, providing access to different **AWS resources**, design and refine the workflows used to grant access.
* Experienced in using **Hive scripts** to do transformations, event joins, filters and pre-aggregations before storing the data into HDFS.
* Designed and developed the **core data pipeline** code, involving work in Python.
* Implemented and maintained the monitoring and alerting of production and **corporate servers/storage** using **AWS** **Cloud watch.**
* Working experience with data streaming process with **Kafka**, **Apache Spark**, **Hive**.
* Used Spark API over Hadoop YARN as execution engine for data analytics using Hive.
* Designed **AWS Cloud Formation** templates to create custom sized **VPC**, **subnets**, **NAT** to ensure successful deployment of **Web applications** and **database templates**.
* Launched Compute (**EC2**) and **DB** (**Aurora**, **Cassandra**) instances from **Amazon Management Console** and **CLI.**
* Designed and Developed Spark workflows using Scala for data pull from AWS S3 bucket and Snowflake applying transformations on it.
* Developed 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
* Led the migration of on-premises data processing to Databricks, resulting in improved scalability and cost efficiency
* Used Lambda function for Infrastructure automation.
* Hands on experience in customizing **Splunk** dashboards, visualizations, configurations using customized Splunk queries.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production.
* Worked with Spark Ecosystem using Scala and Hive Queries on different data formats like Text file and parquet.
* **Elastic search** experience and capacity planning and **cluster maintenance**. Continuously looks for ways to improve and sets a very high bar in terms of quality.
* Implemented real time log analytics pipeline using **Elastic search**.
* Setup and configured **Elastic search** in a POC test environment to ingest over million records from oracle DB.
* Designed the data models to be used in data intensive **AWS Lambda**applications which are aimed to do complex analysis creating analytical reports for end**-to-end traceability, lineage, definition** of Key Business elements from **Aurora.**
* Deployed applications on **AWS** by using Elastic Beanstalk. Integrated delivery (**CI** and **CD**) using **Jenkins** and puppet.
* Used Spark Streaming to receive real time data from the **Kafka** and store the stream data to **HDFS** using Python and NoSQL database **DynamoDB**
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs with Scala.
* Implemented Workload Management (WML) in **Redshift** to prioritize basic dashboard queries over more complex longer running adhoc queries. This allowed for a more reliable and faster reporting interface, giving sub-second query response for basic queries.
* Responsible for **Designing Logical and Physical data modelling** for various data sources on Confidential **Redshift**
* Wrote scripts and indexing strategy for a migration to Confidential **Redshift** from SQL Server and MySQL databases
* Used Pig Latin at client-side cluster and HiveQL at server-side cluster.
* Worked on **AWS Data Pipeline** to **configure data loads from S3** to into **Redshift**
* Responsible for moving data between different **AWS** compute and storage services by using **AWS Data Pipeline.**
* Design and Develop ETL Processes in **AWS Glue** to relocate Campaign information from outside sources like **S3, ORC/Parquet/Text Files into AWS Redshift.**
* Virtualized the servers using the **Docker** for the test environments and dev-environments needs and configuration automation using **Docker** containers.

**Environment:** Amazon Web Services, IAM, S3, RDS, EC2, VPC, Java, Airflow, Kafka, pig, cloud watch, Bit Bucket, Chef, Sql, Puppet, Ansible, Docker, Apache HTTPD, Apache Tomcat, Snowflake, Scala, JBoss, Junit, Cucumber, Python.

**EM Raaga Informatics LLP, India. Nov 2018 – June 2022**

**AWS 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** and prepared data for Exploratory Data Analysis using data munging.
* 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**, **MapReduce**.
* 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 Matillion from data lake Confidential AWS S3 bucket
* Performed **Exploratory Data Analysis** on features to identify **multi-collinearity** in healthcare data with the help of subject matter experts on various use cases and removed collinear features, which gave us **unique records** and **clean data** to downstream analysis.
* Created various machine learning algorithms and statistical modeling like **Decision Trees**, **Regression models**, **Random Forest**, **Neural Networks**, **SVM**, **Density Based clustering** to identify patient's disease using **Scikit-Learn** package in **Python**.
* Implemented the **Dimensionality Reduction** (**PCA**) on the data to convert high dimensional feature space to lower dimension and to reduce the **Time** and **Speed Complexity**.
* Performed **Feature Selection** technique to select features that are useful for making predictions and dropped features that causing model **Overfitting and Underfitting.**
* Evaluated the model with various performance metrics like **Confusion matrix**, **Accuracy**, **Precision**, **Recall**, **Sensitivity**, and **Specificity.**
* Retrained the machine learning model with optimal parameters by implementing **Hyperparameter tuning** with **Grid Search** and **Random Search**.
* Worked with **Apache NiFi** for Data Ingestion. **Triggered the shell Script** and Schedule them using **NiFi.**
* Deployed the model in **AWS Sage Maker** for making prediction on real-time data in production.
* Implemented **Spark**/**PySpark**, **Python**, **R** for potential healthcare project in the **Hadoop**/**Hive** environment with **Linux/Windows** for big data resources. Used machine learning clustering technique **K-Means** to identify outliers and to classify unlabeled data.
* 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, MapReduce, 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**

**AWS 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
			* Experience in building and architecting multiple Data pipelines, end to end **ETL** and **ELT** process for Data ingestion and transformation in **GCP** and coordinate task among the team.
			* 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 MapReduce 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 MapReduce 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, GCP, Sqoop, Spark, Kafka, Oozie, HBase, Scala, MapReduce.

**Ivy Technologies India. Aug 2016 to June 2017**

**ETL Developer**

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