Name
Email id: xyz09@gmail.com
Mob. No: 91 1212121212

Carrier summary:

- 2+ years of experience in Bigdata and Hadoop ecosystems like HDFS, SparkCore, SparkSql, MapReduce, YARN, Sqoop, Hive and Oozie.
- Load and transform large sets of structured, semi-structured and unstructured data from Relational Database Systems to HDFS and vice-versa using Sqoop tool.
- Data ingestion and refresh from RDBMS to HDFS using Apache Sqoop and processing data through Spark Core and Spark SQL.
- Designing and creating Hive external tables, using shared meta-store instead of Derby, and creating Partitions and Bucketing techniques in Hive using Hive Query Language (HQL)
- Involved in performance tuning of Hive
- Good Knowledge of EC2 instance type and designing and implementing security with VPC, IAM and Security groups, Load Balancer, Internal Gateway and Routing Table etc.
- Experience in EMR, Lambda Functions, AWS Glue, AWS Athena, AWS S3.
- Hands on Experience in python Boto3 library to access the AWS S3 storage
- Experience of creating architecture on AWS, Deploying Hadoop clusters using AWS EC2.
- Hands on Experience in AWS Management tools like CloudFormation, CloudTrail and CloudWatch architecture, monitoring memory, disk Metrics, logs and Graphs, and setting alarms.

- Hands on experience in Server less technologies like AWS GLUE to perform ETL operations, Lambda Functions to trigger the pipeline

Technical Skills:

- **Big Data Ecosystem:** Hadoop, Hive, Flume, Spark, Oozie, Glue.

- **Languages:** Python.

- **Database:** MySQL, Athena, Redshift.

Work Experience:

Currently Working in Autodesk India Pvt Ltd, Bangalore

Project #1:
Project: Internal Project
Project Name: ADP (Autodesk Data Platform)
Environment: HDFS, Oozie, Hive, Spark, Yarn, S3, Redshift, Mysql Database, Athena, Lambda, Kinesis Firehouse, Aws Glue, Qubole, Cloud Watch, ECS, Jenkins, Ambari, **Attunity**, Looker, putty, WinSCP.
Period: July 2018 to till data.
Role: Big Data Developer

Description:

Autodesk Data Platform Projects is to collect structured and semi-structured data from different sources and dumping into ADP. After Collecting the raw data from different source systems running the ETL pipelines to cleansing the data and converting into business needs. Later it is used for Reporting Dashboards and data visualization for internal use to fulfill need of Stakeholders Requirement.

Responsibilities:

- Involved in extracting data from various data sources into Hadoop HDFS. This included data from **Attunity**, Kinesis Firehouse, Aws Glue, Lambda and SDK.
- Responsible for Data Ingestion, Data Cleansing, Data Standardization and Data Transformation.
- Worked on creating Hive managed and external tables based on the requirement.
- Implemented Partitioning and Bucketing on Hive tables for better performance.
- Used Spark-SQL to process the data and to run on Spark engine.
- Worked on Oozie to develop workflows to automate ETL data pipeline.
- Explored with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, and Data Frame.
- Configured Oozie workflow to run multiple Hive and spark jobs which run independently with time and data availability.
- Configuring Hcatsever in Oozie to check the dependency data before the workflow get started.
- Installing, Upgrading and Managing Hadoop Cluster.
- Collaborated with the infrastructure, network, database, application and BI teams to ensure data quality and availability.
- Creating the tables in Athena and integrating with looker dashboard. In Looker Stakeholders will create there-own dashboards with processed final output data for business requirements
- Deploying the oozie workflows by using the automated Jenkins Tool.

Project #2:
Project: Internal Project
Project Name: UCP (Unified Customer Profile)
Client: Autodesk
Environment: HDFS, Sqoop, Hive, EMR, S3, Redshift, MySQL Database, putty, WinSCP. Crontab.
Period: July 2018 to Oct 2018.
Role: Big Data Developer

Description
Purpose of this project is to collect the data from the different source of RDBMS to store it in HDFS. In this project mainly focus on Daily Ingestion, Analytics Pipeline and Reporting Framework. To maintain the day to day into the HDFS from different source of RDBMS. After clubbing the multiple table run the Analytics Pipeline For Stakeholders. Finally Moving the data into Redshift for Reporting Purpose.

Responsibilities:

- Maintain the day to day data in Hive Datawarehouse without lagging from different source systems.
- Writing the Sqoop Jobs to transfer data from SQL Server and Oracle Databases to HDFS.
- Load the processed data into Hive External table and to create Partitioning and Bucketing techniques in hive to improve the performance, involved in choosing different file format's like ORC, Parquet format.
- Scheduling the Sqoop jobs using Crontab.
- To Check the SQL Database for monitoring daily ingestion jobs are successful if not debugging the issue.
- Running Analytics Pipeline for Aggregating and joining the multiple daily ingestion tables to meet the stakeholder's requirements.
- Scheduled the jobs and transferring final output data into Redshift for reporting purpose.
- Creating the cluster group in EMR for running Daily ingestion, Analytics Pipeline jobs.

Education Qualification

BTech from Anna University in 2016 with 68%