BIG DATA ENGINEER

ADDEPALLI YUVA KISHORE

Nellore, Andhra Pradesh, India 524126 | +91 9490910979 | addepalliyuva09@gmail.com

Professional Summary

Dynamic and motivated IT professional with around 4 years of experience as a Big Data Engineer highly effective at developing new programs and fixing problems with existing systems. Proficient in **Python, SQL, Spark Ingestion** and **Agile/Scrum Methodology.** Focused on Usability and performance improvements. Developing Useful, efficient and cost-effective ingestion and analysis projects.

Skills

- Big Data: Hive, Spark, Snowflake
- Confluence: Documentation
- Operating Systems: Linux, Windows
- Word, Excel. PowerPoint. Outlook
- Scrum and Agile Knowledge: Jira
- HDFS and AWS: S3, Athena, EMR, EC2
- AWS: EMR, Lambda, Glue, Athena, IAM, Cloud watch, Redshift
- Database: MYSQL, ORACLE
- Languages: Python, Shell Scripting, SQL
- Performance optimization. Quality assurance

Professional Experience

Organization	Designation	Duration
Legato Health Technologies	Software Engineer	Jan 2021 - Present
Autodesk India Pvt Ltd	Associate Software Engineer	July 2018 – Jan 2021

Project Experience

Data Engineer **01/2021 to Current**

Legato Health Technologies – Bangalore, India

❖ **#3 Project**: Anthem Project

Project Name: National Consumer Cost-Calculation Tool (NCCT)

Environment: HDFS, Hive, Spark, Python, Yarn and Shell script

Period: Jan 2021 to till date.
Role: Data Engineer

Description:
A BCBSA mandate that requires all the Blues plan to be submitted twice a year for a set number of procedures for PPO, Exchange Network, and Alternative network. Blues plans must display this cost data along with an out of pocket estimate for all National Account members.

Roles & Responsibilities:

● Working as Big Data developer, for developing code using Big Data technologies and deploying them in production.
● Working closely with Business Team to understand the requirement and gathering the outcome of changes in cost estimates of National Account members.
● Thoroughly working with testing team to identify any defects on the change and fixing them prior any production deployments.
● Developing NCCT code login using Spark and Scala for different stages: Claim Extraction, Service Type loads and Submission
● Building Data pipelines to transfer data from multiple sources like SQL server, PostgreSQL CSV/Excel etc. to Bigdata Platform (HFDS/Hive)
● Dealing with hive optimization techniques for faster execution of complete processes.
● Transform raw data into meaningful information to support business decisions.
● Apply business rules to data and code as per the requirements.
● Hive, Shell Scripting, Spark, Python used in the complete project.
● Cluster Size: 10 TB (Allocated to our team), Data nodes: 50+. V-Cores: 89 Edge nodes :1 Scheduler: Control-M
● Created a trigger-based script for job execution - on receiving trigger it will execute a process immediately using shell scripting.
● Monitor Production Jobs, in case of failures address cause & provide the required fix / Solution.
● Creation of jobs in Control-M and their migration/execution.
● Bitbucket workflow and usage.
● Work in fast-paced agile development environment to quickly analyze and test potential use cases for business.
● Involved in Agile Methodologies, Daily Scrum Meeting and Sprint Planning.

Associate Data Engineer **07/2018 to 01/2021**
Autodesk India Pvt Limited – Bangalore, India
❖ **#2 Project: ADP (Autodesk Data Platform)**
Environment: HDFS, Oozie, Hive, Spark, Yarn, S3, Redshift, Mysql Database, Athena, Lambda, Aws Glue, Qubole, Cloud Watch, ECS, Jenkins, Ambari, Putty, WinSCP.
Period: July 2018 to Jan 2021.
Role: Big Data Developer
Description:

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

Roles & 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 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 Spark improving the performance and optimization of 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.
● 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 the looker dashboard. In Looker Stakeholders will create their-own dashboards with processed final output data for business requirements
● Deploying the oozie workflows by using the automated Jenkins Tool.

❖ **#1 Project**: **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 data from different sources of RDBMS to store it in HDFS. This project mainly focuses on Daily Ingestion, Analytics Pipeline and Reporting Framework. To maintain the day to day into HDFS from different sources of RDBMS. After clubbing multiple tables, run Analytics Pipeline for Stakeholders. Finally Moving data into Redshift for Reporting Purpose.

Roles & 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 formats like ORC, Parquet format.
● Scheduling Sqoop jobs using Crontab.
● To Check 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 stakeholder's requirements.
● Scheduled jobs and transferred final output data into Redshift for reporting purposes.
● Creating cluster group in EMR for running Daily ingestion, Analytics Pipeline jobs.

Education

Bachelor of Engineering: Mechanical Engineering **2012-2016**
Anna University - Chennai

Accomplishments

- Awarded "Best Employee Award" Appreciation Certification in Autodesk India Pvt Ltd for the quarter ending Jan'2019
- Awarded "Impact Go Above 250" for inspiring a high-performance culture in Legato Health care for the quarter ending Sep'2021
- Awarded “Impact Go Above Award” Appreciation Certification in Legato Health Care for the Quarter ending March 2022.