**Saranya**

[**Tel:+1**](Tel:+1) **617 701 7267**

**Email:saranyam3000@gmail.com**

**Sr. Big Data Engineer/ Data Engineer Consultant**

**PROFESSIONAL SUMMARY**

* Having 9+ years of overall experience as software developer in design, development, deploying, and large scale supporting large scale distributed systems.
* In depth understanding/knowledge of Hadoop Architecture and various components such as HDFS, MR, Hadoop GEN2 Federation, High Availability and YARN architecture and good understanding of workload management, scalability, and distributed platform architectures.
* Assisted Deployment team in setting up Hadoop cluster and services.
* Performeddataengineeringfunctions:dataextraction,transformation,loading,andintegrationtosupportenterprisedatainfrastructures-datawarehouse,operationaldatastores,andmasterdatamanagement.
* Good experience in creating and designing data ingest pipelines using technologies such as ApacheStorm-KafkaandExperiencedinwritingliveReal-timeProcessingusingSparkStreamingwithKafkaasadatapipelinesystem.
* StrongprogrammingcapabilityusingPythonandHadoopframeworkutilizingClouderaHadoopEcosystemprojects(HDFS,Hadoop,Sqoop,Hive,HBase,Oozie,Impala,Zookeeper,etc.)
* Practicing consulting on Snowflake Data Platform Solution Architecture, Design, Development, and deployment focused to bring the data driven culture across the enterprises.
* Experienced in Agile development, specification, and performance tuning on Oracle databases byleveragingexplainplans,tuningSQLqueries,andexcellentexperienceinwritingComplexSQLqueries.

to validatedatamovementbetween differentlayers in thedatawarehouseenvironment.

* Expert in Building reports using SQL Server Reporting Services (SSRS, Crystal Reports, Power BI, andBusiness Objects.
* experience in dimensional modeling such as star schema, snowflake schema, creating facts, dimensions, and measures.
* Experience in designing a Terraform and deploying it in cloud deployment manager to spin up resources like cloud virtual networks, Compute Engines in public and private subnets along with AutoScaler in Google Cloud Platform.
* Experience in Designing, Architecting, and implementing scalable cloud-based web applications using AWS.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the AWS EMR Spark cluster.
* Highly Skilled at Python coding using SQL, NumPy, Pandas, and Spark/Pyspark for Data Analysis and Modelbuilding,deploying,andoperatinghighlyavailable,scalable,andfault-tolerantsystemsusingAmazonWebServices(AWS).
* Hands-onexperience indevelopingwebapplicationsimplementingModelViewControlarchitectureusingDjango, Flask web application frameworks.
* Proficient in Hive Query language and experienced in Hive performance optimization.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data / Hadoop Eco System** | Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Apache Flume, Apache Storm, Apache Airflow, HBase |
| **Programming Languages** | Java, Python, Scala, PySpark, C, C++ |
| **Cluster Mgmt& monitoring** | CDH 4, CDH 5, Horton Works Ambari 2.5 |
| **Data Bases** | MySQL, SQL Server, Oracle 10g/11g/12c, MS Access |
| **NoSQL Data Bases** | MongoDB, Cassandra, HBase, KairosDB |
| **Workflow mgmt tools** | Oozie, Apache Airflow |
| **Visualization & ETL tools** | Tableau, BananaUI, D3.js, Informatica, Talend |
| **Cloud Technologies** | Azure, Kubernetes, AWS(S3, EC2, EMR, Kinesis, Firehose) |
| **IDE’s** | Eclipse, Jupyter notebook, Spyder, PyCharm, IntelliJ |
| **Version Control Systems** | Git, SVN |
| **Operating Systems** | Unix, Linux, Windows |

**PROFESSIONAL EXPERIENCE**

**CareSource - Dayton, OH**

**Sr. Big Data Engineer/ Data EngineerAug 2022 – Current**

* Hands - on experience in Azure Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake.
* Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
* Have good experience working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Extract, transform and load data from source systems to Azure Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL of Azure Data Lake Analytics.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL DB, Azure SQL DW), and processing the data in Azure Databricks.
* Have experience of working on Snow-flake data warehouse.
* Used Hadoop scripts for HDFS (Hadoop File System) data loading and manipulation.
* Performed Hive test queries on local sample files and HDFS files.
* Used Spark Streaming to divide streaming data into batches as an input to spark engine for batch processing.
* Worked on analyzing Hadoop cluster and different Big Data analytic tools including Pig, hive, HBase, Spark and Sqoop.
* Creating Stored Procedure and Scheduled them in Azure Environment
* Experience in using SSIS tools like Import and Export Wizard, Package Installation, and SSIS Package Designer.
* Experience in ETL processes involving migrations and in sync processes between two databases.
* Analyzed Data Profiling Results and Performed Various Transformations.
* Hands on Creating Reference Table using Informatica Analyst tool as well as Informatica Developer tool.
* Hands-on experience with Snowflake utilities, Snow SQL, SnowPipe, Big Data model techniques using Python.
* ETL pipelines in and out of data warehouse using combination of Python and Snowflakes Snow SQL Writing SQL queries against Snowflake.

**Elevance Health - Norfolk, VirginiaMar 2019 to Jul 2022**

**Sr. Data engineer**

* Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
* Worked on closely with the business analysts to convert the Business Requirements into Technical Requirements and prepared low and high-level documentation.
* Experience in performing transformations using Hive, MapReduce, hands on experience in copying .log, snappy files into HDFS from Greenplum using Flume and Kafka, loaded data into HDFS and extracted the data into HDFS from MYSQL using Sqoop.
* Imported required tables from RDBMS to HDFS using Sqoop and used Storm/ Spark streaming and Kafka to get real time streaming of data into HBase.
* Worked extensively with importing metadata into Hive and migrated existing tables and applications to work on Hive and AWS cloud.
* Used AWS Redshift, S3, Spectrum and Athena services to query large amount data stored on S3 to create a Virtual Data Lake without having to go through ETL process.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Exported data from HDFS to RDBMS via Sqoop for Business Intelligence, visualization, and user report generation.
* Developed views and templates with Python and Django's view controller and templating language to create a user-friendly website interface.
* Experience in Writing Map Reduce jobs for text mining and worked with predictive analysis team and Experience in working with Hadoop components such as HBase, Spark, Yarn, Kafka, Zookeeper, PIG, HIVE, Sqoop, Oozie, Impala and Flume.
* Wrote HIVE UDF's as per requirements and to handle different schemas and xml data.
* Developed data pipeline using Python, Hive to load data into data link. Perform data analysis data mapping for several data sources.
* 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.
* 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.
* Designed new Member and Provider booking system which allows providers to book new slots, with sending out the member leg and provider Leg directly to TP through DataLink.
* Open SSH tunnel to Google DataProc to access to yarn manager to monitor spark jobs.
* Analyze various type of raw file like Json, Csv, Xml with Python using Pandas, NumPy etc.
* Developed Spark applications using Scala for easy Hadoop transitions. And Hands on experienced in writing Spark jobs and Spark streaming API using Scala and Python.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
* Installed Oozie workflow engine to run multiple Hive and Pig jobs.
* Designed and developed User Defined Function (UDF) for Hive and Developed the Pig UDF’S to pre-process the data for analysis as well as experience in (UDAFs) for custom data specific processing.
* Implement One time Data Migration of Multistate level data from SQL server to Snowflake by using Python.
* Day to-day responsibility includes developing ETL data pipelines in and out of data warehouse, develop major regulatory and financial reports using advanced SQL queries in snowflake.
* Designed and developed the core data pipeline code, involving work in Python and built on Kafka and Storm.
* AWS CI/CD data pipeline and AWS data lake using EC2, Glue and Lambda.
* Goal to build an enterprise datahub on amazon s3 storage platform to able to provide better enterprise-wide data services, analytics.
* Developing CI/CD pipelines to automate build and deploy to Dev, QA and production environments.
* Good experience on Partitions, bucketing concepts in Hive and designed both Managed and External tables in Hive for optimized performance.
* Managed the PostgreSQL database that stores metadata and state information for Airflow, which includes creating and maintaining tables, managing backups and restores, and optimizing performance.
* Used Excel formulas, pivot tables, etc. to present reports for ease of understanding to engineers and analysts; used PowerPivot to create quick I dashboards linked to DB2 tables and flat files.
* Created cloud-based software solutions written in Scala Spray IO, Akka, and Slick.
* Experience on fetching the live stream data from DB2 to HBase table using Spark Streaming and Kafka.
* Worked on NoSQL databases including HBase and Cassandra.
* Populated HDFS and Cassandra with huge amounts of data using Apache Kafka.

**Fintech Financial Information Services – Jacksonville, FL**

**Hadoop / Data Engineer Nov 2016 – Feb 2019**

* Designed solution for Streaming data applications using Apache Storm.
* Create and maintain optimal data pipeline architecture in cloud Microsoft Azure using Data Factory and Azure Databricks
* Create an Architectural solutions that leverages the best Azure analytics tools to solve our specific need in Chevron use case.
* Design and present technical solutions to end users in a way the is easy to understand and buy into
* Educating client/business users on the pros and cons of various Azure PaaS and SaaS solutions ensuring the most cost-effective approaches are taken into consideration.
* Worked in a real-time data streaming data using AWS kinesis, EMR and AWS Glue.
* Transfer data in logical stages from System of records to raw zone, refined zone and produce zone for easy translation and denormalization.
* Create custom logging framework for ETL pipeline logging using Append variables in Data factory.
* Enabling monitoring and azure log analytics to alert support team on usage and stats of the daily runs
* Took proof of concept projects ideas from business, lead, developed and created production pipelines that deliver business value using Azure Data Factory
* Kept our data separated and secure across national boundaries through multiple data centers and regions.
* Implement Continuous integration/continuous development best practice using Azure Devops, ensuring code versioning.
* Worked in a SAFE (Scaled Agile Framework) team with daily standups, sprint planning, quarterly planning.
* Worked on creating tabular models on Azure analysis services for meeting business reporting requirements.
* Have good experience working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).

**Evoke Technologies - Hyderabad, India Jun 2013 to Apr2016**

**Data/ Cloud Engineer**

* Designed database Model, APIs and Views utilizing Python to build an interactive web - based application.
* Thoroughly used Python Libraries Beautiful Soup, NumPy, Pandas data frame, Matplotlib, python-twitter, urllib2.
* Developed new functionality keeping in mind OOP Principles, performance, scalability, and robustness deployed within Docker containers.
* Used python libraries like Beautiful Soap, NumPy and SQL Alchemy and Wrote Python scripts to parse JSON documents and load the data in database.
* Designed and managed API system deployment using fast http server and Amazon AWS architecture.
* Worked in DevOps group running Jenkins in a Docker container with EC2 slaves in Amazon AWS cloud configuration. Also gained familiarity with supporting technologies like Kubernetes and Mesos.
* Integrated services like Bitbucket AWS code pipeline to create deployment pipeline.
* Build Cassandra queries for performing various CRUD operations like create, update, read and delete, also used Bootstrap as a mechanism to manage and organize the html page layout.
* Used Ansible, Vagrant, and Docker for managing the application environments.
* Used regular expression for faster search results in combination with Angular2 built-in and custom pipes.
* Mentor and guide analyst on building purposeful analytics tables in dbt for cleaner schemas.
* Worked on user controls, master pages for data caching and used ADO to implement data layer to help communicate with database.
* Working in multithreading, AWS Serverless Lambda, AWS Serverless Lambda, Node JS, S3, DynamoDB, CloudWatch environment for execution Golang of multiple orders in the trading platform.
* Developed backend web services using Node.js and stored dependencies using Node Package Manager (NPM).
* Managed, developed, and designed a dashboard control panel for customers and Administrators using Django, Oracle DB, PostgreSQL, and VMWare API calls.
* Used SOAP and Restful API for information extraction.
* Added support for Amazon AWSS3 and RDS to host static/media files and the database into Amazon Cloud.
* Deployer async jobs monitoring system using Celery flower and deployed an API that asynchronously distributes tasks using RabbitMQ and Celery.
* Designed python scripts to load transformed data into cloud services (AWS, Azure, GCP).
* Used several python libraries like python as well as SVN implementation also involving GCP.
* Engineered a data processing pipeline with GCP, API, CSS to preprocess large scale.
* Automated the existing scripts for performance calculations using NumPy and SQL Alchemy.
* Used Git source control to manage simultaneous development.
* Created the Python Django forms to record data of online users and used PyTest for writing test cases.