**Santosh santhoshdataengineer89@gmail.com**

**(719) 960-0309**

**Senior Big Data Engineer**

**Professional Summary:**

* Over 8+ years of professional experience in IT, working with various Legacy Database systems, which includes work experience in **Big Data** technologies as well.
* Hands on experience on complete Software Development Life Cycle **SDLC** for the projects using methodologies like Agile and hybrid methods.
* Expert in using **Python** for **Data Engineering** and **Modeling**.
* Excellent working experience on **Amazon Web Services**, **Azure Services and** Management.
* Data importing and exporting by using **Sqoop** from **HDFS** to Relational Database Systems and vice-versa.
* Proficient in data mining tools like **R, SAS, Python, SQL Alchemy, Excel, BigDataHadoop** eco-systems.
* Good knowledge on converting complex **RDMS (Oracle, MySQL & Teradata**) queries into **Hive** query language.
* Transforming and retrieving the data by using **Pig, Hive, SSIS** and **Map Reduce**.
* Excellent Knowledge of **RalphKimball** and **BillInmon's** approaches to **Data Warehousing**.
* Extensive experience in Text Analytics, developing different Statistical **Machine Learning**, **Data Mining solutions** to various business problems and generating data visualizations using **R, Python**.
* Good knowledge in **OLAP, OLTP**, **Business Intelligence** and **Data Warehousing** concepts with emphasis on **ETL** and Business Reporting needs.
* Excellent knowledge of **MicrosoftOffice** with an emphasis on **Excel**.
* Hands on experience for Data Integrity process and **Data Modelling** concepts.
* Experience in Data mining with querying and mining large datasets to discover transition patterns and examine financial reports.
* An excellent experience in generating ad-hoc repots using **Crystal Report**
* Experience in Performance Tuning and query optimization techniques in transactional and **Data Warehouse** Environments.
* Responsible for designing and building a **Data Lake** using **Hadoop** and its ecosystem components.
* Strong background in various **Data modeling** tools using **ERWIN**, **ER Studio** and **PowerDesigner**.
* Expertise in **DataReporting**, **Ad-hocReporting**, Graphs, Scales, PivotTables and **OLAP** reporting.
* Hands on experience on data modeling with **Star schema** and **Snowflake schema.**
* Perform structural modifications using **Map - Reduce, HIVE** and analyze data using visualization/ reporting tools.
* Experience in developing customized **UDF's** in **Python** to extend **Hive** and **PigLatin** functionality.
* Experience in extracting source data from Sequential files, **XML** files, **CSV** files, transforming and loading it into the target **Data warehouse**.
* Good experience with use-case development, with Software methodologies like **Agile** and **Waterfall.**
* Proven ability to manage all stages of project development Strong Problem Solving and Analytical skills and abilities to make Balanced and Independent Decisions

**Technical Skills:**

* **Languages:** SQL, PL/SQL, PYTHON, Java, Scala, R, T-SQL, Unix, Linux
* **Operating Systems:** UNIX, Windows, Linux
* **Concepts and Methods:** Business Intelligence, Data Warehousing, Data Modeling, Requirement Analysis
* **ETL Tools:** Streamsets, AWS Redshift-snowflake Matillion, Alteryx, Informatica PowerCenter, Ab Initio
* **Big Data:** HDFS, Map Reduce, Spark, Airflow, Yarn, NiFi, HBase, Hive, Pig, Flume, Sqoop, Kafka, Oozie, Hadoop, Zookeeper, Spark SQL.
* **Cloud Platform:** AWS (Amazon Web Services), Microsoft Azure
* **RDBMS:** Oracle 9i/10g/11g/12c, Teradata, My SQL, MS SQL
* **NO SQL:** MongoDB, HBase, Cassandra
* **Data Modeling Tools:** ERwin, Power Designer, Embarcadero ER Studio, IBM Rational Software Architect, MS Visio, ER Studio, Star Schema Modeling, Snowflake Schema Modeling, Fact and Dimension tables
* **Application Servers:** Apache Tomcat, Web Sphere, Web logic, JBoss
* **Other Tools:** Azure Databricks, Azure Data Explores, Azure HDInsight, Power BI
* **Data warehouses (SaaS):** Amazon Redshift, snowflake – SnowPipe, SnowSQL, etc.

**Project Work Experience:**

**Senior Big Data Engineer Jun 2020 to till now**

**Client: Travel Port, Englewood, CO**

* Developed **HIVE UDFs** to incorporate external business logic into **Hive** script and developed join data set scripts using **HIVE** join operations.
* Lead Migration activities from On Premise to cloud warehouse Snowflake on AWS Platform including database and models with all securities in place around data in wire as well as at rest with all the best practices in place. Created DevOps process, application infrastructure and deployed application codes utilizing AWS-based cloud services for HA/DR strategies.
* Migrate data into RV Data Pipeline using **Databricks, Spark SQL** and **Scala**.
* Streamed the Data from SQLserver to AWS Elastic Search service using the Streaming ETL tool Streamsets for the search, fuzzy search capabilities.
* Involved in creating **HDInsight** cluster in **Microsoft Azure Portal** also created **Events hub** and **Azure SQL Databases.**
* Configured **Spark streaming** to get ongoing information from the **Kafka** and store the stream information to **HDFS**.
* Extracted and updated the data into **HDFS** using **Sqoop** import and export.
* Worked with the **Spark** for improving performance and optimization of the existing algorithms in **Hadoop** using **Spark Context, Spark-SQL, PySpark, Impala,** **Tealeaf, Pair RDD's, DevOps, Spark YARN**
* Worked with **PowerShell** and **UNIX** **scripts** for file transfer, emailing and other file related tasks.
* Delivered de normalized data for **Power BI** consumers for modeling and visualization from the produced layer in **Data lake**
* Developed and designed **data integration** and **migration solutions** in **Azure**.
* 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
* Installed **Kafka** Producer on different severs and Scheduled to produce data for every 10 seconds Implemented Data Quality in **ETL Tool Talend** and having good knowledge in Data Warehousing Developed **Apache Spark** applications by using spark for data processing from various streaming sources.
* Took proof of concept projects ideas from business, lead, developed and created production pipelines that deliver business value using **Azure Data Factory**
* Good Exposure on **Map Reduce** programming using Java, PIG Latin Scripting and Distributed Application and **HDFS**.
* Worked with **NoSQL databases** like **HBase** in creating **HBase** tables to load large sets of semi-structured data coming from various sources
* Create **Spark** Vectorized panda user defined functions for data manipulation and wrangling
* Transfer data in logical stages from System of records to raw zone, refined zone and produce zone for easy translation and de normalization
* Create and maintain optimal data pipeline architecture in cloud **Microsoft Azure** using **Data Factory** and **Azure Databricks**
* 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**
* Create custom logging framework for **ELT** pipeline logging using Append variables in **Data factory**
* Developed various **Oracle SQL scripts, PL/SQL packages**, procedures, functions, and java code for data
* Strong Knowledge on architecture and components of Tealeaf, and efficient in working with **Spark Core,** **SparkSQL.** Designed and developed RDD Seeds using Scala and Cascading. Streaming data to **Spark** streaming using **Kafka**
* Built real time pipeline for streaming data using **Events hub/Microsoft Azure Queue** and **Spark** **streaming**.
* Responsible to manage data coming from different sources through **Kafka**
* Setting up **Azure infrastructure** like storage accounts, integration runtime, service principal id, app registrations to enable scalable and optimized utilization of business user analytical requirements in **Azure**.
* Worked on a clustered **Hadoop** for **Windows Azure** using **HDInsight** and **Hortonworks** Data Platform for Windows.
* Writing **PySpark** and spark **Sql transformation** in **Azure Databricks** to perform complex transformations for business rule implementation
* Implemented **Kafka** producers create custom partitions, configured brokers and implemented High level consumers to implement data platform.
* Implement IOT streaming with **Databricks Delta tables** and **Delta lake** to enable ACID transaction logging
* Exposed transformed data in **Azure Spark Databricks** platform to parquet formats for efficient data storage
* Creating Data factory pipelines that can bulk copy multiple tables at once from relational database to **Azure data lake gen2**
* Used **Airflow** to monitor and schedule the work
* Utilized **Ansible playbook** for code pipeline deployment
* Used **Delta Lake** as it is an open-source data storage layer which delivers reliability to **data lakes.**
* Worked in a SAFE (Scaled Agile Framework) team with daily stand-ups, sprint planning, quarterly planning.

**Environment:**  Hadoop, Spark, MapReduce, Kafka, Scala, JAVA, Azure Data Factory, Data Lake, Databricks, Azure DevOps, PySpark, Agile, Power BI, Python, R, PL/SQL, Oracle 12c, SQL, No SQL, HBase, Scaled Agile team environment

**Big Data Engineer Oct 2017 to Jun 2020**

**Client: QVC, West Chester, PA**

* Used Agile methodology in developing the application, which included iterative application development, weekly Sprints, stand up meetings and customer reporting backlogs.
* Optimized **Map Reduce** Jobs to use **HDFS** efficiently by using various compression mechanisms.
* Development of efficient **pig** and **hive** scripts with joins on datasets using various techniques.
* Automating the ETL tasks and data work flows for the data pipeline of the ingest process through **UC4** scheduling tool.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, Spark and loaded data into **HDFS.**
* Developed UDFs in **Java** as and when necessary to use in **PIG** and **HIVE queries.**
* Automated the cloud deployments using chef, python and **AWS Cloud Formation** Templates.
* Design and build ETL workflows, leading the efforts of programming data extraction from various sources into Hadoop file system, implement end to end ETL workflows using **Teradata, SQL, TPT, SQOOP** and load to **HIVE** data stores.
* Monitoring resources and Applications using **AWS Cloud Watch,** including creating alarms to monitor metrics such as **EBS, EC2, ELB, RDS, S3, SNS** and configured notifications for the alarms generated based on events defined.
* Solved performance issues in **Hive** and **Pig** scripts with understanding of Joins, Group and aggregation and how does it translate to **MapReduce** jobs.
* Used ORC and Parquet file formats in **Hive**.
* Understanding of AWS Product and Service suite primarily **EC2, S3, VPC, Lambda, Redshift, Spectrum,**
* **Athena, EMR(Hadoop)** and other monitoring service of products and their applicable use cases, best practices and implementation, and support considerations
* Experience in change implementation, monitoring and troubleshooting of **AWS Snowflake** databases and cluster related issues
* Analyze and develop programs by considering the extract logic and the data load type using Hadoop ingest processes using relevant tools such as **Sqoop, Spark, Scala, Kafka, Unix shell scripts** and others.
* Design the incremental, historical extract logic to load the data from flat files into Massive Event Logging Database (MELD) from various servers.
* Writing technical design document based on the data mapping functional details of the tables.
* Extracting batch and Real time data from **DB2, Oracle, Sql server, Teradata, Netezza** to Hadoop (HDFS) using **Teradata TPT, Sqoop, Apache Kafka, Apache Storm.**
* Assist with the analysis of data used for the **tableau** reports and creation of dashboards.
* Design and implement large scale distributed solutions in **AWS.**
* Developing **Apache Spark** jobs for data cleansing and pre-processing.
* Writing spark programs to improve the performance and optimization of the existing algorithms in Hadoop using **spark context, spark-sql, data frame, pair RDD's, spark yarn.**
* Using **Scala** language to write programs for faster testing and processing of data.
* Writing code and creating **hive** jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Created and managed cloud VMs with **AWS EC2** Command line clients and AWS management console.
* Migrated on premise database structure to Confidential Redshift data warehouse. Worked on AWS Data Pipeline to configure data loads from **S3** into **Redshift**
* Created functions and assigned roles in **AWS Lambda** to run python scripts, and **AWS Lambda** using **java** to perform event driven processing. Created **Lambda** jobs and configured Roles using **AWS CLI.**
* Write documentation of program development, subsequent revisions and coded instructions in the project related **GitHub** repository.
* Deployment support including change management and preparation of deployment instructions.
* Prepare release notes, validation document for user stories to be deployed to production as part of release.

**Environment**: RHEL, HDFS, Map-Reduce, Hive, AWS, EC2, S3, Lambda, Redshift, Pig, Sqoop, Oozie, Teradata, Oracle SQL, UC4, Kafka, GitHub, Hortonworks data platform distribution, Spark, Scala.

**Big Data Engineer Nov 2016 to Sep 2017**

**Client: MetLife, Cary, NC**

* Developed SSRS reports, SSIS packages to Extract, Transform and Load data from various source systems
* Implementing and Managing ETL solutions and automating operational processes.
* Was responsible for **ETL** and data validation using **SQL Server Integration Services.**
* Measured Efficiency of Hadoop/Hive environment ensuring SLA is met
* Optimized the **Tensor Flow** Model for efficiency
* Created Entity Relationship Diagrams (ERD), Functional diagrams, Data flow diagrams and enforced referential integrity constraints and created logical and physical models using **Erwin.**
* Defined and deployed monitoring, metrics, and logging systems on **AWS.**
* and **Spark SQL** for ETL jobs and using the right technology for the job to get done.
* Compiled data from various sources to perform complex analysis for actionable results
* Strong understanding of **AWS** components such as **EC2** and **S3**
* Managed security groups on **AWS,** focusing on high-availability, fault-tolerance, and auto scaling using Terraform templates. Along with Continuous Integration and Continuous Deployment with **AWS Lambda** and AWS code pipeline.
* Defined facts, dimensions and designed the data marts using the Ralph Kimball's Dimensional Data Mart modeling methodology using Erwin.
* Worked on Big data on **AWS** cloud services i.e**. EC2, S3, EMR** and **DynamoDB**
* Created ad hoc queries and reports to support business decisions **SQL Server Reporting Services (SSRS).**
* Analyze the existing application programs and tune **SQL queries** using execution plan, query analyzer, SQL Profiler and database engine tuning advisor to enhance performance.
* Optimizing and tuning the **Redshift** environment, enabling queries to perform up to 100x faster for **Tableau** and SAS Visual Analytics.
* Migrated on premise database structure to Confidential Redshift data warehouse

**Environment:** AWS, EC2, S3, SQL Server, Erwin, Oracle, Redshift, Informatica, RDS, NOSQL, Snowflake Schema, MySQL, Dynamo DB, Docker, PostgreSQL, Tableau, Git Hub

**Data Engineer Mar 2015 to Aug 2016**

**Client: Way2Online, Hyderabad, India**

* Written **Hive** jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Worked on **Oozie workflow** engine for job scheduling.
* Worked with **Avro Data** Serialization system to work with JSON data formats.
* Worked on various performance optimizations like using distributed cache for small datasets, partition and bucketing in **Hive**, doing map side joins etc.
* Performed pre-processing on a dataset prior to training, including standardization, normalization.
* Created pipelines to create a processing pipeline including transformations, estimations, evaluation of analytical models.
* **Hive Context**, with transformations and actions (**map, flat Map, filter, reduce, reduce** by Key).
* Developed **PIG scripts** for the analysis of semi structured data.
* Migrated **ETL jobs** to **Pig scripts** to do Transformations, even joins and some pre-aggregations before storing the data onto **HDFS**.
* Worked on different file formats like Sequence files, XML files and Map files using **MapReduce** Programs.
* Developed **PIG UDF'S**for manipulating the data according to Business Requirements and also worked on developing custom **PIG Loaders.**
* Worked with heterogeneous source to Extracted data from **Oracle database, XML** and flat files and loaded to a relational **Oracle warehouse**.
* Evaluated model accuracy by dividing data into training and test datasets and computing metrics using evaluators.
* Performed tuning of **SQL queries** and Stored Procedure for speedy extraction of data to resolve and troubleshoot issues in **OLTP** environment.

**Environment:** Hive, HDFS, Oozie, Map Reduce, Oracle 10g, SQL, OLTP, Windows, MS Office

**Database Developer July 2013 to Feb 2015**

**Client: CTS, Hyderabad, India**

* Analyzed requirement and impact by participating in Joint Application Development sessions with business client online.
* Developed complex parameterized reports which were used for making current and future business decisions.
* Performed and automated **SQL Server** version upgrades, patch installs and maintained relational databases.
* Performed front line code reviews for other development teams.
* Created and modified various Stored Procedures used in the application usin**g T-SQL.**
* Modified and maintained **SQL Server** **stored procedures, views, ad-hoc queries**, and **SSIS** packages used in the search engine optimization process.
* Updated existing and created new reports using **Microsoft SQL Server** Reporting Services. Team consisted of 2 developers.
* Gathered business requirements and converted them into new **T-SQL** stored procedures in visual studio for database project.
* Performed unit tests on all code and packages.
* Scheduled and maintain packages by daily, weekly and monthly using **SQL Server** Agent in SSMS.
* Monitored and tuned database resources and activities for **SQL Server databases.**

**Environment:** MSSQL, Stored Procedures, Views, ad-hoc queries, SSIS, MS Office, T-SQL, Windows.