Sai Teja

**EMAIL:** datasai777@gmail.com

**PHONE:**

**Location:** NJ

**Skills:**

**Operating Systems:** Windows (7/10), Mac (10.4/10.5/10.6), Linux (Red Hat), Ubuntu

**Databases:** Oracle, MS SQL Server, My SQL, Redshift, Snowflake.

**Data Modeling:** Star Schema, Snowflake.

**Reporting Tool:** Tableau, Power BI.

**Scheduling Tools:** Autosys

**Languages:** Python, Java, R, Microsoft SQL Server, Oracle PLSQL, Splunk.

**Hadoop and Big Data Technologies:** HDFS, MapReduce, Flume, Sqoop, Pig, Hive, Morphline, Kafka, Oozie, Spark, Nifi, Zookeeper, Elastic Search, Apache Solr, Talend, Cloudera Manager, R Studio, Confluent, Grafana.

**NoSQL:** HBase, Couchbase, Mongo, Cassandra

**Web Services:** XML, SOAP, Rest APIs

**Web Development Technologies:** JavaScript, CSS, CSS3, HTML, HTML5, Bootstrap, XHTML, JQUERY, PHP

**Databases:** Oracle, DB2, MS-SQL Server, MySQL, MS-Access, Teradata

**Build Tools:** Maven, Scala Build Tool (SBT), Ant

**IDE Development Tools:** Eclipse, Net Beans, IntelliJ, R Studio

**Programming and Scripting Languages:** C, SQL, Python, C++, Shell scripting, R

Work Experience

* 8 years of IT experience in Data warehousing with emphasis on Business Requirements Analysis, Application Design, Development, testing, implementation and maintenance of client/server Data Warehouse and Data Mart systems
* Expertise in Hadoop ecosystem components such as Spark, HDFS, Map Reduce, Yarn, HBase, Pig, Sqoop, Flume, Oozie, Impala, Zookeeper, Hive, NiFi and Kafka for scalability, distributed computing, and high-performance computing.
* Excellent understanding of Hadoop architecture, Hadoop daemons and various components such as HDFS, YARN, Resource Manager, NodeManager, NameNode, Data Node and MapReduce programming paradigm.
* Good understanding of Apache Spark, Kafka, Storm, Nifi, Talend, RabbitMQ, Elastic Search, Apache Solr, Splunk and BI tools such as Tableau.
* Knowledge of Hadoop administration activities using Cloudera Manager and Apache Ambari.
* Experience working with Cloudera, Amazon Web Services (AWS), Microsoft Azure and Hortonworks
* Worked on Import and Export of data using Sqoop from RDBMS to HDFS.
* Have good knowledge in Containers, Docker and Kubernetes for the runtime environment for the CI/CD system to build, test, and deploy.
* Hands on experience in loading data (Log files, Xml data, JSON) into HDFS using Flume/Kafka.
* Experience in pyspark programming language with Spark Core and Spark modules extensively.
* Built ETL data pipelines using Python/MySQL/Spark/Hadoop/Hive/UDFs
* Experience in analyzing data using Hive QL, Pig Latin, HBase, Spark, R Studio and custom Map Reduce programs in python. Extending Hive and Pig core functionality by writing custom UDFs.
* Used packages like Numpy, Pandas, Matplotlib, Plotly in python for exploratory data analysis.
* Hands on experience with cloud technologies such as Azure HDInsight, Azure Data Lake, AWS EMR, Athena, Glue and S3.
* Good knowledge in using Apache NiFi to automate the data movement between different Hadoop systems.
* Experience in performance tuning by using Partitioning, Bucketing and Indexing in Hive.
* Experienced in job workflow scheduling and monitoring tools like Airflow, Oozie, TWS, Control-M and Zookeeper.
* Flexible working Operating Systems like Unix/Linux(Centos, Redhat, Ubuntu) and Windows Environments.
* Hands on development experience with RDBMS, including writing complex SQL scripting, Stored procedure, and triggers.
* Experience in writing Complex SQL Queries involving multiple tables inner and outer joins.
* dimensional
* modelling for OLAP.
* using SQL
* Loader.

**Work Experience**

**Data Engineer**

**Allstate, Chicago, Illinois**

**Oct 2018 – Till Date**

* Executed all phases of Big Data project lifecycle starting from Scoping Study, Requirements gathering, Estimation, Design, Development, Implementation, Quality Assurance and Application Support.
* Working on building frameworks for data curation pipelines using Spark and Hive, and migrating Hive based applications to Spark.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Designed and built data processing applications using Spark on AWS EMR cluster which consumes data from AWS S3 buckets, apply necessary transformations and store the curated business ready datasets onto Snowflake tables.
* Involved in design and analysis of the issues and providing solutions and workarounds to the users and end-clients
* Extensively worked on developing Spark jobs in Python (Spark SQL) using Spark APIs
* Involved in performing Data Screening and Profiling by Accuracy Checks, fixing Missing Data and Outliers removal, examining historical data, detecting patterns/correlations or relationships in the data, and then extrapolating these relationships forward in time
* Involved in performing Exploratory Data Analysis (EDA), Hypothesis Testing and Predictive Analysis using R/R Studio to analyze the customer behavior.
* Experience in writing PySpark scripts and a wrapper shell scripts to automate data validations
* Experience in orchestrating and building schedules/workflows on Tivoli Workload Scheduler (TWS) and Oozie in the environment.
* Developed functionality to perform auditing and threshold checks for error handling for smooth and easier debugging and data profiling
* Built visualizations using the tool, Looker on top of the business ready datasets loaded in Snowflake.
* Worked on preparing test cases for unit testing for development
* Involved in creating Hive tables, loading data in ORC, JSON, CSV format and writing hive queries to analyze data using Spark-SQL
* Build data quality framework to run data rules that can generate reports and send emails of business critical successful and failed job notifications to business users daily.
* Built solution design and implemented Data Quality monitoring and reporting framework in PySpark
* Built pipelines to send data extracts and reports over Data Router, SFTP and to AWS S3 buckets

**Data Engineer**

**Gilead Sciences, CA**

**June 2016 – Sep 2018**

* Involved in analyzing business requirements and prepared detailed specifications that follow project guidelines required for project development.
* Communicate regularly with business and I.T leadership.
* Built and Deployed jobs using Airflow.
* Responsible for data extraction and data ingestion from different data sources into S3 by creating ETL pipelines using Spark and Hive.
* Used Pyspark for data frames, ETL, Data Mapping, Transformation and Loading in complex and high-volume environment
* Extensively worked with pyspark / Spark SQL for data cleansing and generating data frames and RDDs.
* Co-ordinated with the other team members to write and generate test scripts, test cases for numerous user stories.
* Pandas to calculate the moving average and RSI score of the stocks and generated them into data warehouse.
* Worked on EMR clusters of AWS for processing Big Data across a Hadoop Cluster of virtual servers.
* Developed Spark Programs for Batch Processing.
* Developed Spark code using python for pyspark/Spark-SQL for faster testing and processing of data.
* Involved in design and analysis of the issues and providing solutions and workarounds to the users and end-clients.
* Designed and built data processing applications using Spark on AWS EMR cluster which consumes data from AWS S3 buckets, apply necessary transformations and store the curated business ready datasets onto Snowflake analytical environment.
* Developed functionality to perform auditing and threshold checks for error handling for smooth and easier debugging and data profiling.
* Build data quality framework to run data rules that can generate reports and send emails of business critical successful and failed job notifications to business users daily.
* Used spark to build tables that require multiple computations and non equi-joins.
* Scheduled various spark jobs for daily and weekly.
* Modelled Hive partitions extensively for faster data processing.
* Implemented various udfs in python as per the requirement.
* Used Bit Bucket to collaboratively interact with the other team members.
* Involved in Agile methodologies, daily scrum meetings and sprint planning with business users in gathering, analyzing and documenting the business requirements and translate them into technical specifications.

**SQL server Developer**

**IBM, HYD, India**

**Jan 2014 – April 2015**

* Designed and developed a custom database (Tables, Views, Functions, Procedures, and Packages).
* Monitoring existing SQL code and performance Tuning if necessary.
* Extensively involved in new systems development with Oracle 6i.
* Interact with business analysts to develop modeling techniques.
* USED SQLCODE returns the current error code from the error stack SQLERRM returns the error message from the current error code.
* Used Import/Export Utilities of Oracle.
* Wrote UNIX Shell Scripts to automate the daily process as per the business requirement.
* Writing Tuned SQL queries for data retrieval involving Complex Join Conditions.
* Use of EXPLAIN PLAN, ANALYZE, HINTS to tune queries for better performance and also extensive Usage of Indexes.
* Read data from flat files and load into Database using SQL\*Loader.
* Created the External Tables in order to load data from flat files and PL/SQL scripts for monitoring.

**Education**

Masters Stream & Year & University MS In IT University of Cumberland’s 2016

Bachelors Stream & Year & University B. tech in IT from JNTU HYD 2014