**SHANKARI SEETHALAKSHMI MOHANAKRISHNAN**

4698468345|saisri@mastroservices.com|[www.linkedin.com/in/shankari-sm](http://www.linkedin.com/in/shankari-sm)

**SUMMARY Leetcode :**<https://leetcode.com/u/SaiShankari/>

* 7+ years of work experience as Data professional, working with Pyspark, Python, SQL, Tableau.
* Experienced process-oriented Data Engineer and Analyst having excellent analytical, quantitative, and problem-solving skills using Spark, Python, SQL, Tableau, Big data technologies (SPARK,SQOOP, HIVE), CI/CD, GIT, Advanced Excel and R.
* Expertise in creating Data Warehousing including Metadata and Data Marts.
* Working knowledge on creating ETL work flow with PySpark using cloud technologies.
* Working knowledge on cloud technology like AWS & Azure.
* Created complex ETL data pipelines with PySpark by migrating data from multiple source systems to data warehouse.
* Expertise in creating complex PySpark transformations including data generation, data cleansing and manipulation, window functions, joins and date time.
* Working knowledge on creating ETL work flow with PySpark using cloud technologies.
* Created data pipelines with AWS technologies S3, Athena, redshift, Glue and AWS Lambda.
* Working knowledge on data warehousing concepts like SCD1, SCD2, star and snowflake schemas.
* Created and managed multiple intermediate tables (Apache Hive and Apache Iceberg) while building data pipeline.
* Experienced working as a part of EDG, EDW and EDL teams.
* Extensively worked on NumPy and Pandas library in python to transform data based on user requirements.
* Experience in creating control flow, data flow and transformations using the ETL tool SSIS.
* Knowledge in statistical concepts Descriptive & Inferential statistics (hypothesis testing, mean, median, mode, variance, standard deviation)
* Experience in creating control flow, data flow and transformations using the ETL tool SSIS & Adeptia.
* Hands on experience on creating dashboards with Tableau Desktop versions 7/8.2 and PowerBI.
* Experience writing complex SQL queries like analytical (rank, dense\_rank, rownumber) and windowing functions (lag, lead).
* Skilled in visualizing, manipulating, and analyzing large datasets and has ability of designing and developing effective reports. Proficient with MS Office Applicants, Excel (Pivot Table, Lookups, Index, Match, CountIf, dashboards, etc.)
* Experience working in agile environment with agile tools like JIRA.
* Experienced creating dashboards and user story to meet business requirements in Tableau.
* Built dashboards using techniques for guided analytics, interactive dashboard design, and visual best practices.
* Participated in cross-functional teams to reengineer and improve business processes.
* Extensive experience in developing SQL scripts to validate the databases tables and reports data for backend database testing.
* Strong Analytical, Interpersonal, Communication skills and excellent team player.
* Self-motivated with the ability to work both independently and in a team.
* Strong organizational skills. Ability to work well under pressure.
* Self-motivated and able to work on my own initiative.

**EDUCATION TABLEAU** - <https://public.tableau.com/app/profile/shankari.seethalakshmi.mohanakrishnan/vizzes>

**Illinois Institute of Technology, Master’s in information technology management (GPA-3.9/4.0)** 2016 –2018

**College of Engineering Guindy, Bachelors Information Technology (GPA-3.6/4.0)** 2011 – 2015

**SKILLS GITHUB -** <https://github.com/Shankari-Mohanakrishnan>

* **Database**: Oracle, MS Access, SQL Server, MySQL, Azure, Teradata, postgres.
* **Visualization**: Tableau, Power BI, Visio.
* **Statistics**: Descriptive Statistics, Inferential Statistics, Sampling Distribution, Central Limit Theorem, Hypothesis Testing.
* **Statistical Packages**: NumPy, pandas, matplotlib, seaborn & scikit-learn.
* **Analytic Tools**: Python, R, T-SQL, PL/SQL, Excel (Power Pivot, Pivot tables, lookup function) & VBA.
* **Big Data Tools & Techniques**: Hadoop (Map Reduce), HDFS, SQOOP, Hive, Spark,pySpark, ETL, Airflow, Jupyter Labs & Notebooks, PyCharm, VsCode.
* **Business Intelligence Tools**: SSIS, SSAS, SSRS.
* **Modelling Techniques**: Regression, Clustering and Classification.
* **Cloud Technology**: AWS, Azure
* **Programming Language**: Python, SQL, C, C++, Java.
* **Scripting Language**: HTML, XML.
* **Methodology**: Agile, Scrum, Waterfall, GitHub, BitBucket.

**EXPERIENCE Strata scratch -** <https://platform.stratascratch.com/user/SaiShankari>

**Capital One, Virginia - Data Engineer/ Data Analyst** **Jan 2024- present**

**Full Max - Migration project:**

* Ingested data from multiple source systems including relational data base (MYSQL), no SQL (MongoDB), S3 and redshift into a s3 data lake which is a landing zone. Read the data from staging area and performed the necessary transformation using **Spark** and provisioned the data in **redshift** data mart in form of a star schema. Used AWS **Cloud watch** for orchestration.

**Financial Data Management (FDM):**

* Created complex dashboards for adding purpose for auditing team by optimizing existing query and created trend line charts and graphs for matched and unmatched outstanding summary reports..

**Responsibilities:**

* Designed, implemented and developed ETL solutions for data ingestion, cleansing, business rules execution as per the business requirements.
* Used SQL Queries (joins, rank, dense\_rank, lag, lead, nested queries)to create complex data pipelines for transactions.
* Used Tableau to create dashboard to monitor data and monitor trends for multiple transactions.
* Worked on complex transformation like explode(), data functions, string functions, aggregation and window functions.
* Responsible for managing data coming from different sources like RDBMS, Oracle, Mainframe Systems & Web server logs.
* Optimized and reduced line of code for complex sql query from 1000 LOC to less than 420 LOC.
* Developed Python Script to import data SQL Server into HDFS & created Hive views on data in HDFS using Spark.
* Involved in importing and exporting the data from RDBMS to HDFS and vice versa using Sqoop.
* Migrated data from multiple source systems such SQL Server and Oracle to Hadoop distributed file system for data analysis
* Used unix and linux commands to migrate data from local to AWS cloud.
* Developed Hive queries to analyze large datasets.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and dataframes.
* Worked on complex pyspark transformations including joins, window functions, partition by , ranking & aggregate functions.
* Load and transform large sets of structured and semi structured data like XML, JSON format.
* Worked on complex pyspark functions like explode,() concat\_ws(), array\_sort(), split() to convert json to required structure.
* Performed trend analysis of data troubleshoot when there are data issues by cross checking with various sources like snowflake and splunk. Reached out to product teams for fixing data inconsistencies.
* Perform trend analysis to determine data inconsistencies for 115+ transactions with 99.9% accuracy by communicating with cross functional teams by creating tableau dashboards executing queries against source (Snowflake, Splunk, SQL) systems. Used Tableau for Visualization.
* Created dashboards in Tableau to perform trend analysis for determining data inconsistencies on executing queries against source (Snowflake, Splunk) systems. Communicated with product owners to determine root causes for data discrepancy.
* Took responsibility for CTA monthly report generation approved by C-level management to determine the availability of each transactions.
* Maintained repository of queries and contacts for 200+ transactions across multiple line of business (Branded, COAF, SBB, SBC) and channels (POS, ATM, IVR).

**Technical Stack**: PySpark,, AWS S3, AWS redshift, python, pySpark, SQL, Tableau, UNIX, LINUX.

**Salesforce, California - Data Engineer**  Jan 2023 - Jan 2024

**UIP:**

* Working on migrating 66TB of data from snowflake to AWS using PySpark to optimize storage cost and to improve performance of the application for being used by downstream teams. The pipeline is orchestrated with airflow DAGs.

**Bounty Alert - Glue**

* Created **Glue** pipeline to transform data from json format to parquet format. Queried the GLUE catalog table using **Athena**.
* Created a pipeline to trigger a lambda function on a new file being uploaded to S3 using S3 event notification. Used lambda script to create back up of the file in a new bucket.

**Responsibilities:**

* Developed data pipelines incorporating SCD1 (Slowly Changing Dimension Type 1) logic. Implemented data snapshots at key points in the pipeline, enabling BI developers to leverage Hive external tables for snapshot analysis and trend identification.
* Enabled historical data analysis for data science and machine learning teams by implementing SCD2 logic within data pipelines. Leveraged Apache Iceberg tables sourced from original data to ensure efficient access and querying of historical information.
* Created data generation and cleaning transformation including explode() and coalesce() function in Pyspark.
* Worked on complex joins and windowing functions (rank(), dense\_rank()) in PySpark.
* Worked on PySpark date time operations like date\_format() and unix\_timestamp.
* Worked on data partitioning and bucketing functions in PySpark including repartition(), coalesce() and bucketBy().
* Worked on data transformation functions in PySpark like map(), flatmap(), filter().
* Worked on data aggregation function (count, sum, average) in Pyspark with groupBy() clause.
* Used Airflow to execute PySpark jobs using LivyOperator by scheduling the batch jobs based on the same frequency in the source data warehouse.
* Configured Airflow DAGs with dynamic configurations and error handling for robust and scalable ETL pipelines using PySpark jobs.
* Implemented knowledge on data warehousing concepts like SCD1, SCD2, role playing dimensions, date dimension, degenerated dimension, date dimension, hierarchical dimension and so on.
* Create Glue crawler, catalog tables and create and execute jobs on AWS GLUE.
* Used Athena to query tables in S3 that are present in GLUE data catalog.
* Performing complex transformation in GLUE in GLUE STUDIO.

**Technical Stack:** PySpark, GitHub, AWS, EMR, EC2, S3, Amazon redshift, Athena, AWS Glue, AWS Delta lake, AWS Data brew, CI/CD pipelines, SQL,Tableau, Data warehousing & Data modeling concepts.

**Cardinal Health, Ohio – Cloud Big Data Engineer**                                                                **Jun 2021 – Dec 2022**

**Clarity Adherence**

Performed preprocessing and transformation using spark on client data in amazon S3 and persisting the data in Postgres database used by downstream for creating dashboards.

* Obtain client data from **AWS S3** buckets
* Reading the files and creating data frames performing transformations such as filter, join, grouping, aggregation, sorting and writing the data frame back to S3 buckets.
* Used AWS Glue Crawlers which will crawl the processed data in AWS S3 buckets to create the Athena tables on top of it.

**Thundercats**

Automated loading of business excel files into SQL Server and postgresql databases using spark scripts to load data into SQL Server by reading from SFTP location at specified schedules.

* Created Python scripts to read data from SFTP, perform transformations and load data into SQL Server.
* Automated the application by deploying the application in PCF.
* Scheduled run based on business requirement using JSON scripts and got message of success/error through slack.

**DataLabs**

Created data warehouse by migrating 120+ tables from multiple product sources from different databases (SQL Server, MySQL, Postgres) to Amazon **redshift** using **Attunity** to load data near real time.

* Created decoupled log stream task to load data from source to data sink.
* Created two separate processes for current view (CV) and historic view (HV).
* Worked on performing both change data capture (CDC) and full load.
* Validated the data from target to source with counts and sample records.

**Responsibilities:**

* Performed automation using pyspark by ingesting transforming data based on clients requirement and loading into multiple databases like postgres and SSMS.
* Design & Build data flow pipelines using tools like Attunity by migrating data from multiple source systems into redshift.
* Performed POC’s for multiple ETL tools like AWS data migration, Attunity and Adeptia.
* Performed automation using PySpark by ingesting transforming data based on clients requirement and loading into multiple databases.
* Performed file ingestion using framework to load data to redshift that are of multiple formats like csv, json, txt, tsv and so on.
* Used unix commands to work on files and folder structures.
* Migrated data from on-prem database into AWS cloud using S3 and Athena services.
* Designed and built end-to-end Data Warehouse infrastructure from the ground up on Redshift for large scale data handling.
* Worked closely with BI devlepers and data scientist to create dashboards and troubleshoot connection issues in Tableau.
* Improved the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN
* Developed SQL scripts using Spark for handling different data sets and verified the performance over Map Reduce jobs.
* Handled Data Ingestion: Importing and exporting data into HDFS using SQOOP.
* Used python cron jobs to schedule sql scripts to load data from product databases to data warehouse.

**Technical Stack:** PySpark, Python, CI/CD pipelines, GitHub, Attunity, AWS, EMR, EC2, S3, redshift, Athena, AWS Glue, AWS Delta lake, AWS Data brew, Glue, SQL,Tableau.

**FedEx, Tennessee – Data Scientist/Analyst Nov 2019 – May 2021**

**Verde Analysis**

* Extracted package shipment and customer contact information data globally from multiple source systems by segmenting customers into six clusters in **Teradata** to survey customers for every quarter using **clustering** techniques.
* Studied and analyzed sentiments of low hanging customers and derived strategies to convert passive customers into promoters using customer experience management programs that increased profits by 6%.
* Created dashboards in **Tableau** to determine profits generated, customer sentiments (call volumes, case volumes), most serious problem for customers by states and customer group for every quarter to be presented to C-level management.
* Prioritized problems faced by customers for each quarter and compared with baseline incidence and observe the frequency of each issues. Brainstorming **problem resolution strategy** for the most serious issues and observer a 12% decrease in problem frequency over a period of two years.
* Determined **statistically significant volume of customer** to be invited for survey to ensure random sampling based on customer segment population by considering **95% confidence interval, 4% margin of error and estimated response rate** as parameters in **Excel**.

**CASCADE**

* Determined root causes of clearance cases in EDW database to allocate and train human resources based on the requirement that cut human resources by 5%.
* Determined cost to company on calls and cases by regions and service types using **SQL**.

**SDE Healthcare analysis**

* Determined the action triggered and evaluate customer sentiments prior to vs consequence of sending notifications to healthcare customers during natural catastrophe for medical packages shipped in the affected regions by observing case and call volumes from specific customers.
* Performed data analysis on customer and shipment data using **Python Pandas** library.

**Responsibilities:**

* Perform descriptive statistics on the given data like mean, median, mode, variation, central tendency for a given data set in Excel.
* Perform demographic analysis on given customer data by age, sex, region and so on in Excel.
* Perform hypothesis testing to determine if sample size is statistically significant with respect to the population for customer segments in Excel.
* Created presentation in PowerPoint to present to C-level management that are used for decision making.
* Experienced in scripting HIVE QL queries to analyze millions records with different file formats like RCFile, Sequence files and text files using portioning and bucketing concepts on customer data.
* Designed decision driving dashboards, used by program managers for showcasing the markets performances &revenue generation.
* Extract data from Teradata by segments of business type as small business and Ecommerce and determine the package volume and profits generated using SQL.
* Created histograms, pie charts, bar charts, state maps in Tableau to determine profit generated by states, categories, sub-categories, price.
* Created different KPI using calculated key figures and parameters
* Developed Tableau data visualization using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
* Provided Production support to Tableau users and Wrote Custom SQL to support business requirements.

**Technical Stack:**Hadoop, Spark SQL, HQL, Teradata, SQL (DML, DDL), Tableau,AWS, Big data, HIVE,Excel.

**JLL, Portland– Data Analyst Sep 2018 – Sep 2019**

**Aero – Construction & Property Management**

* Maintained integrity of sale and leasedata regarding properties andAvailabilityused by property brokers with 99% accuracyby performing Regression, Dataand Unit Testing with **SQL**and validate the correctness of data between the database and front end (browser API’s).
* Extracted data from multiple source system transformed and loaded the data into the target system using data mapping documentation to satisfy client requirement in compliance withdata governance policies and standards.
* Created reports in **Tableau** to illustrate the correctness of data flow from multiple source systems into the system database that are regularly used by C-level management and heads of the department to determine the root cause of mismatches and percentage of inaccurate data with respect to each source and entity.
* Maintained a detail central repository used by c-level managementthat had mapping information regarding all the entities along with the **ETL** performed for each attributefor these entities that would cut down the time consumption of other departments.

**Responsibilities:**

* Prepared the ETL mapping specifications to extract data from the multiple source systems to the target system.
* Combining data from multiple source system and performed data flow and data transformation using SSIS.
* Data analysis using SQL, data quality validation, data governance (business processes, policies)
* Maintained data consistency and database integrity and attended team meetings to identify requirements for data loading and reporting
* Wrote Test Plan and created Test cases and scenarios for the JIRA items based on business requirements.
* Prepared/Reviewed weekly testing status and reports and performed regression testing in all the environments – DEV, QA, UAT and production using excel, power pivot and power query.
* Performed testing by consuming data from API and comparing with backend relational database.
* Performed data testing once every week over all the environment using SQL and Tableau.
* Generated data extracts in Tableau by connecting to the view using Tableau MySQL connector.
* Created complex Calculations in extracted Tableau data with parameter.
* Using tableau to run complex correlation analysis to identify relationships between measures.
* Combined data from multiple source systems like Excel, oracle, SSMS, MySql to a common staging database using SSIS data flow and control flow tasks.
* Performed various SSIS transformation on the data present in the staging database like data conversion, conditional split, copy column, merge join, derived column transformation and so on.

**Technical Stack:** Python, Anaconda, Jupyter Notebook, SQL, SSMS, PowerBI, Tableau, Excel (lookups, match,index), Data Quality Analysis, ETL - SSIS Testing, source target mapping

**Illinois Institute of Technology (IIT), Chicago – Graduate Analyst Mar 2017 –May 2018**

**Technology Analyst - Associate VP of International Affairs**

* Generated **dashboard** in **Excel** by extracting data from **Salesforce** by using workbench for comparing and analyzing revenue generated from international students and categorizing revenue generated based on country, department, gender and degree at Illinois Institute of Technology.
* Created and documented flow charts and sequence diagram for admission process for research students at IIT using **Visio**.

**Graduate Teaching Assistant – Information Technology & Management Department**

* Assisted professor by guiding and teaching the other fellow students on topics from SQL, RDBMS & Data analytics.
* Independently supervised class of 90 students and conducted interactive discussions weekly 2-3 times to assist them with their assignments and projects resulting in improvement in their grades.
* Provided study materials, lesson plan/structure, assignments that averaged a 92% completion rate.
* Maintained records of 75 students and created grade sheets to track individual performance and overall class performance.

**Temenos Private Limited, Chennai– Data Analyst**  **Jul 2015 - Aug 2016**

**Classifying customers Demanding for Loan**

* Implemented **classification algorithms (K Nearest Neighbors, Logistic Regression)** to **predict** sanctioning of loan for bank customers by analyzing target client’s details by which bank managers can reduce risk up to 20% using technologies like python, Excel, and SQL.

**Responsibilities:**

* Pre-processed the dataset that came with duplicate rows, multiple data formats, null values, etc.
* Visualized the data into pie-charts, bar graphs, histograms, scatter plots and performed analysis.
* Performed prediction analysis on classification models like logistic regression and KNN means.
* Worked on libraries like Scikit Learn, Pandas and NumPy.
* Experienced with A/B tests on the web

**Technical Stack:** SQL, Machine Learning – Classification, Python, Explanatory data analytics, SQL Server Management Service

**National Informatic Center, Chennai – Business Analyst Intern May 2015 – Jun 2015**

* Played a significant role in designing the road map of the One Point system integration project that integrated and streamline three systems customer identity, sales, and services together into one system by extensively communicating with internal customers and clients.
* Followed SDLC Agile methodology in implementing requirements of One Point into Sprints.
* Prepared Functional Requirement Documents, Business Solution Documents, and user interface design documents (UI).
* Reviewed business requirements & revised existing functional specification document and Master Test Plan according to new requirements.
* Participated in functional design sessions, created, and executed SQL test scripts to address data issues.

**CERTIFICATION**

* Databricks Spark Certified Nov 2022
* SQL c3f1c6225eaf4a40ac7ed23af66aaa7a Sep 2019
* Apache SQOOP for CCA UC-U191MDV3 Feb 2018
* Hands on Tableau training for Data Scientist UC-H5G1EXVV Mar 2018
* Hands on R Programming for Data Scientist UC-EX6AYTQR Mar 2018
* Implementation of Data Warehouse with SQL Server 328776229 Jan 2018
* Information Technology Infrastructure Library 04097778-01-N3NJ Dec 2016