**SHANKARI**

**Only C2C positions|AWS Cloud Big Data Engineer | saishri.0110@gmail.com| 4698468345**

**SUMMARY**  LinkedIn: www.linkedin.com/in/seethalak

* 8+ years of professional work experience as Data Engineer, working with Big data, Spark, AWS, Python, R programming, GCP, Machine Learning, SQL & Tableau.
* Have 5+ years of industrial experience in Big Data analytics, Data manipulation using Hadoop Eco system tools Map-Reduce, HDFS, Yarn/MRv2, Pig, Hive, HBase, Spark, Kafka, Flume, Sqoop, Oozie, Avro, AWS, Spark integration with Cassandra, Zookeeper.
* 3+ years of experience working with automation projects with python and pyspark.
* Have 2 years of experience working with Machine Learning algorithms like classification, Regression, Clustering.
* Extensively worked on system analysis, design, development, testing and implementation of projects (SDLC) and capable of handling responsibilities independently as well as proactive team members.
* Hands-on experience in designing and implementing data engineering pipelines and analyzing data using Hadoop ecosystem tools like HDFS, MapReduce, Yarn, Spark, Sqoop, Hive, Pig, Flume, Kafka, Impala, Oozie and HBase.
* Sound knowledge of architecture of Distributed Systems and parallel processing frameworks.
* Sound knowledge on statistical concepts like descriptive and inferential statistics, probability distribution and hypothesis testing.
* Designed and implemented end-to-end data pipelines to extract, cleanse, process and analyze huge amounts of behavioral data and log data.
* Good experience working with various data analytics and big data services in AWS Cloud like EMR, Redshift, S3, Athena, Glue etc.,
* Experience in Developing Spark applications using Spark - SQL in Databricks for data extraction, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Experienced in developing production ready spark applications using Spark RDD APIs, Data frames, Spark-SQL and Spark-Streaming API's.
* Strong experience in using Spark Streaming, Spark SQL and other components of spark like accumulators, Broadcast variables, different levels of caching and optimization techniques for spark jobs
* Proficient in importing/exporting data from RDBMS to HDFS using Sqoop.
* Used hive extensively to perform various data analytics required by business teams.
* Solid experience in working various data formats like Parquet, Orc, Avro, Json etc.,
* Experience automating end-to-end data pipelines with strong resilience and recoverability.
* Worked on Spark Streaming and Structured Spark streaming including Kafka for real time data processing.
* Responsible for developing multiple Kafka Producers and Consumers from scratch as per the software requirement specifications.
* Experience in creating Impala views on hive tables for fast access to data
* Hands-on experience in various Hadoop distributions like Cloudera, Hortonworks and AWS EMR.
* Experienced in using waterfall, Agile and Scrum models of software development process framework.
* Good knowledge in Oracle PL/SQL and shell scripting.
* Experienced in developing production ready spark applications using Spark RDD APIs, Data frames and Spark-SQL.
* Strong knowledge in developing, designing and implementing data mining models using machine learning algorithms.
* 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.)

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

Illinois Institute of Technology Master**’s in information technology management (GPA-3.9/4.0) College of Engineering Guindy Bachelors Information Technology (GPA-3.6/4.0)**

**SKILLS**

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala |
| **Hadoop Distribution** | Cloudera CDH, Horton Works HDP, Apache, AWS |
| **Machine Learning / Data mining algorithms** | Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbour (KNN), Principal Component Analysis |
| **Languages** | Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, pyspark, Regular Expressions |
| **Web Technologies** | HTML, JavaScript, Restful, SOAP |
| **Operating Systems** | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| **Version Control** | GIT, GIT HUB |
| **IDE & Tools, Design** | Eclipse, Visual Studio, Net Beans, Junit, CI/CD, SQL Developer, MySQL, SQL Developer, Workbench, SSIS, Tableau |
| **Databases** | Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB). |
| **Operating Systems** | Windows 98, 2000, XP, Windows 7,10, Mac OS, Unix, Linux |
| **Cloud Technologies** | MS Azure, Amazon Web Services (AWS) |
| **Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools** | Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Salesforce, Google Shell, Linux, Bash Shell, Unix, etc., Tableau, Power BI, SAS, Crystal Reports, Dashboard Design. |

**EXPERIENCE**

**Capital One- AWS Big Data Engineer** January 2023- 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 and provisioned the data in redshift data mart. Used AWS Cloud watch for orchestration.

**Customer Transaction Availability- Creating Backup:**

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.

**Bounty Alert - Glue**

Created Glue pipeline to transform data from json format to parquet format. Queried the GLUE catalog table using Athena.

Responsibilities:

* Designed, implemented and developed ETL solutions for data ingestion, cleansing, business rules execution as per the business requirements.
* 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.
* 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.
* 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
* Developed Hive queries to analyze large datasets.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and dataframes.
* 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.
* Load and transform large sets of structured and semi structured data like XML, JSON format.

Technical Stack: Pyspark, AWS Glue, AWS S3, AWS redshift, Spark, SQOOP, HIVE, HBase,Airflow, MapReduce, Teradata, SQL, python.

**Cardinal Health – Cloud Big Data Engineer** Aug 2021 – Jan 2023

**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 postgres 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 like postgres and SSMS.
* Performed file ingestion using framework to load data to redshift that are of multiple formats like csv, json, txt, tsv and so on.
* Model lift and shift custom SQL into dbt for materialized incremental views.
* Migrated data from on-prem database into AWS cloud using S3 and athena services.
* Created Data Quality Scripts using SQL and Hive to validate successful and quality of the data. Created various types of data visualizations using Python and Tableau.
* Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format Parsed Semi Structured JSON data and converted to Parquet using Data Frames in PySpark.
* Developed a Python Script to load the CSV files into the S3 buckets and created AWS S3 buckets, performed folder management in each bucket, managed logs and objects within each bucket.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Designed and developed data mapping and transformation scripts to support and promote data warehouse development, structural changes of multiple RDBMS and data analytics efforts as well as design effective ETL logic and code as required
* Migrated on premise database structure to Redshift data warehouse.
* Designed and built end-to-end Data Warehouse infrastructure from the ground up on Redshift for large scale data handling.
* Migrated their Big Data Platform from on-premise Hadoop to Google Cloud Platform (GCP) to one of the projects that we are working.
* Experienced in working with spark eco system using Spark SQLand Scala queries on different formats like Text file, CSV file,transformation in GCP.

**Technical Stack**: PySpark, Python, CI/CD pipelines, GitHub, Attunity, AWS, EMR, EC2, S3, redshift, Athena, AWS Glue, AWS Delta lake, AWS Data brew, Glue, dbt, SQL, Google Cloud Platform (GCP).

**FedEx, Tennessee – AWS Cloud Data Engineer** Nov 2019 – Jul 2021

**Responsibilities:**

* Work closely with multiple teams to gather requirements and maintain relationships with those that are heavy users of data for analytics. Used AWS Redshift, S3, Spectrum and Athena services to query large amounts of data stored on S3 to create a Virtual Data Lake without having to go through the ETL process.
* Ingested user behavioral data from external servers such as FTP server and S3 buckets on a daily basis using custom Input Adapters.
* Worked on importing metadata into Hive using Python and migrated existing tables and the data pipeline from Legacy to AWS cloud (S3) environment and wrote Lambda functions to run the data pipeline in the cloud
* Developed various spark applications using pyspark to perform various enrichments of user behavioral data (click stream data) merged with user profile data.
* Involved in data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for downstream model learning and reporting.
* Utilized Spark pyspark API to implement batch processing of jobs
* Bulk loading from the s3 external stage to snowflake cloud using COPY command.
* Write complex snowsql scripts in snowflake cloud data warehouse for business analysis and reporting.
* Fine-tuning spark applications/jobs to improve the efficiency and overall processing time for the pipelines. Troubleshooting Spark applications for improved error tolerance.
* Converted Hive/SQL queries into Spark transformations using Spark RDDs, Python and pyspark.
* Developed Spark scripts using Python on AWS EMR for Data Aggregation, Validation and Adhoc querying.
* Experience working for EMR cluster in AWS cloud and working with S3.
* Used broadcast variables in spark, effective & efficient Joins, transformations, and other capabilities for data processing. Utilized Spark in Memory capabilities, to handle large datasets.
* Experienced in working with EMR cluster and S3 in AWS cloud and developed API for using AWS Lambda to manage the servers and run the code in the AWS.
* Utilized machine learning algorithms such as linear regression, multivariate regression, PCA, K-means, & KNN for data analysis.
* Created Airflow Scheduling scripts in Python.
* Explored the Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark-Context, Spark-SQL, Data Frame and Pair RDD's in python data bricks.
* Creating Hive tables, loading and analyzing data using hive scripts.
* Implemented Partitioning, Dynamic Partitions, Buckets in Hive using HQL.
* 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.
* Have done POC on AWS Athena service.
* Designed end to end pysparkble architecture to solve business problems using various Azure Components like HDInsight, Data Factory, Data Lake, Storage and Machine Learning Studio.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity using Azure Databricks.
* Perform descriptive statistics on the given data like mean, median, mode, variation, central tendency for a given data set in R studio.
* Implemented data migration from SQL Server to snowflake by using Python and SnowSQL.
* 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.

**Technical Stack:** AWS, RedShift, Snowflake, Spark, HBase, Athena, Airflow, pyspark, MapReduce, Azure, Teradata, SQL, python, R Studio, Excel, Power Point, Tableau.

**JLL, Portland – Data Engineer** May 2018 – Oct 2019

**Aero – Construction & Property Management**

**Responsibilities:**

* Developed a data pipeline to ingest property data and availability information into Hadoop cluster for analysis.
* Responsible for implementing a generic framework to handle different data collection methodologies from the client primary data sources, validate transform using spark and load into AWS S3.
* Involved in all phases of Installation and upgradation of Hadoop big data platform. Implementing security for Hadoop big data platform
* Designed the sequence diagrams to depict the data flow into Hadoop.
* Involved in importing and exporting data between HDFS and Relational Systems like Oracle, MySQL and DB2 using Sqoop.
* Analyzed the SQL scripts and designed the solution to implement using PySpark.
* Helped Application and Operations team to troubleshoot the performance issues
* Implemented Partitioning, Dynamic Partitions and bucketing in HIVE for efficient data access using HQL.
* Created final tables in Parquet format. Use of Impala to query and manage Parquet tables.
* Implemented data Ingestion and handling clusters in real time processing using Apache Kafka.
* Involve in creating Hive tables, loading with data and writing Hive queries using HQL.
* Collects data using Spark from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
* Explored the usage of Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark SQL, and Spark Yarn.
* Developed python code for different tasks, dependencies, SLA watcher and time sensor for each job for workflow management and automation using Airflow tool.
* Developed Spark Code using pyspark and Spark-SQL/Streaming for faster testing and processing of data.
* Involved in converting Hive/SQL queries into Spark Transformations using Spark RDDs and pyspark.
* Worked on the Spark SQL and Spark Streaming modules of Spark and used pyspark and Python to write code for all Spark use cases.Responsible for building and developing Data Warehouse system in Horton works Hadoop environment.
* Explored Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark-Context, Spark-SQL, Data Frame and Pair RDD's.
* Wrote and Tuned complex Airflow jobs.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Data Extraction, aggregations and consolidation of Adobe data within AWS Glue using PySpark.
* Migrated historical data to AWS S3 and developed a reliable mechanism for processing the incremental updates.
* Used Oozie workflow engine to manage independent Hadoop jobs and to automate several types of Hadoop such as java MapReduce, Hive and Sqoop as well as system specific jobs
* Used to monitor and debug Hadoop jobs/applications running in production.
* Worked on providing user support and application support on Hadoop infrastructure.
* Designed, developed and created ETL (Extract, Transform and Load) packages using Python to load data into Data warehouse tools (Teradata) from databases such as Oracle SQL Developer, MS SQL Server.
* Performed testing under different environments like Dev, QA & Production.
* Used Python Pandas module to read CSV files to obtain member data and store the data in data structures.
* Automated all the Python jobs using Crontab scheduler.
* Supported the testing team on Hadoop Application Testing.

**Technical Stack:** AWS, Snowflake, Hadoop, HDFS, Pig, Hive, Spark, MapReduce, Python, Control M, Cloudera CDH 4.6, Map Reduce, Sqoop, Oozie, Cassandra, Kafka, Tableau, HQL, SQL, Excel.

**Illinois Institute of Technology (IIT), Chicago – Data 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 (personal data of students in the university).
* Created and documented flow charts and sequence diagrams for the 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.
* Managed personal information of students like name, grades, origin and so on.
* Independently supervised a 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 - Dec 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.
* Technical Stack: SQL, Machine Learning – Classification, Python, Explanatory data analytics, SQL Server Management Service

**National Informatic Center, Chennai –** Data Analyst Feb 2014 – Jun 2015

**Responsibilities:**

* 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.
* Pre-processed the dataset that came with duplicate rows, multiple data formats, null values, etc with python libraries.
* Visualized the data into pie-charts, bar graphs, histograms, scatter plots and performed analysis with matplotlib library in python.
* Performed prediction analysis on classification models like logistic regression and KNN means using sklearn in python.
* Worked on libraries like Scikit Learn, Pandas and Numpy.
* Performed data mining using logistic regression and knn algorithm.
* 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.
* Involved in project management using MS Project.
* 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.
* Performed QA & prod testing.
* Participated in functional design sessions, created, and executed SQL test scripts to address data issues.
* 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.
* Responsible for writing and updating test plan documents for every sprint.
* 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 databases.
* 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 parameters.
* 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 transformations 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:** Machine Learning – Classification, Python, Explanatory data analytics, SQL, Excel, Data mining,Tableau, SQL Server Management Service

**CERTIFICATION**  TABLEAU -<https://public.tableau.com/profile/shankari8590>

* Databricks Spark Developer Certification
* SQL
* Apache SQOOP for CCA
* Hands on Tableau training for Data Scientist
* Hands on R Programming for Data Scientist
* Implementation of Data Warehouse with SQL Server
* Information Technology Infrastructure Library