**VENKATA SAI**

**Data Engineer**

**+1 (479) 388 1660**

**saivenkat.de66@gmail.com**

[**https://www.linkedin.com/in/venkata-sai-s/**](https://www.linkedin.com/in/venkata-sai-s/)

**Professional Summary:**

* Experienced software engineer with overall 9+ years in IT industry specializing in big data engineering and data analytics. Proficient in designing, developing, implementing, and supporting business applications for healthcare, supply chain, banking, and insurance industries.
* Hands on experience in big data Hadoop Ecosystem technologies like MapReduce, HDFS, Hive, Yarn, Spark, Sqoop, Kafka, HBase, Pig, Storm, Oozie, Flume, Matillion, Zookeeper, and Zenbooks.
* Over 6+ years of experience in SQL. And knowledgeable of Python, Java, and Scala programming Skills.
* 5+ years of experience working on public cloud platforms – AWS, Azure, GCP, and Expertise in data visualization and analytical technologies.
* Well-versed with Design and Architecture principles to implement Big Data Systems.
* Experience in AWS cloud services like S3 bucket using Lambda, EC2, EMR, IAM, and Redshift as well as in Orchestration and Data Pipelines like AWS Step functions, Data Pipeline, Glue, and Amazon EventBridge.
* Experience in GCP cloud platform including IAM roles, application migration using Pub/Sub, Cloud Storage, BigQuery, dataflow, Dataproc and Google Composer.
* Knowing Azure cloud services like Azure SQL DW, HD Insights cluster, Blob, Azure data factory, ADLS Gen2, Delta Lake, data bricks, and Snowflake.
* Used Python and several Python packages like Pandas, NumPy, and Matplotlib to build machine-learning statistical models and manipulate data.
* Hands-on experience with Java8, Scala, and Akka frameworks and developed the rest APIs using Scala.
* Experience in developing SQL scripts for automation purposes, involved in testing the SQL, Looker reports, and Snowflake DB Scripts.
* Experienced in importing data in reporting software like Tableau, Looker, Athena, and creating graphs, flowcharts, and control tower internal analytical dashboards.
* Demonstrated expertise in creating robust data workflows and ETL processes, along with deep experience in integrating with NoSQL databases like Cassandra and MongoDB.
* Proficient in working within Palantir Foundry's Ontology layer, enabling structured semantic data integration and relationship modeling for complex networking data analysis.
* Implemented ETL operations using Big Data platforms and Experience in Coding in Java or Scala along with EMR, and Apache Spark. Worked on real-time data integration using Kafka, Spark Streaming, and HBase.
* Enhanced performance using various sub-projects of Hadoop, performed data migration from legacy using Sqoop, handled performance tuning, and conducted regular backups.
* Data cleansing, Data manipulation, and exploratory analysis to identify, analyze, and interpret trends and patterns in large data sets.
* Experience in building tools like Maven, Gradle, Log4j, Junit, Ant, Pipenv, Build, and SBT. Expertise knowledge in relational databases like Oracle, My SQL, and SQL Server.
* Specified the cluster size, allocating Resource pool, and Distribution of Hadoop by writing the specification texts in JSON File format.
* Experience in designing and developing applications in Spark using Scala to compare the performance of Spark with Hive and SQL/Oracle.
* Expertise in writing the Real-time processing application Using spout and bolt in Storm.
* Experienced migrating ETL transformations using Pig Latin Scripts, transformations, and join operations.
* Experience in configuring and using Flume to load data from multiple sources directly into HDFS.
* Solid experience as an RDBMS developer with stored procedures, query performance tuning, and ETL.
* Integrated with Hadoop/HDFS, Real-Time Systems, Data Warehouses, and Analytics solutions.
* Experience in Data Warehousing, Data Lake, and ETL concepts using Informatica Power Center, OLAP, OLTP, and AutoSys.
* Experienced in working with version control systems like GIT and used Source code management client tools like Git Bash, GitHub, and Git Lab.
* Experience in software engineering best practices for the full software development life cycle, including coding standards, code reviews, source control build, automated testing and release processes, continuous integration and continuous deployment concepts (Cl & CD), peer reviews, and maintenance.
* Track and manage Saas application deployment, Win 10 upgrade, Office 2016 upgrade, Testing, and UAT.
* Built real-time data pipelines by developing Kafka procedures and spark streaming applications for consumption.
* Supporting production and technical assistance, problem resolution, and troubleshooting support.
* Experience in implementing projects both in Agile and Waterfall methodologies.
* Strong analytical and problem-solving skills, highly motivated, good team player with good communication skills, and interpersonal skills.

**Technical skills:**

* **Big Data Ecosystem**: Hadoop HDFS, MapReduce, Hive, YARN, Kafka, Pig, Flume, Sqoop, Impala, Oozie, Zookeeper, Spark, Ambari, Databricks, MongoDB, Cassandra, Storm, Parquet, ab initio, Snappy.
* **Hadoop Distributions**: Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR, AWS EMR, Hadoop
* **Programming Languages**: Python, SQL, Java, HTML, DHTML, Scala, JavaScript, and power shell.
* **ETL Tools**: Talend, Informatica, Pentaho, IBM data stage, Stitch, Apatar, AWS Glue, Cloud dataflow
* **No SQL Databases**: Cassandra, MongoDB, Neo4j, and HBase
* **Data base Languages**: MySQL, PL/SQL, PostgreSQL, and Oracle
* **Code build Tools**: Eclipse, Ant, Maven, IntelliJ, PyCharm, Citrix, JUNIT, log4J, Pipenv, Free marker template
* **Python Technologies**: NumPy, Pandas, TensorFlow, Keras, Pyspark, Hadoop streaming, Boto3, Fast API.
* **Data analytical tools**: Tableau, Tableau public, Power BI, RapidMiner, QlikView, SAS, MS Excel, Splunk, and Control Tower dashboard (Internal).
* **Deployment tools**: Git, GitHub, Jenkins, Bamboo, GitLab, XLR, Bitbucket,
* **Scrum methodology:** Agile Jira, Rally, waterfall, Slack, Confluence.

**Professional Experience:**

**Client: American Express, Phoenix, AZ September 2022 to Present**

**Role: GCP Data Engineer/Backend data engineer**

* Involved in requirements gathering, analysis, design, development, and deployment phases of the project.
* Experience in cornerstone data platform for MDM creation and ingestion framework setup in various environments.
* Working with an event engine server for a workload management tool for queuing and load balancing.
* Developed an intraday reporting dashboard for monitoring money movement events and business financial health and analysis.
* Developed XLR CI/CD build pipelines for deploying the code in multiple environments and config files moved to cloud storage buckets.
* Developed a LUMI GCP cloud-based big data ecosystem platform enabling users to securely govern data.
* Migrated the historical data from the cornerstone data platform to the Google Cloud platform using the big query transfer service and cloud pub/sub
* Working closely with LUMI data ingestion, metadata management, balance and control (BnC), account data encryption and decryption, and governance process.
* Create and manage real-time data storage solutions using GCP services such as Big Query, cloud storage, data proc, and cloud SQL Server.
* Delivered Palantir Foundry AIP models for various Gen AI applications, focusing on predictive analytics and automation.
* Implement streaming data pipelines using Pub/Sub for real-time data processing and analytics.
* Monitor and troubleshoot data pipelines and storage(bucket) solutions using GCP operations and cloud monitoring and automate data processing tasks using Python.
* Extensively used SQL, NumPy, Pandas, Seaborn, Matplotlib, Spark, Netezza, and Hive for Data Analysis
* Lumi offers data ingestion, transformation, project space, big query, data proc, and cloud composer airflow.
* Developed complex SQL queries for in-depth data analysis on big query tables.
* Having experience in orchestrating pipelines built on Apache airflow composer that are used to author, schedule, and monitor workflows.
* Developed monitoring and alerting scenarios for manual and scheduling workflow jobs in airflow composer to check the job status.
* Designing, scheduling, and monitoring google composer Airflow jobs to coordinate data processing between multiple technologies.
* Build data pipelines in airflow for ETL-related jobs using different airflow operators.
* Perform multiple tasks and trigger rules in the script, which can be configured by creating DAGs and executing them to determine the status of the airflow composer jobs.
* Working with the LUMI team to gain access across the platform and resolve firewall connection issues.
* Wrote Hive Queries for analyzing data in Hive warehouse using Hive Query Language (HQL).
* Developed multiple views/charts scenarios based on business requirements such as net flows, daily inflows, and outflows, summary, segmentation products, state and territory, age and tenure, and industry.
* Developed security protocols and access controls in GCP IAM to safeguard sensitive banking data, preventing unauthorized access and ensuring data integrity.
* Played a key role in data foundation and modernization efforts by structuring, standardizing, and optimizing data assets in Palantir Foundry.
* Creating and implementing processes to transfer data from various sources SFTP file, JDBC database, Teradata, REST to the Google Cloud Platform.
* Implemented ETL process to streamline data import from various sources into Big Query warehouse.
* Developed and maintained data ingestion pipelines using Python, Apache Airflow, and GCP services.
* Designed and developed robust data solutions for merchant businesses.
* Developed Spark Java application for ingestion framework to ingest data from Kafka topics to Cassandra and Elastic Search.
* Implemented dynamic data processing framework to generate customer-specific reports.
* Implemented a reconciliation process engine for reconciling the data/transactions.
* Developed a scheduling framework to schedule jobs in Oozie by consuming configuration data from Elastic Search. Worked with the team to optimize the spark jobs.
* Developed a job monitoring framework to monitor the jobs and generate job monitor reports.
* Worked with various file formats for consuming and producing the data.
* Have good exposure to debugging and root cause analysis to improve the data quality.
* Worked on production support and on weekends to monitor the jobs. Implemented a logging mechanism in every framework to capture the logs.
* Developed workflow in Oozie to manage and schedule jobs on the Hadoop cluster to trigger daily, weekly, and monthly batch cycles.
* Developed spark applications to handle various file formats like Parquet, AVRO, XML, JSON, CSV, and TEXT.
* Created solutions for data movement and transformation utilizing Python, and Pyspark. And Extensive experience in developing Pyspark, and Python scripts.
* Encoded and decoded JSON objects using Pyspark to create and modify the data frames in Apache Spark
* Developed Restful APIs and integrated them using Python, and Flask with data exchange formats including JSON, CSV, and XML.
* Created and designed diagrams of various applications for non-production, business continuity, production, and DR hosted in the cloud and on-premises infrastructure.
* Develop, maintain, and automate data processing workflows on the Google Cloud Platform (GCP) to ensure efficient data handling and transformation.
* Implemented data validation and performance tuning strategies on Dataproc Spark jobs, leveraging caching and resource optimization to handle complex queries on large datasets.
* Worked on data cleaning and reshaping, generated segmented subsets using NumPy and Pandas in Python.
* Specified the cluster size, allocating Resource pool, and Distribution of Hadoop by writing the specification texts in JSON File format.
* Set up Virtual Machines for production, Non-production, and DR environments in the cloud and On-Premises.
* Interacted with Business Analysts for requirements gathering and attending peer review meetings for code change requests for writing test cases.

**Environment**: Apache Spark, Java, Python, SQL, Apache Kafka, Hive, HDFS, Cassandra, Elastic Search, Kibana, Yarn, cornerstone MDM, Event engine, DBeaver, yellow bricks, Oozie, Git, Cloudera, JUnit, SFTP, data integration, data transformation, data modeling, and data warehousing, XLR template (CI/CD), Pub/Sub, Dataflow, DataProc, Big Query, Google Cloud Storage, Apache airflow, and Agile methodologies (Rally and Jira Tools).

**Agility e-services - Hyderabad, India November 2017 to December 2021**

**AWS Data Engineer**

**Responsibilities:**

* Designed, developed, and tested the ETL process on AWS EMR to migrate FinTech data from external sources, such as AWS S3 buckets and SQL databases, using AVRO, Parquet, and text files into AWS Redshift. Transformed stored procedures into Spark SQL API for enhanced data processing capabilities.
* Involved in designing and optimizing SPARK SQL queries and data frames, importing data from Data sources, performing transformations, and storing the results to output directory into AWS S3.
* Defined and deployed monitoring, metrics, and logging systems on AWS.
* Experience in the processing of data transformations, and actions in Spark by applying Python language. Performance tuning of Pyspark scripts.
* Created and managed QuickSight datasets, including defining the data schema, creating calculated fields, and adding filters and joins.
* Worked with Palantir Foundry for ETL workflows, data lineage, and advanced analytics to enhance enterprise decision-making.
* Built Quick Sight dashboards and reports using drag-and-drop tools to create interactive data visualizations.
* Created automated pipelines in AWS Code Pipeline to deploy Docker containers in AWS ECS using S3.
* Implemented MongoDB for handling large datasets and used for fraud detection and for making decisions for load approval.
* Designing and implementing ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) solutions to automate data flows.
* Configured AWS Cloud Watch to monitor AWS resources, including creating customized AWS Scripts to monitor various applications and system & Instance metrics.
* Data Extraction, aggregations, and consolidation of Adobe data within AWS Glue using Pyspark. Create external tables with partitions using Hive, AWS Athena, and Redshift.
* Designing and implementing Scalable ETL pipelines to process a variety of data types (structured, semi-structured), and file formats JSON, CSV, text, and delimited
* Developed and executed ETL scripts. ETL scripts are the code that is used to extract, transform, and load data. And written in various languages, such as Python, Java, and SQL.
* In-depth knowledge of Snowflake Database, Schema, and Table structures.
* Experience in job/workflow scheduling and monitoring tools like Oozie, AWS Data pipeline & Autopsy’s.
* Developed a reconciliation process to make sure the Elastic Search index document count matches source records using Python Flask specification.
* Skilled in implementing analytical use cases by leveraging network data within Palantir Foundry, with hands-on experience in exploratory data analysis and visualization.
* Wrote data normalization jobs for newly ingested data into Redshift. Designed and developed ETL jobs to extract data from replicas and load it into the data mart in Redshift.
* Designed data model for Dynamo DB and participated in capacity planning. Experience in creating NoSQL document data models for DynamoDB.
* Designed and optimized data pipelines within Palantir Foundry to streamline ETL processes, ensuring efficient data ingestion and processing from diverse network sources.
* For faster access to data, perform ad-hoc queries using HIVE joins partitioning and bucketing techniques.
* Participate in the development improvement and maintenance of Snowflake database applications
* Experience with Snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source systems, including loading nested JSON formatted data into Snowflake table.
* Performed POC for Loading tables from Azure data Lake to Azure blob storage to push them to Snowflake.
* Performed data cleaning and reshaping, generating segmented subsets with NumPy and Pandas.
* Created scripts in Python (Boto) which are integrated with Amazon API to control instance operations.
* Involved in maintaining the reliability, availability, and performance of Amazon Elastic Compute Cloud (Amazon EC2) instances.
* Developed data ingestion pipeline from HDFS into AWS S3 buckets using Nifi. And created external and permanent tables on the AWS data.
* Designed, built, and coordinated an automated build & and released CI/CD process using Gitlab, Jenkins, and Puppet on hybrid IT infrastructure.
* Experience with snowflake utilities SnowSQL and big data model techniques using Python.
* Responsible for collecting, scrubbing, and extracting data from various generated reports, dashboards, and analytical solutions. Helped with debugging the Tableau dashboards.
* Involved in the complete SDLC life cycle - Designing, Coding, Testing, Debugging, and Production Support.
* Worked in an Agile development environment and participated in daily Scrum and other design-related meetings.

**Environment**: Hadoop, HDFS, Spark, AWS Glue, Redshift, Lambda, Databricks, Python, Pyspark, Kafka, PyCharm, ADFV2, AWS EMR, EC2, S3, Data Lake, Snowflake, Hive, Flume, Apache Nifi, Shell-scripting, SQL, Pig, Sqoop, MongoDB, SQL Server, SAS EG 7.1, SAS VA, SAS Viya 3.4, SAS DI Studio, Tableau, and Agile Methodology,

**Client: MRR Software Solutions, India July 2015 to October 2017**

**Role: ETL developer / Hadoop Developer /Database Engineer**

* Worked extensively on Hadoop components such as HDFS, Map Reduce, YARN, HBase, Hive, Sqoop, and Spark, and used Python programming.
* Responsible for building scalable distributed data solutions using the Hadoop framework.
* Experience monitoring the Hadoop cluster using Cloudera Manager, interacting with Cloudera support, logging the issues in the Cloudera portal, and fixing them as per the recommendations.
* Experience in Cloudera Hadoop Upgrades and Patches and Installation of Ecosystem Products through Cloudera Manager and Cloudera Manager Upgrade.
* Responsible for developing, supporting, and maintaining the ETL processes using Informatica Power Center.
* Develop ETL processes to populate a Hadoop data warehouse and Data Lake with large datasets from various sources and integrate Hadoop within an SQL Server data warehousing environment.
* Responsible for tuning ETL procedures and STAR schemas to optimize load and query performance.
* Generated server-side PL/SQL scripts for data manipulation and validation and materialized views for remote instances.
* Used Bulk Collections for better performance and easy data retrieval, by reducing context switching between SQL and PL/SQL engines.
* Involved in loading data from the UNIX file system and FTP to HDFS and Developed Hive queries to analyze the output data.
* Specified the cluster size, allocating Resource pool, and Distribution of Hadoop by writing the specification texts in JSON File format.
* Developed predictive and prescriptive models within Palantir Foundry to provide actionable insights and support data-driven decision-making.
* Developed data pipeline using Spark, Hive, Pig, Python, Impala, and HBase, Sqoop to ingest customer data.
* Worked with Linux systems and RDBMS database regularly to ingest data using Sqoop.
* Wrote MapReduce program for data validation focusing on performance tuning of spark jobs using Cache and fully leveraging the cluster environment.
* Managed and reviewed Hadoop and HBase log files. Create Hive tables and work on them using HiveQL.
* Designed and implemented HIVE queries and functions for evaluation, filtering, loading, and storing of data.
* Involved in creating hive tables loading with data and writing hive queries which will run internally in map-reduce and supported the existing MapReduce Programs running on the cluster.
* Continuous monitoring and managing of the Hadoop cluster through Cloudera Manager.
* Written and Implemented Teradata Fast load, Multiload, and Bteq scripts, DML, and DDL.
* Used Zookeeper to coordinate the servers in clusters and to maintain data consistency.
* Designed and implemented Spark jobs to support distributed data processing.
* Created and automated data extraction, transformation, and loading (ETL) procedures using PL/SQL, stored procedures, functions, and triggers.
* Performed partitioning, data loading, refreshing, logical backup, recovery, and capacity planning.
* Monitored database performance analysis of high-volume transactional ORACLE databases.
* Coordinated system releases and ensured smooth transactions of supported applications and databases from the development to acceptance and production environments.
* Involved in the review of functional and non-functional requirements.
* Followed agile methodology for the entire project.

**Environment**: Apache Hadoop, MapReduce, HDFS, HBase, YARN, Hive, Unix, Shell scripting, REST web Services, Elastic Search, MySQL, PL/SQL, NoSQL, Oozie, Sqoop, Java (JDK 1.5), Python, CentOS 6.4, Flume, Eclipse, Qlik View, Qlik Sense, Quest Tools, Oracle Database, Oracle Forms/Reports, Jenkins, Agile methodology.

**EDUCATIONAL DETAILS:**

* Bachelor’s from Satavahana University Telangana in 2015
* Master’s from Union Commonwealth University Kentucky in 2023