Senior Data Engineer

Name: Sai Krishna Janapala

Email: [Janapalasaikrishna8@gmail.com](mailto:Janapalasaikrishna8@gmail.com)

Phone: (720) 710-3271  
LinkedIn: <https://www.linkedin.com/in/janapala-sai-krishna/>

**PROFESSIONAL SUMMARY**

* 9+ years of IT experience in a variety of industries working on Big Data technology using technologies such as Cloudera and Hortonworks distributions. Hadoop working environment includes Hadoop, Spark, MapReduce, Kafka, Hive, Ambari, Sqoop, HBase, and Impala.
* Fluent programming experience with Scala, Java, Python, SQL, T - SQL, R.
* Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.
* Played a key role in data analysis projects, employing Splunk to extract valuable information and drive data-driven decision-making.
* Expertise in developing and customizing Splunk apps, dashboards, and advanced visualizations.
* Adept at configuring and installing Hadoop/Spark Ecosystem Components.
* Proficient with Spark Core, Spark SQL, Spark MLlib, Spark GraphX and Spark Streaming for processing and transforming complex data using in-memory computing capabilities written in Scala. Worked with Spark to improve efficiency of existing algorithms using Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.
* Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into a data warehouse.
* Prepared and participated in AML audits, ensuring a high level of readiness and compliance with regulatory standards.
* Able to use Sqoop to migrate data between RDBMS, NoSQL databases and HDFS.
* Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
* Hands-on experience with Hadoop architecture and various components such as Hadoop File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Hadoop MapReduce programming.
* Comprehensive experience in developing simple to complex Map reduce and Streaming jobs using Scala and Java for data cleansing, filtering and data aggregation. Also possess detailed knowledge of MapReduce framework.
* Used IDEs like Eclipse, IntelliJ IDE, PyCharm IDE, Notepad ++, and Visual Studio for development.
* Seasoned practice in Machine Learning algorithms and Predictive Modeling such as Linear Regression, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, KNN, Neural Networks, and K-means Clustering.
* Ample knowledge of data architecture including data ingestion pipeline design, Hadoop/Spark architecture, data modeling, data mining, machine learning and advanced data processing.
* Experience working with NoSQL databases like Cassandra and HBase and developed real-time read/write access to very large datasets via HBase.
* Developed Spark Applications that can handle data from various RDBMS (MySQL, Oracle Database) and Streaming sources.
* Proficient SQL experience in querying, data extraction/transformations and developing queries for a wide range of applications.
* Implemented and maintained Anti-Money Laundering (AML) policies and procedures to ensure adherence to regulatory requirements.
* Maintained up-to-date knowledge of AML laws and regulations, ensuring the organization's practices align with current standards.
* Capable of processing large sets (Gigabytes) of structured, semi-structured or unstructured data.
* Experience in analyzing data using HiveQL, Pig, HBase and custom MapReduce programs in Java 8.
* Experience working with GitHub/Git 2.12 source and version control systems.
* Strong in core Java concepts including Object-Oriented Design (OOD) and Java components like Collections Framework, Exception handling, I/O system.

TECHNICAL SKILLS

Big Data Technologies : Spark, Cloudera, Hive, Impala, HBase, Oozie, Kafka, Databricks, Airflow

Languages : SQL, Python and Scala

Cloud : Azure Data factory, Azure Stream Analytics, Databricks, Azure PowerShell, Azure HD Insight, ADLS, Blob storage, AWS, S3, EMR, Glue, Athena, Redshift, GCP, GCS, BigQuery, DataProc, DataStudio, DataFusion

Databases : Oracle, MySQL, SQL Server, Teradata

IDE and Notebooks : Eclipse, IntelliJ, PyCharm, Jupiter, Databricks notebooks

Data formats : Json, Parquet, AVRO, XML and CSV

Search and BI tools : PowerBI, DataStudio, Tableau

Web Technologies : JDBC, JSP, Servlets, Struts (Tomcat, JBoss)

EDUCATION DETAILS:

Bachelor’s: JNTUK, 2014

**PROFESSIONAL EXPERIENCE**

**VISA Inc, Atlanta, GA Sep 2021 to Present**

**Sr Data Engineer**

**Responsibilities:**

* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in AWS.
* Implemented a Continuous Delivery pipeline with Docker and Git Hub
* Worked with Lambda function to load Data into Redshift on arrival of csv files in S3 bucket.
* Devised simple and complex SQL scripts to check and validate Dataflow in various applications.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python.
* Performed data engineering functions: data extract, transformation, loading, and integration in support of enterprise data infrastructures - data warehouse, operational data stores and master data management.
* Responsible for data services and data movement infrastructures good experience with ETL concepts, building ETL solutions and Data modeling.
* Architected several DAGs (Directed Acyclic Graph) for automating ETL pipelines.
* Worked closely with business stakeholders to understand their requirements and translate them into effective Splunk queries and visualizations.
* Hands on experience on architecting the ETL transformation layers and writing spark jobs to do the processing.
* Gather and process raw data at scale (including writing scripts, web scraping, calling APIs, write SQL queries, writing applications)
* Spearhead data analysis initiatives, utilizing Splunk to uncover patterns and trends that contribute to strategic decision-making.
* Communicated effectively with various departments, fostering a collaborative approach to AML compliance throughout the organization.
* Imported data from AWS S3 into Spark RDD and performed actions/transformations on them.
* Created Partitions, Bucketing and Indexing for optimization as part of Hive data modeling.
* Involved in developing Hive DDLs to create, alter and drop Hive tables.
* Worked on different RDDs to transform the data coming from different data sources and transform the data into required formats.
* Created data frames in SPARK SQL from data in HDFS and performed transformations, analyzed the data and stored the data in HDFS.
* Worked with Spark Core, Spark Streaming and Spark SQL modules of Spark for faster processing of data.
* Developed Spark code and Spark SQL for faster testing and processing of Realtime data.
* Worked on Talend ETL to load data from various sources to Data Lake.
* Conduct in-depth analysis of Splunk logs to identify potential issues, anomalies, and security threats.
* Used amazon Elastic Beanstalk with Amazon EC2 to deploy project in AWS.
* Used spark for interactive queries, processing of streaming data and integrating with popular NOSQL databases for a huge volume of data.
* Consumed the data from Kafka queue using Spark.
* Involved in regular standup-meetings, status calls, Business owner meeting with stake holders.

Environment: Spark, Python, AWS, S3, Glue, Redshift, DynamoDB, Hive, SparkSQL, Docker, Kubernetes, Airflow, ETL workflows, Splunk

**Bank of America, Raleigh, NC Feb 2020 to August 2021**

**Sr Data Engineer**

**Responsibilities:**

* Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
* Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines.
* Performed data extraction, transformation, loading, and integration in data warehouse, operational data stores and master data management.
* Implemented Copy activity, Custom Azure Data Factory Pipeline Activities
* Primarily involved in Data Migration using SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB).
* Migration of on-premises data (Oracle/ SQL Server/ DB2/ MongoDB) to Azure Data Lake and Stored (ADLS) using Azure Data Factory (ADF V1/V2) & Azure, self-hosted Integration runtime.
* 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.
* Implement ad-hoc analysis solutions using Azure Data Lake Analytics/Store, HDInsight Cluster
* Experience in developing splunk queries and dashboards targeted at understanding application performance and capacity analysis.
* Experienced in ETL concepts, building ETL solutions and Data modeling.
* Worked on architecting the ETL transformation layers and writing spark jobs to do the processing.
* Aggregated daily sales team updates to send report to executives and to organize jobs running on Spark clusters.
* Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the snowflake.
* Loaded application analytics data into data warehouse in regular intervals of time.
* Designed & build infrastructure for the Google Cloud environment from scratch.
* Implemented Dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
* Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines.
* Worked on confluence and Jira.
* Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python.
* Compiled data from various sources to perform complex analysis for actionable results.
* Measured Efficiency of Hadoop/Hive environment ensuring SLA is met.
* Optimized the Tensorflow Model for efficiency.
* Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes.
* Implemented a Continuous Delivery pipeline with Docker, and Git Hub
* Built performant, scalable ETL processes to load, cleanse and validate data.
* Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies.
* Collaborate with team members and stakeholders in design and development of data environment.
* Preparing associated documentation for specifications, requirements, and testing

**Environment:** Azure, Azure Data Factory, Lambda Architecture, Stream Analytics, Snowflake, MySQL, SQL Server, Python, Scala, Spark, Hive, Spark -SQL, Pandas, NumPy, Splunk

**Zurich Insurance, Schaumburg, IL Jul 2018 to Jan 2020**

**Data Engineer**

**Responsibilities:**

* Worked extensively on performance tuning of Spark jobs and Hive queries.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python.
* Created Hive tables as per requirement as managed or external tables, intended for efficiency.
* Proactively and continuously drive system wide quality improvements by undertaking thorough root cause analysis for major incidents with component engineering teams.
* Implemented Schema extraction for parquet and Avro file formats in Hive.
* Used Spark-SQL to Load JSON data and create Schema RDD and loaded it into Hive tables and handled structure data using SparkSQL.
* Hands of experience in GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Pub/sub cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver.
* Implement Hive UDF’s for evaluation, filtering, loading, and storing of data.
* Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
* Worked on developing Pyspark script to encrypting the raw data by using hashing algorithms concepts on client specified columns.
* Responsible for Design, Development, and testing of the database and Developed Stored Procedures, Views, and Triggers
* Developed Python-based API (RESTful Web Service) to track revenue and perform revenue analysis.
* Compiling and validating data from all departments and Presenting to Director Operation.
* KPI calculator Sheet and maintain that sheet within SharePoint.
* Created Tableau reports with complex calculations and worked on Ad-hoc reporting using PowerBI.
* Creating data model that correlates all the metrics and gives a valuable output.
* Worked on the tuning of SQL Queries to bring down run time by working on Indexes and Execution Plan.
* Performing ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Design, develop, and test dimensional data models using Star and Snowflake schema methodologies under the Kimball method.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Ensure deliverables (Daily, Weekly & Monthly MIS Reports) are prepared to satisfy the project requirements cost and schedule.
* Worked on a direct query using PowerBI to compare legacy data with the current data and generated reports and stored and dashboards.
* Designed SSIS Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for the SSAS cubes (OLAP)
* SQL Server reporting services (SSRS). Created & formatted Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Subreports, ad-hoc reports, parameterized reports, interactive reports & custom reports
* Created action filters, parameters and calculated sets for preparing dashboards and worksheets using PowerBI
* Used ETL to implement the Slowly Changing Transformation, to maintain Historically Data in Data warehouse.
* Performing ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Created dashboards for analyzing POS data using Power BI

**Environment:** Spark, Python, ETL, Power BI, Hive, Power BI, GCP, BigQuery, DataProc Data Pipeline, IBM Cognos 10.1, Data Stage, Cognos Report Studio 10.1, Cognos 8 & 10 BI, Cognos Connection, Cognos office Connection, Cognos 8.2/3/4, Data stage and Quality Stage 7.5, MS SQL Server 2016, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel

**Verizon, Dallas, TX Jan 2017 to Jun 2018**

**Data Engineer**

**Responsibilities:**

* Involved in Analysis, Design, and Implementation/translation of Business User requirements.
* Worked on collection of large sets of Structured and Unstructured data using Python Script.
* Worked on creating DL algorithms using LSTM and RNN.
* Actively involved in designing and developing data ingestion, aggregation, and integration in the Hadoop environment.
* Developed Sqoop scripts to import export data from relational sources and handled incremental loading on the customer, transaction data by date.
* Experience in creating Hive Tables, Partitioning and Bucketing.
* Performed data analysis and data profiling using complex SQL queries on various source systems including Oracle 10g/11g and SQL Server 2012.
* Identified inconsistencies in data collected from different sources.
* Designed object model, data model, tables, constraints, necessary stored procedures, functions, triggers, and packages for Oracle Database.
* Wrote Spark applications for Data validation, cleansing, transformations, and custom aggregations.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Worked on installing cluster, commissioning & decommissioning of Data node, Name node high availability, capacity planning, and slots configuration.
* Developed Spark applications for the entire batch processing by using Scala.
* Stored the time-series transformed data from the Spark engine built on top of a Hive platform to Amazon S3 and Redshift.
* Facilitated deployment of multi-clustered environments using AWS EC2 and EMR apart from deploying Dockers for cross-functional deployment.
* Visualized the results using Tableau dashboards and the Python Seaborn libraries were used for Data interpretation in deployment.
* Created PDF reports using Golang and XML documents to send it to all customers at the end of month.
* Applied various data mining techniques: Linear Regression & Logistic Regression, classification, clustering.

**Environment:** R, SQL server, Oracle, HDFS, HBase, AWS, MapReduce, Hive, Impala, Pig, Sqoop, NoSQL, Tableau, RNN, LSTM, Unix/Linux.

**Accenture Solutions PVT LTD, India July 2014 to Dec 2016 Data Analyst**

**Responsibilities:**

* Devised simple and complex SQL scripts to check and validate Dataflow in various applications.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python.
* Devised PL/SQL Stored Procedures, Functions, Triggers, Views and packages. Made use of Indexing, Aggregation and Materialized views to optimize query performance.
* Developed logistic regression models (using R programming and Python) to predict subscription response rate based on customers variables like past transactions, response to prior mailings, promotions, demographics, interests and hobbies, etc.
* Created Tableau dashboards/reports for data visualization, Reporting and Analysis and presented it to Business.
* Created Data Connections, Published on Tableau Server for usage with Operational or Monitoring Dashboards.
* Knowledge in Tableau Administration Tool for Configuration, adding users, managing licenses and data connections, scheduling tasks, embedding views by integrating with other platforms.
* Worked with senior management to plan, define and clarify dashboard goals, objectives and requirement.
* Responsible for daily communications to management and internal organizations regarding status of all assigned projects and tasks.

**Environment:** SQL, Tableau, R, Python, Excel, Lookups, Access