**Sai Vishwanath**

**Sr. Data Engineer**

**Mail: saivishwanath.kanchibhotla@gmail.com**

**Phone: 9012649473**

**Professional Summary:**

* Over 9+ years of IT industry knowledge with hands on working experience in Data Engineering & Data Analysis.
* Good knowledge in Data Quality & Data Governance practices & processes.
* Well versed with Agile with Scrum, Waterfall Model and Test-driven Development (TDD) methodologies.
* Proficient in SQLite, MySql and SQL databases with Python.
* Practical understanding of the Data modelling (Dimensional & Relational) concepts like Star - Schema modelling, Snowflake Schema Modelling, Fact and Dimension tables.
* Experience in handling python and spark context when writing PySpark programs for ETL.
* Strong knowledge in data visualization using Power BI and Tableau.
* Hands in experience on NoSQL database like Snowflake, HBase, Cassandra and Mongo-DB.
* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Experience in GCP Dataproc, GCS, Cloud functions, Big Table and Big Query.
* Experience with Apache Spark ecosystem using Spark-Core, SQL, Data Frames and RDD's.
* Experienced in data manipulation using python.
* Hands on experience working Amazon Web Services (AWS) using Elastic Map Reduce (EMR), Redshift, and EC2 for data processing.
* Proficient in installing, configuring and using Apache Hadoop ecosystems such as Map Reduce, Hive, Pig, Flume, Yarn, HBase, Sqoop, Spark, Storm, Kafka, Oozie, and Zookeeper.
* Strong experience on designing Big data pipelines such as Data Ingestion, Data Processing (Transformations, enrichment and aggregations) and Reporting.
* Experience in integrating Kafka with Spark streaming for high speed data processing.
* Experience in implementing Azure data solutions, provisioning storage account, Azure Data Factory, Azure Data bricks, Azure Blob Storage, Azure Synapse and Azure Cosmos DB.

**Technical Skills:**

* **Hadoop/Spark Ecosystem**: Hadoop, Map Reduce, Hive/impala, YARN, Kafka, Flume, Sqoop, Oozie, Zookeeper, Spark
* **Databases**: Oracle, MySql, SQL Server, Postgre SQL, HBase, Snowflake, Cassandra, MongoDB
* **Cloud computing** : Amazon Web Services(AWS), Amazon Redshift, MS Azure, Azure blob storage, Azure Data Factory, Azure Synapse & Google cloud Platform(BigQuery, Big Table, Dataproc )
* **BI Tools**: Business Objects XI, Tableau 9.1, Power BI
* **Query Languages**: SQL, PL/SQL, T-SQL
* **Scripting Languages**: Unix, Python
* **Operating Systems**: Linux, Windows, Ubuntu, Unix
* **SDLC Methodology**: Agile, Scrum, Waterfall, UML

**Professional Experience:**

**Client: Cummins Columbus, Indiana September 2021 to Present**

**Sr. Data Engineer**

**Responsibilities:**

* As a Sr. Data Engineer, I am responsible for building scalable distributed data solutions using Hadoop.
* Involved in Agile Development process (Scrum and Sprint planning).
* Handled Hadoop cluster installations in Windows environment.
* Migrated on premise environment in GCP (Google Cloud Platform)
* Migrated data warehouses to Snowflake Data warehouse.
* Defined virtual warehouse sizing for Snowflake for different type of workloads.
* Involved in porting the existing on-premise Hive code migration to GCP (Google Cloud Platform) Big Query.
* Involved in migration an Oracle SQL ETL to run on Google cloud platform using cloud Dataproc & Big Query, cloud pub/sub for triggering the Apache Airflow jobs.
* Extracted data from data lakes, EDW to relational databases for analyzing and getting more meaningful insights using SQL Queries and PySpark.
* Experirnce in creating Data Governance Policies,Business Glossary,Data Dictionary,Reference Data,Metadata,Data Lineage and data Quality Rules.
* Designed, developed and did maintenance of data integration programs in a Hadoop and RDBMS environment with both traditional and non-traditional source systems.
* Developed Map Reduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
* Wrote Sqoop Scripts for importing and exporting data from RDBMS to HDFS.
* Set up Data Lake in Google cloud using Google cloud storage, Big Query and Big Table.
* Developed scripts in Big Query and connecting it to reporting tools.
* Designed workflows using Airflow to automate the services developed for Change data capture.
* Carried out data transformation and cleansing using SQL queries and PySpark.
* Used Kafka and Spark streaming to ingest real time or near real time data in HDFS.
* Involved in various project related to Data modeling , Data Analysis, Design and development for both OLTP and data warehousing environment
* Worked related to downloading Big Query data into Spark data frames for advanced ETL capabilities.
* Built reports for monitoring data loads into GCP and drive reliability at the site level.
* Participated in daily stand-ups, bi-weekly scrums and PI panning.

**Environment:** Hadoop, GCP, Big Query, Big Table, Spark 3.0, Sqoop 1.4.7, ETL, HDFS, Snowflake DW, Oracle Sql, Map Reduce, Kafka 2.8 and Agile process.

**Client: Quotient, Mountain view, CA May 2019 to August 2021**

**Sr. Data Engineer**

**Responsibilities:**

* As a Sr. Data Engineer designed and deployed scalable, highly available, and fault tolerant systems on Azure.
* Lead the estimation, review the estimates, identify the complexities and communicate to all the stakeholders.
* Involved in complete SDLC life cycle of big data project that includes requirement analysis, design, coding, testing and production.
* Defined the business objectives comprehensively through discussions with business stakeholders, functional analysts and participating in requirement collection sessions.
* Implemented end-to-end systems for Data Analytics, Data Automation and integrated with custom visualization tools.
* Exposure to implementation and operations of data governance, data strategy, data management and solutions
* Migrated on-primes environment on Cloud using MS Azure.
* Designed the business requirement collection approach based on the project scope and SDLC (Agile) methodology.
* Moved data to Azure Data Lake to Azure data warehouse using Poly Base.
* Created external tables in ADW with 4 compute nodes and scheduled.
* Extensively used Agile Method for daily scrum to discuss the project related information.
* Worked with data ingestions from multiple sources into the Azure SQL data warehouse
* Transformed and loading data into Azure SQL Database.
* Configured Spark streaming to receive real time data from the Kafka and store the stream data to HDFS.
* Development and maintenance of data pipeline on Azure Analytics platform using Azure Data bricks.
* Developed a data pipeline using Kafka to store data into HDFS.
* Implemented Kafka producers create custom partitions, configured brokers and implemented High level consumers to implement data platform.
* Created Airflow Scheduling scripts in Python.
* Maintained NoSQL database to handle unstructured data, clean the data by removing invalidate data, unifying the format and rearranging the structure and load for following steps.
* Wrote Python scripts to parse XML documents and load the data in database.
* worked on data analysis, data profiling, source to target mapping, Data specification for the conversion process
* Written DDL and DML statements for creating, altering tables and converting characters into numeric values.
* Performed data cleaning and data manipulation activities using NOSQL utility.
* Worked on Data load using Azure Data factory using external table approach.
* Automated recurring reports using SQL and Python.
* Developed purging scripts and routines to purge data on Azure SQL Server and Azure Blob storage.
* Resolved the data type inconsistencies between the source systems and the target system using the Mapping Documents.
* Developed Python Scripts for automation purpose and Component unit testing using Azure Emulator.
* Involved in T-SQL queries and optimizing the queries in Azure SQL Server.
* Maintaining data storage in Azure Data Lake.
* Written and executed customized SQL code for ad hoc reporting duties and used other tools for routine report generation.

**Environment:** Spark, Kafka, Apache Airflow, Azure SQL DB, Azure DW, Azure Data Lake, Azure Data factory, Python, XML, Azure Data bricks, T-SQL and Agile process.

**Client: Merck Pharma, Branchburg, NJ January 2017 to April 2019**

**Data Engineer**

**Responsibilities:**

* As a Data Engineer I was responsible to build a data lake as a cloud based solution in AWS using Apache Spark and Hadoop.
* Involved in Agile methodologies, daily Scrum meetings, Sprint planning.
* Objective of this project is
* Installed and configured Hadoop and responsible for maintaining cluster and managing and reviewing Hadoop log files.
* Used AWS Cloud and On-Premise environments with Infrastructure Provisioning/ Configuration.
* Contributed to the development of key data integration and advanced analytics solutions leveraging Apache Hadoop.
* Wrote complex Hive queries to extract data from heterogeneous sources (Data Lake) and persist the data into HDFS.
* Developed Big Data solutions focused on pattern matching and predictive modeling.
* Strong ungerstanding of Data modeling(Relational, dimensional, Data analysis, implementations of Data warehousing using Windows and UNIX.
* Developed the code for Importing and exporting data into HDFS and Hive using Sqoop
* Developed a data pipeline using Kafka, HBase, Spark and Hive to ingest, transform and analyzing customer behavioral data.
* Developed Spark jobs and Hive Jobs to summarize and transform data.
* Participated in the Data Governance working gorup sessions to create Data Governance policies
* Developed reconciliation process to make sure elastic search index document count match to source records.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster processing of data.
* Implemented Sqoop to transform the data from Oracle to Hadoop and load back in parquet format
* Developed incremental and complete load Python processes to ingest data into Elastic Search from oracle database
* Used Hive to analyze data ingested into HBase by using Hive-HBase integration and compute various metrics for reporting on the dashboard
* Created Hive External tables to stage data and then move the data from Staging to main tables
* Pulled the data from data lake (HDFS) and massaging the data with various RDD transformations.
* Load the data through HBase into Spark RDD and implement in memory data computation to generate the output response.
* Continuously tuned Hive UDF's for faster queries by employing partitioning and **bucketing**.

**Environment**: Hadoop, Spark, Hive, Sqoop, AWS, HBase, Kafka, Python, HDFS, Elastic Search & Agile Methodology

**Client: Brio Technologies Private Limited Hyd India August 2015 to September 2016**

**Data Engineer**

**Responsibilities:**

* Worked on automating the flow of data between software systems using Apache NiFi.
* Prepared workflows for scheduling the load of data into Hive using IBIS Connections.
* Worked on a robust automated framework in Data Lake for metadata management that integrates various metadata sources, consolidates, and updates podium with latest and high-quality metadata using the big data technologies like Hive and Impala.
* Responsible for penetration testing of corporate networks and simulated virus infections on computers to assess network security & presented a report to the Development Team to assess the intrusions.
* Used Spark-SQL to load JSON data and create schema RDD and loaded it into the Hive tables and handled structured data using Spark SQL.
* Loaded the data into Spark RDD and did the memory data computation to generate the Output response.
* Developed Spark scripts by using Scala shell commands as per the requirement.
* Worked on the Pub-Sub system using Apache Kafka during the ingestion process.
* Coordinated with product leads to identify problems with Norton products and acted as a liaison between the Development Team and Quality Team to ascertain the efficiency of the product
* Handled network intrusion data and manipulated spark jobs on the same data to identify most common threats and analyze the aforesaid issues.
* Used Azure Data bricks for fast, easy, and collaborative spark-based platform on Azure.
* Performed complex data analysis in support of adf-hoc and standing customer requests
* Used Data bricks to integrate easily with the whole Microsoft stack.
* Wrote spark SQL and spark scripts (pyspark) in data bricks environment to validate the monthly account level customer data.
* Creating Spark clusters and configuring high concurrency clusters using Azure Data bricks (ADB) to speed up the preparation of high-quality data.
* Spun up HDInsight clusters and used Hadoop ecosystem tools like Kafka, Spark and data bricks for real-time analytics streaming, Sqoop, pig, hive, and Cosmos DB for batch jobs.
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Data bricks.

**Environment:** Hadoop (Cloudera Stack), Hue, Azure, Data bricks, Spark, Kafka, HBase, HDFS, Hive, Pig, Sqoop

**Client: Yana Software Private Limited Hyderabad, India April 2013 to July 2015**

**Hadoop Developer**

**Responsibilities:**

* As a Data Engineer I was responsible for gathering data migration requirements.
* Identified problematic areas and conduct research to determine the best course of action to correct the data.
* Analyzed problem and solved issues with current and planned systems as they relate to the integration and management of order data.
* Involved in Data Mapping activities for the data warehouse.
* Analyzed reports of data duplicates or other errors to provide ongoing appropriate inter-departmental communication and monthly or daily data reports.
* Monitor for timely and accurate completion of select data elements.
* Collected, analyse and interpret complex data for reporting and/or performance trend analysis.
* Monitor data dictionary statistics.
* Involved in analysing and adding new features of Oracle 10g like DBMS\_SHEDULER, Create Directory, Data pump, CONNECT\_BY\_ROOT in existing Oracle 10g application.
* Archived the old data by converting them in to SAS data sets and flat files.
* Extensively used Erwin tool in Forward and reverse engineering, following the Corporate Standards in Naming Conventions, using Conformed dimensions whenever possible.
* Colibra workflow development and configuration based on MS Data Governance approach & Requirements
* Enhance smooth transition from legacy to newer system, through change management process.
* Planned project activities for the team based on project timelines using Work Breakdown Structure.
* Compare data with original source documents and validate Data accuracy.
* Used reverse engineering to create Graphical Representation (E-R diagram) and to connect to existing database.
* Generate weekly and monthly asset inventory reports.
* Created Technical Design Documents, Unit Test Cases.
* Written SQL Scripts and PL/SQL Scripts to extract data from Database to meet business requirements and for Testing Purposes.
* Written complex SQL queries for validating the data against different kinds of reports generated by Business Objects XIR2
* Involved in Test case/ data preparation, execution and verification of the test results
* Created user guidance documentations.
* Created reconciliation report for validating migrated data.

**Environment**: UNIX, Shell Scripting, XML Files, XSD, XML, SAS, PL/SQL, Oracle 10g, Erwin, Autosys.