**Sai Yashwanth**

**saiyashwanth1409@gmail.com 2248012369**

**Professional Summary:**

* Having 5+ years of professional experience in information technology with an expert hand in the areas of Big Data, Hadoop Spark, Hive, Impala, Sqoop, Flume, Kafka, SQL tuning, ETL development, report development, SAS, database development, data modeling and strong knowledge of oracle database architecture.
* Experience in Big Data analytics, Data manipulation, using Hadoop Eco system tools Map - Reduce, HDFS, Yarn/MRv2, Pig, Hive, HDFS, HBase, Spark, Kafka, Flume, Sqoop, Flume, Oozie, Avro, Sqoop, GCP, Azure, Spring Boot, Spark integration with Cassandra, Avro, Solr and Zookeeper.
* Strong experience in migrating other databases to Snowflake.
* Managing Database, Azure Data Platform services (Azure Data Lake(ADLS), Data Factory(ADF), Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes.
* Extensive experience in Text Analytics, generating data visualizations using R, Python and creating dashboards using tools like Tableau, Power BI.
* Created **Snowflake Schemas** by normalizing the dimension tables as appropriate, and creating a Sub Dimension named Demographic as a subset to the Customer Dimension.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy. Experience in working on creating and running Docker images with multiple micro services.
* Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like Map Reduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.
* 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.
* Having extensive experience in Microsoft Azure Cloud Computing, GCP and SQL BI Technologies.
* Hands-on experience in Azure Cloud Services (PaaS&SaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis Services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake.
* Good experience in tracking and logging end to end software applications build using Azure DevOps.
* Used SQL Azure extensively for database needs in various applications.
* Experience in development and design of various scalable systems using Hadoop technologies in various environments mostly using Python. Extensive experience in analyzing data using Hadoop Ecosystems including HDFS, MapReduce, Hive & PIG.
* In - depth knowledge of Apache Hadoop Architecture (1.x and 2.x) and Apache Spark 1.x Architecture.
* Experience building data pipelines using Azure DataBricks and PySpark.
* Developed SQL/HSQL queries inside AzureDatabricks.
* Extensive knowledge in programming languages such as Scala, Python, Java.
* Experience in understanding the security requirements for Hadoop.
* Developed Spark applications using PySpark and Scala in Azure DataBricks Environment.
* Used Azure Pipelines for CI/CD production devops.
* Strong skills in visualization tools Power BI, Confidential Excel - formulas, Pivot Tables, Charts and DAX Commands.
* Experienced with Docker and Kubernetes on multiple cloud providers, from helping developers build and containerize their application (CI/CD) to deploying either on public or private cloud.

**Technical Skills:**

|  |  |
| --- | --- |
| **Big Data/Hadoop Technologies** | Map Reduce, Spark, Spark SQL, Azure, Spark Streaming, Kafka, PySpark, ,Pig, Hive, HBase, Flume, Yarn, Oozie, Zookeeper, Hue, Ambari Server |
| **Languages** | Java, Scala, Python (NumPy, SciPy, Pandas, Genism, Keras), Shell Scripting |
| **NO SQL Databases** | Cassandra, HBase, MongoDB, Maria DB |
| **Web Design Tools** | HTML, CSS, JavaScript, JSP, jQuery, XML |
| **Development Tools** | Microsoft SQL Studio, IntelliJ, Azure Data bricks, Eclipse, NetBeans. |
| **Public Cloud** | EC2, IAM, S3, Auto scaling, Cloud Watch, Route53, EMR, Redshift, GCP |
| **Development Methodologies** | Agile/Scrum, UML, Design Patterns, Waterfall |
| **Build Tools** | Jenkins, Toad, SQL Loader, PostgreSQL, Talend, Maven, ANT, RTC, RSA, Control-M, Oozie, Hue, SOAP UI |
| **Reporting Tools** | MS Office (Word/Excel/Power Point/ Visio/Outlook), SSRS, Cognos |
| **Databases** | Microsoft SQL Server, MySQL, Oracle 11g, 12c, DB2, Teradata, Netezza |
| **Operating Systems** | All versions of Windows, UNIX, LINUX, Macintosh HD, Sun Solaris |

**Education: Masters in computer science from University of Bridgeport**

**Professional Experience:**

**Senior Data Engineer Jan 2022 – Present**

**Ensemble Health Partners, Cincinnati, OH**

**Responsibilities:**

* Import the data from different sources like HDFS/HBase into Spark RDD and developed a data pipeline using Kafka and Storm to store data into HDFS.
* Created yaml files for each data source and including glue table stack creation
* Designed and developed Azure Cloud relational servers and databases analyzing current and future business requirements.
* Worked on migration of data from On-Prem SQL server to Cloud database (Azure Synapse Analytics (DW) & Azure SQL DB).
* Created tabular models on Azure analysis services for meeting business reporting requirements.
* Used Azure services like ADLS and Synapse analytics for small data sets.
* Experience in data transformations using Azure HDInsight, HIVE for different file formats.
* Developed Spark and SparkSQL code to process the data in Apache Spark on Azure HDInsight to perform the necessary transformations based on the STMs developed.
* Worked on Version control systems like Subversion, Azure DevOps, GIT, CVS.
* Developed business intelligence solutions using SQL Server data tools to load data to SQL & Azure Cloud databases.
* Analyzed, designed and built modern data solutions using Azure PaaS services to support visualization of data.
* Converted Talend Job lets to support the snowflake functionality
* Importing data from source flies like flat files using Teradata load utilities like Fast Load, Multiload, and TPump.
* Used Airflow for scheduling the Hive, Spark and Map Reduce jobs.
* Developed Oozie work processes for planning and arranging the ETL cycle. Associated with composing Python scripts to computerize the way towards extricating weblogs utilizing Airflow DAGs.
* Transformed date related data into application compatible format by developing apache Pig UDFs.
* Configure Zookeeper to coordinate and support Kafka, Spark, Spark Streaming, HBase and HDFS
* Programmer analysts with expertise Tableau Servers in ETL, Teradata and other EDW data integrations and developments.
* Integrated Kafka with Spark Streaming for real time data processing.
* Proposed an automated system using Shell script to sqoop the job.
* Worked on Implementation of a log producer in Scala that watches for application logs transform incremental log and sends them to a Kafka and Zookeeper based log collection platform
* Consulting on Snowflake Data Platform Solution Architecture, Design, Development and deployment focused to bring the data driven culture across the enterprises
* Worked on Oracle Databases, Redshift and Snowflakes.
* Documented the requirements including the available code, which should be implemented using Spark, Hive, HDFS, HBase and Elastic Search.
* Transformed and aggregated data for analysis by implementing work flow management of Sqoop, Hive and Pig scripts.
* Involved in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, HBase, Oozie, Zookeeper, SQOOP, flume, Spark, Impala, and Cassandra with Horton work Distribution.
* Developed Kafka consumer API in Scala for consuming data from Kafka topics.
* Developed Map Reduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW
* Developed a strategy for Full load and incremental load using Sqoop.
* Designed and implemented Sqoop for the incremental job to read data from DB2 and load to Hive tables and connected to Tableau for generating interactive reports using Hive server2.
* Used Sqoop to channel data from different sources of HDFS and RDBMS.
* Developed Scala scripts using both Data frames/SQL and RDD/Map Reduce in Spark for Data Aggregation, queries and writing data back into OLTP system through SQOOP.
* Experienced in writing live Real-time Processing and core jobs using Spark Streaming with Kafka as a data pipeline system.
* Integrated Oozie with Hue and scheduled workflows for multiple Hive, Pig and Spark Jobs.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats.
* Involved in modeling different key risk indicators in Splunk and building extensive Hive & SPL quires to understand behavior across the customer life cycle
* Created and Managed Splunk DB connect identities, connections inputs, outputs lookups and access controls.
* Created dashboards, reports and alerts for real time monitoring in Splunk. Tableau and Jasper soft.

**Environment**: Hadoop, Oracle, Scala, Pig, Spark-SQL, PySpark, Python, Kafka, SQL, Oozie, Cosmos, Pig, Sqoop, MS Access, Splunk, MS Azure, HBase, Snowflake, Airflow, Tableau.

**Data Engineer May 2020 -Dec 2021**

**Enquero, Milpitas, CA**

**Responsibilities:**

* Source Data from various feeds-Semi/Structured data including flat files, Parquet, JSON &XML
* Created Snow pipe for continuous data load from staged data residing on cloud gateway server Implemented data transformations on Snowflake using various built-ins like Snow SQL, Snowpipe
* Use DBT cloud to connect to Snowflake and set the development, deployment environments
* Use Flatten object table to produce output in several formats – variant, object & array columns
* Work on Snowflake pipeline to configure data loads from AWS S3 to into Snowflake Virtual DW
* Use JSON &XML for serialization and deserialization to load JSON,XML data into Hive tables.
* Imported data from AWS S3 into Spark RDD’s and performed Transformations & Actions on RDD’s.
* Query performance optimization in Snowflake and involved other teams for user experience
* Extensively used jinja syntax to bring programmability on SQL before writing to Snowflake
* Workon UNIX Shell scripting and batch execution in Putty
* Use SparkSQL to load JSON data & create schema RDD & loaded it into Hive tables
* Responsible for Design, Development, testing of the database and developed Stored Procedures, Views, and Triggers, wrote shell script to trigger DataStage jobs
* Workon integration engagements in services such as EC2, S3, RDS.
* Attending Stand-ups daily and providing updates on assigned stories
* Defining user stories out of features in Program meetings, iteration planning & track them in JiraSprint planning, Story Demonstration and PI planning,
* Worked on PySpark programming daily to perform transformations for applying business logic
* Converting data load pipeline algorithms written in python and SQL to Scala spark and PySpark
* Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries
* Involved in developing PySpark script to encrypting the raw data by using hashing algorithms concepts on client specified columns
* Worked with AWS ecosystem (EC2, S3, Lambda, RDS, Redshift)
* Worked on AWS Data pipeline to configure data loads from S3 to into Redshift
* Using AWS Redshift, Extracted, transformed & load data from various data sources &destinations
* Setting up HBase column-based storage repository for archiving data on daily basis
* Worked with SQL across several dialects (we commonly write HBase, MySQL, PostgreSQL, Redshift, SQL Server, and Oracle), Like Access, Excel, CSV, Oracle, flat files using connectors, tasks and transformations provided by AWSData Pipeline
* Worked in the CI/CD pipeline management for managing the weekly releases
* Provide capacity planning, performance tuning and monitoring
* Attending Stand-ups daily and providing updates on assigned stories
* Defining user stories out of features in Program meetings, iteration planning & track them in Jira
* Sprint planning, Story Demonstration and PI planning.

**Environment:** PySpark, SQL, AWS (EC2, S3, Lambda, RDS, Redshift), Snowflake, Oracle, Unix, PL/SQL, PostgreSQL, REST APIs, dbt, UNIX, Snowflake Web UI, Snow SQL Snowflake – Streams, Snowpipes, ETL, AWSEC2, S3, RDS, Tasks, Snow SQL (Snowflake CLI), SQL, Python, JSON, XML, PowerShell, UNIX

**Big Data Engineer Jan 2019 -May 2020**

**Barclays, Whippany, NJ**

**Responsibilities:**

* Worked in Agile environment, and used rally tool to maintain the user stories and tasks.
* Implemented Apache Sentry to restrict the access on the hive tables on a group level.
* Designed and implemented by configuring Topics in new Kafka cluster in all environment.
* Created multiple dashboards in tableau for multiple business needs.
* Implemented Partitioning, Dynamic Partitions and Buckets in HIVE for efficient data access.
* Designed SSIS Packages to extract, transfer, load (ETL) existing data into SQL Server from different environments for the SSAS cubes (OLAP)
* Extract Transform and Load data from sources systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, Azure Data Lake analytics.
* Designed & implemented database solutions in Azure SQL Data Warehouse, Azure SQL.
* Implemented Composite server for the data virtualization needs and created multiples views for restricted data access using a REST API.
* Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team Using Tableau.
* Migrated Map reduce jobs to Spark jobs to achieve better performance.
* Involved in converting Map Reduce programs into Spark transformations using Spark RDD's using Scala and Python.
* Developed Apache Spark applications by using spark for data processing from various streaming sources.
* Developed data pipeline using Spark, Hive, Pig, python, Impala, and HBase to ingest customer
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Queried and analyzed data from Cassandra for quick searching, sorting and grouping through CQL.
* Joined various tables in Cassandra using spark and Scala and ran analytics on top of them.
* Applied Spark advanced procedures like text analytics and processing using the in-memory processing.
* Implemented Apache Drill on Hadoop to join data from SQL and No SQL databases and store it in Hadoop.
* Brought data from various sources in to Hadoop and Cassandra using Kafka.
* SQL Server reporting services (SSRS). Created & formatted Cross-Tab, Conditional, Drill-down, Top N, Summary, Form, OLAP, Sub reports, ad-hoc reports, parameterized reports, interactive reports & custom reports
* Designing and Developing Oracle PL/SQL and Shell Scripts, Data Import/Export, Data Conversions and Data Cleansing

**Environment:** Hadoop, Map Reduce, HDFS, Hive, Pig, Impala, Kafka, Cassandra, Spark, Scala, Python, SQL, Zookeeper, Sqoop, Azure