**Saiswarup Darshnam**

**Email:** saiswar100@gmail.com

**PH: (754)600-9473**

**Sr Big Data Engineer**

**PROFESSIONAL SUMMARY**

* 9+ years of technical experience in **Analysis, Design, Development** with **Big Data** technologies **like Spark, MapReduce, Hive, Kafka** and **HDFS** including programming languages such as **Python**, **Scala** and **Java**.
* Strong experience in Software Development Life Cycle (**SDLC**) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies.
* Data Engineering professional with solid foundational skills and proven tracks of implementation in a variety of data platforms.
* Strong experience in writing scripts using Python API, **PySpark** API and Spark API for analyzing the data.
* Extensively used Python Libraries PySpark, Pytest, Pymongo, cxOracle, PyExcel, Boto3, Psycopg, embedPy, NumPy and Beautiful Soup.
* Worked in developing a **Nifi** flow prototype for data ingestion in **HDFS**.
* Expertise in Python and Scala, user-defined functions (UDF) for Hive and Pig using Python.
* Experienced in creating shell scripts to push data loads from various sources from the edge nodes onto the **HDFS**.
* Experience in developing **Map Reduce** programs using Apache Hadoop for analyzing the big data as per the requirement.
* Good Experience in implementing and orchestrating data pipelines using Oozie and Airflow.
* Experience working with NoSQL databases like **Cassandra**, **HBase and MongoDB**.
* Good working knowledge of Amazon Web Services (AWS) Cloud Platform which includes services like EC2, S3, VPC, ELB, IAM, DynamoDB, Cloud Front, Cloud Watch, Route 53, Elastic Beanstalk (EBS), Auto Scaling, Security Groups, EC2 Container Service (ECS), Code Commit, Code Pipeline, Code Build, Code Deploy, Dynamo DB, Auto Scaling, Security Groups, Red shift, CloudWatch, CloudFormation, CloudTrail, Ops Works, Kinesis, IAM, SQS, SNS, SES.
* Worked with **Cloudera** and **Hortonworks** distributions.
* Extensive experience working on **spark** in performing **ETL** using **Spark-SQL**, **Spark Core** and **Real-time data** processing using **Spark Streaming**.
* Strong experience working with various file formats like **Avro**, **Parquet**, **Orc**, **Json**, **Csv** etc.
* Experience in developing customized UDF's in Python to extend Hive and **Pig Latin** functionality.
* Extensively worked with **Teradata** utilities Fast export, and Multi Load to export and load data to/from different source systems including flat files.
* Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server, Hive, and Mongo DB using **Python**.
* Proficiency in **SQL** across several dialects (we commonly write MySQL, PostgreSQL, Redshift, SQL Server, and Oracle).
* Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and controlling, granting database access and Migrating On premise databases to Azure Data Lake store using Azure Data factory.
* Worked extensively on **Sqoop** for performing both batch loads as well as incremental loads from relational databases.
* Experience in designing star schema, Snowflake schema for Data Warehouse and ODS architecture.
* Hands on expertise in writing different **RDD** (Resilient Distributed Datasets) transformations and actions using **Scala**, **Python**.
* Proficient **SQL** experience in data extraction, querying and developing queries for a wide range of applications.
* Experience working with **GitHub**, **Jenkins,** and **Maven**.
* Performed importing and exporting data into **HDFS** and Hive using Sqoop.
* Strong experience in analyzing large amounts of data sets writing **PySpark** scripts and **Hive** queries.
* Highly motivated, self-learner with a positive attitude, willingness to learn new concepts and accepts challenges.

**Certifications**

A picture containing application

Description automatically generated

**AWS Certified SysOps Administrator** – Associate

**Certification ID** M2EZX7YKHMB4QJGB

**TECHNICAL SKILLS:**

* **Big Data Ecosystem:** Hive, Spark, MapReduce, Hadoop**,** Yarn, HDFS, Hue, Impala, HBase, Oozie, Sqoop, Pig, Flume, Airflow
* **Hadoop Distribution:** Cloudera, Hortonworks
* **Programming Languages:** Python, Scala, Shell Scripting and Java
* **Methodologies:** Agile/Scrum development, Waterfall model, RAD
* **Build and CICD:** Maven, docker, Jenkins, GitLab
* **Cloud Management:** Amazon Web Services (AWS)- EC2, EMR, S3, Redshift, Lambda, Athena, Microsoft Azure.
* **Databases:** MySQL, Oracle, Teradata
* **NO SQL Databases:** Cassandra, MongoDB and HBase
* **IDE and ETL Tools:** IntelliJ, Eclipse, Informatica 9.6/9.1, Tableau prep.
* **Operating System:** Windows, Unix, Sun Solaris

**PROFESSIONAL EXPERIENCE:**

**Client: Independent Purchasing Cooperative (Subway), Miami, FL October 2019 to Present**

**Sr. Big Data Engineer**

**Responsibilities**

* Responsible for ingesting large volumes of user behavioural data and customer profile data to Analytics Data store.
* Developed custom multi-threaded Java based ingestion jobs as well as **Sqoop** jobs for ingesting from FTP servers and data warehouses.
* Developed **PySpark** and **Scala** based Spark applications for performing data cleaning, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
* Wrote Spark-Streaming applications to consume the data from Kafka topics and write the processed streams to **HBase** and **MongoDB**.
* Worked on troubleshooting spark application to make them more error tolerant.
* Worked on fine-tuning spark applications to improve the overall processing time for the pipelines.
* Wrote **Kafka** producers to stream the data from external rest APIs to Kafka topics.
* Experienced in handling large datasets using Spark in Memory capabilities, using broadcasts variables in **Spark**, effective & efficient Joins, transformations, and other capabilities.
* Good experience with continuous Integration of application using Bamboo.
* Worked extensively with **Sqoop** for importing data from Oracle.
* Created private cloud using **Kubernetes** that supports DEV, TEST, and PROD environments.
* Written HBase bulk load jobs to load processed data to Hbase tables by converting to Hfiles.
* Designing and customizing data models for Data warehouse supporting data from multiple sources on real time.
* Experience working for EMR cluster in **AWS cloud** and working with **S3, Redshift, Snowflake**.
* Wrote **Glue** jobs to migrate data from **hdfs** to **S3** data lake.
* Involved in creating Hive tables, loading, and analysing data using **hive** scripts.
* Implemented Partitioning, Dynamic Partitions, Buckets in **Hive**.
* Designed, documented operational problems by following standards and procedures using **JIRA**.
* Experience in Developing **Spark** applications using **Spark - SQL** in **Databricks** for data extraction.
* Used Reporting tools like **Tableau** to connect with Impala for generating daily reports of data.
* Collaborated with the infrastructure, network, database, application, and BA teams to ensure data quality and availability.

**Environment:** Hadoop, Spark, Scala, Python, Hive, HBase, MongoDB, Sqoop, Oozie, Kafka, Snowflake, Amazon EMR, Glue, YARN, JIRA, amazon S3, Shell Scripting, SBT, GITHUB, Maven.

**Client: ASG, Matawan, NJ May 2016 to September 2019**

**Big Data Engineer**

**Responsibilities:**

* Worked on developing **ETL** processes (Data Stage Open Studio) to load data from multiple data sources to **HDFS** using **FLUME** and **SQOOP**, and performed structural modifications using **Map Reduce**, **HIVE**.
* Worked collaboratively to manage build outs of large data clusters and real time streaming with **Spark**.
* Developed ETL data pipelines using **Spark**, Spark streaming and **Scala**.
* Responsible for loading Data pipelines from web servers using **Sqoop,** **Kafka** and **Spark** Streaming API.
* Have experience of working on **Snowflake** data warehouse.
* Creating Databricks notebooks using SQL, Python and automated notebooks using jobs.
* Used **Spark** for interactive queries, processing of streaming data and integration with popular **NoSQL** database for huge volume of data.
* Implemented large Lambda architectures using **Azure Data** platform capabilities like Azure Data Lake, Azure Data Factory, Azure Data CatLog, HDInsight, Azure SQL Server, Azure ML and Power BI.
* Using **Azure Databricks**, created Spark clusters and configured high concurrency clusters to speed up the preparation of high-quality data.
* Used **Azure Databricks** for fast, easy, and collaborative spark-based platform on Azure.
* Implemented Spark using Scala and **SparkSQL** for faster testing and processing of data.
* Developed various UDFs in **Map-Reduce** and **Python** for Pig and Hive.
* Defined job flows and developed simple to complex **Map Reduce** jobs as per the requirement.
* Developed **PIG UDFs** for manipulating the data according to Business Requirements and worked on developing custom PIG Loaders.
* Installed **Oozie** workflow engine to run multiple Hive and Pig Jobs.
* Designing and Developing **Apache NiFi** jobs to get the files from transaction systems into data lake raw zone.
* Developed **PIG Latin** scripts for the analysis of semi structured data
* Experienced in **Databricks** platform where it follows best practices for securing network access to cloud applications.
* Used Hive and created Hive tables and involved in data loading and writing Hive UDFs.
* Installed and configured Hive, Pig, Sqoop, Flume and Oozie on the **Hadoop** cluster.
* Analysed the SQL scripts and designed it by using **PySpark** SQL for faster performance.
* Used Azure Data Factory, SQL API and MongoDB API and integrated data from MongoDB, MS SQL, and cloud (Blob, Azure SQL DB, cosmos DB)

**Environment:** Spark, Spark Streaming, Apache Kafka, Apache NiFi, Hive, Tez, Azure, Azure Databricks, Azure data grid, Azure Synapse analytics, Azure data catalog, ETL, PIG, PySpark, UNIX, Linux, Tableau, Teradata, Pig, Snowflake, Sqoop, Hue, Oozie, Java, Scala, Python, GIT, GIT HUB

**Client: 3M, Murray, Utah**

**Data Engineer January 2015 to April 2016**

**Responsibilities:**

* Responsible for building an Enterprise Data Lake to bring ML ecosystem capabilities to production and make it readily consumable for data scientists and business users.
* Processing and transforming the data using **AWS EMR** to assist the Data Science team as per business requirement.
* Developing **Spark** applications for cleaning and validation of the ingested data into the **AWS cloud**.
* Working on fine-tuning Spark applications to improve the overall processing time for the pipelines.
* Implement simple to complex transformation on Streaming Data and Datasets.
* Work on analysing **Hadoop** cluster and different big data analytic tools including Hive, Spark, Python, Sqoop, flume, Oozie.6
* Use **Spark** Streaming to stream data from external sources using **Kafka** service and responsible for migrating the code base from **Cloudera Platform** to Amazon EMR and evaluated Amazon eco systems components like RedShift, Dynamo DB.
* Perform configuration, deployment, and support of cloud services in Amazon Web Services (AWS).
* Designing and building multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Confidential **Redshift**.
* Design Develop and test ETL Processes in **AWS Glue** to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Migrate an existing on-premises application to AWS.
* Build and configure a virtual data centre in the Amazon Web Services cloud to support Enterprise Data Warehouse hosting including Virtual Private Cloud, Security Groups, Elastic Load Balancer.
* Implement data ingestion and handling clusters in real time processing using **Kafka**.
* Develop Spark Programs using Scala and Java API's and performed transformations and actions on RDD's.
* Develop Spark application for filtering **Json** source data in AWS S3 and store it into HDFS with partitions and used spark to extract schema of Json files.
* Develop Terraform scripts to create the AWS resources such as EC2, Auto Scaling Groups, ELB, S3, SNS and Cloud Watch Alarms.
* Developed various kinds of mappings with collection of sources, targets and transformations using Informatica Designer.
* Develop Spark programs with **PySpark** and applied principles of functional programming to process the complex unstructured and structured data sets. Processed the data with Spark from Hadoop Distributed File System **(HDFS).**
* Implement Serverless architecture using AWS Lambda with Amazon S3 and Amazon Dynamo DB.

**Environment:** Apache Spark, Scala, Java, PySpark, Hive, HDFS, Hortonworks, Apache HBase, AWS EMR, EC2, AWS S3, AWS Redshift, Redshift Spectrum, RDS, Lambda, Informatica Center, Maven, Oozie, Apache NiFi, CI/CD Jenkins, Tableau, IntelliJ, JIRA, Python and UNIX Shell Scripting

**Client: Dhakshita Sri Software Solutions, Hyderabad, India June 2012 to November 2014**

**Data Engineer**

**Responsibilities:**

* Worked on development of data ingestion pipelines using **ETL** tool, Talend & bash scripting with big data technologies including but not limited to Hive, Impala, Spark, Kafka, and Talend.
* Experience in developing scalable & secure data pipelines for large datasets.
* Gathered requirements for ingestion of new data sources including life cycle, data quality check, transformations, and metadata enrichment.
* Importing data from **MS SQL** server and **Teradata** into **HDFS** using **Sqoop**.
* Supported data quality management by implementing proper data quality checks in data pipelines.
* Enhancing Data Ingestion Framework by creating more robust and secure data pipelines.
* Implemented data streaming capability using **Kafka** and **Talend** for multiple data sources.
* Responsible for maintaining and handling data inbound and outbound requests through big data platform.
* Working knowledge of **cluster security components** like Kerberos, Sentry, SSL/TLS etc.
* Worked with multiple storage formats (**Avro, Parquet**) and databases (**Hive, Impala, Kudu**).
* Involved in the development of agile, iterative, and proven data modeling patterns that provide flexibility.
* Created **Oozie** workflows to automate and productionize the data pipelines.
* Troubleshooted user's analyses bugs (JIRA and IRIS Ticket).
* Involved in developing **spark** applications to perform ELT kind of operations on the data.
* Worked with SCRUM team in delivering agreed user stories on time for every Sprint.
* Worked on analyzing and resolving the production job failures in several scenarios.
* Implemented **UNIX** scripts to define the use case workflow and to process the data files and automate the jobs.
* Knowledge on implementing the JILs to automate the jobs in production cluster.
* Involved in creating **Hive** external tables to perform **ETL** on data that is produced on daily basis.
* Utilized **Hive** partitioning, bucketing and performed various kinds of joins on Hive tables

**Environment**: Spark, HDFS, Hive, Pig, Sqoop, Scala, Kafka, Shell scripting, Linux, Jenkins, Eclipse, Git, Oozie, Talend, Agile Methodology, Teradata.

**Education:**

Bachelor of Technology in Electronics and Communication (2012) from TKR College of Engineering, Hyderabad, India.

Master of Science in Information Systems (2017) from Monmouth University, NJ, USA.