**Yeshasvini**

**yeshasvinirt@gmail.com| +1(972)-696-9417|**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Sr. Data Engineer**

Senior Data Engineer offering Over 8 years of experience in Data Engineering, Data Pipeline Design, Development and Implementation as a Sr. Data Engineer/Data Developer and Data Modeler.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**PROFESSIONAL SUMMARY:**

* Around 8 years of professional experience involving project development, implementation, deployment, and maintenance using Big Data technologies in designing and implementing complete end-to-end Hadoop based data analytical solutions using **HDFS, MapReduce, Spark, Scala, Yarn, Kafka, PIG, HIVE, Sqoop, Flume, Oozie, Impala, HBase.**
* Strong working experience in planning and carrying out of **Teradata** system extraction using **Informatica**, Loading Process and Data warehousing, Large-scale Database Management and Reengineering.
* Highly experienced in creating complex **Informatica mappings** and workflows working with major transformations.
* Worked on **NoSQL** databases including **HBase and Mongo DB**.
* Experience in Implementing and building **CI/CD** pipelines with **Jenkins and AWS.**
* Experience in using **PL/SQL** to write Stored **Procedures, Functions and Triggers**.
* Having proficient experience in various Big Data technologies like **Hadoop, Apache NiFi, Hive Query Language, HBase NoSQL database, Sqoop, Spark, Scala, OOZIE and Pig. Oracle Database and Unix shell Scripting technologies.**
* Experience developing **Pig Latin** and **HiveQL** scripts for **Data Analysis and ETL** purposes and extended the default functionality by writing User Defined Functions (**UDFs**), User Defined Aggregate Function (**UDAFs**) for custom data specific processing.
* Good experience in creating data ingestion **Pipelines**, Data Transformations, Data Management, Data Governance, and real time streaming at an enterprise level.
* Good Understanding of Data ingestion, **Airflow** Operators for Data Orchestration and other related python

libraries

* Hands on expertise with **AWS** Databases such as **RDS** (Aurora), **Redshift**, **Dynamo** DB and **Elastic Cache** (Memcached & Redis).
* Involved in designing and deploying multi-tier applications using all the AWS services like (**EC2, Route53, S3, RDS, Dynamo DB**) focusing on high-availability, fault tolerance, and auto-scaling in **AWS Cloud Formation.**
* Involved in data migration team from on prem **SQL** cluster to **Azure** cloud using Azure Data Factory and did transformations, standardizations and cleaning in Azure Databricks and stored in Azure SQL DB.
* Performed Visualizations using **Tableau**, **Power** **BI** and **Domo** for creating dashboards and gave valuable insights related to business uses.
* Experience in using **SDLC** methodologies like **Waterfall, Agile** Scrum for design and development.
* Experience developing iterative Algorithms using **Spark Streaming** in **Scala** and **Python** to builds near Realtime dashboards.
* Experience with migrating data to and from **RDBMS** and unstructured sources into **HDFS** using **Sqoop**.
* Experience in writing real time query processing using **Cloudera Impala.**
* Profound understanding of Partitions and Bucketing concepts in **Hive** and designed both Managed and External tables in Hive to optimize performance.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**TECHNICAL SKILLS:**

**Databases:** MySQL, Oracle, MongoDB, PostgreSQL, DB2, MS-SQL Server, HBASE

**Big Data Ecosystem:** HDFS, MapReduce, HBase, Pig, Hive, Sqoop, Kafka, Flume, Cassandra, Impala, Oozie, MapR

**Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools:** Databricks, Datalakes, Hadoop Distributed File System (HDFS), Hive, Sqoop, MapReduce, MLlib, Oozie, Linux, PuTTY, Bash Shell, Unix, Tableau, Denodo, Power BI, SAS

**Cloud Technologies:** Microsoft Azure, AWS, Google Cloud Platform

**Programming / Query Languages**: SQL, Python Programming, NoSQL, Spark, SAS, PL/SQL, Linux shell scripts, Scala

**IDE’s:** IntelliJ, Eclipse.

**Development Tools:** Visual Studio, Eclipse

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**PROFESSIONAL EXPERIENCE:**

**Micron - Boise, Idaho Nov 2020 to Till date**

**Sr. Data Engineer**

**Responsibilities:**

* Collecting data from various **Flume** agents that are imported on various servers using Multi-hop flow.
* Ingest real-time and near-real time (NRT) streaming data into **HDFS** using **Flume**.
* Experienced with handling administration activations using **Cloudera** manager.
* Involved in developing **Impala** scripts for extraction, transformation, loading of data into data warehouse.
* Create and maintain optimal data pipeline architecture in **cloud Microsoft Azure** using **Data Factory** and **Azure Databricks**
* Create an Architectural solution that leverages the best **Azure analytics** tools to solve our specific need in Chevron use case
* Develop **Nifi workflow** to pick up the data from rest API server, from data lake as well as from SFTP server and send that to **Kafka** broker.
* Implement **Spark Kafka** streaming to pick up the data from Kafka and send to Spark pipeline.
* Experience in working with different join patterns and implemented both **Map and Reduce** Side Joins.
* Wrote **Flume** configuration files for importing streaming log data into HBase with **Flume.**
* Imported several transactional logs from web servers with **Flume** to ingest the data into **HDFS.**
* Installed and configured **pig**, written Pig Latin scripts to convert the data from Text file to Avro format.
* Created Partitioned **Hive** tables and worked on them using **HiveQL.**
* Involved in data ingestion into **HDFS** using **Sqoop** for full load and Flume for incremental load on variety of sources like web server, **RDBMS** and Data API’s.
* Built the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using **SQL** and ‘big data’ technologies like **Hadoop Hive, Azure Data Lake storage**
* Installed **Oozie** workflow engine to run multiple **Hive** and **Pig** jobs which run independently with time and data availability.
* Loading Data into **HBase** using Bulk Load and Non-bulk load.
* Implemented the workflows using **Apache Oozie** framework to automate tasks.
* Involved in migrating tables from **RDBMS** into **Hive** tables using **SQOOP** and later generate visualizations using Tableau.
* Built pipelines to move hashed and un-hashed data from XML files to **Data lake.**
* Developed **Spark scripts** using **Python** on **Azure HDInsight** for Data Aggregation, Validation and verified its performance over MR jobs.
* Extensively worked with **Spark-SQL** context to create data frames and datasets to pre-process the model data.
* Writing **pyspark** and **spark Sql** transformation in **Azure Databricks** to perform complex transformations for business rule implementation
* Used **Kafka** functionalities like distribution, partition, replicated commit log service for messaging systems by maintaining feeds and Created applications using Kafka, which monitors consumer lag within **Apache Kafka** clusters.
* Worked with **NoSQL** databases like **HBase** in making **HBase** tables to load expansive arrangements of semi structured data.
* Involved in transforming data from Mainframe tables to **HDFS**, and **HBase** tables using Sqoop.
* Acted for bringing in data under **HBase** using HBase shell also **HBase** client API.
* Worked on UDFS using **Python** for data cleansing
* Exporting of a result set from **HIVE** to **MySQL** using **Sqoop** export tool for further processing.
* Designed and implemented Incremental Imports into **Hive** tables and writing Hive queries to run on **TEZ**.
* Created **ETL** Mapping with Talend Integration Suite to pull data from Source, apply transformations, and load data into target database.

**Environment:** Hadoop, Cloudera, Flume, HBase, HDFS, MapReduce, Kafka, YARN, Hive, Pig, Sqoop, Oozie, Tableau, Azure, Data Factory, Databricks, HDInsight, PL/SQL, MySQL, Oracle, TEZ

**Bluecoat systems - Sunnyvale, CA May 2018 to Nov 2020**  **Sr. Data Engineer**

**Responsibilities:**

* Analyzed **SQL** scripts and designed the solutions to implement using **PySpark**.
* Developing data processing tasks using **PySpark** such as reading data from external sources, merge data, perform data enrichment and load in to target data destinations, used Pandas, NumPy, Spark in Python for developing Data Pipelines, also for performing Data Cleaning, features scaling, features engineering • Worked collaboratively to manage build outs of large **data clusters** and real time streaming with **Spark**.
* Implement end-end data flow using **Apache NiFi**.
* Responsible for loading Data **pipelines** from web servers using **Kafka** and **Spark** Streaming API.
* Used **Spark** for interactive queries, processing of streaming data and integration with popular **NoSQL** database for huge volume of data.
* Developed the batch scripts to fetch the data from **AWS S3** storage and do required transformations in **Scala** using **Spark** framework.
* Implemented **Spark** using **Scala** and **SparkSQL** for faster testing and processing of data.
* Monitoring the **Hive** Meta store and the cluster nodes with the help of **Hue**.
* Created **AWS EC2** instances and used JIT Servers.
* Worked with **Snowflake** cloud data warehouse and **AWS** **S3** bucket for integrating data from multiple source system which include loading nested JSON formatted data into Snowflake table.
* Install and configure **Apache Airflow** for S3 bucket and **Snowflake** data warehouse and created **dags** to run the Airflow.
* Automated resulting scripts and workflow using **Apache Airflow** and **shell scripting** to ensure daily execution in production.
* Hands-on experience with **Snowflake** utilities, Snow SQL, SnowPipe, Big Data model techniques using **Python** and Java and Data Integrity checks have been handled using **Hive queries, Hadoop, and Spark.**
* Developed ETL Pipelines in and out of data warehouse, developed major reports using advanced SQL queries in **Snowflake**.
* **ETL** **pipelines** in and out of data warehouse using combination of Python and Snowflakes **SnowSQL** writing **SQL** queries against Snowflake.
* Worked on performing transformations & actions on **RDDs** and **Spark** Streaming data with Scala.
* Responsible in handling Streaming data from web server console logs.
* Defined job flows and developed simple to complex **Map** Reduce jobs as per the requirement.
* Optimized **Map**/Reduce Jobs to use **HDFS** efficiently by using various compression mechanisms.
* Developed **PIG** **UDFs** for manipulating the data according to Business Requirements and worked on developing custom **PIG** **Loaders**, also developed **PIG** **Latin** Scripts for the analysis of semi structured data.
* Installed **Oozie** workflow engine to run multiple **Hive** and **Pig** Jobs
* Used **Hive** and created **Hive** Tables and involved in data loading and writing **Hive** **UDFs**.
* 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**.
* Involved in **NoSQL** database design, integration, and implementation. Loaded data into NoSQL database HBase.
* Developed **Kafka** producer and consumers, **HBase** clients, Spark, and Hadoop MapReduce jobs along with components on HDFS, Hive.
* Following **Scrum** methodology to track the project details and updating the development status every day in the SCRUM.

**Environment**: Spark, Spark Streaming, Snowflake SQL, Airflow, Apache Kafka, Apache NiFi, Hive, Tez, AWS, ETL, PIG, UNIX, Linux, Tableau, Teradata, Pig, Sqoop, Hue, Oozie, Java, Scala, Python, GIT

**Travelport, Englewood, CO. Sept 2015 to May 2018**

**Big Data Engineer**

**Responsibilities**:

* Written **Hive** queries for data analysis to meet the business requirements
* Migrated an existing on premises application to **AWS**,
* Developed **PIG** Latin scripts to extract the data from the web server output files and to load into HDFS
* Created many Sparks **UDF** and **UDAFs** in Hive for functions that were not preexisting in **Hive** and **Spark** **SQL**.
* Involved in converting **Hive/SQL** queries into Spark transformations using Spark **RDDs** and **Scala**.
* Implementing different performance optimization techniques such as using distributed cache for small datasets, partitioning, and bucketing in hive, doing map side joins etc.
* By using **Zookeeper** implementation in the cluster, provided concurrent access for **Hive** Tables with shared and exclusive locking.
* Using Sqoop to import and export data from **Oracle** and **PostgreSQL** into **HDFS** to use it for the analysis.
* Migrated Existing **MapReduce** programs to **Spark** Models using **Python**, also Involved in converting the **HQL's** in to spark transformations using Spark **RDD** with support of python and **Scala**
* Used **Spark** Data Frame API over **Cloudera** platform to perform analytics on hive data, migrating the data from Data Lake (**hive**) into S3 Bucket and validated data present in **Data Lake** and **S3** bucket
* Used Apache **Kafka** for real time data ingestion and created different topic for reading the data in **Kafka**.
* Moved data from **S3** bucket to **Snowflake** Data Warehouse for generating the reports.
* Created database design, converted loaded scripts from **Teradata** to **vertica**.
* Extensively used **Informatica** Workflow manager and Workflow monitor for creating and monitoring workflows, Worklets and sessions.
* Worked with SQL Server Stored Procedures and experienced in loading data into Data Warehouse/Data Marts using **Informatica**, SQL Loader, Export/Import utilities.

**Environment**: Linux, Apache Hadoop Framework, Snowflake, HDFS, YARN, HIVE, HBASE, AWS (S3, EMR), Scala, Spark, SQOOP, Vertica.

**TCS, India April 2013 to June 2015 Data Analyst**

**Responsibilities:**

* Document the complete process flow to describe program development, logic, testing, and implementation, application integration, coding and recommended structural changes and enhancements to systems and Databases.
* Used **MS Excel, MS Access, and SQL** to write and run various queries.
* Worked extensively on creating tables, views, and **SQL** queries in **MS SQL** Server.
* Worked with internal architects and assisting in the development of current and target state data architectures.
* Created **Hive** target tables to hold the data after all the **PIG ETL** operations using **HQL**.
* Analyzed the data by performing **Hive** queries and running **Pig** scripts to know user behavior.
* Integrated **Hadoop** with **Tableau** and **SAS** analytics to provide end users analytical reports
* Coordinate with the business users in providing appropriate, effective, and efficient way to design the new reporting needs based on the user with the existing functionality.
* Remain knowledgeable in all areas of business operations to identify systems needs and requirements.

**Environment**: SQL, SQL Server, MS Office, and MS Visio

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **Education:**

**Jawaharlal Nehru Technological University August 2009- June 2013**