**Name: Haripriya V
Email:phari29056@gmail.com**

**Contact:7205153764**

**PROFESSIONAL SUMMARY:**

Around **10+ years** of IT experience in AWS, Cloud, Analysis, design, development, implementation, Big Data Warehouse, NOSQL, GCP, Hadoop, maintenance and support with experience in developing strategic methods for deploying big data technologies to efficiently solve Big Data processing requirement

* Having Hands-on experience in Talend
* Strong experience in BigData using HADOOP framework and related technologies such as HDFS, Teradata, HBASE, MapReduce, PySpark, Spark, Snowflake, Python, HBase, HIVE, PIG, FLUME, OOZIE, SQOOP, and ZOOKEEPER
* Expertise in Extraction Transformation Loading (ETL) process using Informatica consisting of administration, data transformation, data sourcing, mapping, conversion and loading.
* Performance tuning in Informatica Mappings by identifying the bottlenecks and implemented effective transformation Logic.
* Having excellent experience in Data Pipeline, ETL Tool, Apache Airflow, AWS Glue, Talend, Informatica.
* Hands on experience with AWS services including EMR, EC2, Redshift, Glue, Lambda, SNS, SQS, CloudWatch, Kinesis, Step functions, Managed Airflow instances, Storage & Compute.
* Hands on experience with Azure services including Synapse, Azure Data Factory, Azure functions, EventHub, Stream Analytics, Key Vault, Storage & Compute.
* Hands on experience with GCP services including DataProc, VM, Big Query, Dataflow, Cloud functions, Pub/Sub, Composer, Secrets, Storage & Compute.
* Hands on experience with Databricks services including Notebooks, Delta Tables, SQL Endpoints, Unity Catalog, Secrets, Clusters.
* Have Extensive Experience in IT data analytics projects, Hands on experience in migrating on-premises data & data processing pipelines to cloud including AWS, Azure & GCP.
* Expert level experience in designing, building and managing applications to process large amounts of data in a Hadoop/DevOps (GCP) ecosystem.
* Extensive experience with performance tuning applications on Hadoop/GCP and configuring Hadoop/GCP systems to maximize performance
* Experience in data analysis using HIVE, PIG LATIN, HBASE and custom Map Reduce programs in Java
* Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, NOSQL, Data Bricks and Azure SQL Data warehouse and Controlling and granting database access and Migrating On premise databases to Azure Data lake store using Azure Data factory
* Experience in writing custom UDFs in JAVA and SCALA for HIVE and PIG to extend the functionality.
* Experience with Cloudera and Horton works distributions.
* Experience on SPARK, SCALA, Python, HBASE and KAFKA.
* Developed analytical components using KAFKA, SCALA, SPARK, HBASE and SPARK STREAM.
* Experience in working with Flume to load the log data from multiple sources directly into HDFS
* Excellent experience with data processes, data quality reports
* Performed Hadoop backup Strategy to take the backup of hive, HDFS, HBase, oozie etc.
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems (RDBMS) and from RDBMS to HDFS
* Extensive experience in Strategic development of a Data Warehouse and in Performing Data Analysis and Data Mapping from a Operational Data Store to a Enterprise Data Warehouse
* Involved in creating HDINSIGHT cluster in MICROSOFT AZURE PORTAL also created EVENTSHUB and AZURE SQL DATABASES.
* Worked on a clustered Hadoop for Windows Azure using HDInsight and HORTON WORKS Data Platform for Windows.
* Built real time pipeline for streaming data using EVENTSHUB/MICROSOFT AZURE Queue and SPARK STREAMING
* Loaded the aggregated data into HBase for reporting purpose!
* Read the data from HBase to Spark to perform Join on different tables.
* Created the HBase tables for validation, audit and offset management table.
* Created logical view instead of tables in order to enhance the performance of hive queries.
* Involved in developing Hive DDLS to create, alter and drop Hive tables
* Pretty Good Knowledge on hive Optimization techniques like Vectorization and column-based optimization.
* Written oozie workflow to invoke the Jobs in predefined Interval.
* Expert in scheduling Oozie coordinator based on input data events it starts Oozie workflow when input data is available.
* On Other Hand working on POC with Kafka and NIFI to pull the real-time events into Hadoop Box.
* Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, SPARK-SQL, DATA FRAME, PAIR RDD'S and YARN.
* Experienced in managing Hadoop Cluster using HORTONWORKS AMBARI
* experience in design, develop and maintain BI Solutions using, Amazon S3 & Redshift.
* Strong Experience in the Business Requirements Analysis, design, conversion of business requirements into high level functional specifications.
* Ability to interact with various groups ranging from Business Users, Business IM teams, Analysts. Had long lasting relationships with Clients in all previous engagements.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Technologies** | Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Zookeeper, Apache Flume, Apache Airflow, Cloudera, HBase |
| **Programming Languages** | Python, PL/SQL, SQL, Scala, C, C#, C++, T-SQL, Power Shell Scripting, JavaScript |
| **Cloud Services** | Azure Data Lake Storage Gen 2, Azure Data Factory, Blob storage, Azure SQL DB, Databricks, Azure Event Hubs, AWS RDS, Amazon SQS, Amazon S3, AWS EMR, Lambda, AWS SNS, GCP Big Query, Dataproc, Data Flow, Cloud Composure, Cloud Functions. |
| **Databases** | MySQL, SQL Server, Oracle, MS Access, Teradata, and Snowflake |
| **NoSQL Data Bases** | MongoDB, Cassandra DB, HBase |
| **Development Strategies** | Agile, Lean Agile, Pair Programming, Waterfall, and Test-Driven Development. |
| **Visualization & ETL tools** | Tableau, Informatica, Talend, SSIS, and SSRS |
| **Version Control & Containerization tools** | Jenkins, Git, and SVN |
| **Operating Systems** | Unix, Linux, Windows, Mac OS |
| **Monitoring tool** | Apache Airflow, Control-M |

**PROFESSIONAL EXPERIENCE**

**Fifth Third Bank, Cincinnati, OH July 2021 - Present**

**AWS Data Engineer**

**Responsibilities:**

* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats.
* Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift
* Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra
* Develop Spark streaming application to read raw packet data from Kafka topics, format it to JSON and push back to Kafka for future use case’s purpose.
* Collected data using Spark Streaming from AWS S3 bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in HDFS.
* Designed and Developed ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
* Involved in file movements between HDFS and AWS S3 and extensively worked with S3 bucket in AWS.
* Experience with setting up and managing Databricks clusters for data processing and analysis.
* Proficient in working with Databricks notebooks to develop, test, and deploy ETL pipelines using languages such as Python, Scala, or SQL.
* Strong proficiency in Scala, with experience using it to build data processing pipelines, ETL workflows, and other data engineering tasks.
* Experience with AWS services related to data processing, such as Amazon EMR, Amazon S3, and AWS Glue, as well as proficiency in scripting languages such as Python and Bash for data processing and automation tasks.
* Create Snapshots of EBS Volumes and Monitor AWS EC2 Instances using Cloud Watch and worked on AWS Security Groups and their rules.
* Developed calculated columns and measures using complex DAX functions and power pivot.
* Validated the test data in DB2 tables on Mainframes and on Teradata using SQL queries.
* Knowledge of configuring Databricks jobs and scheduling them to run on a regular basis to ensure the timely processing of data.
* Expertise in integrating Databricks with other tools such as Apache Spark, Apache Hadoop, and cloud services like AWS.
* Authoring Python (PySpark) Scripts for custom UDF’s for Row/ Column manipulations, merges, aggregations, stacking, data labeling and for all Cleaning and conforming tasks.
* Worked on AWS CLI Auto Scaling and Cloud watch monitoring creation and update.
* Handled AWS Management Tools as Cloud watch and Cloud Trail.
* Stored the log files in AWS S3. Used versioning in S3 buckets where the highly sensitive information is stored.
* Integrated AWS Dynamo DB using AWS lambda to store the values of items and backup the DynamoDB streams.
* Prepared scripts to automate the Ingestion process using Pyspark and Scala as needed through various sources such as API, AWS S3, Teradata and Redshift.
* Used SQL Server Management Tool to check the data in the database as compared to the requirement given.
* Played a lead role in gathering requirements, analysis of entire system and providing estimation on development, testing efforts.
* Involved in designing different components of system like Sqoop, Hadoop process involves map reduce & hive, Spark, FTP integration to down systems.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS EMR
* Strong experience and knowledge of real time data analytics using Spark Streaming, Kafka, and Flume Configured Spark streaming to get ongoing information from the Kafka and store the stream information to HDFS.
* Successfully completed a POC on GCP services such as Big Query, Dataflow, Pub/Sub, and Cloud Storage, demonstrating the ability to quickly learn and work with new cloud platforms.
* Leveraged existing data engineering experience in AWS to identify similarities and differences between AWS and GCP services and make informed decisions regarding the most appropriate GCP services for the POC.
* Enforced standards and best practices around data catalog, data governance efforts.
* Created DataStage jobs using different stages like Transformer, Aggregator, Sort, Join, Merge, Lookup, Data Set, Funnel, Remove Duplicates, Copy, Modify, Filter, Change Data Capture, Change Apply, Sample, Surrogate Key, Column Generator, Row Generator, Etc.
* Expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Excellent problem-solving skills to identify and resolve issues related to data pipelines, performance, and scalability on Databricks.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Created Unix Shell scripts to automate the data load processes to the target Data Warehouse.
* Responsible for implementing monitoring solutions in Ansible, Terraform, Docker, and Jenkins.

**Environment:** Python, Spark, AWS EC2, AWS S3, AWS EMR, AWS Redshift, AWS Glue, AWS RDS, AWS SNS, AWS SQS, AWS Athena, Snowflake, Data warehouse, Airflow, Data Governance, Kafka, ETL, Terraform, Docker, SQL, Tableau, Git, REST, Bitbucket, Jira.

**Fidelity National Information Services Inc, Jacksonville, FL Mar 2019 – July 2021**

**Azure Data Engineer**

**Responsibilities:**

* Proficient in designing and implementing end-to-end data solutions on Azure, leveraging services such as Azure Data Factory, Azure Databricks, Azure Synapse Analytics, and Azure Data Lake Storage.
* Extensive experience working with Apache Spark, including Spark programming using PySpark, Spark SQL, and Scala, for large-scale data processing and analytics.
* Deep knowledge of Apache Hive, including HiveQL, Hive partitioning, and performance tuning, to enable efficient data querying and processing on big data sets.
* Strong understanding of Kafka and its integration with Azure services, facilitating real-time data streaming and event-driven architectures.
* Developed scalable and fault-tolerant data pipelines using Apache Airflow, enabling workflow orchestration, scheduling, and monitoring of data processing tasks.
* Proficient in Hortonworks Data Platform (HDP) and its components, including Hadoop Distributed File System (HDFS) and YARN, for managing and processing large volumes of data.
* Expertise in building and optimizing Data Warehouse solutions on Azure, utilizing Azure SQL Data Warehouse or Azure Synapse Analytics for high-performance analytics and reporting.
* Skilled in working with various databases, including Teradata, Oracle, and MySQL, for data extraction, transformation, and loading (ETL) processes.
* Experience with Snowflake, a cloud-based data warehousing platform, for building scalable and elastic data storage and analytics solutions.
* Proficient in using Python and Scala for data manipulation, analysis, and scripting tasks within the Azure ecosystem.
* Strong understanding of data modeling principles and best practices, including both relational and dimensional models, to design efficient and performant data structures.
* Hands-on experience with Azure Data Bricks, an Apache Spark-based analytics platform, for collaborative data exploration, machine learning, and advanced analytics.
* Proficient in using Visual Studio and Azure DevOps for version control, continuous integration, and deployment of data engineering solutions.
* Skilled in Unix Shell Scripting, enabling automation and orchestration of data processing tasks and system administration.
* Solid understanding of data governance and security principles, including Azure Active Directory integration, role-based access control (RBAC), and data encryption.
* Experience in implementing data quality checks and validation processes to ensure the accuracy and reliability of data pipelines.
* Strong troubleshooting and debugging skills to identify and resolve issues within complex data engineering workflows.
* Proficient in working with structured, semi-structured, and unstructured data formats, including JSON, XML, Parquet, Avro, and CSV.
* Hands-on experience with data integration and transformation using Azure Data Factory, enabling seamless data movement across different sources and targets.
* Knowledge of NoSQL databases, such as Azure Cosmos DB or MongoDB, for handling unstructured and schema-less data storage and retrieval.
* Expertise in performance tuning and optimization techniques for big data processing workloads, including query optimization, partitioning, and indexing strategies.
* Familiarity with data governance frameworks and data cataloging tools to ensure compliance, data lineage, and metadata management.
* Experience in implementing real-time data analytics solutions using Azure Stream Analytics and event-driven architectures.
* Strong understanding of cloud computing concepts and hands-on experience in deploying data engineering solutions on Azure cloud platform.
* Excellent communication and collaboration skills, with a proven track record of working effectively in cross-functional teams and managing stakeholder expectations.
* Proactive approach to learning and keeping up to date with the latest advancements in Azure services and data engineering technologies.
* Demonstrated ability to analyze business requirements and translate them into technical data engineering solutions.
* Solid knowledge of Agile and DevOps methodologies, with experience in working in Agile development environments.
* Strong problem-solving and analytical thinking skills, with a focus on delivering high-quality and scalable data solutions.
* Track record of successful project delivery, meeting deadlines, and achieving business objectives in data engineering projects.

**Environment:**  Azure Data Factory, Azure Databricks, Azure Synapse Analytics, and Azure Data Lake Storage, Azure Blob storage, Azure Cosmos DB, Spark, Hive, PySpark, Spark SQL, Kafka, Scala, Airflow, Horton Works, Hive, Python, Scala, Teradata, Snowflake, Data Warehouse, Data bricks, oracle, mysql, Visual Studio, Unix Shell Scripting.

**Mastercard, Saint Louis, MO Aug 2017- Mar 2019**

**Big Data Engineer**

**Responsibilities:**

* Import data from sources like HDFS/HBase into Spark RDD.
* Usage of Spark Streaming and Spark SQL API to process the files.
* Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems/mainframe and vice-versa loading data into HDFS.
* Stored data in AWS S3 like HDFS and performed EMR programs on data stored in S3.
* Worked on Big Data Hadoop cluster implementation and data integration in developing large-scale system software.
* Developing UDFs in java for hive and pig and worked on reading multiple data formats on HDFS using Scala.
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Hive.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
* Involved in Migrating the platform from Cloudera to EMR platform.
* Developed analytical component using Scala, Spark and Spark Streaming.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run MapReduce jobs in the backend.
* Extensively involved in developing Restful API using JSON library of Play framework.
* Developed Storm topology to ingest data from various source into Hadoop Data Lake.
* Developed web application using HBase and Hive API to compare schema between HBase and Hive tables.
* Played a vital role in Scala/Akka framework for web based applications
* Connected to AWS s3 using SSH and ran spark-submit jobs
* Developed Python Script to import data SQL Server into HDFS & created Hive views on data in HDFS using Spark.
* Expert in Troubleshooting MapReduce Jobs.
* Created scripts to append data from temporary HBase table to target HBase table in Spark.
* Developed complex and Multi-step data pipeline using Spark.
* Worked on Big Data Integration and Analytics based on Hadoop, SOLR, Spark, Kafka, Storm and web Methods technologies.
* Populated HDFS and Cassandra with huge amounts of data using Apache Kafka.
* Monitoring YARN applications. Troubleshoot and resolve cluster related system problems.
* Upgrading the Hadoop Cluster from CDH3 to CDH4, setting up High Availability Cluster and integrating HIVE with existing applications.
* Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
* Involved in creating ETL flow using Pig, loading with data and writing Pig Latin queries which will run internally in Map Reduce way.
* Involved in writing Unix/Linux Shell Scripting for scheduling jobs and for writing pig scripts and hive QL.
* Involved in creating Hive Tables, loading with data and writing Hive queries which will invoke and run MapReduce jobs in the backend.
* Assisted in exporting data into Cassandra and writing column families to provide fast listing outputs.
* Used Zookeeper for providing coordinating services to the cluster.
* Worked with Hue UI in scheduling jobs with ease and File browsing, Job browsing, Meta store management.
* Developed and designed system to collect data from multiple portal using Kafka and then process it using spark.

**Environment**: Hadoop, HDFS, Hive, Core Java, Sqoop, Spark, Scala, Hive, Cloudera CDH4, Oracle, Elastic search, Kerberos, SFTP, Impala, Jira, Wiki, Alteryx, Teradata, Shell/Perl Scripting, Kafka, AWS EC2, S3, EMR, Cloudera.

**Arena ITsoft Consultancy pvt ltd, India Aug 2014 – Feb 2017**

**ETL Developer**

**Responsibilities:**

* Analyzed business requirements and worked closely with the various application teams and business teams to develop ETL procedures that are consistent across all applications and systems.
* Involved in the Data Analysis for source and target systems and good understanding of Data Warehousing concepts, Staging Tables, Dimensions, Facts and Star, Snowflake Schemas.
* Installation of Informatica PowerCenter 9.x and setup services and client tools.
* Creation of ETL framework and review of high-level design document
* Creation of Logical Data Model and Physical Data Model
* Extensively worked in the performance tuning of the programs, ETL Procedures and processes
* Responsible for kinds ETL activities such as normalization, cleansing, scrubbing, aggregation, summarizing, and integration.
* Worked in importing and cleansing of data from various sources like Teradata, Oracle, flat files, SQL Server, semi structured and unstructured files, XML files and MS WORD with high volume data.
* Involving in Supporting ETL Applications production jobs and changing existing code based on business needs.
* Develops logical and physical data flow models for ETL applications. Translates data access, transformation and movement requirements into functional requirements and mapping designs.
* Documented both High level and Low-level design documents, Involved in the ETL design.
* Implemented Type 4 and Type 6 slowly changing dimensions.
* Developed complex Multi load and Fast Load scripts for loading data into Teradata tables from legacy systems.
* Involved in various phases of the software development life cycle right from Analysis, Design, Development, and Testing and Production support.
* Performed and documented the unit testing for validation of the mappings against the mapping specifications documents.
* Performed production support activities in Data Warehouse (Informatica) including monitoring and resolving production issues, pursue information, bug fixes and supporting end users.

**Environment:** Informatica PowerCenter 8.5, UNIX, Oracle SQL.

**Anblicks, India Sep 2012 – Aug 2014**

**Data Analyst**

**Responsibilities:**

* Migrated Excel reports and Pyramid analytics reports to Power BI.
* Developed and maintained multiple Power BI dashboards/reports.
* Created Datasets using different data sources like Microsoft SQL Server, Microsoft Azure and Excel
* Created Workspace and Apps in Power BI and giving access to end users to Apps.
* Configured Gateways in the Power BI service.
* Implemented Data security (Row level security) on Dataset and Reports in Power BI.
* Responsible for Scheduling jobs for weekly, daily data refresh on different Power BI reports.
* Responsible for building Power BI Dashboards using Tabular and Multi-Dimensional Models.
* Designed a Power BI Data Model with multiple Fact Tables and Dimensions depending on the business requirements.
* Migrated existing Tableau Visualizations to Power BI using Power BI desktop.
* Developing and Maintaining of Data Dashboards, Reports and Data Visualization tools and Products using SSRS, Power BI, Visio, and Excel.
* Created Calculated Columns and Measures in Power BI depending on the requirement using DAX queries.
* Created complicated DAX measures in the Power BI reports as per the requirements of the Business Users.
* Generated Monthly, Yearly, Quarterly and Daily reports in Power BI.
* Worked on Power BI, Power Pivot and Power View to visualize reports.
* Understanding the existing data and create the required data models for the development.
* Created Databases, Data Mart Tables, Views, Stored Procedures, Triggers, Cluster/Non-Cluster Index, Unique/Check Constraints.
* Involved in complete SSIS life cycle in creating SSIS packages, building, deploying, and executing the packages in both the environments (Development and Production).
* Designed SSIS packages to transfer the data from flat files, Access database and excel documents to the Staging Area in SQL Server 2016.
* Developed VB.NET and C# application to load data from multiple excel files.
* Developed master ETL packages for loading and migrating historical data from Legacy system into D/W.
* Using SQL Server Integration Services (SSIS) as an ETL Tool, the data has Extracted,
* Transformed and Loaded (ETL) from Excel and Flat file to MS SQL Server.
* Developed SSIS packages using looping, data formatting, check points, error handling, email.
* notification etc.
* Create and optimize complex T-SQL queries and stored procedures to encapsulate reporting logic.
* Developed complex reports with multiple datasets and sources in SSRS.
* Optimized SSRS reports with snapshots, caches, execution log views, embedded queries and stored procedures.
* Involved in Data scrubbing/Cleansing/validating/Translating, and scheduling SSIS packages.

**Environment:** MS SQL Server 2012, Visual Studio 2012, SSIS, SSRS, Power BI, JIRA, Power Pivot, Skype for Business, Tableau, Agile Scrum, Web Services, T-SQL, MS Excel.

B.Tech , Velagapudi Ramakrishna Siddhartha Engineering College 2012