**HARISH**

**Email:** [**harishk2559@gmail.com**](mailto:harishk2559@gmail.com) **PH: 913-391-3137**

**LinkedIn:** [**https://www.linkedin.com/in/harish-k-2559hk**](https://www.linkedin.com/in/harish-k-2559hk)

**Senior Data Engineer**

**Professional Summary**

* 10 years of experience as a Data Engineer and Data Analyst, with an expertise in Data Mapping, Data Validation, and dealing with statistical data analysis such as transforming business requirements into analytical models, designing algorithms, machine learning, and strategic solutions that scale across massive volumes of data. Experience with a various BIGDATA technologies, tools, and databases **Spark, Hive, python, SQL, AWS, Snowflake, Hadoop, Sqoop, CDL(Cassandra), Teradata, Tableau, and Redshift**, but always making sure of living in the world, I cherish most i.e., DATA WORLD.
* Strong experience in Software Development Life Cycle (**SDLC**) including Requirements Analysis, Design Specification, application development, application migration and maintenance and Testing as per Cycle in both Waterfall and Agile methodologies.
* Strong experience in writing scripts using **Python API, Spark API** for analyzing the data; Hands-On experience on **Spark Core, Spark SQL, Spark Streaming,** and creating the Data Frames handle in SPARK with Scala.
* Involved in setting up **Jenkins** Master and multiple slaves for the entire team as a CI tool as part of Continuous development and deployment process
* Used AWS glue catalog with crawler to get the data from S3 and perform sql query operations.
* Installed and configured **Apache Airflow** for workflow management and created workflows in python, created the DAG’s using Airflow to run jobs sequentially
* 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.
* Hands on expertise with AWS Databases such as RDS(Aurora), Redshift, DynamoDB and Elastic Cache.
* Performed Raw data ingestion into S3 from kinesis firehouse which would trigger a lambda function and pit refined data into another S3 bucket and write to SQS queue as aurora topics.
* Very Good experience working in **Azure Cloud, Azure DevOps, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos NO SQL DB, Azure HD Insight, Big data Technologies (Hadoop and Apache Spark), and Data bricks.**
* In-depth knowledge of **Google Cloud Platform, GCP Data processing, Cloud functions and Google Big Query**. Built data pipelines in airflow in GCP for ETL-related jobs using different airflow operators.
* Conducted linear regression to predict teh transaction volume, and distinguished frequent claimers based on MapReduce using Hadoop and R studio.
* Used python Boto 3 to configure the services AWS glue, EC2, S3.
* Designed and implemented end-to-end systems for DataAnalytics and Automation, integrating custom visualization tools using R Studio, Hadoop, MongoDB, Tableau, and Power BI.
* Strong experience in the Analysis, design, development, testing, and implementation of Business Intelligence solutions using Data Warehouse/Data Mart Design, ETL, BI, Client/Server applications and writing ETL scripts using Regular Expressions and custom tools (**Informatica**, **Pentaho**, and **Sync Sort**) to ETL data.
* Strong experience in migrating other databases to **Snowflake**. In-depth knowledge of Snowflake Database, **Schema** and Table structures. Skilled in designing and implementing ETL pipelines for snowflake schema databases, employing T-SQL and SSIS to ensure robust data integration and transformation processes.
* Solid experience of creating PL/SQL packages, Procedures, Functions, Triggers and Views to Retrieve, Manipulate and Migrating Complex Data sets in Oracle Databases. Extensive working noledge in developing Database Triggers, Stored Procedures, Functions for developing Forms and Reports.
* Worked on advanced PL/SQL constructs like Oracle-supplied packages, nested tables, arrays, records, types, dynamic SQL and analytical functions.
* Thorough knowledge of Features, Structure, Attributes, Hierarchies, and Star and Snowflake Schemas of Data Marts.
* Shared data outside using Snowflake to quickly set up to share data without transferring or developing pipelines.
* Administered and troubleshoot PostgreSQL databases for critical problems.
* Big data on AWS cloud services i.e. EC2, S3, Glue, Anthena, DynamoDB and RedShift.
* Hands on Experience with dimensional modeling using star schema and snowflake models. Experienced in Optimizing the Pyspark jobs to run on **Kubernetes** Cluster for faster data processing
* Firm understanding of Hadoop architecture and various components including **HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce programming.**
* Design, develop, and deploy ETL processes using Pentaho Data Integration (PDI) to extract, transform, and load data from various sources into data warehouses and data marts.
* Defined user stories and driving the agile board in **JIRA** during project execution, participate in **sprint demo** and retrospective.
* Experience in development of DataStage job sequencers for complex processes.
* Experience in writing distributed Scala code for efficient big data processing.
* Experience building distributed high-performance systems using Spark and Scala.
* Develop ETL mappings for various Sources (.TXT, .CSV, XML) and load the data from these sources into relational tables with Talend Enterprise Edition.
* Extensively Used Talend components tMap, tDie, tConvertType, tFlowMeter, tLogCatcher, tRowGenerator, tOracleInput, tOracleOutput, tfileList etc.
* Good understanding of web design based on **HTML5, CSS3, and JavaScript**.
* Good knowledge of Big Data ecosystem **like Hadoop 2.0 (HDFS, Hive, Pig, Impala), Spark (Spark SQL, Spark Streaming)**.
* Excellent performance in building, publishing customized interactive reports and dashboards with customized parameters including producing tables, graphs, listings using various procedures and tools such as **Tableau** and user-filters using Tableau.
* **TECHNICAL KNOWLEDGE:**

|  |  |
| --- | --- |
| **BigData/Hadoop Technologies** | MapReduce, Spark, SparkSQL, Azure, Spark Streaming, AWS, Kafka, PySpark, Airflow, Pig, Hive, Oozie, Spark |
| **Scripting languages** | HTML5, CSS3, XML, JAVA, Scala, Python, Shell Scripting |
| **Development Tools** | Microsoft SQL Studio, Azure, Databricks, Eclipse, Azure Data Lake |
| **Cloud Technologies** | Microsoft AZURE, AWS (Amazon Web Services) |
| **Public Cloud** | EC2, S3, Autoscaling, CloudWatch, RedShift |
| **Reporting Tools** | MS Office (Word/Excel/PowerPoint/ Visio/Outlook), Power BI, Tableau |
| **Databases** | Microsoft SQL Server, MySQL, Oracle, Teradata, Netezza, Teradata |
| **ETL Tools** | Informatica, Pentaho |
| **Operating Systems** | All versions of Windows, UNIX, LINUX. |

**Professional Experience**

**UBS Weehawken, NJ Nov 2021 to Present**

**Senior Data Engineer**

**Project Description:**

UBS is a financial services company that uses data and technology to support its operations and provide innovative solutions to its clients. my role is to identify a sizable amount of data, including information from financial transactions, credit reports, and news items, that needs to be gathered, processed, and analyzed. I have to collaborate with the risk management team for their specific data needs and design and implement data pipelines that collect, process, and store this data using **Azure Blob Storage** and **ADLS**.

**Responsibilities:**

* Very Good experience working in **Azure Databricks, Azure Cloud, Azure DevOps, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Azure Analytical services, Azure Cosmos NO SQL DB, Azure HD Insight Big data Technologies (Hadoop and Apache Spark).**
* Experience utilizing **Azure Databricks, Azure SQL, PostgreSQL, and SQL Server** to Extract, Transform, and Load data from a wide variety of sources into target databases
* Used **Spark SQL** to create ETL solutions in Azure Databricks for data extraction, transformation, and aggregation from a variety of file formats and data sources for analyzing and transforming the data to reveal insights on client usage patterns.
* Worked on Azure Data Factory to integrate data of both on-premises (MYSQL, Cassandra) and cloud (Blob storage, Azure SQL DB) and applied transformations to load back to snowflake.
* Hands-on working knowledge of creating and deploying Databricks-based big data analytics apps, and optimizing the Data Analytics system on the Azure platform.
* Migrated some of the existing pipelines to Azure Databricks using **PySpark Notebooks** for the analytical team.
* Performed systems integration design and development in cloud architecture design (Azure)
* Created the automated build and deployment process for application, application setup for better user experience, and leading up to building a **continuous integration system**.
* Worked on developing a Pyspark script to encrypt the raw data by using Hashing algorithms concepts on client-specified columns.
* Worked with Azure Logic Apps administrators to monitor and troubleshoot issues related to process automation and data processing pipelines.
* Performed review and analysis of the detailed system specifications related to the DataStage ETL and related applications to ensure they appropriately address the business requirements.
* Evaluated impact of proposed changes on existing DataStage ETL applications, processes and configurations
* Developed Autosys jobs for scheduling and running the DataStage jobs in Production Environment.
* Worked on analyzing the Hadoop cluster using different big data analytic tools including **Pig, Hive, and Map Reduce**.
* Worked with different source data file formats like **JSON, CSV, TSV**, etc.
* Experience in importing data from various data sources like **MySQL and Netezza using Sqoop**, and SFTP performed transformations using Hive, Pig and loaded data back into HDFS.
* Developed spark applications in Python (PySpark) on distributed environment to load huge number of CSV files with different schema into Hive ORC tables.
* Creating PL/SQL objects like procedures, Packages, triggers wherever required.
* Rectified the poorly written SQL & PL/SQL statements and resolved the performance Issues.
* Azure PaaS Solutions like Azure Web Apps, Web Roles, Worker Roles, SQL Azure, and Azure Storage.
* Experience in loading the tables from **Azure Data Lake to Azure blob storage** for pushing them to Snowflake.
* Developed ETL pipelines into and out of the data warehouse and created significant financial and regulatory reports utilizing Snowflake's sophisticated SQL queries.
* Creating PostgreSQL databases, Tables and doing user adminstration/security by creating and deleting users.
* Backing up and restoring databases in PostgreSQL.
* Processed and loaded various gold layer tables from **Delta Lake** into Snowflake & integrating Azure Databricks with **Snowflake**.
* Experience developing Scala applications for loading/streaming data into NoSQL databases (MongoDB) and HDFS.
* Used Spark and Scala for developing machine learning algorithms which analyses click stream data.
* Developed and implemented R Studio which showcases machine learning for business forecasting.
* Generated graphs and reports using ggplot package in R Studio for analytical models.
* Converted applications that were on MapReduce to PySpark which performed the business logic.
* Designing and Developing **Oracle PL/SQL and Shell Scripts**, Data Import/Export, Data Conversions, and Data Cleansing.
* Responsible for importing data from **PostgreSQL to HDFS, HIVE using SQOOP tool, HBase using Spark.**
* Integrated Flume with **Kafka**, and Worked on monitoring and troubleshooting the **Kafka-Flume-HDFS** data pipeline for real-time data ingestion in HDFS

**Environment**: Azure Databricks, Spark, Hive, HBase, Sqoop, Flume, MapReduce, HDFS, SQL, Apache Kafka, Scala, Apache Airflow, Snowflake, Spark, Azure Cloud, Azure Data Factory, Azure Data Lake Storage, Azure Synapse Analytics, Python.

**Kroger, Cincinnati, OH December 2019 to October 2021**

**Data Engineer**

**Responsibilities:**

* Implemented Bigdata technologies such as **Hadoop, Map Reduce Frameworks, HBase, and Hive** for ingesting data from diverse sources and processing Data-at-Rest.
* Optimizing of existing algorithms in Hadoop using **Spark Context, Spark-SQL, Data Frames and Pair RDDs**.
* Used Hadoop technologies like spark and hive Including using the PySpark library to create spark data frames and converting them to normal Pandas data frames for analysis.
* Used various python libraries like **Pandas, NumPy, Matplotlib, SciPy**, etc., for Data cleaning, features scaling, and features engineering.
* Developed Proof of Concept POC for DataStage to SSIS migration.Worked with DataStage Designer for importing metadata from repository, new job categories and creating new data elements
* Extensively performed large data read/write to and from CSV and Excel files using pandas.
* Handled multiple operations of data sets such as sub setting, slicing, filtering, group by, re-ordering, and re-shaping using python libraries like Pandas.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage and experienced in maintaining the Hadoop cluster on AWS EMR. Designed and build a Data Lake using Hadoop and its ecosystem components.
* Developed Airflow jobs to ingest data from RDBMS Systems like Teradata and Oracle database into S3 buckets.
* Experienced in configuring Apache Airflow for S3 bucket and Snowflake data warehouse and created DAGS to run the Airflow
* Was responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using Cloudwatch.
* Used AWS Glue for the data transformation, validate and data cleansing.
* Automatically scale-up the EMR instances based on the data. And stored the time-series transformed data from the Spark engine built on top of a Hive platform to Amazon S3 and Redshift.
* Worked extensively with AWS services like **EC2, S3, VPC, ELB, Auto Scaling Groups, Route 53, IAM, CloudTrail, CloudWatch, CloudFormation, CloudFront, SNS, and RDS.**
* Designed the data models to be used in data intensive AWS Lambda applications which are aimed to do complex analysis creating analytical reports for end-to-end traceability, definition of Key Business elements from Aurora
* Developed Python scripts to parse XML, and JSON files and load the data in AWS Snowflake Data warehouse.
* Experience integrating data from various source systems, including importing nested JSON formatted data into Snowflake tables, using cloud data warehouses like Snowflake and AWS S3 buckets.
* Showed dynamic visualization of clustering of user& coverage and network usage based on area by using R Studio.
* Involved in continuous enhancements and fixing of production issues. Generated server-side PL/SQL scripts for data manipulation and validations.
* Created PL/SQL scripts to extract data from operational database into simple flat text files using UTL FILE package.
* Used Bulk Collections for better performance and easy retrieval of data, by reducing context switching between SQL and PL/SQL engines.
* Designed and Implemented the ETL process using Talend Enterprise Big Data Edition to load the data from Source to Target Database. Using Talend to load the data into our warehouse systems.
* Load and transform data into HDFS from large set of structured data /Oracle/Sql server using Talend Big data studio.
* Work on developing events-based data processing pipeline using AWS Lambda, SNS and DynamoDB streams
* Worked with data investigation, discovery, and mapping tools to scan every single data record from many sources.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS.
* Designed and Developed Scala workflows for data pull from cloud based systems and applying transformations on it.
* Worked on Machine Learning Algorithms Development for analyzing click stream data using Spark and Scala.
* Exported and Extracted the data from Teradata into HDFS using Sqoop.
* Analyzed the data by performing Hive queries and running Pig scripts to know user behavior.
* Monitoring and managing the Hadoop cluster through Cloudera Manager.
* Installed **Oozie workflow engine** to run multiple Hive
* Developed Hive queries to process the data and generate the data cubes for visualizing.

**Environment**: Spark, Hive, HBase, Sqoop, Cosmos DB, MapReduce, HDFS, Cloudera, Scala, Spark, Aurora, SQL, Apache Kafka, AWS, Lambda, Redshift, S3, Kubernetes, Python, Pandas, NumPy, Boto3, Unix

**Change Healthcare, Bridgeton MO Jun 2017 to November 2019**

**Data Engineer**

**Responsibilities:**

* Installed, configured and maintained Data Pipelines, Developed Data Pipeline with Kafka and Spark.
* Developed processes for loading the data into snowflakes. Designed data modeling on the data and joined them with other DIM tables DataStage for tableau reporting.
* Implementation of Azure cloud solution using HDInsight, Event Hubs, CosmosDB, cognitive services and KeyVault.
* Ingest and Prep business ready data by building ELT/ETL data pipelines using Azure Data Factory, Azure Data bricks (Spark, Scala, Python) into Azure SQL Data Warehouse.
* Authored Python (PySpark) Scripts for custom UDF's for Row/ Column manipulations, merges, aggregations, stacking, data labeling and for all Cleaning and conforming tasks.
* Evaluated Snowflake Design considerations for any change in the application, and built the Logical and Physical data model for snowflake as per the changes required.
* Redesigned the Views in snowflake to increase the performance and Unit tested the data between and Snowflake.
* Developed data warehouse model in snowflake for over 100 datasets using whereScape and Created Reports in Looker based on Snowflake Connections
* Developed solutions to leverage ETL tools and identify opportunities for process improvements using Scheduling tool and Python.
* Developed and modified triggers, packages, functions and stored procedures for data conversions and PL/SQL procedures to create database objects dynamically based on user inputs.
* Involved in continuous enhancements and fixing of production issues. Generated server-side PL/SQL scripts for data manipulation and validations.
* Suggested latest PostgreSQL updates and evaluated database issues.
* Worked on scheduling all jobs using Airflow scripts using python added different tasks to DAG, LAMBDA, DataStage.
* Designed and implemented 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.
* Designed and Developed data flows (streaming sources) using Azure Databricks features available.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation and aggregation from multiple file formats.
* Used AWS services like EC2 and S3 for small data sets processing and storage. Experienced in Maintaining the Hadoop cluster on AWS EMR.
* Worked on Dimensional and Relational Data Modeling using Star and Snowflake Schemas, OLTP/OLAP.
* Developed Automation Regressing Scripts for validation of ETL process between multiple databases like AWS, SQL Server using Python.

**Environment**: AWS, EC2, ETL, Pyspark, Snowflake, Kafka, Spark, DataStage, Lambda, Hadoop, Tableau, Python, Hive, SQL, Oracle, scheduling tool, Shell scripting.

**IBing Software Solutions Private Limited Hyd India October 2015 to April 2017**

**Big Data Engineer**

**Responsibilities:**

* Analyzed and gathered business requirements specifications by interacting with client and understanding business requirement specification documents.
* Worked in an Agile environment for the project, tools used were JIRA, GIT. Following up with the activities on JIRA regarding issues, reports, related documents, coordinating with the development, business, and testing teams from various locations.
* Creating Scrum projects, discussing about sprints to be added, standup meetings, burndown charts, summary reports.
* Used GIT to maintain repository, creating and merging branches, commit changes, checking out, moving, and removing files.
* Created data models, stored procedures, queries for data analysis and manipulations, views, functions. Maintain, upgrade databases and creating backups in SQL.
* Analyzed the client’s snapshot pages in the web interface in HTML and CSS to spot inconsistencies.
* Developed automated python scripts for repetitive task like delimiters splitting, characters joining, stray values filtering, date and data format conversions, regex operations (like code matching, replacing, pattern matching).
* The data received after all the tests were done would be parsed to see that there are no inconsistencies and save the data to the database.
* Involved in importing and exporting data from local and external file system and RDBMS to HDFS
* Managed datasets using Panda data frames and MySQL, queried MYSQL database queries from python using Python-MySQL connector and MySQL dB package to retrieve information. Designed and developed a data management system using MySQL.
* Managed large data sets using Pandas data frames and MYSQL. Responsible for debugging and troubleshooting the web application. Automated most of the daily task using python scripting.

**Environment**: Python, Flask, Azure, SDLC, GIT, Agile, MySQL, RDBMS, SOAP, Shell Script, HTML, CSS, JIRA.

**Yana Software Private Limited Hyderabad, India Nov 2013 to Sept 2015**

**Data Engineer**

**Responsibilities:**

* Used MS Excel, MS Access, and SQL to write and run various queries.
* Worked extensively on creating tables, views, and SQL queries in MySQL.
* Worked with internal architects and assisted in the development of current and target state data architectures.
* Perform troubleshooting, fixed, and deployed many Python bug fixes of the two main applications that were the main source of data for both customers and the internal customer service team.
* Write Python scripts to parse JSON documents and load the data in the database.
* Generating various capacity planning reports (graphical) using Python packages like NumPy, and MatPlotLib.
* Analyzing various logs that are been generating and predicting/forecasting the next occurrence of an event with various Python libraries.
* Performed Exploratory Data Analysis, trying to find trends and clusters.