**ABHISHEK**

**Email:** [**Abhishekit461@gmail.com**](mailto:Abhishekit461@gmail.com) **PH: 816-237-0630**

**Sr. Data Engineer**

**Professional Summary**

* 8 years of overall IT experience in a variety of industries, this includes hands-on experience in Big Data Analytics and development.
* Experience in collecting, processing, and aggregating large amounts of streaming data using Kafka, Spark Streaming.
* Good Knowledge on Apache NiFi for automating and managing the data flow between systems.
* Experience in designing Data Marts by following Star Schema and Snowflake Schema Methodology.
* Highly skilled in Business Intelligence tools like Tableau, PowerBI, Plotly and Dataiku.
* Experience in managing and analyzing massive datasets on multiple Hadoop frameworks like Cloudera and Hortonworks.
* Experience in Spark-Scala programming with good knowledge on Spark Architecture and its In-memory Processing
* Experience in designing and developing applications in Spark using Python to compare the performance of Spark with Hive.
* Hands-on Experience in Service Oriented Architecture (SOA), Event Driven Architecture, Distributed Application Architecture and Software as Service (SAS).
* Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, Cloud Front, CloudWatch, SNS, SES, SQS and other services of the AWS family.
* Good work experience with the cutting-edge technologies like Kafka, Spark, Spark streaming.
* Partnered with cross functional teams across the organization to gather requirements, architect, and develop proof of concept for the enterprise Data Lake environments like MAPR, CLOUDERA, HORTONWORKS, AWS, and AZURE.
* Strong Experience in analyzing data using HIVE, Impala, Pig Latin, and Drill.
* Experience in writing custom UDFs in Hive and Pig to extend the functionality.
* Experience in writing MAPREDUCE programs in java for data cleansing and preprocessing.
* Excellent understanding/knowledge on Hadoop architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Resource manager,Node manager.
* Hands on experience on Data Analytics Services such as Athena, Glue Data Catalog & Quick Sight
* Good working experience with Hive and HBase/MapRDB Integration.
* Excellent understanding and knowledge of NOSQL databases like HBase, and Cassandra.
* Experienced in converting Hive/SQL queries into Spark transformations using Spark Data Frames and Python.
* Experience setting up instances behind Elastic Load Balancer in AWS for high availability and cloud integration with AWS using ELASTIC MapReduce (EMR).
* Experience in working in Hadoop eco-system integrated to the Cloud platform provided by AWS with several services like Amazon EC2 instances, S3 bucket and RedShift.
* Good experience working with Azure Cloud Platform services like Azure Data Factory (ADF), Azure Data Lake, Azure Blob Storage, Azure SQL Analytics,HDInsight/Databricks.
* Expose to various software development methodologies like Agile and Waterfall.
* Extensive experience working with spark distributed Framework involving Resilient Distributed Datasets (RDD) and Data Frames using Python, Scala and Java8.
* Involving in developing applications on Windows, UNIX, and Linux Platforms.

**Professional Experience**

**Sr. Data Engineer**

**Chewy Dania Beach, FL April 2021 to Present**

**Responsibilities:**

* Migrate the existing data from Teradata/SQL Server to Hadoop and perform ETL operations on it.
* Responsible for loading structured, unstructured, and semi-structured data into Hadoop by creating static and dynamic partitions.
* Worked on different data formats such as JSON and performed machine learning algorithms in Python.
* Performing statistical data analysis and data visualization using Python and R
* Implemented data ingestion and handling clusters in real time processing using Kafka.
* Imported real time weblogs using Kafka as a messaging system and ingested the data to Spark Streaming.
* Created a task scheduling application to run in an EC2 environment on multiple servers.
* Strong knowledge of various Data warehousing methodologies and Data modeling concepts.
* Developed Hadoop streaming Map/Reduce works using Python.
* Created Hive partitioned tables using Parquet & Avro format to improve query performance and efficient space utilization.
* Responsibilities include Database Design and Creation of User Database.
* Moving ETL pipelines from SQL server to Hadoop Environment.
* Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Implemented a CI/CD pipeline using Jenkins, Airflow for Containers from Docker and Kubernetes.
* Used SSIS, NIFI, Python scripts, Spark Applications for ETL Operations to create data flow pipelines and involved in transforming data from legacy tables to Hive,HBase tables, and S3 buckets for handoff to business and Data scientists to create analytics over the data.
* Support current and new services that leverage AWS cloud computing architecture including EC2, S3, and other managed service offerings.
* Implemented data quality checks using Spark Streaming and arranged bad and passable flags on the data.
* Used advanced SQL methods to code, test, debug, and document complex database queries.
* Design relational database models for small and large applications.
* Designed and developed Scala workflows for data pull from cloud-based systems and applying transformations on it.
* The ability to develop reliable, maintainable, efficient code in most of SQL, Linux shell, and Python.
* Implemented Apache-spark code to read multiple tables from the real-time records and filter the data based on the requirement.
* Stored final computation result to Cassandra tables and used Spark-SQL, spark-dataset to perform data computation.
* Used Spark for data analysis and store final computation results to HBase tables.
* Troubleshoot and resolve complex production issues while providing data analysis and data validation.

**Environment**: Teradata, SQL Server, Hadoop, ETL operations, Data Warehousing, Datamodelling, Cassandra, AWS Cloud computing architecture, EC2, S3, Advanced SQL methods, NiFi, Python, Linux, Apache Spark, Scala, Spark-SQL, HBase.

**Data Engineer**

**Empower Retirement, Greenwood Village, CO June 2019 to March 2021**

**Responsibilities:**

* Create and maintain reporting infrastructure to facilitate visual representation of manufacturing data for purposes of operations planning and execution.
* Extract, Transform and Load data from source systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and Azure Data Lake Analytics.
* Implemented Restful web service to interact with Redis Cache framework.
* Intake happens through Sqoop and Ingestion happens through Map Reduce, HBASE.
* Extensively worked on Spark Streaming and Apache Kafka to fetch live stream data.
* Responsible for performing Machine-learning techniques regression/classification to predict the outcomes.
* Constructed product-usage SDK data and data aggregations by using PySpark, Scala.
* Spark SQL and Hive context in partitioned Hive external tables maintained in AWS S3 location for reporting, data science Dashboarding, and ad-hoc analyses.
* Involved in data processing using an ETL pipeline orchestrated by AWS Data Pipeline using Hive.
* Installed Kafka manager for consumer lags and for monitoring Kafka Metrics also this has been used for adding topics, Partitions etc.
* Experience in creating configuration files to deploy the SSIS packages across all environments.
* Experience in writing queries in SQL and R to extract, transform and load (ETL) data from large datasets using Data Staging.
* Implemented CI/CD pipelines using Jenkins and built and deployed the applications.
* Worked on developing Restful endpoints to cache application specific data in in-memory data clusters like redis and exposed them with Restful endpoints.
* Creating Databricks notebooks using SQL, Python and automated notebooks using jobs.
* Interacting with other data scientists and architected custom solutions for data visualization using tools like Tableau, packages in R
* Developed predictive models using Python & R to predict customers churn and classification of customers.
* Documenting the best practices and target approach for CI/CD pipeline.
* Coordinated with QA team in preparing for compatibility testing of Guidewire solution.
* Familiar with data architecture including data ingestion pipeline design, Hadoop information architecture, data modelling and data mining, machine learning and advanced data processing.
* Designed and implemented by configuring Topics in the new Kafka cluster in all environments.

**Environment**: Hadoop, ETL operations, Data Warehousing, Data Modelling, Cassandra, AWS Cloud computing architecture, EC2, S3, Advanced SQL methods, NiFi, Python, Linux, Apache Spark, Scala, Spark-SQL, HBase

**Data Engineer**

**Broadridge, Lake Success, NY January 2017 to May 2019**

**Responsibilities:**

* Migrate the existing data from Teradata/SQL Server to Hadoop and perform ETL operations on it.
* Responsible for loading structured, unstructured, and semi-structured data into Hadoop by creating static and dynamic partitions.
* Worked on different data formats such as JSON and performed machine learning algorithms in Python.
* Performing statistical data analysis and data visualization using Python and R
* Implemented data ingestion and handling clusters in real time processing using Kafka.
* Imported real time weblogs using Kafka as a messaging system and ingested the data to Spark Streaming.
* Created a task scheduling application to run in an EC2 environment on multiple servers.
* Strong knowledge of various Data warehousing methodologies and Data modeling concepts.
* Developed Hadoop streaming Map/Reduce works using Python.
* Created Hive partitioned tables using Parquet & Avro format to improve query performance and efficient space utilization.
* Responsibilities include Database Design and Creation of User Database.
* Moving ETL pipelines from SQL server to Hadoop Environment.
* Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Implemented a CI/CD pipeline using Jenkins, Airflow for Containers from Docker and Kubernetes.
* Used SSIS, NIFI, Python scripts, Spark Applications for ETL Operations to create data flow pipelines and involved in transforming data from legacy tables to Hive,HBase tables, and S3 buckets for handoff to business and Data scientists to create analytics over the data.
* Support current and new services that leverage AWS cloud computing architecture including EC2, S3, and other managed service offerings.
* Implemented data quality checks using Spark Streaming and arranged bad and passable flags on the data.
* Used advanced SQL methods to code, test, debug, and document complex database queries.
* Design relational database models for small and large applications.
* Designed and developed Scala workflows for data pull from cloud-based systems and applying transformations on it.
* The ability to develop reliable, maintainable, efficient code in most of SQL,Linux shell, and Python.
* Implemented Apache-spark code to read multiple tables from the real-time records and filter the data based on the requirement.
* Stored final computation result to Cassandra tables and used Spark-SQL, spark-dataset to perform data computation.
* Used Spark for data analysis and store final computation results to HBase tables.
* Troubleshoot and resolve complex production issues while providing data analysis and data validation.

**Environment**: Teradata, SQL Server, Hadoop, ETL operations, Data Warehousing, Data Modelling, Cassandra, AWS Cloud computing architecture, EC2, S3, Advanced SQL methods, NiFi, Python, Linux, Apache Spark, Scala, Spark-SQL, HBase

**Data Engineer**

**Couth Infotech Pvt. Ltd, Hyderabad, India November 2014 to August 2016**

**Responsibilities:**

* Experience in building distributed high-performance systems using Spark and Scala.
* Experience developing Scala applications for loading/streaming data into NoSQL databases (MongoDB) and HDFS.
* Perform T-SQL tuning and optimizing queries for and SSIS packages.
* Designed Distributed algorithms for identifying trends in data and processing them effectively.
* Creating an SSIS package to import data from SQL tables to different sheets in Excel.
* Used Spark and Scala for developing machine learning algorithms that analyze clickstream data.
* Used Spark SQL for data pre-processing, cleaning, and joining very large data sets.
* Performed data validation with Redshift and constructed pipelines designed over 100TB per day.
* Co-developed the SQL server database system to maximize performance benefits for clients.
* Assisted senior-level data scientists in the design of ETL processes, including SSIS packages.
* Database migrations from traditional data warehouses to spark clusters.
* Ensure the data warehouse was populated only with quality entries by performing regular cleaning and integrity checks.
* Used Oracle relational tables and used them in process design.
* Developed SQL queries to perform data extraction from existing sources to check format accuracy.
* Developed automated tools and dashboards to capture and display dynamic data.
* Installed a Linux operated Cisco server and performed regular updates and backup and used MS excel functions for data validation.
* Coordinated data security issues and instructed other departments about secure data transmission and encryption.

**Environment**: T-SQL, MongoDB, HDFS, Scala, Spark SQL, Relational Databases, Redshift, SSIS, SQL, Linux, Data Validation, MS Excel.

**Data Analyst and Quality Analyst**

**Cybage Software Private Limited Hyd India June 2013 to October 2014**

**Responsibilities:**

* Collaborated with business leaders, IT groups and vendors to ensure issues communicated clearly and accurately.
* Perform data analysis to evaluate the data quality and resolve the data related issues
* Responsible for improving data quality and for designing or presenting conclusions gained from analyzing data using Microsoft Excel as statistical tool.
* Performed Data reconciliation of source, Referential Integrity, Null and Default value checks and Business Rules applied.
* Debugged data issues between source systems and EDW.
* Profiled data in the sources prior to defining the data cleaning rules.
* Verified the data loads as well as the reports.
* Involved in identifying reporting needs and helping the business in creating specification, and worked closely with report development team to achieve the reporting goal.
* Assisted the project manager in setting up the direction for business analysis team.
* Developed custom routines for generating test cases.
* Created the Test Specifications, Test Scripts and Test Categories for the testing of data in the EDW.
* Attended defect triage meetings with the end users and developers.