Dileep Kumar

 **Email:** marneni.dk@gmail.com

 **Tel: +1(**636)-556-1014

***Over 9 years of diversified IT experience in analysing, designing, developing, and implementing big data solutions and experience with building and maintaining data lakes to increase data processing efficiency, implementing data security and compliance measures and leading cross-functional teams in the design and implementation of real-time data pipelines.***

**Professional Summary**

* 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 and granting database access and Migrating On premise databases to Azure Data Lake store using Azure Data factory.
* Experience in Developing Spark applications using Spark - SQL In Databricks for data extraction, transformation, and aggregation from multiple file formats for analysing & transforming the data to uncover insights into the customer usage patterns.
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, Delta Lake with Python, Delta Lake with Spark SQL.
* Good understanding of Spark Architecture with Databricks, Structured Streaming. Setting up AWS and Microsoft Azure with Databricks, Databricks workspace for Business Analytics, manage clusters in Databricks, managing the machine learning lifecycle.
* Expertise in using major components of the Hadoop ecosystem, including HDFS, YARN, MapReduce, Hive, Impala, Pig, Sqoop, HBase, Spark, Spark SQL, Kafka, Spark Streaming, Flume, Oozie, and Zookeeper.
* Expertise in building Spark applications using Python and have a strong ability to troubleshoot and fine-tune long-running Spark applications. Proficient in using Spark RDD API, Spark Data Frame/Dataset API, Spark-SQL, and Spark ML frameworks to build end-to-end data pipelines. Furthermore, I have experience working with real-time streaming pipelines using Kafka and Spark Streaming.
* Working with NoSQL databases like HBase, Azure, MongoDB, and Cassandra.
* Experienced in designing and implementing enterprise data warehouse, business intelligence, analytical, batch/real-time/near-real-time streaming big data solutions.
* Worked with ETL methods for data extraction, transformation and loading in corporate-wide ETL Solutions and Data Warehouse tools for reporting and data analysis.
* Skilled in data modelling, data mining, and improving data reliability, efficiency, and quality.
* Experience in running queries using Impala and using BI tools to run ad-hoc queries directly on Hadoop.
* I am well-versed in Agile and Waterfall methodologies and have experience defining user stories, driving the agile board in JIRA, and taking part in sprint demos and retrospectives. I have also conducted proof-of-concept projects on newly adopted technologies such as Apache Airflow, Snowflake, and GitLab.
* Proficient in Databricks MLflow for running machining learning models on distributed platforms.
* Furthermore, I have a strong understanding of NoSQL databases like HBase, Cassandra, and MongoDB, and I have proficient knowledge and hands-on experience in writing shell scripts in Linux. I am experienced in the full software development lifecycle (SDLC+) and have collaborated with Python technologies for requirement analysis, application development, migration, and maintenance.
* I have ability in configuring and installing PostgreSQL, Postgres Plus Advanced Server, and have detailed exposure to Azure tools such as Azure Data Lake, Azure Data Bricks, Azure Data Factory, HDInsight, Azure SQL Server, and Azure DevOps. I am knowledgeable in setting up and keeping Postgres master-slave clusters utilizing streaming replication. I also have exposure to Cloudera installation on Azure Cloud instances.

**Information Technology Skill set**

|  |  |
| --- | --- |
| **Big Data Technologies** | Kafka, Apache Spark, Spark Streaming, HBase, Impala, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Flink, Delta Lake, Apache Beam, Apache NiFi Registry, Apache Airflow, Google Cloud Dataflow, Google Big Query, Snowflake, Azure Data Lake, Azure Stream Analytics |
| **Hadoop Distribution**  | Cloudera CDP, Hortonworks HDF, Apache Hadoop, Amazon EMR, Microsoft HDInsight, Google Cloud Dataproc |
| **Programming Languages**  | Oracle SQL, PL/SQL, Python, Scala, Shell Scripting, Java |
| **Data Warehousing** | Informatica PowerCenter, ETL Development, Amazon Redshift |
| **Cloud Infrastructure** | AWS, Azure, GCP, IBM Cloud |
| **Databases**  | PL/SQL- RDBMS, MySQL, SQL Server, Teradata, NoSQL Databases (MongoDB, Cassandra, DynamoDB) |
| **Data Modelling** | Erwin |
| **Version Control** | CVS, SVN, Clear Case, GIT, Bitbucket, GitHub, GitLab |
| **Data Governance** | Informatica Axon, Alation |
| **Containerization Tools**  | Kubernetes, Docker, Docker Swarm, Apache Mesos, Apache Aurora |
| **Reporting Tools** | JUnit, Eclipse, Visual Studio, NetBeans, Azure Databricks, UNIX, Power BI, SAS, Tableau, Mode Analytics, Quick Sight |

**IT Work Experience**

**GDIT, Rensselaer-NY**  (June 2021– Till Date)

**Role: Sr. Data Engineer**

**Responsibilities:**

* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and IJ-SQL Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from various sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Experience in leveraging Spark to handle large-scale data processing, including batch and real-time/streaming data. Hands-on experience on developing SQL Scripts for automation purposes.
* Ability to perform complex data transformations, data cleaning, and exploratory data analysis using Spark.
* Experience integrating Spark with other tools and frameworks such as Hadoop, Hive, Kafka, or cloud services (AWS, Azure, Google Cloud Platform) to build end-to-end data pipelines.
* Designed and developed Flink pipelines to consume streaming data from Kafka and applied business logic to massage and transform raw data.
* Automated resulting scripts and workflow using Apache Airflow and shell scripting to ensure daily execution in production.
* Use SQL, Spark SQL, and NoSQL databases like MongoDB and Cassandra for data analysis and querying.
* proficiency in handling large volumes of structured and unstructured data, implementing data processing techniques efficiently using Hadoop ecosystem tools.
* Building data warehousing solutions on Hadoop, integrating ETL processes and ensuring data quality and consistency.
* Develop data pipelines using Azure services like Azure Data Factory, Azure Data bricks, or Azure Synapse Analytics.
* Setting up Azure active directory permissions for team and managing all infrastructure support services.
* Orchestrate the flow of data between various stages of the pipeline, including extraction, transformation, and loading (ETL).
* Collaborate with cross-functional teams, including data scientists, analysts, and business users, to understand their data needs.
* Used Hibernate ORM tools which automate the mapping between SQL databases and objects in Java.
* Maintaining quality reference data: Ensuring the integrity and accuracy of reference data in the source systems by performing operations such as data cleaning, transformation, and validation.
* Importing and exporting databases: Using SQL Server Integration Services (SSIS) and Data Transformation Services (DTS Packages) to efficiently import data from external sources into databases and export data to external systems.
* Identifying and resolving performance bottlenecks, using techniques such as indexing, query optimization, and caching.
* Developing reusable frameworks and automation solutions for future data migrations. Creating frameworks that use technologies like Spark Data Sources and Hive data objects to streamline the Extract, Transform, Load (ETL) processes from relational database management systems (RDBMS) to Data Lakes.
* Data blending and preparation: Conducting data blending and preparation activities using tools like Alteryx and SQL to prepare data for consumption in Tableau.
* Design end-to-end data pipeline architectures that align with business requirements and data processing needs.
* Load transformed data into target data storage solutions, including Azure SQL Database, Azure Data Lake Storage, and Azure Cosmos DB.
* Implement CI/CD practices for data pipelines to ensure smooth deployment, testing, and updates.
* Manage and automate deployment pipelines for Azure Data Bricks workspaces.
* Publishing data sources to the Tableau server and Power BI for analysis and reporting purposes.
* Integrating Apache Airflow with cloud platforms like AWS to check and manage multi-stage machine learning (ML) workflows.
* collaborated with cross-functional teams to successfully implement Spark-based solutions that deliver business value.

**Truist, Charlotte-NC**  (Feb2019– May 2021)

**Role: Data Engineer**

**Responsibilities:**

* Designed and Configured Azure Cloud relational servers and databases analysing current and future requirements.
* Experience in creating pipeline jobs, scheduling triggers, mapping data flows using Azure Data Factory (V2).
* Designing and implementing data models and schemas using NoSQL databases such as MongoDB, Cassandra, and DynamoDB.
* Creating and managing data pipelines using tools such as Apache Beam, Apache Flink, and Apache Spark Streaming for real-time data processing.
* Developed business intelligence solutions using SQL server data tools and loaded data to SQL & Azure cloud databases.
* Developing and implementing data visualization solutions using tools such as Tableau, Power BI, and Apache Superset for reporting and dashboarding.
* Utilizing containerization technologies such as Docker and Kubernetes to deploy and manage big data applications in a scalable and efficient manner.
* Developing and implementing data governance policies and best practices to ensure compliance with data privacy regulations such as GDPR and CCPA
* Prepared ETL design document which consists of the database structure, change data capture, Error handling, restart, and refresh strategies.
* Worked in Production Environment which involves building CI/CD pipeline using Jenkins with various stages starting from code checkout from GitHub to Deploying code in specific environment.
* Developed AWS cloud formation templates and setting up Auto scaling for EC2 instances and involved in the automated provisioning of AWS cloud environment using Jenkins.
* Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries.
* Developed data pipeline using Spark, Hive, python to ingest customer.
* Worked on migrating Map Reduce programs into Spark transformations using Python.
* Worked on Spark Data sources (Hive, JSON files, Spark Data frames, Spark SQL and Streaming using Python.
* Developed Spark scripts by writing custom RDDs in Python for data transformations and perform actions on RDDs.
* Implemented Kafka, spark structured streaming for real time data ingestion.
* Analysed substantial amounts of data sets to figure out best way to aggregate and report on it using Map Reduce programs.
* Involved in designing optimizing Spark SQL queries, Data frames, import data from Data sources, perform transformations and stored the results to output directory into AWS S3.
* Configured Spark streaming to receive real time data from Kafka and store the stream data to HDFS.
* Created spark jobs to apply data cleansing/data validation rules on new source files in inbound bucket and reject records to reject-data S3 bucket.
* Involved in converting Hive/SQL queries into transformations using Python and performed complex joins on tables in hive with various optimization techniques.
* Created Hive tables as per requirements, internal or external tables defined with proper static and dynamic partitions, intended for efficiency.
* ETL (Extract, Transform, Load) development: Supporting the development of ETL processes to extract data from source systems, transform it to meet business requirements, and load it into the target data storage or processing systems. This may involve working with tools like Apache Airflow, Talend, or SSIS.
* Assisting in implementing data quality checks and validations to ensure the accuracy, completeness, and consistency of data. This may involve developing data quality rules or using data profiling tools.
* Assisting in designing and implementing data models or database schemas based on business requirements. This may involve working with relational databases like MySQL, PostgreSQL, or NoSQL databases like MongoDB or Cassandra.

**Merk Pharma, Branchburg-NJ**  (Dec2017 – Nov 2019)

**Role: Data Engineer**

**Responsibilities:**

* Proficient in parsing and generating JSON data using programming languages such as Python, Java, or C# to facilitate seamless data communication and integration.
* Strong programming acumen in Python, SQL, and Java for intricate data manipulation tasks and scripting. Utilized these skills to develop efficient data processing solutions and automate repetitive tasks.
* Designed, developed, and supported end-to-end data pipelines using Python and AWS services for efficient data extraction, transformation, and loading (ETL) processes.
* Implemented data engineering solutions on the Azure cloud platform, using services such as Azure Data Factory and Azure Databricks, ensuring high availability and scalability.
* Created automated and scheduled reports using Power BI's subscription and distribution features, enabling stakeholders to receive up-to-date insights without manual intervention.
* Demonstrated ability to connect, clean, and transform data from various sources into usable formats within Power BI, ensuring data accuracy and consistency for meaningful visualizations.
* Demonstrated ability in Python programming for data processing, transformation, and ETL pipelines.
* Implemented dynamic parameterization within Airflow workflows, allowing for dynamic scheduling and execution of tasks based on changing data sources and conditions.
* Developed and deployed AWS Lambda functions to process real-time data streams, enabling efficient data transformation and manipulation.
* Ability in working with JSON data formats for structured data representation and exchange between systems.
* Ability in using Matillion ETL for designing and orchestrating complex ETL workflows on cloud platforms.
* Skilled in designing and generating interactive and visually appealing reports using SSRS, supplying valuable insights to stakeholders and enabling data-driven decision-making.
* Experienced in handling complex data transformations, error handling, and data cleansing within SSIS packages, ensuring data quality and integrity.
* Demonstrated mastery in leveraging Prophecy to develop robust data engineering pipelines, enabling data extraction, transformation, and loading (ETL) with an emphasis on automation and scalability.
* Designed and implemented Python scripts to process large volumes of raw data from various sources, ensuring data quality, integrity, and consistency.
* Developed robust ETL pipelines using industry-leading tools such as Apache Spark, Apache Airflow, and custom Python scripts.
* Implemented end-to-end data governance solutions using AXON, EDC (Enterprise Data Catalog), and IDQ (Informatica Data Quality).
* Implemented data profiling and validation rules using Informatica IDQ to find and rectify anomalies, inconsistencies, and errors within datasets.
* Leveraged Azure Databricks to design, develop, and implement robust ETL pipelines for processing large-scale and complex datasets. Utilized Databricks' powerful cluster computing capabilities to efficiently transform and aggregate data, ensuring best performance and scalability.
* Employed Python scripts to clean and preprocess data by managing missing values, outliers, and inconsistencies, enhancing the overall quality of the datasets.
* applications on AWS EKS, using Docker and Kubernetes.
* Designed and implemented data transformation logic using DataStage's transformation stages, handling data cleansing, validation, aggregation, and enrichment.
* Environment: Spark, AWS S3, Python, Google Cloud Platform (GCP), Sqoop, Hive, Kafka, Tableau, Hadoop, HDFS, Agile, Unix UHC, AWS Redshift, AWS Quick sight Snowflake.

**Esko, Hyderabad-India**  (Aug 2016 – Sep 2017)

**Role: Hadoop Developer**

**Responsibilities:**

* Installed and configured Hive, HDFS and the NIFI, implemented HDP Hadoop cluster. Helped with performance tuning and monitoring.
* Involved in loading and transforming large sets of structured data from router location to EDW using a NIFI data pipeline flow.
* Developed PySpark code and Spark-SQL for faster testing and processing of data.
* Worked on Data serialization formats for converting complex objects into sequence bits by using Parquet, ORC, AVRO, JSON, and CSV formats.
* Created Hive tables to load large data sets of structured data coming from WADL after transformation of raw data.
* Created reports for the BI team using SQOOP to export data into HDFS and Hive.
* Managed and reviewed Hadoop log files.
* Evaluated suitability of Hadoop and its ecosystem to project and implemented various proof of concept applications to eventually adopt them to receive help from the Hadoop initiative.
* Data documentation and metadata management: Documenting data warehouse structures, data lineage, transformations, and business rules. Maintaining metadata repositories to track data sources, data definitions, and mappings.
* Monitoring and troubleshooting: Monitoring data warehouse processes, including ETL jobs, data loads, and query performance. Finding and resolving issues, such as data load failures or data discrepancies, on time.
* Collaboration with stakeholders: Collaborating with business analysts, data architects, and other stakeholders to understand data requirements, define data warehouse solutions, and ensure alignment with business goals.

**Sanovi, Hyderabad-India**  (June 2014 – July 2016)

**Role: Data Warehouse Developer**

**Responsibilities:**

* Front-end development: Developing user interfaces and user experiences (UI/UX) using HTML, CSS, and JavaScript within the Hadoop framework. Implementing responsive design principles to ensure best usability across different devices.
* Back-end development: Building and supporting server-side components and functionalities using programming languages like Java, C#, or Python. Developing APIs and integrating them with the Hadoop framework to enable data exchange and application logic.
* Database management: Designing and managing databases, including schema design, data modelling, and optimization. Utilizing relational databases like MySQL or PostgreSQL, or NoSQL databases like MongoDB, depending on the project requirements.
* Application deployment and maintenance: Deploying Hadoop applications to production environments, managing version control, and monitoring application performance. Conducting regular maintenance activities, bug fixing, and troubleshooting issues.
* Integration with third-party systems: Integrating Hadoop applications with external systems, such as payment gateways, CRM systems, or marketing automation tools. Implementing secure data transfer and supporting data integrity during integration.
* Security and data protection: Implementing security measures to protect Hadoop applications and data. Ensuring compliance with security standards and regulations, such as GDPR or PCI-DSS, and conducting regular security audits.
* Performance optimization: Perfecting Hadoop applications for performance, scalability, and efficiency. Conducting code reviews, finding, and resolving performance bottlenecks, and implementing caching mechanisms or code optimizations.
* Collaboration and teamwork: Collaborating with team members, including project managers, designers, and quality assurance professionals, to deliver high-quality Hadoop solutions. Participating in agile development processes, attending meetings, and supplying technical guidance to junior developers.

**Environment**: Autosys, Oracle 9i, PL/SQL, SQL, Unix Shell Scripts, Sun Solaris, Windows 2000, IBM WebSphere Data Stage EE/7.0/6.0 (Manager, Designer, Director, Administrator), Ascential Profile Stage 6.0, Ascential Quality Stage 6.0, Erwin, TOAD.

**Academic Credentials**

|  |  |
| --- | --- |
| **Bachelor of Commerce in Computers** | Osmania University, 2012 |