**DISHA**

**Data Engineer**

**Phone Number : 469-251-5116**

**Email id:** [**ndisha2022@gmail.com**](mailto:ndisha2022@gmail.com)

|  |
| --- |
| **SUMMARY:** |

Overall, 7+ years of IT experience, result oriented **big data consultant** possessing a proven track record of effectively administering Hadoop ecosystem components & architecture and managing file distribution systems in the Big Data arena. Proficient in collaborating with key stakeholders to conceptualize & execute solutions for resolving systems architecture-based technical issues. Highly skilled in processing complex data designing Machine Learning modules for effective data mining &modelling. Adept at Hadoop cluster management & capacity planning for end-to-end data management & performance optimization.

|  |
| --- |
| **PROFESSIONAL SUMMARY:** |

* Data Engineer with experience in implementing various Big Data/ Cloud Engineering, Snowflake, Data Warehouse, Data Modelling, Data Mart, Data Visualization, Reporting, Data Quality, Data virtualization and Data Science Solutions.
* Experience in Data transformation, Data mapping from source to target database schema, Data Cleansing procedures
* Deep knowledge and strong deployment experience in Hadoop and Big Data ecosystems- HDFS, MapReduce, Spark, Pig, Sqoop, Hive, Oozie, Kafka, zookeeper, and HBase.
* Expertise in working with Hive data warehouse infrastructure-creating tables, data distribution by implementing Partitioning and Bucketing, developing and tuning the HQL queries.
* Strong experience in tuning Spark applications and Hive scripts to achieve optimal performance.
* Developed Spark Applications using Spark RDD, Spark-SQL and Data frame APIs.
* Strong experience building end-to-end data pipelines on the Hadoop platform.
* Developed Simple to complex MapReduce streaming jobs using Python language that are implemented using Hive and Pig.
* Capable of processing large sets of structured, semi-structured, and unstructured data and supporting systems application architecture.
* Expertise in OLTP/OLAP System Study, Analysis and E-R modelling, developing Database Schemas like Star schema and Snowflake schema used in relational, dimensional and multidimensional modelling
* Experience in creating separate virtual data warehouses with difference size classes in AWS Snowflake
* Hands-on experience in bulk loading &amp; unloading data into Snowflake tables using COPY command
* Experience with data transformations utilizing SnowSQL in Snowflake
* Developed Spark Applications that can handle data from various RDBMS (MySQL, Oracle Database) and Streaming sources.
* Solid understanding of AWS, Redshift, S3, EC2 and Apache Spark, Scala process, and concepts, configuring the servers for auto scaling and elastic load balancing
* Hands on experience in machine learning, big data, data visualization, R and Python development,
* Linux, SQL, GIT/GitHub
* Experienced in python data manipulation for loading and extraction as well as with python libraries such as NumPy, SciPy and Pandas for data analysis and numerical computations
* Extensive working experience with Python including Scikit-learn, SciPy, Pandas, and NumPy developing machine learning models, manipulating and handling data
* Extensive experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, Python
* Experience in extracting, transforming and loading (ETL) data from spreadsheets, database tables and other sources using Microsoft SSIS and Informatica.
* Developed mapping spreadsheets for (ETL) team with source to target data mapping with physical naming standards, data types, volumetric, domain definitions, and corporate meta-data definitions.
* Performed data visualization and Designed dashboards with Tableau, and generated complex reports, including charts, summaries, and graphs to interpret findings to team and stakeholders
* Developed Snow pipes for continuous injection of data using event handler from AWS (S3 bucket)
* Design and developed end-to-end ETL process from various source systems to Staging area, from staging to Data Marts and data load
* Strong Understanding of dimensional and relational modelling techniques. Well versed normalization/ Denormalization techniques for optimum performance in relational and dimensional DB.

|  |
| --- |
| **TECHNICAL SKILLS:** |

|  |  |
| --- | --- |
| **Operating Systems** | Linux (Ubuntu, CentOS), Windows, Mac OS |
| **Hadoop Ecosystem** | Hadoop, MapReduce, Yarn, HDFS, Pig, Oozie, Zookeeper |
| **Big Data Ecosystem** | Spark, SparkSQL, Spark Streaming, Spark MLlib, Hive, Impala, Hue, Airflow |
| **Cloud Ecosystem** | Azure, AWS, Snowflake cloud data warehouse |
| **Data Ingestion** | Sqoop, Flume, NiFi, Kafka |
| **NOSQL Databases** | HBase, Cassandra, MongoDB, CouchDB |
| **Programming Languages** | Python, C, C++, Scala, Core Java, J2EE |
| **Scripting Languages** | UNIX, Python, R Language |
| **Databases** | Oracle 10g/11g/12c, PostgreSQL 9.3, MySQL, SQL-Server, Teradata, HANA |
| **IDE** | IntelliJ, Eclipse, Visual Studio, IDLE |
| **Tools** | SBT, Putty, Win SCP, Maven, Git, Jasper reports, Jenkins, Tableau, Mahout, UC4,  Pentaho Data Integration, Toad |
| **Methodologies** | SDLC, Agile, Scrum, Iterative Development, Waterfall Model |



|  |
| --- |
| **PROFESSIONAL EXPERIENCE:** |

**Client: Pure Storage, CA Nov 2020 – Till Date**

**Role: Sr Data Engineer**

**Responsibilities:**

* Used Spark and **Spark-SQL** to read the parquet data and create the tables in hive using the Scala API.
* Implemented Spark using Scala and utilizing **Data frames**and Spark SQL API for faster processing of data.
* Implemented **Spark** using Scala and Spark SQL for faster testing and processing of data.
* The **hive**tables are created as per requirement were Internal or External tables defined with appropriate static, dynamic partitions and bucketing, intended for efficiency.
* Good Exposure for performance **tuning** hive queries, **Spark jobs**.
* Worked extensively on **AWS** Components such as **Elastic Map Reduce (EMR)**
* Created data pipeline for different events of ingestion, aggregation and load consumer response data in **AWS** S3 bucket into Hive external tables in HDFS location to serve as feed for Tableau dashboards.
* Created monitors, alarms and notifications for **EC2** hosts using **Cloud Watch, Cloud trail** and **SNS**.
* Experience in data ingestions techniques for batch and stream processing using **AWS Batch, AWS Kinesis, AWS Data Pipeline**
* Building data pipeline ETLs for data movement to S3, then to **Redshift**.
* Written **AWS Lambda** code in **Python** for nested Json files, converting, comparing, sorting etc.
* Installed and configured **Apache airflow** for workflow management and created workflows in python for cloud infrastructure and for the on premise used UC4 schedular for automation of workflows and scheduling jobs
* Developed **python** code for different tasks, dependencies, SLA watcher and time sensor for each job for workflow management and automation using **Airflow** tool.
* Virtualized the servers using **Docker**for the test environments and dev-environments needs, also configuration automation using **Docker**containers.
* Managed **Kubernetes**charts using Helm. Created reproducible builds of the **Kubernetes**applications, managed Kubernetes manifest files and Managed releases of Helm packages.
* Created **Snow pipe** for continuous data load from staged data residing on **cloud gateway servers**.
* Used **COPY** to bulk load the data.
* Using **FLATTEN** table function to produce lateral view of VARIENT, OBJECT and ARRAY column.
* Working with both **Maximized** and **Auto-scale** functionality while running the multi-cluster warehouses.
* Using **Temporary** and **Transient** tables on different datasets.
* Sharing sample data using grant access to customer for **UAT/BAT**.
* Used **Snowflake time travel** feature to access historical data.
* Heavily involved in testing Snowflake to understand best possible way to use the cloud resources.
* Encoded and decoded json objects using **Pyspark** to create and modify the data frames in **Apache Spark**.
* Implement **Continuous Integration and Continuous Delivery process** using GitLab along with Python and Shell scripts to automate routine jobs, which includes synchronize installers, configuration modules, packages and requirements for the applications
* Developed **Restful** and **soap API’s** using Spring boot.

**Environment:** Snowflake Web UI, Snow SQL, Airflow, Hadoop MapR 5.2, Hive, Hue, Toad 12.9, Share point, Control-M, Tidal, ServiceNow, Teradata Studio, Oracle 12c, Tableau, Hadoop Yarn, Spark Core, Spark Streaming, Spark SQL, Spark MLlib

**Client: Blue Cross Blue Shield, MI July 2018 – Oct 2020**

**Role: Data Engineer**

**Responsibilities:**

* Created **Hive** tables for loading and analysing data.
* Handled importing of data from various data sources, performed transformations using Hive, MapReduce, loaded data into HDFS and extracted data from MYSQL into HDFS vice-versa using **Sqoop**.
* Used **Spark** API over Cloudera Hadoop Yarn to perform analytics on data in Hive.
* Designed and implemented an ETL framework using Scala and Python to load data from multiple sources into Hive and from **Hive** to **Vertica**
* Used **HBase** on top of HDFS as a non-relational database.
* Load the data into **Spark RDD**, perform advanced procedures like text analytics and processing using in-memory data Computation’s capabilities of Spark using Scala to generate the Output response.
* Developed **Scala** scripts using both **Data frames**/SQL and RDD/**MapReduce** in Spark for Data Aggregation, queries and writing data back into the OLTP system through **Sqoop**.
* Handled large datasets using **Partitions**, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and others during the ingestion process itself.
* Implemented Partitions, Buckets, and developed Hive query to process the data and generate the data cubes for visualizing.
* Optimizing existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD.
* Used Spark Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real-time and persists into Cassandra.
* Extracted Fingerprint image Data stored on local network to Conduct Exploratory Data analysis (EDA), Cleaning and organizing. Ran NFIQ algorithm to ensure data quality by collecting the high score images. Finally Created histograms to compare distributions of different datasets.
* Loaded the data in GPU and achieved Half Precision FP16 training on Nvidia Titan RTX and Titan V GPU for TensorFlow 1.14.
* Setup alerting and monitoring using Stack driver in GCP
* Optimized TFRecord data ingestion pipeline using tf. Data API and made them scalable by streaming over network, thus enabling training of models with Datasets which were bigger than CPU memory.
* Worked extensively on AWS Components such as Elastic Map Reduce (EMR)
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards
* Loaded data using AWS Glue
* Used Athena for data analytics.
* Worked with the data science team in automating and productional zing various models like logistic regression, k-means using Spark MLlib.
* Created various reports using Tableau based on requirements with the BI team.
* Used DataStax Spark-Cassandra connector to load data into Cassandra and used CQL to analyse data from Cassandra tables for quick searching, sorting, and grouping.
* Used Spark API over Cloudera Hadoop Yarn to perform analytics on data in Hive

**Environment:** Hadoop, HDFS, Hive, Oozie, Sqoop, Kafka, Elastic Search, Shell Scripting, HBase, Tableau, Oracle, MySQL, Teradata and AWS.

**Client: Alvi Group, Bangladesh Sep 2014 – Dec 2017**

**Role: Data Analyst**

**Description:**

**Responsibilities:**

* Experienced in Data modelling performing business area analysis and logical and physical data modelling using Erwin and data warehouse/ data mart applications as well as operational applications enhancements and new development. Data warehouse/ data marts design was implemented using Ralph Kimball methodology.
* Highly maintained the stage and production conceptual, logical, and physical data models along with related documentation for a large data warehouse project. This included confirming migration of data models from oracle designer to Erwin and updating the data models to correspond to the existing database structures.
* Excellent SQL programming skills and developed Stored Procedures, Triggers, Functions, Packages using SQL/PL SQL. Performance tuning and query optimization techniques in transactional and data warehouse environments.
* Involved with DBA group to create Best-Fit Physical Data Model from the logical Data model using Forward engineering using Erwin tool.
* Enforced referential integrity in the OLTP data model for consistent relationship between tables and efficient database design.
* Conducted design walk through sessions with the business intelligence team to ensure that reporting requirements are met for the business.
* Developed Data Mapping, Data Governance, and Transformation and cleansing rules for the Master Data Management Architecture involving OLTP, ODS.
* Served as a member of a development team to provide business data requirements analysis services, producing logical and physical data models using Erwin.
* In depth analyses of data report was prepared weekly, biweekly, monthly using MS Excel, SQL UNIX.

**Environment:** MS Excel, SQL, UNIX, Data Mapping, Data Model, OLTP