**Narendra L   
Email: narendra.bigdata27@gmail.com  
PH: +1 (575) 650-7820**

**Big Data Engineer**

**Professional Summary**

* Experience in Big Data, Hadoop ecosystem for data processing, Data Warehousing and Data Pipeline design and implementation.
* Expertise in using Cloud based managed services for data warehousing in Confidential Azure (Azure Data Lake Storage, Azure Data Factory).
* Big data processing using Hadoop technologies Map Reduce, Apache Spark, Apache Hive and Pig.
* Have experience in Dimensional Modeling using Star and Snowflake schema methodologies of Data Warehouse and Integration projects
* Worked with the Apache Airflow engine that can easily schedule and run my complex data pipelines which will make each task to get executed in a correct order
* Proficient in data governance.

|  |  |
| --- | --- |
| **Languages** | Python, R, SQL, PL/SQL, Java |
| **Data Visualization** | Tableau, Informatica, Spotfire, Cognos |
| **Hadoop Eco System** | Hadoop, MapReduce, Spark, HDFS, Sqoop, YARN, Oozie, Hive, Impala, Apache Flume, Apache Storm, Apache Airflow, HBase |
| **Cloud Platform** | AWS, Azure, Cloud Stack/Open Stack |
| **ETL/Data warehouse Tools** | Informatica 9.6/9.1, SAP Business Objects XIR3.1/XIR2, Talend, Tableau, and Pentaho. |

* Recreated existing application logic and functionality in the Azure Data Lake, Data Factory, Data Bricks, SQL Database and SQL data warehouse environment.
* Analyzed, designed and built Modern data solutions using Azure Pass service to support visualization of data. Understand current Production state of application
* 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
* Created indexes both non clustered and clustered indexes in order to maximize the query performance in T-SQL.
* Managed Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services
* Developed Spark code using Scala and Spark-SQL for faster testing and data processing.

**Environment**: Hadoop, Azure Data Lake, Spark, Scala, APIs, Pig, Python, Kafka, HDFS, Hive, SSRS, Sqoop, Scala, HBase, SSIS.

* Creating pipelines, data flows and complex data transformations and manipulations using Azure Data Factory(ADF) and PySpark with Databricks.
* Developed streaming pipelines using Apache Spark with Python.
* Develop Azure Databricks notebooks to apply the business transformations and perform data cleansing operations.
* Develop Databricks Python notebooks to Join, filter, pre-aggregate, and process the files stored in Azure data lake storage.
* Ingested huge volume and variety of data from disparate source systems into Azure DataLake Gen2 using Azure Data Factory V2.
* Created reusable pipelines in Data Factory to extract, transform and load data into Azure SQL DB and SQL Data warehouse.
* Experienced in developing audit, balance and control framework using SQL DB audit tables to control the ingestion, transformation and load process in Azure.
* Used Azure Devops to build and release different versions of code in different environments.
* Automated jobs using Scheduled, Event based, Tumbling window triggers in ADF.
* Created External tables in Azure SQL Database for data visualization and reporting purpose.
* Well-versed with Azure authentication mechanisms such as Service principal, Managed Identity, Key vaults.

**Environment:** Azure Data Factory (ADF v2), SQL server, AZURE PowerShell, DatabricksAzure Cosmos DB, Azure Stream Analytics, Azure Event Hub.

* Implemented Installation and configuration of multi-node cluster on Cloud using Amazon Web Services (AWS) on EC2.

**Environment**: Apache Spark 2.3, Hive 2.3, Informatica, HDFS, MapReduce, Scala, Apache Nifi 1.6, Yarn, HBase, PL/SQL, Netezza 7.2, Pig 0.16, Sqoop 1.2, Flume 1.8

* Design and develop efficient PySpark programs using cloud-based data platforms (EMR) to extract/transform/load data in between various data warehouse applications.
* Worked on Cloud Platform like AWS and possess good knowledge on different types of instances for optimal usage of clusters based on the requirement.
* Engineered a solution to optimize the ETL process of Alteryx to Snowflake ingestion process which takes around
* Designed Hive tables over the parquet and csv files for loading and analyzing data.
* Involved in developing DAGS using Airflow orchestration tool and monitored the weekly processes.
* Writing data to parquet tables both non-partitioned and partitioned tables by adding dynamic data to partitioned tables using Spark.
* Wrote User Defined functions (UDFs) for special functionality for Spark.