**Naresh D**

**Hadoop / Big Data Engineer**

**Professional Summary**

* 10 Years of professional IT experience in analyzing requirements, designing, building, highly distributed mission critical products and Applications.
* 5+ years of experience in Hadoop, Spark, PySpark, ScalaSpark, Python, Azure Databricks, Azure Data Factory, Azure Data Lake Store, SQL server, Blob Storage, Rest API calls, Azure Log Analytics and CI/CD.
* Hands - on experience in Azure Cloud Services (PaaS & IaaS), Talend Big Data, Storage, Web Jobs, Active Directory, Application Insights, Logic Apps, Data Factory, Azure Monitoring, Key Vault and SQL Azure, Snowflake, TAC and Power BI
* Educate customers of all size on the value proposition of managed services on Azure and participate in architectural discussions to ensure solutions are designed for successful deployment in the cloud.
* Experience in Data warehouse design using Star Schemas and Snowflake dimensional models.
* Experience in importing and exporting the data using Sqoop/Scala to HDFS from Relational Database systems/Vertica and vice-versa.
* Strong understanding of the principles of Data Warehousing using fact tables, dimension tables and star/snowflake schema modelling, Dimensional modelling, Data migration, Data cleansing, ETL Processes for data warehouses.
* Hands on experience on major components in Hadoop Echo systems like HDFS, YARN, Hadoop Map Reduce and Spark, Scala, Python along with Hive, Impala, Sqoop, Control M and NoSQL databases like HBase and Dynamo DB.
* Analyze or transform stored data by writing MapReduce, Spark, Hive jobs based on business requirements.
* Hands on experience in establishing data pipelines to load large datasets from various data sources like SQL Server, Oracle, DB2, Files, etc to HDFS, Hive using Sqoop and Spark.
* Analyze or transform stored data by writing MapReduce, Spark, Hive jobs based on business requirements.
* Configured Spark Streaming to receive real time data from Scala and store the stream data to HDFS and process it using Spark and Scala.
* Participated in entire Software Development Life Cycle (SDLC) including Requirement Analysis, Design, Development, Testing, Implementation, Production support and post implementation
* Experience on NoSQL databases such as HBase, Cassandra, MongoDB, and DynamoDB.
* Excellent Knowledge in understanding Big Data infrastructure, distributed file systems –HDFS, parallel processing – MapReduce framework and complete Hadoop ecosystem – Hive, Pig, Sqoop, Spark, Kafka, Hbase, NoSQL, Oozie and Flume.
* Configured and administered Azure like Resource Group, Storage Account, Blob Storage, Delta Lake, Cluster config, Event Hub, Cosmos DB for different zones in development, testing and production environments.
* Experience in working with Azure Code Pipeline and creating Cloud Formation JSON templates to create custom sized VPC & migrate a production infrastructure into an Azure.
* Hands-on experience with building Azure notebooks, dbutils functions using Visual Studio Code and creating
* deployment using GIT.
* Automated to build the Azure infrastructure using Terraform and Azure cloud formation.
* Hands on experience on Spark Framework with Spark core, Spark Streaming, Spark SQL for data processing by using Scala programing language.
* Experienced Hadoop developer using Scala Spark, Pyspark having end to end experience in developing applications in Hadoop ecosystem.
* Experience with agile/scrum methodologies to iterate quickly on product changes, developing user stories and working through backlog.
* Hands on experience in writing HiveQL queries to do data cleansing and processing and experienced in hive
* performance optimization using Partitioning and Bucketing and Parallel Execution concepts.
* Excellent understanding and knowledge on NOSQL databases like HBase, Cassandra and MongoDB.
* Proficient in developing Sqoop scripts for the extractions of data from various RDBMS databases into HDFS.
* Good Experience in ETL, Data Integration and Migration, extensively used ETL, ELT methodologies for supporting Data Extraction, transformations and loading in HDFS, Hive, and HBase, Redshift, S3.
* Good experience in scripting for automation, and monitoring using Linux Shell scripts.
* Strong experience working on Version control tools like SVN and GIT revision control systems such as GitHub and JIRA to track issues.
* Adequate knowledge and working experience in Agile and Waterfall methodologies including Scrum methodology.

**Technical Skill Set**

|  |  |
| --- | --- |
| **Hadoop/Big Data** | Cloudera Hadoop, Map Reduce, NOSQL (HBase), Apache Spark, Apache Flume, HDFS, Hive, Sqoop, Spark, Scala, Hive, Hue, Sqoop, Pig, Scala, Python, Kafka |
| **Programming languages** | Python, PySpark, ScalaSpark, HiveQL and Cobol |
| **Cloud Services** | **Azure:** Azure Cloud, Azure Databricks, Azure Data Lake Store, Azure Data Factory, Azure Log Analytics, Azure SQL, Blob Storage, ADLS**AWS** – IAM, VPC, EC2, S3, Lambda, RDS, EMR, Glue, Data Pipeline, Step Functions, Aurora DB, SNS, Redshift, SQS, CloudWatch |
| **Scripting** | Unix shell scripting, JavaScript |
| **Databases** | Hive, HBase, IMSDB, IBM DB2, My SQL, SQL Server, Oracle, RedShift, AWS Aurora DB and DynamoDB |
| **Operating Systems** | UNIX, Linux, RedHat Linux, Ubuntu Linux and Windows |
| **Tools and IDE** | Eclipse, Eclipse, IntelliJ, PyCharm, Maven, DB Visualizer |
| **Version Control** | Endeavor, SVN, GitHub |
| **IDE** | Eclipse, IntelliJ, PyCharm |
| **Tools & Utilities** | PyCharm, Databricks, SQL server management studio, Airflow |
| **Data -Streaming** | Batch Processing & Real-time streaming using KAFKA |

**Professional Experience**

**Client: Hilton, Dallas, TX Jan 2023 – Till Date**

**Role: Azure Data Engineer/ /Hadoop**

**Environment:** Azure Cloud, Azure Databricks, Azure Data Lake Store, Azure Data Factory, Azure Log Analytics, Azure SQL, Blob Storage, SQL server, Spark, PySpark, Python, ScalaSpark, Terraform, Kafka, Log Analytics and CI/CD.

**Roles and Responsibilities:**

* Integrated in Infrastructure Development and Operations involving Azure Cloud platforms, Firewall setup, Blob Storage, Resource Groups, Network etc.…,
* Create Notebooks using Databricks, Scala and spark and capturing the data from Delta tables in Delta lakes.
* Created Azure Data Factory and managing policies for Data Factory and Utilized Blob storage for storage and backup on Azure. Extensive knowledge in migrating applications from internal data storage to Azure.
* Experience in building streaming applications in Azure Notebooks using Kafka and Spark.
* Captured SCD2 and updated or inserted or deleted based on Business requirement using Databricks.
* Develop up the Framework for creation of new snapshots and deletion of old snapshots in Azure Blob Storage and worked on setting up the life cycle policies to back the data from delta lakes.
* Expert in building the Azure Notebooks functions by using Python, Scala and Spark.
* Built and configured a virtual data center in the Azure cloud to support Enterprise Data Warehouse hosting including Virtual Private Cloud (VPC), Public and Private Subnets, Security Groups, Route Tables.
* Integrated both framework and CloudFormation to automate Azure environment creation along with ability to deploy on Azure, using build scripts (Azure CLI) and automate solutions using terraform.
* Created GIT repositories and specified branching strategies that best fitted the needs of the client.
* Migrate Events system from Cosmos to Azure Databricks and ADLS Gen2. Design and rchestrate the Migration
* Analysis and Design for the Migration, CI/CD and Engineering items.
* Ingestion of the Data from different source by creating Data Factory pipelines and Logic Apps WorkFlows.
* Develop powershell scripts and ARM templates to automate the provisioning and deployment process
* Job tracking and Monitoring of SMDP platform and handling production issues and hot fixes.
* Troubleshoot and identify performance, connectivity and other issues for the applications hosted in Azure platform
* Experience in developing Spark applications using Spark-SQL in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Maintain Telemetry of the all the jobs and create Dashboards using Log Analytics.
* As a part of DevOps taking DRI role and support for the Entire Sales and Marketing system.
* Designing and implementing database solution in Azure SQL

**Client: Capital One, Plano, TX Sep 2021 to Dec 2022**

**Role: Azure Data Engineer/Big Data**

**Environment**: Azure Cloud, Azure Data Lake, Azure Data factory, Power Bi, Azure Databricks, Spark, Python, Azure SQL, Azure HDInsights for HDP Spark, Azure Powershell. Scope, USQL, Cosmos, Scope, SQL Server, Hadoop Distribution (HDP) 2.4, Azure, Hive, Confluent Kafka, Control-M, Hive 8.1, MySQL, Spark 2.3, Scala 2.1, HDFS, Unix, Hbase, GIT, Apache Parquet, Cloudera, Impala, Oozie, SQL, GitHub, JIRA, Confluence, Control-M and Zena.

**Roles and Responsibilities:**

* Performed Data ingestion, Batch Processing, Data Extraction, Transformation, Loading and Real Time Streaming using Hadoop Frameworks.
* Worked in Azure Cloud IaaS stage with components Delta Lakes, Azure blob storage, Notebooks, DBFS, Spark, Scala, Data Factory and CosmosDB.
* POC in Powershell to improve the quality of the MetaData which got appreciation and moved to Prod.
* Job tracking and Monitoring of SMDP platform and handling production issues and hot fixes.
* Data Transfer from Cosmos to HDinsights Spark cluster by creating Hive schema automatically using Azure Powershell.
* Interacted with multiple teams (Business Analyst, Project Management and Upstream development teams) and progressively tracking the issues and solving them.
* Assemble a scalable Data Lakes and Data Warehousing staff of the highest caliber. To comprehend and meet data demands, establish cross-functional connections with product owners, engineers, and data analyst.
* Design and build production data pipelines from ingestion to consumption within a big data architecture, using Java, Python, Scala
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark Databricks cluster
* Transform data by running a Python activity in Databricks.
* Implementing large scale systems using Azure Data factory, Azure Data Lake storage.
* Building complex systems which involve in huge data handling and collecting the metrics building Data Pipelines
* Design and implement data engineering, ingestion and curation functions on Azure cloud using Azure native or custom programming.
* Perform detail assessments of current state data platforms and create an appropriate transition path to Azure cloud.
* Performed Data profiling, preliminary data analysis and handle anomalies such as missing, duplicates, outliers, and imputed irrelevant data.
* Designed front end and backend of the application utilizing Python on Django Web Framework.
* Developed consumer-based features and applications using Python and Django in test driven Development.
* Reading, Processing and parsing the source data files through Spark/Scala and ingesting to Hive tables.
* Kafka streaming with Spark framework is implemented for ingesting and analyzing the huge volumes of BMO data which is coming as 25TPS from the source systems.
* Loading the data into Hadoop ecosystem (Hive and HDFS) by using Hive ETL tool (Spoon).
* Extensively worked on Hive tables partitions and buckets for analyzing large volumes of data.
* Version controlling by using GIT Hub, Bitbucket tools and document maintenance by using JIRA, Confluence tools.
* Expert in Performance tuning and optimization of the Hive jobs and SQL queries.
* Scheduling and monitoring the Hive jobs on daily basis by using Control-M.
* Reading, Processing and Parsing CSV source data files through Spark/Scala and ingesting to Hive tables.
* The CSV source data is read through the Scala programming language from core (creating RDDs, Data Frame, Dataset, Scala Methods, Scala Classes and Objects, Pattern Matching, Working with Lists, Collections, Etc.)
* Knowledge transition to the end users and junior developers to understand the hive queries and the business requirements.
* Perform data validation, filtering, sorting or other transformations for every data change in HBase table and load the transformed data to another data store.
* Worked on ingesting the source data into the Hadoop datalake (OLONA) from various databases by using Sqoop tool.
* Reading, Processing and parsing CSV source data files through HQL script and ingesting to Hive and Impala tables.
* Extensively worked on Hive, Hbase, Impala tables, partitions and buckets for analyzing large volumes of data.
* Scheduled the Hive queries daily by using oozie coordinator and by writing an oozie workflow.
* I also worked on database testing and QA validation to make sure that the product is bug free.
* Worked in Agile environment and participated in daily Stand-ups/Scrum Meetings.
* Proactively involved in ongoing maintenance, support and improvements in Hadoop cluster.

**Client: Kellogg’s, Battle Creek, Michigan Mar 2020 to Aug 2021**

**Role: Big Data Developer.**

**Roles and Responsibilities:**

* Involved in reviewing business requirements and analyzing data sources form Excel/Oracle SQL Server for design, development, testing, and production rollover of reporting and analysis projects within Tableau Desktop.
* Worked on AWS services like Lambda, Glue and EMR for ingesting data from different source systems like relational and non-relational to
* meet business functional requirements.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB
* Used AWS cloud watch monitors the services within the application and to analyze the real-time logs.
* Optimizing ETL flows and SQL scripts.
* Storing the data for further analysis in Amazon redshift before loading the data into end database.
* Extensively used Amazon S3 for storing the ingested data and also the queried data after the EMR process.
* Event logging through the AWS event bridge on the stored data in S3.
* Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD&#39;s, Spark YARN
* Worked on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, Spark and Map Reduce programming
* Extract Real time feed using Scala and Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Converting the existing relational database model to Hadoop ecosystem.
* Experience in developing data processing tasks using PySpark such as reading data from external sources, merging data, performing data enrichment and loading in to target data destinations.
* Used Spark and Spark-SQL to read the parquet data and create the tables in hive using the pyspark API.
* Created pipeline for processing structured and unstructured streaming data using spark streaming and stored the filtered data into S3 as parquet files.
* Worked with Linux systems and RDBMS database on a regular basis to ingest data using Sqoop.
* Managed and reviewed Hadoop and HBase log files.
* Writing and maintaining SQL scripts for the creation of database objects.
* Integrated data from various sources (homogenous, heterogenous, legacy systems) into SQL server databases.
* Worked extensively with importing metadata into Hive and migrated existing tables and applications to work on Hive and AWS cloud.
* Designed and implemented HIVE queries and functions for evaluation, filtering, loading and storing of data.
* Involved in ETL process from design, development, testing and migration to production environments.
* Have good understanding of Teradata MPP architecture such as Partitioning, Primary Indexes,
* Created HBase tables to store variable data formats of data coming from different portfolios.
* Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Creating Hive tables and working on them using HiveQL.
* Building and creating scripts for data modelling, mining for easier access to Azure Logs, App Insights to Developed data pipeline using Scala to store data into HDFS.
* Used Spark API over Hadoop YARN as execution engine for data analytics using Hive.
* Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.

**Client: Texas Instrumental, Richardson, TX Dec 2017 – Sep 2019**

**Role: Data Engineer**

**Roles and Responsibilities:**

* Involved in reviewing business requirements and analyzing data sources form Excel/Oracle SQL Server for design, development, testing, and production rollover of reporting and analysis projects within Tableau Desktop.
* Created Heat Map showing current services by color that was broken into regions allowing business user to understand.
* Experience with other BI tools like Congas, Business Objects and MicroStrategy.
* Projected and forecasted future growth in terms of number of subscribers developing Area Maps to show details at the county level.
* Converted charts into Crosstabs for further underlying data analysis in MS Excel.
* Blended data from multiple databases into one report by selecting primary keys from each database for data validation.
* Combined views and reports into interactive dashboards in Tableau Desktop that were presented to Business Users, Program Managers, and End Users.
* Proficiency in SQL Querying, Database Development and Debugging.
* Intermediate working Knowledge on SharePoint Reporting.
* SSRS REPORTS Development and Debugging.
* Converted older access reports into running SQL Reports.
* Worked on modifying and debugging Crystal Reports.
* Perform data extraction, sampling, advance data mining and statistical analysis using linear and logistic regression, time series analysis and multivariate analysis within R and Python.
* Reviewed basic SQL queries and edited inner, left, and right joins in Tableau Desktop by connecting live/dynamic and static datasets.
* Developed story telling dashboards in Tableau Desktop and published them on to Tableau Server which allowed end users to understand the data on the fly with the usage of quick filters for on demand needed information.
* Scheduled data refresh on Tableau Server for weekly and monthly increments based on business change to ensure that the views and dashboards were displaying the changed data accurately.
* Tested dashboards to ensure data was matching as per the business requirements and if there were any changes in underlying data.
* Participated in meetings, reviews, and user group discussions.

**Client: CSX, Jacksonville, FL Aug 2016 – Dec 2017**

**Role: Data Engineer**

**Roles and Responsibilities:**

* Involved in full life cycle Implementation of the project from initial phase to the production release (requirement gathering, design, development, testing and release).
* Designed and developed Tableau visualization solutions and Created Business requirement documents and plans for creating dashboards.
* Developed Key Performance Indicators (KPI) dashboards that provides Trends of specific period
* Created actions, filters, parameters, hierarchies, Level-of-Detail (LOD), calculated fields, sorting, groupings, and live connections in Tableau.
* Created views in Tableau Desktop that were published to internal team for review and further data analysis and customization using filters and actions.
* Created Heat Map showing current service subscribers by color that were broken into regions allowing business user to understand where we have most users vs. least users.
* Created drill down and drill up in worksheets.
* Projected and forecasted future growth in terms of number of Members, developed Area Maps to show details on which states had most members.
* Converted charts into Crosstabs for further underlying data analysis in MS Excel.
* Created Bullet graphs to determine profit generation by using measures and dimensions data from SQL Server and MS Excel.
* Blended data from multiple databases into one report by selecting primary key's from each database for data validation.
* Reviewed basic SQL queries and edited inner, left, and right joins in Tableau Desktop by connecting live/dynamic and static data sets.
* Worked with Front End Team for Embedding Tableau dashboards into website for internal teams.
* Involved in testing dashboards to ensure data was matching as per the business requirements and if there were any changes in underlying data.
* Performed unit testing, coordinated system integration testing and UAT to ensure reliable and quality dashboard, reports and visualizations.

**Client: Telstra, Hyderabad, India June 2014 to Mar 2015**

**Role: Developer**

**Environment:** IBM S/390, TSO/ISPF, MVS, COBOL, JCL, DB2, CICS, ESP, ESP WORKLOAD, SDSF, SSO, Super Session, ChangeMan, File-Aid, SPUFI and AOTS

**Roles and Responsibilities:**

* Analysis of COBOL programs.
* Unit testing of Programs/JCL/PROC.
* Preparing UTPs and UTRs as per the client requirement.
* Raise the Change Requests in AOTS for the implementation
* All the L2 Capabilities including Operational Support, Incident Management, Problem Management, Change Management as per the standards.
* Problem solving for abended jobs.
* Coordinating in the Release plan meetings and code deployments.
Troubleshoot application problems. Review of COBOL programs.

**Educational Details**

|  |  |  |
| --- | --- | --- |
| **Examination** | **University/Board** | **Year of Passing** |
| **Bachelor in Computer Science** | **JNTU** | **2012** |
| **Masters in Information Technology** | **Virginia International University** | **2016** |