**Mohammed Syed**

**Sr. Data Engineer**

**Email:** syed.engineer02@gmail.com **Phone: (847) 510-5880**

**Summary:**

* Having 8+ years of experience in Analysis, Design, Development, Testing, Maintenance, and User training of software application which includes around 5+ Years in Big Data, Hadoop Framework and HDFS, Hive, MapReduce, Sqoop, Oozie, Spark, Kafka, Data Warehousing (Snowflake, Hive, Teradata), Cassandra, AWS (S3, Glue, Redshift, Lambda, EMR), ETL, HDP/Cloudera environment.
* Excellent understanding of Hadoop architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce programming paradigm.
* Experienced on data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling and data mining, machine learning and advanced data processing.
* Experience in developing Spark Programs for Batch and Real-Time Processing and developing Spark Streaming applications for Real Time Processing and good experience in writing Spark applications using Python and Scala.
* Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, Amazon Elastic Load Balancing, Auto Scaling, Cloud Front, CloudWatch, SNS, SQS and other services of the AWS family.
* Extensively worked on Sqoop, Hadoop, Hive, Spark, Cassandra to build ETL and Data Processing systems having various data sources, data targets and data formats and expertise in optimizing Map Reduce algorithms using Mappers, Reducers, and combiners to deliver the best results for the large datasets.
* Experience in building ETL scripts in different languages like PLSQL, Informatica, Hive, Pig and PySpark and expertise in Creating, Debugging, Scheduling and Monitoring jobs using Airflow and Oozie.
* Strong Experience in writing Pig scripts and Hive Queries and Spark SQL queries to analyze large datasets and troubleshooting errors and Expertise in loading and transforming large sets of structured, semi-structured and unstructured data.
* Excellent working experience on designing and implementing the complete end-to end life cycle of Hadoop Infrastructure including Pig, Hive Sqoop, Oozie, Flume and zookeeper and Excellent working experience with Hadoop distributions such as Hortonworks, Cloudera, and IBM Big Insights.
* Experience in Creating, Scheduling and Debugging Spark jobs using Python and experience in Parsing both structured and unstructured data files using Data Frames in PySpark.
* Experienced in working with Teradata Utilities like Fast load, Multi load, Tpump and Fast Export Teradata Query submitting and processing tools like Bteq and Teradata SQL Assistant.
* Proficient in Hive Query language and experienced in hive performance optimization using Static-Partitioning, Dynamic-Partitioning and Bucketing and Parallel Execution concepts.
* Expertise in Normalization to 3NF/De-normalization techniques for optimum performance in relational and dimensional database environments and experienced on ER Modeling, Dimensional Modeling (Star Schema, Snowflake Schema) and Data warehousing and OLAP/OLTP tools.
* Well versed in Relational Database Design/Development with Database Mapping, PL/SQL Queries, Stored Procedures and Packages using Oracle, DB2, Teradata and MySQL Databases.
* Have extensive knowledge and working experience on Software Development Life Cycle (SDLC), Service-Oriented architecture (SOA), Rational Unified Process (RUP), Object Oriented Analysis and Design (OOAD), UML and J2EE Architecture.
* Expertise in development of various reports, dashboards using various Tableau Visualizations and Excellent understanding of best practices of Enterprise Data Warehouse and involved in Full life cycle development of Data Warehousing.
* Highly Self-motivated and goal oriented team player with strong analytical, debugging and problem solving skills, Strong in object oriented analysis and design. Diversified knowledge and ability to learn new technologies quickly.

**Technical Skills:**

* **Hadoop/Big Data:** HDFS, MapReduce, Hive, Pig, HBase, Sqoop, Pig, Impala, Oozie, Kafka, Spark, Zookeeper, Yarn, AWS, AWS S3, AWS EMR, Redshift, PySpark, Nifi, AWS Glue.
* **Data Warehouse:** Erwin, OLTP, OLAP, Hive, Snowflake and Teardata
* **Programming languages:** Java, Scala, Python, Unix & Linux shell scripts and SQL
* **Databases:** Oracle, SQL Server, Cassandra, MongoDB, Teradata and DB2.
* **Web Servers:** Web Logic, Web Sphere, Apache Tomcat
* **Web Technologies:** HTML, XML, JavaScript, AJAX, SOAP, WSDL
* **ETL Tools:** Informatica, Talend, AWS Glue and Databricks.
* **Cloud Technology:** AWS S3, Redshift, Glue, Lambda, Athena and RDS.

**Sr. Data Engineer**

**Target, Chicago IL Jan 2019 to Present**

**Responsibilities:**

* Build Hadoop Datalakes/Deltalake and developed the architecture and used in implementations within the organization and Ingest Legacy datasets into HDFS using Sqoop Scripts and populate Enterprise DataLake by importing tables from Oracle and other databases and Mainframe Sources and store them in partitioned hive tables using different file compression techniques.
* Created data pipeline of gathering, cleaning and optimizing data using Hive, Spark and Building data pipeline ETLs for data movement to S3, then to Redshift and Developed automated data pipelines from various external data sources (web pages, API etc) to internal data warehouse (SQL server, AWS), then export to reporting tools.
* Installed and Setup Hadoop CDH clusters for development and production environment and installed and configured Hive, Pig, Sqoop, Flume, Cloudera manager and Oozie on the Hadoop cluster.
* Implement AWS Data Lake leveraging S3, terraform, EC2, Lambda, and IAM in performing data processing and storage while writing complex SQL queries, analytical and aggregate functions on views in Snowflake data warehouse to develop near real time visualization using Tableau.
* Used Spark Streaming APIs to perform transformations and actions on the fly for building common learner data model which gets the data from Kafka in near real time and persist it to Cassandra.
* Responsible for writing Hive Queries for analyzing data in Hive warehouse using Hive Query Language (HQL) and Hive UDF's in Python.
* Involved in installing EMR clusters on AWS and used AWS Data Pipeline to schedule an Amazon EMR cluster to clean and process web server logs stored in Amazon S3 bucket and Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using Cloudwatch
* Perform data masking and ETL process using S3, Glue, Azure Databricks to support Snowflake Data warehousing solution in the cloud.
* Extract Real time feed using Kafka 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.
* Performed an upgrade in development environment from CDH 4.2 to CDH 4.6 and automated end to end workflow from Data preparation to presentation layer for Artist Dashboard project using Shell Scripting.
* Developed dashboard prototypes using Cloud Dashboard Tools Looker and AWS Quicksight managing all aspects of the technical development.
* Develop framework for converting existing PowerCenter mappings and to PySpark (Python and Spark) Jobs and implemented ETL frame work using Spark with Python and loaded standardize data into Hive and Hbase tables and Create Glue jobs to process the data from S3 staging area to S3 Persistence area.
* Working on creating different tables in Hive to incorporate CDC logics by writing some Pig Scripts and HiveQl scripts which will perform CDC logics
* Converted Informatica ETLs to Spark Scala ETLs and stored data in Hive External tables for end user / analyst requirements to perform ad hoc analysis and Design and create the Complete "ETL" process from end-to-end using Talend jobs and create the test cases for validating the Data in the Data Marts and in the Data Warehouse.
* Used AWS Glue crawler and ETL jobs implementation to catalog the data available on S3 and Data Cataloging using AWS Glue and AWS Athena.
* Developed PySpark and SparkSQL code to process the data in Apache Spark on Amazon EMR to perform the necessary transformations based on the STMs developed
* Design & Develop ETL workflow using Oozie for business requirements, which includes automating the extraction of data from MySQL database into HDFS using Sqoop scripts.
* Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS. Implemented a Python-based distributed random forest via Python streaming.
* Creating Spark clusters and configuring high concurrency clusters using Azure Databricks to speed up the preparation of high-quality data.
* Developed Map reduce program which were used to extract and transform the data sets and result dataset were loaded to Cassandra and vice-versa using kafka and using Kafka messaging system registered to Cassandra brokers and pulled the data to HDFS and Implemented Kafka consumers to move data from Kafka partitions into Cassandra for near real time analysis.
* Implemented continuous integration & deployment (CICD) through Jenkins and Docker for Hadoop jobs.

**Environment:** Cloudera Hadoop, MapReduce, Informcatica, Python, PySpark, HDFS, Nifi, Hive, Pig, Sqoop, Oozie, Zookeeper, LDAP, Cassandra, HBase, Erwin, Spark, Spark-SQL, Scala, AWS EMR, S3, AWS Glue, Redshift, AWS Lambda, Snowflake, Databircks, Quicksight, Kafka, SQL, Data Warehousing, Java, PL/SQL, RDBMS (Oracle, Teradata).

**Sr. Data Engineer**

**Bank of Montreal, Chicago IL Feb 2017 to Dec 2018**

**Responsibilities:**

* Involved in start to end process of Hadoop jobs that used various technologies such as Sqoop, PIG, Hive, MapReduce, Spark and Shel lscripts (for scheduling of few jobs) extracted and loaded data into DataLake environment (AmazonS3) by using Sqoop which was accessed by business users and data scientists.
* Responsible to manage data coming from various sources and involved in HDFS maintenance and loading of structured and unstructured data and visualize the HDFS data to customer using BI tool with the help of Hive ODBC Driver.
* 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 and responsible for creating on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Generation of business reports from DataLake using Hadoop SQL (Impala) as per the Business Needs and automation of Business reports using Bash scripts in UNIX on Datalake by sending them to business owners.
* Analyzed the SQL scripts and designed it by using PySpark SQL for faster performance and Encoded and decoded JSON objects using PySpark to create and modify the data frames in Apache Spark
* Utilized Apache Spark with Python to develop and execute Big Data Analytics and Machine learning applications, executed machine learning use cases under Spark ML and Mllib.
* Used Erwin Data Modeler and Erwin Model Manager to create Conceptual, Logical and Physical data models and maintain the model versions in Model Manager for further enhancements.
* Used ScalaAPI for programming in Apache Spark and imported data using Sqoop from Teradata using Teradata connector and developed multiple POCs using Scala and Pyspark and deployed on the Yarn cluster, compared the performance of Spark, and SQL.
* Developed Spark scripts by using Scala shell commands as per the requirement and analyzed the data using Amazon EMR.
* Developed export framework using Python, Sqoop, Oracle & Mysql and Created Data Pipeline of Map Reduce programs using Chained Mappers.
* Used PySpark-SQL to load JSON data and create schema RDD, Data Frames and loaded it into Hive Tables and handled structured data using Spark-SQL.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis and implemented Optimized join base by joining different data sets to get top claims based on state using Map Reduce.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's and Improved the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in HDFS and created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Created Hive Generic UDF's to process business logic that varies based on policy and Imported Relational Data base data using Sqoop into Hive Dynamic partition tables using staging tables.
* Worked on custom Pig Loaders and storage classes to work with variety of data formats such as JSON and XML file formats.
* Used Spark as an ETL tool to remove Duplicates, Joins and aggregate the input data before storing in a Blob and extensively worked on developing Informatica Mappings, Mapplets, Sessions, Worklets and Workflows for data loads.

**Environment:** Hadoop, HDFS, HBase, Spark, MapReduce, Teradata, Informatica, MySQL, Java, Python, Hive, Pig, Data Warehousing, Sqoop, Flume, Oozie, SQL, Cloudera Manager, Erwin, , Cassandra, Scala, Python, AWS( EMR, S3, EC2, Athena, Glue, Redshift), SQL, Elastic Search, Kafka, Tableau, ETL.

**Data Engineer/Analyst**

**Health First, New York NY Jan 2015 to Jan 2017**

**Responsibilities:**

* Provided Data Architecture solutions for multiple relational and dimensional models and involved in Data Warehouse Support –StarSchema and Dimensional modeling to help design DataMart and data warehouse.
* Involved in architecting system interfaces, understand interface requirement and design data model (Logical and Physical) using ERwin and Deliverables are PDM, DDL scripts and STTM documents.
* Imported Data from Different Relational Data Sources like RDBMS, Teradata to HDFS using Sqoop and imported Bulk Data into HBase Using Map Reduce programs and perform analytics on Time Series Data exists in HBase using HBaseAPI.
* Designed and implemented Incremental Imports into Hive tables and used Rest API to Access HBase data to perform analytics.
* Forward engineering the data model and creating the development Database modify and maintain all changes to the Data model and Database scripts.
* Importing of data from various data sources, performed transformations using Hive, MapReduce, and loaded data into HDFS& Extracted the data from MySQL into HDFS using Sqoop.
* Worked in Loading and transforming large sets of structured, semi structured and unstructured data
* Written Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
* Work in team using ETL tool Informatica to populate the database, data transformation from the old database to the new database using Oracle and SQL Server.
* Worked with Cassandra for non-relational data storage and retrieval on enterprise use cases and wrote MapReduce jobs using Java API and Pig Latin.
* Created reports for the BI team using Sqoop to export data into HDFS and Hive and involved in creating Hive tables and loading them into dynamic partition tables.
* Involved in Teradata utilities (BTEQ, Fast Load, Fast Export, Multiload, and Tpump) in both Windows and Mainframe platforms.
* Involved in managing and reviewing the Hadoop log files and migrated ETL jobs to Pig scripts to do Transformations, even joins and some pre-aggregations before storing the data to HDFS.
* Worked on Talend ETL tool and used features like context variable and database components like input to oracle, output to oracle, tFile compare, tFile copy, to oracle close ETL components.
* Developed several behavioral reports and data points creating complex SQL queries and stored procedures using SSRS and Excel and developed different kind of reports such as Drill down, Drill through, Sub Reports, Charts, Matrix reports, Parameterized reports and Linked reports using SSRS.
* Worked on NoSQL databases including HBase and MongoDB. Configured MySQL Database to store Hive metadata.
* Deployment and Testing of the system in Hadoop MapR Cluster and worked on different file formats like Sequence files, XML files and Map files using Map Reduce Programs.
* Developed multiple MapReduce jobs in Java for data cleaning and preprocessing and imported data from RDBMS environment into HDFS using Sqoop for report generation and visualization purpose using Tableau.
* Developed the ETL mappings using mapplets and re-usable transformations, and various transformations such as source qualifier, expression, connected and un-connected lookup, router, aggregator, filter, sequence generator, update strategy, normalizer, joiner and rank transformations in Power Center Designer.

**Environment:** Erwin, SQL, Oracle, SSIS, Hadoop, Teradata, HDFS, Map Reduce, Hive, HBase, Oozie, Sqoop, Pig, Tableau, Rest API, Maven, Strom, SQL, ETL, PySpark, JavaScript, Shell Scripting.

**Data Analyst**

**GenRe, Chicago IL Nov 2012 to Dec 2014**

**Responsibilities:**

* Involved in architecting system interfaces, understand interface requirement and design data model (Logical and Physical) using ER Studio and Deliverables are PDM, DDL scripts and STTM documents.
* Performed Reverse Engineering of the legacy application using DDL scripts in ER Studio, and developed Logical and Physical data models for Central Model consolidation.
* Defining data governance process: processes for data quality rules definition, review process, communication plan, templates, etc.
* Developed MapReduce programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the EDW.
* Involved in performing extensive Back-End Testing by writing SQL queries and PL/SQL stored procedures to extract the data from SQL Database using Hadoop
* Strong ability in developing advanced ANSI SQL queries to extract, manipulate, and/or calculate information to fulfil data and reporting requirements including identifying the tables and columns from which data is extracted.
* Closely worked with BI team to write DAX queries for building the Cube in Azure Analysis Services.
* Analyzed, Designed, and Developed OBIEE Metadata repository (RPD) that consists of Physical Layer, Business Mapping and Model Layer and Presentation Layer.
* Responsible for all metadata relating to the EDW's overall data architecture, descriptions of data objects, access methods and security requirements and developed and automated multiple departmental Reports using Tableau and MS Excel.
* Review & analyzing SQL Server databases in terms of finding performance bottlenecks & prepare optimization strategies as well.
* Involved in designing and maintaining Oracle Database in Clustered environment (RAC) on both Windows and UNIX.
* Worked on all types of transformations that are available in Power BI query editor and wrote calculated columns, Measures query’s in Power BI desktop to show good data analysis techniques.
* Design and document the complete ETL specification document detailing the queries involved for extraction, transformation logic and loading methods.
* Used the Agile Scrum methodology to build the different phases of Software development life cycle.
* Create logical and physical data models using best practices to ensure high data quality and reduced redundancy, Define, document, and articulate design goals and standards.
* Wrote and executed SQL queries to verify that data has been moved from transactional system to DSS, Data warehouse, data mart reporting system in accordance with requirements.
* Worked in importing and cleansing of data from various sources like Teradata, Oracle, flatfiles, SQLServer2010 with high volume data.
* Identifying functional and technical gaps and writing complex SQL's for generating reports and done performance Tuning of SQL and Database server with huge data volume across all the clients.
* Generated and DDL (Data Definition Language) scripts using ER Studio and assisted DBA in Physical Implementation of data Models.
* Created and maintained partitions for large tables, used 11g partitioning feature automatic partitions creations.

**Environment:** ER Studio, MS Visio, Oracle 11g,Oracle Designer, CRM, Hadoop, Power BI, Teradata, GIT, DynamoDB, SQL Server 2010, Oracle, SQL, PL/SQL, JIRA, Windows7, ERP, and UNIX, and SSIS ETL Tool.