**Sr Data Engineer**

**Sameer Pasha Syed**

**Email:** [smrpshsyed0116@gmail.com](mailto:smrpshsyed0116@gmail.com)

**Contact: +1 (469) 305-1697**

**Linkedin:** www.linkedin.com/in/sameer-syed-85a05a273

**Professional Summary:**

* + - * Around 9+ years of professional experience in IT which includes experience in Big Data Ecosystem with Ingestion, Query processing and Analysis of big data.
* Experience with processing large sets of structured, unstructured, and semi-structured data using **Spark** and **Scala**.
* Excellent hands-on with importing and exporting data from different Relational Database Systems like **MySQL** and **Oracle** into **HDFS** and **Hive** using **Sqoop**.
* Knowledge of job workflow scheduling and monitoring tools using **Control-M** and **Zookeeper**.
* Experienced in writing **Spark** applications in Scala and Python (**Pyspark**).
* Experienced in writing and implementing **Shell** **Scripts** for automating jobs.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and **Pyspark** concepts
* Experienced in handling messaging services using **Apache Kafka**.
* Designed and implemented database solutions in **SQL Data Warehouse** and **Azure SQL.**
* Hands-on with real-time data processing using distributed technologies **Kafka** and **Storm**.
* Experience in designing & developing applications using Big Data technologies HDFS, Map Reduce, Sqoop, Hive, PySpark & Spark SQL, Hbase, Python, Snowflake, S3 storage, Airflow.
* Experience in doing performance tuning for map reduce jobs & hive complex queries.
* Experience in efficiently doing ETL’s using Spark - in memory processing, Spark SQL and Spark streaming using Kafka distributed messaging system.
* Understanding of structured data sets, data pipelines, ETL tools, data reduction, transformation and aggregation technique, Knowledge of tools such as DBT, DataStage
* Good understanding of various Hadoop distribution platforms Cloudera, Hortonworks, and cloud platforms like Amazon AWS
* 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 accessandMigrating On premise databases to **Azure Data lake store** using Azure Data factory.
* Have good knowledge in Job Orchestration tools like Oozie, Zookeeper & Airflow.
* Very capable at using Amazon Web Services utilities such as EMR, S3 and CloudWatch to run and monitor Hadoop/Spark jobs on AWS.
* Written PySpark job in AWS Glue to merge data from multiple tables and in Utilizing Crawler to populate AWS Glue Data Catalog with metadata table definitions.
* Generated a script in AWS Glue to transfer the data and utilized AWS Glue to run ETL jobs and run aggregation on PySpark code.
* Having knowledge in Amazon EC2, S3, VPC, RDS, Elastic Load Balancing, Autoscaling, IAM, SQS, SWF, SNS, Security Groups, Lambda, Cloud Watch services
* Having very good experience in MS Azure offerings like Azure Data Factory, Azure Power BI, Azure SQL, Azure Virtual Machines, and Azure Data Lake Store.
* Ability in Azure ADF having hands-on experience in both ADF v1 and ADF v2.
* Hands-on experience in handling database issues and connections with SQL and NoSQL databases such as MongoDB, HBase, Cassandra, SQL server, and PostgreSQL. Created Java apps to handle data in MongoDB and HBase. Used Phoenix to create SQL layer on HBase.
* Experience in designing and creating RDBMS Tables, Views, User Created Data Types, Indexes, Stored Procedures, Cursors, Triggers and Transactions.
* Expert in designing ETL data flows using creating mappings/workflows to extract data from SQL Server and Data Migration and Transformation from Oracle/Access/Excel Sheets using SQL Server SSIS.
* Expert in designing Parallel jobs using various stages like Join, Merge, Lookup, remove duplicates, Filter, Dataset, Lookup file set, Complex flat file, Modify, Aggregator, XML.
* Hands-on experience with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SES, SQS, Lambda, EMR and other services of the AWS family.
* Created and configured new batch job in Denodo scheduler with email notification capabilities and Implemented Cluster setting for multiple Denodo node and created load balance for improving performance activity.
* Instantiated, created, and maintained CI/CD (continuous integration & deployment) pipelines and apply automation to environments and applications. Worked on various automation tools like GIT, Terraform, Ansible.
* Experienced in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
* Experienced with JSON based RESTful web services, and XML/QML based SOAP web services and also worked on various applications using python integrated IDEs like Sublime Text and PyCharm.
* Experience with Agile development methodologies and Azure GitHub repository.
* Developed web-based applications using Python, DJANGO, QT, C++, XML, CSS3, HTML5, DHTML, JavaScript and jQuery.

**TECHNICAL SKILLS**

**Big Data/Hadoop Technologies:** HDFS, MapReduce, YARN, Hive, Pig, HBASE, Impala, Zookeeper, Sqoop,

OOZIE, Apache Cassandra, Flume, Spark, AZURE, AWS, EC2

**Languages:** C, Java, HTML5, DHTML, CSS3, SQL, Json, PL/SQL, Scala, Shell Scripts

**Databases:** NoSQL, Oracle, DB2, MySQL, SQL Server, MS Access, HBase

**NO SQL Databases:** Cassandra, MangoDB, HBase

**Cloud:** Azure(Azure Lake, Azure storage), Aws

**Application Servers:** WebLogic, WebSphere, Apache Tomcat, JBOSS

**Build Tools:** Jenkins, PostgreSql, Oozie, SOAP UI

**Reporting Tools:** Jaspersoft, Qlik Sense, Tableau, JUnit

**Professional Experience**

**Client: Walmart - Arkansas Oct 2021 - Till Date**

**Azure Data Engineer**

**Responsibilities:**

* Worked closely with business teams, transforming business requirements to technical requirements as part of Design Reviews & Daily Project Scrums.
* Created **Sqoop** jobs to import data from Oracle to Hive tables.
* Worked on exporting the data from Hiveto **Teradata** using Hadoop batch jobs.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of **Azure Data Factory**, **Spark SQL,** **T-SQL**, and **Azure Data Lake** Analytics.
* Experience in working with **Spark SQL**, which involves transferring data from Teradata to Azure.
* Data Ingestion to one or more Azure Services - (**Azure Data Lake, Azure Storage, Azure SQL, Azure SQL Data Warehouse**) and processing the data in **Azure Databricks**.
* Hands-on experience setting up, configuring, and managing Azure Data Factory and ingestion pipelines to pull data to Azure Data Lake Store and Azure Blob Storage.
* Worked with the Azure copy process to transfer data from On-Premise SQL Server to Azure SQL Data Warehouse.
* Worked on reading and writing multiple data formats like JSON, ORC, and Parquet on HDFS using **Pyspark.**
* Developed Hive queries to pre-process the data required for running the business process.
* Worked on Various tools like **Apache NiFi**, **Kafka**, **Stream Sets,** and **Navigator** for POC’s.
* Involved in the development and implementation of **SSIS** and **SSAS** applications.
* Implemented Spark using **Pyspark** and **Spark SQL** for faster testing and processing of data.
* Developed **Oozie** workflows to schedule **Sqoop**, **Hive**, and **Spark** jobs.
* Used **Spark** to process the data before ingesting the data into the HBase. Both Batch and real-time spark jobs were created using Scala.
* Designed **SSIS** packages to transfer data from flat files, Excel, SQL Server using Business Intelligence Development Studio.
* Developed JSON Scripts for deploying the Pipeline in **Azure Data Factory** (ADF) that process the data using the Cosmos Activity.
* Analyzed large volumes of structured data using **Spark SQL.**
* Used **SSIS** transformations such as Lookup, Aggregate, Conditional split, SQL task and Send Mail task etc.
* Creating ETL **Informatica** mappings, workflows and enhancing existing mappings to facilitate the data load in system.
* Developed a data pipeline for data processing using Kafka-Spark API.
* Worked on performance tuning for **Spark** applications.
* Used **GitHub** and **GitLab** for version control.
* Implemented **Spark** RDD’s transformations and performed actions to implement the business analysis.
* Used **SSIS** and **T-SQL** Stored Procedures to load data into staging area from OLTP databases.
* Involved in testing at the database end and reviewing the **Informatica** Mappings as per the business logic.
* Used **Jenkins** to deploy scripts to other environments by following CI/CD process.
* Created Hive **Partitions** and **Buckets** to optimize the performance of join operations.
* Created **ETL** (Informatica) jobs to generate and distribute reports from **MySQL** database.
* Implemented POC to migrate map reduce jobs into **Spark RDD** transformations.
* Worked in an Agile development environment in sprint cycles of two weeks by dividing and organizing tasks.

**Environment**: MS SQL Server 2016, T-SQL, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), Management Studio (SSMS), Advance Excel (creating formulas, pivot tables, Hlookup, Vlookup, Macros), Spark, Python, ETL, Data Bricks, Power BI, Tableau, Hive/Hadoop, Snowflakes, Power BI, Data Pipeline, IBM Cognos 10.1, Data Stage

**BCBS - Chicago, IL Jan 2020 – Oct 2021**

**Azure Data Engineer**

**Responsibilities:**

* Worked with a team of 15 to Support the entire Microsoft Agile BI Operations in Azure Data Factory, Azure Data bricks, SQL Server databases, Analysis Services Cubes (SSAS), Power BI, Azure Synapse Analytics, Agent Jobs, SQL Server Virtual Machines environments.
* Created Azure Data factory pipelines for loading the data to the Azure SQL database from different platforms and sources.
* Designed ADF pipelines with multiple chaining activities, using parameters, applying incremental and delta loads, and automated triggers.
* Worked on migration of pipelines from ADFv1 to ADFv2 connecting to multiple sources and destinations.
* Developed multiple apps using the Power Apps platform with SQL, and SharePoint as databases. Created and managed the Power Platform solutions across different environments.
* Extract Transform and Load (ETL) data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-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 Azure Databricks.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load (ETL) data from various sources like Azure SQL, Blob storage, and Azure SQL Data warehouse.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Developed SQL Scripts for automation purposes.
* Migrated On-Premises database to snowflake database via shift and load method in ADF.
* Have extensive experience in creating pipeline jobs, scheduling triggers, Mapping data flows using Azure Data Factory(V2), and using Key Vaults to store credentials.
* Have good experience working with Azure BLOB and Data Lake storage and loading data into Azure SQL Synapse analytics (DW).
* Identified the root cause for the long-running processes, Dead-lock processes, stuck processes, and time-out issues, and provided the fixes in SSAS, Power BI, and Azure Data warehouses, agent jobs, and ETL packages.
* Handled various production failures on Agent jobs, Databases, Data bricks, Cube Processing, Windows Servers, Services, and Azure VMs.
* Performed the requirement gathering, requirement analysis, design, and development of Azure Analysis Services Tabular models (SSAS) and Power BI visualizations.
* Automated the processing of Azure Analysis Services Tabular models using Azure Automation.
* Implemented role-based security in SSAS tabular and power BI models for various azure active directory groups based upon the company location.
* Implemented code check-in/check-out and managed multiple versions of complicated code within TFS.
* Handled critical PRODUCTION requests and analyzed and resolved SQL job failures and AZURE pipelines.

**Environment:** Azure Analysis Services, Power BI, Azure SQL Database, Python, Numpy, Pandas, Keras, Tensor Flow, Azure CLI, Azure HD Insights, Eclipse, IntelliJ

**Client: Express Scripts - Austin, TX Mar 2017 – Dec 2019**

**Aws Data Engineer**

**Responsibilities:**

* Evaluated suitability of Hadoop and its ecosystem to the project and implementing, validating with various proof of concept (POC) applications to eventually adopt them to benefit from the Big Data Hadoop initiative.
* Estimated the Software & Hardware requirements for the Name node and Data nodes in the cluster.
* Experience in migrating existing databases from on premise to AWS Redshift using various AWS services.
* Developed Columnar Data Storage, Advanced Compression, and Massive Parallel Processing using Redshift Multi-node.
* Utilizing Amazon Batch's TWS scheduler to import data from DynamoDB to Redshift in batches.
* Experience implementing and deploying (Elastic Map Reduce) EMR clusters using Amazon Web Services and EC2 instances.
* Understanding of AWS compute services such as EC2, EMR, and EBS, as well as accessing metadata for instances.
* Developed the Pysprk code for AWS Glue jobs and for EMR.
* Installed and configured Hadoop Map Reduce, HDFS, developed multiple Map Reduce jobs in java and Scala for data cleaning and preprocessing.
* Utilize Hadoop infrastructure for data storage and analysis, such as MapReduce, Hive, HBase, Sqoop, Spark RDDs.
* Developed Java Map Reduce programs for the analysis of sample log file stored in cluster
* Implemented Spark using Python and Spark SQL for faster testing and processing of data.
* Imported data using Sqoop to load data from MySQL to HDFS on regular basis.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala
* Used IAM to create new accounts, roles and groups and polices and developed critical modules like generating amazon resource numbers and integration points with S3, Dynamo DB, RDS, Lambda and SQS Queue
* Reviewing the explain plan for the SQLs in snowflake
* Developed ETL parsing and analytics using Python/Spark to build a structured data model in Elastic search for consumption by the API and UI.
* Developed ETL jobs using Spark -Scala to migrate data from Oracle to new Cassandra tables.
* Used Spark -Scala (RDD’s, Data frames, Spark Sql) and Spark - Cassandra -Connector API's for few tasks (Data migration, Business report generation etc.)
* Created Partitions, Buckets based on State to further process using Bucket based Hive joins.
* Created an e-mail notification service upon completion of job for the team which requested for the data.
* Implemented security to meet PCI requirements, using VPC Public/Private subnets, Security Groups, NACLs, IAM roles, policies, VPN, WAF, Trust Advisor, Cloud Trail etc. to pass penetration testing against infrastructure
* Defined job work flows as per their dependencies in Oozie.
* Played a key role in productionizing the application after testing by BI analysts.

**Environment**: MapReduce, Hive, Sqoop 1.4.4, Oozie 4.2, Python, Scala, Spark 2.3, Kafka, Ambari, Cassandra, Linux, AWS EMR, S3, Storm

**Client: Black Knight – Jacksonville, FL. May 2015 – Feb 2017**

**Aws Data Engineer**

**Responsibilities:**

* Written Spark applications using Scala to interact with the PostgreSQL database using Spark SQL Context and accessed Hive tables using Hive Context.
* Involved in designing different components of system like big-data event processing framework Spark, distributed messaging system Kafka and SQL database PostgreSQL.
* Implemented Spark Streaming and Spark SQL using Data Frames.
* I have integrated product data feeds from Kafka to Spark processing system and store the order details in PostgreSQL data base.
* Created functions and assigned roles in AWS Lambda to run python scripts, and AWS Lambda using java to perform event driven processing
* Created multiple Hive tables, implemented Dynamic Partitioning and Buckets in Hive for efficient data access.
* Designed tables and columns in Redshift for data distribution across data nodes in the cluster keeping columnar database design considerations
* Create, modify and execute DDL in table AWS Redshift and snowflake tables to load data
* Involved in creating Hive External tables, also used custom SerDe's based on the structure of input file so that Hive knows how to load the files to Hive tables.
* Managed large datasets using Panda data frames and MySQL
* Monitor Resources and Applications using AWS Cloud Watch, including creating alarms to monitor metrics such as EBS, EC2, ELB, RDS, S3, SNS and configured notifications for the alarms generated based on events defined
* Monitor System health and logs and respond accordingly to any warning or failure conditions.
* Worked on scheduling all jobs using Oozie.

**Environment**: AWS EMR 5.0.0, EC2, S3, Oozie 4.2, Kafka, Spark, Spark SQL PostgreSQL, Shell Script, SQOOP1.4, Scala, Kafka

**Birla Sun Life Insurance - Hyderabad, India July 2013 – May 2014**

# Role: SQL Server BI Developer

**Responsibilities:**

* Analyzing and understanding the existing reporting environment.
* Worked closely with the Business analysts for requirements gathering, designing the database and creating workflow for report.
* Created one central SSIS Master Package to execute multiple child packages with Control Flow and Data Flow. Utilized Parent Package and XML Configuration to dynamically pass variable values on runtime.
* Participated in identifying data migration issues and resolved them.
* Extensively used variables, break point, check point, logging, package configuration and event handler in SSIS packages to meet the business needs.
* Involved in performance tuning to optimize queries and enhance the performance of databases, SQL queries, and stored procedures using SQL Profiler, Execution Plan and Index Tuning Wizard.
* Created Complex SSAS Cubes with multiple fact measure groups, and multiple hierarchies based on the OLAP reporting needs.
* Built MDX queries for Analysis Services & Reporting Services.
* Designed, developed, created and tested PivotTable/PivotChart reports based on OLAP cubes and offline cubes.
* Created multiple partitions and aggregations for the different measure groups for improving performance of the cubes.
* Resolving the SSAS cube connectivity and data issues as and when needed.
* Interacted with Business Users to help them understand to generate reports/look at the business data with various drill down options with Excel via connecting to SQL Server Analysis Services (SSAS).

**Environment:** SQL Server 2012, SQL Server Integration Services (SSIS), Reporting Services (SSRS), TFS, Pivot Tables, MS Visual Studio.Net, C#, SQL Profiler, Windows 2003/2007 Server OS.

**Educational Details:**

* Bachelor’s completed in Electronics and communication engineering, Jawaharlal Nehru Technological University, Hyderabad- 2013.