Soyed Khan

Professional Summary

- Having 12+ years of IT experience in analysis, design, development, testing, delivery, and production support in Python along with Big Data experience in Hadoop ecosystem such as Hive, Pig, Flume, Sqoop, HBase, SPARK, Kafka, Python, AWS, Azure, EC2, Dynamo DB, S3, Kinesis, APP FLOW and Cloudwatch.
- Hands-on experience in Data Modeling, Dimensional Modeling, implementation, and support of various applications in OLTP and Data Warehousing.
- Experience in dealing with Apache Hadoop components like HDFS, HiveQL, Sqoop, Big Data and Big Data Analytics.
- Strong experience in migrating to other databases to snowflake.
- Experience with snowflake multi-cluster warehouses.
- Experience in building snow pipe, snowflake cline and time travel.
- Hands on experience on AWS, S3, EMR, GLUE and knowledge in Microsoft Azure, ADF, ADLS.
- Hands-on experience in writing shell scripts for application utilities.
- Experienced in working with Dev-Ops tools like Jenkins to perform Continuous integration and Continuous delivery. Worked on Jenkin files using Groovy.
- Strong working experience in all phases of development including Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses and Data Marts using IICS Informatica Cloud (CIH,CDI,CAI) and Power Center (Repository Manager, Designer, Server Manager, Workflow Manager, and Workflow Monitor).
- Good Experience in using Informatica Power Center 10.1/9.1 and Informatica Cloud for extraction, transformation and loading mechanism.
- Expertise also in NoSQL databases like MongoDB, Map R-DB and Cassandra.
- Experience working on different Hadoop distributions like Cloudera.
- Proficient in Alteryx Designer for data blending, data cleansing, and advanced data analytics tasks.
- Proficient in SSIS for designing and executing robust ETL packages to migrate data from heterogeneous sources to SQL Server.
- Expertise in SSRS for creating interactive, tabular, graphical, or free-form reports from relational, multidimensional, or XML-based data sources.
- Proficient in utilizing Microsoft Power BI to design and distribute interactive and analytical visual reports.
- Proficient in designing and implementing ETL processes to facilitate data integration, migration, and warehousing.

- Experienced in implementing Spark RDD transformations, actions to implement business analysis.
- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
- 3+ years of experience as Azure Cloud Data Engineer in Microsoft Azure Cloud technologies including Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), Azure Synapse Analytics (SOL Data warehouse), Azure SQL Database, Azure Analytical services, Polybase, Azure Cosmos NoSQL DB, Azure Key vaults, Azure HDInsight Big Data Technologies like Hadoop, Apache Spark, and Azure Data bricks.
- Proficient in Semarchy xDM for designing and deploying master data management (MDM) models.
- Proficient in Tableau Desktop for building dynamic and interactive dashboards to visualize complex data sets.
- Experience in designing Azure Cloud Architecture and Implementation plans for hosting complex application workloads on MS Azure.
- Spearheaded the implementation of end-to-end Master Data Management solutions, driving data consistency and accuracy across enterprise systems.
- Proficient in using Salesforce CRM to streamline sales processes, boost productivity, and enhance customer relationships.
- Demonstrated knowledge in developing varied data loads (incremental, history/full, survivorship rules) with dedicated experience; additionally, adept in implementing change data capture (CDC) methodologies.
- Proficient in version control using Git, including branching, merging, and commit strategies.
- Strong in Data Warehousing concepts, Star schema and Snowflake schema methodologies, understanding Business process/requirements.
- Experienced in ETL TALEND Data Fabric components and used features of Context Variables, MySQL, Oracle, Hive Database components.Hands-on experience in Hadoop Framework and its ecosystem like Map Reduce Programming, Spark, Hive, Impala, Pig, Sqoop, HBase, Oozie, and Python.
- Developed UDFs using both Data Frames/SQL/Data sets and RDD in Spark for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
- Experience in importing and exporting data using stream processing platforms like Flume and Kafka.
- Hands on databases like Oracle, MS SQL, DB2 and developing in RDBMS that includes SQL queries, Stored procedures, and triggers.
- Demonstrated expertise in crafting intricate SQL queries to extract, manipulate, and analyze large datasets, ensuring data accuracy and meeting business requirements.
- Experience in developing applications on different platforms like Windows, UNIX, and LINUX.
- Experience in using GIT, BITBUCKET for version controlling and error reporting and project management tools like JIRA, RALLY.

Professional Experience

Lead Data Engineer| Highmark Health, Pittsburgh, PA – Remote | Oct 2020 – Present

Responsibilities

- Writing Pig scripts to generate MapReduce jobs and performing ETL procedures on data in Azure Data Lake Storage.
- Worked on Azure DevOps and Azure Boards.
- Wrote PowerShell scripts to copy or move data from local file system to Azure Blob Storage.
- Performed the migration of Azure Data Lake Analytics and Azure Databricks jobs from on-premise MapR to Azure cloud using Azure HDInsight and Azure Databricks.
- Involved in Azure Data Factory implementation, which helps in loading data from various RDBMS sources to Azure services and vice versa.
- Designed and implemented a configurable data delivery pipeline for scheduled updates to customer-facing data stores built with Python using Azure Data Factory.
- Involved in cluster maintenance, cluster monitoring, adding and removing cluster nodes, and installed and configured Azure HDInsight, Data Lake Storage, and Data Factory, and developed multiple MapReduce jobs in Java for data cleaning and pre-processing.
- Created and maintained various Shell and Python scripts for automating various processes, and optimized MapReduce code, Pig scripts, and performance tuning and analysis.
- Implemented Azure Data Factory for authoring, scheduling, and monitoring data pipelines.
- Designed several DAGs (Directed Acyclic Graph) using Azure Data Factory for automating ETL pipelines.
- Utilized Azure Data Lake Storage, Event Hubs, Blob Storage, and Data Lake Analytics for implementing data ingestion pipelines.
- Responsible for data services and data movement infrastructures in Azure.
- Worked on architecting the ETL transformation layers and writing Spark jobs in Azure Databricks for data processing.
- Implemented and data integration in developing large-scale system software experience with HDInsight components like HBase, Data Factory, Storage, SQL, Data Lake Analytics, and Databricks.
- Implemented Azure Databricks streaming to pick up data from Azure Event Hubs and send it to Spark pipeline.
- Writing PySpark and SparkSQL transformations in Azure Databricks to perform complex transformations for business rule implementation.

- Designed and implemented incremental imports and delta imports using Azure Data Factory on tables without primary keys and dates from Teradata and SAP HANA, and append directly into Azure Data Lake Storage.
- Developed MapReduce programs in Java for parsing raw data and populating staging tables in HDInsight.
- Developed Spark code using Scala and Spark SQL/Streaming for faster testing and processing of data in Azure Databricks.
- Analyzed SQL scripts and designed solutions to implement using Scala in Azure Databricks.
- Used Spark SQL to load JSON data, create schema RDD, and load it into Hive tables, handling structured data using Spark SQL in Azure Databricks.
- Implemented Spark scripts using Scala, Spark SQL to access Hive tables in Azure Databricks for faster processing of data.
- Extracted, transformed, and loaded data from source systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL, and U-SQL in Azure Data Lake Analytics. Data ingestion to one or more Azure services such as Data Lake Storage, Blob Storage, SQL Database, and Synapse Analytics, and processing the data in Databricks.
- Tested Apache Tez for building high-performance batch and interactive data processing applications on Pig and Hive jobs in Azure HDInsight.
- Involved in creating Hive tables, loading them with data, and writing Hive queries which run internally in a map-reduce way in Azure HDInsight.
- Automated all jobs for pulling data from FTP server to load data into Hive tables using Azure Data Factory pipelines.
- Analyzed the system for new enhancements/functionalities and performed impact analysis of the application for implementing ETL changes using Azure services.
- Integrated Apache Spark with Azure Event Hubs to perform web analytics. Uploaded clickstream data from Azure Event Hubs to Data Lake Storage, HDInsight, and SQL Database by integrating with Spark.
- Built performant, scalable ETL processes to load, cleanse, and validate data using Azure Data Factory and Azure Databricks.
- Collaborated with team members and stakeholders in the design and development of the data environment using Azure services.
- Developed Spark/Scala and Python code for regular expression (regex) projects in the Azure HDInsight/Hive environment with Linux/Windows for big data resources.
- Installed and configured Azure HDInsight ecosystem components.
- Wrote Apache Flume configuration files for importing streaming log data into Azure HDInsight and Data Lake Storage with Flume.

- Imported several transactional logs from web servers with Flume to ingest the data into Azure Data Lake Storage.
- Using Flume and spool directory for loading the data from local system (LFS) to Azure Data Lake Storage.
- Installed and configured Pig and written Pig Latin scripts to convert data from a text file to Avro format in Azure HDInsight.
- Created partitioned Hive tables and worked on them using HiveQL in Azure HDInsight.
- Loaded data into Azure HBase using bulk load and non-bulk load techniques in HDInsight.
- Worked with Tableau and integrated Hive with Tableau Desktop reports and published them to Tableau Server in HDInsight.
- Used Azure Data Factory extensively for ingesting data from disparate source systems.
- Used Azure Data Factory as an orchestration tool for integrating data from upstream to downstream systems.
- Automated jobs using different triggers (Event, Scheduled and Tumbling) in ADF.
- Used Cosmos DB for storing catalog data and for event sourcing in order processing pipelines.
- Designed and developed user defined functions, stored procedures, triggers for Cosmos DB
- Analyzed the data flow from different sources to target to provide the corresponding design Architecture in Azure environment.

Environment: Azure Cloud, Azure Data Factory (ADF v2), Azure functions Apps, Azure DataLake, Azure BLOB Storage, SQL server, SSIS, SSRS, Teradata Utilities, Windows remote desktop, UNIX Shell Scripting, AZURE PowerShell, Data bricks, Python, Erwin Data Modelling Tool, Azure Cosmos DB, Azure Stream Analytics, Azure Event Hub, Azure Machine Learning, Alteryx, Tableau, GITHUB, CDC, Power BI, Salesforce, CRM, Informatica MDM, IDQ/CDQ.

Sr. Data Engineer | Bank Of America, Charlotte, NC | June 2018 – Sep 2020

Responsibilities:

- Utilized Apache Spark with Python to develop and execute Big Data Analytics.
- Worked with PySpark, improving the performance and optimized of the existing applications running on EMR cluster to AWS Glue.
- Performed Transformation and Loading using AWS Glue.
- Configured Hadoop Framework on EC2 instances to make sure application that was created is up and running, troubleshoot issues to meet the desired application state.
- Configured Glue Dev Endpoints to point Glue Job to specify EMR cluster or EC2 instance.
- Worked on PySpark SQL where the task is to fetch the NOTNULL data from two different tables and loads.
- Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using Cloudwatch.

- Designed and Implement test environment on AWS.
- Experience in AWS EC2, configuring the servers for Auto scaling and Elastic load balancing.
- Implemented error handling and data validation techniques to ensure data accuracy and completeness in AWS AppFlow pipelines.
- Collaborated with cross-functional teams to gather requirements and design data integration solutions using AWS AppFlow.
- Worked on ETL Migration services by creating and deploying AWS Lambda functions to provide a serverless data pipeline that can be written to Glue Catalog and queried from Athena.
- Proficient in ETL job scheduling and data warehouse loading, combined with robust management skills overseeing multiple ETL tools. Adept at mentoring peers in ETL development best practices and methodologies.
- Experience in designing and orchestrating ETL job schedules, ensuring timely data extraction, transformation, and loading processes, and optimizing workflow sequences to meet business needs and system requirements.
- Demonstrated proficiency in utilizing ETL tools, especially with SQL Server, to load data warehouses efficiently over a span of 2 years, ensuring data consistency, integrity, and readiness for analytical processing and reporting.
- Skilled in repository management with GitHub, utilizing features like pull requests, issues, and GitHub Actions.
- Create data ingestion modules using AWS Glue for loading data in various layers in S3 and reporting using Athena and Quicksight.
- Working knowledge of Amazon's Elastic Cloud Compute (EC2) infrastructure for computational tasks and Simple Storage Service (S3) to store objects.
- Imported data from AWS S3 into Spark Data frames.
- Optimized CloudWatch configurations to improve performance, reduce costs, and align with best practices.
- Developed CloudWatch Events rules to trigger automated responses to events and changes in the AWS environment, improving operational efficiency and reducing manual intervention.
- Performed transformations and actions on Data frames.
- Creating AWS Lambda functions using python for deployment management in AWS and designed, investigated and implemented public facing websites on Amazon Web Services and integrated it with other applications infrastructure.
- Creating different AWS Lambda functions and API Gateways, to submit data via API Gateway that is accessible via Lambda function.
- Familiar with DAX (Data Analysis Expressions) to create custom formulas and enhance data modeling capabilities.

- Designed intuitive and interactive dashboards with a variety of visualizations, including charts, graphs, maps, and matrices to aid business decision-making.
- Set up an AWS Lambda function that runs every 15 minutes to check for repository changes and publishes a notification to an Amazon SNS topic.
- Extracted data using spark from AWS redshift and performed data Analysis.
- Designing and building multi-terabyte, full end-to-end Data Warehouse infrastructure from the ground up on Redshift for large scale data handling Millions of records every day.
- Responsible for Designing Logical and Physical data modeling for various data sources on Confidential Redshift.
- Developed POC to perform ETL operations using AWS glue to load Kinesis stream data into S3 buckets.

Environment: AWS, Snowflake, Hadoop, Hive, HDFS, Spark, Spark-SQL, SSIS, SSRS, Hive, Sqoop, Oozie, EC2, S3, IAM, VPC, Athena, Glue, Data catalog, CloudTrail, Alteryx, Tableau, GITHUB, CDC, Power BI, SalesForce, CRM Informatica MDM, IDQ/CDQ.

Sr. Data Engineer | Microsoft, Redmond, WA | Apr 2015 – May 2018

Responsibilities:

- Involved in developing New Spark Application using Scala Framework to migrate data from traditional databases and data warehousing, process and transform the data as per the business needs.
- Used Apache airflow in GCP composer environment to build data pipelines and used various airflow operators like bash operator, Hadoop operators and python callable and branching operators.
- Worked on architecting and configuring secure VPC, Subnets, and Security Groups through private and public networks.
- Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators both old and newer operators.
- Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
- Used Google Cloud Platform (GCP) to build, test and deploy applications on Google's very adaptable and solid framework for web, portable and backend arrangements.
- Worked on a migration project to migrate data from different sources (Teradata, Hadoop, and DB2) to Google Cloud Platform (GCP) using UDP framework and transforming the data using Spark Scala scripts.
- Worked on creating data ingestion processes to maintain Global Data Lake on the GCP cloud and Big Query.
- Normalized the data according to the business needs like data cleansing, modifying the data types and various transformations using Spark, Scala and GCP Dataproc.
- Built a system for analyzing the column names from all tables and identifying personal information columns of data across on-premises Databases (data migration) to GCP

- Worked on Apache Dataflow to migrate the data from SQL server, ERP systems and FTP to BigQuery in GCP.
- Writing big query to get data wrangling for with halp of data flow in GCP cloud.
- Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
- Setup Alerting and monitoring using stack driver in GCP.
- Design and implement large scale distributed solutions in AWS and GCP clouds.
- Designed and implemented both on - premise and cloud based data analytic on Azure, GCP and AWS solutions
- Used Azure Data warehouse, Hive, Presto, Dremio, Snowflake, Azure blobs, AWS Redshift, GCP BigQuery to feed BI reporting.
- Designed Azure Data warehouse, Azure blobs, Redshift, GCP BigQuery to feed BI reporting.
- Developed data extraction / pipeline jobs using Informatica Cloud, Talend, and SSIS that loads data into Redshift, Azure Data Warehouse and GCP BigQuery.
- Experience in Google Cloud components, Google container builders, and GCP client libraries.
- Developed ETL solution for GCP Migration using GCP Dataflow, GCP Composer, Apache Airflow and GCP Big Query.
- Working on Google Cloud Platform (GCP) services like cloud storage, cloud SQL, stack driver monitoring.
- Cloud Platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
- Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
- Involved in designing and deploying multitude of applications utilizing almost all the GCP stack (Including EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM) focusing on high-availability, fault tolerance, and auto-scaling in GCP cloud formation.

Environment: GCP, Pyspark, GCPs Data Proc, BigQuery, Hadoop, Hive, GCS, Python, Snowflake, Dynamo DB, Oracle Database, Power Bi, SDK'S, Data Flow,Glacier,EC2, EMR Cluster, SQL Database, Synapse, Data Bricks.

Data Engineer | Tesla, Amsterdam, NY | Feb 2012– Mar 2015

Responsibilities:

- Developed data warehouse model in snowflake for over 100 datasets using whereScape.
- Creating Reports in Looker based on Snowflake Connections.
- Loaded the tables from the azure data lake to azure blob storage for pushing them to snowflake.

- Validating the data from SQL Server to Snowflake to make sure it has Apple to Apple match.
- Consulting on Snowflake Data Platform Solution Architecture, Design, Development and deployment focused to bring the data driven culture across the enterprises.
- Develop stored procedures/views in Snowflake and use in Talend for loading Dimensions and Facts.
- Performed ETL data translation using informatica of functional requirements to Source to Target Data Mapping documents to support large datasets (Big Data) out to the AWS Cloud databases; Snowflake and Aurora.
- Performed logical and physical data structure designs and DDL generation to facilitate the implementation of database tables and columns out to the DB2, SQL Server, AWS Cloud (Snowflake) and Oracle DB schema environment using ERwin Data Modeler Model Mart Repository.
- Snowflake data engineers will be responsible for architecting and implementing very large-scale data intelligence solutions around Snowflake Data Warehouse.
- A solid experience and understanding of architecting, designing and operationalization of large-scale data and analytics solutions on Snowflake Cloud Data Warehouse is a must.
- Need to have professional knowledge of AWS Redshift. Developing ETL pipelines in and out of data warehouse using combination of Python and Snowflakes Snows Writing SQL queries against Snowflake.
- Involved in migration from on prem to Cloud AWS migration.
- Process Location and Segments data from S3 to Snowflake by using Tasks, Streams, Pipes, and stored procedures.
- Led a migration project from Teradata to Snowflake warehouse to meet the SLA of customer needs.
- Responsible for Migration of key systems from on-premises hosting to Azure Cloud Services Snow SQL Writing: SQL queries against Snowflake.
- Designing and implement a fully operational production grade large scale data solution on Snowflake Data Warehouse.
- Experience on Migrating legacy data warehouse and other databases (SQL Server / Oracle Database 10g/11g, Teradata 15.0, DB2) to Snowflake.
- Experience in learning architecting data intelligence solutions around Snowflake Data Warehouse and architecting snowflake solutions as developer.
- Build, create and configure enterprise level Snowflake environments. Maintain, implement, and monitor Snowflake Environments.
- Hands-on experience with Snowflake utilities, Snowflake SQL, Snow Pipe, etc.
- Worked in Snowflake advanced concepts like setting up Resource Monitors, Role Based Access Controls, Data Sharing, Virtual Warehouse Sizing, Query Performance Tuning, Snow Pipe, Tasks, Streams, Zero-copy cloning etc.

- Worked on Snowflake Schema, Data Modeling and Elements, and Source to Target Mappings, Interface Matrix and Design elements. Performed data quality issue analysis using Snow SQL by building analytical warehouses on Snowflake.
- Migrated the data from Amazon Redshift data warehouse to Snowflake.
- Involved in code migration of quality monitoring tool from AWS EC2 to AWS Lambda and built logical datasets to administer quality monitoring on snowflake warehouses.
- Responsible for loading data into S3 buckets from the internal server and the Snowflake data warehouse.
- Developed Snowflake views to load and unload data from and to an AWS S3 bucket, as well as transferring the code to production.
- Developed ETL pipelines in and out of data warehouse using a combination of Python and Snowflakes SnowSQL Writing SQL queries against Snowflake.
- Loaded, transformed the data continuously using Snowpipe from Amazon S3 buckets to Snowflake and used Spark Connector.
- Integrated data sources from Kafka (Producer and Consumer API) for data stream-processing in Spark using AWS network.
- Implemented COPY command to unload the data from Snowflake data warehouse to Azure Data Lake Storage Gen 2.

Environment: Snowflake, Pyspark, Hadoop, Hive, Python, Dynamo DB, Oracle Database, Power Bi, SDK'S, Data Flow,Glacier,EC2, EMR Cluster, SQL Database, Synapse, Data Bricks.

Area of Expertise

Big Data/ Hadoop Technologies	Hadoop, MapReduce, HBase, Apache Pig, HDFS, Hive, Sqoop, Apache Spark, Apache Flume, Apache Nifi, Yarn, Zookeeper and Apache Kafka.
Programming Languages	Python, Java, SQL, Scala.
Web Servers	Web Logic, Web Sphere, Apache Tomcat.
AWS Stack	IAM, S3, EC2, VPC, EMR, Glue, Dynamo DB, RDS, Redshift, Cloud Watch, Cloud Trail, Cloud Formation, Kinesis, Lambda, Athena, EBS, DMS, Elastic Search, SQS, SNS, KMS, QuickSight, ELB, Auto Scaling XML,XSL,XSLT,EJB 2.0/3.0,Struts1.x/2.x,Spring2.5,Hibernate3.2.

Azure Stack	Azure Data Factory v2, Azure Data Lake Store, Azure Storage Accounts, T - SQL, Power-Shell, Azure Active Directory, Azure key Vault, Azure Batch, SQL Data Warehouse, Visual Studio Team Services, ARM, Azure Automation, AZURE Runbook, Azure Logic Apps, Azure Analysis Services, Azure Data lake, Azure Cosmos DB
GCP & Snowflake Stack	GCP Cloud Storage, Big Query, Composer, Cloud Dataproc, Cloud SQL, Cloud Functions, Cloud Pub/Sub, Dataflow. Snowflake, SnowSQL, SnowpipeAWS.
Scripting Languages	Shell Scripting, Java script.
Application Build Tools	Apache Ant, Apache Maven.
Databases	Oracle, SQL Server, MySQL, Teradata, NoSQL, SQL, MongoDB Cassandra.
Version Control	Git, SVN, CVS