**Sathwik**

**Azure Data Engineer**

**Phone**: **860-375-5660**

**Email**: **sathwikp.it@gmail.com**

**PROFESSIONAL SUMMARY**

* Over 10+ years of experience in the IT industry as a Data Engineer, showcasing expertise in Data Modelling, Data Validation, Data Source Evaluation, and proficient skills in Data Warehouse, Data Mart Design, ETL, BI, OLAP, and Client/Server applications on AWS and Azure and GCP cloud Services..
* Experience in working with Azure Cloud and its components like **Data Factory**, **Azure Databricks**, **Azure Synapse Analytics**, **Azure Stream Analytics**, **Logic Apps**, **Function Apps** and **Azure DevOps** services.
* Hands-on experience on **Databricks** Workspace UI, Managing **Databricks** notebooks, **Delta Lake** with Pyspark and SparkSQL.
* Designed and implemented complex data transformations using mapping data flows in **Azure Data Factory**, optimizing data processing, and enhancing overall efficiency.
* Experience in working with GCP Cloud and its components like Cloud Data Fusion, Cloud Dataproc, BigQuery, Dataflow, Cloud Functions, and Cloud Build services.
* Experience in IICS (Informatica Intelligent Cloud Services).
* Extensive knowledge of AWS services such as EMR, Redshift, S3, EC2, including server configuration for auto-scaling and elastic load balancing. Skilled in utilizing AWS S3 for data staging, transfer, and archival, with hands-on experience in large-scale data migrations using AWS DMS and implementing CDC in AWS Redshift.
* Implemented data pipelines using ADF, Synapse, Azure functions, Logic apps, stream sets, ADLS Gen2 and snowflake.
* Conducted performance tuning and optimization of queries and data processing jobs within **Azure Synapse Analytics**, enhancing overall system performance and reducing processing time.
* Experience with **Azure Event Hubs, Spark Streaming** and **Apache Kafka** for messaging and streaming applications.
* Experience in creating and managing **Azure DevOps** tools for continuous integration and deployment (CI/CD) pipelines.
* Strong knowledge of data migration strategies and experience in migrating on-premises storage systems to **Azure Storage**.
* Experience in working with Various Big Data Components like **HDFS**, **YARN**, **MapReduce**, **Spark**, **Sqoop**, **Oozie**, **Pig**, **ZooKeeper**, **Hive**, **HBase**, **Kafka** and **Airflow**.
* Experience in building sophisticated metadata models from various data sources using CognosFramework Manager and used Cognos Transformer to build Multidimensional Cubes. Maintaining andSupport Analyst Models and Contributor Applications using the Cognos Planning environment
* Expert in writing SQL queries and optimizing the queries in Oracle, SQL Server 2008, Netezza
* Executed data extraction, transformation, and loading from source systems to Azure Data Storage using Azure Data Factory, T-SQL, and Spark SQL. Well-versed in Big Data technologies, Hadoop, Spark, Python, SQL, Power BI, Tableau, and other Data Engineering tools on both Azure and AWS platforms.
* Hands on experience in working with Azure Cloud and its components like Azure Data Factory, Azure Databricks, Logical Apps, Azure function Apps, snowflake, and Azure DevOps services.
* Good Understanding of Spark Architecture including **Spark Core**, **Spark SQL**, Dataframes, **Spark Streaming**, Driver Node, Worker Node, Stages, Executors and Tasks.
* Implemented the machine learning algorithms using python to predict the quantity a user might want to order for a specific item so we can automatically suggest using kinesis firehose and S3 data lake.
* Implemented data pipelines using SnowSQL, Snowflake Integrated services and snow pipe.
* Good Understanding of Big Data Hadoop and YARN Architecture along with various hadoop demons such as Job Tracker, Task Tracker, Name Node, Data Node, Resource/Cluster Manager.
* Experience on Palantir Foundry and Data warehouses (SQL Azure and Con dential Redshift/RDS).
* Experience in importing and exporting the data using **SQOOP** from **HDFS** to Relational Database systems and vice versa.
* Experience in optimizing query performance in **Hive** using bucketing and partitioning techniques.
* Optimized **Spark** jobs and workflows by tuning **Spark** configurations, partitioning and memory allocation settings.
* Leveraged the power of **PySpark** and **Scala** to execute complex data transformations and aggregations.
* Experience in building, maintaining multiple **Hadoop** Clusters of different sizes and configurations and setting up the rack topology for large clusters.
* Well-versed in Amazon Web Services (AWS) components, including EC2, S3, EMR, Step functions, Lambda, Redshift, and DynamoDB.
* Strong understanding of Data Modelling (Relational, dimensional, Star and Snow flake Schema), Data analysis, Palantir Foundry, implementations of Data warehousing using Windows and UNIX.  Developed mappings in Informatica to load the data from various sources into the Data warehouse different transformations like Source Quali er, JAVA, Expression, Talend, Aggregate, Update Strategy and Joiner.
* Integrated on-premises (MySQL, Cassandra) and cloud-based (Blob storage, Azure SQL DB) data using Azure Data Factory, applying transformations, and loading data into Snowflake.
* Involved in performance tuning the Spark Sql and analyzing the Spark logs and Dag on Palantir Foundry.
* Experience in setting up workflow using **Apache Oozie** workflow engine for managing and scheduling **Hadoop** jobs.
* Optimized data ingestion, Data modelling, data Encryption and performance by tuning **ETL** workflows.
* Experience in NOSQL databases like **HBase**, MongoDB, **Azure Cosmos DB**.
* Processed data into HDFS by developing solutions, analyzed data using **MapReduce**, **Hive**, Pig and produced summary results from Hadoop to downstream systems.
* Executed data extraction, transformation, and loading from source systems to Azure Data Storage using Azure Data Factory, T-SQL, and Spark SQL. Well-versed in Big Data technologies, Hadoop, Spark, Python, SQL, Power BI, Tableau, and other Data Engineering tools on both Azure and AWS platforms.
* Proficient in writing Python scripts, optimizing SQL queries in Oracle, SQL Server, and Teradata. Demonstrates strong familiarity with the Software Development Life Cycle (SDLC) and a comprehensive understanding of testing methodologies
* Good understanding of JavaScript design patterns and MVC frameworks (i.e., Angular.JS, Node.js). Knowledge of common browser developer tools and performance tuning techniques.
* Designed and Developed **ETL** pipelines in and out of **snowflake** using SnowSQL and Snow pipe.
* Extensive experience in developing, maintaining and implementation of EDW, Data Marts, ODS and **Data warehouse** with Star schema and snowflake schema.
* Worked on Snowflake Schema, Data Modeling, Source to Target Mappings, Interface Matrix, and Design Elements.
* Created report models from cubes as well as relational **data warehouse** to create ad-hoc reports and chart reports.
* Performed Database testing on **Data warehouse** by using complex **SQL Queries** in **SQL Server** to validate the data.
* Developed data ingestion workflows to read data from various sources and write it to Avro, Parquet, Sequence, JSON, and ORC file formats for efficient storage and retrieval.
* Implemented complex business logic through T-SQL stored procedures, Functions, views and advanced query optimizations.
* Developed serialization and deserialization routines to convert complex data structures into a compact binary format, reducing storage footprint and improving data transfer efficiency.
* Experience with data governance and data security practices to ensure data privacy and compliance.
* Experience with version control systems like Git, GitHub to push the code for maintaining versions.
* Comprehensive knowledge of Software Development Life Cycle and worked on Agile Methodology.

**EDUCATION**

* Bachelor’s in computer science, Gitam University, India
* Master’s in business Analytics and Project Management, University of CT

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Azure Services | Azure data Factory, Azure Databricks, Logic Apps, Functional App, Azure Synapse Analytics, Azure Stream Analytics, Snowflake, Azure DevOps, Azure Event Hub, Azure Machine Learning, Azure Cosmos DB |
| Hadoop Distribution | Cloudera, Horton Works |
| Big Data Technologies | MapReduce, Hive, Tez, Python, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop, Zookeeper, HDFS, YARN, Airflow. |
| Cloud Platform | AWS (Lambda, S3, EC2, EMR, RDS) |
| Reporting and ETL Tools | AWS GLUE, Tableau, Power BI. |
| Languages | Java, SQL, PL/SQL, Python, HiveQL, Scala. |
| Web Technologies | HTML, CSS, JavaScript, XML, JSP, Restful, SOAP |
| Operating Systems | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| Build Automation tools | Ant, Maven |
| Version Control | GIT, GitHub. |
| Methodology | Agile, Scrum. |
| IDE &Build Tools, Design | Eclipse, Visual Studio. |
| Databases | MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB |

**PROFESSIONAL EXPERIENCE**

**Client: Optum, Eden Prairie, MN Nov 2021 - Till Now**

**Role: Sr Azure Data Engineer/ Snowflake**

**Responsibilities:**

* Developed and maintained end-to-end operations of ETL data pipeline and worked with large data sets in **azure data factory**.
* Developed custom activities using **Azure Functions**, **Azure Databricks** and PowerShell scripts to perform data transformations, data cleaning, and data validation.
* Advised on all emerging technologies to include Artificial Intelligence/Machine Learning, Big Data Analytics, Cloud Architecture, and Mobility.
* Extensive experience as a User Interface/Front End Developer in developing web applications using HTML, XML, CSS, Java Script, jQuery, AJAX, AngularJS Kafka.
* Rigorously used Spark -Scala (RRD’s, Data frames, Spark SQL) and Spark - Cassandra -Connector APIs for various tasks (Data migration, Business report generation etc.).
* Extensive experience in developing, maintain and implementation of EDW, Data Marts, ODS and Data warehouse with Star schema and snowflake schema.
* Responsible for investigating advanced Cognos Analytics V11.1.6 to provide new functionality and performance improvements to current Cognos Reports
* Effectively supported the enterprise reporting solutions and provided continuous support to business users in leveraging self-service Cognos Analytics capabilities and improved efficiency and productivity.
* Promoted best practices, prepared and implemented Cognos BI Standard Operating Procedures (SOPs)to streamline the reporting process.
* Strengthened Cognos Framework Manager modeling experience due to the utilization of Cognos BIBest Practices. This included activities such as adding Determinants into Framework model where necessary, eliminating loop joins, custom SQL queries, modeling metadata as a star/snowflake schema, and so forth.
* Delivered solutions focusing on service offerings of: Event Hub/Grid, IoT hub, Kafka, Azure Data Factoryv2, Azure Data Lake Storage-GEN2, Azure SQL Data Warehouse/Synapse Analytics, Cosmo DB, HDInsight (Hadoop/Hive/Spark/Storm), Machine Learning Studio, Cognitive Services, as well as third-party AI/ML tooling in Azure (PyTorch, SciKt, etc.).
* Loading bulk data into Hadoop environment from netezza using SQOOP.
* Analyzed the sources, transformed the data, mapped the data and loaded the data into targets using PowerCenter Designer.
* Utilized Terraform to automate infrastructure provisioning and management, ensuring consistent and reproducible deployments in a GCP environment.
* Experience in creating Docker containers and Docker images for managing the application life cycle. Good knowledge on Docker components like Swarm manager for Docker swarm clusters Prometheus server
* for machine-centric monitoring as well as monitoring of highly dynamic service-oriented
* Utilized **Azure Synapse analytics** platform to perform advanced data analytics, including complex data transformations, aggregations, and statistical analysis, enabling data-driven decision-making and business insights.
* Created Pipelines in **Azure Data Factory** using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data.

from different sources like **Azure Data Lake**, **Azure SQL**, **Blob storage**, **Azure SQL Pools**, write-back tool and backwards.

* Implemented **Azure EventHub** for real-time data ingestion, enabling efficient streaming and processing of high-volume data.
* Utilized **Terraform** to automate infrastructure provisioning and management, ensuring consistent and reproducible deployments in an Azure environment.
* Did Data Synchronization and Masking tasks using IICS.
* Integrated on-premises (MySQL, Cassandra) and cloud-based (Blob storage, Azure SQL DB) data using Azure Data Factory, applying transformations, and loading data into Snowflake.
* Built **ETL** solutions using **Databricks** by executing code in notebooks on data in data lake using **delta lake** and loading data into Azure SQL Pools.
* Effectively used IICS Data integration to create mapping templates to bring data into staging layer fromhybrid data source systems like Sql Server, Oracle, AWS.
* Automated **dataflows** using **Logic apps** and **power automate** which connects different GCP services and Function apps for customization.
* Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory Palantir Foundry, 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 InAzure Data bricks
* Designed and implemented data engineering solutions using AWS services like EC2, S3, Lambda, and Glue.
* Integrated Kafka with other AWS services to establish end-to-end data pipelines.
* Ensured scalability and flexibility of data solutions on AWS to accommodate growing data volumes.
* Developed data warehouse model in snowflake for over 100 datasets using whereScape.
* Creating Reports in Looker based on Snowflake Connections.
* Code review of all the Kafka Unit test case documents, Palantir Foundry, Talend, EQA documents, completed by team with proper review check list. Also do development for same.
* Utilized Palantir Foundry and Docker for the runtime environment of theCI/CDsystem to build and test deploy. Analyzed, design and build Modern data solutions using Azure PaaS service to support visualization of data. Understand current Production state of application and determine the impact of new implementation on existing business processes.
* Involved in architecting cloud data analytics solutions in Azure and migrating on premise DWH to Azure cloud using services ADLS Gen 2, Azure Data Factory, Azure Databricks, and Azure Synapse
* Installed Kafka on the cluster and configure producer and consumer coding part in java to establish a connection from source to HDFS with popular hash tags.
* Built ETL solutions in **Azure Databricks** by executing code in notebooks against data in data lake and delta lake and loaded data into Azure DW following the Bronze, silver, and Gold Architecture.
* Ingested data from multiple sources like Azure Data Lake, Azure DW, Azure SQL and processed the data in **Databricks**.
* Created python functions to transform the data from Azure storage to Azure SQL on Azure **Databricks** platform.
* Developed various Spark Applications using Pyspark and Spark-SQL in **Azure Databricks** for data extraction, data transformation and aggregation from multiple file formats to uncover insights into the customer usage pattern.
* Experience in capacity planning for Azure Storage, ensuring optimal resource utilization and cost-effectiveness.
* Developed Python, PySpark, Bash scripts to transform and load data across on-premises and cloud platforms.
* Designed and implemented data partitioning strategies within **Azure Synapse Analytics** to optimize query performance and enable parallel processing, resulting in significant reduction in query execution time.
* Designed and developed a new solution to process the NRT (Near Real Time) Data by using **Azure Stream Analytics**, **Azure** Event Hub and Service Bus queue.
* Analyzed data from GCP data storages using Spark jobs in Cloud Dataproc extracting insights from large datasets.
* Developed and executed comprehensive migration strategies for traditional systems levarazing GCP lift and shift approach, Polybase and Cloud Migrate service.
* Implemented and managed AWS-based Continuous Integration/Continuous Deployment (CI/CD) pipelines, automating the build, test, and deployment processes for diverse applications.
* Developed techniques for data warehousing, data cleansing, Slowly Changing Dimension (SCD) handling, surrogate key assignment, and change data capture for Snowflake modeling.
* Built SCD Type2 Dimensions and Facts using Delta Lake in **Azure Databricks**.
* Designed and implemented data orchestration workflows using **Azure Logic Apps** to automate complex data processing tasks, including data validation, enrichment, transformation, and aggregation, ensuring data accuracy and consistency.
* Developed **Azure function apps** as API services to communicate with various Databases.
* Designed and developed user defined functions, stored procedures, and triggers for **Cosmos DB**.
* Implemented End to End Migration of 800+ Objects with 4TB size from **SQL Server** to **Snowflake**.
* Experience in building real-time pipelines using **Azure Stream Analytics** on Near Real time data for continuous data monitoring and anomaly detection and performing time-based aggregations.
* Automated dataflows using **Logic apps** and power automate which connects different azure services and **Function apps** for customization.
* Proficient in ETL (Extract - Transform - Load)using SQL Server Integration Services 2012(SSIS) and Informatica Power Center tool.
* Responsible for the contour graphs using spark data frames wif in Palantir Foundry.
* Used Hadoop and Palantir Foundryto push the messages for the business statistical analysis of the customers related information.
* Developed ELT/ETL pipelines using Python and Snowflake Snow SQL for seamless data movement to and from Snowflake data store.
* Worked on optimizing and tuning the Netezza SQLs to improve the performance of batch. 
* Troubleshoot specific issue related to the optimization in Netezza, including but not limited to dataintegrity and additional queries.
* Familiar with data architecture including data ingestion pipeline design, Hadoop information architecture, data modeling and data mining, machine learning, and advanced data processing.
* Designed and implemented by configuring Topics in the new Kafka cluster in all environments.
* Implemented Java Message Services (JMS) using JMS API.
* Configured data pipeline orchestration using YAML pipelines in **Azure DevOps**, ensuring efficient and reliable execution of data workflows.
* Leveraged **Azure DevOps** for continuous integration and deployment (CI/CD) of data pipelines and applications, streamlining the development and deployment processes.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Azure EventHub ,Powercenter, Azure Stream Analytics, spark streaming, data pipeline, terraform, azure DevOps, Oracle, Spark, SQL,GCP, Python, Scala, PySpark, GIT, Cassandra, Kafka, Cloud DataWare, JavaScript, ADF Pipeline, Power BI, Azure Cosmos DB,IICS, Azure Synapse Analytics, Machine learning, Data Flow, Palantir Foundry, Function App, Snowflake, SQL Server, Azure Data Lake, Netezza, Azure BLOB, Azure SQL Database.

**Client: Marsh McLennan, LA, California Jan 2021 – Oct 2021**

**Role: Azure Data Engineer/ Snowflake**

**Responsibilities:**

* Developed and implemented end-to-end data integration solutions using **Azure Data Factory**, encompassing activities, pipelines, data sets, linked services, and triggers.
* Lead data migration projects, leveraging **Azure Cloud Data Warehouse** to migrate on-premises data warehouses to the cloud, resulting in improved scalability, cost savings, and enhanced data accessibility for business users.
* Worked on optimizing **Azure Storage (ADLS**, **Azure BLOB**) performance by implementing caching, compression, and other performance tuning techniques.
* Analyzed data from **Azure data storages** using spark jobs in **Databricks** extracting insights from large datasets.
* Developed and executed comprehensive migration strategies for traditional systems levarazing azure lift and shift approach, Polybase and **Azure Migrate** service.
* Built the data pipeline using GCP Services like Data Fusion to load the data from Legacy SQL server to Cloud SQL using Data Fusion, API Gateway Services, SSIS Packages, custom .Net and Python codes
* Involved in scheduling the Palantir Foundry jobs to run as trigger event or on schedule time.
* Designed AWS EC2 instance architecture for high-availability application architecture and security parameters.
* Created AWS S3 buckets, managed policies, and utilized S3 buckets and Glacier for storage and backup. Worked on Hadoop cluster and data querying tools for data storage and retrieval.
* Designed and developed SSIS Packages for importing and exporting data from MS Excel, SQL Server, and flat files.
* Collaborated with cross-functional teams, including data scientists, analysts, and engineers.
* Identified and resolving issues related to Kafka and AWS components, minimizing downtime and optimizing performance.
* 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.
* Designed and developed the complex Cognos Framework Manager Metadata Model to support futuregrowth for reporting requirements.
* Created Cognos Transformer Models and generate various Power Cubes from Packages, Reports, andIQD’s as data sources
* Loaded data directly from Oracle to Netezza without any intermediate files.
* Collaborated with the NLP and ML team to design efficient HBase schemas for machine learning applications.
* Processed streaming data for specific use cases using Kafka and Spark Streaming
* Involved in generating the Pyspark framework for generating the Dataframes in Palantir Foundry.
* Migrated on premise data (Oracle, SQL Server, MongoDB) to **Azure Data Lake Storage** (ADLS) using **Azure Data Factory**.
* Developed and Designed fact and dimension tables in **Azure SQL Data Warehouse** including CTAS statements for both full and incremental load into these tables.
* Built the data pipeline using Azure Service like **Data Factory** to load the data from **Legacy SQL server** to **Azure Database** using **Data Factories**, API Gateway Services, SSIS Packages, custom .Net and Python codes.
* Written complex PySpark and Spark SQL transformations in **Azure Databricks** for business rule implementation.
* Integrated **Azure DevOps** for continuous integration and continuous deployment (CI/CD) of data pipelines and applications.
* Utilized **Terraform** for infrastructure provisioning and automation, enabling seamless deployment and management of cloud resources.
* Written several UNIX shell scripts for moving several Staging files into Netezza database using bulkloads.
* Designed and implemented robust data models and schemas using technologies like Apache Hive, Apache Parquet, or Snowflake for efficient data storage, retrieval, and analysis
* Created Spark Clusters and configured high concurrency clusters in **Azure Databricks** to speed up the preparation of high-quality data.
* Designed and deployed automated ETL workflows using AWS Lambda, organized and cleansed data in S3 buckets using AWS Glue, and processed data using Amazon Redshift.
* Used Java Collection classes, interfaces, Spring Boot JAX-RS API in backend for building custom REST API’s.
* Created Partitions, Buckets based on State to further process using Bucket based **Hive** joins.
* Worked on **RDD’s** & **Data frames (SparkSQL)** using PySpark for analyzing and processing the data.
* Used **Spark Streaming** to divide streaming data into batches as an input to **Spark engine** for batch processing.
* Constructed end-to-end data transformation pipelines utilizing **Hadoop**, **Hive**, **Python**, and **Spark** technologies.
* Fine-tuned and optimized **Spark jobs** to maximize data processing efficiency and performance.
* Implemented Slowly Changing Dimensions to incrementally load data into the **Azure SQL Data Warehouse**.
* Used HTML5, CSS3, JavaScript, Bootstrap, Angular, Node.js as frontend building Environment.
* Implemented data governance and security measures to ensure compliance with industry regulations and protect sensitive data.

**Environment**: Azure Databricks, Data Factory, spark streaming, data pipeline, AWS EC2, AWS S3 buckets, Hadoop, PySpark, AWS Lambda, AWS Glue terraform, azure DevOps, yaml, Spark,GCP, Hive, Palantir Foundry,SQL, Python, Scala, PySpark, GIT, Kafka, ADF Pipeline,Snowflake, Power Bi,Cassandra, Azure SQL Data Warehouse, SQL Server, BLOB, ADLS.

**Client:** **SMBC Bank, Newyork City, NY Oct 2018 – Dec 2020**

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

**Responsibilities:**

* Utilized **Sqoop** to import data from MySQL, Oracle to Hadoop Distributed File System (**HDFS**) on a regular basis, ensuring seamless data integration.
* Performed aggregations on large volumes of data using **Apache Spark** and **Scala** and stored the processed data in **Hive warehouse** for further analysis.
* Created registration and configured in PowerCenter to load data in real-time mode
* Worked extensively with Data Lakes and big data ecosystems, including **Hadoop**, **Spark**, Hortonworks, and Cloudera, to leverage their capabilities for efficient data processing.
* Developed Hive queries to analyze data and meet specific business requirements, utilizing Hive Query Language (**HiveQL**) to simulate **MapReduce** functionalities.
* Create load Flat file data into target region as Dim and fact tables. Legacy PowerCenter jobsare re-written using latest version of Data Modelation..
* Built **HBase** tables by leveraging **HBase** integration with **Hive**, facilitating efficient storage and retrieval of data.
* Applied **Kafka** and **Spark** Streaming to process streaming data in specific use cases, enabling real-time data analysis and insights generation.
* Rigorously used **Spark** -**Scala** (RRD’s, Data frames, Spark SQL) and Spark - Cassandra -Connector API's for various tasks (Data migration, Business report generation etc.)
* Extensively worked on creating combiners, Partitioning, distributed cache to improve the performance of **MapReduce** jobs.
* Designed and implemented a data pipeline using **Kafka**, **Spark**, and **Hive**, ensuring seamless data ingestion, transformation, and analysis.
* Shared data outside using Snowflake to quickly set up to share data without transferring or developing pipelines.
* Developed custom scripts and tools using Oracle's PL/SQL language to automate data validation, cleansing, and transformation processes, ensuring data accuracy and quality.
* Implemented Continuous Integration and Continuous Deployment (**CI/CD**) pipelines to build and deploy projects in the **Hadoop** environment, ensuring streamlined development and deployment processes.
* Worked with **Spark** using Python (PySpark) and **Spark SQL** for faster data testing and processing, enabling efficient data analysis and insights generation.
* Created and maintained HiveQL scripts and jobs using tools such as **Apache Oozie** and **Apache Airflow**.
* Employed Spark Streaming to divide streaming data into batches as input to the **Spark** engine for batch processing, facilitating real-time data processing and analytics.
* Utilized **Zookeeper** for coordination, synchronization, and serialization of servers within clusters, ensuring efficient and reliable distributed data processing.
* Integrated serialization techniques with distributed caching systems, improving data retrieval speeds by implementing efficient **serialization** and **deserialization** of cached data.
* Worked on **Oozie** workflow engine for job scheduling, enabling seamless execution and management of data processing workflows.
* ​​Implemented partitioning in a **data lake** architecture, leveraging Apache **Hive** and Apache **Parquet**, to improve data query performance and enable efficient data exploration and analysis.
* Developed Python Code to gather data from **HBase** and design the solutions using **Pyspark**.
* Analyzed the sql scripts and designed it by using **PySpark** SQL for faster performance.
* Created **Hive Queries** and **Pig scripts** to study customer behaviour by analyzing the data.
* Developed spark applications in **Pyspark** in distributed environments to load huge numbers of CSV files with different schema into **Hive** ORC tables.
* Leveraged **Git** as a version control tool to maintain code repositories, ensuring efficient collaboration, version tracking, and code management.
* Improved system resilience by implementing partition-level backup and recovery strategies, enabling efficient restoration of specific partitions in case of data corruption or failure.
* Utilized **JIRA** to manage project workflows, track issues, and collaborate effectively with cross-functional teams.

**Environment**: Sqoop, MYSQL, HDFS, Apache Spark, Scala, Hive,Snowflake, powercenter, Hadoop, Cloudera, HBASE, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, Python, PySpark, Ambari, JIRA, YARN, Pig, Git.

**Client: Nationwide, Columbus, OH Feb 2017 – Sept 2018**

**Role: Data warehouse Developer**

**Responsibilities:**

* Designed tables, complex **ETL** mappings and workflows in Informatica and **SSIS** to integrate new billing systems into existing data warehouses.
* Source Controlling, environment specific script deployment tracking using **TFS**.
* Created indexed views, UDF and stored procedures to be used by BI Teams for creating reports.
* Developing, Administering, and Managing corresponding databases: Consolidated Data Store, Reference Database (Source for the Code/Values of the Legacy Source Systems), and Actuarial Data Mart.
* Written Triggers, Stored Procedures, Functions, Coding using Transact-SQL (**TSQL**), creating, and maintaining Physical Structures.
* Extensively used fuzzy lookup, fuzzy grouping, slowly changing dimension wizard as well as custom **T-SQL** Code to extend ETL packages.
* Worked on shell scripts to automate data integration process and worked with DBA to resolve performance issues.
* Developed and deployed **SSIS/SSRS** packages for data extraction, transformation, and reporting, resulting in improved data accuracy and timely delivery of business insights.
* Deployment of Scripts in different environments according to Configuration Management, Playbook requirements Create / Manage Files/File group - Table/Index association Query Tuning, Performance Tuning.
* Defect tracking and closing by using Quality Center Maintain Users / Roles / Permissions.

**Environment**: SQL Server 2008/2012 Enterprise Edition, SSRS, SSIS, T-SQL, Windows Server 2003, PerformancePoint Server 2007, Oracle 10g, Visual Studio 2010.

**Client: HSBC Bank, Hyderabad, India Aug 2013 – Oct 2016**

**Role: Data Warehouse Developer**

**Responsibilities:**

* Proficient in dimensional data modelling for Data Mart design, identifying facts and dimensions, and developing fact tables and dimension tables using Slowly Changing Dimensions (SCD) techniques.
* Developed complex stored procedures, efficient triggers, and necessary functions, along with creating indexes and indexed views to optimize performance in **SQL Server**.
* Extensive experience in monitoring and tuning SQL Server performance, employing best practices to ensure optimal database performance.
* Developed and implemented **SSIS** and **SSRS** packages to extract, transform, and load data from various sources, including DB2, SQL, Oracle, flat files (CSV, delimited), APIs, XML, and JSON.
* Skilled in error and event handling techniques, such as precedence constraints, breakpoints, check points, and logging, ensuring reliable and robust **ETL** processes.
* Experienced in building cubes and dimensions with different architectures and data sources for business intelligence purposes, including writing MDX scripting.
* Expertise in designing **ETL** data flows using **SSIS**, creating mappings and workflows for extracting data from SQL Server, as well as performing data migration and transformation from Access/Excel sheets using SQL Server SSIS.
* Proficient in developing **SSAS cubes**, implementing aggregations, defining KPIs (Key Performance Indicators), managing measures, partitioning cubes, and creating data mining models. Deploying and processing **SSAS objects**.
* Experience in creating ad hoc reports and reports with complex formulas, utilizing querying capabilities of the database for business intelligence purposes.
* Expertise in developing parameterized, chart, graph, linked, dashboard, scorecards, and drill-down/drill-through reports on **SSAS** cubes using **SSRS** (SQL Server Reporting Services).

**Environment**: MS SQL Server, Visual Studio, SSIS, Share point, MS Access, Team Foundation server, Git.