PRAJEET
E-mail: prajeetp.chavhan@gmail.com
Ph: 248-973-5987
Sr. GCP Data Engineer/Data Analyst/Spark Developer

PROFESIONAL SUMMARY:

- **12+** years of IT experience in a variety of industries working on Big Data technology using technologies such as Cloudera and Hortonworks distributions. Hadoop working environment includes Hadoop, Spark, MapReduce, Kafka, Hive, Ambari, Sqoop, HBase, and Impala.
- Fluent programming experience with Scala, Java, Python, SQL, T - SQL, R.
- Experienced data professional with expertise in data warehousing, data aggregation, and data modeling using ERwin, facilitating effective data management and analytics.
- Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like MapReduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.
- Adept at configuring and installing Hadoop/Spark Ecosystem Components.
- Proficient with Spark Core, Spark SQL, Spark MLlib, Spark GraphX and Spark Streaming for processing and transforming complex data using in-memory computing capabilities written in Scala. Worked with Spark to improve efficiency of existing algorithms using Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.
- Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into a data warehouse.
- Designed and executed data strategies for healthcare payer projects, leveraging 5 years of experience to optimize data solutions.
- Excellent knowledge on Hadoop Architecture such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce programming paradigm.
- Experience tuning spark jobs for efficiency in terms of storage and processing.
- Experience in creating and executing Data pipelines in GCP and AWS platforms.
- Hands on experience in GCP, Big Query, GCS, cloud functions, Cloud dataflow, Pub/Sub, cloud shell, GSUTIL command- line utilities, Data Proc.
- Experience in developing the big data applications and services using in Amazon Web Services (AWS) platform using EMR, S3, EC2, Lambda, CloudWatch and cloud computing using AWS RedShift.
- Experienced in integrating Hadoop with Kafka, experienced in uploading Clickstream data from to HDFS.
- Experienced in loading dataset into Hive for ETL (Extract, Transfer and Load) operation.
- Experience in analysing data using HQL, Pig Latin and custom MapReduce programs in Python.
- Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and process of data.
- Hands on experience in developing ETL jobs in Hadoop eco-system using Oozie & Stream sets.
- Proficient in usage of tools like Erwin (Data Modeler, Model Mart, navigator), ER Studio, IBM Meta Data Workbench, Oracle data profiling tool, Informatica, Oracle Forms, Reports, SQL*Plus, Toad, Crystal Reports.
- Proficient in using Hive optimization techniques like Buckets, Partitions, etc.
- Very keen in knowing newer techno stack that Google Cloud platform (GCP) adds.
- Extensive experience using MAVEN as a Build Tool for the building of deployable artifacts from source code
- Involved in writing data transformations, data cleansing using PIG operations and good experience in data retrieving and processing using HIVE.
- Experience in creating Spark Streaming jobs to process huge sets of data in real time.

- Experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, SAS and Python and creating dashboards using tools like Tableau.
- Created reports using visualizations such as Bar chart, Clustered Column Chart, Waterfall Chart, Gauge, Pie Chart, Tree map etc. in Power BI.
- Able to use Sqoop to migrate data between RDBMS, NoSQL databases and HDFS.
- Good understanding and hands on experience with MS Azure, AWS S3 and EC2.
- Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
- Experience on Migrating SQL database to Azure Data Lake, Azure data Lake Analytics, Azure SQL Database, Databricks and Azure SQL Data warehouse and controlling, granting database access, and migrating on premise databases to Azure Data Lake store using Azure Data factory.
- Can work parallelly in both GCP and Azure Clouds coherently.
- Proficient and Expert in Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, Spark SQL.
- Proficient in designing and managing data warehousing solutions, ensuring data storage, retrieval, and analysis for business needs.
- Skilled in determining the appropriate level of data aggregation to support reporting and analytics requirements, optimizing performance and efficiency.
- Proficient in using ERwin Data Modeler for data modeling, database design, and documentation, streamlining data-related processes.
- Experienced in database management tasks, including schema design, indexing, and query optimization.
- Strong command of SQL (Structured Query Language), enabling efficient data retrieval, manipulation, and reporting.
- Adept at analyzing data to identify trends, patterns, and insights that support decision-making.
- Experienced in ensuring data quality and integrity through data validation, cleansing, and error handling processes.
- Hands-on experience with Hadoop architecture and various components such as Hadoop File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Hadoop MapReduce programming.
- Comprehensive experience in developing simple to complex Map reduce and Streaming jobs using Scala and Java for data cleansing, filtering and data aggregation. Also possess detailed knowledge of MapReduce framework.
- Used IDEs like Eclipse, IntelliJ IDEA, PyCharm IDE, Notepad ++, and Visual Studio for development.
- Seasoned practice in Machine Learning algorithms and Predictive Modeling such as Linear Regression, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, KNN, Neural Networks, and K-means Clustering.
- Ample knowledge of data architecture including data ingestion pipeline design, Hadoop/Spark architecture, data modeling, data mining, machine learning and advanced data processing.
- Experience working with NoSQL databases like Cassandra and HBase and developed real-time read/write access to very large datasets via HBase.
- Developed Spark Applications that can handle data from various RDBMS (MySQL, Oracle Database) and Streaming sources.
- Proficient SQL experience in querying, data extraction/transformations and developing queries for a wide range of applications.
- Capable of processing large sets (Gigabytes) of structured, semi-structured or unstructured data.
- Experience in analyzing data using HiveQL, Pig, HBase and custom MapReduce programs in Java 8.
- Experience working with GitHub/Git 2.12 source and version control systems.

- Strong in core Java concepts including Object-Oriented Design (OOD) and Java components like Collections Framework, Exception handling, I/O system.

TECHNICAL SKILLS

Google Cloud Platform:	GCP Cloud Storage, Big Query, Composer, Cloud Dataproc, Cloud SQL, Cloud Functions, Cloud Pub/Sub.
Azure Cloud Services (PaaS & IaaS):	Azure Blob Storage, Azure Monitoring, Azure Search, DataFactory, Azure SQL, Azure Analysis Services, Azure Synapse Analytics (DW), Azure Data Lake.
Hadoop/Big Data Technologies:	HDFS, Hive, Pig, Sqoop, Yarn, Spark, Spark SQL, Kafka
Hadoop Distributions:	Horton works and Cloudera Hadoop
SQL/ NO SQL:	MongoDB, HBase, Cassandra
Languages:	C, C++, Python, Scala, UNIX Shell Script, COBOL, SQL and PL/SQL
Tools:	Teradata SQL Assistant, Pycharm, Autosys
Operating Systems:	Linux, Unix, ZOS and Windows
Databases:	Teradata, Oracle 9i/10g, DB2, SQL Server, MySQL 4.x/5.x
ETL Tools:	IBM InfoSphere Information Server V8, V8.5 & V9.1
Reporting:	Tableau

PROFESSIONAL EXPERIENCE

Client: Mayo Clinic Rochester, MN **September 2022 to Present**

GCP Data Engineer

Responsibilities:

- Developed Python scripts, UDF's using both Data frames/SQL and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop.
- Created real-time data streaming pipelines using GCP Pub/Sub and Dataflow for financial data.
- Experience in GCP Dataproc, GCS, Cloud functions, Big Query and moving data between GCP and Azure using Azure Data Factory.
- Collaborated with technical and non-technical stakeholders for 1 year, effectively communicating data strategies and insights.
- Worked with GCP services like Cloud Storage, Compute Engine, App engine, Cloud SQL, Cloud Bigtable and Pub/Sub to process data for the downstream customers.
- Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
- Worked extensively on spark and MLlib to develop a regression model for cancer data.
- Loaded and transformed large sets of structured, semi structured and unstructured data using PIG by importing data using Sqoop to load and export data from My SQL to HDFS and NoSQL Databases on regular basis for designing and developing PIG scripts to process data in a batch to perform trend analysis of data.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query
- Designed a number of partitions and replication factor for Kafka topics based on business requirements and worked on migrating MapReduce programs into Spark transformations using Spark and Scala, initially done using python (PySpark).
- Orchestrated infrastructure as code using Terraform, automating the deployment and scaling of data processing environments for pharmaceutical data analytics.
- Collaborated with cross-functional teams to design and implement secure and compliant cloud infrastructure for sensitive healthcare data in AWS.

- Managed version control for data engineering codebase using GitHub, ensuring traceability and collaboration among team members in the pharmaceutical data analytics domain.
- Implemented branching strategies and pull request workflows to facilitate code reviews and maintain a reliable and efficient development process.
- Implemented Terraform modules for provisioning and managing resources, optimizing infrastructure efficiency for data pipelines and analytics workloads.
- Conducted regular infrastructure audits and updates, ensuring alignment with pharma industry regulations and security standards.
- Designed & developed various departmental reports by using SAS, SQL, PL/SQL, and MS Excel.
- Launched multi-node Kubernetes cluster in Google Kubernetes Engine (GKE) and migrated the
- Dockerized application from AWS to GCP.
- Collaborated with Business Analysts, SMEs across departments to gather business requirements, and identify workable items for further development.
- Partnered with ETL developers to ensure that data is well cleaned and the data warehouse is up-to-date for reporting purpose by Pig.
- Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
- Experience in GCP Dataproc, GCS, Cloud functions, Big Query.
- Experience in moving data between GCP and Azure using Azure Data Factory.
- Ingested huge volume and variety of data from disparate source systems into Azure Data Lake using Azure Data Factory.
- Orchestrated data pipelines using Apache Airflow to interact with services like Azure Databricks, Azure Data Factory, Azure Data Lake, and Azure Synapse Analytics.
- Experience in building power bi reports on Azure Analysis services for better performance.
- Processed some simple statistical analysis of data profiling like cancel rate, var, skew, kurt of trades, and runs of each stock every day group by 1 min, 5 min, and 15 min.
- Used PySpark and Pandas to calculate the moving average and RSI score of the stocks and generated them into data warehouse.
- Designed and implemented data warehousing solutions, aligning data structures with business needs and ensuring efficient data storage and retrieval.
- Determined the appropriate aggregation levels for data in various reporting and analytical contexts, optimizing performance and resource utilization.
- Utilized ERwin Data Modeler for data modeling and database design, creating structured and efficient data storage systems.
- Managed database schemas, indexing, and query optimization, enhancing database performance and responsiveness.
- Executed SQL queries and scripts to extract, transform, and load data, supporting reporting and analytics initiatives.
- Conducted data analysis to uncover insights, trends, and anomalies, contributing to data-driven decision-making.
- Developed and maintained ETL pipelines to extract data from multiple sources and load it into Azure Synapse Analytics, and integrated Azure Synapse with other data processing tools such as Apache Spark and Azure Databricks for real-time data processing and analysis.
- Used cloud shell SDK in GCP to configure the services Data Proc, Storage, Big Query
- Exploring with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, postgreSQL, Data Frame, Open Shift, Talend, pair RDD's
- Migrated previously written cron jobs to airflow/composer in GCP
- Worked on creating POC for utilizing the ML models and Cloud ML for table Quality Analysis for the batch process.

- Involved in integration of Hadoop cluster with spark engine to perform BATCH and GRAPHX operations.
- Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines, such as AWS, GCP.
- Performed data preprocessing and feature engineering for further predictive analytics using Python Pandas.
- Developed and validated machine learning models including Ridge and Lasso regression for predicting total amount of trade.
- Developed and maintained data catalogs using Unity Catalog to organize and manage pharmaceutical datasets, facilitating easy discovery and access for data scientists.
- Collaborated with domain experts to define metadata standards, ensuring comprehensive documentation of pharmaceutical data assets in the catalog.
- Implemented data governance policies within Unity Catalog, promoting data quality and compliance with regulatory requirements in the pharma industry.
- Conducted training sessions for end-users on how to effectively utilize Unity Catalog for data discovery and collaboration.
- Was involved in setting up of Apache airflow service in GCP.
- Develop and deploy the outcome using spark and Scala code in Hadoop cluster running on GCP.
- Boosted the performance of regression models by applying polynomial transformation and feature selection and used those methods to select stocks.
- Build data pipelines in airflow in GCP for ETL related jotis using different airflow operators.
- Experienced in importing data from various sources using StreamSets.
- Created Streamsets pipelines for collecting Logs, Alerts and metrics from customer Vpod s.
- Configured Streamsets to attach the VpodID to each data flowing through and create topics in kafka.
- Generated report on predictive analytics using Python and Tableau including visualizing model performance and prediction results.
- Utilized Agile and Scrum methodology for team and project management.
- Used Git for version control with colleagues.

Environment: Spark (PySpark, SparkSQL, SparkMLIib), Hadoop (Cloudera), GCP, SAS, Python 3.x (Scikit-learn, Numpy, Pandas), Tableau 10.1, GitHub, AWS EMR/EC2/S3/Redshift, Pig, GCP, Stream sets, Google BigQuery, Cloud DataProc, Extract, Transform, Load (ETL), Apache Spark, Azure Synapse Analytics, SQL Database, Azure Data Lake Storage (ADLS), and Azure Data Factory, Azure Data Lake, Azure Data Factory, Teradata, Azure Data Share, MySQL, Jenkins, Git, PySpark with Databricks, Apache Airflow.

Client: Empower Retirement, Greenwood Village, CO January 2020 to August 2022
Data Engineer
Responsibilities:

- Involved in analyzing business requirements and prepared detailed specifications that follow project guidelines required for project development.
- Written Terraform scripts to automate AWS services which include ELB, Cloud Front distribution, RDS, EC2, database security groups, Route 53, VPC, Subnets, Security Groups, and S3 Bucket and converted existing AWS infrastructure to AWS Lambda deployed via Terraform and AWS Cloud Formation.
- Used Pyspark for data frames, ETL, Data Mapping, Transformation and Loading in complex and high-volume environment
- Utilized GitHub for version control and collaboration in a team of power engineers, ensuring coordinated development and deployment of data processing infrastructure and workflows.
- Implemented Git branching strategies to manage concurrent development efforts and streamline the integration of new features and improvements.
- Utilized Unity Catalog to create a centralized repository for power-related datasets, streamlining data discovery and fostering collaboration among power engineers and analysts.

- Defined and enforced metadata standards within Unity Catalog, ensuring consistency and accuracy of power grid data documentation.
- Implemented automated processes to update Unity Catalog entries, reflecting changes in the underlying data sources and maintaining catalog accuracy.
- Employed Terraform to automate the deployment and configuration of cloud resources, enhancing the scalability and reliability of financial data processing systems.
- Collaborated with security teams to implement Terraform best practices, ensuring compliance with financial industry regulations and security standards.
- Designed and maintained Terraform scripts for infrastructure versioning, enabling efficient tracking and management of changes.
- Implemented continuous integration and delivery (CI/CD) pipelines for Terraform code, ensuring rapid and reliable infrastructure updates.
- Process and store parquet files in the Data Lake using GCS in GCP for easy access and analysis.
- Read CSV and JSON files from Google Cloud Storage in GCP to get the information required for the client and partners using lambda functions on an event driven architecture.
- Extensively worked with Avro and Parquet, XML, JSON files and converted the data from either format
- Process the data from Kafka pipelines from topics and show the real time streaming in dashboards
- Worked on AWS Elastic Beanstalk for fast deploying of various applications developed with Java, PHP, Node.js, Python on familiar servers such as Apache.
- Exported the analysed data to the relational databases using Sqoop for visualization and to generate reports for the BI team Using Tableau.
- Developed an equivalent Spark Scala code for existing SAS code to extract summary insights on the hive tables.
- Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python
- ETL process in End-To-End Pipelines using python and GCP.
- Deposit clean parquet files into the Google Cloud Storage in GCP to provide the information for the partner and client.
- Implemented business use case in Hadoop/Hive and visualized in Tableau
- Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
- Designed several DAGs (Directed Acyclic Graph) for automating ETL pipelines
- Performed data extraction, transformation, loading, and integration in data warehouse, operational data stores and master data management
- Strong understanding of AWS components such as EC2 and S3
- Performed Data Migration to GCP
- Responsible for data services and data movement infrastructures
- Proficient in AWS Lambda, Glue, and CloudWatch and Utilized AWS Glue for ETL (Extract, Transform, Load) processes to prepare and move data from various sources into databases.
- Experienced in ETL concepts, building ETL solutions and Data modeling.
- Worked on Ingesting the data using StreamSets from various sources like JDBC to Hive by Sqoop jobs.
- Extensive working experience in ETL tools such as Informatica BDM, Data stage, Stream sets/Oracle Warehouse Builder.
- Designing streaming ETL Data pipelines using Kafka as Staging area using Stream Sets.
- Worked on designing and developing the Real-Time Tax Computation Engine using Oracle, Stream Sets, Kafka, Spark Structured Streaming.
- Developing data pipelines for loading data in to MongoDB, Elastic Search and Influx database through the change data capture using Stream Sets

- Designing and developing the mappings, re-usable components (Address Validation, Parsing, Standardization, etc.) based on the business requirements using Informatica BDM and Stream Sets.
- Worked on architecting the ETL transformation layers and writing spark jobs to do the processing.
- Implemented event-driven architecture with AWS Lambda and Amazon MSK to keep databases synchronized in real time.
- Implemented data quality measures, including data validation checks and cleansing processes, to maintain data accuracy and reliability.
- Collaborated with cross-functional teams to gather data requirements and ensure data solutions aligned with business objectives.
- Developed documentation and data dictionaries using ERwin, facilitating clear communication and understanding of data models and structures.
- Aggregated daily sales team updates to send report to executives and to organize jobs running on Spark clusters
- Loaded application analytics data into data warehouse in regular intervals of time
- Designed & build infrastructure for the Google Cloud environment from scratch
- Experienced in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
- Leveraged cloud and GPU computing technologies for automated machine learning and analytics pipelines, such as AWS, GCP
- Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with Python
- Proficient in Machine Learning techniques (Decision Trees, Linear/Logistic Regressors) and Statistical Modeling
- Compiled data from various sources to perform complex analysis for actionable results
- Measured Efficiency of Hadoop/Hive environment ensuring SLA is met
- Optimized the Tensor flow Model for efficiency
- Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes
- Implemented a Continuous Delivery pipeline with Docker, and Git Hub and AWS
- Built performant, scalable ETL processes to load, cleanse and validate data
- Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies
- Collaborate with team members and stakeholders in design and development of data environment
- Preparing associated documentation for specifications, requirements, and testing

Environment: AWS, GCP, Bigquery, Gcs Bucket, G-Cloud Function, SAS, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Bq Command Line Utilities, Dataproc, Cloud Sql, MySQL, Posgres, Sql Server, Stream sets, Python, Scala, Spark, Hive, Spark –Sql

Client: Ace hardware, Oak Brook, IL **August 2017 to December 2019**

Spark Developer

Responsibilities:

- Created and executed Hadoop Ecosystem installation and document configuration scripts on Google Cloud Platform.
- Transformed batch data from several tables containing tens of thousands of records from SQL Server, MySQL, PostgreSQL, and CSV file datasets into data frames using PySpark.
- Researched and downloaded jars for Spark-Avro programming.

- Developed a PySpark program that writes data frames to HDFS as Avro files.
- Utilized Spark's parallel processing capabilities to ingest data.
- Automated the provisioning of scalable data processing clusters on cloud platforms using Terraform, enhancing the efficiency of power grid data analytics.
- Implemented Unity Catalog as a comprehensive metadata management solution for financial datasets, improving data discoverability and lineage tracking.
- Collaborated with data stewards to establish and enforce metadata governance policies within Unity Catalog, enhancing data quality and compliance.
- Implemented version control best practices on GitHub for financial data engineering codebase, ensuring code integrity, collaboration, and compliance with regulatory requirements.
- Utilized GitHub Actions for continuous integration, automating code validation and ensuring the reliability of financial data processing pipelines.
- Integrated Unity Catalog with data lineage tools to provide a holistic view of financial data flows, supporting regulatory reporting requirements.
- Collaborated with power engineers to design and implement infrastructure solutions that support real-time data processing for monitoring power grid performance.
- Utilized Terraform for infrastructure as code to deploy and manage resources, enabling seamless integration with Databricks and other data processing tools.
- Implemented version control for infrastructure code, ensuring traceability and reproducibility of deployments.
- Created and executed HQL scripts that create external tables in a raw layer database in Hive.
- Developed a Script that copies avro formatted data from HDFS to External tables in raw layer.
- Created PySpark code that uses Spark SQL to generate dataframes from avro formatted raw layer and writes them to data service layer internal tables as orc format.
- Imported required modules such as Keras and NumPy on Spark session, also created directories for data and output.
- Read train and test data into the data directory as well as into Spark variables for easy access and proceeded to train the data based on a sample submission.
- Created SAS Datasets, Data manipulation, developed data marts for the preparation of reports, tables, listings & graphs.
- The images upon being displayed are represented as NumPy arrays, for easier data manipulation all the images are stored as NumPy arrays.
- Created a validation set using Keras2DML in order to test whether the trained model was working as intended or not.
- Defined multiple helper functions that are used while running the neural network in session. Also defined placeholders and number of neurons in each layer.
- Created neural networks computational graph after defining weights and biases.
- Created a TensorFlow session which is used to run the neural network as well as validate the accuracy of the model on the validation set.
- After executing the program and achieving acceptable validation accuracy a submission was created that is stored in the submission directory.
- Executed multiple SparkSQL queries after forming the Database to gather specific data corresponding to an image.
- In charge of PySpark code, creating data frames from tables in data service layer and writing them to a Hive data warehouse.
- Installed Airflow and created a database in PostgreSQL to store metadata from Airflow.
- Configured documents which allow Airflow to communicate to its PostgreSQL database.
- Developed Airflow DAGs in python by importing the Airflow libraries.
- Utilized Airflow to schedule automatically trigger and execute data ingestion pipeline.

Client: Dhruvsoft Services Private Limited, Hyderabad, India October 2015 to May 2017

Role: Data warehouse Developer

Roles & 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.
- Develop, administer, and managed corresponding databases: Consolidated Data Store, Reference Database (Source for the Code/Values of the Legacy Source Systems), and Actuarial Data Mart.
- Wrote Triggers, Stored Procedures, Functions, and 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.
- Worked on 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**.
- Monitor 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.
- Worked on error and event handling techniques, such as precedence constraints, breakpoints, check points, and logging, ensuring reliable and robust **ETL** processes.
- Built cubes and dimensions with different architectures and data sources for business intelligence purposes, including writing MDX scripting.
- Designed **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.
- Developed **SSAS cubes**, implementing aggregations, defining KPIs (Key Performance Indicators), managing measures, partitioning cubes, and creating data mining models. Deploying and processing **SSAS objects**.
- Created adhoc reports and reports with complex formulas, utilizing querying capabilities of the database for business intelligence purposes.
- Developed parameterized, chart, graph, linked, dashboard, scorecards, and drill-down/drill-through reports on **SSAS** cubes using **SSRS** (SQL Server Reporting Services).
- 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, MS SQL Server, Visual Studio, SSIS, Share point, MS Access, Team Foundation server, GIT, Visual Studio 2010, Python, PySpark, Spark, Spark ML Lib, Spark SQL, TensorFlow, NumPy, Keras, PowerBI, Python 3, Django 1.6, Tableau 8.2, Beautiful soup, HTML5, CSS/CSS3, Bootstrap, XML, JSON, JavaScript, JQuery, Angular JS, Backbone JS, Restful Web services, Apache spark, Linux, Git, Amazon s3, Jenkins, MySQL, Mongo DB, T-SQL, Eclipse.

Client: Maisa Solutions Private Limited Hyderabad, India August 2011 to September 2015

Role: Data Analyst

Roles & Responsibilities:

- Worked on dimensional data modelling for Data Mart design, identifying facts and dimensions, and developing fact tables and dimension tables using Slowly Changing Dimensions (SCD) techniques.
- Participated in collaborative sessions with business users and sponsors to conduct requirement gathering, comprehensively documenting business needs.
- Created Power BI reports from inception, encompassing requirements gathering, data modeling, and report generation.
- Enhanced Power BI reports by incorporating new visualizations and integrating additional data sets.
- Designed ETL jobs in Data Stage for extracting, transforming, and loading data from diverse source systems into Data Marts.
- Ensured consistency across Power BI reports throughout the company by revamping the visualization and appearance of existing reports.
- Implemented DAX table functions, including FILTER, ALL, VALUES, DISTINCT, and RELATEDTABLE.
- Developed reports and dashboards within Power BI, employing various visualizations such as Stacked Bar Chart, Clustered Bar Chart, Scatter Chart, Pie Chart, Donut Chart, Line & Clustered Column Chart, Map, Slicer, and Time Brush.
- Automated the schedule refresh of Power BI reports using gateways and utilized DAX functions for creating calculations and measures.
- Stored different versions of Power BI reports on One Drive for collaborative purposes.
- Implemented data source filters to manage large volumes of data from databases.
- Created parameterized, linked, matrix, drill-down, and aggregation reports using SSRS.
- Proficient in creating Power Pivots, Power Views, and SSRS reports using Tabular Model (DAX), Cubes (MDX), and SQL queries.
- Utilized Power BI Data Gateway to ensure the up-to-date status of Dashboards and Reports.
- Developed Hierarchies and KPIs for enhanced data insights and applied Z-order to visualizations in reports.
- Converted SQL query code into DAX code for Power BI reports.
- Installed and configured Enterprise Gateway and Personal Gateway in Power BI Services.
- Conducted tuning of datasets (queries) to optimize refresh times on Power BI web.
- Created Calculated columns and measures using DAX Expressions.
- Implemented data blending, filters, and actions features effectively in Power BI.
- Established Row Level Security in Power BI for enhanced data security.
- Designed and developed T-SQL stored procedures for data extraction, aggregation, transformation, and insertion.
- Utilized a forward engineering approach for designing and creating databases for OLAP models.
- Employed Teradata utilities such as Fast Export and MLOAD for various tasks.
- Developed and scheduled various reports, including cross-tab, parameterized, drill-through, and sub-reports using SSRS.
- Wrote SQL scripts for loading data from staging areas to confidential tables and worked on SQL and SAS script mapping.
- Identified and recorded defects with detailed information for the development team to reproduce the issues.
- Demonstrated flexibility to work extended hours for effective coordination with offshore teams.
- Developed complex stored procedures, efficient triggers, and necessary functions, along with creating indexes and indexed views to optimize performance in SQL Server.
- Monitor 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.

- Worked on error and event handling techniques, such as precedence constraints, breakpoints, check points, and logging, ensuring reliable and robust ETL processes.
- Built cubes and dimensions with different architectures and data sources for business intelligence purposes, including writing MDX scripting.
- Designed 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.
- Developed SSAS cubes, implementing aggregations, defining KPIs (Key Performance Indicators), managing measures, partitioning cubes, and creating data mining models. Deploying and processing SSAS objects.
- Created adhoc reports and reports with complex formulas, utilizing querying capabilities of the database for business intelligence purposes.
- Developed 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, Software Development Life Cycle (SDLC), Database Development, ETL (Extraction, Transformation, Loading), Data Visualization, Data Warehousing, Data Migration, Virtualization, Project Management, Coding Standards, Automation, GIT, Snowflake, AWS S3, GitHub, Service Now, HP Service Manager, EMR, Nebula, Teradata, SQL Server, Apache Spark, Sqoop

Educational Details

University	Year
Bachelors of Computer Science, Thakur College of Engineering, Mumbai University	2004 to 2008