A blue and white logo

Description automatically generated**Manideep Gongalla**

manideepgongi@gmail.com

+1(716)449-8884

**Data Engineer | Azure | Python | Spark | Power BI | Hadoop | ETL Developer |**

**SQL Advanced Analytics Data Engineer MSBI | Snowflake | AWS | GCP**

**ABOUT ME:**

Highly skilled and experienced **Senior** **Azure Data Engineer** with around 10 years of expertise in building data solutions using SQL Server, MSBI, and Azure Cloud technologies. Proficient in **Azure Data Factory, Data Lake Storage, Azure Synapse Analytics, Python, and Spark**. **Extensive experience in** **ETL development, data migration, and data visualization using Power BI, AWS, GCP.** Strong understanding of data warehousing concepts and data modeling. Possess a successful track record in delivering data solutions for various domains, including **E-Commerce, Financials, and Data Analytics.**

**PROFESSIONAL SUMMARY:**

* Proficient in various Azure Cloud Services (PaaS & IaaS) including Storage, **Web Apps, Active Directory, Application Insights, Logic Apps, Data Factory, Service Bus, Traffic Manager, Azure Monitoring, OMS, Key Vault**, Visual Studio Online (VSO), and Cognitive Services (LUIS).
* Designed and developed the **Dynamics-AAA (Access, Authorize & Audit) Portal, enabling secure access to Azure resources with customized roles**. This Portal became a standard for granting access and ensured compliance with MSIT standards.
* Expertise in leveraging **Python data science packages** for data processing, visualizations, and building machine learning models.
* Skilled in utilizing BI tools such as **Power BI and QlikView** to enhance reporting capabilities and develop BI applications according to client requirements.
* Demonstrated ability to **collaborate with key stakeholders**, delivering compelling business value to senior leadership and clients.
* Well-versed in Data warehouse concepts, including **normalization/de-normalization techniques for optimal performance in relational and dimensional database environments**, as well as building Referential Integrity Constraints.
* Strong knowledge in **Data Modeling and Data analysis in MS SQL Server**, with expertise in query optimization.
* Extensive experience in **developing** **complex Stored Procedures, Functions, Triggers, Views, Cursors, Indexes, CTEs, Joins**, and Subqueries with T-SQL.
* Experienced in managing **Azure Data Lakes (ADLS) and Data Lake Analytics**, with a solid understanding of integration with other Azure Services.
* Proficient in migrating existing Active Directory and Exchange users to Microsoft Azure AD using Azure AD Connect, Active Directory Federation Service (ADFS), and DirSync tools.
* Implemented robust security measures for Azure Service Fabric applications, including encryption, authentication, and access control mechanisms, to safeguard sensitive data and ensure compliance with industry regulations such as GDPR, HIPAA, and PCI DSS.
* Experienced in dealing with **Windows Azure IaaS, including Virtual Networks, Virtual Machines, Cloud Services, Resource Groups, Express Route**, VPN, Load Balancing, Application Gateways, Auto-Scaling, and Traffic Manager.
* Experienced in creating Ad-hoc reports and data-driven subscription reports using Report Builder in SSRS.
* Good knowledge of Data Marts, **OLAP, Dimensional Data Modeling, Star Schema Modeling, and Snow-Flake Modeling** for FACT and Dimensions Tables using Analysis Services.
* An expert in Data Warehouse development, from inception to implementation and ongoing support, with a strong understanding of BI application design and development principles.
* Highly experienced in the whole cycle of DTS/SQL Server Integration Services (SSIS) Packages, including developing, deploying, scheduling, troubleshooting, and monitoring for data transfers and ETL purposes across different servers.
* Utilized **PySpark, NumPy, and pandas** to handle extensive datasets efficiently.
* Possess expertise in **Python and Scala**, with a focus on creating user-defined functions (UDF) for Hive and Pig using Python.
* Proficient in utilizing AWS services for data warehousing and analytics, including Amazon Redshift for large-scale data storage and processing.
* **Advanced Excel / Power BI / MS SQL**
* Demonstrated experience in utilizing Excel Pivot and VBA macros to address diverse business scenarios effectively.
* Possess expertise in SQL Server Analysis Services (SSAS) and SQL Server Reporting Services (SSRS).
* Worked extensively on Data Marts, OLAP, and Dimensional **Data Modeling** using Ralph Kimball Methodology, implementing **Star Schema Modeling and Snow-Flake Modeling for FACT and Dimensions Tables** through Analysis Services. Proficient in creating various types of Tabular, Matrix, Drill-Down, Chart reports, Sub reports, and Cross Tab Reports in multiple formats using SSRS.
* Experience in building data pipelines on **AWS** using services like **AWS Glue and AWS Data Pipeline** to extract, transform, and load data from various sources into data warehouses.
* Implemented serverless data processing solutions using **AWS Lambda, AWS Step Functions**, and **AWS Batch** to efficiently handle data transformations and data integration tasks.
* Worked with a wide range of file formats, including CSV, JSON, XML, ORC, Avro, and Parquet.
* Proficient in designing Schemas **(Star and Snowflake)** tailored to fit reporting, querying, and business analysis requirements.
* Utilized Azure Service Fabric's support for stateful and stateless services to build real-time data processing solutions, enabling rapid insights generation and actionable intelligence from streaming data sources.
* Hands-on experience in data warehousing and Business Intelligence Technologies, with expertise in **SQL Server development, Tableau, MSBI stack (TSQL, SSIS, SSAS, SSRS), Azure, Power BI, and Power Apps.** Skilled in building, deploying, and managing applications and services through a global network of Microsoft-managed Data Centers.
* Utilized Python for designing, coding, debugging, and deploying operations, reporting, data analysis, and web applications.
* Experience in managing and monitoring **AWS resources using AWS CloudWatch, AWS CloudTrail, and AWS Config** for tracking and auditing data-related activities and resource utilization.
* Worked with **AWS Glue to create and manage ETL workflows**, automating data preparation and transformation tasks to support data analytics and reporting requirements.
* Developed Python scripts to automate the data sampling process, ensuring data integrity through checks for completeness, duplication, accuracy, and consistency.

**Technical Skills:**

|  |  |
| --- | --- |
| **SDLC** | Agile, Scrum, Waterfall, Kanban |
| **Big Data Ecosystem** | Hadoop, MapReduce, Pig, Hive, HBase, YARN, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, Spark, Ambari, Elastic Search, Parquet, Snappy, Airflow, NiFi. |
| **Hadoop Distributions** | Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapReduce, Apache EMR |
| **Cloud Platforms** | MS Azure , Amazon Web Services (AWS), |
| **MS Azure Services** | Azure SQL Database, Azure Data Lake (ADL), Azure Data Factory (ADF), Azure SQL Data Warehouse, Azure Service Bus, Azure Key Vault, Azure Analysis Service (AAS), Azure Blob  Storage, Azure Search, Azure App Service, and Azure Data Platform Services |
| **ETL/BI Tools** | Informatica Power Center 10.x/9.x/8.x (Source Analyzer, Repository Manager, Transformation Developer, Mapplet Designer, Mapping Designer, Workflow Manager, Workflow Monitor, Warehouse Designer and Informatica Server) Informatica Data Transformation B2B (Parser, Mapper & Serializer), |
| **CI/ CD** | Azure DevOps, Jenkins, Ant, Maven |
| **Ticketing Tools** | JIRA |
| **Operating Systems** | Linux, Windows, Ubuntu, Unix |
| **Databases (RDBMS/ NoSQL)** | Oracle, SQL Server, Cassandra, Teradata, PostgreSQL, HBase, MongoDB, Milus Vector DB |
| **Programming**  **Languages/Scripting** | Scala, SQL, PL/SQL, R, Python (Pandas, NumPy, SciPy, ScikitLearn, Seaborn, Matplotlib, NLTK), Shell Scripting. |
| **DWH Schemas** | Star Schema, Snowflake Schema |
| **Data Modeling Tools** | Erwin, MS VISO |
| **Web/ Application Server** | Apache Tomcat, WebLogic, WebSphere, JBoss |
| **Version Control** | Git, Subversion, GitHub |
| **CI/CD Tools** | Jenkins, Azure DevOps |
| **Reporting / BI Tools** | MS Excel, Tableau, Tableau Server and Reader, Power BI,  QlikView, Crystal Reports, SSRS, Splunk |
| **Machine Learning Techniques** | Principal Component Analysis, Single Value Decomposition, Data Standardization Techniques, L1 and L2 regularization, RMS prop, Hyper parameter tuning, KL Divergence, Resampling Techniques like SMOTE, Cluster Centroid Methods, Ensemble Methods, Feature selection and Feature Engineering, Cross Validation Methods(K-fold), Bleu Score. |

PROFESSIONAL EXPERIENCE

# Client: US Bank – Alpharetta, GA AUG 2021 – Till Date

**Domain: banking**

**Sr Azure Data Engineer**

**Project Description:** The project involves implementing authentication mechanisms using AWS Identity and Access Management (IAM) and AWS Directory Service, and working with various AWS services like EMR, S3, Glue, and Lambda. Data processing pipelines are being developed in AWS Glue using PySpark scripts, and data warehousing solutions are being implemented using Amazon Redshift. Additionally, the project focuses on efficient file storage using formats like Parquet and JSON with compression methods such as GZip and Snappy. Spark is being utilized for data processing, and data modeling concepts like star and snowflake schema are being employed. The codebase is managed using Git repositories, and CI/CD pipelines are set up with AWS CodePipeline for application deployment in the AWS cloud.

**Key Contributions:** Developed a data set process for data mining and data modeling and recommended ways to improve data quality, efficiency, and reliability. Extracted, transformed, and loaded data from various sources to AWS data storage services using a combination of AWS Glue, SQL, PySpark, and AWS Data Pipeline.

**Responsibilities:**

• Involved in data warehouse implementations using Amazon Redshift, RDS, AWS Glue, AWS Data Pipeline.

• Involved in creating specifications for ETL processes, finalized requirements, and prepared specification documents.

• Migrated data from on-premises SQL Server to Amazon Redshift using AWS Data Pipeline, designed optimized database architecture.

• Created AWS Glue jobs for copying data from Amazon S3 to Amazon Redshift

• Implement ad-hoc analysis solutions using AWS Glue, EMR/Databricks

• Work with similar AWS on-prem data platforms, specifically SQL Server and SSIS, SSRS, and SSAS

• Create Reusable AWS Glue pipelines to call REST APIs and consume Kafka Events.

• Used AWS CodePipeline for scheduling AWS Glue jobs and used AWS Lambda for scheduling Databricks jobs.

• Developing and configuring Build and Release (CI/CD) processes using AWS CodePipeline, along with managing application code using AWS CodeCommit with required security standards for .Net and java applications.

• Worked with Amazon Redshift’s stored procedures, used procedures with corresponding DDL statements, used JavaScript API to easily wrap and execute numerous SQL queries.

• Used AWS PaaS Solutions like AWS Elastic Beanstalk, AWS Lambda, AWS RDS, and AWS S3 in configuring and deploying the Operations Management Suite (OMS) to monitor and track changes. Worked on AWS Fabric, Microservices, IoT & Docker containers in and involved in setting up Terraform continuous build integration system. Used AWS Internal Load Balancer to provide high availability for IaaS VMs & PaaS role instances.

• Integrated Kafka with Spark Streaming for real-time data processing

• Implemented large Lambda architectures using AWS Data platform capabilities like AWS Glue, AWS Data CatLog, Amazon Redshift, ML, and Power BI.

• Using AWS Databricks, created Spark clusters and configured high concurrency clusters to speed up the preparation of high-quality data.

• Used AWS Databricks for fast, easy, and collaborative spark-based platform on AWS.

• Implemented Spark using Scala and Spark SQL for faster testing and processing of data.

**Environments:** Python, Hadoop (HDFS, MapReduce), YARN, Spark, Spark Context, Spark-SQL, PySpark, Pair RDD's, Spark Data Frames, Spark YARN, Hive, Pig, HBase, Oozie, Hue, Sqoop, Flume, Oracle, NIFI, AWS, Kafka, Amazon Redshift.

# Client: Comcast – Addison, TX Nov 2018 – Aug 2021

**Title: Azure Data Engineer**

**Domain: Retail**

**Project Description:** The project involved implementing data warehouse solutions using Microsoft

Azure cloud services, such as Azure SQL Data Warehouse, SQL Database, and Azure Data Lake Storage (ADLS). Responsibilities included designing and optimizing ETL processes, creating data pipelines with Azure Data Factory, and implementing ad-hoc analysis solutions with HDInsight/Databricks. The project also involved working with on-premises data platforms like SQL Server, using Databricks for data analysis, and integrating Kafka with Spark Streaming for real-time data processing. Additionally, Azure DevOps was used for CI/CD, and Azure Databricks was utilized for collaborative Spark-based data processing.

**Key Contributions**

* **Involved in data warehouse implementations** using Azure SQL Data warehouse, SQL Database, Azure Data Lake Storage (ADLS), Azure Data Factory v2.
* **Involved in creating specifications for ETL processes,** finalized requirements and prepared specification documents.
* **Migrated data from on-premises SQL Database to Azure Synapse** Analytics using Azure Data Factory, designed optimized database architecture.
* Created Azure Data Factory for copying data from Azure BLOB storage to SQL Server
* Implement ad-hoc analysis solutions using Azure Data Lake Analytics/Store, HDInsight/Databricks
* Work with similar Microsoft on-prem data platforms, specifically SQL Server and SSIS, SSRS, and SSAS
* Create Reusable ADF pipelines to call REST APIs and consume Kafka Events.
* Work on SQL queries – create, modify the queries to obtain an abstract set of data and connect to Tableau and create dashboards.
* Designed and Developed ETL jobs to extract data from different sources and load it in data mart in **Snowflake** and managed Snowflake clusters such as launching the cluster by specifying the nodes and performing the data analysis queries.

**Responsibilities:**

* Involved in data warehouse implementations using Azure SQL Data warehouse, SQL Database, Azure Data Lake Storage (ADLS), Azure Data Factory v2.
* Involved in creating specifications for ETL processes, finalized requirements and prepared specification documents.
* Migrated data from on-premises SQL Database to Azure Synapse Analytics using Azure Data Factory, designed optimized database architecture.
* Created Azure Data Factory for copying data from Azure BLOB storage to SQL Server
* Implement ad-hoc analysis solutions using Azure Data Lake Analytics/Store,
* HDInsight/Databricks
* Work with similar Microsoft on-prem data platforms, specifically SQL Server and SSIS, SSRS, and SSAS
* Create Reusable ADF pipelines to call REST APIs and consume Kafka Events.
* Used Control-M for scheduling DataStage jobs and used Logic Apps for scheduling ADF pipelines.
* Developing and configuring Build and Release (CI/CD) processes using Azure DevOps, along with managing application code using Azure GIT with required security standards for .Net and java applications.
* Worked with Snowflake’s stored procedures, used procedures with corresponding DDL statements, used **JavaScript API** to easily wrap and execute numerous SQL queries.
* Used Azure PaaS Solutions like Azure Web Apps, Web Roles, Worker Roles, SQL Azure and Azure Storage and in configuring and deploying the Operations Management Suite (OMS) to monitor and track changes. Worked on Azure Fabric, Microservices, IoT & Docker containers in and involved in setting up Terraform continuous build integration system. Used Azure Internal Load Balancer to provide high availability for IaaS VMs & PaaS role instances.
* Integrated **Kafka** with **Spark Streaming** for real time data processing
* Implemented large Lambda architectures using **Azure Data** platform capabilities like Azure Data Lake, Azure Data Factory, Azure Data CatLog, HDInsight, Azure SQL Server, ML and Power BI.
* Using **Azure Databricks**, created Spark clusters and configured high concurrency clusters to speed up the preparation of high-quality data.
* Used **Azure Databricks** for fast, easy, and collaborative spark-based platform on Azure.
* Implemented Spark using Scala and **Spark SQL** for faster testing and processing of data.

**Environments:** Python, Hadoop (HDFS, MapReduce), YARN, Spark, Spark Context, Spark-SQL, Pyspark, Pair RDD's, Spark Data Frames, Spark YARN, Hive, Pig, HBase, Oozie, Hue, Sqoop, Flume, Oracle, NIFI, Kafka, Erwin9.8, BigData3.0, Hadoop3.0, Oracle12c, Pig0.17, Sqoop1.4, Azure, Snowflake.

# Client: Lowe’s - Sutherland Jun 2017 –Oct 2018

**Title: Data Engineer**

**Domain: Data Analytics**

**Project Description:** The project was intended to build a data engineering and analytics solution using Python, Spark, Snowflake, and Google Cloud Platform (GCP). I managed data cleaning, preprocessing, and real-time processing using Spark Streaming and Kafka on GCP. We used SSIS packages for data movement between Snowflake and other environments, optimizing performance and queries through Snowflake's features and GCP's BigQuery.

Our data insights came to life with Python, Tableau, Snowflake, and GCP's visualization tools. Automation was achieved via Python/R scripts, with additional development of analytical models and algorithms, tapping into GCP's machine learning capabilities.

Crucially, we seamlessly integrated Snowflake with GCP services like Hadoop, Hive, Pig, and more, ensuring smooth data exchange throughout the ecosystem.

**Responsibilities:**

* Worked on Data Cleaning, Data preparation, Data reduction, features scaling, features engineering using native libraries in Python 2.x/3.,x.
* Implemented Spark using Python and Spark SQL for faster testing and processing of data. Loaded processed data into Snowflake for advanced data warehousing and analytics capabilities.
* Worked on real-time data processing using Spark Streaming and Kafka using Python. Integrated real-time data into Snowflake to enable real-time analytics and insights.
* Involved in the complete SSIS life cycle in creating SSIS packages, building, deploying, and executing the packages in both environments (Development and Production) with Snowflake as the data source and destination.
* Involved in ETL architecture enhancements to increase performance using Snowflake's query optimizer and best practices for data loading and transformation.
* Enhanced and wrote triggers, existing stored procedures, views, functions, database objects using Transact-SQL (TSQL) in Snowflake to meet additional requirements of the latest releases and improve data processing efficiency.
* Working with Dynamic Management Views, System Views in Snowflake for Indexing, and other Performance Problems to optimize queries and improve overall system performance.
* Integrated Milvus Vector DB with Azure services.
* Enhanced data retrieval efficiency through smart indexing.
* Process automation using Python/R scripts with Snowflake to generate and write the results in the production environment every week, enabling stakeholders to access up-to-date information easily.
* Used Python 3.X (NumPy, SciPy, pandas, Scikit-learn, seaborn), Spark 2.0 (PySpark, MLLib), and Snowflake to develop a variety of models and algorithms for analytic purposes, leveraging Snowflake's data warehousing capabilities for scalable and high-performance analytics.
* Utilized Snowflake's integration with various tools and platforms such as Hadoop (HDFS, MapReduce), YARN, Hive, Pig, HBase, Oozie, Hue, Sqoop, Flume, Oracle, NIFI, Azure, and Kafka to enable seamless data exchange and processing across the entire data ecosystem.

**Environments:** Python, Hadoop (HDFS, MapReduce), YARN, Spark, Spark Context, Spark-SQL, PySpark, Pair RDD's, Spark Data Frames, Spark YARN, Hive, Pig, HBase, Oozie, Hue, Sqoop, Flume, Oracle, NIFI, Azure, Kafka, Snowflake.

# Client: Micro Intellect Pvt Ltd, India Jan 2013 – June 2017

**Title: ETL Developer**

**Domain: Data Analytics**

**Project Description:** Micro Intellect Pvt Ltd is a digital automation technology-consulting firm with deep expertise in Data Analytics, Application Development, Robotic Process Automation, AI, DevOps, and Test Automation Services. It has headquarters in Denver, Colorado, with a wholly owned delivery center in Mumbai, India. The objective of the project is to create a database for a New Program and design the schemas in the data warehouse relational and dimensional, providing end-to-end technology and operational fulfillment to individual consumers, small businesses, middle-market businesses, and large corporations.

**Key Contributions:**

* Developed views and templates using Python and created a user-friendly website interface using Django's view controller and template language.
* Used Python libraries like NumPy/SciPy, Pandas for market analysis and conducted machine learning tasks using Python.
* Involved in building database Model, APIs, and Views utilizing Python to build an interactive webbased solution. Used Django Database APIs to access database objects.
* Worked on converting Hive/SQL queries into Spark transformations using Spark RDDs, Python, and OOP with Python. Developed and executed shell scripts to automate jobs.
* Developed Python OOD code for quality, logging, monitoring, and debugging code optimization. **Responsibilities:**
* Developed ETL framework using Spark and Hive, including daily runs, error handling, and logging.
* Worked extensively on analyzing data cubes and reports used in Pentaho reporting.
* Worked on Mondrian OLAP Database and deployed it in the cloud and exposed as an Azure service.
* Used SMOTE to treat highly imbalanced data before prediction to improve model accuracy for symptom prediction.
* Implemented NLP methods using Python NLTK and SpaCy to process client data like prescriptive data and customer comments data to improve customer satisfaction.
* Utilized data analysis packages (e1071, caTools, scikit-learn) in programming languages (R, Python) as well as big data tools (Spark SQL) for query, extraction, and manipulation of data to validate data quality and prepare data.
* Implemented advanced geographic mapping techniques and used custom images and geocoding to build spatial visualizations of non-geographic data using Java API scripting.
* Created Tableau dashboards with advanced charts, drill downs, stack bars, bar graphs, scattered plots, and geographical maps.
* Scheduled data refreshes (Complete refresh and Incremental refresh) on extracted source data in Tableau.

**Environments:** Azure, AWS, Tableau, Oracle, SQL Server, Toad, Informatica Power Designer.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EDUCATION** |  |  |  |  |

**BTECH –**Osmania university-Computer Science. April 2013.