**Karthik PE**    
Contact : **978-483-9246  
Mail :** [**kaarthikp823@gmail.com**](mailto:kaarthikp823@gmail.com)

**PROFESSIONAL SUMMARY**

* Over 9 years of experience in data engineering, **ETL, Azure Data Factory, Informatica, SQL, Azure cloud, python, UNIX, data bricks, Spark** and Hadoop development substantially experienced in designing and executing solutions for complex business problems involving large data warehousing, real time analytics solutions. My core competency is problem solving with a knack of picking up new technologies and techniques quickly.
* Experience in building and architecting multiple Data pipelines, end to end **ETL and ELT** process for Data ingestion and transformation in **Azure data factory, Azure synapse analytics, Azure data bricks** and **Informatica power center.**
* Hands on experience with building stored procedures in **T-SQL** as per business requirements.
* Hands on experience with automating stored procedures by using ADF pipeline by using time based and event-based triggers.
* Used ETL (**SSIS**) to develop jobs for extracting, cleaning, transforming and loading data into data warehouse.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD and PySpark concepts
* Designed and orchestrated multiple **Azure data factory** pipelines to ingest data from multiple sources like SAP ECC, SAP BW, SAP HANA, Oracle DB, Sales force to **Azure data warehouse** and **Azure data lake gne2 (ADLS Gen2)**.
* I have expertise in writing **Spark RDD** transformations, actions, Data Frames, Persistence (Caching), Accumulators, Broadcast Variables, Optimizing Broadcasts, case classes for the required input data and performed the data transformation using **Spark-core**.
* Experience in translating business requirements into actionable insights through data analysis and visualization using **Tableau**.
* Implemented advanced visualization techniques, such as storytelling and guided analytics , to facilitate better data comprehension.
* Experience working with **Milvus**, an open source **vector database**, designed for efficient storage and retrieval of high dimensional vectors.
* Experience in converting processed data into dashboards by using **PowerBI**.
* Hands on experience in converting multiple complex **Informatica power center** workflows into **Azure data factory** pipelines and automated them by using **Azure LogicApps** and **Azure service bus**.
* Hands on experience in writing UNIX shell scripts& **PMCMD** commands for FTP of files from remote server.
* Experience working with **Azure active directory (AD)** and **Azure key vault** services to protect client-side server credentials in secret key.
* Hands on experience on Unified Data Analytics with Databricks, Databricks Workspace User Interface, Managing Databricks Notebooks, **Delta Lake** with Python, **Delta Lake with Spark SQL**.
* Involved in implementing Delta lake architecture to provide reliable,scalable, and performant data storage and processing capabilities.
* Developed **Python scripts** to manage **AWS resources** from **API** calls using BOTO3 SDK and worked with **AWS CLI**.
* Experience in **AWS S3 buckets** creation, **policies** and IAM role-based policies and configuring buckets with permissions and logging.
* Experience in resolving ongoing maintenance issues and bug fixes, monitoring Informatica sessions as well as performance tuning mappings and sessions.
* Experience in Dimensional modeling using star and snowflake schema, identifying Facts and Dimensions, physical and logical modeling.
* Various stages of the software development life cycle were built using the Agile Scrum methodology.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Data Warehousing: | Azure Synapse Analytics (DW), Snowflake, Amazon Redshift,BigQuery, Oracle,  Teradata. |
| ETL Tools: | Azure Data Factory,SQL Server Integration Services(SSIS), Informatica power center ,IBM DataStage |
| Databases: | SQL Server,MySQL, PostgreSQL, Oracle, DB2,Cassandra |
| Languages: | Python, Java, Scala, C, C#, R, JavaScript, PHP |
| Visualization Tools: | Power BI, Tableau |
| Data Modeling: | ER diagrams, Dimensional data modeling, Star and Snowflake Schema |
| Scripting: | Shell scripting (Linux/Unix),Python scripting |
| Version Control: | Git, GitHub,Azure DevOps and TFS. |
| Other Tools and Technologies | Microsoft Visual Studio, Jupyter Notebook,WinSCP, Filezilla, SQL developer ,Toad,Anaconda, PyCharm, Apache  Airflow, Docker, Kubernetes |

**Certificates:** Microsoft azure **AZ-900.**

**EXPERIENCE**

**Senior Data Engineer**

**AmerisourceBergen May 2019-present**

**Projects:**

**Enterprise analytics platform enhancements : August2022-present**

* Designed and orchestrated Extract, Transform, and Load (ETL) pipelines to migrate data from multiple types of data sources (SQL Server, SAP ECC, SAP HANA, SAP BW, DB2, Oracle, Excel, Flat Files, SharePoint, Salesforce, Azure event hub etc.) to destination database server by using **Azure Data Factory** (ADF)and **Informatica Power center, Azure data bricks and SSIS**.
* Built Real-Time Event based Streaming Service to ingest Data into Snowflake using Java, Python, and PySpark
* Involved in migrating SQL database to **Azure synapse analytics**, **Azure data Lake**, **Dedicated SQL pool, Data Bricks** and **Azure SQL Data warehouse** and controlling and granting database accessandMigrating On premise databases to **Azure Data Lake store** using Azure Data factory.
* Designed **robust views, complex queries, functions,** **stored procedures**, **clustered and non-clustered indexes and triggers**. Ensure consistency and integration with existing data warehouse structure.
* Developed **Azure** **LogicApps** and **Azure** **Servicebus** to set up event-based triggers and hands on experience in creating time-based triggers in ADF.
* Built Data pipeline to extract data from Snowflake using **Databricks PySpark**, AWS Lambdas.
* Developed **Informatica Power Center** (10.5.1) , Informatica IICS workflows using task developer, worklets designer, and workflow designer in workflow manager and monitored the results using workflow monitor.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/to Extract, Transform, and load data from different sources like **Azure SQL, Blob storage, Azure SQL Data warehouse, Oracle, salesforce, Amazon S3, Azure Data Lake Storage, FTP** and **SFTP**.
* Involved in set up and deploy **Milvus on Azure virtual machines**, utilizing cloud infrastructure to optimize performance and scalability.
* Developed Spark applications using Hive-metastore, **Pyspark**, **Spark-SQL and RDD’s** for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Good understanding of **Spark Architecture** including Spark Core, Spark SQL, Data Frames, **Spark structured Streaming, Kafka**, Driver Node, Worker Node, Stages, Executors and Tasks.
* Hands on experience in using **Azure databricks** framework , **Jupyter notebooks** to write **Pyspark,SparkSQL** and automate them by using Azure Data Factory.
* Configured and managed **Auto loader** systems to automate data ingestion from various sources,including files,databases,APIs, and streaming platforms.
* Involved in designing data lake solutions using **Delta lake** on Azure , leveraging Azure Data Lake Storage Gen2 as the underlying storage layer.
* Developed **Databricks** and **PySpark** jobs to join Cassandra tables, ETL workflows to load Data into various Targets databases
* Utilized **Apache Spark** with **Delta Lake** to perform data transformations and merge operations effectively.
* Worked on the **Teradata** stored procedures and functions to confirm the data and have load it on the table.
* Worked on optimizing and tuning the **Teradata** views and SQL to improve the performance of batch and response time of data for users.
* Created Build and Release for multiple projects (modules) in production environment using Visual Studio Team Services (VSTS) and **Azure DevOps** **Git hub repository** for implementing CI/CD.
* Involved in complete Software Development Life Cycle (SDLC) process by analyzing business requirements and understanding the functional workflow of information from source systems to destination systems.
* Wrote **Unix shell scripts** &**PMCMD** commands for **FTP** of files from remote server and backup of repository.
* Worked with cross functional teams in an **Agile/Scrum** environment to ensure a quality product is delivered.

Tools: Azure data factory(ADF),Azure Databricks, azure synapse analytics, azure data lake gen2,SSIS,UNIX,WinSCP, Informatica power center(10.5.1,10.4.0),azure DevOps,

Pyspark, azure logic apps, azure service bus, azure active directory, azure key vault,azure data bricks, Pyspark,SparkSQL.

**Data engineer March2021-August2022**

**Supply chain analytics :**

* Designed and orchestrated end-to-end data integration solutions, integrating data from diverse sources including ERP systems,IoT devices and external APIs, using **Azure Data factory**.
* Involved in designing and implemented data architecture for the supply chain using Azure services like **Azure Data Lake Storage**, **Azure Synapse Analytics**, and **Azure Data factory**.
* Involved in Developing data models to support various supply chain processes,such as inventory management ,demand forecasting.
* Assisted data architects to build data warehousing solutions using **Azure Synapse Analytics** to enable **real-time analytics** and reporting for supply chain stakeholders.
* Developed Carbon Black Endpoint Data normalizer to transform the ingested data using Golang, Protobuf, AWS SDK for Golang, AWS S3, Kafka, AWS AthenaAWS EMR, AWS Athena, PySpark, and Python to query the normalized data.
* Implemented data security solutions with industry regulations by implementing Azure data security features like **Azure Key Vault** and **Azure Active Directory**.
* Managed and optimized **Azure SQL Databases** to ensure efficient data storage, retrieval, and indexing.
* Performed query performance tuning and indexing strategies to improve database performance.
* Developed a custom Milvus client application that leveraged **Milvus** API/SDK for vector retrieval and similarity search functionalities.
* Developed Microsoft Defender for Endpoint Data normalizer to transform the ingested data using Goland, Protobuf, Azure Event Hub, Azure SDK for Golang, AWS S3, Kafka, AWS EMR, AWS Athena, PySpark, and Python to query the normalized data.
* Implemented **continuous Integration and continuous deployment(CI/CD)** pipelines for Azure data solutions using **Azure DevOps**.
* Implemented monitoring and altering mechanisms using **Azure Monitor** and **Log Analytics** to proactively identify and resolve issues in data pipelines.
* Worked with the team to deliver components using **agile** software development principles.

Tools: Azure data factory,Pyspark ,T-SQL,Azure synapse analytics, Azure data lake gen2(ADLS Gen2), azure data bricks,pyspark,SparkSQL.

**Informatica Developer May2019-March2021**

**Business Technology Transfer:**

* Designed and orchestrated **Informatica Power Center** (10.5.1) workflows using task developer, worklets designer, and workflow designer in workflow manager and monitored the results using workflow monitor.
* Led Data engineer’s team to migrate data from SAP\_ECC,SAP\_BW,Oracle and on prem data to **Azure synapse analytics DW** by using **Informatica power center**.
* Responsible for **Extract , transform and load** (ETL) pipeline design in Informatica and orchestrated to help team to complete project in time.
* Took whole responsibility for Deploying Informatica objects **mappings,sessions,worklets** and **workflows** from dev to stage and production environment and monitor jobs by using **workflow monitor** as part of post production validation.
* Implemented **partitioning** ,**pushdown optimization** ,and **session optimization** techniques to enhance ETL processing speed.
* Developed  **UNIX** **shell script** to support file transfers from one path to another path (**SFTP**)
* Effectively worked in Informatica version based environment and used deployment groups to migrate the objects.
* Performed data manipulations using various Informatica transformations like **Filter,Expression,Lookup (connected and unconnected), Aggregate ,update strategy, Normalizer,joiner, router , sorter and union**.
* Used **slowly changing dimensions (SCD)** and incremental loading techniques for efficient data warehousing.
* Documented **mappings ,transformations**, and **workflow** to facilitate knowledge sharing and collaboration.
* Worked in Agile methodology to track daily and weekly progress to give updates to stakeholders.

Tools: Informatic power center, UNIX, Azure DevOps, Azure synapse data warehouse, T-SQL,Oracle DB,WinSCP.

**Data Engineer,**

**Lab Corp, Worcester, MA.**  **August 2016 - May2019**

* Designed and orchestrated Informatica **Power Center** (10.4.0) workflows using task developer, worklets designer, and workflow designer in workflow manager and monitored the results using workflow monitor.
* Led Data engineer’s team to migrate data from SAP\_ECC,SAP\_BW and on prem data to Azure synapse analytics DW by using Informatica power center.
* Used **Azure Data Factory** to extract data from different source systems and load it into **Azure Data Lake**, **Azure Storage**, **Azure SQL**, and **Azure DW**.
* Developed ETL pipelines using **T-SQL** and **U-SQL** in **Azure Data Lake Analytics** for transforming and processing the data.
* Implemented Apache Airflow for authoring , scheduling and monitoring data pipelines using DAGs.
* Optimized **PySpark** performance on **Azure Databricks**.
* Created **linked servers, integration runtimes, datasets, and triggers** in **Azure Data Factory** to automate and schedule data pipelines, ensuring timely and accurate data loading.
* Design and create hive **Meta store tables, partition, archive**, and **purge tables** for efficient load strategies.
* Design and architecture ADLS data in **Azure Synapse** to expose as external tables and internal tables.
* Configured **SQL Server jobs** using **SQL Server Agent** to optimize data loading by scheduling jobs during off-peak hours.
* Involved in developing and running Map-Reduce jobs on YARN and Hadoop clusters to produce daily and monthly reports as per user’s need.
* Improved query performance by analyzing **execution plans** and **statistics**, identifying and implementing optimization techniques to enhance data retrieval efficiency.
* Collaborated with the IT team to support ETL processes and used **Postgre SQL** to transfer historical data to a data lake
* Implemented various tasks and transformations in **SSIS packages**, ensuring data quality through data cleansing and performance tuning.
* Built **tabular models** on **Azure Analysis Services** to fulfill business reporting needs
* Perform analyses on data quality and apply business rules in all layers of data extraction transformation and loading process.
* Implemented **logic apps** and email notifications to keep users informed about data load successes and failures.

**Technologies:** Microsoft Azure Cloud (Azure Data Factory, Azure Data Lake, Azure Storage, Azure SQL, Azure DW, Azure Databricks), SQL Server (SQL Server Agent, T-SQL, SSIS), Python (Scikit-learn, Keras, TensorFlow, Scikit-Image, OpenCV), Postgre SQL, ML model training environment

**SQL Developer, Value labs, India****January 2014 – August 2015**

* Preformed database back-ups and proactive maintenance. Responsible for ensuring space availability, monitoring activity, disaster recovery, and documenting problems, changes, and solutions.
* Created and managed database objects ,including tables ,views, indexes, and constraints. Using SQL Server Management studio (SSMS).
* Designed and implemented database schemas, tables, and indexes for new applications, ensuring efficient data storage and retrieval.
* Developed and maintained SQL Server Integration services for designing and deploying data integration packages and workflows.
* Utilized joins and sub-Queries to simplify complex queries involving multiple tables while optimizing procedures and triggers to be used in production.
* Assisted in UAT Testing and provided necessary reports to the business users.

**EDUCATION**

Bachelor of Technology in Electronics and Communication Engineering; JNTUH, India

Master of science in Computer information sciences, New England college, Henniker, NH.