`

YAMINI

817-630-4415 | Irving, TX |yamini.chiguru13@gmail.com| LinkedIn

PROFESSIONAL SUMMARY

With 5 years of experience in Data Engineering, I possess extensive expertise in designing data systems, developing ETL pipelines, and modeling data. My strong proficiency in SQL, MySQL, Oracle, and SQL Server is enhanced by my deep understanding of AWS and Azure technologies. I am skilled in using AWS tools such as Redshift for data warehousing, Lambda for serverless computing, and S3 for effective data storage. My background also includes applying machine learning algorithms to drive business insights. I am adept at problem-solving, communicating effectively, and maintaining data integrity.

EDUCATION

The University of Texas at Arlington **August 2023**

Master of Science in Data Science.

Coursework: Data Science, Probability and Statistics, Machine Learning , Data Science Project Management, Foundation of Computing, Operational Research, Database Systems, Web Data Management, Data Mining.

GITAM University, India **July 2018**

Bachelor of Technology in Electronics and communications Engineering.

CERTIFICATIONS:

- AWS Certified Solutions Architect – Associate

TECHNICAL SKILLS:

Programming Language	Python, Pyspark, Spark, Scala, Shell script, SQL, PL/SQL, APIs
Databases	Snowflake(cloud), Teradata, Oracle, SQL Server, MySQL, NoSQL, MongoDB, PostgreSQL
Tools	SQL Server Management Studio, Informatica, Oracle, Query Analyzer, Query Optimizer, SQL Profiler, Performance Monitor, PyCharm, Eclipse, Visual Studio, SQL*Plus, SQL Developer, TOAD, SQL Navigator, Query Analyzer, SQL Assistance, Eclipse, Postman, Airflow, Kafka
AWS Services	Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation.
Microsoft Azure Services	Azure Virtual Machines, Azure SQL Database, Azure Cosmos DB, Azure Data Lake Storage, Azure Functions, Azure Databricks, Azure Stream Analytics, Azure Data Factory, Azure Logic Apps, Azure DevOps, Azure Active Directory, Azure ExpressRoute, Azure synapse Analytics
Operating Systems	Windows Server 2000/2003/2008/2012, MS DOS, UNIX, Linux, Mac OS-X, CentOS
Database Modeling	Dimension Modelling, ER Modelling, Star Schema Modelling, Snowflake Modelling
Visualization/Reporting	Tableau, ggplot2, matplotlib, SSIS, SSRS and Power BI
Version Control	Git, GitHub, SVN, CVS
Machine Learning & Analytics Tools	Supervised Learning (Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Classification), Unsupervised Learning (Clustering, KNN, Factor Analysis, PCA), Natural Language Processing, Google Analytics Fiddler

PROFESSIONAL EXPERIENCE:

Client name : **MEDISTREAM** **May 2023 - Present**
Role : Data Engineer

- Collected data using **Spark** Streaming from **AWS S3** bucket in near-real-time and performs necessary Transformations and Aggregation on the fly to build the common learner data model and persists the data in **HDFS**.
- Subscribe to **Kafka** topics with a **Kafka** consumer client to process events in real-time using **Spark**. Design a Kafka producer client using **Confluent Kafka** to produce events into **Kafka topics**. Develop **Spark Streaming** jobs to consume data from **Kafka topics** from different source systems and push the data into **HDFS** locations.
- Led the migration of a multi-terabyte data warehouse from **Redshift** to **Snowflake**, achieving a seamless transition with zero downtime and enhancing query performance by 40%.
- Integrated Snowflake with various data sources like **S3**, **Kafka** using **Snowpipe** and external stages, enabling real-time data analytics and reporting capabilities.
- Write **SQL** scripts for data mismatches and load historical data from **Teradata SQL** to **Snowflake**.
- Create **Snowpipe** for continuous data loading from staged data residing on cloud gateway servers.
- Use **AWS Lambda** Functions and **AWS Glue** with **Python** to create on-demand tables on **S3** files.
- Utilize **SparkSQL** for loading **JSON** data, creating Schema **RDDs**, and loading them into **Hive Tables**.
- Convert Hive/SQL queries into Spark transformations using **Spark RDDs** and **Pyspark.**
- Develop Spark code using **Scala** and **Spark-SQL**/Streaming for efficient data processing.
- Utilized **SparkSQL** to ingest and process **JSON** data into Hive Tables using **Schema RDDs**.

- Transformed Hive/SQL queries into Spark transformations with **RDDs** and **Pyspark** for efficiency.
- Developed efficient data processing applications with **Scala**, **Spark-SQL**, and **Spark** Streaming.
- Apply principles of functional programming with **Python** to process complex structured datasets.
- Filter data using **Scala** code and **SQL** queries. Join various tables in **Cassandra** using **Spark** and **Scala** for analytics.
- Utilize **Apache Spark** with **Python** for Big Data Analytics and Machine Learning applications.
- Design and develop real-time stream processing applications using **Spark**, **Kafka**, **Scala**, and Hive to perform streaming **ETL** and apply **Machine learning**.
- Use **Airflow** for scheduling Hive, Spark, and MapReduce jobs.
- Implement installation and configuration of a multi-node cluster on Cloud using **AWS EC2**.
- Effectively communicated best practices and optimization strategies to team members and stakeholders.

Environment: Spark, Scala, MapReduce, Snowflake, Hive, Python, AWS, EC2, S3, Lambda, Cloud Watch, Flat files, MS SQL Server database, XML files, JSON, Cassandra, Kafka, Airflow

Client name : **OMEGA ENGINEERING** **Oct 2020 – Dec 2021**
Role : Data Engineer
Company name : Cigniti Technologies Limited, Hyderabad, India (*IT services*)

Roles & Responsibilities:

- Engaged in the full **Software Development Life Cycle**, including analysis, design, and development, while collaborating with the team using **Agile methodologies**.
- Utilized **Pyspark** for data processing to manage data from various relational databases and streaming sources, integrating with **Redshift** for warehousing.
- Developed a comprehensive data pipeline and performed analytics across the **AWS** ecosystem, including **EMR, EC2**, **S3, RDS, Lambda, Glue, SQS,** and **Redshift**.
- Established end-to-end data quality controls and monitoring systems for **ETL** pipelines to reduce data processing delays.
- Independently managed the **Amazon Redshift** cluster and **AWS** architecture, automating data pipelines and supporting data lakes.
- Constructed data pipelines that extracted data from diverse sources, transformed it according to business needs, and loaded it into **Redshift** for analysis.
- Leveraged **AWS Kinesis** for real-time data capture from various sources, processing terabytes of data.
- Implemented a **CI/CD** framework using **Git** and **Jenkins** to configure and manage big data solutions on the **AWS cloud** platform.
- Accountable for writing unit tests and deploying production-level code using **Git** for version control.

Environment: AWS cloud services, XML files, JSON, flat files, snowflake, Python, SQL, Git, Agile methodology

Client name : **ATOM BANK** **Dec 2018 – Oct 2020**
Role : Associate Engineer
Company name : Cigniti Technologies Limited, Hyderabad, India (*IT services*)

Roles & Responsibilities:

- Involved in Requirement gathering, Design and Development, testing and implementation of business rules.
- Understand business use cases, integration business, write business & technical requirements documents, logic diagrams, Pro charts, and other application related documents.
- Used **Pandas** in **Python** for Data Cleansing and validating the source data.
- Designed and developed **ETL** pipeline in **Azure cloud** which gets customer data from **API** and process it to **Azure SQL DB**.
- Orchestrated all **Data pipelines** using **Azure Data Factory** and built a custom alerts platform for monitoring.
- Created custom alerts queries in **Log Analytics** and used **Web hook** actions to automate custom alerts.
- Created **Databricks** Job workflows that extracts data from **SQL server** and upload the files to sftp using **Pyspark** and **python**.
- Implemented secure secret management using **Azure Key Vault**, integrating with **Azure Data Factory** and **Databricks** for enhanced **data security** and streamlined operations.
- Built a common sftp download or upload framework using **Azure Data Factory**. Maintain and support **Teradata** architectural environment for **EDW** Applications.
- Involved in **logical modeling**, physical database design, data sourcing and data transformation, data loading, **SQL**, and performance tuning.
- Develop conceptual solutions & create proof-of-concepts to demonstrate viability of solutions.
- Technically guide projects through to completion within target timeframes.(Mention GIT, Agile)
- Collaborate with application architects and **DevOps**.
- Design Setup maintain Administrator the **Azure SQL Database, Azure Analysis Service, Azure SQL Data warehouse, Azure Data Factory, Azure SQL Data warehouse.**
- Build Complex distributed systems involving huge amount data handling, collecting metrics building **data pipeline**, and Analytics.

Environment: Azure cloud services, XML files, Snowflake, REST APIs, JSON, flat files, snowflake, Python, SQL, Git, Agile methodology, Unix, Linux

Client name : **BYJUS** **Nov 2017 - Dec 2018**
Role : Data Engineer

Roles & Responsibilities:

- Built and maintained a wide range of **ETL** pipelines to extract, transform, and load data from various sources, including **flat files, databases, and APIs**.
- Developed **ETL** as per mapping requirements to load data into staging from multiple sources such as csv, xml, xlsx.
- Utilized **SQL** and **Python** for data manipulation and analysis, experience with libraries such as **Spark and Pyspark.**
- Wrote Python script for manipulating and looping through different **user defined objects**.
- Normalized existing **OLTP** systems to speed-up the execution time in **Data Modeling** logical statements.
- Developed multitude of **Database objects** such as tables, views, stored procedures, user-defined functions to support various database applications per business requirements.
- Designed & maintained databases using Python, troubleshot, fixed, and deployed many **Python** bug fixes successfully.
- Designed **ETL packages** on **SSIS** to migrate data and make it available for the reporting applications.
- Implemented **Agile** for frequent changes to client requirements and following parallel development and testing.
- Writing SQL Scripts to extract the data from **Database** and for **Testing Purposes**.
- Developed test cases, performed **Unit testing** and **Integration testing**.
- Handle the migration process from Development, Test and Production Environments.
- Developed ETL jobs to extract information for **Enterprise Data Warehouse**.
- Extensive use of **ETL** process to load data from different **RDBMS**, **XML**, and **flat files.**
- Experience in Performance tuning and writing Complex queries, stored procedures, Views, Cursors, **SQL** Joins.
- Design and develop **spark job** with **Scala** to implement end to end data pipeline for batch processing.
- Developed a fully automated continuous integration system using **Git, Jenkins, MySQL**, and custom tools developed in **Python** and **Bash**.
- Developed Data mapping data governance, transformation and cleaning rules for the **Master Data Management** architecture involving **OLTP** and **OLAP**.
- Developed complex mappings, Mapplets in **Informatica** workflow designer to integrate data from varied sources like **Teradata, Oracle, SQL Server, Flat files** and loaded into target.
- Worked on performance tuning by identifying the bottlenecks in Sources, Targets, and Mapping. Enhanced Performance on **Informatica** sessions using large data files by using partitions.
- Created **Unix Shell** Scripts for **ETL** jobs, session log cleanup, dynamic parameter and maintained **shell** scripts for data conversion.
- Validate the data in source and target systems using **PL/SQL** queries.
- Created reports and data visualization dashboards using **complex SQL** logic and **BI** tool **Tableau.**
- Leveraged **Informatica** to facilitate data migration and integration processes, achieving high data quality and consistency across multiple platforms and systems.

Environment**:** Spark, Python, SQL, PL/SQL, Unix, Informatica, Snaplogic, OLAP, OLTP, RDBMS, Agile and Toad.

ACADEMIC PROJECTS

DBLP Data Analysis Using Graph Characteristics **August 2022 - December 2022**

- Performed Data cleaning and Preprocessing on the DBLP dataset to prepare it for graph analysis.
- Created a graph representation of the DBLP network, with nodes representing authors, articles, journals, and conferences, and edges representing relationships between them.
- Calculated various graph characteristics, such as centrality, clustering coefficient, and degree distribution, and used them to analyze the structure and behavior of the network.
- For Data visualization, used tools like NetworkX to visualize the graph and to find patterns.

Tech stack: Python(Libraries- Pandas, NumPy, NetworkX, matplotlib)

IMDb Database and Analysis (MySQL) **June 2022 – August 2022**

- Analyzed 10000+ records of IMDB dataset from a relational database which is having different types of data tables like movies, genres, directors, release dates, and ratings.
- Analyzed the IMDb database using SQL queries to identify patterns and trends in the entertainment industry.
- In this project, Disk-based data analysis techniques is used to analyze the IMDb database, which is too large to fit into memory.
- Used Python and its data analysis libraries, such as Pandas, to load the dataset from disk, and use SQL queries to extract and analyze the data from the database.
- Visualizations of the data used tools like Matplotlib, Tableau and Excel.
- Evaluated year-on-year growth of different types of genres and critics along with the director's trend pattern of successes according to their releases by writing SQL queries through 6 tables in relational databases.

Tech stack: SQL, Python, Libraries- Pandas, NumPy, Matplotlib, seaborn, PowerBI, Tableau and Excel