# Mounika Yandapalli

# mounikayandapalli001@gmail.com

# [www.linkedin.com/in/mounika-yandapalli](http://www.linkedin.com/in/mounika-yandapalli)

# Summary

# Over 10+ years of relevant experience in building and consulting solutions on Databricks, DBT, Python, Big Data, cloud-based applications spanning across technologies and business domains. Application Development using Data engineering, DevOps, Big Data Technologies and Cloud (Azure & AWS).

**Work Experience Highlights**

* Lead Data Engineer: Extensively worked as a Lead for a team (12 members onshore & offshore) on designing and developing Data pipelines using Azure Data Factory, Databricks Notebooks using PySpark and DBT.
* Handson in implementing incremental data processing using Delta Live Tables (DLT).
* Experienced in enabling Unity Catalog and Delta sharing in Databricks.
* Implemented Observability in a data engineering system to have an ability to monitor, troubleshoot, and understand the system’s performance and behavior. Visualize anomalies identified in varied data sources using Grafana.
* Handson in designing and developing Azure DevOps pipelines for Databricks. Have Clients recognized record in reducing the turnaround time for DevOps pipelines execution.
* Implemented Disaster recovery management in Databricks.
* Extensive hands-on in designing and developing ETL integration patterns using Python, Pyspark, Airflow, Azure Data Factory, Glue, Lambda, Step Functions, Snowflake.
* Have mentored the team on various new technologies based on “Big Data” such as Pyspark, Databricks, ADF.
* Gained knowledge of Insurance, Banking, Pharma and FMCG domains/business.
* Certified in technologies such as Hadoop, Python and Apache Spark, Lakehouse Fundamentals and Gen AI by Databricks.

# Specialization/Certification

* Databricks Certified Data Engineer Associate
* Databricks Generative AI Fundamentals
* Databricks Certified Associate Developer for Apache Spark 3.0 Python
* Databricks Lakehouse Fundamentals
* MTA: Introduction to Programming Using Python - Certified 2022
* Cloudera Certified Hadoop Developer (CCD-410)
* Cloudera Certified Specialist in Apache HBase (CCB-400)

# Professional Experience Summary

* Lead Data Engineer, Persistent (May 2024 – Till Date)
* Design and develop Data Orchestration pipelines/workflows in DLT, Azure Data Factory, Databricks and DBT models. Enable and setup Delta sharing with several down streams. Manage the entire Databricks account, including users, cloud resources, and account usage. Designed and developed Disaster recovery management in Databricks. Enhanced Azure DevOps pipelines to reduce the turnaround time tremendously from 2 hours minutes.
* Lead Data Engineer, PepsiCo (Oct 2023 – Apr 2024)
* Lead the team (12 members onshore and offshore) to design, develop and deploy scalable and reusable metadata-driven data ingestion frameworks using Azure Data Factory, Azure Databricks, Pyspark, Azure Deltalake and ADLS Gen2. Implementing Data Observability which allows the ability to monitor,
* troubleshoot, and understand system’s performance and behavior. Designing, architecting, and improving our CI/CD pipelines using DevOps
* Senior Staff Data Engineer, Novartis through Altimetric India Pvt. Ltd (June 2022 - Sept 2023)
* Lead the team (7 members) to Design and develop integration patterns to bring data from different sources to form a Unified Data Lake in AWS (S3, Lambda, Step Functions, Glue, Athena) & Azure Cloud that allows Data Analysts and ML team to further analyze data and take business critical decisions. A medallion architecture was designed and developed to transform the data according to business needs and downstream demand. Also, I worked as a Data Governance Lead for a Data Domain in the organization.
* Data Engineer, Bank of America (Apr 2015 - Jun 2022)
* Design big data frameworks/platforms in on-premises clusters. Define patterns and frameworks for migration of on-perm Hadoop applications with data to cloud-based data lake to calculate Credit Risk Capital.

# Education

# Master of Computer Applications, 2010, Sri Venkateswara University, India.

* Bachelor of Science, 2007, Sri Venkateshwara University, India.

**Technical Skills**

|  |  |
| --- | --- |
| Area | Skillset |
| Operating Systems | Linux – Centos, Ubuntu, Windows |
| Cloud Platforms | Azure, AWS, Databricks |
| BigData Technologies | Hadoop Ecosystem, Spark |
| Hadoop Ecosystem | HDFS, MapReduce, HBase, Hive, Pig, Oozie, Sqoop, Impala |
| Hadoop Distributions | Cloudera, Hortonworks |
| Languages | Python, Pandas |
| SQL Databases | PostgreSQL, MySQL, Oracle, Teradata |
| NoSQL Datastores | HBase |
| Amazon Web Services | S3, Glue, RDS-PostgreSQL, IAM, Lambda, Athena, Step Functions |
| Azure Services | Azure Data Factory, Azure Synapse Analytics, ADLS Gen2, Blob, Azure Delta Lake, SQL Database, Azure Log Analytics, Monitor, Azure DevOps |
| Reporting Tools | Power BI |
| IDEs, Frameworks, Tools | Ansible, Airflow, Jenkins |
| Servers | Tornado, Flask |
| ETL services | DBT |

**PROJECT DETAILS**

**Client: Parexel (Persistent) (May 2024 – Till Date)**

**Role: Lead Data Engineer**

**Team size: 7**

**Project Description:**

* Parexel is a clinical research organization (CRO) that designs and conducts clinical trials for biopharmaceutical companies. They offer a range of services from Phase I to IV clinical development, and work with patients, biopharmaceutical leaders, and innovators to increase access to clinical research. The project Enterprise Data Lake (EDL) focus is to use Microsoft Azure data PaaS services to design, build, modify, and support data pipelines leveraging Databricks and PowerBI in a medallion architecture.
* Responsibilities:
* Using Microsoft Azure data PaaS services, design, build, modify, and support data pipelines leveraging Databricks, DLT and Power BI in a medallion architecture setting.
* Responsible for developing Databricks Notebooks and workflows using PySpark and Spark SQL.
* Responsible for developing incremental data loads using DLT & Autoloader.
* Responsible for developing DBT Models for implementing the transformations and orchestrate them using Databricks workflows.
* Accountable for access management in Databricks.
* Designed and developed Azure Data Factory Pipelines to pull data from different data sources to ADLS GEN2 storage.
* Enhanced Azure DevOps pipeline to reduce the build time from 2 hours to 20 minutes.
* Designed and developed Databricks Disaster Recovery management.
* Technologies: Azure Databricks, PySpark, DLT, DBT, Python, Azure Data Factory, Azure Delta Lake, ADLS Gen2, YAML
* Tools: Microsoft Power BI, Azure DevOps

**Client: PepsiCo (Oct 2023 – Apr 2024)**

**Role: Lead Data Engineer**

**Team size: 12**

**Project Description**:

* The Enterprise Data Foundation is the single data platform for PepsiCo that drives digital transformation across sectors and markets and acts to uplift the capability available to all markets and sectors. The Data Foundation is built alongside an operating philosophy that directly supports the Enterprise Data Strategy Enterprise Data Strategy objectives whilst also representing the most modern thinking in how large enterprises should work with data. PepsiCo is familiar with product management approaches across food and beverage portfolios yet applying this to data represents a new way of working.

**Responsibilities:**

* Lead a team responsible to design and develop Azure Data factory pipelines to perform data ingestion from varied data sources to Bronze, Silver & Gold Data Lake layers.
* Design and develop Databricks Notebooks using PySpark and Spark SQL and include them as an activity in Azure Data factory pipelines.
* Responsible for designing and developing incremental pipelines using DLT.
* Responsible for developing and reviewing (developed by team members) the Databricks notebooks to implement transformations in silver layer data assets.
* Leading and coordinating with the technical team to meet project deadlines.
* Deliver a POC to implement Incremental Refresh in Power BI reporting.
* I lead a team and work as an individual contributor on implementing Data Observability which allows the ability to monitor, troubleshoot, and understand system’s performance and behavior.
* Design Grafana Dashboards to visualize the Databricks logs that are stored in the Time series Database - Azure Log Analytics (LGA), Monitor. This identifies Usage tracking, performance metrics, frequent
* Databricks job failures, Resource utilization trend etc.
* Visualize failures and exceptions of Databricks workflows, ADF pipelines in the past 24 hours.
* Technologies: Python, Pyspark, Azure Databricks, DLT, Azure Data Factory, Azure Log Analytics, Azure Delta Lake, ADLS Gen2, DevOps
* Tools: Azure managed Grafana, Microsoft Power BI, Azure Devops

**Client: Novartis through Altimetric India Pvt Ltd (June 2022 - Sept 2023)**

**Role: Senior Data Engineer**

**Team size: 8**

**Project Description:**

* The Formula One platform is a future-ready enterprise big data and analytics platform developed and managed by NBS IT. The Formula One platform will ingest and unify data to accelerate reusability and availability at scale. Through adherence to the FAIR principles for data management and stewardship, the platform will deliver trustworthy data that will realize bold business outcomes. Formula One invests in the latest advanced analytics and machine learning technology. The Formula One platform provides AI models which are re-usable across the enterprise.

**Responsibilities:**

* Responsible for driving a team of 8 members to create and schedule reusable Airflow pipelines to orchestrate Databricks notebooks in AWS.
* Develop and orchestrate ETL pipelines using Lambda, Glue, Step Functions with multiple data sources and having target as Databricks Hive tables.
* Coordinate with multiple stakeholders to gather requirements, design and reviews.
* Involved in Data Extraction, Data Refinement, and Feature transformation of data using Pyspark.
* Building efficient, optimized & re-usable processing frameworks on distributed technologies such as PySpark.
* Implementation of Snowflake in the refinement layer for data storage, processing, and building analytic solutions.
* Environment: Python, Pyspark, Spark SQL, ADLS, Azure Data Factory (ADF), AWS Databricks, S3, EC2, Glue Catalog, Step Functions, Lambda, Athena, IAM, Snowflake, Airflow
* Tools: GIT, Bitbucket, DevOps - Ansible, Jenkins and JIRA.

**Client: BA Continuum India Pvt. Ltd (Apr 2019 - Jun 2022)**

**Role: Senior Data Engineer Team Size: 6**

**Project Description:**

* BA Continuum India Pvt. Ltd. ('BACI/Company') is a non-banking subsidiary of Bank of America Corporation (BOFA). RAP is a common integrated Risk Analytics Platform (RAP) to facilitate End to End model executions utilizing provisioned data sources from GRA and providing advanced analytical and reporting capabilities to downstream users. RAP provides end to end automated workflow orchestration which includes approval, audit capability, entitlements, accountability and attestation. Overlay management provides an ability to overlay model results at various levels of granularity. The purpose of applying overlay is to forecast the loss in various scenarios to account for Lower loss rates.

**Responsibilities:**

* Sole Owner for end-to-end automated workflow orchestration.
* Coordinate with onshore counter parties to gather requirements, design review. Responsible for design, code, test, debug, and document solution.
* Received awards for accountability shown.
* Providing a solution to extend existing system functionality to support multi-tab excel sheet upload from UI.
* Implemented UDF's to read multi-tab excel sheets into Spark data frames.
* Introduced window specification functions in Pyspark to obtain optimal solutions with less execution time.
* Performance tuning in existing overlays management for other models. Responsible for regression testing.
* Creation and load Delta tables in Databricks.
* Support reuse and sharing of platform components and technologies within software engineering teams.
* Deliver software components to enable delivery of platforms, applications and services.
* Gained expertise creating Continuous Integration builds and deployment automation, for example CI/CD Pipelines (DevOps)
* Being the first point of contact in responding to production issues and conducting any technical analysis to arrive at solutions.
* Environment: Pyspark, Spark SQL, Shell scripting, Hive, Impala, Azure Databricks, ADLS
* Tools: GIT, Bitbucket, Ansible and JIRA

**Client: BA Continuum India Pvt. Ltd (Apr 2015 - Apr 2019)**

**Role: Data Engineer / Bigdata Engineer Team size: 8**

**Description:**

* Mortgage Loss -level Model (MLM) is an enhancement of New Mortgage Model (NMM). The prior model's data sources were SQL Server, Vertica, and Files. The requirement of RAP is to source the Model's output data to our Downstream. This enables Downstream users to perform risk analysis with the MLM model's output data.
* Project Responsibilities:
* Created complete solution for Financial Data model pipelines to source AML(Anti Money Laundering) data to downstream, CRP.
* Creation of staging, final and reporting tables in Hive and Impala Tables. Implemented Trigger table logic on Oracle.
* Implementation of append/delete logic of existing data. Making sure the existing model runs without getting disturbed.
* Enhance existing NMM framework to support new MLM model sourcing. Implemented DI/DQ logic
* w.r.t source and target tables in hive.
* Scheduling Autosys jobs by creating JIL scripts and calendar.
* Environment: Sqoop, Oozie, Shell scripting, Hive, Impala
* Tools: Autosys, GIT, Bitbucket, Ansible and JIRA