# Lokesh Mothukuri

**Email Id:** mothukurilokesh6@gmail.com **|Mobile:** (412)-376-7363

**LinkedIn**

# PROFESSIONAL SUMMARY

* **Data Engineer** with **7+ Plus** years of experience in Designing, Developing, Documentation, and integrating applications using **Big Data** platforms like **Cloudera, Hortonworks,** and **Map Reduce,** using cloud technologies Mainly **AWS.**
* Strong experience in **Data Analyst, Data mining** with large data sets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, Statistical modeling, Data modeling, Data Visualization, Web Crawling, Web Scraping.
* Adept in statistical programming languages like **R** and **Python, SAS, Apache Spark, Matlab** including Big Data technologies like **HDFS**, **Map Reduce**, **Hive**, **HBase**, **Pig**, **Spark**, **Spark Streaming**, Impala, Oozie, Flume, Catalog, **Kafka**, and **Sqoop.**
* Experience with all **SDLC** and **Agile** Development model stages right from requirement gathering to deployment and production support
* Experience in configuring the **Zookeeper** to coordinate the servers in clusters and to maintain data consistency
* Experience on **Apache Oozie** for scheduling and managing the Hadoop Jobs
* Experience with extracting Real-time feed using **Kafka** and **Spark Streaming**, converting it to RDD, processing data in the form of Data Frame, and saving the data as Parquet format in **HDFS**
* Experience in working with **MapReduce** programs using **Apache Hadoop** for working with Big Data.
* Extensively worked with **Teradata** utilities Fast export, and Multi Load to export and load data to/from different source systems including flat files
* Experience with visualizing the data using BI and services amp tools such as **Power BI, Tableau and Matplotlib.**
* Experience working on various file-formats including delimited text files, clickstream log files, **Apache log files**, **Parquet files**, Avro files, JSON files, XML files, and others.
* Experience in data mart life cycle development, performed **ETL** procedure to load data from various sources into Data marts and Data warehouses using **Informatica Power Center.**
* Expertise in NOSQL databases and its integration with Hadoop cluster to store and retrieve huge amount of data.
* Experience in writing **SQL queries** and **optimizing** the databases like **MySQL**, **SQL Server**, and **Oracle.**
* Hands-on experience using **Airflow DAG** for Managing and scheduling Jobs on a Hadoop cluster.
* Good understanding of data modeling (Dimensional & Relational) concepts like **Star-Schema Modeling, Schema Modeling, Fact and Dimension tables**.
* Strong knowledge of NoSQL Databases like **MongoDB**, **HBase**, **Oracle**, **Cosmos DB**, and **Dynamo DB**
* Experience in **UNIX** Shell Scripting and Deployment of Applications in Server.
* Involved in **Agile** methodologies, daily Scrum meetings, Sprint Planning, and sprint retrospective meetings
* Experience in creating Reports and Dashboards using **Power BI** on the requirement of the team
* Experienced in AWS Cloud Computing services, such as **EC2, S3, Lambda, API, DynamoDB, EBS, VPC, ELB, Route53, Cloud Watch, Security Groups, Cloud Trail, IAM, Cloud Front, EMR, RDS** and **Glacier** also worked on **DNS, SSL** and **Firewalls.**

# Technical Skills:

|  |  |
| --- | --- |
| **Azure services** | Azure Data Factory, Databricks, Azure Active Directory, Blob Storage, Data Lake, SQL Database, SQL Data Warehouse |
| **AWS services** | EC2, S3, EMR, Glue, RDS, Elasticsearch, SQS, EBS, Lambda, Athena, Kinesis, ECS, DynamoDB,Quick Sight, Redshift |
| **GCP services** | Big query, Datapost, pub/subs, datasets, Data proc, Airflow, Kubernetes, cloud functions, g shell |
| **Big Data Ecosystem** | HDFS, Hive, Yarn, Spark, MapReduce, HBase, Spark, Zookeeper, Airflow, Stream Sets, Oozie,Sqoop, Flume, Pig, Snowflake |
| **Databases** | MySQL, Oracle, Teradata, MS SQL, DynamoDB |
| **ETL/BI Tools** | Talend Open Studio, Informatica, Tableau, Power BI |
| **Hadoop Distribution** | Horton Works, Cloudera and AWS EMR |

|  |  |
| --- | --- |
| **Scripting Languages** | Python, PySpark, Scala, SQL, PowerShell Scripting |
| **Operating systems** | Linux (Ubuntu, Centos, RedHat), Windows (XP/7/8/10) |
| **Languages** | T-SQL, C#, Visual Basic, Shell Script, Python, C, Java, JavaScript, Node JS, Angular, J2EE, ASP.Net, J Query |
| **Methodologies** | SDLC, Agile, Waterfall |
| **IDE and Build Tools** | Jupiter Notebook, Terraform, Anaconda, Jira |

**Work Experience:**

**Vizio, Dallas, Texas**

**Senior Data Engineer/Python Developer**

**Jan 2024 - Present**

I work on building and scaling data pipelines that power Smart TV analytics and personalized ad experiences for millions of users. I use PySpark, Kafka, and Airflow to process more than 3TB of data daily, combining real-time and batch data sources across 100M+ devices.

* Implementation of end-to-end data solution on Azure using Azure Databricks, ADF, DW and PowerBI.
* Designed a robust data modelling environment using Databricks on Azure, enabling consumers to easily operate highly descriptive Notebooks in a fully governed environment.
* Engineered scalable, end-to-end data pipelines using Apache Airflow, Python, and SQL to support analytics for millions of smart TV users.
* Built real-time data ingestion systems using Kafka and Spark Streaming to process streaming events from Vizio devices with sub-minute latency.
* Collaborated cross-functionally with Product, Analytics, and Data Science teams to deliver high-impact datasets for personalization and ad-tech initiatives.
* Implemented robust data quality frameworks and anomaly detection mechanisms, improving data reliability and trust across teams.
* Migration large data sets to Databricks (Spark), create and administer cluster, load data, configure data pipelines, loading data from ADLS Gen2 to Databricks using ADF pipelines
* Extensive hands-on experience of writing notebooks in data bricks using python/Spark SQL for complex data aggregations, transformations, schema operations. Good familiarity with Databricks delta and data frames concepts.
* Extensive hands-on experience in designing and implementing scalable ETL pipelines to process variety of data types (structured, unstructured), file formats (Json, csv, text delimited…)
* Collaborate with analysts to prepare complex data sets that can be used to solve data needs
* Built pipelines to take data from various telemetry streams and data sources to craft a unified data model for analytics and reporting.
* Creating Temporary views and loading curating data in destination tables.
* Databricks job configuration, Refactoring of ETL Databricks notebooks
* Created various pipelines to load the data from Azure data lake into Staging SQLDB and followed by to Azure SQL DB.
* Knowledge in retrieve, analyze and present data using Azure Data Explorer/Kusto
* Experience in Lift and shift existing SSIS packages using ADF.
* Created pipelines to load data from Lake to Databricks and Databricks to Azure SQL DB.
* Created Databricks Notebooks to streamline data and curate data for various business use cases.
* Designed various azure data factory pipelines to pull data various data sources and load the data into Azure SQL database.
* Triggers have been created for the pipelines to run on a day-to-day basis.
* Various tabular model cubes are created on top of the azure SQL database which are consumed for various reporting needs.
* Migrated the data from existing on prem database into Azure SQL.
* Worked on Huge Data Transfers from & to SQL Server Databases using utilities / tools like DTS, SSIS, BULK Inserted and used configuration files and variables for production deployment.
* Created SSIS packages to transform source data into dimension and fact tables.
* Developed complex Stored Procedures and views, ingested them in to SSIS packages. implemented slowly changing dimensions while transforming data in SSIS.
* Designed and loaded data into tabular model cubes to connect power bi reports.
* Created pipelines using ADF to run SQL scripts.
* Created database tables and stored procedures as required for reporting and ETL needs.
* Created/modified the existing SQL views as needed.
* Created Power BI reports and dashboards as per the business requirement using different data sources.
* Extensively used DAX to create complex calculated measures and columns in power bi and cubes.
* Implemented row level security within power bi to create user specific view of the data within report based on the role.

**Environment:** Azure Data Factory, Azure Data Lake Storage (ADLS), Azure Databricks, Power BI, PySpark, Python, Spark SQL, PowerShell scripting, ETL, GIT, Kanban, Jira, data airflow, Apache airflow, MongoDB, Kafka, Informatica, Blob, Azure SQL DB, MS SQL, Cosmos DB, XML.

**IPX Retirement |New York, NY Sr. AWS Data Engineer**

**Aug 2022 – Jan 2024**

I’ve been designing and building robust data pipelines to manage large volumes of retirement plan data across multiple plan types, including 403(b), 457, and IRA plans. I work with Python, SQL, and PySpark to automate the ingestion and transformation of financial data, processing over 2 TB of data daily from multiple sources like payroll systems and custodians.

* Designed and deployed scalable, cloud-native data pipelines on **AWS (Glue, Lambda, S3, Redshift)** to ingest and transform high-volume retirement investment data with 99.9% reliability.
* Reduced ETL processing time by through pipeline optimization and migration from on-prem systems to AWS Glue- based workflows.
* Engineered a centralized **data lake** using **S3**, integrating structured and semi-structured financial data from over 10 internal and third-party sources.
* Developed reusable ETL templates using **AWS Glue + PySpark**, accelerating onboarding of new financial products and improving team efficiency.
* Implemented robust **data validation and monitoring** mechanisms using **CloudWatch**, **SNS**, and **Lambda**, reducing downstream data issues.
* Partnered with compliance and finance teams to deliver automated **regulatory reporting** pipelines, ensuring timely delivery of IRS and SEC filings.
* Integrated **AWS Athena** and **Redshift Spectrum** for ad-hoc querying and analytics, enabling near real-time access to historical and transactional retirement data.
* Leveraged **Terraform** and **CloudFormation** for IaC, ensuring consistent, repeatable infrastructure deployments across dev and prod environments.
* Built and maintained **data cataloging and lineage tracking** using **AWS Glue Data Catalog** and **Apache Atlas**, improving data discoverability and governance.
* Collaborated with data analysts and product owners to define **star/snowflake schemas**, accelerating dashboard development in **Tableau** and **QuickSight**.

**Environment:** AWS, Python, Django, Restful API, SSIS, ETL, Spark, Kafka, CI/CD, AWS Lambda, AWS Glue, EC2, S3, IAM, CLI, EKS, CloudWatch, DynamoDB, JIRA, Git, GitHub, PySide, PyMongo, PySpark, Pickle, Panda’s data frame, MySQL, Docker, Jenkins, Ansible, VM Ware ESX, Docker, Kubernetes, SQL, Snowflake, Tableau, SVN, Informatica, Teradata, ETL, Agile, RHEL, Linux

**RMSI Pvt Ltd |Hyderabad, India Azure Data Engineer**

**Jan 2020 – Dec 2021**

I built data pipelines with **Azure Data Factory** and developed real-time dashboards using **Power BI** to support transportation projects. I automated workflows with **Azure Functions** and ensured data security with **Key Vault**, helping improve decision-making and operational efficiency in transportation planning.

* Designed and managed data pipelines to handle geospatial data for projects in transportation, city planning, and environmental analysis.
* Built ETL workflows using Azure Data Factory to bring in data from satellite imagery, IoT sensors, and third-party APIs.
* Stored and organized large volumes of spatial data using Azure Data Lake Gen2 and Azure Blob Storage for easy access and cost efficiency.
* Worked with Azure SQL Database and Cosmos DB to support fast querying of both structured and unstructured geospatial data.
* Collaborated with GIS experts and data scientists to build real-time dashboards using Azure Synapse Analytics and Power BI that helped clients make informed decisions.
* Automated routine checks and validation steps using Azure Functions and Logic Apps, improving data quality and reducing manual tasks.
* Ensured data security and compliance by setting up monitoring tools like Azure Monitor, managing secrets with Key Vault, and enforcing access controls with RBAC.
* Contributed to a cloud-based data environment that supported both real-time and batch data processing at scale.
* Regularly tuned pipelines and storage for better performance, cost savings, and reliability.
* Played a key role in cross-functional teams, making sure our data systems aligned with both technical needs and client goals.

**Environment:** Azure Data Factory, Azure Data Lake Gen2, Azure Blob Storage, Azure SQL Database, Cosmos DB, Azure Synapse Analytics, Power BI, Azure Functions, Logic Apps, Azure Monitor, Azure Key Vault, Event Hubs, Stream Analytics, Azure DevOps, Python, SQL, REST APIs, GeoJSON, Shapefiles, Satellite and IoT Data

**Wipro Hyderabad, India Data Engineer**

**Aug 2017 – Nov 2019**

When I joined Wipro in 2018, I had the opportunity to work with a top-tier banking client to build and optimize their data pipelines. This involved automating the processing of financial data, ensuring compliance, and supporting business-critical reporting systems.

* Engineered and deployed secure, scalable ETL pipelines using AWS Glue, Lambda, and Step Functions, streamlining ingestion and transformation of banking data from legacy systems to modern cloud architecture.
* Designed data lakes on Amazon S3 for structured and unstructured banking data, enabling seamless integration with downstream analytics tools and regulatory reporting platforms.
* Developed and optimized Redshift data warehouses to support advanced financial analytics, customer segmentation, and credit risk scoring models.
* Worked with banking data models (e.g., customer onboarding, transactions, loans, credit, KYC) to deliver end-to- end data solutions aligned with compliance and business rules.
* Integrated real-time transaction data using Kinesis Data Streams and Kafka, enabling fraud detection systems and customer behavior analytics in near real-time.
* Automated data workflows and orchestrations using Apache Airflow and AWS Step Functions, reducing data processing time.
* Created and maintained Terraform scripts for infrastructure-as-code deployment of AWS services, improving infrastructure consistency and version control.
* Implemented data quality checks, lineage tracking, and governance using AWS Glue Data Catalog, ensuring audit- readiness and compliance with banking regulations (e.g., RBI, GDPR).
* Partnered with cross-functional teams including business analysts, QA, and DevOps to deliver robust data solutions supporting banking operations, risk analysis, and reporting.
* Built data visualization dashboards using Amazon QuickSight and Tableau, enabling stakeholders to track key KPIs like customer churn, NPA trends, and transaction volumes.

**Environment:** AWS, Python, Django, Restful API, SSIS, ETL, Spark, Kafka, CI/CD, AWS Lambda, AWS Glue, EC2, S3, IAM, CLI, EKS, CloudWatch, DynamoDB, JIRA, Git, GitHub, PySide, PyMongo, PySpark, Pickle, Panda’s data frame, MySQL, Docker, Jenkins, Ansible, VM Ware ESX, Docker, Kubernetes, SQL, Snowflake, Tableau, SVN, Informatica, Teradata, ETL, Agile, RHEL, Linux

I worked closely with business analysts and other teams to define requirements, ensuring our solutions aligned with regulatory needs. I ensured data security by applying encryption and masking techniques, ensuring compliance with industry standards. Throughout the project, I thrived in an Agile environment, contributing to high-quality, scalable solutions that met the bank’s evolving business needs.

# Education & Certifications:

Master of Science in Information Systems – La Roche University, Pissburgh

* Snowflake Certified snow pro core
* AWS Certified Developer-Professional