Khasim Mohamed

Email:Khasimmohamed1976@gmail.com

Indian Institute of Technology, Kharagpur, India. **Mobile: +1 703 340 9757**

Professional Summary:

- Overall, 12 plus years of IT experience with Data Models, full-stack applications, extensively engaged with Data Migrations, Validated Data Security Solutions. For the last twelve years, heavily engaged with AWS Services, Snowflake and DBT platforms.
- Engaged with domains of Hospitality, Insurance, Healthcare, Life Sciences, Power, Manufacturing, and Automotive Industries.
- Gap analysis of AS-IS and TO-BE Data Models and present refined data models to Enterprise Applications Steering Committee.
- Worked with Gen AI frameworks (OpenAI, Hugging Face, or LangChain) for data-driven automation and insights generation. Implemented AI-powered data pipelines to optimize data quality, enrichment, and predictive analytics.
- Championed best practices in data security, privacy, data architecture, warehousing, and governance.
- Worked on Data Replication Solutions to meet business needs of application and data availability. Processed the data depending on downstream application needs. Built Batch Pipelines and Streaming Pipeline. Deployed different ETL tools.
- Rich experience in Amazon Web Services like S3, IAM, EC2, EMR, Kinesis, VPC, Dynamo DB, RedShift, Amazon RDS, Lambda, Athena, Glue, DMS, Quick Sight, Amazon Elastic Load Balancing, Auto Scaling, CloudWatch, SNS, SQS and other services of the AWS family. Hands on experience in Data Analytics Services such as Athena, Glue, Data Catalog & Quick Sight.
- Hands on expertise with AWS Databases such as RDS(Aurora), Redshift, DynamoDB and Elastic Cache (Memcached & Redis).
- Wrote AWS Lambda functions in python for AWS's Lambda which invokes python scripts to perform various transformations.
- Experience in building and optimizing AWS data pipelines, architectures, and data sets.
- Expertised in using Cloud based managed services for data warehousing in Confidential Azure (Azure Data Lake Storage, Azure Data Factory). MDM Tools were analyzed by comparing Master Data Management features required for Data Quality, Data Integrity, and Data Governance. Worked with Collibra Data Intelligence Platform.
- Integrated and utilized performance monitoring tools to track and troubleshoot bottlenecks in data processes.
- Hands on experience on tools like Hive for data analysis and Sqoop for data ingestion and Oozie for scheduling.
- Experience in scheduling and configuring the oozie and also having good experience in writing Oozie workflow and coordinators.
- Worked on different file formats like JSON, XML, CSV, ORC, Paraquet. Experience in processing both structured and semi structured Data with the given file formats.
- Worked on Apache Spark transformations on RDDs and Data Frames using spark SQL and Spark streaming contexts.
- Extensively worked with SQL queries with Snowflake SnowSQL and Snowpipe tools
- Created internal and external stages and transformed data using DBT. Implemented SCD type1 and type2 Loads using DBT.
- Consulted on Snowflake Data Platform Solution Architecture, Design, Development, and deployment focused to bring the data driven culture across the enterprises. Implemented Change Data Capture technology in Snowflake using stream.
- Executed multiple end to end Enterprise data warehousing projects. Designed and developed Data Pipelines for loading data and building transformations. Ensured Data Quality and Data Integrity across source and destination datamarts. Enhanced performance of ETL Scripts. Accountable for Enhancements, Upgrades, Migration, Maintenance and Production support projects.
- Seasoned professional to enforce Agile and DevOps Processes. Collaborated with cross-functional Agile-Scrum teams.
- Worked on Informatica IPASS solutions to automate ETL Transformations using Web Transformations (REST APIV2), Python, and SQL Transformations.
- Developed Python scripts, using Data frames/SQL and RDD/Map Reduce in Spark for Data Aggregation, queries.

Academic Qualifications:

- Master of Technology in Mechanical Engineering from Indian Institute of Technology, Kharagpur India
- Bachelor of Technology in Mechanical Engineering from Acharya Nagarjuna University, India.

Technical Expertise:

Database	AWS RDS, Redshift, Snowflake, Oracle, MS SQL Server, Postgres SQL.
Cloud Services	AWS, Azure, and Snowflake
Storage	AWS, Azure Data Factory, Snowflake, Hive, and Hadoop
Data Modelling	Dimension Modeling, ER Modeling, Star Schema Modeling, Snowflake Modeling
ETL Tools	**DBT, Informatica** IICS, and DataStage
Query Tools	Spark SQL, SnowSQL, and SQL.
Scripting	Python and JavaScript, JSON, CURL
Languages	SQL, Python, PL/SQL, Shell Scripting, Java,
Platforms	Unix, Linux, GitHub

CUSTOMERS:

Lifesciences – Insmed, Merk & Medtronic

Healthcare – Centro Hospitalar CHTMAD, Portugal, Medical Card Systems Holdings, General Electric Healthcare and Johnson & Johnson

Power – Eaton, General Electric Power & Transportation Systems.
Retail – Target
Manufacturing – Daimler Truck North America, Johnson Automotive, General Electric Healthcare.

KEY PROJECTS:

#1. Sales Data Analysis for a Hospitality Industry, Chain of Steakhouses at Northern Virginia. Client: Bourbon Blvd, VA;

Duration: 7 Months; June 2024 to Present; Role: Lead Data Engineer, Work Location: Northern Virginia, VA.

- Created Kinesis Data streams, Kinesis Data Firehose and Kinesis Data Analytics to capture and process the streaming data and then output into S3, and Redshift for storage and analyzation.
- Used AWS services including S3, EC2, AWS Glue, Athena, RedShift, EMR, SNS, SQS, Kinesis. Defined custom IAM policies.
- Deployed various function calls from Boto3, OS, Sys, and Requests modules to enhance custom logic inside AWS Lambda Hander function. Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs using CloudWatch.
- Design and Develop ETL Processes in AWS Glue to migrate data from external sources like S3, Parquet/Text Files into AWS Redshift. Validated Queries using AWS Athena against AWS Glue Database Tables.
- Change data capture (CDC) with AWS Glue has been implemented. Written PySpark job in AWS Glue to merge data from multiple sources using AWS Glue Data Catalog metadata table definitions.
- Designed and setup Data Lake to provide support for various uses cases for voluminous, rapidly changing data by using multiple AWS Services.

#2. Supply Chain Upgrade for a Biopharmaceutical Company in New Jersey, USA. Client: Insmed Inc, New Jersey

Duration: Seventeen months; August 2022 to May 2024, Role: Lead Data Engineer, Work Location: Austin Texas.

- Worked on setting up local development environment to develop Data Engineering Applications using Snowflake.
- Used Snowflake Snowsite interface to manage files, jobs, clusters. Extensively worked with Snowflake Jobs and Clusters.
- Deployed and Run Data Engineering Jobs on Snowflake Job Clusters as Applications using Notebooks.
- Recreated existing application logic and functionality in the Azure Data Lake, Data Factory, Data Bricks, SQL Database and SQL data warehouse environment.
- Developed Python scripts using Spark Data frames/SnowSQL and RDD/Map Reduce function calls for Data Aggregation, queries.
- Implemented Change Data Capture (CDC) in Snowflake. Addressed complex issue using Snowflake optimization features.
- Implemented Data Cloning, Data Masking, and Row-Level Security features as per business processes and data security.

#3. Amazon Supply Chain Implementation for Automotive Giant in North America: Client: Daimler Truck North America, Portland, Oregon;

Duration: Eighteen months; February 2021 to July 2022; Role: Lead Data Engineer

- Worked with Snowflake Database Objects: Warehouses, Roles, Databases, Schemas, Constraints, Clustered keys, Material views.
- Developed User Defined functions and Sequences. Worked with Standard SQL Window functions.
- Engaged with Data Sharing and Data Exchange functionalities with Snowflake Data Market place.
- AWS Glue was deployed for ETL Scripts design and development for transformations. Loaded and Queried JSON data.
- Worked with Micro partitions and Clustering keys. Re-engineered Data Models using sophisticated features from Table designs offered by Snowflake and AWS Redshift service. For BI/Analytics, AWS IoT Analytics and Quicksight have been considered.
- CDC is performed through log-based CDC or trigger-based CDC. Extensively worked with JSON, AVRO, and PARQET data through data pipelines. Validated Integration Approaches/ Solutions at high level and determined appropriate one.
- Involved in purchasing options and cost control activities for the Snowflake platform services.
- Developed Snowpipe scripts, using Data frames/SQL and RDD/Map Reduce for Data Aggregation, queries.

#4. Farm management data platform: Client: Chetu Software Solutions, Tempe, AZ;

Duration : 6 months; August 2020 to January 2021 Role: Lead Data Engineer

- Used local development environment for Snowflake platform.
- Setup Single node and multi node clusters.

- Installed and configured Snowflake platform tools.
- Integrated Snowflake cluster with AWS Glue Catalog.
- Mounting into Snowflake clusters
- Installed and Configured Snowflake CLI.
- Implemented CDC solutions with the AWS Glue Service.

#5. MercyOne Dubuque Medical Center, Iowa:

Client: MercyOne Dubuque Medical Center, Iowa;
Duration: 10 months; October 2019 to July 2020; Role: Lead Data Engineer

- Amazon API Gateway is used to create a web service to support XDS transactions such as ITI-41 (HL7 V3 standard).
- Amazon HealthLake FHIR-based APIs are deployed.
- Worked with Enterprise Architects to set up AWS Infrastructure (S3, EC2, VPC Network, and Databases)
- Processed sensitive patient data using Amazon EMR clusters.
- Engaged in Data Modelling for scalable applications.
- Engaged in customization of data exchange data pipelines.
- Utilized AWS tools for Data Flow automation and performance enhancement.
- Architected pipelines using snowflake & AWS cloud. Automated pipelines within AWS & Snowflake ecosystems.
- Designed solutions for AWS IAM Policy Permission sets for restricting unwanted access to the data.
- Developed Python scripts, using Data frames/SQL and RDD/Map Reduce in Spark for Data Aggregation, queries.
- Integrated real time streaming data and data orchestration with Airflow and Snowflake.

#6. For Merck Corporation, pharmaceutical company New Jersey, USA;

Duration: 28 Months; June 2017 to September 2019; Role: Lead Data Engineer.

- Loaded Data Incrementally to Target Table on Databricks. Handled S3 Events using AWS Services on Databricks.
- Incremental Load using cloudFiles File Notifications on Databricks.
- Worked with AWS Services for cloudFiles Event Notifications on Databricks.
- Used Databricks SQL Editor to develop scripts or queries. Reviewed Metadata of Tables using Databricks SQL Platform
- Configured Databricks CLI to push data into the Databricks Platform. Analyzed JSON Data using Spark APIs
- Analyzed Delta Table Schemas using Spark APIsExtensively worked with AWS services such as CloudFormation, S3, Glue, EMR/Spark, RDS, DynamoDB, Lambda, Step Functions, IAM, KMS.
- Designed and implemented ETL Transformations using Python Collections, Panda and PySpark Data Frames APIs.
- Planned and implemented infrastructure, frameworks, and platforms for data ingestion, aggregation, integration, and analysis

#7. Implementation of Amazon Redshift Data Wearhouse solutions (Healthcare):

Client: Centro Hospitalar de Trás-os-Montes e Alto Douro (CHTMAD); Duration: 12 Months; June 2016 to May 2017; Role: Lead Data Engineer

- Securely stored, transformed, transacted, and analyzed health data using Amazon Healthlake.
- Queried HealthLake FHIR data using SQL through AWS Athena query engine and create metric tracking care gap dashboards.
- Deployed AWS EMR Service for Big Data requirements.
- Engaged in ETL Development and Data Warehouse Development
- Leveraged the elasticity of the cloud with Databrics.
- Spark - Distributed Computing with Databrics is adopted in multiple projects.
- Built capabilities such as inserting, updating, and deleting the data from files in Data Lake.
- CloudFiles - Get the files in an incremental fashion in the most efficient way.
- Databricks SQL interface is used for running queries submitted for reporting and visualization.
- Worked on Seting up local development environment to develop Data Engineering Applications using Databricks.
- Used Databricks CLI to manage files, jobs, clusters.
- Developed complex SQL scripts to transform and load data (as per business needs) and addressed performance issues.

#8 Data Replication Project: (Engman)
Client: MCS Healthcare Holdings, LLC Puerto Rico
Duration : 39 months; March 2013 to May 2016; Role: Lead Data Engineer

- Stored patient medical history from multiple data sources in the normalized common data model (FHIR-based) format and leverage FHIR APIs to build transactional applications.
- Leverage HealthLake Patient Access APIs and Bulk FHIR APIs with built-in support for US CORE and CARIN BB profile validation to meet the 21st Century Cures Act for patient access and interoperability requirements.
- Developed Python scripts, using Data frames/SQL and RDD/Map Reduce in Spark for Data Aggregation, queries.
- Architected pipelines using Snowflake & AWS cloud. ETL Scripts design and development, designing technical solution for data pipelines including data ingestion, integration, transformations, and storage.
- Worked with Micro partitions and Clustering keys. Re-engineered Data Models using sophisticated features from Table designs offered by AWS Redshift and Snowflake. Adopted SQL Managed instance for reporting and analytics.

#9. For Eaton Corporation, USA;
Duration : 5 years ; March 2008 to February 2013; Role: Data Engineer.

- Worked with Enterprise Architects for AWS Infrastructure (S3, EC2, VPC, & IAM services) sizing and deployment.
- Proof of Concepts performed to introduce Amazon EMR Service for Big Data Management.
- Developed Python scripts, using Data frames/SQL and RDD/Map Reduce in Spark for Data Aggregation, queries.
- Data Analysis and Data Modelling using ER Diagram and Erwin and Toad tools.
- Heavily used AWS Database services Amazon Relational Database Service (RDS) and DynamoDB.
- Planned and implemented Data Transfer activities from On Premise database Oracle, SQL Server to Cloud Managed Database Instances AWS RDS.