**Sr. Data Engineer**

**Mahesh H**

**(980) 3048265**

**scott@metaitsystems.com**

**PROFESSIONAL SUMMARY:**

* Around 10+ years of expertise as a Data Engineer in the analysis, design, development, and implementation of data, architecture, and frameworks, as well as the ability to operate on both the GCP and Azure cloud platforms concurrently and effectively.
* Diverse experience in all phases of software development life cycle (SDLC) especially in Analysis, Design, Development, Testing and Deploying of applications.
* Hands on experience with GCP stacks including Big Query, Cloud Data Proc, Cloud Storage (GCS), Cloud Pub/Sub, Dataflow, Cloud SQL, IAM, Cloud Functions, Stack driver, Cloud Logging, Data Studio and Composer.
* Proficient in utilizing Azure Data Factory, Azure Databricks, Azure Synapse Analytics, and other Azure services to build robust data pipelines, perform advanced data analytics, and drive data-driven decision-making.
* Proficient in various Big Data Ecosystem, Hadoop ecosystem, Design, Developing, Maintenance of various applications like HDFS, YARN, MapReduce, Strom, Spark, Pig, Sqoop, Hive, Impala, HBase, Kafka, and Crontab tools and have a strong background in data visualization, reporting, and data quality solutions.
* Solid experience and understanding of implementing large scale data warehousing programs and E2E data integration solutions on Snowflake Cloud, Informatica Intelligent Cloud Services (IICS-CDI) & Informatica Power Center Integrated with multiple relational databases (MYSQL, Teradata, Oracle, SQL Server DB2).
* Expertise in Database design and development with Business Intelligence using SQL Server 2014/2016, Integration Services (SSIS), DTS Packages, SQL Server Analysis Services (SSAS), DAX, OLAP Cubes, Star Schema and Snowflake Schema.
* Strong proficiency in SQL concepts, Presto SQL, Hive SQL, Python (Pandas, NumPy, SciPy, Matplotlib), Scala, Java, and PySpark to cope with the increasing volumes of data.
* Implemented new features and optimized existing code for the platform using Kubernetes and Docker.
* Instantiated, created, and maintained CI/CD pipelines and applied automation to environments and applications. Worked on various automation tools like GIT and Terraform.
* Developed mappings to extract data from SQL Server, Oracle, Flat files, and load into DataMart using the Power Center.
* Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing and moving data from various sources using Flume, Kafka, Power BI, and Microsoft SSIS.
* Strong knowledge in data preparation, data modeling and data visualization using Power BI and had experience in developing various reports, dashboards using various visualizations in Tableau.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| **Big Data Technologies** | Apache Hadoop 2.x/1.x, MapReduce, HDFS, Sqoop, Pig,Hive,HBase, Impala, Oozie,Flume, NiFi, Kafka, Zookeeper, Yarn, Apache Spark, Mahout, Sparklib, Cloudera CDP, Hortonworks HDP |
| **Databases** | Oracle,MySQL,SQLServer, MS SQL,PostgreSQL,Teradata, Spark-Redis, DB2. |
| **NoSQL Databases** | HBase, Cassandra, MongoDB,DynamoDB, and Cosmos DB. |
| **GCP Cloud Platform** | Big Query, Cloud Data Proc, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Shell, GSUTIL, BQ Command Line, GCP Cloud Storage, Composer, Cloud SQL, Cloud Pub/Sub, Cloud Data Flow. |
| **Azure Cloud Platform** | Azure HDInsight, Data Bricks, Data Lake, Blob Storage, Data Factory ADF, SQL DB, SQL DWH, Snowflake, Azure AD. |
| **Programming Languages** | C, C++, Java, Python, PySpark, Scala, R, SQL, PL/SQL, Pig Latin, HiveQL, Unix, JSON, Shell Scripting. |
| **Web Technologies** | Java, J2EE, Struts, Spring, JSP, Servlet, JDBC, EJB, JavaScript, XTML, XML |
| **Versioningtools** | SVN,Git,GitHub, Git Bucket |
| **ETL/BI Tools** | Tableau,Data Studio, Matplotlib,SSRS, Power Bl, Informatica, SSIS, SSAS, QlikView, Erwin, and Arcadia |
| **DatabaseModelling** | DimensionModeling, ERModeling,StarSchemaModeling,SnowflakeModeling |
| **Operating Systems** | Windows7/8/XP/2008/2012, UbuntuLinux,MacOS, Unix |
| **Build Tools** | Jenkins, Maven, ANT, ApacheAirflow, Tivoli. |
| **Development Tools** | PyCharm, Microsoft SQL Studio, Eclipse, NetBeans, Toad, IntelliJ Ide. |
| **Methodologies** | System Development Life Cycle (SDLC), Jira, Agile, Confluence, RAD, JAD, UML, and Waterfall Model. |

**WORK EXPERIENCE:**

**Client: Co-op (Retail), Cucamonga, CA May 2022 – Present**

* **Sr. Data Engineer**
* Responsibilities:
* Orchestrated the full Software Development Life Cycle (SDLC) for complex data engineering projects, ensuring high-quality deliverables.
* Emphasize your ability to design and create effective graph data models and schemas.
* Orchestrated end-to-end data mapping and modeling for complex retail data engineering projects, ensuring high-quality deliverables in line with the Software Development Life Cycle (SDLC).
* Contributed to the architecture and optimization of data solutions, emphasizing effective data modeling for complex retail data engineering projects.
* Specify the graph database technologies you have hands-on experience with, such as Neo4j, Amazon Neptune, or JanusGraph.
* Successfully executed data migration projects, transferring data from various sources to PostgreSQL databases.
* Used SSMS performance monitoring tools to analyze and optimize database performance.
* Designed, developed, and maintained REST APIs in Python for seamless data access and integration.
* Architected and optimized data solutions in AWS, utilizing services such as EC2, S3, Lambda, Step Functions, and IAM for secure and scalable data pipelines.
* Implemented data lakes with AWS Lake Formation, enabling efficient data storage and retrieval.
* Managed and optimized relational (MySQL) and NoSQL (DynamoDB) databases, ensuring data availability and performance.
* Proficient in designing and implementing database schemas using PostgreSQL.
* Implemented version control for database scripts and changes using SSMS.
* Spearheaded end-to-end ETL (Extract, Transform, Load) processes using AWS Glue, transforming data from diverse sources into actionable insights.
* Processed and transformed data in JSON and XML formats, leveraging Pandas and NumPy libraries for data manipulation.
* Highlight your proficiency in designing, modeling, and implementing graph databases to address specific business requirements.
* Implemented Snowflake schema for optimized data warehousing and querying.
* Managed and optimized Hadoop and HDFS clusters using PySpark, Hive, Pig, and Spark for large-scale data processing.
* Maintained comprehensive documentation for database configurations, schemas, and processes using SSMS.
* Implemented version control and code repositories using Git, ensuring collaborative development practices.
* Established CI/CD pipelines with Jenkins, automating testing, integration, and deployment.
* Provisioned and managed infrastructure as code (IaC) using Terraform for AWS resources.
* Containerized applications and services using Docker and managed orchestration with Amazon Kubernetes Services (EKS).
* Created data visualizations and dashboards in Tableau, providing actionable insights to stakeholders.
* Implemented log analysis and monitoring solutions with Splunk, ensuring the health and performance of data pipelines.
* Implemented version control for database scripts and changes using SSMS.
* Collaborated within Agile and Scrum methodologies, participating in sprint planning and retrospectives.
* Facilitated comprehensive insurance product modeling initiatives, contributing to the full Software Development Life Cycle (SDLC) for complex data engineering projects in the retail sector.
* Collaborated in the design, development, and maintenance of REST APIs in Python to enhance data access and integration for insurance-related products.
* Proven ability to identify and resolve performance bottlenecks in PostgreSQL databases.
* Describe how you have integrated graph databases into existing data architectures.
* Tracked project tasks and progress using Jira, ensuring alignment with project goals and timelines.
* Environment:Python, REST API, AWS, EC2, S3, Lambda, Step Functions, IAM, AWS Lake Formation, MySQL, DynamoDB, ETL, AWS Glue, JSON, XML, Pandas, NumPy, Snowflake, Hadoop, HDFS, PySpark, Hive, Pig, Spark, Git, Jenkins, Terraform, Docker, EKS, Tableau, Splunk, Agile, Scrum, Jira.
* Client: Next Health technologies, Denver, CO Feb 2019 to April 2022 Role: Cloud Data Engineer

**Responsibilities:**

* Migrated peta byte-scale data warehouse from Azure Synapse to Big Query to improve query performance and lower costs. Redesigned schema and ETL process.
* Created scalable and cost-effective data storage solutions on GCP, utilizing Cloud Storage and Big Query depending on the data requirements.
* Migrated Azure Data Factory pipelines to Cloud Composer for orchestrating data movement and transformations. Redesigned workflows using best practices.
* Implemented CI/CD for Big Query and Dataproc using Cloud Build and Ansible. Automated testing and deployment of infrastructure as code.
* Stored data files in Google Cloud S3 buckets daily basis using Data Proc, Big Query to develop and maintain GCP cloud-based solutions.
* Used Informatica power center 9.6.1 to Extract, Transform and Load data into Netezza Data Warehouse from various sources like Oracle and flat files.
* Highlight any ETL (Extract, Transform, Load) processes you've implemented to populate and synchronize data in the graph database.
* Rebuilt Azure Data bricks notebooks in Cloud Dataflow for more scalable and server less big data processing. Reduced runtimes by 40%.
* Developed Spark Programs in Scala to perform data transformations, created data frames and running applications by using Scala.
* Developed and modified the complex Informatica mappings using BDM/DEI, PowerCenter, IICS for business rules integration and changes.
* Used Cloud Functions to support the data migration from Big Query to the downstream applications.
* Designed new data lake architecture on Google Cloud leveraging storage tiers for cost optimization. Established data lifecycle management policies.
* Led migration of 500TB of unstructured data from Azure Blob Storage to Cloud Storage. Optimized data layout and leveraged advanced storage features like object versioning.
* Worked on the migration of 500TB of unstructured data, optimizing data layout and leveraging advanced storage features, for a global perspective.
* Demonstrate your skills in optimizing graph database performance through indexing, query tuning, and other performance-enhancing techniques.
* Kept up-to-date with the latest advancements in PostgreSQL and related technologies.
* Collaborated within Agile and Scrum methodologies, participating in sprint planning and retrospectives.
* Developed real time data pipeline using Kafka and Spark stream to ingest client data from their web log server and apply transformation.
* Migrated databases from Azure SQL to Cloud Spanner utilizing schema conversion tools. Optimized data model for horizontal scalability.
* Provide examples of how you've addressed scalability challenges in graph databases.
* Contributed to the design and implementation of scalable and cost-effective data storage solutions on GCP, utilizing Cloud Storage and BigQuery based on specific data mapping requirements.
* Collaborated with cross-functional teams to define and implement efficient data mapping and modeling strategies.
* Developed data integration workflows in Data Fusion using visual drag-and-drop interface, enabling rapid development and deployment of data pipelines.
* Led intricate data mapping initiatives for the migration of a petabyte-scale data warehouse from Azure Synapse to BigQuery, focusing on optimizing schema designs and ETL processes.
* Implemented CI/CD for Big Query and Dataproc using Cloud Build for enabling seamless code deployment and testing in a controlled environment.
* Participated in forums, conferences, and training sessions to enhance PostgreSQL expertise.
* Played a pivotal role in the migration of a petabyte-scale data warehouse, focusing on optimizing schema designs and ETL processes for improved data modeling.
* Utilized Spark programs in Scala for advanced data transformations, showcasing expertise in modeling techniques for comprehensive data analysis.
* Environment: Google Big Query, Cloud Storage, Azure Databricks, Cloud DataProc, Apache Kafka, Cloud Data Fusion, Apache Beam, SSIS, Azure Blob Storage, Azure SQL, Teradata, Hadoop, SSMS, Python, Power BI, Apache Spark.
* Client: Vanguard , Charlotte, NC  **May 2018 to Jan 2019**

**Role: Azure Data Engineer**

**Responsibilities:**

* Implemented end-to-end data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from diverse sources into Snowflake.
* Used Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics, extract, transform, and load data from source systems to Azure Data Storage services.
* Utilized Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics for data mapping and modeling.
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure SQL, Azure DW) and processing the data in Azure Databricks.
* To meet specific business requirements wrote UDF’s in Scala and PySpark.
* Utilizes Azure ETL, Azure Data Factory (ADF) services to ingest data from legacy disparate data stores- SFTP Servers and Cloudera Hadoop’s HDFS to Azure Data Lake Storage (GEN2).
* Executed meticulous data mapping processes using Azure Data Factory, ensuring seamless extraction, transformation, and loading (ETL) of diverse data sources into Snowflake.
* Played a key role in the comprehensive redesign of schema structures, enhancing data modeling efficiency and scalability.
* Developed Spark applications using PySpark and Spark SQL in Data Bricks for data extraction, transformation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage pattern.
* Implemented Proof of concepts for SOAP & REST APIs and utilized REST APIs to retrieve analytics data from different data feeds.
* 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, write-back tool, and backward.
* Implemented data integration solutions using Azure Data Factory and third-party tools like Informatica or Talend, ensuring seamless data movement across on-premises and cloud environments.
* Implemented data loading processes using Snowflake's native data ingestion methods, integrating with Azure Data Factory for seamless data pipeline management.
* Worked on Dimensional Data Modeling (Star schema, Snowflake schema), Transactional Modeling, and SCD (Slowly Changing Dimension).
* Implemented data serialization and deserialization techniques for efficient data transfer and compatibility between various components in the Kafka ecosystem.
* Involved in designing data warehouse and creating fact and dimension tables with Star schema and Snowflake schema.
* Developed report using Tableau that keeps track of the dashboards published to Tableau Server, which helps us find the potential future clients in the organization.

**Environment:** Azure Data bricks, Azure Data Factory (ADF), SQL Server, Kafka, Unix, LINUX, Shell Script, Snowflake, Azure Synapse, Hadoop, Tableau, Informatica, Python.

**Client:** **Dream orbit, Bangalore, India Sep 2016 – Nov 2017**

**Data Engineer**

**Responsibilities:**

* Assisted in requirements gathering and analysis for data projects, collaborating with cross-functional teams to define data needs.
* Utilized Python, R, SQL, and PL/SQL to manipulate, clean, and transform data, ensuring its accuracy and reliability.
* Implemented data processing and analysis tasks using libraries such as NumPy, Pandas, and Seaborn, contributing to data-driven insights.
* Supported ETL (Extract, Transform, Load) processes to extract data from various sources, transform it into a usable format, and load it into databases.
* Worked with tools like SciPy and SSIS to optimize data processing and streamline workflows.
* Maintained and optimized SQL Server databases to ensure data availability and performance.
* Created and automated Excel data reports and dashboards, providing valuable stakeholder insights.
* Assisted in managing and configuring the ServiceNow platform to support data integration and workflow automation.
* Collaborated with cross-functional teams and participated in requirements gathering and analysis for global data projects.
* Engaged in extensive data mapping and modeling activities to support insurance product modeling, employing Python, R, SQL, and PL/SQL.
* Employed Python, R, SQL, and PL/SQL to manipulate, clean, and transform data for international clients.

**Environment:** Python, R, SQL, PL/SQL, NumPy, Pandas, Seaborn, ETL, SciPy, SSIS, SQL Server, Excel.

**Client: Careator Technologies Pvt Ltd Hyderabad, India Nov 2014 to August 2016 Role: Data Analyst**

**Responsibilities:**

* Developed and examined business needs to create technically proficient data solutions that can be put into practice.
* Analyzing classified data items for data profiling and mapping from source to target data environments and created.
* Working documents to back up results and assign responsibilities.
* Used complicated SQL to analyze and profile data from a variety of sources, including Teradata and Oracle.
* Participated in meetings for gathering information and JAD sessions to deliver a business requirement document.

and a draft logical data model.

* Specify how the data will be sourced from sources and loaded into DWH tables by defining the ETL mapping.

Specification and designing the ETL procedure.

* Created mappings using the transformations Source Qualifier, Expression, Filter, Lookup, Update Strategy, Sorter,

Joiner, Normalizer, and Router.

* Carrying out data administration tasks and completing ad-hoc requests in accordance with user requirements using
* Data management software and tools like Perl, Toad, MS Access, Excel, and SQL.
* Utilized PL/SQL to write, test, and implement triggers, stored procedures, and functions at the database level.
* Created number of reports using a collection of data from a SQL query and a set of pivot tables built up in Excel.
* Extracted data from a variety of sources such as MySQL, Oracle 11g, MS SQL server, etc., and analyzed large data sets using tools such as SQL, SAS, and JMP.
* Using Tableau, I created and delivered reports that had drill-down, drill-through, and drop-down menu options, as well as parameterized and linked reports.

**Environment:** MS SQL Server, Teradata, PL/SQL, IBM Control Center, TOAD, Microsoft Project Plan, Repository Manager, Workflow Manager, Oracle 10g/9i, TOAD, UNIX, and Shell scripting.

**Brio Technologies Private Limited Hyderabad, India Sept 2013 to Nov 2014**

**Role: ETL Developer**

**Responsibilities:**

* Creating mappings, sessions, and workflows to use with Informatica PowerCenter to load data for Ultima tix projects from source to target database.
* Creating Informatica positions for SCD Types 1, 2, and 3. Transferring the processes and mappings from development to QA, QA to production, and unit testing the procedure.
* Participated in the seamless migration of the mappings, sessions, workflows, and repositories from Informatica 9.0

to Informatica 9.6.

* Created the Informatica Mappings by leveraging the Aggregator, SQL overrides in Lookups, source qualifiers, and
* Router to govern data flow into different targets.
* Stored the data in JSON format and developed and tested many features of dashboards using Python.
* With the Informatica Power Center Designer, strong mappings were created using a variety of transformations, including Filter, Expression, Sequence Generator, Update Strategy, Joiner, Router, and Aggregator.
* PL/SQL stored procedures were created, and push-down optimization tuning was used to minimize execution times by 30%.
* Created materialized views, planned jobs, ETL workflows, and reporting that allowed data to move between eight ERP systems.
* Conducted unit testing at various ETL stages and participated actively in team code reviews.
* Working knowledge of tools for SQL Server 2005, 2008 R2, and 2012, including Management Studio, Query Analyzer,

**Environment:** Oracle, Informatics 9.6, PL/SQL, UNIX, Shell Scripting, Python, ALM, Agile Methodology.

**Education:**

BV Raju Institute of Technology, India 2009- 2013

Computer Science Engineering