**Name**: Dinesh Kumar

**Phone:** **571-999-7753**

**Email:** **naveenp@molinatek.com**

**Role:** Sr. Data Engineer

**PROFESSIONAL SUMMARY:**

* Seasoned IT professional with over 9 years of diversified experience in E2E Data Analytics Platforms (ETL – BI-Java) as **Bigdata, Hadoop, Java/J2EE Development, Informatica, Data Modelling**, and **System Analysis** across Banking, Finance, and Health Care domains.
* Proficient in deploying and managing cloud solutions across **AWS, GCP**, and **Azure**, enhancing data management and application performance.
* Implemented **AI** and **machine learning** models across various sectors, including healthcare, finance, and retail, driving data-driven decision-making.
* Hands-on Experience **Hadoop Framework** and its ecosystem like **Distributed file System (HDFS), MapReduce, Pig, Hive,** and **Spark.**
* Developed automated data ingestion processes using tools like **GCP Dataflow** and **Informatica**, ensuring efficient data pipeline management.
* Created dynamic, interactive dashboards and reports using tools like **Tableau, Power BI**, and **Looker**, enhancing data accessibility, and understanding.
* Effectively oversaw real-time data streaming and processing operations, leveraging platforms such as **GCP Pub/Sub** and **AWS Kinesis** to deliver timely insights and analytics.
* Experienced senior data engineer proficient in managing data clusters, setting up Hive tables, and implementing effective Star/Snowflake schema. Skilled in handling Slowly Changing Dimension (SCD) tables, PowerShell scripting, and utilizing **Git & CI/CD** for efficient code management. Familiar with **Azure Synapse** and adept at SQL assessment for database optimization.
* Extensive experience in Developing Data Warehouse applications using **Hadoop, Informatica, Oracle, Teradata, MS SQL Server** on Unix, and Windows platforms.
* Proficient in databases including **Oracle, MySQL, SQL Server, PostgreSQL,** and **Teradata** for data storage and management.
* Experienced in programming languages such as **Python, Scala, Java** and **Shell script** for developing data processing algorithms and applications.
* Familiar with cloud technologies like **Databricks, Kubernetes, AWS,** and **Microsoft Azure** for scalable and cost-effective data solutions.
* Skilled in frameworks such as **Django REST** framework for web development.
* Experienced in using development tools like **PyCharm, Eclipse, Visual Studio, SQL Developer,** and **TOAD**.
* Proficient in versioning tools like **SVN, Git,** and **GitHub** for code management.
* Expertise in using **Core Java, J2EE, Multithreading, JDBC, Shell Scripting** and proficient in using Java API’s Collection, Servlets, JSP for application development.
* Good Work Experience with **UNIX/Linux** Commands, Scripting and Deploying the applications on the servers.
* Knowledgeable in operating systems like **Windows, Ubuntu Linux,** and **MacOS**.
* Experience in **Dimensional Data Modelling Star Schema, Snow-Flake Schema**, Fact and Dimensional Tables, Concepts like **Lambda Architecture,** and **Batch Processing**.
* Worked Closely to review Pre- and Post-Processed data to ensure data accuracy and integrity with Dev and QA teams.
* Experience in **Java, J2EE, JDBC, Collections, Servlets, JSP, Struts, Spring, Hibernate, JSON, XML, REST, SOAP Web Services, Groovy, Eclipse, WebLogic, WebSphere**, and **Apache Tomcat Servers**.
* Experience in Working as part of fast-paced Agile Teams, exposure to testing in **Scrum Teams, Test-Driven Development**.

**SKILLS:**

* **Big Data Technologies:** Apache Hadoop (HDFS), Spark
* **Databases**: Snowflake, SQL Server, Oracle Database, Teradata Database, MySQL, PostgreSQL, SQLite
* **Programming**: Python, SQL, Scala
* **Cloud Technologies**: AWS (Amazon Web Services), Azure, Google Cloud Platform (GCP)
* **Frameworks**: Django, Flask
* **Tools**: Informatica PowerCenter, SQL Alchemy, Beautiful Soup, Scrapy, Pandas, Matplotlib, NumPy, SciPy, Click, Fabric, Git, PY test, Unit Test
* **Versioning tools**: Git
* **Operating Systems**: Unix/Linux
* **Testing Methodologies**: Test-driven development (TDD)
* **Database Modelling**: ORM (Object-Relational Mapping)
* **Visualization/ Reporting**: Tableau, Power BI, SAS Analytics
* **Machine Learning Techniques**: Predictive modeling, Statistical analysis
* **Data Warehousing**: Informatica PowerCenter
* **Streaming Platforms**: Apache Kafka, AWS Kinesis

**WORK EXPERIENCE :**

**Client: TJX, Northborough, MA Nov 2022 - Present**

**Role: GCP Data Engineer**

**Responsibilities:**

* Spearheaded the utilization of Google Cloud Platform services to deploy and optimize applications, enhancing system performance and scalability while managing and optimizing Snowflake environments for efficient data storage and querying for analytics.
* Developed and implemented AI models for accurate demand forecasting, significantly improving inventory management, and machine learning algorithms for customer segmentation, aiding in personalized marketing strategies.
* Designed and managed robust data warehousing solutions, supporting comprehensive data analysis,and orchestrated efficient data movement processes, ensuring timely and accurate data availability for decision-making.
* Performed complex data transformations to extract meaningful insights into sales and customer behavior and implemented automated data ingestion pipelines to enhance the speed and reliability of data flow into systems.
* Developed sophisticated data models tailored to business needs, facilitating effective data-driven strategies, and ensured high levels of data security through rigorous encryption and security protocols.
* Managed real-time data streaming services on GCP, enabling immediate analysis and response to trends, and utilized predictive analytics to identify and capitalize on emerging sales trends, boosting revenue.
* Applied NLP techniques to analyze customer feedback, enhancing satisfaction and created dynamic, interactive dashboards using business intelligence tools, providing actionable insights to management.
* Optimized ETL pipelines in the cloud, improving data processing efficiency and accuracy, and developed AI-powered chatbots for enhancing customer service and engagement.
* Established and maintained robust data governance frameworks, ensuring integrity and compliance in operations.
* Leveraged big data analytics to extract meaningful insights, driving strategic decisions.
* Analyzed customer behavior using machine learning models, enabling targeted marketing and sales strategies, and created algorithms for inventory optimization, reducing overheads and improving stock turnover.
* Integrated various APIs to enhance functionality and interoperability of systems and applications, and expertly managed both SQL and NoSQL databases, catering to diverse data needs.
* Utilized Docker for containerization of applications, ensuring consistency across different environments, and developed, managed datalake architecture, providing a scalable data storage solution.
* Implemented continuous data quality improvement processes, ensuring high-quality data for analytics, and effectively managed cloud costs, ensuring cost-effective operations.
* Developed supply chain optimization models, enhancing efficiency and reducing costs in logistics, and designed scalable data architectures to support growth and expansion.
* Led cross-functional teams in data projects, driving collaboration and innovation to achieve project objectives and enhance data-driven decision-making processes.

**Environment** : Google Cloud Platform (GCP), Snowflake, AI models, ML algorithms, Natural Language Processing (NLP) techniques, BI tools, Big data analytics, Data governance frameworks, ETL pipelines, API integration, SQL databases, NoSQL databases, Docker, DataLake architecture.

**Client: Homesite Insurance, Boston, MA Dec 2020 – Oct 2022**

**Role: Azure Data Engineer**

**Responsibilities:**

* + Led the design and implementation of data processing pipelines using Azure Data Factory, ensuring efficient extraction, transformation, and loading (ETL) of data from various sources into Azure data lakes and warehouses.
	+ Utilized Azure SQL Database and Azure SQL Data Warehouse (now known as Azure Synapse Analytics) for storing and analyzing structured and semi-structured data, implementing data models, and optimizing query performance.
	+ Implemented Azure Blob Storage and Azure Data Lake Storage Gen1 (now replaced by Gen2) for scalable and cost-effective storage of big data, facilitating data lake architecture and analytics.
	+ Developed and deployed Azure Databricks clusters for big data analytics and machine learning tasks, leveraging Apache Spark for distributed data processing and Azure Machine Learning for predictive analytics.
	+ Orchestrated data workflows and data processing tasks using Azure Data Factory pipelines and Azure Scheduler, automating data integration, transformation, and movement processes.
	+ Configured and managed Azure HDInsight clusters for running Apache Hadoop, Apache Spark, and Apache HBase workloads in the cloud, enabling scalable and cost-effective big data processing.
	+ Utilized Azure Event Hubs for real-time event ingestion and processing, capturing streaming data from various sources and feeding it into Azure data lakes and analytics services.
	+ Designed and implemented data security and access control measures using Azure Active Directory, Azure Key Vault, and Azure Role-Based Access Control (RBAC), ensuring data privacy and compliance with industry regulations.
	+ Implemented Azure Stream Analytics for real-time data analytics and monitoring, analyzing streaming data streams from IoT devices and sensors for operational insights and predictive maintenance.
	+ Developed and deployed Azure Functions for serverless data processing and automation tasks, enabling event-driven data processing and integration with other Azure services.
	+ Created data visualizations and reports using Power BI, enabling stakeholders to gain insights and make data-driven decisions based on analytics and business intelligence.

**Environment:** Azure Data Factory, Azure SQL Database, Azure SQL Data Warehouse (now Azure Synapse Analytics), Azure Blob Storage, Azure Data Lake Storage Gen1 (now Azure Data Lake Storage Gen2), Azure Databricks, Azure Machine Learning, Azure Scheduler, Azure HDInsight, Azure Event Hubs, Azure Active Directory, Azure Key Vault, Azure RBAC, Azure Stream Analytics, Azure Functions, Power BI

**Client: Fiserv, Omaha, Nebraska Mar 2018 – Nov 2020**

**Role: AWS Data Engineer**

**Responsibilities:**

* + Led cross-functional teams in the successful implementation of Agile methodologies, including Scrum, fostering collaboration and ensuring efficient project execution.
	+ Guided and mentored junior data engineers, providing technical leadership and fostering their professional growth within the organization.
	+ Worked on end to end Machine Learning workflow, written Python code for gathering the data from AWS Snowflake.
	+ Facilitated Scrum meetings and stand-ups, ensuring effective communication, task prioritization, and timely issue resolution, resulting in increased team productivity and project success.
	+ Constructed and enhanced data processing pipelines, responsible for extracting, transforming, and loading data from a multitude of sources into diverse data storage systems.
	+ Collaborated with different data storage technologies such as AmazonS3, Redshift, RDS, DynamoDB, and other similar platforms as necessary.
	+ Utilized Kubernetes and Docker for runtime environment for the CI/CD system to built, test, and deploy.
	+ Optimized data models and schemas in data bricks on AWS with Terraform.
	+ Adept in MDM processes, Data cleaning and extractions, Data governance and Data profiling.
	+ Used Kafka for activity and Log aggregation.
	+ Launched Compute (EC2) and DB (Aurora, Cassandra) instances form Amazon Management Console and CLI.
	+ Used AWS Glue Catalog with crawler to get the data from S3 and perform SQL query operations using AWS Athena.
	+ Executed process improvements in data workflows using Alteryx processing engineer and SQL.
	+ Written Terraform scripts to automate AWS services which include ELB, Cloud Front distribution, RDS, EC2 and S3 bucket and converted existing AWS infrastructure to AWS Lambda deployed via Terraform and AWS Cloud Formation.
	+ Created analytical reports for end-to-end traceability, Lineage, definition of key Business elements from Aurora.
	+ Integrated Apache Airflow with AWS to monitor Machine Learning workflows on Sage Maker With Terraform.
	+ Collaborated with the Data Science team building machine learning models on Spark EMR cluster to deliver the data needs under business requirements.
	+ Developed MapReduce/Spark Python modules for Machine Learning in Hadoop on AWS.
	+ Guided and migrated PostgreSQL and MySQL databases to AWS Aurora.
	+ Demonstrated proficiency in designing and implementing ETL workflows, utilizing AWS services like AWS Glue, AWS Data Pipeline, or AWS Step Functions.
	+ Developed Spark applications leveraging Pyspark and Spark-SQL to perform data extraction, transformation, and aggregation tasks across multiple file formats.
	+ Written AWS Lambda code in Python for nested Json files, converting, comparing, sorting, etc.
	+ Created a Snowflake warehouse strategy and set it up to use PUT scripts to migrate a terabyte of data from S3 into Snowflake.
	+ Used coding/scripting pipelines and APIs to uncover and turn data into assets using AWS services like Athena Data Catalogue.
	+ Proficiently utilized Looker to generate and tailor dashboards, reports, and data visualizations, empowering data-driven decision-making processes.
	+ Demonstrated skill in utilizing Elastic search to perform efficient searching, analysis, and visualization of substantial volumes of both structured and unstructured data, catering to diverse use cases.
	+ Written PySpark job in AWS Glue to merge data from multiple tables and in Utilizing Crawler to populate AWS Glue data Catalog with metadata table definitions.
	+ Generated a script in AWS Glue to transfer the data and utilized AWS Glue to run ETL jobs and run aggregation on PySpark code.
	+ Design and Develop ETL Processes in AWS Glue to migrate Campaign data from external sources like S3, ORC/Parquet/Text Files into AWS Redshift.
	+ Spearheaded the migration of legacy data systems to Snowflake, implementing data modeling strategies and schema design to optimize query performance and enable efficient data storage and retrieval.
	+ Collaborated with cross-functional teams to design and develop complex data pipelines, integrating various data sources into Snowflake and implementing data transformations, ensuring high data quality and consistency throughout the analytics ecosystem.
	+ Implemented Spark Streaming to ingest real-time data from Kafka and subsequently stored the streaming data to HDFS using Python, while effectively utilizing NoSQL databases like HBase and Cassandra.
	+ Implemented data quality and data governance frameworks to ensure data accuracy, consistency, and integrity.
	+ Written Pyspark job in AWS Glue to merge data from multiple tables and in Utilizing Crawler to populate AWS Glue Data Catalog with metadata table definitions.
	+ Leveraged Spark Streaming to collect data in near-real-time from an AWSS3 bucket, conducting on-the-fly transformations and aggregations to construct the common learner data model, and subsequently persisted the processed data in HDFS.
	+ Implemented the installation and configuration of a multi-node cluster on the cloud using Amazon Web Services (AWS), specifically on EC2 instances.
	+ Demonstrated hands-on experience in implementing DevOps practices, including Infrastructure as Code, Continuous Integration and Deployment (CI/CD), and automated testing.
	+ Proficiently worked with containerization technologies like Docker and Kubernetes.
	+ Proficiently managed AWS Management Tools, specifically CloudWatch and Cloud Trail, to monitor and track activities within the AWS environment.
	+ Implemented versioning in S3 buckets to ensure enhanced security and data integrity for highly sensitive information stored within them.
	+ Leveraged different file formats Parquet, Avro, ORC, and Flat files. Used Snappy and GZIP compression codec to optimize storage and processing.
	+ Developed and implemented scalable data pipelines using Databricks on AWS to ingest, transform, and load large volumes of data into data lakes and data warehouses.
	+ Designed and optimized data models and schemas for efficient data storage and retrieval in Data bricks on AWS.
	+ Designing and Implementing Splunk – based best practice solutions.
	+ Used Kubernetes to orchestrate the deployment, scalling and management of Docker containers.
	+ Involved in Design, Implementation and modifying the Python Code and MySQL database schema on the Back-End.
	+ Integrated Apache Airflow with AWS to monitor multi-stage Machine Learning workflows with the tasks running on Amazon Sagemaker.
	+ Leveraged Data bricks Delta Lake for managing and maintaining structured and versioned data in a cloud-based environment.
	+ Implemented real-time streaming data processing using Data bricks Streaming on AWS to enable near real-time analytics and insights.
	+ Processed the data from Kafka pipelines from topics and show the real time streaming in dashboards.
	+ Worked on migration data from Teradata to AWS using Python and BI tools like Alteryx.
	+ Install and configured Splunk clustered search head and Indexer, Deployment servers, Deployed.
	+ Worked on POC to integrate Amazon S3 with Informatica Intelligent Cloud Services (IICS) for processing multiple Parquet files, CSV files to Snowflake.
	+ Designed various Jenkins jobs to continuously integrate the processes and executed CI/CD pipeline using Jenkins.
	+ Sound knowledge in developing highly scalable and resilient Restful APIs, ETL solutions, and third-party platform integrations as part of the Enterprise Site platform.
	+ Sound experience in building production ETL pipelines between several source systems and Enterprise Data Warehouse by leveraging SSIS, SSAS, and SSRS.
	+ Experience in different phases of Data Warehouse development like requirements gathering, design, development, implementation, testing, and documentation.
	+ Solid knowledge of Dimensional Data Modeling with Star Schema and Snowflake for FACT and Dimensions Tables using Analysis Services.
	+ Successfully integrated AWS Dynamo DB with AWS Lambda to seamlessly store item values and create backups of Dynamo DB streams.
	+ Demonstrated expertise in managing and maintaining the cloud infrastructure that supported data pipelines, encompassing a wide range of AWS services including EC2, VPC, IAM, and others.
	+ Expertly installed and configured Apache Airflow for AWSS3 bucket, establishing the necessary infrastructure to create Directed Acyclic Graphs (DAGs) and execute them seamlessly within the Airflow environment.
	+ Developed automation scripts using Pyspark and Scala to streamline the ingestion process, enabling seamless data integration from diverse sources such as APIs, AWSS3, Teradata, and Redshift.
	+ Used Git Version control to manage the source code and integrating Git with Jenkins to support build automation.
	+ Used public or private Kubernetes frameworks for scaling data and services infrastructure.
	+ Proficiently crafted multiple automation scripts using Pyspark to facilitate both ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) processes, efficiently handling data from various sources.
	+ Developed Pyspark scripts leveraging SQL and RDD in Spark for comprehensive data analysis, enabling efficient processing and subsequently storing the results back into the S3 storage system.
	+ Experience in designing interactive dashboards, reports, performing ad-hoc analysis, and visualizations using Tableau, Power BI, SSRS, Crystal reporting.

**Environment:** AWS (EC2, S3, EBS, Splunk, Machine Learning, Snowflake, ELB, RDS, SNS, SQS, VPC, Cloud

formation, Cloud Watch, ELK Stack, Docker, Airflow, AWS Glue), Bit bucket, Ansible, Python, Shell Scripting,

GIT, Jira, AWS Glue data Catalog, Jenkins, Terraform, My SQL, MDM, Aurora, Kubernetes, Redshift,

Unix/Linux, IICS, Alteryx, Dynamo DB, Kinesis.

**Client: Cadila Pharmaceuticals Ltd, India Jan 2016 to Aug 2017**

**Role: Data Engineer**

**Responsibilities:**

* Led the design and implementation of data processing pipelines using Informatica PowerCenter, facilitating efficient extraction, transformation, and loading (ETL) of healthcare data from disparate sources into enterprise data warehouses.
* Utilized SQL Server Integration Services (SSIS) for data integration and ETL tasks, orchestrating complex data workflows and automating data movement processes to ensure data accuracy and consistency.
* Developed and maintained data models and schemas using Microsoft SQL Server for storing and analyzing structured healthcare data, ensuring data integrity and optimized query performance.
* Implemented Oracle Database for storing and managing electronic health records (EHRs), patient demographics, medical histories, and clinical data, enabling secure and scalable storage of healthcare information.
* Utilized IBM Infosphere Information Server for data integration, data quality, and data governance, ensuring compliance with healthcare regulations and standards such as HIPAA.
* Designed and developed data warehousing solutions using Teradata Database, enabling advanced analytics, reporting, and decision support for healthcare providers and stakeholders.
* Implemented IBM Cognos for healthcare analytics and reporting, creating interactive dashboards and visualizations to monitor key performance indicators (KPIs) and track healthcare outcomes.
* Leveraged SAS Analytics for healthcare predictive modeling and statistical analysis, identifying trends, patterns, and insights from healthcare data to improve patient outcomes and healthcare delivery.
* Developed and maintained data pipelines using Python and Apache Hadoop (HDFS), enabling scalable and distributed processing of healthcare data for analytics, research, and population health management.
* Collaborated with healthcare professionals, clinicians, and data scientists to understand data requirements, develop data-driven solutions, and support clinical decision-making and healthcare research initiatives.
* Provided technical expertise, guidance, and support to cross-functional teams, including data analysts, healthcare informaticians, and IT professionals, to ensure successful implementation and utilization of data solutions.

**Environment:** Informatica PowerCenter, SQL Server Integration Services (SSIS),Microsoft SQL Server, Oracle Database, IBM Infosphere Information Server, Teradata Database, IBM Cognos, SAS Analytics, Python, Apache Hadoop (HDFS).

**Client: HSBC Software Development, India** **Jun 2014 to Dec 2015**

**Role: Python Developer**

**Responsibilities:**

* Developed web applications and backend systems using Python programming language and the Django web framework, enabling rapid development and deployment of scalable software solutions.
* Utilized Flask, a lightweight web framework for Python, to build RESTful APIs and microservices for integrating with third-party systems and external services.
* Implemented object-relational mapping (ORM) using SQL Alchemy to interact with relational databases such as MySQL, PostgreSQL, and SQLite, enabling seamless data manipulation and persistence.
* Developed and maintained web scraping scripts using Beautiful Soup and Scrapy, extracting data from websites and web pages for data analysis, content aggregation, and business intelligence purposes.
* Implemented data analysis and visualization tasks using Pandas and Matplotlib, processing and analyzing structured and unstructured data to derive insights and generate visualizations for stakeholders.
* Utilized NumPy and SciPy libraries for scientific computing and numerical analysis, solving complex mathematical and statistical problems in software applications and research projects.
* Developed and maintained command-line interfaces (CLIs) and automation scripts using Click and Fabric, enabling efficient system administration and DevOps tasks such as deployment, configuration management, and testing.
* Integrated version control using Git for source code management and collaboration, following best practices for branching, merging, and code reviews to ensure code quality and project integrity.
* Implemented unit testing and test-driven development (TDD) using PY test and unit test frameworks, ensuring software reliability, scalability, and maintainability through automated testing.
* Collaborated with cross-functional teams including software engineers, QA testers, and product managers to define requirements, design software architectures, and deliver high-quality software products and features.
* Provided technical support, troubleshooting, and bug fixes for existing software applications, ensuring timely resolution of issues, and minimizing downtime for end users.

**Environment** : Python, Django, Flask, SQL Alchemy, Beautiful Soup, Scrapy, Pandas, Matplotlib, NumPy, SciPy, Click, Fabric, Git, PY test, Unit Test.