MANOJ KUMAR

Dallas, TX 75039 | 469-629-9272 | reachatmanojg@gmail.com

Professional Summary

- Certified Databricks Developer Professional With over 10 years of experience, there is a strong backgroundin working with Databricks Lake House, migrating Hadoop and SQL databases to Azure Data Lake, Azure Data Lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data Warehouse.
- Excel at managing and granting data access, as well as migrating on-premises databases to the Azure Data Lakestore using Azure Data Factory. Expertise in developing Spark applications using Spark-SQL in Databricks.
- Successfully extracted, transformed, and aggregated data from multiple file formats, enabling analysis and transformation to uncover valuable insights into customer usage patterns.
- Possesses a solid understanding of Spark Architecture, including its core components such as Spark SQL, Data Frames, Spark Streaming, Driver Nodes, Worker Nodes, Stages, Executors, and Tasks. Knowledge extends to BigData Hadoop and Yarn architectures, along with various Hadoop demons like Job Tracker, Task Tracker, Name Node, Data Node, Resource/Cluster Manager, and Kafka (distributed stream-processing).
- Strong grasp of Hadoop cluster architecture, encompassing distributed file systems, parallel processing, scalability, fault tolerance, and high availability. Within Databricks, I gained extensive experience in utilizing various Hadoop ecosystem tools, including HDFS, MapReduce, Yarn, Spark, Kafka, Hive, Impala, HBase, Sqoop,Pig, Airflow, Oozie, Zookeeper, Ambari, Flume, and Nifi.
- Skilled in designing and implementing Dimensional Data Models for FACT and Dimension Tables using StarSchema and Snowflake in Databricks. Optimizing Hadoop cluster performance, debugging, monitoring, datatransformations, mapping, cleaning, and troubleshooting are areas of demonstrated expertise in Databricks.
- Leverage AWS cloud services to architect and deploy scalable data storage and processing solutions that seamlessly integrate with Databricks.
- Proficient in Python programming for Databricks, utilizing features such as Lists, Dictionaries, Tuples, the Pandasframework, and the boto3 package to read files from S3.
- Have a strong technical skill set, including proficiency in programming languages such as Python, Java, and SQL.
- Demonstrated ability to quickly grasp and apply new technologies, frameworks, and tools to solve complex technical challenges.
- Build ETL (Extract, Transform, Load) processes using Databricks notebooks and AWS Glue, orchestrating data workflows efficiently.
- Additionally, an excellent communicator with strong work ethics proactive, a team player, and maintains a positive attitude. Possesses domain knowledge in Finance, Logistics, and Health insurance.
- Skills extend to visualization tools such as Power BI and Excel, including formulas, Pivot Tables, Charts, and DAX Commands. Experience in Agile software development methodologies, including Scrum and Sprint, as wellas traditional models like Waterfall and TDD for Databricks.
- Extracted, transformed, and loaded data from various source systems into Azure Data Storage services usingAzure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics
- Well-versed in UNIX, Linux, and Windows operating systems for Databricks. Have a solid understanding of the Software Development Lifecycle (SDLC) and various techniques like Waterfall and Agile for Databricks.
- Proficient in utilizing Databricks collaborative workspace for team-based software development, data engineering, and analysis.
- Lead and manage technical projects, overseeing all aspects of planning, development, and implementation.

- Ensured code quality and reliability by incorporating automated testing and code review processes into Databricks-based software development workflows.
- Excel in analyzing and solving technical problems, whether related to software development, system optimization, or data analysis.

Skills

- Cloud Technologies and Services
- Amazon AWS- EMR, EC2, ENS, RDS, S3, Athena, Glue, Elastic search, Lambda, SQS, DynamoDB, Redshift, Kinesis, Microsoft Azure- Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Google Cloud Platform
- Big Data Ecosystems
- Databricks Lakehouse, Apache Spark, HDFS, YARN, Map-reduce, Sqoop, Hive, Oozie, Pig, Spark, Zookeeper, Cloudera Manager, Kafka, Flume, NiFi, Connect, Airflow, Stream Sets, Kafka connect.
- Hadoop Distributions
- Apache Hadoop 2. x, Cloudera CDP, Hortonworks HDP
- Cassandra, MongoDB, HBase
- Database, Integration,
- MySQL, Oracle, Teradata, MSSQL SERVER, PostgreSQL, DB2
- Database Management
- Version Control
- Git, SVN, Testing
- BI tools, Software Development
- Tableau, PowerBI
- Scripting language Python, PySpark, SparkSQL, SQL, Scala, R, shellscripting, HiveQL.

Work History

Client: - TriWest Healthcare Alliance – Phoenix, AZ **08/2022 – Present**
Role:- Senior Data Engineer.

- Demonstrated expertise in implementing robust data security measures to safeguard sensitive informationthroughout the ETL pipeline process in Azure Databricks
- Utilizing encryption protocols, access controls, and data masking techniques to protect data integrity and privacy.
- Proficiency in developing and maintaining data validation processes is a strength, utilizing Spark SQL andDatabricks workflows.
- Possesses the skills to design and maintain Azure data warehouses to efficiently store business datasets.
- Capabilities extend to leveraging Python programming skills to build Azure Databricks ETL pipelines and meetspecific business requirements.
- Proficient in writing and maintaining Python unit tests for Azure applications, utilizing Bitbucket and Jenkins
- Successfully integrated Adobe Analytics with other marketing and analytics tools using Azure Databricks
- Skilled in designing ETL Data Pipeline flows for data ingestion from RDBMS sources to Azure, includingMySQL.
- Experienced in creating and maintaining data workflows with dependencies for efficient big data processing on Azure Databricks.
- Strong background in adhering to data governance best practices, utilizing Unity Catalog and version control onAzure.
- Prepared and preprocessed data using Python libraries like Pandas and NumPy, ensuring it was suitable for training machine learning models.
- Developed and deployed predictive models in production using Databricks machine learning pipelines.

- Utilized Azure Databricks to design, construct, and optimize data pipelines, utilizing distributed computing and advanced analytics for faster data processing, resulting in improved data-driven decision-making inside the enterprise.
- Proficient in providing datasets for building models developed by the data science team using Azure Databricks
- Skilled in data modeling, statistical analysis, and data visualization on Azure Databricks.
- Excellent knowledge of machine learning methods, model creation, and data analysis.
- Capable of enhancing Tableau visualization load times by caching data with Data Reflections using the AzureData Lake House Engine
- Experienced in optimizing resource-intensive SQL database queries to achieve significantly faster performance. using Azure best practices.
- Skilled in designing reports using Azure Databricks warehouse endpoint to identify the latest hacking techniquesprevalent in the market.
- Capable of designing complex SQL queries to validate data and establish a reliable data warehouse using Azure. Databricks.
- Experienced in creating Azure Databricks workflows to ensure continuous job execution for efficient dataprocessing.
- Proficient in building machine learning models using Databricks' MLlib and other relevant libraries.
- Implemented monitoring and logging solutions for DataProc clusters, proactively identifying and resolving performance bottlenecks and issues.
- Implemented Git for source code version control in data integration and transformation workflows, enforcing best practices for code collaboration.
- Proficient in developing and maintaining data catalogs in Azure Databricks to document data sources, metadata, and lineage information

Environment: Azure data bricks, Teradata, SQL Server, ETL pipelines, Terraform, Apache Presto, Apache drill.

Client: - Sonder – San Francisco, CA **11/2020 - 08/2022**
Role: - Databricks Data Engineer

- Designed and setup Enterprise Data Lake to provide support for various uses cases including Storing, processing, Analytics and Reporting of voluminous, rapidly changing data by using various AWS Services.
- Used various AWS services including S3,EC2, AWS Glue, Athena, RedShift, EMR,SNS, SQS, DMS, Kenesis.
- Extracted data from multiple source systems $3, Redshift, RDS and Created multiple tables/databases in Glue Catalog by creating Glue Crawlers.
- Created AWS Glue crawlers for crawling the source data in S3 and RDS.
- Created multiple Glue ETL jobs in Glue Studio and then processed the data by using different transformations and then loaded into S3, Redshift and RDS.
- Created multiple Recipes in Glue Data Brew and then used in various Glue ETL Jobs.
- Design and Develop ETL Processes in AWS GIue to migrate data from external sources like S3, Parquet/Text Files into AWS Redshift.
- Used AWS glue catalog with crawler to get the data from S3 and perform SQL query operations using AWS Athena.
- 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.
- Used AWS Glue for transformations and AWS Lambda to automate the process.
- Designed and implemented a secure and scalable AWS architecture for the migration, leveraging services like Amazon EC2, Amazon RDS, and Amazon S3.
- Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs using CloudWatch.
- Performed end- to-end Architecture & implementation assessment of various AWS services like Amazon EMR, Redshift and S3.

- Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
- To analyze the data Vastly used Athena to run multiple queries on processed data from Glue ETL Jobs and then used Quick Sight to generate Reports for Business Intelligence.
- Used AWS EMR to transform and move large amounts of data into and out of AWS $3.
- Used DMS to migrate tables from homogeneous and heterogenous DBs from On-premise to AWS Cloud.
- Created Kinesis Data streams, Kinesis Data Firehose and Kinesis Data Analytics to capture and process the streaming data and then output into S3, Dynamo DB and Redshift for storage and analyzation.
- Created Lambda functions to run the AWS Glue job based on the AWS S3 events.
- Utilized Jira reporting and dashboards to provide stakeholders with real-time project status updates and performance metrics.

Environment: AWS Glue, S3, IAM, EC2, RDS, Redshift, EC2, Lambda, Boto3, DynamoDB, Apache Spark, Kinesis, Athena, Hive, Sqoop, Python.

Client: - Barclays- New York, NY **01/2019 - 11/2020**
Role:- Senior Data Engineer

- I was responsible for analyzing the business requirement and estimating the tasks and preparing the design documents for the existing Abinitio and Teradata code for converting into hive/spark SQL.
- Develop the Spark Sal logics which mimics the Teradata ETL logics and point the output Delta back to Newly Created Hive Tables and as well the existing TERADATA Dimensions, Facts, and Aggregated Tables.
- Imported data from Abinitio LDR(Load Ready Files) and into Spark RDD and performed transformations and actions on RDD's.
- Experienced in designing and deployment of Hadoop cluster and different big data analytic tool including Pig, Hive, Flume, HBase and Sqoop.
- Primary responsibilities include building scalable distributed data solutions using Hadoop ecosystem.
- Loaded the CDRs from relational DB using Sqoop and other sources to Hadoop cluster by Flume.
- Implementing quality checks and transformations using Spark.
- Developed simple and complex MapReduce programs in Hive, Pig and Python for Data Analysis on
- different data formats.
- Performed data transformations by writing MapReduce and Pig scripts as per business requirements.
- Implemented Map Reduce programs to handle semi/unstructured data like xml, Json, Avro data files and sequence files for log files.
- Primary responsibilities include building scalable distributed data solutions using Hadoop ecosystem.
- I was responsible for analyzing the business requirement and estimating the tasks and preparing the design documents for the existing Abinitio and Teradata code for converting into hive/spark SQL.
- Develop the Spark Sal logics which mimics the Teradata ETL logics and point the output Delta back to Newly Created Hive Tables and as well the existing TERADATA Dimensions, Facts, and Aggregated Tables.
- Imported data from Abinitio LDR(Load Ready Files) and into Spark RDD and performed transformations and actions on RDD's.
- Experienced in designing and deployment of Hadoop cluster and different big data analytics tool
- including Pig, Hive, Flume, HBase and Sqoop.
- Loaded the CDRs from relational DB using Sqoop and other sources to Hadoop cluster by Flume.
- Implementing quality checks and transformations using Spark.
- Developed simple and complex MapReduce programs in Hive, Pig and Python for Data Analysis on different data formats.
- Performed data transformations by writing MapReduce and Pig scripts as per business requirements.
- Implemented Map Reduce programs to handle semi/unstructured data like xml, Json, Avro data files and sequence files for log files.
- Developed various Python scripts to find vulnerabilities with SQL Queries by doing SQL injection, permission checks and analysis.
- Experienced in Kerberos authentication to establish a more secure network communication on the cluster.

- Analyzed substantial data sets by running Hive queries and Pig scripts.
- Managed and reviewed Hadoop and HBase log files.
- Experience in creating tables, dropping and altered at run time without blocking updates and queries using Spark and Hive.
- Experienced in writing Spark Applications in Scala and Python.
- Used Spark SQL to handle structured data in Hive.
- Imported semi-structured data from Avro files using Pig to make serialization faster
- Processed the web server logs by developing Multi-hop ifume agents by using Avro Sink and loaded into MongoDB for further analysis.
- Experienced in converting Hive/SQL queries into Spark transformations using Spark RDD, Scala andPython.
- Experienced in connecting Avro Sink ports directly to Spark Streaming for analyzation of weblogs.
- Involved in making Hive tables, stacking information, composing hive inquiries, producing segments and basins for enhancement.
- Managing and scheduling Jobs on a Hadoop Cluster using UC4(Confidential preoperatory scheduling tool) workflows.
- Continuous monitoring and managing the Hadoop cluster through Hortonworks (HDP) distribution.
- Configured various views in Yarn Queue manager.
- Involved in review of functional and non-functional requirements.
- Indexed documents using Elastic search.
- Responsible for using Flume sink to remove the date from Flume channel and deposit in No-SQL database like MongoDB.
- Involved in loading data from UNIX file system and FT to HDFS.
- Developed workflow in Oozie to automate the tasks of loading the data into HDFS and pre-processing with Pig.
- Loaded JSON-Styled documents in NoSQL database like MongoDB and deployed the data in cloud service Amazon Redshift.
- Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in Amazon EMR, AZURE.
- Used Zookeeper to provide coordination services to the cluster.
- Created Hive queries that helped market analysts spot emerging trends by comparing fresh data with reference tables and historical metrics.
- Involved in migrating tables from RDBMS into Hive tables using SQOOP and later generated data visualizations using Tableau.
- Designed and implemented Spark jobs to support distributed data processing.
- Experience in optimizing Map Reduce Programs using combiners, partitioners and custom counters for delivering the best results.
- Written Shell scripts to monitor the health check of Hadoop daemon services and respond accordingly to any warning or failure conditions.
- Involved in Hadoop cluster task like Adding and Removing Nodes without any effect to running jobs and
- Followed Agile methodology for the entire project.
- Experienced in Extreme Programming, Test-Driven Development and Agile Scrum

Environment: Big Data Ecosystem (Hadoop, Hive, Pig, Spark, Sqoop), Data Ingestion, ETL, Data Analysis, Elastic Search, AWS, EC2, S3, Pig, Hive, Mysql, Python, MapReduce, Flume, Kerberos Authentication, Streaming, Hortonworks, NoSQL (MongoDB), Oozie, Query Optimization, Abinitio, Shell Scripting, Agile Methodology

Client: - Accenture INDIA **06/2013 - 12/2017**
Role:- Big Data Engineer

- Understand the requirements and prepared architecture document for the Big Data project.
- Worked with Horton Works distribution
- Supported MapReduce Java Programs those are running on the cluster.
- Optimized Amazon Redshift clusters, Apache Hadoop clusters, data distribution, and data processing

- Developed MapReduce programs to process the Avro files and to get the results by performing some calculations on data and also performed map side joins.
- Imported Bulk Data into Base Using MapReduce programs.
- Used Rest Apl to Access HBase data to perform analytics.
- Designed and implemented Incremental Imports into Hive tables.
- Involved in creating Hive tables, loading with data and writing Hive queries that will run internally in MapReduce way
- Involved in collecting, aggregating and moving data from servers to HDFS using Flume.
- Imported and Exported Data from Different Relational Data Sources like DB2, SQL Server, Teradata to HDFS using Sqoop.
- Migrated complex map reduce programs into in memory Spark processing using Transformations and actions.
- Worked on POC for IOT devices data, with spark.
- Used SCALA to store streaming data to HDFS and to implement Spark for faster processing of data.
- Worked on creating the RDD's, DF's for the required input data and performed the data transformations using Spark Python.
- Involved in developing Spark SQL queries, Data frames, import data from Data sources, perform transformations, perform read/write operations, save the results to output directory into HDFS. ten Hive jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
- Developed PIG UDF'S for manipulating the data according to Business Requirements and also worked on developing custom PIG Loaders.
- Worked on Oozie workflow engine for job scheduling.
- Developed Oozie workflow for scheduling and orchestrating the ETL process.
- Experienced in managing and reviewing the Hadoop log files using Shell scripts.
- Migrated ETL jobs to Pig scripts to do Transformations, even joins and some pre-aggregations before storing the data onto HDFS.
- Worked on different file formats like Sequence files, XML files and Map files using MapReduce Programs.
- Worked with Avro Data Serialization system to work with JSON data formats.
- Used AWS S3 to store large amounts of data in identical/similar repository.
- Involved in build applications using Maven and integrated with Continuous Integration servers like Jenkins to build jobs.
- Used Enterprise Data Warehouse database to store the information and to make it access all over organization.
- Responsible for preparing technical specifications, analyzing functional Specs, development and maintenance of code.
- Worked with the Data Science team to gather requirements for various data mining projects
- Wrote custom Python scripts for various data processing and automation tasks.
- Managed version control using Git for code collaboration and codebase maintenance.

Environment: Big Data Architecture, HortonWorks, MapReduce, Amazon Redshift Optimization, Hadoop Clusters, Avro Data Processing, HBase, Hive, Flume, Sqoop, Spark, Scala, RDDs, Data Frames, Spark SQL, Pig, Oozie, Shell Scripting, ETL, Data Serialization (Avro), AWS S3, Maven, Data Warehousing, Technical Specification, Data Mining, Shell Script Automation.

Education

Master of Science

Texas A&M University - Commerce
Commerce, TX

Manoj Kumar

- Irving, TX, US

Contact Information

- 7si-w40-1ma@mail.dice.com
- 4696299272

Work History

Total Work Experience: 11 years

- **Senior Data Engineer Triwest Healthcare Alliance**
  Aug 01, 2022
- **Data Engineer Sonder Inc**
  Nov 01, 2020
- **Data Engineer Barclays**
  Nov 01, 2019
- **Data Analyst Accenture**
  Jun 01, 2013

Skills

- **testing** - 6 years
- **business requirements** - 31 years
- **data engineering** - 31 years
- **python** - 31 years
- **software** - 31 years
- **sql** - 31 years
- **reporting** - 28 years
- **analytics** - 27 years
- **apache spark** - 27 years

- **data modeling** - 27 years
- **microsoft sql server** - 27 years
- **microsoft windows azure** - 27 years
- **database** - 24 years
- **jenkins** - 24 years
- **data governance** - 23 years
- **data science** - 23 years
- **tableau** - 20 years
- **data processing** - 19 years
- **data visualization** - 19 years
- **etl** - 19 years
- **meta-data management** - 19 years
- **decision-making** - 5 years

Work Preferences

- Desired Work Settings: No Preference
- Likely to Switch: False
- Willing to Relocate: True
- Travel Preference: 0%
- Work Authorization:
    - US
- Work Documents:
    - Have H1 Visa
- Security Clearance: False
- Third Party: True
- Employment Type:
    - Contract - Corp-to-Corp
    - Contract to Hire - Corp-to-Corp

Profile Sources

- Dice: https://www.dice.com/employer/talent/profile/04570efca6bd0a3757609c279d9d9abf