M.Manisha
Role : GCP Data Engineer
Manisha1992mano@gmail.com
Phno : 510-730-1664
https://www.linkedin.com/in/manisha-manyam-03baa32a6/

Professional Summary:

- 10+ years of IT experience in a variety of industries working on Big Data technology using technologies such as **Cloudera** and Hortonworks distributions and Web Programming using Java and **Big data** technologies.
- Hadoop working environments include Hadoop, Spark, Map Reduce, Kafka, Hive, Ambari, Sqoop, HBase, and Impala.
- Hands-on experience in developing and deploying enterprise-based applications using major Hadoop ecosystem components like Map Reduce, YARN, Hive, HBase, Flume, Sqoop, Spark MLlib, Spark GraphX, Spark SQL, Kafka.
- Proven expertise in deploying major software solutions for various high-end clients meeting business requirements such as **Big Data** Processing, Ingestion, Analytics and Cloud Migration from On-prem to **Cloud**.
- Adept at configuring and installing Hadoop/Spark Ecosystem Components.
- Proficient in **cloud-based data technologies**, including **GCP** services such as **Big Query, Dataflow, and Pub/Sub**.
- Worked with Spark to improve efficiency of existing algorithms using Spark Context, Spark SQL, Spark MLlib, Data Frame, Pair RDD's and Spark YARN.
- Experience in application of various data sources like Oracle SE2, SQL Server, Flat Files and Unstructured files into data warehouses.
- Strong knowledge of data warehousing concepts and methodologies, including dimensional modeling and ETL processes.
- Proficient with Spark Core, Spark SQL, Spark MLlib, Spark GraphX and Spark Streaming for processing and transforming complex data using in-memory computing capabilities written in **Scala**.
- Familiarity with data governance principles and practices, ensuring compliance and data security.
- Proficient in data visualization tools, such as **Tableau or Power BI**, to create insightful reports and dashboards.
- Able to use Sqoop to migrate data between RDBMS, **NoSQL** databases and HDFS.
- Experience in Extraction, Transformation and Loading (**ETL**) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating and moving data from various sources using Apache Flume, Kafka, PowerBI and Microsoft SSIS.
- Hands-on experience with Hadoop architecture and various components such as Hadoop File System HDFS, Job Tracker, Task Tracker, Name Node, Data Node and **Hadoop** Map Reduce programming.
- Comprehensive experience in developing simple to complex Map reduce and Streaming jobs using Scala and Java for data cleansing, filtering and data aggregation. Also possess detailed knowledge of Map Reduce framework.
- Used IDEs like Eclipse, IntelliJ IDE, PyCharm IDE, Notepad++, and Visual Studio for development.
- Seasoned practice in Machine Learning algorithms and Predictive Modeling such as Linear Regression, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, KNN, Neural Networks, and K-means clustering.
- Ample knowledge of data architecture including data ingestion pipeline design, Hadoop/Spark architecture, data modeling, data mining, machine learning and advanced data processing.
- Experience working with **NoSQL** databases like Cassandra and HBase and developed real-time read/write access to very large datasets via HBase.
- Developed Spark Applications that can handle data from various RDBMS (**MySQL**, Oracle Database) and Streaming sources.
- Ample knowledge on Apache Kafka, Apache Storm to build data platforms, pipelines, and storage systems; and search technologies such as Elastic search.
- Proficient with complex workflow orchestration tools namely Oozie, Airflow, Data pipelines and Azure Data Factory, Cloud Formation & Terraforms.
- Implemented Data warehouse solution consisting of ETLS, On-premises to **Cloud Migration** and good expertise building and deploying batch and streaming data pipelines on cloud environments.

- Worked on Airflow 1.8(Python2) and Airflow 1.9(Python3) for orchestration and familiar with building custom Airflow operators and orchestration of workflows with dependencies involving multi-clouds.
- Spark for ETL follower, Data bricks Enthusiast, Cloud Adoption & Data Engineering enthusiast in Open-source community.
- Proficient with Azure Data Lake Services (ADLS), Data bricks & iPython Notebooks formats, Data bricks Delta lakes & Amazon Web Services (**AWS**).
- Orchestration experience using Azure Data Factory, Airflow 1.8 and Airflow 1.10 on multiple cloud platforms and able to understand the process of leveraging the Airflow Operators.
- Knowledge in automated deployments leveraging Azure Resource Manager Templates, DevOps, and Git repository for Automation and usage of Continuous Integration (**CI/CD).**
- Experienced in data processing and analysis using Spark, HiveQL, and SQL.
- Extensive experience in Writing User-Defined Functions (UDFs) in Hive and Spark.
- Worked on **Apache Sqoop** to perform importing and exporting data from HDFS to **RDBMS/NoSQL DBs** and vice-versa.
- Working experience on **NoSQL** databases like **HBase, Azure, MongoDB**, and **Cassandra** with functionality and implementation.
- Worked extensively over semi-structured data (fixed length & delimited files) for data sanitation, report generation, and standardization.
- Excellent understanding of **Zookeeper** for monitoring and managing Hadoop jobs.
- Experience with NumPy, Matplotlib, Pandas, Seaborn, and Plotly Python libraries. Worked on large datasets by using PySpark, NumPy, and pandas.
- Utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis.
- Experience in HANA SQL Script Stored Procedures, Table functions, and dynamic privileges by **SQL** query in information models.
- Experience in using Kafka and Kafka brokers to initiate spark context and processing live streaming.
- Extensive experience across both relational databases and non-relational databases Oracle, **PL/SQL, SQ**L Server, **MySQL**, and DB2.
- Holds a strong ability to handle multiple priorities and workloads and can understand and adapt to new technologies and environments faster.

Technical Skills:

Hadoop/Spark Ecosystem	Hadoop, Map Reduce, Pig, Hive/impala, YARN, Kafka, Flume, Oozie, Zookeeper, Spark, Airflow, MongoDB, Cassandra, HBase, and Storm.
Hadoop Distribution	Cloudera distribution and Horton works
Programming Languages	Scala, Hibernate, JDBC, JSON, HTML, CSS, SQL, R, Shell Scripting.
Script Languages:	JavaScript, jQuery, Python.
Databases	Oracle, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL, HBase, MongoDB
Cloud Platforms	GCP, Azure, AWS.
Distributed Messaging System	Apache Kafka
Data Visualization Tools	Tableau, Power BI, Looker Pro, SAS, Excel, ETL
Batch Processing	Hive, Map Reduce, Pig, Spark
Operating System	Linux (Ubuntu, Red Hat), Microsoft Windows
Reporting Tools/ETL Tools	Informatica Power Centre, Tableau, Pentaho, SSIS, SSRS, Power BI

Professional Experience:

Kroger, Cincinnati, Ohio **Nov 2022 - Current**

GCP Data Engineer

Responsibilities:

- Leveraged **GCP** services for scalable data infrastructure, storage, processing, and analytics, ensuring efficient handling of large datasets and complex analytics tasks.
- Built and deployed data pipelines using **Cloud Compose**r and **Cloud Function**s, enabling seamless integration with other GCP services such as **BigQuery**, **Pub/Sub**, and Cloud Storage.
- Demonstrated expertise in leveraging GCP services such as **Compute Engine**, **Kubernetes Engine**, **Cloud Storage**, **BigQuery**, and **Cloud SQL** for seamless migration and efficient operation of workloads.
- Utilized **Cloud Data Fusion** for building and managing data pipelines, simplifying integration and processing tasks with its graphical interface and pre-built connectors.
- Work related to downloading **BigQuery** data into **pandas** or **Spark** data frames for advanced **ETL** capabilities.
- Implemented **CI/CD** practices for automating data pipeline deployment, ensuring reliability and consistency in pipeline operations, and facilitating frequent updates.
- Defined migration strategies, including lift-and-shift, re-platforming, and re-architecting, based on specific business and technical requirements, and conducted risk assessments with mitigation strategies for potential challenges during migration.
- Utilized Infrastructure as Code (IaC) tools like **Terraform** or Deployment Manager to automate the provisioning of **GCP** resources, streamlining the migration process, and developed custom scripts and automation tools for optimization and acceleration of migration tasks.
- Implemented monitoring and alerting mechanisms using Stackdriver for proactive issue identification and resolution in **GCP** data pipelines.
- Designed and executed end-to-end testing strategies for **GCP** data pipelines, ensuring data accuracy and completeness from ingestion to analysis.
- Employed **DevOps** practices and tools such as **Jenkins**, **Terraform**, and **Ansible** to automate **GCP** infrastructure deployment and configuration, resulting in a 50% reduction in deployment time.
- Moreover, I have expertise in **Terraform** key features such as Infrastructure as Code, Execution plans, and Resource Graphs. I have also created and automated Windows-based tasks, scripts, and processes using PowerShell and utilized **Ansible** for configuration management, application deployment, and task automation.
- I have experience implementing Puppet for configuration automation and infrastructure management to ensure system consistency and compliance across diverse environments.
- I have designed and developed **Spark** jobs with **Scala** for end-to-end data pipelines for batch processing, utilized **Flume**, **Kafka**, and **Spark Stream** for data ingestion and transformation, developed data validation scripts in **Hive** and **Spark**, and performed validation using **Jupyter Notebook**.
- Implemented spark programs/applications in **Scala** using Spark APIs for Data Extraction, Transformation, and Aggregation.
- I have also utilized **PySpark** for data extraction, loading, and performing **SQL** queries, developed PySpark scripts for data encryption using hashing algorithms, and was responsible for the design, development, and testing of cloud databases.
- Proficiently worked with both **SQL** and **NoSQL** databases for various data storage and retrieval tasks, selecting the appropriate technology based on specific data requirements.
- I have developed Python-based **APIs (RESTful Web Services)** for revenue tracking and analysis.
- Developed visually compelling reports and dashboards using **PowerBI** and **Tableau**, incorporating advanced features such as custom visuals and natural language(NLP) queries to provide actionable insights to business users across the organization.
- Applied **R programming** language to create advanced statistical graphics and visualizations, including heatmaps, scatter plots, and time series plots, facilitating in-depth data analysis and hypothesis testing for actionable insights.

Environment: *GCP, G-Cloud Function,Cloud Data Fusion, Cloud Dataflow, Cloud Shell, Cloud SQL, Big Query, Cloud Dataproc CI/CD (Continuous Integration/Continuous Deployment), Pub/Sub, GCS, Cloud SQL, Cloud Composer, Talend for Big Data, Airflow, Hadoop, Hive, SAS, Spark, Python, SQL Server, Kubernetes, Terraform, R, PowerBI, Tableau.*

Discover Financial, Riverwoods, Illinois **Mar 2020 – Oct 2022**

GCP Data Engineer
Responsibilities:

- Conducted in-depth analysis of business domain data, ensuring a deep understanding of its usage for metrics and analytical data designs.
- Utilized expert **SQL** skills and data exploration techniques to perform data profiling and analysis, facilitating data-driven decision-making.
- Collaborated with business analysts to gather domain data requirements and translate them into detailed design deliverables, including source-to-target mapping documentation for data engineering.
- Leveraged Erwin for data warehouse data modeling, creating comprehensive **Data Flow, ER Diagram, Conceptual, Logical, and Physical** data models.
- Build data pipelines in airflow in **GCP** for **ETL** related jobs using different airflow operators.
- Demonstrated a thorough understanding of data flows, data taxonomy, organization, and data lineage, ensuring data integrity and quality throughout the architecture.
- Designed end-to-end data solutions encompassing both batch and real-time processing, enabling efficient data processing and analysis.
- Implemented metadata and documentation management practices for **Erwin** modeling and data cataloging, ensuring accurate and up-to-date documentation of the data environment.
- Created high-level and detailed data architecture design documentation, providing clear and comprehensive guidelines for data solution implementation.
- Collaborated closely with clients, developers, and architecture teams to understand requirements and translate them into practical and effective data solutions.
- Effectively communicated complex technical concepts to non-technical stakeholders through excellent verbal and written communication skills.
- Migrated an entire oracle database to **BigQuery** and using **Power BI** for reporting.
- Integrate data visualization tools like **Tableau** and **Looker Pro** with data sources to create interactive dashboards and reports for business stakeholders.
- Use data visualization tools to communicate insights, trends, and key performance indicators derived from data analysis and processing activities.
- Hands-on experience in **GCP Dataproc**, **GCS, Cloud functions**, **BigQuery** and moving data between **GCP** and **Azure** using **Azure Data Factory**.
- Experience in building **Power BI** reports on Azure Analysis services for better performance.
- Used **cloud shell** SDK in **GCP** to configure the services **Data Proc**, Storage, **BigQuery**
- Coordinated with team and Developed framework to generate Daily adhoc reports and Extracts from enterprise data from **BigQuery**.
- Designed and Co-ordinated with the Data Science team in implementing Advanced Analytical Models in Hadoop Cluster over large Datasets.
- Wrote scripts in **Hive SQL** for creating complex tables with high performance metrics like partitioning, clustering and skewing
- Work related to downloading **BigQuery** data into pandas or Spark data frames for advanced **ETL** capabilities.
- Worked with Google data catalog and other Google cloud API's for monitoring, query and billing related analysis for BigQuery usage.
- Worked on creating **POC** for utilizing the **ML** models and **Cloud ML** for table Quality Analysis for the batch process.
- Knowledge about cloud dataflow and **Apache** beam.
- I have Good knowledge in using cloud shell for various tasks and deploying services.
- Created **BigQuery** authorized views for row level security or exposing the data to other teams.
- Expertise in designing and deployment of Hadoop cluster and different Big Data analytic tools including **Pig**, **Hive**, **SQOOP**, **Apache Spark**, with Cloudera Distribution.
- Intensively used **Python**, **JSON** scripts coding to deploy the Stream Sets pipelines into the server.
- Build pipeline solutions to integrate data from multiple heterogeneous systems using Stream Sets data collectors.
- Responsible for maintaining quality reference data in Oracle by performing operations such as cleaning,

transformation, and ensuring Integrity in a relational environment.
- Written shell scripts to extract data from Unix servers into **Hadoop** HDFS for long-term storage.

Environment: *GCP, Machine learning, SQL, Oracle, PL/SQL, GCP Cloud Storage,Cloud shell, Cloud ML, Big Query, Azure Data Factory, GCP Dataproc, GCS, Cloud functions, Power BI, Apache Spark, Tableau and Looker Pro, Airflow.*

Target, Minneapolis, Minnesota **Jun 2018 - Feb 2020**
Data Engineer
Responsibilities:

- Collaborated with business stakeholders to gain a deep understanding of the business domain data and its utilization for metrics and analytical data designs.
- Conducted extensive data exploration and analysis to identify data relationships and ensure accurate data profiling using expert **SQL** skills and data exploration techniques.
- Applied strong data design experience and knowledge to architect big data architectures and cloud data lake environments, specifically utilizing **Teradata** and **GCP** platforms for curation and analytics use cases.
- Ensured a deep understanding of data flows, data taxonomy, organization, and data lineage for effective data governance and management.
- Architected end-to-end data solutions encompassing both batch and real-time designs, leveraging appropriate technologies and frameworks.
- Used Pandas in Python for Data Cleansing and validating the source data.
  Designed and developed ETL pipeline in Azure cloud which gets customer data from **API** and processes it to Azure **SQL DB**.
- Orchestrated all Data pipelines using **Azure Data** Factory and built a custom alerts platform for monitoring.
  Created custom alerts queries in Log Analytics and used Web hook actions to automate custom alerts.
- Created Data bricks Job workflows which extracts data from SQL server and upload the files to **sftp** using **PySpark** and **Python**.
- Used **Azure** Key vault as central repository for maintaining secrets and referenced the secrets in **Azure Data Factor**y and also in Data bricks notebooks
- Managed metadata and documentation through Erwin modeling and data cataloging, ensuring accurate and up-to-date information.
- Developed high-level and detailed data architecture design documentation, encompassing the entire data ecosystem.
- Collaborated with clients, developers, and architecture teams to gather requirements, design data solutions, and oversee implementation.
- Communicated effectively through strong verbal and written skills, facilitating clear and concise communication across teams and stakeholders.
- Involved in developing Spark scripts for data analysis in both **Python** and **Scala**. Designed and developed various modules of the application with **J2EE** design architecture.
- Implemented modules using **Core Java API**s, Java collection and integrating the modules.
- Experienced in transferring data from different data sources into **HDFS** systems using Kafka producers, consumers and Kafka brokers.
- Installed Kibana using salt scripts and built custom dashboards that can visualize aspects of important data stored by Elastic search.
- Analyzed data where it lives by Mounting **Azure Data Lake** and Blob to Data bricks.
- Used Logic App to take decisional actions based on the workflow.
- Developed custom alerts using **Azure** Data Factory, **SQLDB** and Logic App
- Used File System Check (FSCK) to check the health of files in HDFS and used Sqoop to import data from SQL server to **Cassandra**.
- Streaming the transactional data to **Cassandra** using Spark Streaming/Kafka
- Implemented a distributed messaging queue to integrate with Cassandra using Apache Kafka and Zookeeper.
- ConfigMap and Daemon set files to install File beats on Kubernetes PODS to send the log files to Log stash or Elastic search to monitor the different types of logs in **Kibana**.
- Created Database in Influx DB also worked on Interface, created for Kafka also checked the measurements on Databases.

- Installed Kafka manager for consumer lags and for monitoring Kafka Metrics also this has been used for adding topics, Partitions etc. Successfully Generated consumer group lags from Kafka using their API.
- Used **Oozie** and **Zookeeper** operational services for coordinating cluster and Scheduling workflows.
- Implemented **Flume**, **Spark**, and **Spark** Streaming framework for real time data processing.

Environment: *Hadoop, Python, HDFS, Hive, Scala, Map Reduce, Agile, Cassandra, Kafka, Storm, Azure, YARN, Spark, ETL, Teradata, NoSQL, Oozie, Java, Cassandra, Talend, LINUX, Kibana, HBase*

Liberty Mutual, Boston, MA **Mar 2016 - May 2018**

ETL Developer / Teradata Developer

Responsibilities:

- Maintain and support the Teradata architectural environment for EDW Applications.
- Interacted with the business community and gathered requirements based on changing needs.
- Designed the logical Data Model using Erwin and transformed the Logical model to Physical database using Power Designer. Data Mart and Dimensional Modeling, Star and **Snowflake** Schema Modeling for Data Warehouse.
- Developed mappings/scripts to extract data from **Oracle, Flat files, SQL Server, DB2** and load into data warehouse using the Mapping Designer, **BTEQ**, Fast Load and Multiload.
- Exported data from Teradata database using Fast Export and **BTEQ**.
- Wrote **SQL Queries**, Triggers, Procedures, Macros, Packages and Shell Scripts to apply and maintain the Business Rules. Coded and implemented PL/SQL packages to perform batch job scheduling.
- Performed Teradata and Informatica tuning to improve the performance of the Load.
- Performed error handling using error tables and log files.
- Used Informatica Designer to create complex mappings using different transformations like Filter, Router, Connected & Unconnected lookups, Stored Procedure, Joiner, Update Strategy, Expressions and Aggregator transformations to pipeline data to Data Warehouse.
- Performed DML and DDL operation with the help of **SQL** transformation in Informatica.
- Collaborated with Informatica Admin in the process of Informatica Upgradation from Power Center 7.1 to Power Center 8.1. Used **SQL** Transformation to sequential loads in Informatica Power Center for ETL processes.
- Worked closely with the business analyst's team to solve the Problem Tickets, Service Requests. Helped the 24/7 Production Support team.
- Developed mappings in Informatica to load the data from various sources into the Data Warehouse, using different transformations like Source Qualifier, Expression, Lookup, aggregate, Update Strategy, and Joiner.
- Worked on Informatica Advanced concepts & also Implementation of Informatica Push down Optimization technology and pipeline partitioning.
- Performed bulk data load from multiple data sources (ORACLE 8i, legacy systems) to **TERADATA RDBM**S using BTEQ, Multiload and Fast Load.
- Used various transformations like Source qualifier, Aggregators, lookups, Filters, Sequence generators, Routers, Update Strategy, Expression, Sorter, Normalizer, Stored Procedure, Union etc.
- Used Informatica Power Exchange to handle the change data capture (CDC) data from the source and load into Data Mart by following slowly changing dimensions (SCD) type II process.

Environment: Teradata, Oracle, DB2, SQL, RDBMS, Snowflake

Menlo Technologies Inc, Hyderabad, India **Nov 2013 - Jan 2016**

Data Analyst

Responsibilities:

- Work with Project Management in the creation of project estimates.
- Analysis of the data identifying source of data and data mappings of HCFG.
- Worked extensively in documenting the Source to Target Mapping documents with data transformation logic.
- Interact with the SMEs to analyze the data extracts from Legacy Systems Mainframes and COBOL Files and determine the
- element source, format and its integrity within the system
- Transformation of requirements into data structures which can be used to efficiently store, manipulate and retrieve

information.

- Collaborate with data modelers and **ETL** developers in creating Data Functional Design documents.
- Ensure that models conform to established best practices including normalization rules and accommodate change in a cost effective and timely manner.
- Enforce standards to ensure that the data elements and attributes are properly named.
- Work with the business and the **ETL** developers in the analysis and resolution of data related problem tickets.
- Support development teams creating applications against supported databases.
- Provide 24 x 7 problem management support to the development team.
- Document various Data Quality mapping documents, audit and security compliance adherence.
- Perform small enhancements SOR element additions, data cleansing/data quality.
- Create various Data Mapping Repository documents as part of Metadata services EMR.
- Provide inputs to the development team in performing extraction, transformation and load for data marts and data warehouses.
- Provide support in developing and maintaining ETL processes that extract data from multiple SOR's residing on various technology platforms then transport the data to various delivery points such as data marts or data warehouses. Collaborate with data modelers and ETL developers in creating Data Functional Design documents.

Environment: *MS Excel, MS Access, Oracle 10g, UNIX, Windows XP, SQL, PL/SQL*