Priyanka

padinampriya@gmail.com

(617) 708-2803

Big Data Engineer

**PROFESSIONAL SUMMARY:**

● Over 10+ years of experience as a Big Data Engineer and ETL developer, specializing in designing, developing, and implementing data models for enterprise-level applications. Knowledgeable about the architecture and components of Spark, with expertise in Spark Core, Spark SQL, and Spark Streaming.

● Extensive knowledge of AWS services such as EMR, Redshift, S3, EC2, including server configuration for auto-

scaling and elastic load balancing. Skilled in utilizing AWS S3 for data staging, transfer, and archival, with hands- on experience in large-scale data migrations using AWS DMS and implementing CDC in AWS Redshift.

● Proficient in writing Python scripts, optimizing SQL queries in Oracle DB2, SQL Server, and Teradata. Demonstrates

strong familiarity with the Software Development Life Cycle (SDLC) and a comprehensive understanding of testing methodologies.

● Adapt at Data Analysis, Data Validation, Data Cleansing, and Data Verification, with a keen eye for identifying

data discrepancies. Conducted data analysis and profiling using complex SQL on various systems, including

Oracle and Teradata.

● Executed data extraction, transformation, and loading from source systems to Azure Data Storage using Azure

Data Factory, T-SQL, and Spark SQL ,Snowflake and DB2. Well-versed in Big Data technologies, Hadoop, Spark, Python, SQL, Power

BI, Tableau, and other Data Engineering tools on both Azure and AWS platforms.

● Hands-on experience with Azure Cloud Technologies, including Azure SQL Database, Azure Data Factory, Azure Databricks, Azure Data Lake store, Azure Synapse Analytics, Azure storage accounts, Azure Key Vault, and IoT Hub.

● Expertise in Data Migration, Data Profiling, Data Ingestion, and Data Cleansing using ETL tools such as Informatica Power Centre. Proficient in Spark RDD, Data frame API, Data set API, Data Source API, Spark SQL, and Spark Streaming.

● Skilled in data exporting/importing using Sqoop between HDFS and Relational Database systems, with expertise in optimizing Hive Queries. Knowledgeable about partitions, bucketing concepts in Hive, and designing Managed and External tables for performance optimization.

● Experience with business intelligence and data warehouse software, including SSAS/SSRS/SSIS, Business

Objects, Amazon Redshift,DB2, Azure Data Warehouse, Snowflake, and Teradata.

● Proven expertise in creating real-time data streaming solutions using Apache Spark/Spark Streaming, Kafka, and

Flume.

● Responsible for implementing new SQL Server features to enhance query performance, including T-SQL

language enhancements, in-memory optimization, and performance tuning with Extended Events.

● Utilized Apache NiFi for automating data movement between different Hadoop Systems and demonstrated proficiency in handling messaging services using Apache Kafka.

● Experience in data mining and warehousing using ETL tools, coupled with proficiency in building reports and dashboards in Tableau.

● Worked concurrently in GCP and Azure clouds, with extensive hands-on experience in migrating on-premise ETLs to Google Cloud Platform (GCP) using native tools such as BigQuery, Cloud Data Proc, Google Cloud Storage, and Composer.

● Strong knowledge of job workflow scheduling and locking tools/services like Oozie and Zookeeper.

● Solid understanding and knowledge of NoSQL databases like HBase and Cassandra.

● Well-versed in Amazon Web Services (AWS) components, including EC2, Snowflake ,DB2,S3, EMR, Step functions, Lambda,

Redshift, and DynamoDB.

● Hands-on experience with Microsoft Azure services such as HDInsight Clusters, BLOB, ADLS, Data Factory, and

Logic Apps.The source system is tables from Claim Center of Guidewire, and Policy/Insurance management system.

● Experience in working with NoSQL databases like HBase and Cassandra.

● Proficient in using defect tracking/issue tracking/ Bug tracking tools like Atlassian Jira, and Bugzilla.

● Excellent interpersonal and communication skills, efficient time management, and organization skills, ability to handle multiple tasks, and work well in a team environment.

**CATEGORY TECHNOLOGIES**

Hadoop Components / Big

Data

PySpark, Airflow, HDFS, MapReduce, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Kafka, Yarn.

Programming Languages Python, Scala, SQL, PySpark, PowerShell, T-SQL. Cloud Platform - AWS (Lambda, S3, EC2, EMR, RDS)

- Microsoft Azure (Azure Databricks, Azure Data Factory, Azure Data

Explorer, Azure HDInsight, ADLS)

- GCP

Reporting and ETL Tools AWS GLUE, Tableau, Power BI.

Oracle, SQL Server, MS Access, NoSQL Database (HBase, MongoDB,

Databases

Big Data Technologies

DynamoDB)

Hadoop, HDFS, Hive, Oozie, Sqoop, Spark, Machine Learning, Pandas, NumPy, Seaborn, Impala, Zookeeper, Airflow, Informatica, Snowflake, Data Bricks, Kafka, Cloudera, Snowflake

Data Analysis Libraries Pandas, NumPy, SciPy, Scikit-learn, Matplotlib

Containerization Docker, Kubernetes CI/CD Tools Jenkins, GitLab, Bamboo. Software Methodologies Agile, Scrum, Waterfall

Eclipse, PyCharm, IntelliJ, SSMS, Microsoft Office Suite (Word, Excel,

Development Tools

PowerPoint)

Version Control Tools GitHub and Azure DevOps

Java, J2EE, RMI, Sockets, JDBC, Servlets, JSP, JMS, Java Beans, JSTL,

J2SE/J2EE Technologies

Jakarta Struts, JSF, EJB, Spring, Hibernate, JTA, JNDI, JPA, JMS.

**WORK EXPERIENCE:**

**Comcast, PA**

**Senior Data Engineer**

**Responsibilities:**

**March 2022 - Present**

● Collaborated with business/user groups to gather requirements, playing a key role in the creation and development of pipelines.

● Utilized Airflow for building batch data pipelines and Kafka for developing streaming data pipelines.

● Developed Spark Streaming scripts for real-time processing, consuming topics from Kafka distributed messaging sources.

● Created and managed Azure Data Factory, implementing policies and utilizing Blob storage for storage and backup on Azure.

● Ingested data from web services into Azure cloud, loading it to Azure SQL DB,DB2, Snowflake and developed front-end interfaces using various technologies.

● Employed PySpark to load high-volume files into PySpark DataFrames, processing and reloading them into Azure

SQL DB ,DB2 , Snowflake tables. Involved extensively in code and design of various Guide wire components like PCF’s, plug - ins, Workflows, Batch jobs, product model and different components of Guide wire Suite ● Implemented and maintained Iceberg-based data storage solutions, ensuring efficient and scalable storage of structured data for analytics purposes.

● Developed and optimized ETL processes leveraging Iceberg, facilitating the seamless extraction, transformation,

and loading of data from various sources into a structured format.

● Designed and developed pipelines using Databricks, Snowflake automating ETL processes and workload maintenance.

● Created ETL packages using SSIS to extract data from various sources and maintain data in SQL Server. Good knowledge of Relational Databases like SQL Server, Snowflake, Oracle, DB2.

● Utilized Azure Databricks for ETL operations, connecting to relational databases using Kafka, Snowflake and employed Informatica for session and workflow management.

● Implemented Infrastructure as Code (IaaC) using Terraform with Terragrunt for consistent and automated infrastructure provisioning.Implementing One time Data Migration of Multistate level data from SQL server to Snowflake by using Python and SnowSQL.

● Facilitated code collaboration by configuring and optimizing workflows within Bitbucket, enabling seamless

Performed day-to-day integration with the Database Administrators (DBA) DB2, SQL Server, Oracle DB2,and AWS Cloud teams toensure the insertion of database tables, columns and its metadata have been successfully implemented out to the DEV, QUAL and PROD region environments in AWS Cloud and Snowflake

integration with CI/CD pipelines.

● Led migration processes and upgrades for Bitbucket instances, ensuring a smooth transition and improved performance.

● Automated data ingestion into the Lakehouse, leveraging Apache Spark and storing data in Delta Lake.

● Ensured data quality and integrity using Azure SQL Database and automated ETL deployment.

● Utilized Data bricks, Scala,DB2 and Spark for data workflows, capturing data from Delta tables in Delta Lakes.

● Performed streaming of pipelines using Azure Event Hubs and Stream Analytics to analyze data from data-driven workflows.

● Oversaw and manage G Suite email services, including Gmail, ensuring reliable and secure email communication

within the organization.

● Developed and maintained interactive and visually appealing dashboards using Superset, providing stakeholders with insightful and accessible data visualizations.

● Collaborated with cross-functional teams to understand data requirements and translated them into effective

Superset charts and dashboards that supported data-driven decision-making.

● Configured and optimized Superset settings, ensuring seamless integration with various data sources and databases for efficient data exploration and analysis.

● Administered email filtering, routing, and archiving policies to enhance email security and compliance.

● Configured and optimized collaboration tools within G Suite, such as Google Drive, Google Calendar, Google

Docs, Sheets, and Slides, to enhance team productivity.

● Implemented and managed sharing settings to ensure secure and efficient collaboration on documents and projects.

● Worked with Delta Lakes for consistent unification of streaming, processed data, and implemented ACID

transactions using Apache Spark and Snowflake.

● Managed Azure Blob Storage, developing frameworks for the implementation of large volumes of data and

system files.

● Implemented and managed metadata repositories using a Metastore, ensuring efficient storage, retrieval, and organization of metadata associated with various data assets.

● Collaborated with data engineers and analysts to define and enforce metadata standards, ensuring consistency and

accuracy in data descriptions, classifications, and lineage.

● Conducted regular audits and quality checks on metadata entries in the Metastore, addressing discrepancies and maintaining data governance standards.

● Identified and implemented new SQL Server features to improve query performance, including T-SQL

enhancements, in-memory optimization, and performance tuning.

● Acted as the resident SME for performance tuning stored procedures, functions, T-SQL scripts, Snowflake, indexes, and SSIS packages.

● Planned and executed data migration strategies when transitioning to G Suite, ensuring a smooth transition of email, calendar, and other collaboration data.

● Integrated G Suite with other enterprise systems and applications to streamline workflows and data exchange.

● Implemented tasks and transformations for data cleansing and performance tuning of packages using SSIS.

● Deployed applications in a clustered SOA 12c environment, tuning SOA middleware performance.

● Implemented a distributed stream processing platform for low-latency integration with data and analytics services inside and outside Azure.

● Used PowerShell scripting for maintaining and configuring data, automating and validating data using Apache

Airflow.

● Implemented and managed AWS-based Continuous Integration/Continuous Deployment (CI/CD) pipelines, automating the build, test, and deployment processes for diverse applications.

● Developed and maintained shell scripts in Linux/Unix environments, streamlining repetitive tasks and enhancing system automation.

● Identified and resolved performance bottlenecks in Unix/Linux processes, conducted in-depth analysis of query plans, and implemented optimizations for enhanced system efficiency.

● Optimized Hive queries using best practices, parameters, and technologies such as Hadoop, YARN, Python, and

PySpark.

● Utilized Sqoop to extract data from Teradata into HDFS and export analyzed patterns back to Teradata.

● Worked on Kafka for data ingestion from various sources, storing it in HDFS for filtering.

● Used Accumulators and Broadcast variables to tune Spark applications and monitored created analytics and jobs.

● Tracked Hadoop cluster job performance, conducted capacity planning, and tuned Hadoop performance for high availability and recovery.

● Contributed to a Data Migration project, migrating data from different sources to Google Cloud Platform (GCP)

using the UDP framework and transforming data with Spark Scala scripts.

● Developed data ingestion processes to maintain a Global Data Lake on the GCP cloud and Big Query.

● Collaborated with Tableau for report generation, creating dashboards, pie charts, and heat maps as per business

requirements.

● Engaged in all phases of the Software Development Life Cycle and employed Agile methodology for development.

**Environment**: Python, SQL, Cassandra DB, Azure Data Lake Storage Gen 1, Azure Data Factory, Azure SQL DB, Spark, Databricks, SSIS, SQL Server, Kafka, Informatica, Apache Spark, Delta Lake, Azure Event Hubs, Stream Analytics,DB2, Snowflake

Azure Blob Storage, PowerShell, Apache Airflow, Hadoop, YARN, PySpark, Hive, Teradata, Sqoop, HDFS, Spark, Agile.

Amazon,WA

**Senior Data Engineer**

**Responsibilities:**

**May 2020 – Feb 2022**

● Led the complete Software Development Life Cycle (SDLC) process, analyzing business requirements and understanding functional workflows for information transfer from source to destination systems.

● Applied analytical, statistical, and programming skills to collect, analyze, and interpret large datasets, developing data-driven solutions using SQL and Python.

● Designed AWS EC2 instance architecture for high-availability application architecture and security parameters.

● Created AWS S3 buckets, managed policies, and utilized S3 buckets and Glacier for storage and backup. Worked

on Hadoop cluster and data querying tools for data storage and retrieval.

● Designed and developed SSIS Packages for importing and exporting data from MS Excel, SQL Server, and flat files.

● Designed and implemented data ingestion pipelines into Apache Druid, ensuring efficient and timely processing

of large-scale streaming and batch data.

● Configured and optimized Druid clusters to achieve high performance, low-latency queries, and effective data storage and retrieval.

● Collaborated with data scientists and analysts to design and implement Druid data models, facilitating interactive and exploratory analytics.

● Integrated Artifact Registry seamlessly into CI/CD pipelines, automating the retrieval and deployment of artifacts during the build and release processes.

● Implemented and maintained dependency management strategies within the artifact repository, facilitating the tracking and resolution of dependencies for development projects.

● Contributed to the design, development, and testing of J2EE components, including Java Beans, Java, XML,

Collection Framework, JSP, Servlets, and JDBC.

● Developed UI pages using HTML5, Bootstrap, CSS3, JavaScript, jQuery, AJAX, React.js, and Node.js.

● Worked on Data Integration for extracting, transforming, and loading processes within the designed packages.

● Designed and deployed automated ETL workflows using AWS Lambda, organized and cleansed data in S3 buckets using AWS Glue, and processed data using Amazon Redshift.

● Enhanced ETL architecture to improve performance using query optimizer.

● Implemented data extraction using Spark, Hive, and managed large datasets using HDFS.

● Worked on streaming data transfer, extracting data from various sources into HDFS and NoSQL databases.

● Created ETL mappings with Talend Integration Suite to pull data from sources, apply transformations, and load

data into the target database.

● Collaborated on projects involving Hadoop and Spark, gaining exposure to big data processing and analytics in distributed computing environments.

● Developed and optimized stored procedures in relational databases, enhancing data processing efficiency and

supporting application functionality.

● Applied relational database management system (RDBMS) knowledge to optimize data storage and retrieval, contributing to efficient data management practices.

● Actively contributed to the development and support of Java/Spring Boot applications, ensuring robust and scalable solutions for business needs.

● Optimized SQL queries using query analyzer for performance tuning.

● Scripted with Python in Spark for transforming data from various file formats like text files, CSV, and JSON.

● Loaded data from different relational databases like MySQL and Teradata using Sqoop jobs.

● Processed and tested data using Spark SQL and real-time processing with Spark Streaming and Kafka using

Python.

● Created detailed diagrams and visualizations for clear communication.

● Illustrated complex workflows, system architectures, and data flows.

● Managed Agile project processes, including sprint planning and task tracking.

● Ensured efficient collaboration and transparency within the development team.

● Collaborated on comprehensive project documentation and knowledge base articles.

● Maintained a centralized repository for team members.

● Utilized as a central hub for real-time communication and collaboration.

● Streamlined team interactions and enhanced responsiveness.

● Scripted using Python and PowerShell for setting up baselines, branching, merging, and automation processes across the process using GIT.

● Implemented ETL architecture enhancements, optimized workflows by building Directed Acyclic Graphs (DAGs) in Apache Airflow to schedule ETL jobs, and utilized additional components in Apache Airflow such as Pools, Executors, and multi-node functionality.

● Utilized various transformations in SSIS Dataflow, Control Flow using for loop Containers, and Fuzzy matching.

● Created SSIS packages for data conversion and produced advanced extensible reports using SQL Server

Reporting Services.

● Deployed applications to GCP using Spinnaker (rpm based), launched multi-node Kubernetes clusters in Google

Kubernetes Engine (GKE), and migrated dockerized applications from AWS to GCP.

**Environment**: Python, SQL, AWS EC2, AWS S3 buckets, Hadoop, PySpark, AWS Lambda, AWS Glue, Amazon Redshift, Spark Streaming, GCP, Apache Kafka, SSIS, Informatica, ETL, Hive, HDFS, NoSQL, Talend, MySQL, Teradata, Sqoop, PowerShell, GIT, Apache Airflow.

**State ofMaryland,MD**

**Data Engineer**

**Responsibilities:**

**Oct 2019 – April 2020**

● Developed a data set process for data mining and modeling, recommending enhancements for data quality, efficiency, and reliability.

● Executed Extract, Transform, and Load (ETL) processes from source systems to Azure Data Storage services

using Azure Data Factory, T-SQL, Spark SQL, and U-SQL in Azure Data Lake Analytics. Processed data in Azure

Databricks.

● Implemented Spring Model View Controller (MVC) Architecture and coded Java Beans.

● Utilized Web Logic, Apache Tomcat servers, and JBoss servers for deploying and configuring Java APIs and

front-end modules.

● Created pipelines in Azure Data Factory for ETL tasks, incorporating Linked Services, Datasets, and Pipelines for data extraction, transformation, and loading from various sources.

● Developed Spark applications with PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats, analyzing and transforming data to reveal insights into customer usage patterns.

● Wrote Hive Queries and developed Hive DDLs for data analysis in Hive warehouse using Hive Query Language

(HQL).

● Extracted and updated data into Hadoop Distributed File System (HDFS) using Sqoop Import from diverse sources like Oracle, Teradata, SQL Server, etc.

● Created Hive staging tables, external tables, and implemented Dynamic Partitioning, Static Partitioning, and

Bucketing.

● Installed and configured Hadoop MapReduce, Hive, HDFS, Pig, Sqoop, Flume, and Oozie on the Hadoop cluster.

● Worked with Microsoft Azure services such as HDInsight Clusters, BLOB, ADLS, Data Factory, and Logic Apps, conducting a Proof of Concept (POC) on Azure Data Bricks.

● Implemented Sqoop jobs for data ingestion from Oracle to Hive.

● Handled various file formats like delimited text files, clickstream log files, Apache log files, Avro files, JSON

files, and XML Files. Proficient in using columnar file formats like RC, ORC, and Parquet.

● Developed custom Unix/BASH SHELL scripts for pre and post validations of master and slave nodes before and

after configuring the name node and data nodes.

● Developed job workflows in Oozie for automating tasks related to loading data into HDFS.

● Implemented efficient file storage of big data using various formats like Avro, Parquet, and JSON, along with

compression methods like GZip and Snappy.

● Explored Spark for performance improvement and optimization of existing algorithms in Hadoop using Spark

Context, Spark-SQL, Data Frame, and Pair RDDs.

● Utilized Spark with both Python and Scala, as well as Spark SQL for faster testing and data processing.

● Applied various data modeling concepts, including star schema and snowflake schema.

● Extensively used Stash, Bit-Bucket, and GitHub for code control purposes.

● Migrated MapReduce jobs to Spark jobs for enhanced performance.

**Environment:** Azure Data Factory, T-SQL, Spark SQL, Azure Databricks, Sqoop, Hadoop MapReduce, Hive,

Pig, Flume, Oozie, Spring Model View Controller (MVC), Web Logic, Apache Tomcat, JBoss, HDInsight Clusters, Apache Hadoop (HDFS, MapReduce), Stash, Bit-Bucket, GitHub, Avro, Parquet, JSON, Unix/BASH SHELL scripting, Star Schema, Snowflake Schema, PySpark, XML-Files.

American express,NY

**Data Engineer**

**Responsibilities:**

**Mar 2018 – Sept 2019**

● Prepared specification documents (BRD / FRD) based on business rules provided by the Business.

● Created Business Process and Data Process Models using MS Visio.

● Demonstrated extensive expertise in developing Stored Procedures, Triggers, Functions, Cursors, Views, Materialize Views, and Analytical Functions.

● Executed ETL processes to prepare data for creating business analysis visuals, aiding the leadership team in

making informed business decisions.

● Managed and administered artifact repositories, overseeing the storage and versioning of artifacts in a secure and compliant manner.

● Orchestrated the storage and distribution of container images, ensuring seamless integration with containerized

applications and services.

● Orchestrated and managed complex workflows and pipelines using Cloud Composer, facilitating the efficient execution of data processing tasks.

● Developed and deployed serverless functions using Cloud Functions, enabling the execution of event-triggered, scalable, and cost-effective backend processes.

● Orchestrated and deployed containerized applications on a fully managed, serverless platform, optimizing scalability and resource utilization.

● Orchestrated and managed Apache Spark and Hadoop clusters for efficient data processing and analysis, ensuring

optimal performance and resource utilization.

● Worked on visuals such as Employee Info, project details, skills, leaves, calendar, and time sheet.

● Conducted a Proof of Concept (POC) for on-premise to Cloud Azure data migration.

● Created and shared dashboards within the organization, facilitating updates and edits based on business requirements.

● Developed system standards, architecture, scenarios, detailed screen specifications, and documented logical and physical data models.

● Implemented various DAX functions for efficient data visualization in Power BI.

● Utilized Power BI gateway to maintain dashboards and reports up-to-date with on-premise data.

● Created Power BI dashboards and generated reports using SQL server tables as the source, incorporating locked objects on the reports.

● Performed Incremental load with several Dataflow tasks and Control Flow Tasks using SSIS.

● Deployed, tested, and scheduled SSIS packages.

● Prepared layouts by placing fields in appropriate positions according to the final report requirements.

● Conducted data extraction, transformation, and loading (ETL) between systems using SQL tools such as SSIS. **Environment**: Power BI, SQL Server, Azure, Google Big Query, Tableau, Excel, MS SQL Server, Server Integration Services (SSIS), SQL Server Reporting Services (SSRS), SQL Server Analysis Services (SSAS), SQL Server Management Studio (SSMS), and Oracle 11g.

**Sunmicro Technologies India**

**Data Engineer**

**Responsibilities:**

**Sep 2016 -Feb 2018**

● Architect, design, and implement robust and scalable data pipelines for ingesting, processing, and transforming large volumes of data using technologies like Apache Spark and AWS Glue.

● Collaborate with cross-functional teams to understand data requirements and ensure the seamless flow of

information between different systems.

● Manage and optimize Amazon Redshift clusters for efficient data storage and retrieval, ensuring high performance and responsiveness for analytical queries.

● Design and implement data models for the data warehouse, adhering to best practices for data warehousing and dimensional modeling.

● Establish and enforce data quality standards by implementing data validation processes and conducting regular audits to identify and rectify inconsistencies in data stored across different platforms.

● Utilize big data technologies such as Apache Cassandra to manage and store large-scale distributed datasets efficiently.

● Implement solutions for real-time data processing and analysis, leveraging technologies like Apache Kafka for streaming data.

● Identify and address performance bottlenecks in data processing and storage, optimizing queries, indexing strategies, and partitioning schemes to enhance overall system performance.

● Implement comprehensive monitoring and alerting systems for data-related processes and infrastructure components.

● Document data engineering processes, including data flows, transformations, and system configurations, using

tools like Confluence.

● Conduct knowledge-sharing sessions with team members to enhance overall understanding of data engineering best practices and methodologies.

● Implement and enforce data security measures, ensuring compliance with relevant regulations and industry

standards.

● Collaborate with security teams to implement encryption, access controls, and auditing mechanisms to safeguard sensitive data.

● Explore and implement innovative solutions to enhance the efficiency and capabilities of data processing and

analytics systems.

**Environment:** AWS (EMR/EC2/S3/Redshift), Apache Spark, AWS Glue, Apache Cassandra, Apache Kafka, Jenkins, JIRA, Maven, GIT, Python, SQL, Kubernetes, Docker, Confluence.

Sita corp,India

**Hadoop Developer**

**Responsibilities:**

● Conducted requirement discussions and designed effective solutions.

● Estimated Hadoop cluster requirements to ensure optimal performance.

**Nov 2014 – Aug 2016**

● Selected appropriate Hadoop components (Hive, Pig, MapReduce, Sqoop, Flume, etc.) based on project needs.

● Built scalable distributed data solutions using Hadoop technology.

● Established Hadoop clusters and managed data ingestion using Sqoop.

● Imported streaming logs to HDFS through Flume for effective data collection.

● Utilized Flume to collect, aggregate, and store web log data from various sources, including web servers, mobile devices, and network devices, pushing it to HDFS.

● Developed use cases and conducted technical prototyping for implementing Hive and Pig.

● Analyzed data using Hive, Pig, and custom MapReduce programs in Java.

● Implemented partitioning, dynamic partitions, and buckets in Hive.

● Installed and configured Hive, Sqoop, Flume, and Oozie on the Hadoop cluster.

● Scheduled Oozie workflow engine to run multiple Hive and Pig jobs efficiently.

● Tuned Hadoop clusters and monitored memory management and MapReduce jobs.

● Managed cluster maintenance, including adding and removing cluster nodes, monitoring, and troubleshooting.

● Developed a custom framework to address the small files problem in Hadoop.

● Deployed and administered a 70-node Hadoop cluster and administered two smaller clusters.

**Environment**: MapReduce, HBase, HDFS, Hive, Pig, Java, SQL, Cloudera Manager, Sqoop, Flume, Oozie, Java (JDK

1.6), Eclipse.

Medplus,India

**Data Analyst**

**Responsibilities:**

**July 2013 – Oct 2014**

● Documented system requirements, ensuring alignment with end-state objectives, and compiled Software

Requirement Specification and Use Case documents.

● Prepared ETL (Extract, Transform, Load) standards, established naming conventions, and documented ETL flow

procedures.

● Utilized Microsoft SharePoint for uploading, managing, and version controlling all project-related documents.

● Analyzed business requirements, segregating them into low-level and high-level categories, and converted them into Functional Requirements Documents.

● Created Dashboards in Tableau, incorporating calculations and parameters, to generate KPI reports for management analysis.

● Executed SQL queries for effective data manipulation.

● Facilitated weekly team meetings to assign testing tasks and gather status reports from individual team members.

● Managed change effectively by employing change management techniques, including Change Assessment, Impact Analysis, and Root Cause Analysis.

● Applied advanced Excel functions for spreadsheet generation and pivot table creation.

● Presented solutions through written reports, actively participating in analyzing, designing, testing, and monitoring systems following a waterfall methodology.

**Environment:** Microsoft (Word, Excel, PowerPoint, SharePoint, Visio, Teams), Confluence, Tableau, SQL Server

Management Studio (SSMS), DBeaver, Slack, Jira, Trello