PRANAY REDDY

Sr. Data Engineer

Phone: +1 443-406-8983

Mail: pranayreddy9799@gmail.com

Objective

Innovative Data Engineer with **9+ years of IT expertise**, including **5+ years in Big Data Analytics** utilizing the Hadoop ecosystem, Spark, and Azure cloud services with Scala and Python. **With 4+ years specializing in ETL and Data Warehousing**, I excel in architecting scalable data solutions, streamlining ETL pipelines, and driving high-performance data processing. I am eager to leverage my technical skills and creative problem-solving abilities to transform complex data into actionable insights, fueling strategic growth in a forward-thinking organization.

Professional Summary

- Over extensive experience in designing and implementing large-scale data solutions across both cloud (primarily Azure) and on-premises environments.
- Proven expertise in architecting and optimizing big data ecosystems using Hadoop, Spark, and Azure services—including Azure Blob Storage, Data Lake, and Synapse Analytics—for robust data warehousing and efficient ETL/ELT pipelines.
- Adept at developing scalable, real-time data streaming solutions leveraging Apache Kafka, Spark Streaming, and Azure Event Hubs, ensuring timely ingestion and processing of live data.
- Proficient in multiple programming languages such as Python, Scala, and Java; skilled in writing complex PySpark UDFs, Spark SQL, and HiveQL to drive efficient data transformations.
- Extensive experience managing large-scale databases (Hive, Oracle, SQL Server, SQL, PL/SQL, and T-SQL), ensuring optimized performance and rapid, reliable data retrieval.
- Strong background in UNIX shell scripting for automating Hadoop workflows and streamlining in-house development processes.
- Demonstrated expertise in building end-to-end data pipeline infrastructures for machine learning models by leveraging Python, PySpark, and Azure Data Factory.
- Hands-on experience with a comprehensive suite of Azure services—including Databricks, Data Lake Analytics, SQL Database, Logic Apps, and Functional App—to deliver integrated, cloud-based data solutions.
- Proficient in cloud data warehousing solutions including Snowflake, Amazon Redshift, and Google BigQuery.
- Experience with modern data transformation tools such as dbt (data build tool).
- Skilled in data vault modeling techniques for cloud data warehouses.
- Skilled in developing data pipelines using Hive and Sqoop for extracting and processing weblogs data into HDFS, as well as converting Hive/SQL queries into Spark transformations using Java.
- Proficient in developing Spark applications using a variety of tools (RDD transformations, Spark core, Spark Streaming) and in writing MapReduce jobs in Java for complex data processing tasks.
- Adept at implementing CI/CD practices with tools such as Jenkins, Bitbucket, GitHub, and Azure DevOps, ensuring smooth and reliable integration and deployment cycles.
- Strong expertise in troubleshooting and performance tuning Spark and Hive applications, resulting in consistently optimized system performance.
- Excellent communicator with robust client-facing experience and a proven track record in both Agile and Waterfall environments, translating complex business requirements into scalable, efficient data solutions.
- Committed to continuous learning and staying current with emerging technologies, ensuring best practices in data quality, governance, and modern data engineering methodologies.

Education

- Masters in data science at University of North Carolina Chapel Hill (Sep 2018 – Dec 2019)
- Bachelors in computer science at JB Institute of Technology, Telangana (Aug 2010 – May 2014)

Technical Skills

Big Data Technologies	Apache Hadoop (HDFS, MapReduce, YARN), Apache Spark (Core, Streaming, MLlib), Apache Hive, Sqoop, Oozie, Zookeeper, Kafka
Hadoop Distribution	Cloudera, Hortonworks
Azure Services	Azure Data Factory, Azure Data Bricks, Azure Logic Apps, Azure Functions, Azure Synapse Analytics, Azure DevOps EventHub, Apache Airflow, Snowflake, Azure DevOps, PowerBI
Cloud Data Warehousing	Snowflake, Amazon Redshift, Google Bigquery, DBT, Stitch, Looker
Web Technologies	HTML, CSS, JavaScript, XML, JSP, RESTful APIs, SOAP
Operating Systems	Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS.
Build Automation tools	Ant, Maven, Gradle
Version Control	GIT, GitHub.
IDE &Build Tools, Design	Eclipse, Visual Studio, Visual Studio Code, IntelliJ IDEA
Databases	MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB
Programming Languages	Scala, Python, Java, SQL (including PL/SQL & HiveQL)

Professional Experience

Sr. Data Engineer **Nov 2023 – Current**
Client: Google
Responsibilities:

- Architect and implement end-to-end big data solutions, from data ingestion to processing and analysis, utilizing HDFS and Google Cloud components.
- Design and develop scalable data pipelines using Google Cloud Dataflow, Dataproc, and Cloud Composer, integrating various data sources and destinations.
- Implement real-time data streaming solutions using Apache Kafka, Google Cloud Pub/Sub, and Dataflow for processing high-volume, high-velocity data.
- Optimize performance of Spark applications, including Spark SQL and PySpark, for large-scale data processing and analytics on Google Cloud Dataproc clusters.
- Develop and maintain ETL/ELT processes using a combination of Google Cloud services, including Cloud Storage, BigQuery, and Dataprep.
- Implemented and optimized data pipelines using Google Cloud Dataflow and Dataproc, integrating with BigQuery for efficient data processing and analytics.
- Utilized Cloud Composer (managed Apache Airflow) for orchestrating complex data workflows and ensuring efficient execution of data processing tasks.
- Design and implement data models and schemas for various data stores, including Hive, Parquet, and Cloud SQL, optimizing for query performance and storage efficiency.
- Leverage advanced Hive features such as partitioning, bucketing, and SerDe implementations for efficient data storage and retrieval.
- Develop and maintain Spark applications using Scala and Python for batch and stream processing, implementing RDD, Dataset, and DataFrame transformations.
- Implement data quality checks, data profiling, and data governance practices across the data pipeline to ensure data integrity and compliance.
- Orchestrate complex data workflows using Cloud Composer (managed Apache Airflow), ensuring efficient execution and monitoring of data processing tasks.
- Collaborate with data scientists and analysts to provide optimized datasets for machine learning and business intelligence applications.

- Implement and maintain CI/CD pipelines for data engineering projects using Google Cloud Build, Cloud Source Repositories, and JIRA for version control and project management.
- Conduct performance tuning and optimization of data pipelines, Spark jobs, and Google Cloud services to ensure optimal resource utilization and cost-efficiency.
- Mentor junior engineers on best practices in big data processing, cloud architecture, and data engineering principles.
  Environment: Google Cloud Platform (Dataproc , Cloud Storage, Cloud Dataflow, BigQuery , Cloud Composer, Cloud Pub/Sub), Apache Hadoop ecosystem (HDFS, YARN, MapReduce, Hive, Sqoop, Kafka), Apache Spark (Spark SQL, Spark Streaming), Python, Scala, PySpark , Git, JIRA, JUnit, Mockito, Shell scripting, Cloudera.

Sr. Data Engineer **Aug 2022 to Oct 2023**
Client: IT Shoulders, Inc
Responsibilities:

- Designed and developed proof of concepts (POCs) in Apache Spark using Scala to benchmark performance against MapReduce and Hive, showcasing expertise in big data processing frameworks.
- Architected and implemented complex data solutions leveraging Azure cloud services, including Azure Synapse Analytics, Azure Data Factory, Azure Data Lake Storage, and Azure Databricks.
- Developed and optimized batch and streaming data pipelines in Azure Data Factory, incorporating data extraction, transformation, and loading (ETL) processes from various sources into Azure Data Lake Storage and Delta tables.
- Engineered Spark Streaming applications for real-time analytics, integrating with event-driven architectures such as Azure Functions and Azure Logic Apps.
- Designed and implemented advanced Hive solutions, including custom Hive User-Defined Functions (UDFs), complex HiveQL queries for data analysis, and migration of ETL processes from Oracle to Hive.
- Developed and optimized Spark applications using PySpark and Spark SQL for large-scale data processing, transformation, and aggregation across multiple file formats.
- Implemented data ingestion and processing solutions using Sqoop for RDBMS to HDFS/Hive transfers, and Flume for real-time data streaming into Spark.
- Architected and developed solutions for handling complex data formats, including JSON and XML, using appropriate SerDe libraries in Hive and Spark.
- Led the design and implementation of CI/CD pipelines using Azure DevOps, incorporating Git, Maven, and Jenkins plugins for streamlined development and deployment processes.
- Spearheaded the migration of on-premises data warehouses to cloud-based solutions, leveraging Azure Synapse Analytics and Azure Databricks for enhanced performance and scalability.
- Implemented infrastructure-as-code practices using Azure Terraform modules to automate and manage cloud resources efficiently.
- Implemented data quality checks and data profiling using dbt (data build tool) to ensure data integrity and compliance across the data pipeline.
- Designed and implemented data models using data vault methodology for scalable and flexible data warehousing solutions.
- Optimized data processing workflows by leveraging Spark's in-memory computation capabilities and integrating SparkSQL with Hive metastore for improved performance.
- Mentored junior engineers on best practices in big data processing, cloud architecture, and data engineering principles.
  Environment: Apache Spark, Hadoop, HDFS, Hive, Pig, Sqoop, Flume, Azure Synapse Analytics, Azure Data Factory, Azure Data Lake Storage, Azure Databricks, Azure Functions, Azure Logic Apps, Java, Scala, Python, PySpark, Spark SQL, HiveQL, Git, Maven, Jenkins, Terraform, Linux, MySQL, Oracle DB, IntelliJ, CI/CD, Agile Methodologies.

Data Engineer. **April 2021 to July 2022**
Client: Capital One, Mclean, VA.
Responsibilities:

- Architected and implemented cloud data analytics solutions in Azure, focusing on migrating on-premises data warehouses to Azure cloud using services such as Azure Data Lake Storage Gen2, Azure Data Factory, Azure Databricks, and Azure Synapse Analytics.

- Developed and optimized ETL processes using PySpark and Spark SQL in Azure Databricks, transforming data from multiple file formats to derive business insights.
- Designed and implemented data pipelines using Azure Data Factory (V2), including creating jobs, scheduling triggers, and developing mapping data flows.
- Utilized Azure Key Vault to securely store and manage credentials for various data sources and services.
- Implemented data migration strategies to transfer on-premises data to Azure Blob Storage and Azure Data Lake Store using Azure Data Factory.
- Created and maintained Delta tables for ACID transactions using Spark DataFrames and temporary views.
- Developed Spark applications using Python (PySpark) API for complex data transformations and analysis.
- Orchestrated multi-cloud workflows using Apache Airflow, including building custom operators and managing dependencies.
- Managed Azure DevOps pipelines for continuous integration, continuous delivery (CI/CD), and release management.
- Created and optimized various data visualizations including charts, pivot tables, and straight tables to meet client requirements.
- Monitored and managed development processes, including enhancements, support, change requests, and testing, using JIRA and maintaining documentation in repositories.
- Performed repository management activities, including version control of dashboards and merging branches to maintain a centralized repository.
- Collaborated with cross-functional teams to gather requirements and translate them into technical specifications for data engineering solutions.

  Environment: Azure Synapse Analytics, Azure Data Factory, Azure Data Lake Storage, Azure Databricks, Azure Storage Account, Cosmos DB, Azure Automation, Oracle DB, PowerShell, Python, SQL, PySpark, Spark SQL, Power BI, Tableau, SSRS, Git, Azure DevOps

Hadoop Developer. **Jan 2020 to Mar 2021**

Client: CHICAGO BOARD OPTIONS EXCHANGE (CBOE) - Chicago, IL.

Responsibilities:

- Designed and implemented data pipelines using Hadoop ecosystem tools such as Flume, Sqoop, and Pig to extract data from various sources and store it in HDFS.
- Developed efficient MapReduce programs in Java for processing and analyzing large-scale unstructured data.
- Created and optimized Hive queries for data sampling, analysis, and reporting to support business intelligence needs.
- Implemented custom Pig User-Defined Functions (UDFs) to preprocess and transform data for analysis.
- Worked with various HDFS file formats including Avro and Sequence File, and utilized compression techniques like Snappy and bzip2 to optimize storage and processing.
- Developed Sqoop scripts to import and export data between Hadoop and relational databases.
- Utilized Oozie to create and schedule complex data processing workflows, orchestrating Sqoop, MapReduce, and Pig jobs.
- Collaborated with data engineers to configure and maintain Hive metastore for efficient metadata management of Hive tables and partitions.
- Optimized query performance by leveraging Tez as an execution engine for Hive and Pig jobs.
- Developed and maintained data models based on recognized standards to support analytical requirements.
- Created custom shell scripts for data validation, preprocessing, and cluster health checks.
- Implemented data ingestion processes using Flume for efficiently collecting and aggregating large volumes of log data.
- Utilized HBase for real-time random read/write access to big data when required by applications.
- Worked with Zookeeper for distributed coordination of Hadoop services and maintaining configuration information.
- Leveraged Hue as a web interface for interacting with Hadoop clusters and developing queries.
- Participated in proof of concept (POC) projects to demonstrate the advantages of migrating legacy systems to Hadoop.
- Collaborated with the operations team to troubleshoot issues and optimize Hadoop cluster performance.

  Environment: Hadoop, HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, HBase, Zookeeper, Tez, Hue, Java, Python, Shell scripting, Maven, Red Hat Linux, MS SQL Server, MongoDB, Oracle.

ETL Developer /Data Warehouse Developer **June 2014 to Aug 2018**

Client: iMedX Information Services Pvt. Ltd.

Responsibilities

- Gather and analyze business requirements to design and develop ETL processes and data integration solutions.
- Develop and maintain ETL workflows using SSIS packages for data extraction, transformation, and loading from various sources into SQL Server databases.
- Create and optimize complex SQL queries and stored procedures for data manipulation and reporting purposes.
- Implement and maintain data warehousing solutions, including designing star schemas and snowflake schemas.
- Develop and deploy SSRS reports to report servers, manage report subscriptions, and create various report types including tables, matrices, charts, and drill-down reports.
- Perform database administration tasks in SQL Server, including regular maintenance, backups, and performance tuning.
- Utilize SQL Server profiling tools such as Execution Plan and SQL Profiler to optimize query performance and enhance overall database efficiency.
- Design and implement data quality checks and data cleansing processes to ensure data integrity and accuracy.
- Develop and maintain documentation for ETL processes, data models, and reporting solutions.
- Collaborate with business analysts and stakeholders to understand reporting requirements and translate them into technical specifications.
- Troubleshoot and resolve issues related to ETL processes, data integration, and reporting services.
- Implement data security measures and ensure compliance with data privacy regulations.
- Perform data migrations and integrate data from various sources, including Excel, Access, flat files, and other database systems.
- Optimize ETL processes for large-scale data processing and implement best practices for data warehousing.
- Stay updated with the latest trends and technologies in ETL, data warehousing, and business intelligence.
  Environment: Microsoft SQL Server, SSIS, SSRS, T-SQL, SQL Server Management Studio, Visual Studio, Windows Server, XML, HTML.