**Sr DATA ENGINEER**

**Name: Shashank Palusa| Contact: (469) 905-1948**

**Email Id: p835159@gmail.com**

**Profile Summary:**

* **Experienced Data Engineer** with extensive expertise in designing, developing, and optimizing data pipelines using **Azure Databricks**, **Apache Spark**, and **Python/Scala**. Proven ability to process large-scale data, ensuring scalability, performance, and data quality across various enterprise-level projects.
* Strong command of **SQL**, with advanced knowledge in working with **Delta Lake** for building reliable, high-performance data lakes. Hands-on experience with **Azure Data Services** (ADLS, ADF, Synapse) to architect and manage robust data solutions that integrate and automate data flows across platforms.
* Adept at implementing **CI/CD & DevOps** practices using tools like **Terraform**, **Git**, and **Azure DevOps** to streamline deployment pipelines, infrastructure automation, and version control. Expertise in **performance tuning and troubleshooting**, ensuring the efficient execution of Spark jobs and resolving data pipeline issues to maintain high system availability.
* **Experienced Data Engineer** with strong expertise in **Google Cloud Platform (GCP)** services, specializing in building scalable, high-performance data pipelines and leveraging tools like **BigQuery**, **Dataflow**, and **Pub/Sub** for efficient data processing and analytics.
* Proficient in **Apache Spark** for distributed data processing, utilizing both batch and stream processing techniques to handle large datasets and deliver insights at scale in cloud environments.
* **Advanced Python/Scala skills**, applying these languages to develop and optimize data engineering solutions, create custom data transformations, and automate workflows within GCP-based architectures.
* **Strong SQL expertise** in designing and optimizing complex queries for data extraction, transformation, and analysis, ensuring efficient and reliable data operations across both structured and unstructured datasets.
* Adept at integrating data from various sources into cloud-based environments, collaborating with cross-functional teams to deliver data-driven insights, and continuously optimizing performance for high-volume, low-latency processing.
* Strong expertise in leveraging **Google Cloud Platform (GCP)** services such as **BigQuery**, **Google Cloud Storage**, and **Dataflow**, enabling cloud-native data engineering solutions and streamlining data workflows in a cloud-first environment.
* Skilled in **real-time and batch data processing** using **Spark Streaming** and **Python**, optimizing performance and ensuring timely delivery of insights for business-critical decisions.
* **Exceptional communicator** with the ability to translate complex technical concepts into clear solutions for both technical and non-technical stakeholders, enabling collaboration across teams and ensuring successful project delivery.

**Technical Skills:**

**Big Data Ecosystems**: Spark, HDFS and Map Reduce, Pig, Hive, YARN, Oozie, Apache Spark, Apache NiFi, Apache Kafka

**Cloud Technologies**: Google Cloud Platform, Pub/Sub, Dataflow, Big Query, Dataproc, AWS EC2, AWS s3 buckets, Azure, Azure Databricks, Snowflake

**Scripting Languages**: Python, Shell.

**Programming Languages**: Python, Scala

**Databases**: MongoDB, Netezza, SQL Server, MySQL, Postgres, ORACLE, DB2, Cassandra

**Reporting tools**: Tableau, Power Bi

**Operating systems**: UNIX, Linux, and Windows Variants

**Tools**: Putty, SecureCRT, Jenkins, PyCharm, Spyder, Anaconda, Collibra, Pelican, Eclipse, GIT and, Info works, Informatica, Looker, Talend

**Version Control Tools**: Git

**Methodologies**: Agile, Waterfall.

**PROFESSIONAL EXPERIENCE:**

**Client: CVS Health, Texas| Sep 2021 –Till Date**

**Sr Data Engineer.**

* Developed end-to-end data pipelines using **Azure Databricks, Apache Spark, and Python/Scala** for large-scale data processing, enabling seamless **ETL** processes and data integration.
* Optimized **Spark** performance by tuning configurations, partitioning strategies, and caching mechanisms, resulting in faster job execution and reduced data processing times.
* Implemented Delta Lake on top of **Azure Data Lake Storage** (ADLS) to provide scalable, ACID-compliant storage and efficient management of data lakes for batch and streaming workloads.
* Leveraged **Azure Data Factory (ADF)** to design, orchestrate, and automate data workflows, ensuring smooth data ingestion, transformation, and movement across multiple data sources.
* Designed and maintained **Azure Synapse** Analytics pipelines for data warehousing, enabling seamless integration of structured and unstructured data into analytics platforms.
* Built and optimized **SQL** queries for data extraction, transformation, and aggregation, enhancing performance and improving reporting and analytical processes.
* Utilized **Apache Spark** to process real-time data streams and batch jobs, delivering high-performance analytics and insights for critical business decision-making.
* Implemented CI/CD pipelines using **Git, Azure DevOps (ADO),** and **Terraform** to automate data pipeline deployment, version control, and infrastructure provisioning, ensuring consistent and reliable deployments.
* Developed robust data monitoring and alerting systems within **Databricks** to identify pipeline failures, job performance bottlenecks, and data quality issues in real-time.
* Automated infrastructure management using **Terraform** to provision and manage Azure resources, ensuring repeatable and scalable infrastructure deployment for data engineering solutions.
* Collaborated with data scientists and business analysts to gather requirements, integrate machine learning models, and support advanced analytics within the data pipeline.
* Utilized **Azure Key Vault** to securely store and manage sensitive data and credentials used within **Databricks and ADF** pipelines, ensuring data security and compliance with organizational policies.
* Troubleshot performance issues across large **Spark** jobs by analyzing logs, resource utilization, and optimizing queries, reducing job runtime by up to 30%.
* Created detailed documentation for data engineering processes, pipeline architectures, and best practices to ensure knowledge transfer and facilitate team collaboration.
* Enabled real-time data analytics by integrating **Apache Kafka** with **Spark Streaming** for processing high-throughput data streams in near real-time, enhancing decision-making capabilities for business users.
* Designed, developed, and maintained scalable and efficient ETL (Extract, Transform, Load) pipelines using **PySpark** on Hadoop and **GCP Dataproc** to process large datasets from various data sources.
* Implemented and optimized data transformations using **PySpark** for performance tuning and better data processing.
* Built and optimized **Hadoop** **MapReduce** and **PySpark** jobs for processing large volumes of structured and unstructured data.
* Utilized **HDFS (Hadoop Distributed File System)** for storing and managing large datasets in a distributed environment.
* Designed and implemented data workflows on Google Cloud **Dataproc** for running big data processing tasks, leveraging **PySpark** and Hadoop clusters.
* Automated data pipelines to ingest, transform, and load data into **BigQuery** for analytics and reporting, ensuring scalability and performance.
* Developed complex **SQL** queries for aggregating, transforming, and retrieving data from Big Query for analytical insights.
* Tuned and optimized SQL queries and table schemas to improve performance in **BigQuery**.
* Orchestrated and automated end-to-end data workflows using **Python** and **Shell** Scripting to schedule jobs, monitor pipeline health, and handle failures.
* Created shell scripts to automate routine tasks such as log management, data transfers, and job scheduling on **Unix** systems.
* Ensured data quality by developing validation scripts to check for missing, inconsistent, or corrupted data during the **ETL** process.
* Conducted extensive data cleansing and enrichment processes using **PySpark** and **SQL** to ensure data accuracy and integrity.
* Tuned **PySpark** jobs by leveraging caching, partitioning, and efficient joins for optimizing processing performance on **Dataproc** clusters.
* Identified and resolved performance bottlenecks in **Hadoop** jobs and **SQL** queries using profiling tools and query optimization techniques.
* Designed and implemented data models and schemas in **BigQuery** for efficient storage and retrieval, ensuring optimal performance for analytics and reporting.
* Worked on data partitioning, clustering, and indexing to optimize query performance on large datasets.
* Collaborated with data scientists, business analysts, and stakeholders to understand business requirements and translate them into technical solutions.
* Documented the end-to-end data pipeline architecture, workflows, and operational procedures to ensure maintainability and scalability.
* Utilized **Git** for version control, ensuring proper management of code versions, and implemented CI/CD pipelines for deploying **PySpark** jobs and scripts to GCP **Dataproc**.
* Monitored the performance and health of data pipelines and clusters using GCP monitoring tools and logs, identifying and resolving issues related to job failures or performance.
* Developed automated alerts and logging systems to notify in case of failures or performance degradation.
* **Developed and optimized SQL queries** for data extraction, transformation, and analysis from multiple data sources, enabling efficient reporting and decision-making through automated data pipelines.
* **Utilized Python** for data cleaning, statistical analysis, and automation of data processes, applying libraries like Pandas, NumPy, and Matplotlib to derive actionable insights from large datasets.
* **Created interactive dashboards and visualizations** in **Tableau**, transforming complex data into intuitive and impactful insights that guided strategic decisions across business units.

**Client: CapitalOne, Remote| Jul 2020 – Aug 2021**

**Sr Data Engineer.**

* **Developed end-to-end data pipelines** using **Python**, **Scala**, and **SQL** to automate data extraction, transformation, and loading (ETL) processes, improving data flow efficiency across various platforms.
* **Implemented data models** using **AWS Redshift** and **Snowflake**, creating optimized schema designs and ensuring high performance for large-scale analytical queries and data aggregations.
* **Built and maintained real-time and batch data processing systems** using **AWS Lambda**, **Python**, and **Apache Kafka**, enabling high-throughput data streaming and real-time analytics.
* **Designed and optimized complex SQL queries** for data extraction, transformation, and aggregation on both **Open Source RDBMS** (e.g., PostgreSQL, MySQL) and **NoSQL** databases (e.g., MongoDB, DynamoDB), improving query performance and reducing execution time.
* **Leveraged AWS services** (e.g., **S3**, **Glue**, **EMR**) to build scalable, serverless ETL pipelines, enhancing data storage, transformation, and analysis capabilities.
* **Integrated and managed data across multiple storage and database platforms** (**RDBMS, NoSQL, Cloud**) to ensure consistency, availability, and security in a hybrid cloud environment.
* **Migrated on-premise databases to AWS Redshift and Snowflake**, optimizing performance by leveraging cloud-based architecture and ensuring seamless integration of legacy systems with modern data warehouse solutions.
* **Developed automated data ingestion frameworks** using **Python** and **AWS Glue** to seamlessly extract, clean, and load data from various sources into cloud-based warehouses, reducing manual intervention and errors.
* **Implemented data security and access control** in **AWS** environments, using **IAM** roles and encryption methods to secure sensitive data stored in **S3**, **Redshift**, and **Snowflake**.
* **Utilized Snowflake's multi-cluster architecture** to optimize concurrent queries and enable high performance for large data workloads, ensuring faster reporting and analytics capabilities for business users.
* **Optimized cloud-based data warehouses** (**Redshift and Snowflake**) through partitioning, indexing, and query optimization techniques, resulting in a 30% improvement in query performance.
* **Created data quality and validation frameworks** in **Python** to monitor and flag issues in the data pipeline, ensuring high-quality, reliable data for downstream analytics and reporting.
* **Collaborated with data scientists** to provide clean and preprocessed data for machine learning models, ensuring efficient feature extraction and model training processes on cloud infrastructure.
* **Implemented and maintained CI/CD pipelines** using **AWS CodePipeline** and **Git** to automate testing, deployment, and integration of data engineering components across the cloud platform.
* **Troubleshot and resolved performance issues** in cloud data warehouses (**Redshift** and **Snowflake**) and databases (**RDBMS** and **NoSQL**), improving overall data pipeline reliability and system performance.
* **Extracted, transformed, and loaded (ETL) data** using **SQL** from various databases and integrated it into **Power BI** and **Tableau** dashboards, enabling stakeholders to access real-time, actionable business insights.
* **Utilized Python** to automate data cleaning, preprocessing, and analysis tasks, leveraging libraries such as Pandas, NumPy, and SciPy to enhance the efficiency and accuracy of data-driven decisions.
* **Designed and developed interactive visualizations** in **Tableau** and **Power BI**, creating dynamic reports that helped business leaders track KPIs, monitor trends, and drive performance improvements across departments.

**Walmart, Remote, USA| Jan2020 – Jun 2020**

**Big Data Engineer.**

* **Designed and developed end-to-end data pipelines** using **Scala**, **Apache Spark**, and **Python**, enabling the seamless extraction, transformation, and loading (ETL) of large datasets across distributed platforms like **Hadoop** and **Azure/GCP**.
* **Implemented highly scalable Big Data solutions** leveraging **Apache Spark** and **Hadoop**, processing massive volumes of structured and unstructured data efficiently to meet business requirements.
* **Optimized performance** of **Spark** jobs by tuning memory configurations, partitioning strategies, and using caching, reducing data processing times by 40% for large-scale datasets.
* **Built and maintained real-time data streaming pipelines** using **Apache Spark Streaming** and **Kafka**, ensuring near real-time data processing and analytics for business-critical applications.
* **Developed complex ETL workflows** using **Scala** and **Apache Spark**, automating the transformation of raw data into actionable insights, improving data accessibility and reporting speed.
* **Integrated cloud platforms** such as **Azure/GCP** to create scalable data architectures, leveraging cloud-based storage (e.g., **Azure Blob Storage**, **Google Cloud Storage**) and data warehousing (e.g., **Azure Synapse**, **BigQuery**) for optimal data management.
* **Collaborated with cross-functional teams** to design data architectures that ensure data pipelines are robust, scalable, and support analytics at scale, facilitating data-driven decision-making across departments.
* **Built and deployed data models** on **Hadoop** ecosystem tools (**HDFS, Hive, and Pig**), ensuring optimized and reliable data storage solutions that improved query performance and business intelligence insights.
* **Used Scala and Spark to implement data transformation logic** for batch and streaming data jobs, enhancing data pipeline reliability and reducing processing times.
* **Developed data ingestion frameworks** to integrate various data sources (structured and unstructured) into distributed storage systems like **HDFS** and **cloud-based platforms**, ensuring efficient data movement across systems.
* **Created automated monitoring solutions** for data pipelines, leveraging tools like **Spark UI** and **Azure Monitor** to proactively identify and troubleshoot performance bottlenecks and system failures.
* **Utilized advanced SQL and SparkSQL** to perform complex data analysis and querying, optimizing query performance and delivering insights to business users.
* **Communicated complex technical solutions** to both technical and non-technical stakeholders, ensuring alignment on project goals and enabling informed decision-making through clear, concise presentations and documentation.

**Wipro, Hyderabad, India| Jan2018 – Jul 2019**

**Big Data Engineer.**

* **Developed and optimized data pipelines** using **Hadoop**, **Hive**, and **Spark**, enabling efficient processing of large-scale datasets across distributed environments.
* **Built ETL workflows** using **Python** and **Hive** to extract, transform, and load structured and unstructured data into HDFS, ensuring data quality and consistency across multiple data sources.
* **Designed and implemented complex SQL queries** on **Hive** to perform data analysis and aggregation, improving reporting and decision-making processes by enabling quicker data retrieval from large datasets.
* **Leveraged Apache Spark** for real-time data processing, optimizing data transformations and aggregations, resulting in a 40% reduction in processing time compared to traditional MapReduce jobs.
* **Optimized Hadoop cluster performance** by tuning **Spark** jobs, managing memory, and adjusting configurations to improve job execution times and resource utilization.
* **Developed Python-based scripts** for automating data ingestion, extraction, and cleansing tasks within the Hadoop ecosystem, streamlining data pipeline management and reducing manual intervention.
* **Collaborated with data scientists** to implement machine learning algorithms on large datasets using **Spark MLlib**, significantly enhancing predictive analytics capabilities.
* **Integrated Hive with HDFS** to create an efficient data warehouse solution for structured data storage and optimized query performance, reducing report generation time by 25%.
* **Troubleshot and resolved performance bottlenecks** in Hadoop and Spark clusters, leveraging monitoring tools and logs to ensure optimal system uptime and data pipeline reliability.
* **Implemented data partitioning and indexing strategies** in **Hive** and **Spark** to improve query performance and ensure scalability as data volumes continued to grow.

**Timon Global Solutions, India| Jun 2015 – Dec 2017**

**Hadoop Developer**

* Handled importing of data from various sources. Loading data into HDFS and Extracted the data from MYSQL into HDFS using Sqoop.
* Involved in loading data from edge node to HDFS using shell scripting.
* Involved in analyzing system failures, identifying root causes and recommended course of actions.
* Managing and Scheduling jobs on Hadoop Cluster using Oozie workflows.
* Used Spark stream processing to get data into in-memory, implement RDD transformations, actions to process as units.
* Created and worked with Sqoop jobs with incremental load to populate Hive External tables