**Ankuraj Gokul Sampath Kumar**

**Sr Data Engineer**

**E-mail: krishna@adven-it.com  
 Phone: +1- 9728538231**

**Professional Summary:**

* Accumulated over **10 years** of professional expertise in the IT sector, specializing in **Senior Data Engineer/Analyst roles.**
* Skilled in executing comprehensive end-to-end data engineering processes, including **Data Extraction, Transformation,** and **Ingestion** into structured data tables.
* Experienced in working with cloud platforms like Amazon Web Services, Azure, and Databricks, integrating with **Databricks** on both **Azure**, **GCP** and **AWS**.
* Strong background in migrating databases to Snowflake, with a deep understanding of **Snowflake** Database schema and table architecture.
* Extensive expertise in **Hadoop architecture** and ecosystems, covering **HDFS**, **Hive**, **Pig**, **Sqoop**, **Job** **Tracker**, **Task Tracker**, **Name Node**, and **Data Node**.
* Skilled in installing, configuring, managing, supporting, and monitoring **Hadoop clusters** using distributions like **Apache Hadoop**, Cloudera Hortonworks, and cloud services such as **AWS**, **Azure** and **GCP**.
* Expertise in building **Azure** native enterprise applications, migrating applications to the Azure environment, and implementing **Data lakes** and **Business Intelligence** tools in Azure.
* Proficient in crafting custom **Kafka** consumer code, adapting existing producer code in **Python** for data transmission to Spark streaming jobs.
* Hands-on experience with **Elasticsearch** for rapid storage, retrieval, and real-time analysis of large data volumes.
* Executed a data warehouse solution, involving **ETL** processes, on-premises to **cloud migration**, and deploying batch and streaming data pipelines in cloud environments.
* Utilized **Spark** **DataFrames**, **Spark SQL**, and **Spark's RDD API** for diverse data transformations and dataset construction.
* Proficient in integrating various data sources **(Oracle SE2, SQL Server, flat files, unstructured files)** within a data warehouse environment.
* Hands-on experience with Azure's features for diagnosing and resolving end-user issues, providing effective solutions.
* Practical exposure to **NoSQL databases** like **HBase**, **Azure Cosmos DB**, **MongoDB**, and **Cassandra**, understanding their functionality and implementation.
* Extensive hands-on experience using **AWS** **Cloud services** and SDKs for interactions with services like **AWS API Gateway**, **Lambda, S3, IAM,** and **EC2**.
* Creation and launch of numerous **AWS Lambda** functions, using **AWS Lambda Libraries** and deploying **Lambda Functions** in **Scala** with custom **Libraries**.
* Leveraged various features within **Teradata**, including **BTEQ, Fast Load, Multi Load, SQL Assistant, and DDL/DML commands, demonstrating** expertise in Teradata concepts.
* Proficiency in developing real-time data streaming solutions with **Apache Spark, Spark Streaming**, **Apache Storm, Kafka**, and **Flume**.
* Creation of RESTful web services for extracting, transforming, and aggregating data into **Hadoop** from various endpoints, including **HBase** and **Solr**.
* Implemented a **CI/CD Jenkins pipeline using Groovy scripts**.
* Engaged in **Data Lake** implementation, crafting data pipelines, and applying business logic using **Apache Spark**.
* Proficient in **AWS DNS Services** using **Route53**, with knowledge of routing policies like Simple, **Weighted, Latency, Failover**, and **Geolocation**.
* Skilled in utilizing **Spark Core, Spark SQL, Spark MLlib, Spark GraphX**, and **Spark Streaming** for processing and transforming data in-memory, implemented in **Scala**.
* Hands-on experience in optimizing **HIVE queries** and **MapReduce** programs for improved scalability and faster execution.
* Extensive expertise in utilizing scheduling tools for batch job automation.
* Practical experience with **Apache Solr** and **Lucene**.
* Proficiency in **SQL Server, SQL, query development**, **stored procedures**, and **functions**.
* Solid experience in **Agile/Scrum** **methodologies**, actively participating in Scrum calls for project analysis and development discussions.
* Significant expertise in enhancing Hive and Pig core functionality through custom **User-Defined Functions (UDFs)**.
* Extensive engagement with semi-structured data, including fixed-length and delimited files, for **data cleansing, report generation, and standardization**.
* Hands-on experience with **GCP** cloud services such as **GCP** **Cloud Dataflow**, **GCP BigQuery, GCP Cloud Storage, and GCP Pub/Sub**.
* Comprehensive expertise in both relational and non-relational databases, **including Oracle, PL/SQL, SQL Server, MySQL, and DB2.**
* Strong experience in extending **Hive and Pig core functionality** with custom **UDFs**.
* Exceptional comprehension of **Zookeeper** for supervising and administering **Hadoop jobs**.
* Utilized prior data engineering expertise in **AWS** to analyze commonalities and distinctions between **AWS and GCP services** for well-informed choices in selecting GCP services for Proof of Concept (POC). Proficient in **GCP Dataproc, GCS, Cloud Functions**, and **BigQuery**.
* Creation of **Power BI & Tableau** reports and impactful dashboards by collecting and translating end-user requirements.
* Proficient in constructing **Docker images** for various data applications intended for operation within containers.
* Highly skilled in developing, troubleshooting, configuring, and deploying **ETL packages** using **Microsoft SQL Server Integration Services (SSIS)**. Proficient in configuring SQL Azure firewall for enhanced security.
* Collaboration with application architects to transition from **Infrastructure as a Service (IaaS)** to **Platform as a Service (PaaS) models**. Deployment of **Azure Resource Manager** JSON Templates using **PowerShell**.
* Demonstrates a strong capacity for managing diverse tasks and responsibilities, with quick understanding and adaptation to new surroundings and technology.

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming Languages** | Python, Scala, SQL, Java, Shell Scripting |
| **Python Libraries** | NumPy, Pandas, Scikit-Learn, SciPy, Seaborn, Matplotlib, |
| **Web Technologies** | HTML, CSS, JavaScript, PHP |
| **Hadoop Technologies** | Hive, Spark, HDFS, Map Reduce, YARN, Kafka, Pig, Hive, Sqoop, HBase, Oozie, Zookeeper, Cloudera Manager and Horton Works |
| **Cloud Platforms** | Azure Cloud, AWS Cloud architecture, AWS Big Data computing, Microsoft Azure, GCP |
| **Databases** | Oracle, DB2, MySQL, Sybase, PostgreSQL, MS SQL Server |
| **AWS** | EC2, S3, Glacier, Redshift, RDS, EMR, Lambda, Glue, CloudWatch, Recognition, Kinesis, CloudFront, Route53, DynamoDB, Code Pipeline, EKS, Athena, Quick Sight |
| **Azure** | Azure Storage, Azure SQL Server, Azure Data Factory, Azure Databricks, Azure Machine Learning, Azure Cosmos DB, Azure Data Lake Storage, Azure HDInsight, Azure Stream Analytics, Azure Logic Apps, Azure Functions, Azure Kubernetes Service (AKS), Azure Synapse Analytics, Azure DevOps |
| **GCP** | Google Cloud Storage, Google BigQuery, Google Cloud Dataprep, Google Cloud Dataflow, Google Cloud Pub/Sub, Google Cloud Dataproc, Google Cloud Composer, Google Cloud Machine Learning Engine, Google Cloud AutoML, Google Kubernetes Engine (GKE), Google Cloud Bigtable, Google Cloud Firestore, Google Cloud Functions, Google Cloud Run, Google Cloud SQL, Google Cloud Datastore, Google Cloud Pub/Sub, Google Cloud Vision AI, Google Cloud Natural Language AI |
| **NoSQL Databases** | MongoDB, Cassandra, HBase |
| **Visualization Tools** | Tableau, Power BI, Advanced Excel |
| **Version Control Systems** | Git, GitHub, CVS, SVN |
| **IDEs** | PyCharm, IntelliJ IDEA, Jupyter Notebooks, Eclipse, Visual Studio |
| **Operating Systems** | Unix, Linux, Windows |
| **Methodologies** | Agile/Scrum, Waterfall |

**Professional Experience:**

**Client –** **Cigna, Atlanta, GA Sep 2023 – Present**

**Role – Sr GCP Data Engineer**

**Responsibilities:**

* Applied extensive expertise in **Snowflake** and **ETL** **methodologies** to devise and enhance Data Warehousing solutions, empowering informed decision-making.
* Developed and optimized complex **SQL queries** in Google **BigQuery** to analyze large-scale datasets efficiently, enabling **data-driven decision-making**.
* Designed and implemented **data ingestion pipelines** to **load structured and semi-structured data** into **Google BigQuery** using **batch** and **streaming mechanisms**.
* Utilized **BigQuery's** partitioning and clustering features to improve query performance and reduce costs associated with **data storage and processing**.
* Integrated **Google** **BigQuery** with visualization tools like **Data Studio and Looker** to create interactive **dashboards** and **reports** for business stakeholders.
* **orchestrated** and managed **Apache Hadoop** and **Apache Spark clusters** on **Google Cloud Dataproc** for large-scale data processing and analysis.
* Leveraged **Dataproc's** integration with other **GCP** services such as **BigQuery** and **Cloud Storage** to ingest, process, and store data across the entire analytics pipeline.
* Developed and deployed data processing pipelines using **Google Cloud Dataflow** to process streaming and **batch data** at scale, enabling real-time analytics and insights.
* Designed and implemented data transformations and aggregations using **Apache Beam SDK** in **Python and Java** for flexible and expressive pipeline development.
* Leveraged **Dataflow's** unified programming model to seamlessly transition between batch and streaming processing modes, ensuring consistency and **reliability in data processing workflows**.
* Configured **Pub/Sub topics** and subscriptions for reliable and durable message delivery across asynchronous communication channels.
* Integrated Pub/Sub with various GCP services such as **Dataflow, Cloud Functions**, and **App Engine for event-driven data processing, analytics, and automation**.
* Implemented message ordering and **deduplication** mechanisms in Pub/Sub to ensure message integrity and idempotent processing in distributed systems.
* Developed and maintained scalable **ETL** **pipelines** employing **Python**, **SQL**, and **ETL** principles, ensuring smooth **Data Extraction, Transformation,** and **Loading processes**.
* Maximized the capabilities of **Snowflake** to deliver superior Data Warehousing solutions meeting diverse business requirements.
* Collaborated with interdisciplinary teams to design and deploy end-to-end **Data Pipelines**, **streamlining** **data movement** across various stages.
* Leveraged **Snowflake** proficiency to establish secure, reliable, and high-performing Data Warehouses.
* Implemented **GIT** for **version control** to ensure code and data pipeline changes were traceable and maintainable.
* Teamed up with **Databricks** and **AWS** **Services** to efficiently orchestrate Data Processing workflows.
* Designed, developed, and managed robust and scalable data pipelines supporting **real-time** **and batch data processing needs**.
* Utilized **ETL principles** to create efficient processes for Extracting, Transforming, and Loading data into **Snowflake** from diverse sources.
* Promoted the adoption of Data Engineering best practices to uphold **data quality, integrity, and accessibility**.
* Collaborated closely with **Data Scientists** and Analysts, furnishing them with well-prepared data for actionable insights.
* Proactively monitored and resolved data pipeline performance issues to ensure **uninterrupted** **data** **flow** and **processing**.
* Engaged in **Data Ingestion** to **Azure Services** and processing data in **Azure Databricks**.
* Continuously pursued opportunities to enhance **Data Pipeline Efficiency, Reliability, and Scalability**.
* Worked with cross-functional teams to define **Data Integration strategies** and optimize solutions with Snowflake.
* Ingested API data into Azure Data Storage services using **PySpark** in Databricks and further transformed it into a structured format stored in **DataLake gen2**.
* Designed, established, and maintained **Extract, Transform, Load (ETL) pipelines** using **Databricks** and **AWS Glue.**
* Designed and optimized data cataloging workflows in **AWS Glue** to automate the discovery, organization, and management of data assets across various data sources.
* Leveraged **AWS Glue Data Catalog** to provide a centralized metadata repository for enhanced data governance, lineage tracking, and data discovery capabilities.
* Utilized **AWS Glue Crawlers** to automatically discover and catalog metadata from diverse data sources such as Amazon S3, RDS, Redshift, and DynamoDB.
* Integrated **AWS Glue** with **AWS Lambda** functions and **Step Functions** to create **serverless data pipelines** for seamless **data processing and transformation**.
* Performed **PySpark** and **Spark SQL transformations** in **Azure Databricks** for complex business rule implementations.
* Developed **Databricks Spark jobs** with **PySpark** for various table operations.
* Implemented actions and transformations on **RDDs**, data frames, and datasets using Spark SQL and Spark Streaming Contexts connected to **Apache Spark**.
* Transferred data from legacy systems to **Snowflake** using **SnowSQL** and **Snowpipe** for efficient data processing.
* Adhered to **SDLC** principles throughout the development process.
* Scheduled Spark processes in **AWS EMR cluster** and created Spark Streaming tasks in Python.
* Created ETL pipelines to load data into **Databricks Delta Lake** in a **multi-layer architecture**.
* Built complex **ETL** jobs transforming data visually with data flows or compute services in Azure Databricks.
* Customized **Talend** Job lets for seamless integration with **Snowflake** functionality.
* Implemented **Snowpipe** for continuous data loading and utilized **COPY** command for bulk loading into Snowflake.
* Automated **CI/CD data pipelines** for marketing platforms data transfer to AWS S3 data lake.
* Developed a customized **Snowflake** utility function in Scala for data transfer from **AWS S3 bucket** to Snowflake.
* Created, scheduled, and monitored **Data Pipelines** with Apache Airflow.
* Predicted and monitored **Spark** **Databricks** cluster size.
* **Automated EMR** Spin jobs deployment using **Lambda** and other AWS services.
* Facilitated data sharing between multiple **Snowflake** accounts.
* Designed and optimized internal and external stages for efficient **data loading** and **transformation**.
* Redesigned and optimized Views in **Snowflake** for enhanced performance.
* Conducted thorough unit testing of data between **Snowflake** and other platforms like **Redshift**.
* Created Looker reports based on Snowflake connections and validated them using **Redshift data**.
* Provided expertise in **Snowflake**, offering guidance on **solution architecture, design, development, and deployment**.
* Created pipelines, data flows, and complex transformations using **ADF, PySpark, and Databricks**.
* Conducted data blending and preparation for **Tableau** consumption using **Alteryx and SQL**.
* Created a data pipeline using **Spark, Scala, and Apache Kafka** to ingest data and store it in an HDFS secured folder.
* Developed user-friendly website interfaces using **Python and Django**.
* Developed **UNIX** **shell** **scripts** for task scheduling and **PL/SQL scripts** for index maintenance.

**Environment:** Snowflake, Tableau, AWS Services, Redshift, Python, SQL, ETL, Databricks, GIT, Data Warehousing, EMR, Amazon S3, AWS Lambda, Data Pipelines, Legacy Systems, Teradata, Hadoop, HDFS, Sqoop, Apache Airflow, Eclipse.

**Client:** **T- Mobile,** **Madison, OH**  **Mar 2023 – Aug 2023**

**Role:** **Data Engineer**

**Key Responsibilities:**

* Crafted, engineered, and enhanced **Data Pipelines** in **Python** to ensure streamlined Data Extraction, Transformation, and Loading Processes.
* Employed **NumPy** and **Pandas** for sophisticated **Data Manipulation** and Analysis, guaranteeing Data Accuracy and integrity.
* Produced compelling **Data Visualizations** using **Matplotlib**, effectively communicating insights through Charts and **Graphs**.
* Developed Interactive **Visualizations** and **Dashboards** in **Tableau**, empowering stakeholders with Data-Driven Insights.
* Contributed to **Agile Project Methodologies**, actively engaging in collaborative sprints for efficient data solution development.
* Managed version control and facilitated collaborative development using **GITHUB**, ensuring code traceability and maintainability.
* Proficient in **Amazon AWS services** including **RedShift, EC2, S3, and EMR,** for rapid and efficient Big Data processing.
* Utilized **Sqoop** to import and export data to **HDFS, Pig, Hive, and HBase**.
* Oversaw and analyzed **Hadoop log files for control** and examination purposes.
* Configured and set up **Hadoop MapReduce, HDFS**, and developed various **Java MapReduce** tasks for data cleaning and processing.
* Leveraged **S3, Lambda, Glue, DynamoDB, Elasticsearch, CloudWatch, and Athena** to establish and maintain a comprehensive **data lake** across **AWS**.
* Handled large sets of organized, **semi-structured, and unstructured data** within the **Hadoop system**, focusing on loading and transformation.
* Developed **Java MapReduce** programs for both simple and complex data analysis.
* Utilized Flume to ingest data from diverse sources into **HDFS**.
* Developed **MapReduce** scripts to process raw data, populate staging tables, and store cleaned data in partitioned tables in the EDW.
* Utilized Spark SQL to prepare and store data in **AWS S3**, along with data frames imported from **Hive** databases.
* Designed FTP programs to store **DB2 Sqoop data** in **AWS** in **Avro format**.
* Conducted examinations of **Hadoop clusters** and various big data analysis tools such as **MapReduce, Hive, and Spark**.
* Engineered **extract, transform, and load (ETL)** software for **DB2** fact and dimension tables.
* Implemented **Sqoop** to connect to **MySQL** databases, create **Oozie Workflows**, and transform **MySQL** data into **AVRO** before writing it to **HDFS**.
* Developed **Hive** queries by comparing recent data with EDW reference tables and historical measures, aiding market analysts in identifying new trends.
* Engaged in significant development projects utilizing codebases featuring P**ython, Django, R, MySQL, MongoDB, and jQuery.**
* Exported data from **Sqoop** into **HDFS** and **Hive** for report analysis purposes.

**Environment:** Python, NumPy, Pandas, Matplotlib, NLTK, Tableau, Agile, GITHUB, Windows.

**Client: Bridge Global, India**  **Jun 2018 – Jul 2022**

**Role: Sr. Data Engineer/Azure DataBricks**

**Responsibilities:**

* Spearheaded the creation, **optimization**, and **management** of scalable **Data Warehousing** solutions on **Snowflake**, ensuring smooth data access for **stakeholders**.
* Directed end-to-end **ETL Processes** proficiently, employing **SQL** for **Data Extraction**, **Transformation, and Loading**, resulting in efficient **Data Workflows**.
* Automated Data Loading, Versioning, and Historical Analysis utilizing **Snowflake's** innovative features like **Snowpipe** and **Time Travel**, enhancing **Data Accuracy** and **Accessibility**.
* Managed and automated tasks within **Snowflake** to effectively handle Data Processing and maintenance Workflow.
* Integrated **AWS services seamlessly** (**EC2, AWS Glue, Athena, Redshift, Lambda, S3, RDS, SNS, IAM**), constructing resilient and scalable data solutions tailored to diverse business needs.
* Leveraged advanced **Python Libraries** (pandas, NumPy, Seaborn, SciPy, matplotlib, scikit-learn) for **Data Manipulation, Analysis, and Machine Learning**, facilitating data-driven decisions.
* Ensured code traceability and maintainability through version control with **GIT**, fostering efficient collaboration among team members.
* Flourished in **Agile methodologies**, promoting cross-functional collaboration and iterative development, consistently delivering solutions aligned with evolving business requirements.
* Designed and implemented scalable, efficient **data pipeline** processes for data ingestion, cleansing, transformation, and integration using **Sqoop, Hive, Python, and Impala**.
* Participated in requirements **gathering** and analysis to ensure a clear understanding of project objectives and stakeholder needs.
* Configured and integrated necessary **AWS services** to establish Infrastructure as a code (Iaas) in the AWS cloud platform from scratch.
* Developed data processing models using **PySpark and Spark SQL** in **Databricks** to derive insights for specific purposes.
* Conducted data processing using Spark Jobs with **PySpark** in **Databricks** to clean information and integrate it into the Enterprise Data Warehouse in **Big Query**.
* Extracted data from multiple well-logging systems to develop machine learning algorithms for problem-solving.
* Utilized **Spark-SQL** & **PySpark** based models on **Azure Databricks ML workspace** to build & deploy ML models at scale for various purposes.
* Created ETL pipelines using a combination of Python and Snowflake's **SnowSQL** and wrote **SQL queries** against **Snowflake**.
* Performed **Data Cleaning**, feature scaling, and feature engineering using pandas and NumPy packages in Python.
* Designed and deployed multi-tier applications with high availability, fault tolerance, and auto-scaling on **AWS Cloud Formation** utilizing various AWS services.
* Set up storage and data analysis tools in **AWS cloud** computing infrastructure.
* Updated Python scripts to match training data with database stored in **AWS Cloud Search** for further classification.
* Deployed **Snowflake** following best practices and provided expertise in data warehousing, particularly with Snowflake.
* Utilized advanced Snowflake features to optimize **data workflows** and enhance performance.
* Collaborated with **cross-functional** teams to seamlessly integrate Snowflake within the overall technology landscape.
* Implemented data modeling techniques to ensure **optimal storage, organization, and retrieval** of data in Snowflake.
* Conducted thorough testing and validation of the **ETL pipeline** to ensure data integrity and adherence to business requirements.
* Utilized **Terraform** scripts from **Jenkins** for provisioning AWS infrastructure.
* Leveraged **Docker** and **Kubernetes** for secure shipping, running, and deploying applications in containers to accelerate the build and release engineering process.
* Configured Sqoop tasks to import data from **DB2 to HDFS**.
* Implemented **business logic**, processed **XML**, exchanged data, and created visuals using **Python** and **Django**.
* Implemented **Apache Kafka** to efficiently aggregate web log data from diverse servers, enabling downstream data analysis and engineering, while also contributing to security enhancements and performance optimization.
* Mastered the intricacies of architecting data intelligence solutions centered around the **Snowflake Data Warehouse**, actively contributing as a developer to enhance Snowflake-based solutions, while integrating **Azure Synapse** for advanced data warehousing capabilities.
* Demonstrated proficiency in handling a variety of data formats, including **Avro, Parquet, RCFile, and JSON**, by developing User-Defined Functions (**UDFs**) in **Hive**, enhancing data processing capabilities.
* Designed and implemented custom **UDFs** in **Python**, streamlining data sorting and preparation tasks, further enhancing data processing efficiency.
* Leveraged the power of the **Spark SQL API** within **PySpark** to streamline data extraction, loading, and **SQL queries,** optimizing data processing pipelines, and incorporating Azure Synapse for data analytics and insights.
* Played a key role in developing Custom Loaders and Storage Classes in **PIG** to process various data formats such as **JSON, XML, and CSV,** while also generating Bags for processing using Pig. Additionally, developed **Sqoop** and **Kafka** Jobs for efficient data loading into **HDFS and Hive**.
* Developed Oozie coordinators to effectively schedule **Hive scripts**, creating streamlined data pipelines to optimize data **workflows**.
* Implemented multiple **MapReduce** Jobs using **PySpark** and **NumPy**, integrating Jenkins for Continuous Integration, ensuring code quality and reliability.

**Environment:** Snowflake, SQL, ETL, Snowpipe, Time Travel, Task, AWS (EC2, AWS Glue, Athena, Redshift, Lambda, S3, RDS, SNS, IAM),Python **(pandas**, NumPy, Seaborn, SciPy, matplotlib, sci-kit-learn**)**, GIT, Agile, Tableau, Shell Scripting, JSON, Jira, Agile, Spark, Docker, Azure Storage, Blob Storage, Azure SQL Server, Spark SQL API, PySpark, ETL, Apache Storm, Kafka, Python-based API.

**Client:** **vInnovate Technologies, India**  **Jun 2015 – May 2018**

**Role:** **Big Data Engineer**

**Responsibilities:**

* Transforming business problems into **Big Data solutions** and defining Big Data strategy and Roadmap.
* **Installing**, **configuring**, and **maintaining** **Data Pipelines** to ensure optimal performance and reliability.
* Designing the business requirement collection approach based on the project scope and **SDLC methodology** to ensure alignment with organizational objectives and standards.
* Extracting files from **Hadoop** and **transferring** them on a daily or hourly basis into **S3** for efficient data storage and accessibility.
* Authoring **Python (PySpark)** Scripts for custom **UDFs (User Defined Functions)** for versatile data manipulations, merges, aggregations, and data cleaning tasks, ensuring data quality and consistency.
* Developing Spark applications using **PySpark** and **Spark-SQL** for seamless data extraction, transformation, and aggregation from various file formats, facilitating insightful analysis and decision-making.
* Utilizing **Spark Streaming** to process real-time data from **Kafka** and store it in **HDFS**, along with employing NoSQL databases such as **HBase** and **Cassandra** for efficient data handling.
* Developing **Kafka** **Automation** **Regression** **Scripts** to validate **ETL** processes across multiple databases, ensuring **data integrity** and reliability across systems.
* Creating and managing batch jobs for processing data received from diverse sources like **Adobe, Kafka, AWS Glue, and No-SQL databases**, ensuring timely and accurate data processing.
* Utilizing Spark Streaming to collect data from **AWS S3 bucket** in near-real-time, performing necessary transformations and aggregations dynamically, and persisting data in **HDFS** for further analysis and reporting.
* Experience in developing **Spark** applications using **Spark-SQL** in **Databricks** for data extraction, transformation, and aggregation from multiple file formats, facilitating comprehensive data analysis and insights generation.
* Conducting various **AWS Glue performance** optimizations such as distributed cache usage, Partitioning, **Bucketing** in **Hive**, and Map Side joins to enhance data processing efficiency and performance.
* Creating **Kafka** **Pipelines** in **ADF** using Linked **Services/Datasets/Pipelines** to efficiently Extract, Transform, and Load data from different sources, ensuring seamless data integration and processing.
* Implementing automated deployment processes and configuration management using **CI/CD pipelines** with tools like **Azure DevOps and Google Cloud Build**, ensuring efficient and consistent deployment of data solutions.
* Orchestrating complex data workflows and **ETL pipelines** using **Apache Airflow**, incorporating features like branching, scheduling, and task dependencies for efficient data processing and management.
* Collaborating with data engineering teams to design and implement real-time data processing pipelines using streaming technologies such as **Apache Kafka and Google Cloud Pub/Sub,** ensuring robust and scalable data processing capabilities.
* Implementing data governance policies and access controls using **Azure Active Directory** and **Google Cloud Identity** and **Access Management (IAM)** for secure data management and compliance with regulatory requirements.
* Conducting performance tuning and optimization of **SQL** and **NoSQL** **databases** including **Azure SQL** **Database, Google Cloud SQL, and Cosmos DB** to meet stringent SLAs and business requirements, ensuring optimal data processing efficiency and performance.
* Leading the migration of on-premises data infrastructure to **Google Cloud Platform**, leveraging services like BigQuery, Dataflow, and Dataproc for scalable data processing and analytics, ensuring smooth transition and minimal disruption.
* Implementing automated data ingestion pipelines using tools like **Apache NiFi** and **Google Cloud** **Dataflow** to efficiently ingest and process large volumes of data from diverse sources, ensuring timely and accurate data processing.
* Designing and implementing data lake architectures on cloud storage platforms such as **Azure Blob Storage and Google Cloud Storage,** incorporating data partitioning and optimization techniques for efficient data storage and retrieval.
* Developing **Apache Spark applications** for data processing from various streaming sources, ensuring real-time data analysis and insights generation.
* Working in **Azure environment** for development and deployment of Custom **Hadoop Applications,** leveraging the power of cloud computing for scalable and flexible data solutions.
* Implementing **Apache Airflow** for authoring, scheduling, and monitoring **Data Pipelines**, ensuring efficient orchestration and management of data workflows.
* Conducting extensive analysis on large and critical datasets, leveraging a suite of cutting-edge technologies including **Hadoop, HBase, MapReduce, Hive, Zookeeper, and Spark,** to derive actionable insights and drive informed decision-making processes.
* Spearheaded the migration of 20 on-premises servers and 5 databases to the **AWS cloud**, encompassing more than 100 terabytes of data, ensuring minimal downtime and zero data loss.
* Demonstrated mastery in **test-driven development (TDD)** and hands-on experience with **Continuous Integration/Continuous Deployment (CI/CD)** systems, focusing on **Gitlab CI**, ensuring reliable code deployment.
* Proficient in handling diverse datasets comprising structured and unstructured data, utilizing various data **manipulation** techniques and tools to extract valuable insights efficiently.
* Utilized **NoSQL databases** like **Cassandra** for real-time data storage and analysis.
* Demonstrated a consistent track record in **ETL** (**Extracting, Transforming, and Loading**) of data for analytical discovery and production solutions across multiple platforms, ensuring seamless integration and data flow along the **entire pipeline**.
* Developed and maintained serverless applications using **AWS Lambda, DynamoDB, and S3**.
* Engineered scalable, cloud-based data lakes **on AWS and GCP, leveraging S3, Glue, and Dataflow**.
* Implemented Informatica PowerCenter to streamline **ETL processes** for a high-volume data warehouse, resulting in a **50% reduction** in data processing time.
* Executed complex **data migration** projects from legacy systems to modern data platforms using Informatica, ensuring data integrity and minimizing downtime.
* Developed custom scripts and mappings in **Informatica** to handle intricate **data transformations** and business logic, improving **data accuracy** by 30%.
* Conducted performance tuning and optimization of Informatica workflows, reducing resource consumption and enhancing system scalability.
* Installed and configured **Hadoop, MapReduce, HDFS, and AWS** for multiple **MapReduce** jobs in Hive.
* Implemented Spark using **Python** and **Spark SQL API** for accelerated data processing.

**Environment:** Python, AWS Services (EC2, Route53, S3, EMR, RDS, Lambda, Dynamo DB, SNS, SQS, IAM, CloudFormation, SageMaker, AWS CLI), Hadoop, Hive, HBase, HDFS, Pig, Apache Nifi, Sqoop, Spark, PySpark, Spark SQL, Scala, Kafka, Airflow, Snowflake, ETL Informatica, SQL, Unix Shell Scripting, Tableau, Jira

**Client -** **Codash Solutions, India Jun 2013 – May 2015**

**Role: Data Engineer**

**Responsibilities:**

* Developed **Nifi** data types in a **Docker** **container** environment.
* Designed data models for **ODS, OLTP, and OLAP** systems using **ER Studio**.
* Wrote **UNIX scripts** and **automated ETL** processes using **UNIX scripting**.
* Consumed and processed data from various sources using **Hadoop, MapReduce Frameworks, HBase, and Hive.**
* Worked on batch processing using **Apache Spark** and **Elasticsearch**.
* Ported **PIG scripts** and **MapReduce** programs to **Spark Data framework APIs and Spark SQL APIs**.
* Created external **Hive tables** and implemented data pipeline’s using **Oozie**.
* Worked with **NoSQL** databases like **HBase** for real-time data analysis.
* Collaborated with stakeholders for **data mapping**, planning, and auditing.
* Used Sqoop for importing data into **HDFS**/**Hive** from **DB2**.
* Developed scripts for **data import** using **Talend**.
* Handled data transfer from servers to **HDFS** using **Apache Flume**.
* Improved Pig scripts for **data analysis** and refined data storage in **DB2**.
* Worked on **Hadoop log management** and auditing.
* Participated in project team meetings, provided technical support, and solved **test problems**.
* **Orchestrated** data integration projects using a range of tools such as **Talend, Apache NiFi,** ETL processes ensuring seamless connectivity between disparate systems.
* Conducted assessments of various data integration tools, recommending optimal solutions based on project requirements and scalability needs.

**Environment**: Python, UNIX, Hadoop, PL/SQL, Docker, Nifi, Elasticsearch, HDFS, Hive, DB2, Talend, Oozie, HBase