**Yashwanth**

**Sr. Big Data Engineer**

[**@gmail.com**](mailto:Nvarunkumar2905@gmail.co9013167572m)

**+1\*\*\*\*\*\*\*\*\*\*\*\*\***

**PROFESSIONAL SUMMARY :**

Overall 10 years of accomplished experience as a seasoned Big Data Engineer, I possess adept proficiency in data extraction, transformation, and analysis, employing a spectrum of robust big data tools and technologies. My expertise extends to a deep understanding of Hadoop Architecture, its ecosystem components, as well as the intricate intricacies of PySpark Architecture and its elements including GCP, Data Warehouse, Google Big Query. HDFS, Hive, Spark Context, Spark SQL, and Spark Streaming.

* Mastery in orchestrating the data modeling of comprehensive Enterprise Data Warehouses, actively participating in end-to-end developmental phases.
* Profound grasp of data architecture design encompassing Hadoop/Spark pipelines, data modeling, data mining, machine learning, and data processing.
* Diverse industry engagement, proficiently utilizing Big Data technologies including Hadoop, Spark, MapReduce, Kafka, Hive, Sqoop, and HBase within a Hadoop working environment.
* Lead data engineer with emphasis on client/server, real time transaction processing and large data warehousing projects covering all the major roles from determining business & software requirements to developing and testing them. This also includes hands on experience on data migration and conversion of large enterprise scale projects.
* Experience as a Google Cloud Data Architect used Google related services, Cloud Storage, Cloud SQL, Big query, Cloud Pub/Sub, Cloud Composer, Dataflow, Data proc.
* Good experience in utilizing GitLab DevOps.
* Diversified domain experience in healthcare, insurance, telecom and financial.
* Have clear understanding of Data Warehousing concepts with emphasis on ETL and Life Cycle Development including requirement analysis, design, development, testing, implementation, maintenance, and support.
* Mastery in AWS services encompassing S3, EC2, SQS, RDS, Neptune, EMR, Kinesis, Lambda, Event Bridge, Glue, Redshift, Athena, DynamoDB, Elasticsearch, Service Catalog, CloudWatch, and IAM.
* Proficient application of AWS and Big Data technologies, leveraging DynamoDB, Kinesis, S3, HIVE/Spark, to architect robust infrastructure facilitating seamless data extraction, transformation, and loading from diverse data sources, encompassing both NoSQL and SQL.
* Accomplished in Python-based data manipulation, leveraging libraries like NumPy, SciPy, and Pandas for data analysis and numerical computations.
* Proficiency in Spark, utilizing Spark Core, Spark Context, Spark SQL, Data Frame, Pair RDDs, and Spark Streaming for intricate data processing.
* Adept at crafting Spark applications within Databricks for comprehensive data extraction, transformation, and aggregation across diverse file formats, revealing insights into customer usage trends.
* Extensive experience in optimizing and fine-tuning Hive and Spark operations.
* Competence in business and data analysis, data profiling, migration, integration, governance, and metadata management.
* Proficiency in designing insightful reports and dashboards using visualization tools like Power BI and Tableau.
* Proficient scripting skills in Python, and Shell.
* Exceptional communication, interpersonal, problem-solving skills, coupled with strong teamwork capabilities. Quick adaptability to new environments and emerging technologies.

**Education Details:**

* Masters in information systems from Christian Brothers University.
* Bachelor of technology from Archarya University, India.

**Technical Skills:**

**Big Data Technologies:** HDFS, MapReduce, Hive, Sqoop, Oozie, Zookeeper, Kafka, Apache Spark, Spark Streaming

**Cloud Services :** GCP, Big Query, Dataflow, Pub/Sub, Dataprep, Storage, Dataproc, Power BI, Airflow, Azure, Snowflake

**Languages:** Java, SQL, PL/SQL, Python, HiveQL, Scala

**Web Technologies:** HTML, CSS, JavaScript, XML, JSP, Restful, SOAP

**Operating Systems:** Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS

**IDE & Build Tools:** Eclipse, Visual Studio

**Databases:** MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse, MS Excel, MS Access, Oracle 11g/12c, Cosmos DB

**PROFESSIONAL EXPERIENCE:**

**Client : UBS, NY Oct 2022 – Present**

**Role: GCP Data Engineer.**

**Responsibilities:**

* Led the design and implementation of data pipelines using Google Cloud services such as Dataflow, Big Query, and Pub/Sub, improvement in data processing efficiency.
* Implemented data governance best practices, including metadata management, lineage tracking, and data quality monitoring.
* Collaborated with data scientists to deploy machine learning models in a production environment on Google Cloud.
* Architected and implemented a real-time analytics platform using Apache Spark and Scala, enabling the processing of large-scale data streams.
* Implemented robust data security measures and access controls on GCP, ensuring compliance with industry standards and safeguarding sensitive information.
* Orchestrated migration to GCP, reducing infrastructure costs by 25%.
* Implemented serverless functions with Cloud Functions for enhanced scalability. Configured IAM roles and permissions for secure resource access. Set up proactive monitoring and logging using Stack driver for improved reliability.
* Executed complex analytical queries on large datasets for strategic insights.
* Gained hands-on experience with GCP services like Big Query, Cloud Dataflow, and Cloud Storage.
* Developed and maintained ETL pipelines using Google Cloud Dataflow, Cloud Composer, and Apache Beam for real-time and batch data processing.
* Optimized SQL queries, reducing execution time by 30%.
* Proficiently managed Databricks clusters and SQL data warehousing, optimizing query performance and resource utilization, and effectively handled Delta and Parquet files to support data lake operations.
* Led the development of data engineering pipelines on Databricks, encompassing data extraction, transformation, and loading (ETL) processes to meet business requirements and analytics needs.
* Demonstrated extensive expertise in SQL and PL/SQL, adept at handling complex joins, aggregation functions, and DBT (Data Build Tool) for efficient data transformation and modeling.
* Utilized Python, data frames, and Spark for advanced data processing, enabling complex data manipulations and insights generation from large datasets.
* Successfully transitioned ETL pipelines to Azure Databricks and Azure SQL Data Warehouse. Managed access privileges and permissions for the databases.
* Designed data models for efficient storage in a centralized data warehouse.
* Developed and maintained ETL pipelines for timely data updates.
* Implemented strategies for managing historical data for trend analysis.
* Established data quality checks for accuracy and integrity assurance.
* Developed comprehensive Azure Data Factory components, Integration Run Time (IR), File System Data Ingestion, and Relational Data Ingestion.
* Implemented a centralized Azure Key Vault to securely manage and reference secrets within Azure Data Factory and Azure Databricks, with added logic apps for email notifications.
* Utilized Azure Data Factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics for Extract, Transform, and Load (ETL) operations to various Azure services.
* Utilized Bigtable and Cloud SQL for NoSQL and relational data storage solutions.
* Conducted performance tuning, cost optimization, and capacity planning for GCP resources.
* Formulated advanced Hive queries to extract meaningful insights from Spark job data through intricate table joins.
* Developed the end-to-end data pipeline in spark using python to ingest, transform and analyses data.
* Worked on proof of concept in the Google cloud platform (GCP) to migrate the data from On-premises environment into the Big Query.
* Developed PySpark scripts in Azure Databricks to construct data frames and datasets, catering to business-specific enrichment and transformation requirements.
* Assisted in the development of data pipelines for batch and real-time data processing using GCP technologies.
* Established robust data pipelines via Azure Data Factory, along with a custom alerts mechanism for comprehensive monitoring.
* Integrated GCP services with third-party tools and applications to streamline data flows.
* Trained client teams on GCP data engineering best practices and provided ongoing support.
* Designed and implemented Snow pipe for continuous data loading from Azure Data Lake to Snowflake.
* Implemented Spark streaming configuration for real-time data ingestion from Apache Flume, storing the data in Azure Table Storage, and utilizing Delta Lake for versatile processing and analytics.
* Contributed to the establishment of a CI/CD pipeline using Jenkins, Terraform, and Azure DevOps, ensuring efficient code deployment.
* Proficiently utilized Kusto Query Language (KQL) for Azure Log Analytics, creating custom alerts to facilitate proactive monitoring.
* Designed interactive Power BI dashboards and reports, incorporating robust security measures such as row-level access based on business requirements.

**Environment:** GCP, Big Query, Dataflow, SQL, GCP, Azure SQL Database, Data Storage Explorer, Azure Synapse, SQL Server Management Studio (SSMS), Azure Studio, Terraform, PySpark, Python, Hive, Kusto Query Language (KQL), Apache Airflow, Apache Flume, Delta Lake, Power BI.

**Client : CVS, Woonsocket, RI Jan 2021 – Sep 2022**

**Role: GCP Data Engineer.**

**Responsibilities:**

* Developed and maintained ETL processes using Google Cloud Data prep and Apache Beam, resulting in improved data quality and consistency.
* Build data pipelines in airflow in GCP for ETL related jobs using different airflow operators.
* Create firewall rules to access Google data procs from other machines, processed and load bound and unbound Data from Google pub/subtopic to big query using cloud Dataflow with Python.
* Setup GCP Firewall rules to ingress or egress traffic to and from the VM's instances based on specified configuration and used GCP cloud CDN (content delivery network) to deliver content from GCP cache locations drastically improving user experience and latency.
* Responsible for creating a Delta Lake to support the ODM model, utilizing Databricks and its features for data versioning and management. This involved ensuring data quality, reliability, and accessibility.
* Integrated Big Query with Tableau for dynamic data visualization.
* Set up consumption patterns using the Databricks Unity catalog, enabling efficient data discovery and access for end-users, data scientists, and analysts. This required expertise in cataloging and data organization.
* Created and managed Airflow Directed Acyclic Graphs (DAGs) for job orchestration and scheduling of data pipeline jobs, ensuring the seamless execution of data workflows. This involved coordinating data movements and transformations across various components of the pipeline.
* Developed and tested Spark applications in Scala to compare Spark performance with MapReduce and Hive, leveraging solid Java skills in the design, implementation, and support of big data solutions in Hadoop using Hive, Spark, Drill, Impala, and HBase.
* Implemented batch and streaming pipelines using Azure Data Factory, Databricks, and Spark Streaming, while also applying hands-on experience with UNIX, Teradata, and other relational databases.
* Created RESTful APIs using Flask for efficient communication between frontend and backend systems, enhancing data accessibility.
* Worked on Google Cloud Platform applications like Cloud Storage, Pub/Sub, Cloud functions, Big Query, Cloud SQL and Data flow in designing the Cloud Data warehouse
* Implemented partitioning and clustering, reducing query costs by 20%.
* Designed database schemas, optimized SQL queries, and managed Oracle databases to ensure efficient data storage and retrieval.
* Developed Azure Data Factory pipelines for data extraction, transformation, and loading from various sources, including integration with Snowflake.
* Created Azure Logic Apps for triggering processes based on received emails and file attachments, automating workflow orchestration.
* Built Spark Streaming applications for real-time analytics and data processing, utilizing Java and Scala.
* Designed and developed custom Hive UDFs, using JSON and XMLSerDe for data serialization and deserialization, enhancing data processing capabilities.
* Engineered and maintained data pipelines using GCP services such as Dataflow, Dataprep, BigQuery, and Cloud Storage.
* Worked on migration of ETL processes from Oracle to Hive for data manipulation, optimizing data workflows.
* Implemented Kafka consumers in Scala for consuming data from Kafka topics, enabling efficient data ingestion.
* Collaborated with the Linux server admin team for server maintenance and administration, ensuring a stable infrastructure.
* Participated in agile development methodologies and CI/CD practices, contributing to efficient and collaborative development processes.
* Worked on implementing scalable infrastructure and platform for large amounts of data ingestion, aggregation, integration, and analytics in Hadoop using Spark, Snowflake, and Hive.
* Involved in migrating on Hadoop system to using GCP (Google Cloud Platform), ensuring seamless transition and efficient data management.
* Designed Azure ETL to handle large volumes of data and supports scalable data processing, leveraging Azure's cloud infrastructure to parallelize data processing.
* Integrated on-premises (MYSQL, Cassandra) and cloud data (Blob storage, Azure SQL DB) using Azure Data Factory, applying transformations for loading into Snowflake.
* Developed Spark jobs in Python to perform data transformation, creating Data Frames and Spark SQL, Snowflake, streamlining data processing.
* Experience in building and architecting multiple Data pipelines, end-to-end ETL and ELT processes for Data ingestion and transformation in GCP.
* Created applications logic and functionality in the Azure Databricks, SQL Database, and SQL Datawarehouse environment, ensuring efficient data handling.
* Developed Spark applications using spark libraries to perform ETL transformations, eliminating the need for utilizing ETL tools.
* Worked on a proof of concept in the Google Cloud Platform (GCP) to migrate data from On-premises environments into Big Query, facilitating data accessibility and analysis.
* Designed and implemented streaming solutions using Kafka or Azure Stream Analytics, enabling real-time data processing.
* Monitored end-to-end integration using Azure Monitor, ensuring the reliability and performance of data processing pipelines.

**Environment:** GCP, Big Query**,** Snowflake, Azure Data Lake, Azure Databricks, Azure Data Factory, SQL, Azure SQL Database, Data Storage Explorer, Azure Synapse, SQL Server Management Studio (SSMS), Azure Studio, Terraform, PySpark, Python, Hive, Kusto Query Language (KQL), Apache Airflow, Apache Flume, Delta Lake, Power BI

**Client : Citi Bank / Dallas, Tx Aug 2019 – Dec 2020**

**Role: Hadoop developer**

**Responsibilities:**

* Developing data ingestion pipelines using Talend and bash scripting, incorporating prominent big data technologies such as Hadoop, Hive, Spark, and Kafka.
* Crafting scalable and secure data pipelines designed to accommodate extensive datasets.
* Defining requirements for new data source ingestion, encompassing lifecycle management, data quality checks, transformations, and metadata enrichment.
* Leveraging Flume, Sqoop, Pig, and Java MapReduce to create data pipelines for ingesting customer behavioral data into HDFS, enabling subsequent analysis.
* Playing a pivotal role in data quality management by integrating rigorous data quality checks within the data pipelines.
* Continuously enhancing the Data Ingestion Framework to establish more resilient and secure data pipelines.
* Implementing data streaming capabilities through Kafka and Informatica for diverse data sources.
* Executing SQOOP implementations to facilitate seamless data transfers between various RDBMS sources and Hadoop systems.
* Skillfully handling diverse storage formats (Avro, Parquet) and databases (Hive, Azure SQL).
* Employing optimization techniques, including bucketing and partitioning, to enhance query performance within Hive.
* Strategically creating and managing partitions and buckets within Hive tables.
* Managing data inbound and outbound requests via the big data platform, ensuring smooth data flow.
* Leveraging Sqoop for the efficient transfer of data between relational databases and Hadoop.
* Implementing JILs to automate jobs within the production cluster.
* Collaborating closely within SCRUM teams to consistently deliver user stories within specified Sprint timeframes.
* Analyzing and resolving production job failures across a range of scenarios.
* Designing and implementing UNIX scripts to define workflow use cases and automate job processes.

**Environment:** Talend, Hadoop, Hive, Spark, Kafka, Flume, Sqoop, Pig, Java MapReduce, Informatica, Avro, Parquet, Azure SQL, Hive, Kafka, JIL (Job Information Language), UNIX scripting.

**Client : Rmsi Pvt Ltd, INDIA Apr 2015 – Dec 2017**

**Role: Data Engineer/Analyst**

**Responsibilities:**

* Spearheaded a research study to identify gaps, ensure data accuracy and integrity in data modeling, ingestion, evaluation, and analytical procedures which resulted in improving productivity 40%.
* Built real-time distributed data pipelines that optimize data processing and analysis using Hadoop, Hive, PySpark, Kafka, and Airflow.
* Identified opportunities for Unix system automation and process improvement.
* Coordinated hardware/software installations and upgrades to ensure proper Unix system performance.
* Installed and maintained Unix security patches on the development systems.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure Data warehouse, write-back tool and backwards.
* Developed Pyspark applications using Databricks, Synapse and orchestrated them using ADF pipelines.
* Developed a Spark Streaming application that consumes 3GB of data per hour from Kafka in real-time, processes the data, and loads it into Synapse tables which resulted in saving 150 hours of manual intervention.
* Built interactive dashboards and created reports using Tableau to analyze the water quality and track various
* KPIs which helped to gain deeper insights and improve the decision-making process by 50%.
* Involved in Analysis, Design, and Implementation/translation of Business User requirements.
* Worked on collection of large sets of Structured and Unstructured data using Python Script.
* Worked on creating DL algorithms using LSTM and RNN.
* Actively involved in designing and developing data ingestion, aggregation, and integration Cloudera.
* Developed Sqoop scripts to import export data from relational sources and handled incremental loading on the customer, transaction data by date.
* Experience in creating Hive Tables, Partitioning and Bucketing.
* Identified inconsistencies in data collected from different sources.
* Designed object model, data model, tables, constraints, necessary stored procedures, functions, triggers, and packages for Oracle Database.
* Wrote Spark applications for Data validation, cleansing, transformations, and custom aggregations.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Worked on installing cluster, commissioning & decommissioning of Data node, Name node high availability, capacity planning, and slots configuration.
* Developed Spark applications for the entire batch processing.
* Stored the time-series transformed data from the Spark engine built on top of a Hive platform to S3 and Redshift.

**Environment:** R, SQL server, Python, Shell Scripting, Airflow, Oracle, Power BI,HDFS, Scala,HBase, Informatica, ETL, Hive, Impala, Pig, Sqoop, Azure Data Factory, Synapse, Databricks, Tableau.

**Cinar Software Technologies Pvt. Ltd, Hyderabad, India Aug 2012 - March 2015**

**Data Engineer**

* Responsibilities:
* Spearheaded a research study to identify gaps, ensure data accuracy and integrity in data modeling, ingestion, evaluation, and analytical procedures which resulted in improving productivity 40%.
* Built real-time distributed data pipelines that optimize data processing and analysis using Hadoop, Hive, PySpark, Kafka, Airflow.
* Developed a Spark Streaming application that consumes 3GB of data per hour from Kafka in real-time, processes the data, and loads it into Hive tables which resulted in saving 150 hours of manual intervention.
* Built interactive dashboards and created reports using Tableau to analyze the water quality and track various
* KPIs which helped to gain deeper insights and improve the decision-making process by 50%.

**Environment:** python, Informatica, Shell Scripting, Django, Airflow, Scala, RESTful web service, AWS Glue, ETL, Snowflake, MySQL, PostgreSQL, Visio, SQL Server Management Studio, AWS.