**Alapati Yashasvi**

Sr. Cloud Data Engineer

**Phone: +1 201-470-5744 | Email: yashasvi.a22@gmail.com**

**PROFESSIONAL SUMMARY**

Results-driven Sr Cloud Data Engineer with over 11 years of experience architecting, developing, and optimizing data solutions across leading cloud platforms including Gcp and Azure. Expert in designing scalable data pipelines, implementing advanced ETL/ELT workflows, and managing large-scale data warehouse and business intelligence systems. Proficient in cloud-native tools such as Azure Data Factory, Azure Databricks, AWS Redshift.Expert in architecting and implementing cloud-based data pipelines using Azure Data Factory, Databricks, and Azure Synapse Analytics. Adept at collaborating with cross-functional teams-including data analysts, data scientists, and business stakeholders-to deliver high-performance, secure, and scalable data warehouse solutions. Proven track record of building robust data models, maintaining data marts, and supporting advanced analytics and reporting initiatives. Highly skilled in both structured and unstructured data environments, with deep knowledge of data warehousing methodologies (Kimball, Inmon) and hands-on expertise in Snowflake, SQL Server, Oracle, and modern BI tools.

* Results-driven Data Engineer with expertise in designing and implementing scalable data ingestion pipelines using Azure Data Factory.
* Designed and developed end-to-end ETL workflows for ingesting, transforming, and loading data from diverse sources (databases, APIs, flat files) into enterprise data warehouses using Azure Data Factory v2 and IBM DataStage.
* Automated and optimized ETL pipelines, reducing data processing times and improving reliability by implementing best practices for error handling, auditing, and performance tuning.
* Proficient in leveraging Azure Databricks and Spark for distributed data processing and transformation tasks.
* Skilled in ensuring data quality and integrity through validation, cleansing, and transformation operations.
* Adept at designing cloud-based data warehouse solutions using Snowflake on Azure, optimizing schemas, tables, and views for efficient data storage and retrieval.
* Experience with Snowflake Multi - Cluster Warehouses.
* Understanding of SnowFlake cloud technology.Experience in using Snowflake Clone and Time Travel.
* Participates in the development improvement and maintenance of snowflake database applications
* Build the Logical and Physical data model for snowflake as per the changes required
* Define roles, privileges required to access different database objects.In-depth knowledge of Snowflake Database, Schema and Table structures.
* Collaborative approach in working with data analysts and stakeholders to implement appropriate data models and structures.
* Strong expertise in optimizing Spark jobs and leveraging Azure Synapse Analytics for big data processing and analytics.
* Proven track record in performance optimization and capacity planning to ensure scalability and efficiency.
* Experienced in developing CI/CD frameworks for data pipelines and collaborating with DevOps teams for automated pipeline deployment.
* Proficient in scripting languages such as Python and Scala.
* Skilled in working with Hive, SparkSQL, Kafka, and Spark Streaming for ETL tasks and real-time data processing.
* Have a good working experience in Hadoop, HDFS, Map-Reduce, Hive, Teg, Python, PySpark.
* Hands on working experience and developing large scale data pipelines using spark and hiveLed the design, development, and implementation of analytics solutions on Microsoft Azure Synapse Analytics Platform, supporting large-scale data warehousing and analytics needs for enterprise clients.
* Developed robust ETL/ELT pipelines using Azure Data Factory, Azure Synapse Pipelines, and Databricks Notebooks to collect, integrate, and transform high-volume datasets from diverse sources, ensuring data quality and optimal pipeline performance.
* Utilized Azure Data Lake Storage, Azure Synapse Analytics, Azure Spark Pools, Databricks, Cosmos DB, and CDC tools to architect and maintain scalable, secure, and high-performing data solutions.
* Automated repeatable data preparation tasks using Python and Spark, significantly reducing manual intervention and improving operational efficiency.
* Collaborated with business unit IT teams and stakeholders to gather analytics requirements, document solution options, and provide clear recommendations tailored to Oil & Gas industry processes.
* Partnered with cloud service providers, data science, and data governance teams to deliver integrated analytics solutions aligned with business objectives and compliance standards.
* Provided ongoing production support, troubleshooting, and enhancements for existing analytics solutions, ensuring high availability and reliability.
* Participated in scaled agile framework (SAFe) for data engineering product delivery, actively engaging in agile ceremonies and contributing to sprint planning and execution.
* Experience in using Apache Sqoop to import and export data from HDFS and Hive.
* I am experienced in setting up workflow using Apache Oozie workflow engine for managing and scheduling Hadoop jobs, and importing and exporting the data using SQOOP from HDFS to Relational Database systems.
* I have optimized query performance in Hive using bucketing and partitioning techniques, and have extensive hands-on experience tuning spark Jobs.
* Highly proficient in Agile methodologies, including JIRA for project management and reporting.

**EDUCATION**

* Bachelors of Technology in Computer Science.

**TECHNICAL SKILLS**

| Azure Services | Azure data Factory, Azure Data Bricks, snowflake, Logic Apps, Functional App, Snowflake, Azure DevOps |
| --- | --- |
| Big Data Technologies | MapReduce, Hive, Teg, Python, PySpark, Scala, Kafka, Spark streaming, Oozie, Sqoop, Zookeeper |
| Hadoop Distribution | Cloudera, Horton Works |
| Languages | SQL, PL/SQL, Python, HiveQL, Scala,core Java |
| Web Technologies | HTML, CSS, JavaScript, XML, JSP, Restful, SOAP |
| Gcp | Google Cloud Dataflow,Google BigQuery,Google Cloud Storage (GCS),Google Cloud Pub/Sub,Google Cloud Functions,Google Cloud Composer (for orchestration),Google Cloud AI. |
| LLM | machine learning, deep learning, natural language processing, computer vision, data analysis, model deployment, reinforcement learning, and statistical analysis. |
| Aws | Amazon Redshift, S3, Lambda, Glue, RDS, DynamoDB, CloudFormation, IAM, Kinesis, Data Pipeline. |
| Operating Systems | Windows (XP/7/8/10), UNIX, LINUX, UBUNTU, CENTOS. |
| Build Automation tools | Ant, Maven ,DBT , schemas  |
| Version Control | GIT, GitHub. |
| IDE & Build Tools, Design | Eclipse, Visual Studio. |
| Databases | MS SQL Server 2016/2014/2012, Azure SQL DB, Azure Synapse. MS Excel, MS Access, Oracle 11g/12c, Cosmos DB |

**WORK EXPERIENCE**

**Role: Sr Data Engineer | July 2022 - Present**

**Client: Hartford Healthcare, Connecticut**

* Designed and implemented scalable data ingestion pipelines using Azure Data Factory, ingesting data from various sources such as SQL databases, CSV files, and REST APIs.
* Developed data processing workflows using Azure Databricks, leveraging Spark for distributed data processing and transformation tasks.
* Utilized programming languages such as Python and SQL within the Palantir environment to automate data transformations, develop custom logic, and perform advanced analytics.
* Built and optimized data ingestion and preprocessing workflows using BigQuery, Cloud Storage, and Dataflow, enabling seamless integration with Vertex AI for model
* Supported production environments by troubleshooting and resolving issues in Foundry pipelines, ensuring high availability and reliability of mission-critical data solutions.
* Designed and implemented scalable data ingestion pipelines using Azure Data Factory and Databricks, ingesting data from SQL databases, CSV files, and REST APIs.
* Integrated Snowflake with cloud platforms (AWS S3, Azure) and BI tools (Power BI, Tableau) for seamless data processing and visualization.
* Monitored and tuned Snowflake performance using clustering, caching, and virtual warehouses, reducing processing costs and improving efficiency.
* Implemented and managed Snowflake security features, including role-based access control, encryption, and data masking for compliance.
* Oversaw migration of legacy master data systems to Reltio, ensuring data quality, integrity, and minimal business disruption.
* Designed modular dbt projects using Jinja templating to enable reusable, parameterized SQL transformations and macros.
* Integrated dbt with Snowflake and BigQuery, orchestrating ELT workflows and optimizing warehouse resource usage through incremental model materializations.
* Conducted regular dbt run and build activities, ensuring timely data refreshes and accurate downstream reporting.
* Established dbt source freshness testing to monitor data pipeline health and alert on data staleness.
* Led dbt adoption and training initiatives, upskilling team members on best practices and advanced dbt features.
* Documented all dbt model changes and logic using YAML files, ensuring maintainability and auditability of the transformation layer.
* Refactored legacy SQL scripts into modular dbt models, improving code reusability and reducing technical debt.
* Leveraged dbt lineage graphs to visualize and troubleshoot dependencies across complex data pipelines.
* Implemented dbt source and model tagging for better organization, discoverability, and governance of data assets.
* Collaborated with analysts and data scientists to translate business requirements into robust dbt models and tests.
* Utilized dbt snapshots to track historical changes and support slowly changing dimension use cases.
* Automated dbt job scheduling and monitoring using dbt Cloud and Airflow integrations.
* Maintained dbt project documentation and onboarding guides for new team members.
* Applied dbt incremental models to optimize large-scale data transformations and reduce processing time.
* Participated in code reviews and enforced dbt project standards to ensure code quality and reliability
* Utilized Databricks Unity Catalog for centralized data governance, fine-grained access control, and metadata management.
* Built and managed data warehouses and data lakes on Azure Synapse Analytics and Databricks, optimizing for performance and cost.
* Automated ETL/ELT pipelines with Delta Live Tables and Databricks Jobs, integrating CI/CD using Azure DevOps and GitHub.
* Conducted performance tuning, partitioning, and indexing in Databricks and Synapse to reduce query latency and improve scalability.
* Developed and maintained PySpark and Scala scripts for complex data transformations and analytics.
* Implemented data lineage and metadata management solutions to track data flow and ensure compliance.
* Integrated and automated cloud infrastructure provisioning using Terraform and Azure-native tools.
* Developed and enforced data security, privacy, and compliance policies using Unity Catalog and Azure security best practices.
* Designed, developed, and maintained HL7 interfaces (ADT, ORU, ORM, DFT) using integration engines such as Mirth Connect, Cloverleaf, and Rhapsody.
* Led HL7 integration projects for EMR/EHR systems (Epic, Cerner, Allscripts).
* Developer and implemented highly available and scalable data pipelines using AWS services such as AWS Glue, AWS Lambda, AWS S3, and AWS Kinesis to process and deliver data in real time and batch modes.
* Developed and maintained ETL workflows with AWS Glue and Python, ensuring reliable and efficient data transformation and integration across multiple AWS environments.
* Architected robust data lake and data warehouse solutions on AWS, leveraging AWS S3 for storage, AWS Redshift for analytics, and AWS RDS for transactional workloads.
* Automated infrastructure provisioning and configuration management using AWS CloudFormation and Terraform, enabling repeatable and consistent AWS deployments.
* Managed and optimized AWS EC2 instances for compute workloads, including auto-scaling, monitoring, and cost optimization strategies
* Implemented secure networking architectures with AWS VPC, subnets, security groups, and Route 53 for DNS management, ensuring compliance and data .
* Utilized AWS IAM for fine-grained access control, managing user roles, policies, and multi-factor authentication to safeguard AWS resources.
* Built serverless data processing workflows using AWS Lambda and AWS Step Functions, reducing operational overhead and infrastructure costs.
* Collaborated with DevOps teams to automate CI/CD pipelines for Databricks notebooks and jobs.
* Leveraged Databricks certification and ongoing learning to stay current with Azure and Databricks best practices.
* Worked in Agile teams, participating in sprint planning, daily stand-ups, and cross-functional collaboration.

**Environment**: Azure Databricks, Data Factory, Snowflake, Logic Apps, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, shall scripting, GIT, JIRA, Jenkins, kafka, ADF Pipeline, Power Bi.

**Role: Sr Cloud Data Engineer | Nov 2021 – June 2022**

**Client: pacific gas and Electric Company , California .**

* Implemented end-to-end data pipelines using Azure Data Factory to extract, transform, and load (ETL) data from diverse sources into Snowflake.
* Led end-to-end EDI implementations, optimizing data exchange workflows and ensuring compliance with X12, EDIFACT, and HL7 standards.
* Led a cross-functional team to develop a custom library of GCP data pipeline templates, accelerating GCP pipeline development by 35% and standardizing processes.
* Implemented robust GCP backup and disaster recovery strategies, ensuring business continuity and minimal downtime in the GCP environment.
* Designed and implemented robust ETL pipelines using tools like Informatica, Talend, or DataStage, ensuring accurate and timely data integration.
* Integrated advanced GenAI models (e.g., LLMs, vision transformers) into Vertex AI pipelines, achieving a 20% improvement in predictive performance for enterprise applications.
* Implemented robust ETL processes and data integration solutions using Palantir Foundry Pipelines, ensuring high data quality, lineage, and governance across multiple data sources.
* Developed and deployed analytical dashboards and applications using Foundry tools such as Quiver and Contour, enabling real-time business intelligence and KPI monitoring.
* Supported migration of on-premises databases and applications to AWS, ensuring minimal downtime and data integrity throughout the process.
* Collaborated with cross-functional teams to design and deliver end-to-end data solutions on AWS, aligning with business requirements and best practices.
* Conducted performance tuning and cost optimization reviews across AWS resources, identifying opportunities for efficiency improvements.
* Maintained up-to-date expertise in the latest AWS features, certifications, and cloud engineering trends to drive continuous improvement and innovation
* Collaborated with cross-functional teams-including data scientists, analysts, and business users-to gather requirements, translate them into technical solutions, and deliver scalable data products on the Palantir platform.
* Conducted data quality checks, implemented validation routines, and maintained comprehensive documentation of all Palantir Foundry data pipelines.Developed and maintained over 50 dbt models to standardize and automate data transformations, ensuring consistency and scalability across the data warehouse.
* Leveraged dbt to centralize business logic, reducing data silos and enabling cross-team collaboration on data pipeline developmentImplemented dbt testing frameworks, including unique, not null, referential integrity, and custom tests, to proactively identify and resolve data quality issues before production deployment.
* Utilized dbt Docs to generate comprehensive documentation for all dbt models, sources, and lineage, improving transparency and stakeholder trust
* Built scalable and optimized Snowflake schemas, tables, and views to support complex analytics queries and reporting requirements.
* Developed data ingestion pipelines using Azure Event Hubs and Azure Functions to enable real-time data streaming into Snowflake.Leveraged Azure Data Lake Storage as a data lake for storing raw and processed data, implementing data partitioning and data retention strategies.
* Utilized Azure Blob Storage for efficient storage and retrieval of data files, implementing compression and encryption techniques to optimize storage costs and data security.
* Integrated Azure Data Factory with Azure Logic Apps for orchestrating complex data workflows and triggering actions based on specific events.
* Implemented data governance practices and data quality checks using Azure Data Factory and Snowflake, ensuring data accuracy and consistency.
* Developed custom monitoring and alerting solutions using Azure Monitor and Snowflake Query Performance Monitoring (QPM) for proactive identification and resolution of performance issues.
* Integrated Snowflake with Power BI and Azure Analysis Services for creating interactive dashboards and reports, enabling self-service analytics for business users.
* Optimized data pipelines and Spark jobs in Azure Databricks for improved performance, including tuning of Spark configurations, caching, and leveraging data partitioning
* Implemented data cataloging and data lineage solutions using tools like Azure Purview and Apache Atlas to provide a comprehensive understanding of data assets and their relationships.
* Collaborated with cross-functional teams including data scientists, data analysts, and business stakeholders to understand data requirements and deliver scalable and reliable data solutions.

**Environment**: Azure Databricks, Data Factory, Logic Apps, Snowflake, Functional App, Snowflake, MS SQL, Oracle, HDFS, MapReduce, YARN, Spark, Hive, SQL, Python, Scala, PySpark, shell scripting, GIT, JIRA, Jenkins, kafka, ADF Pipeline, Power Bi.

**Role: Big Data Developer | April 2020 – Oct 2021**

**Client: CVS Corporation -Rhode Island**

* Developed and maintained data pipelines using Sqoop, Flume, and Kafka to ingest, transform, and process customer behavioral data for analysis.
* Developed and tested CUBS application code using PIC Basic, ensuring alignment with functional specifications and business rules.
* Enhanced existing CUBS modules by writing efficient PIC Basic code, improving reliability and maintainability.
* Led code reviews and quality assurance for CUBS projects utilizing PIC Basic, maintaining high coding standards.
* Created custom utilities in PIC Basic to automate repetitive CUBS tasks, reducing manual effort and errors.
* Integrated PIC Basic code with UniVerse-based CUBS applications, enabling advanced functionality and reporting.
* Troubleshot and resolved defects in CUBS code written in PIC Basic, ensuring system stability and user satisfaction.
* Performed data aggregation and analysis on large-scale datasets using Apache Spark, Scala, and Hive, resulting in improved insights for the business.
* Utilized big data ecosystems such as Hadoop, Spark, and Cloudera to load and transform large sets of structured, semi-structured, and unstructured data.
* Integrated HBase with Hive on the Analytics Zone, creating and optimizing HBase tables for faster and more efficient querying of data.Utilized Hive queries and Spark SQL to analyze and process data, meeting specific business requirements and simulating MapReduce .
* Developed and optimized complex SQL queries for large-scale databases, improving data retrieval and reporting performance by up to 25%.
* Designed and implemented normalized database schemas and data models to support business operations and analytics.
* Managed database migrations and upgrades, ensuring seamless transitions and zero data loss for critical systems.
* Created and maintained stored procedures, triggers, and user-defined functions to automate business logic and data processing.
* Performed data analysis and generated actionable insights through advanced SQL queries, supporting data-driven decision-making.
* Implemented indexing, partitioning, and query optimization techniques to enhance database performance and scalability.
* Ensured database security by setting user roles, permissions, and compliance protocols, resulting in zero security breaches.
* Supported ETL processes by developing SQL scripts for data extraction, transformation, and loading across multiple systems.
* Provided critical support for disaster recovery planning, including backup strategies and recovery testing for business continuity.
* Collaborated with cross-functional teams to integrate new data sources, improve data quality, and deliver reliable reporting solutions
* Migrated data from RDBMS (Oracle) to Hadoop using Sqoop for processing, enhancing data management and processing capabilities.
* Utilized JIRA to manage issues and project workflow, improving project organization and efficiency.
* Collaborated with team members to identify and resolve JVM-related issues, resulting in better performance and system stability.
* Utilized Git as a version control tool to maintain the code repository, ensuring better code management and tracking of changes.

**Environment**: Sqoop, MYSQL, HDFS, Apache Spark Scala, Hive Hadoop, Cloudera, HBASE, Kafka, MapReduce, Zookeeper, Oozie, Data Pipelines, RDBMS, AWS, EC2, Python, PySpark, shell script, Ambari, JIRA.

**Role: Big data Developer | Mar 2018 – Mar 2020**

**Client: United Health Care, MN**

* Prepared an ETL framework using Sqoop, Pig, and Hive to bring in data from various sources and make it available for consumption.
* Processed HDFS data and created external tables using Hive, along with developing scripts for table ingestion and repair for reuse across the project.
* Developed ETL jobs using Spark and Scala to migrate data from Oracle to new MySQL tables.
* Utilized Spark (RDDs, DataFrames, Spark SQL) and Spark-Cassandra Connector APIs for various tasks, including data migration and business report generation.
* Analyzed source data, efficiently handled data type modifications, and used Excel sheets, flat files, and CSV files to generate Power BI ad-hoc reports.
* Analyzed SQL scripts and designed solutions using PySpark.Extracted data from various data sources into HDFS using Sqoop.
* Handled data import from various sources, performed transformations using Hive and MapReduce, and loaded data into HDFS.
* Extracted data from MySQL into HDFS using Sqoop.Implemented automation for deployments using YAML scripts for streamlined builds and releases.
* Worked with Apache Hive, Apache Pig, HBase, Apache Spark, Zookeeper, Flume, Kafka, and Sqoop.
* Implemented data classification algorithms using MapReduce design patterns.
* Extensively worked on creating combiners, partitioning, and distributed cache to enhance the performance of MapReduce jobs.
* Utilized Git and GitHub repositories to maintain the source code and enable version control.

**Environment**: Hadoop, Hive, spark, PySpark, Sqoop, Spark SQL, Shell script, Cassandra, YAML, ETL.

**Role: Data Developer | Feb 2014 – Feb 2018**

**Client: Citi Financial - Tampa, FL**

* Developed a shell script to create staging, landing tables with the same schema as the source and generate the properties, which are used by Oozie jobs.
* Developed Oozie workflow for executing Sqoop and Hive actions and worked with NoSQL databases like HBase in creating HBase tables to load large sets of semi-structured data coming from various sources.Performance optimizations on Spark and Python. Diagnose and resolve performance issues in Spark.
* Involved in developing a roadmap for migration of enterprise data from multiple data sources like SQL Server, provider databases into S3, which serves as a centralized data hub across the organization.
* Loaded and transformed large sets of structured and semi structured data from various downstream systems.Developed ETL pipelines using Spark and Hive for performing various business specific transformations.Building Applications and automating the pipelines in Spark for Bulk loads as well as Incremental Loads of various Datasets.
* Developed and maintained documentation for Power BI models, data sources, and business logic to support governance and knowledge sharing.Provided training and support to business users, enabling self-service analytics and fostering a data-driven culture.
* Conducted data quality checks and implemented data validation rules in Power BI to ensure accuracy and reliability of insights.
* Participated in Agile development processes, including sprint planning, daily stand-ups, and iterative delivery of Power BI solutions.
* Created ad-hoc and scheduled reports for financial, operational, and executive stakeholders, supporting strategic planning and decision-making.
* Integrated custom visuals and R/Python scripts within Power BI for advanced analytics and data science use cases.Managed and administered Power BI workspaces, user permissions, and data refresh policies to ensure secure and efficient report delivery.
* Utilized Power BI to drive a 30% improvement in reporting efficiency and a 20% increase in user engagement through interactive dashboards and optimized data models
* Developed scripts to run Oozie workflows, capture the logs of all jobs that run on cluster and create a metadata table, which specifies the execution times of each job.Converted existing MapReduce applications to PySpark application as part of overall effort to streamline legacy jobs and create new framework.

**Environment**: Hadoop, HDFS, Map Reduce, Hive, HBase, Oozie, Impala, Java (jdk1.8), Cloudera, Python, UNIX Shell Scripting, Flume, Scala, Spark, Sqoop, Kafka, Oracle.