Build an ETL Pipeline on EMR using AWS CDK and Power BI

Business Overview:

Infrastructure as Code (IaC) is a practice of automating the provisioning and management of IT infrastructure using software engineering techniques. It involves creating and maintaining the infrastructure through code rather than manually configuring individual components.

IaC is typically used in cloud computing environments, where it enables developers and operations teams to automate the deployment and management of resources, such as servers, storage, and networking. It enables organizations to achieve more consistent, reliable, and scalable infrastructure that can be easily replicated across multiple environments.

IaC can be implemented using various tools and platforms, such as Terraform, AWS CloudFormation, Azure Resource Manager, and Google Cloud Deployment Manager. These tools enable the creation of infrastructure templates or scripts that define the desired configuration and state of the infrastructure.

Benefits of IaC include increased speed and agility of infrastructure deployments, improved consistency and reliability of infrastructure, reduced manual errors and overhead, and enhanced collaboration between development and operations teams.

Infrastructure as Code (IaC) is the need of the hour for several reasons:

- **Rapidly changing infrastructure:** In today's fast-paced digital environment, infrastructure needs to be agile and responsive to changing business needs. IaC allows for quick and efficient provisioning, updating, and scaling of infrastructure, which is essential for meeting business demands.
- **Automation:** IaC automates the infrastructure deployment and management process, reducing the likelihood of human error and improving overall efficiency. It also allows for faster and more reliable deployments, which can result in increased productivity and reduced costs.
- **Consistency:** IaC ensures that infrastructure is consistently configured across different environments, reducing the risk of configuration drift and improving overall system reliability.
- **Collaboration:** IaC encourages collaboration between development and operations teams by enabling them to work together on infrastructure code. This

can lead to better communication, faster issue resolution, and improved alignment between different teams.
- **Compliance:** IaC can help organizations maintain compliance with industry standards and regulations by providing a consistent and auditable approach to infrastructure management.

AWS Cloud Development Kit (CDK) is a software development framework that enables developers to define cloud infrastructure using familiar programming languages such as TypeScript, Python, Java, and .NET. It provides a higher-level object-oriented abstraction on top of AWS CloudFormation, which allows for more efficient and expressive code.

Infrastructure as Code (IaC) using AWS Cloud Development Kit (CDK) is a way of defining and deploying AWS resources using code. AWS CDK enables developers to define cloud infrastructure in familiar programming languages such as TypeScript, Python, Java, and .NET, using object-oriented constructs and high-level abstractions. Some benefits of using AWS CDK for IaC include Familiar programming languages, High-level object-oriented abstraction, Consistency and maintainability, and AWS CloudFormation compatibility

Aim:
The objective of this project is to build an ETL Pipeline on Amazon EMR through AWS CDK. The pipeline will involve carrying out data analysis and transformation using Apache Hive on EMR. Additionally, we will create an interactive dashboard on Power BI for the visualization of the results.

Tech Stack
➔
Language: Python
➔
Services: AWS S3, AWS Cloud9, AWS CDK, AWS EMR, Power BI, Apache Hive

Key Takeaways
- Understanding the Sales dataset
- Understanding the AWS CDK
- Installation of AWS CDK and its various commands
- Advantages of Serverless technologies
- Creating an AWS Cloud9 environment
- Understanding the difference between AWS and Azure
- Difference between UUID and NanoID
- Difference between Hive and Spark
- Understanding the S3 Bucket deployment stack

- Understanding the Security stack
- Understanding the Hive script to create tables
- Understanding the Hive script to transform data
- Understanding the EMR cluster stack
- Deployment of AWS CDK stacks
- Debugging of AWS CDK stacks
- Setting up a Virtual Machine in Azure
- Connecting to Hive tables using Power BI
- Creating visualizations using Power BI

Architecture Diagram: