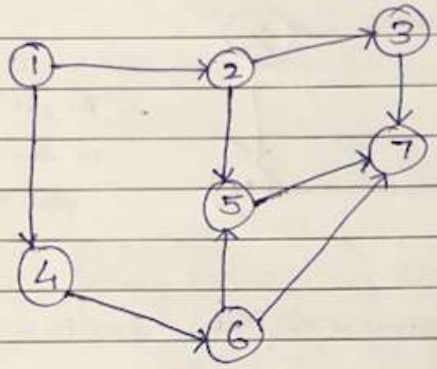
DATA SCIENEE - UNIT 2

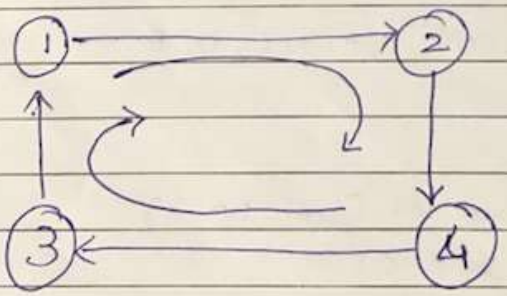
Q. Directed Acylic Graph Scheduling.

ANS DAG known as Directed Acydic Graph, is an fundamental concept in graph theory. DAG's are used to show how things are related or depend on each other in a clear and organized way
A Directly Graph is an Directed Graph that does not contain any cycles.



DAG consists of two main feature → Directed Edges
→ Acyclic

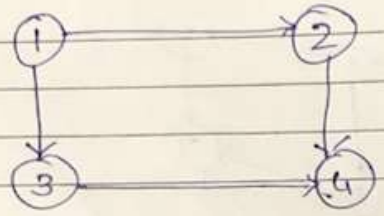
•Directed Edges
In DAG, each edges has a direction, meaning it goes from one vertex to another. This direction signifies a one-way relationship or dependency between nodes.



→ Directed Cyclic Graph
(As their following a cycle)

Acyclic

The term "acyclic" indicates that there are no cycles or closed loops within the graph. In other words, you cannot transverse a sequence of directed edges and return to same node, following the edge directions. Formation of cycles is prohibited in DAG



- Acyclic as there is no proper route

- SCHEDULING

Scheduling involves arranging the execution of tasks so that all dependencies are respected. Tasks without dependencies can be executed in parallel, while dependent tasks must follow a sequence.

Ex: Imagine a project with tasks A, B, C, D, E. Dependencies are as follow: B depends on A

C DP on A

D DP on B & C

E DP on D

SORTING TOPOLOGY : A, B, C, D, E

OUTPUT: A

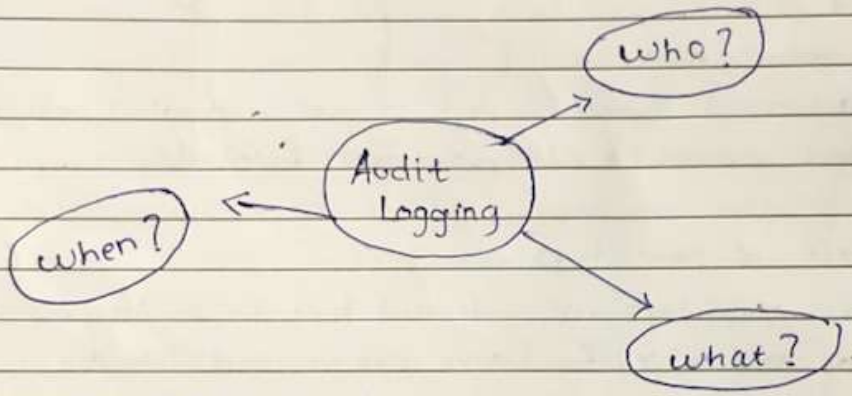
B & C in parallel (both DP A)

D (DP on B & C)

E (Dep. on D)

Q. Explain Different ways of Implementing Built-in Logging in the Audit phase.

ANS Audit logging is the process of documenting activity within the software systems used across your organization. Audit logs record the occurrence of an event, the time at which it occured, the responsible user or a service, and the impacted entity. All of the devices in your network, your cloud services and your application that may be used for auditing purposes.



Approaches in logging:

(i) Identify What to log:
Figure out what events and action you need to track: This includes things like loggin attempts, data change and errors.

(ii) Choose a login Tool
Pick a tool or framework for logging, like Log4j or Frastic Search

(iii) Set Logging Levels
Decide the importance of different events. Use levels like INFO, WARNING and ERROR to classify them.

(iv) Add Logging to your code
Insert logging statements in your application to capture the key events. Like log when a user logs in or when a transaction fails.

(v) Centralize Logs.
Collect logs from all sources into one place for easier analysis. Tools like Splunk or ELK stack help with this.

(vi) Automation Alerts
Setup automated system to look at logs and alert you to any suspicious activities, like multiple failed logins.

(vii) Protect & Store Logs
Make sure your logs are safe and kept for as long as needed, so they are available for future audits and checks.

Ex:
(i) Track who accesses cust data and when
(ii) Use ElasticSearch
(iii) Mark data access as INFO and unauthorized access attempt as ERROR
(iv) Log every code access event in your app
(v) Send all logs to ElasticSearch
(vi) Get alerts on unauthorized access attempts
(vii) Encrypt logs and keep them for a specified period.

This way banks can know who access what data & ensure security and integrity.

Q. Explain Function of Audit, Balance & Control?

Ans AUDIT LAYER:

The audit layer is responsible for tracking and recording data changes ensuring compliance with regulations, and maintaining security.

Functions of AUDIT LAYERS are:

(i) Tracking Data Changes: Keep a detailed log of every modification made to data, including what was changed, who made the change, and when it was made.

(ii) Compliance Monitoring: Ensuring that data handling complies with relevant laws, regulations and internal policies by maintaining comprehensive logs.

(iii) Security: Detect and prevent unauthorized access or alteration to data by monitoring for anomalies and unauthorized access.

BALANCE LAYER:

The Balance layer ensures the accuracy, consistency and integrity of Data

Functions of BALANCE LAYERS ARE:

(i) Data Validation: Checks the data is accurate, consistent and reliable before it is used. This includes verifying data against known standard of rules

(ii) Reconciliation: Compares data from different sources to ensure they match and highlights any discrepancies. This is crucial for maintaining data consistency across systems.

(iii) Integrity Checks: Verifies that all data transactions are complete and correctly processed. This ensures that data remains intact and accurate throughout its lifecycle.

Controll Layer:
The Control Layer manages access to data, automates business process and handles errors in data processing.
Functions of Controll Layer
① Access Management : Controlls who can access or modify data, ensuring that only authorized user's can perform certain actions. This is achieved through role-based access controll and authentic mechanism
② Error Handling
Identifies, reports; and resolve errors in data processing. This ensures that any issues are promtly addressed maintaining the smooth operation of data system.

Q. STATE AND EXPLAIN STEPS TO AVOID DATA SWAMPS.
Ans. A data swamp is a poorly manage data, where stored data became disorganized, inaccessible, or unusable. This happens when the data is ingested without ~~papers~~ proper governance, metadata or quality Checks, leading to swamp of "unstructured", "unclean" info.

STEPS TO AVOID SWAMPS:

① Establish Clear Data Governance
Implement policies and procedures that define how data is ~~managed~~ managed and who is responsible for it. Define ownership, quality standards and access controll's. Regularly Examine your data to ensure it meets government criteria.

Ex: A Health care organization set up a data governance teams to oversee data quality, access permission and comply with regulation

(2) IMPLEMENT DATA QUALITY MEASURES

Ensure data entering the data lake is accurate, consistent and complete. Use tools and processes to clean and validate data before it's ingested. Continously monitor data quality measure and address issues promptly.

Ex: A retail company uses automated scripts to clean and validate customer data before it's stored in data lake, ensuring consistency and accuracy.

(3) MONITOR AND OPTIMIZE DATA USAGE

Continously monitor how data is used and optimize storage and retrieval process. Track data usage patterns and adjust storage strategies to improve performance and cost efficiency.

Ex: An IoT company monitors data access patterns and adjusts its storage strategies, archiving less frequently accessed data to reduce costs and improve performance for active datasets.

By following these steps, Organization can ensure their data lake remains a valuable and maintained well resource with any faulty docs / data / files.

Q. Retrieve Superstep

Ans. The retrieve Superstep is an practical method for importing completely into processing ecosystem a data lake consisting of various external data sources. The Retrieve superstep is the first contact between DS and source system. It refers to a phase in which each vertex retrieves messages sent to it during previous superstep.

Key Aspects:

(i) Message Retrieval

Each vertex collects message that were sent to it during the previous superstep. These messages can contain data or instruction from other devices.

(ii) Computations

Based on retrieved message, the vertex performs computations. This might involve updating its state, performing calculation or preparing messages to sent in their next superstep.

(iii) Message Preparation

After computation, the vertex ~~appears~~ prepares messages to be sent to other vertices in network. These message will be processed in next superstep

(iv) Synchronization

All vertices complete their retrieval and computation tasks before moving on to next step, ensuring that entire system stays in sync.

Ex: Imagine a network Computer (vertices) and connected cables (edges). Each Computer wants to find shortest path to a speific server (the source)

S1 : Initialization
The server starts by sending a message to all directly connected Computers setting their initial distance to 1.

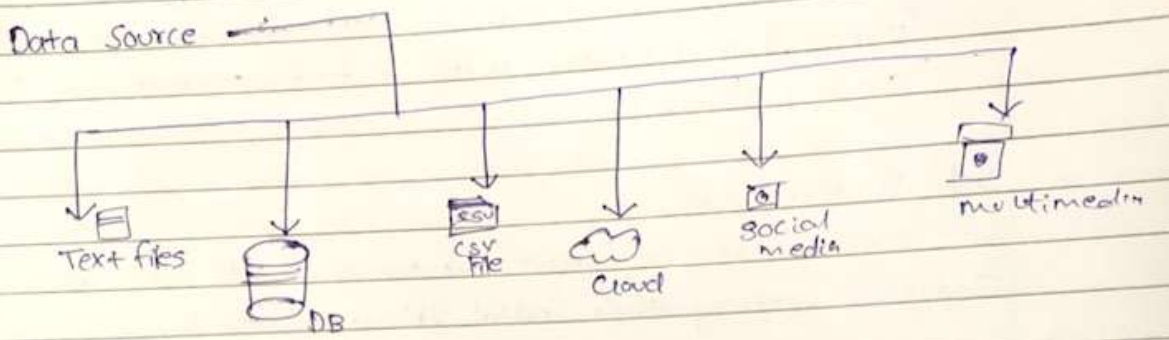
S2 : RS 1
Each Computer Retrieves the message from server and updates its distance to 1. These Computers then send their updated distance to their own neighbours.

S3: RS 2
The next layer of computer retrieves the distance from their neighbours and updates their distance to 2. and info forwards

S4 : Continue SS
This process repeats, with each computer retrieving messages, updating their distances, and forwarding, until all computa knows the shortest path.

Q. List and Explain Diff Data Stores used in Data Science.

Ans A Data store is a repository for storing, managing and retrieving data. It can take or be in various format and structure of data, from structure tabular data to unstructred text & multimedia



① Textfiles
Simple files that containing plain text data. Often used for logs, configuration files and small datasets.
Ex: System log files that record events and errors in a plain text format

② Databases
Structured query language storage system using SQL or NoSQL to manage and query data. Ideal for organized, transact data.
Ex: A customer relationship management (CRM) system storing customer info and transactions

③ CSV Files
Comma-separated files that store tabular data in plain text. Widely used for data exchange and analysis
Ex: A datasets of sales records saved as a CSV file, easily readable by tools like Excel and Pandas

(iv) Cloud Data Warehouse

Scalable storage solutions hosted in cloud, optimize for large-scale data processing and analytics

Ex: Amazon Redshift Storing large volumes of sales and cust data for realtime analytics and reporting

(v) Social Media / API's

Using twitter's API to collect and analyze tweets mentioning a brand for sentiment analysis.

(vi) multimedia

Storage of multimedia files such as audio, vide and images. Requires special data sources that can handle large filess sizes and diverse format

Ex: YouTube.

1. Explain the Foll shipping Terms

Ans.

(i) Seller.

The party selling goods and are responsible for providing products to buyer. The often handle packaging, documentation and initial transportation arrangements.

Ex: A manufacturer selling electronics to a retailer

(ii) Carriey

The company or individual responsible for transporting goods from seller to buyer. This can include trucking companies, shipping line, airlines or rail operation.

Ex: FedEx or Bluedart handling the transportation of goods.

(iii) Port

A harbour where ships load and unload goods. Port plays a critical role in international trade by facilitating the movement of cargo between land & sea.

Ex: Port of Mumbai, one of the busiest ports in world

(iv) Ship

A large vessel used to transport goods across bodies of water. Ships are essential for carrying bulk cargo over long distances.

Ex: Container ships Transporting goods from Asia to US.

(v) Terminal

A facility at port where cargo is transferred btwn different modes of transport, such as truck, & train. Terminal also handle storage and logistic service.

(vi) Named Place

A specific location agreed upon by seller an buyer for delivery of goods. The term is used in shipping agreement to clarify where the responsibility of seller ends.

Ex: Agreement → seller will ship parcel till xyz location.

(vii) Buyer.

The party purchasing goods from seller. Often Responsible for TC & D/T

Ex: A retailer Purchasing electronics from manufacturer to sale in Multi Stores.