Past paper

Q1:explain steps of ETL lifecycle with example?

**Steps of the ETL Lifecycle Explained Individually**

1. **Cycle Initiation**
   * The process begins, triggered by a schedule or specific event.
   * **Purpose:** To set the process in motion.
2. **Build Reference Data**
   * Prepare or update supporting data (like lookup tables or master lists) that provide context for the extracted data.
   * **Purpose:** To standardize and organize the data.
3. **Extract (from Sources)**
   * Collect raw data from different sources (databases, files, APIs, etc.).
   * **Purpose:** To gather all relevant information.
4. **Validate**
   * Check if the extracted data is correct and complete, and identify any missing or incorrect entries.
   * **Purpose:** To ensure the accuracy of data.
5. **Transform**
   * Clean, filter, and modify the data to make it useful (e.g., removing duplicates, converting formats, applying business rules).
   * **Purpose:** To prepare data for analysis.
6. **Stage**
   * Temporarily store the processed data in an intermediate location.
   * **Purpose:** To hold data before loading it into the final destination.
7. **Audit Reports**
   * Create a report to confirm data integrity and process success, or diagnose issues.
   * **Purpose:** To monitor the ETL process.
8. **Publish (to Target Tables)**
   * Load the processed and validated data into the final destination (data warehouse, database, etc.).
   * **Purpose:** To make the data available for use.
9. **Archive**
   * Backup the original and processed data for record-keeping and compliance.
   * **Purpose:** To save data for future reference.
10. **Clean Up**
    * Remove temporary files or intermediate data used during processing.
    * **Purpose:** To optimize system performance and save storage.

**Simple Example of the ETL Lifecycle**

**Scenario:** A school wants to prepare a report of student attendance from multiple sources.

1. **Cycle Initiation**
   * The school decides to run the ETL process at the end of every week.
2. **Build Reference Data**
   * A list of all students, their classes, and subjects is prepared.
3. **Extract (from Sources)**
   * Attendance data is collected from class registers, online attendance systems, and biometric devices.
4. **Validate**
   * Check if each student has valid attendance records and if any data is missing or duplicated.
5. **Transform**
   * Clean the data (remove duplicates), calculate total attendance per student, and mark those with low attendance.
6. **Stage**
   * Temporarily store the processed data in a separate database.
7. **Audit Reports**
   * Create a report showing errors (like missing entries) and confirm the number of records processed.
8. **Publish**
   * Load the cleaned and processed attendance data into the school’s central database.
9. **Archive**
   * Save the original attendance records and the final processed report for future reference.
10. **Clean Up**
    * Delete temporary files created during the ETL process.

**Result:**  
At the end of the week, the school has a clean attendance report, highlighting students with low attendance, ready for teachers and administration to review.

Q2: list 6 data source selection criteria in data science

Here are **6 key data source selection criteria** in data science:

1. **Relevance**
   * The data should align with the objectives of the project or the questions being addressed.  
     *Example:* If analyzing customer preferences, choose data sources containing customer reviews, surveys, or purchase history.
2. **Accuracy**
   * The data should be reliable, precise, and free of errors or inconsistencies.  
     *Example:* Financial analysis should rely on audited financial statements or trusted financial databases.
3. **Completeness**
   * The data should contain all necessary fields and records required for analysis without significant gaps.  
     *Example:* A dataset for patient analysis should include age, gender, medical history, and current diagnosis.
4. **Timeliness**
   * The data should be up-to-date and reflect current trends or situations.  
     *Example:* For stock market analysis, use real-time or latest financial data instead of outdated records.
5. **Accessibility**
   * The data should be easy to obtain and available in a usable format, with proper permissions if needed.  
     *Example:* Public datasets on climate change or open government records may be readily accessible.
6. **Cost and Feasibility**
   * The cost of acquiring the data should be reasonable and within the project budget, and the data should be feasible to process.  
     *Example:* Purchasing expensive market research reports might be infeasible for small businesses, so free alternatives may be preferred.

These criteria ensure the selection of data sources that are suitable and effective for the specific goals of a data science project.