**2. Capture all the new data terms from this class and create a data terms catalog.**

**Ans:**

1. Replayability: The capacity to replay or reprocess previously created or recorded events or data. Replayability in the context of event-driven architecture is the ability to replay events for different uses, such testing, debugging, or reprocessing.

2. Versioning: The process of tracking changes over time in data, schemas, or software components by giving them distinct identities or versions. Versioning in the context of event-driven architecture can refer to keeping event schemas backwards or forwards compatible in order to guarantee interoperability between various iterations of event producers and consumers.

3. Throughput: The speed at which information may be sent or processed inside a system this speed is commonly expressed in terms of data units per second, or messages per second. The term throughput in event-driven architecture describes the system's ability to manage and process a specific number of events in a predetermined amount of time.

4. Latency: The time it takes for data to flow from one location to another or for an action to complete within a system. Latency in event-driven architecture refers to the time it takes for an event to occur and be processed by event consumers. Low latency is frequently preferred in real-time or near-real-time systems.

5. Volume: The amount or quantity of data created, processed, or saved within a system or over time. In event-driven architecture, volume refers to the size or magnitude of the events created and consumed by the system. Handling large numbers of events efficiently is critical for scalability and performance.

6. Velocity: The rate at which data is created, processed or exchanged in a system is often measured by the rate of change or the frequency with which data is updated. In an event-driven architecture, velocity refers to the rate at which events are generated, emitted, and consumed in response to changes or interactions in the system or its environment.

7. Event-Driven Architecture (EDA): A program architecture paradigm in which data flow and behavior are determined by events such as user actions, system events, or messages from other components. In EDA, components communicate asynchronously through events, allowing for loose coupling, scalability, and responsiveness.

8. Veracity: The accuracy, reliability and reliability of the data in the system or dataset. In an event-oriented architecture, veracity refers to the assurance that event information is correct, consistent, and free of errors or inconsistencies that could lead to incorrect decisions or results.

9. Variety: Diversity or heterogeneity of data types, formats, structures, or sources in a system or dataset. In an event-oriented architecture, variety refers to the existence of different types of events, each with its own schema, payload, or metadata. Addressing diversity involves matching and integrating different event formats and structures to ensure interoperability and flexibility.

**3. Read understand and write about 5 V’S of data with example scenario for each.**

**Ans:**

1. Volume: Volume refers to the sheer amount or size of data generated, stored, processed, and analyzed within a given system or environment.

Example: Social media platforms like Facebook generate an enormous volume of data every day, with over 2.9 billion monthly active users collectively generating billions of content pieces (posts, photos, videos) and engaging in billions of interactions (likes, comments, shares), resulting in ingestion of over 600 terabytes of data per day from various sources.

1. Velocity: Velocity refers to the speed or rate at which data is generated, processed, and moved within a system. It often emphasizes the real-time or near-real-time nature of data processing.

Example: In the world of finance and stock trading, velocity is critical as data from global markets needs to be ingested, processed, and analyzed at extremely high speeds to enable automated trading systems to make split-second decisions based on real-time market data feeds, news, and events happening around the world.

1. Variety: Variety refers to the diversity or heterogeneity of data types, formats, structures, and sources within a dataset or system. It encompasses structured, semi-structured, and unstructured data.

Example: Social media platforms like Facebook deal with a huge variety of data types and formats, including structured data from user profiles, semi-structured data like JSON from app logs, and unstructured data such as text from posts, images, videos, and audio recordings, all of which need to be processed and analyzed together to derive insights.

1. Veracity: Veracity refers to the accuracy, reliability, consistency, and trustworthiness of data. It emphasizes the quality and reliability of data, considering factors such as completeness, correctness, and relevance.

Example: For a healthcare organization analyzing patient data to develop predictive models for disease risk, ensuring veracity by having complete, accurate and reliable data from trusted sources like electronic medical records and clinical studies is crucial, as any errors or inconsistencies in the data could lead to flawed models and potentially harmful treatment recommendations.

1. Value: Value refers to the usefulness, relevance, and insights derived from analyzing and interpreting data. It focuses on the practical benefits and outcomes that data-driven insights can provide to organizations and individuals.

Example: Online retailers like Amazon leverage immense volumes of customer data, including browsing history, purchases, reviews, and preferences, to provide highly personalized product recommendations, targeted marketing, and optimized pricing strategies, directly translating the insights from this data into increased sales and customer loyalty, which is the core value proposition of their business model.

**4. What is Data Ingestion vs Data Integration? Explain them with a real-world use-case.**

**Ans:**

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Data Ingestion** | **Data Integration** |
| Definition | The process of collecting and importing raw data from various sources into a centralized data storage system or data lake. | The process of combining data from multiple heterogeneous sources into a unified, consistent, and accurate view |
| Purpose | To consolidate data from disparate sources into a single repository for further processing, analysis or storage. | To provide a comprehensive, high-quality dataset ready for analysis, reporting, and decision-making by reconciling differences across sources. |
| Process | Extract data from sources, optionally transform it into a structured format, and load it into the target storage system. | Extract data from sources, apply complex transformations (cleansing, deduplication, normalization, etc.), and load the integrated data into a data warehouse or analytical system. |
| Data Quality | Does not inherently ensure data quality or consistency. The focus is on acquiring and storing data efficiently. | Focuses on improving data quality by implementing various data quality checks, merging rules, and governance processes. |
| Complexity | Relatively simpler compared to data integration pipelines. | More complex, involving processes like ETL, metadata management, data governance, and other intricate transformations. |
| Example | Amazon ingests raw data from multiple sources, such as website clickstreams, product catalogs, customer reviews, and order management systems, into a centralized data lake or storage system. The ingestion process focuses on efficiently collecting and storing this raw data for further processing. | Amazon then integrates the ingested data from different sources to create a unified view of customer profiles, product information, sales data, and more. This involves complex transformations like cleansing, deduplicating, and normalizing data formats, as well as implementing data quality checks and governance processes. The integrated data is then loaded into a data warehouse or analytical system for reporting, analysis, and decision-making purposes, such as personalized recommendations, inventory management, and targeted marketing campaigns. |