Week 3 - Data Engineering Lifecycle

Data Platforms, Data Stores, and Security

Summary and Highlights

The architecture of a data platform can be seen as a set of layers, or functional components, each one performing a set of specific tasks. These layers include:

- Data Ingestion or Data Collection Layer, responsible for bringing data from source systems into the data platform.
- Data Storage and Integration Layer, responsible for storing and merging extracted data.
- Data Processing Layer, responsible for validating, transforming, and applying business rules to data.
- Analysis and User Interface Layer, responsible for delivering processed data to data consumers.
- Data Pipeline Layer, responsible for implementing and maintaining a continuously flowing data pipeline.

A well-designed data repository is essential for building a system that is scalable and capable of performing during high workloads.

The choice or design of a data store is influenced by the type and volume of data that needs to be stored, the intended use of data, and storage considerations. The privacy, security, and governance needs of your organization also influence this choice.

The CIA, or Confidentiality, Integrity, and Availability triad are three key components of an effective strategy for information security. The CIA triad is applicable to all facets of security, be it infrastructure, network, application, or data security.

Practice Quiz

Question 1

Which one of these steps is an intrinsic part of the “Data Storage and Integration Layer” of a data platform?

- Read data in batch or streaming modes from storage and apply transformations
- **Transform and merge extracted data, either logically or physically**
- Transfer data from data sources to the data platform in streaming, batch, or both modes
- Deliver processed data to data consumers

The Storage and Integration layer in a data platform stores, transforms, and merges extracted data to make it available for data processing.

Question 2

Systems that are used for capturing high-volume transactional data need to be designed for faster response times to complex queries.

- True
- **False**

Systems that are used for capturing high-volume transactional data need to be designed for high-speed read, write, and update operations.

Question 3

What is the role of “Intrusion Detection” and “Intrusion Prevention” in the area of network security?

- Ensure endpoint security by allowing only authorized devices to connect to the network
- **Inspect incoming network traffic for intrusion attempts and vulnerabilities**
- Create silos, or virtual local area networks, within a network so that you can segregate your assets

- Ensure attackers cannot tap into data while it is in transit

Intrusion Detection and Intrusion Prevention systems inspect network vulnerabilities and intrusion attempts and prevent them from happening.

Graded Quiz

Question 1

Which one of these steps is an intrinsic part of the “Data Processing Layer” of a data platform?

- Deliver processed data to data consumers
- Transfer data from data sources to the data platform in streaming, batch, or both modes
- Transform and merge extracted data, either logically or physically
- **Read data in batch or streaming modes from storage and apply transformations**

Question 2

Systems that are used for capturing high-volume transactional data need to be designed for high-speed read, write, and update operations.

- **True**
- False

High-speed read, write, and update operations are essential for systems that need to capture large volumes of transactional data.

Question 3

What is the role of “Network Access Control” systems in the area of network security?

- **To ensure endpoint security by allowing only authorized devices to connect to the network**
- To ensure attackers cannot tap into data while it is in transit
- To create silos, or virtual local area networks, within a network so that you can segregate your assets
- To inspect incoming network traffic for intrusion attempts and vulnerabilities

This is achieved with the help of Network Access Control systems.

Question 4

_____ ensures that users access information based on their roles and the privileges assigned to their roles.

- Authentication
- **Authorization**
- Firewalls
- Security Monitoring

One of the primary controls for data security is to enable access to data through a system of Authorization. It allows access to information based on a user’s role and role-based privileges.

Question 5

Security Monitoring and Intelligence systems:

- Create virtual local area networks within a network so that you can segregate your assets
- **Create an audit history for triage and compliance purposes**
- Ensure users access information based on their role and privileges
- Ensure only authorized devices can connect to a network

Security Monitoring and Intelligence systems create an audit trail and provide reports and alerts that help enterprises react to security violations in time.

Data Collection and Data Wrangling

Summary and Highlights

Depending on where the data must be sourced from, there are a number of methods and tools available for gathering data. These include query languages for extracting data from databases, APIs, Web Scraping, Data Streams, RSS Feeds, and Data Exchanges.

Once the data you need has been gathered and imported, your next step is to make it analytics-ready. This is where the process of Data Wrangling, or Data Munging, comes in.

Data Wrangling involves a whole range of transformations and cleansing activities performed on the data. Transformation of raw data includes the tasks you undertake to:

- Structurally manipulate and combine data using Joins and Unions.
- Normalize data, that is, clean the database of unused and redundant data.
- Denormalize data, that is, combine data from multiple tables into a single table so that it can be queried faster.

Cleansing activities include:

- Profiling data to uncover anomalies and quality issues.
- Visualizing data using statistical methods in order to spot outliers.
- Fixing issues such as missing values, duplicate data, irrelevant data, inconsistent formats, syntax errors, and outliers.

A variety of software and tools are available for the data wrangling process. Some of the popularly used ones include Excel Power Query, Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python, and R, each with their own set of features, strengths, limitations, and applications.

Practice Quiz

Question 1

How is data gathered using Application Programming Interfaces, or APIs?

- APIs are used for aggregating constant streams of data flowing from instruments, IoT devices and applications, and GPS data from cars
- APIs are used for downloading specific data from web pages based on defined parameters
- APIs are used for capturing updated data from online forums and news sites where data is refreshed on an ongoing basis
- **APIs are invoked from applications to access databases, web services, data marketplaces and other such data endpoints for gathering data**

Question 2

What is one of the common structural transformations used for combining data from one or more tables?

- **Joins**
- Cleaning
- Denormalization
- Normalization

Question 3

What tool allows you to discover, cleanse, and transform data with built-in operations?

- **Watson Studio Refinery**
- OpenRefine
- Trifacta Wrangler
- Google DataPrep

Watson Studio Refinery has built-in features that allow you to discover, cleanse, and transform data.

Graded Quiz

Question 1

Web scraping is used to extract what type of data?

- Text, videos, and data from relational databases
- **Text, videos, and images**
- Images, videos, and data from NoSQL databases
- Data from news sites and NoSQL databases

Question 2

_____ focuses on cleaning the database of unused data and reducing redundancy and inconsistency.

- Denormalization
- Data Visualization
- Data Profiling
- **Normalization**

Normalization cleanses the database of unused data and inconsistencies in data that is coming from multiple sources.

Question 3

OpenRefine is an open-source tool that allows you to:

- **Transform data into a variety of formats such as TSV, CSV, XLS, XML, and JSON**
- Automatically detect schemas, data types, and anomalies
- Enforces applicable data governance policies automatically
- Use add-ins such as Microsoft Power Query to identify issues and clean data

Question 4

When you're combining rows of data from multiple source tables into a single table, what kind of data transformation are you performing?

- Denormalization
- Joins
- **Unions**
- Normalization

Unions are a common structural transformation used for combining rows of data from multiple source tables.

Question 5

When you detect a value in your data set that is vastly different from other observations in the same data set, what would you report that as?

- Missing value
- Irrelevant data
- **Outlier**
- Syntax error

Outliers are values in your data set that may be vastly different from other values in the same data field.

Querying Data, Performance Tuning, and Troubleshooting

Summary and Highlights

- In order for raw data to become analytics-ready, a number of transformation and cleansing tasks need to be performed on raw data. And that requires you to understand your dataset from multiple perspectives. One of the ways in which you can explore your dataset is to query it.
- Basic querying techniques can help you explore your data, such as, counting and aggregating a dataset, identifying extreme values, slicing data, sorting data, filtering patterns, and grouping data.

- In a data engineering lifecycle, the performance of data pipelines, platforms, databases, applications, tools, queries, and scheduled jobs, need to be constantly monitored for performance and availability.
- The performance of a data pipeline can get impacted if the workload increases significantly, or there are application failures, or a scheduled job does not work as expected, or some of the tools in the pipeline run into compatibility issues.
- Databases are susceptible to outages, capacity overutilization, application slowdown, and conflicting activities and queries being executed simultaneously.
- Monitoring and alerting systems collect quantitative data in real time to give visibility into the performance of data pipelines, platforms, databases, applications, tools, queries, scheduled jobs, and more.
- Time-based and condition-based maintenance schedules generate data that helps identify systems and procedures responsible for faults and low availability.

Practice Quiz

Question 1

In the video, we used a query function to see how spread out the values in the “Sale Amount” field are. What function did we use?

- Average
- Count
- Maximum Value
- **Standard Deviation**

Question 2

_____ helps you assess if the size of a workload is slowing down the system.

- Monitoring the performance of queries
- Job-level Runtime Monitoring
- **Monitoring the amount of data being processed through a data pipeline**
- Database Monitoring

Governance and Compliance

Summary and Highlights

Data Governance is a collection of principles, practices, and processes that help maintain the security, privacy, and integrity of data through its lifecycle.

Personal Information and Sensitive Personal Information, that is, data that can be traced back to an individual or can be used to identify or cause harm to an individual, needs to be protected through governance regulations.

General Data Protection Regulation, or GDPR, is one such regulation that protects the personal data and privacy of EU citizens for transactions that occur within EU member states.
Regulations, such as HIPAA (Health Insurance Portability and Accountability Act) for Healthcare, PCI DSS (Payment Card Industry Data Security Standard) for retail, and SOX (Sarbanes Oxley) for financial data are some of the industry-specific regulations.

Compliance covers the processes and procedures through which an organization adheres to regulations and conducts its operations in a legal and ethical manner.

Compliance requires organizations to maintain an auditable trail of personal data through its lifecycle, which includes acquisition, processing, storage, sharing, retention, and disposal of data.

Tools and technologies play a critical role in the implementation of a governance framework, offering features such as:

- Authentication and Access Control.
- Encryption and Data Masking.
- Hosting options that comply with requirements and restrictions for international data transfers.
- Monitoring and Alerting functionalities.

- Data erasure tools that ensure deleted data cannot be retrieved.

Practice Quiz

At what stage of the data lifecycle would you establish which third-party vendors in your supply chain will have access to the data you are collecting?

- **Data Sharing**
- Data Acquisition
- Data Processing
- Data Storage

It is in the Data Sharing phase of the data lifecycle that you establish which third-party vendors will have access to your data, and how they will be held accountable to the same regulations you are liable for.

Graded Quiz

Question 1

In which phase of the data lifecycle do you establish the data you need, the amount of data you need, and how you intend to use the data you are collecting.

- Data Processing
- **Data Acquisition**
- Data Sharing
- Data Retention

In the Data Acquisition phase, you establish the data you need to collect, the amount of data you need, and its intended use.

Question 2

The process of _____ abstracts the presentation layer without changing the data in the database physically.

- Encryption
- Data Profiling
- **Anonymization**
- Pseudonymization

Using Anonymization, the presentation layer is abstracted without changing the data in the database itself.