Unit	Topics
I	**Data Science Technology Stack:** Rapid Information Factory Ecosystem, Data Science Storage Tools, Data Lake, Data Vault, Data Warehouse Bus Matrix, Data Science Processing Tools, Spark, Mesos, Akka, Cassandra, Kafka, Elastic Search, R, Scala, Python. **Layered Framework:** Definition of Data Science Framework, CrossIndustry Standard Process for Data Mining (CRISP-DM), Homogeneous Ontology for Recursive Uniform Schema, The Top Layers of a Layered Framework, Layered Framework for High-Level Data Science and Engineering. **Business Layer:** Business Layer, Engineering a Practical Business Layer. **Utility Layer**: Basic Utility Design, Engineering a Practical Utility Layer.
II	**Three Management Layers:** Operational Management Layer, Processing-Stream Definition and Management, Audit, Balance, and Control Layer, Balance, Control, Yoke Solution, Cause-and-Effect, Analysis System, Functional Layer, Data Science Process. **Retrieve Superstep:** Data Lakes, Data Swamps, Training the Trainer Model, Understanding the Business Dynamics of the Data Lake, Actionable Business Knowledge from Data Lakes, Engineering a Practical Retrieve Superstep, Connecting to Other Data Sources.

6640_Shivam

Unit 1 : Ch 1 Data Science Technology Stack

Q. Fundamentals of Data Science

Ans.

1. **Data science**
    1. In 1960, Peter Naur started using the term data science as a substitute for computer science.
    2. Data science is an interdisciplinary science that incorporates practices and methods with actionable knowledge and insights from data in heterogeneous schemas (structured, semi-structured, or unstructured).
    3. It amalgamates the scientific fields of data exploration with thought-provoking research fields such as data engineering, information science, computer science, statistics, artificial intelligence, machine learning, data mining, and predictive analytics.

2.. Data Analytics

1. Data analytics is the science of fact-finding analysis of raw data, with the goal of drawing conclusions from the data lake.

3. Machine learning

1. Machine learning is the capability of systems to learn without explicit software development.
2. It evolved from the study of pattern recognition and computational learning theory.

4. Data mining

1. Data mining is processing data to isolate patterns and establish relationships between data entities within the data lake.

5. Statistics

1. Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

6. Algorithm

1. An algorithm is a self-contained step-by-step set of processes to achieve a specific outcome.

7. Data visualization

1. It consists of the creation and study of the visual representation of business insights.

8. Storyline

1. Data storytelling is the process of translating data analyses into layperson’s terms, in order to influence a business decision or action.

6640_Shivam

Q. Explain the following terms in brief:

1. Data Lake

1. A data lake is a storage repository for a massive amount of raw data.
2. It stores data in native format, in anticipation of future requirements.
3. A data lake uses a less restricted schema-on-read-based architecture to store data.
4. Each data element in the data lake is assigned a distinctive identifier and tagged with a set of comprehensive metadata tags.
5. A data lake is typically deployed using distributed data object storage, to enable the schema-on-read structure.
6. This means that business analytics and data mining tools access the data without a complex schema.
7. Using a schema-on-read methodology enables you to load your data as is and start to get value from it instantaneously.

2. Data Vault

1. Data vault modeling, designed by Dan Linstedt.
2. It is a database modeling method that is intentionally structured to be in control of long-term historical storage of data from multiple operational systems.
3. The data vaulting processes transform the schema-on-read data lake into a schema-on-write data vault.
4. The data vault is designed into the schema-on-read query request and then executed against the data lake.
5. The structure is built from three basic data structures: hubs, inks, and satellites.

3. Hubs

1. Hubs contain a list of unique business keys with low propensity to change.
2. They contain a surrogate key for each hub item and metadata classification of the origin of the business key.
3. The hub is the core backbone of your data vault.

4. Links

1. Associations or transactions between business keys are modeled using link tables.
2. These tables are essentially many-to-many join tables, with specific additional metadata.
3. The link is a singular relationship between hubs to ensure the business relationships are accurately recorded to complete the data model for the real-life business.

5. Satellites

1. Hubs and links form the structure of the model but store no chronological characteristics or descriptive characteristics of the data.
2. These characteristics are stored in appropriated tables identified as satellites.
3. Satellites are the structures that store comprehensive levels of the information on business characteristics and are normally the largest volume of the complete data vault data structure

6640_Shivam

Q. Schema-on-Write and Schema-on-Read (Any one ache se karo)

There are two basic methodologies that are supported by the data processing tools.

1. **Schema-on-Write Ecosystems**
   A traditional relational database management system (RDBMS) requires a schema before you can load the data.
2. **Benefits include the following:**
   a. **In traditional data ecosystems, tools assume schemas and can only work once the schema is described, so there is only one view on the data.**
   b. **The approach is extremely valuable in articulating relationships between data points, so there are already relationships configured.**
   c. **It is an efficient way to store "dense" data.**
   d. **All the data is in the same data store.**
3. **the downsides of this approach are that**
   a. **Its schemas are typically purpose-built, which makes them hard to change and maintain.**
   b. **It generally loses the raw/atomic data as a source for future analysis.**
   c. **It requires considerable modeling/implementation effort before being able to work with the data.**
   d. **If a specific type of data can't be stored in the schema, you can't effectively process it from the schema.**
4. **At present, schema-on-write is a widely adopted methodology to store data.**

1. **Schema-on-Read Ecosystems**
   This alternative data storage methodology does not require a schema before you can load the data.
   Fundamentally, you store the data with minimum structure. The essential schema is applied during the query phase.
2. **Benefits include the following:**
   a. **It provides flexibility to store unstructured, semi-structured, and disorganized data.**
   b. **It allows for unlimited flexibility when querying data from the structure.**
   c. **Leaf-level data is kept intact and untransformed for reference and use for the future.**
   d. **The methodology encourages experimentation and exploration.**
   e. **It increases the speed of generating fresh actionable knowledge.**
   f. **It reduces the cycle time between data generation to availability of actionable knowledge**

Q. Data Processing Tool

Spark, Mesos, Akka, Cassandra, Kafka, Elastic Search, R, Scala, Python.

6640_Shivam

Ch 2 Layered Framework

Q. CrossIndustry Standard Process for Data Mining (CRISP-DM)

CRISP-DM was generated in 1996, and by 1997, it was extended via a European Union project, under the ESPRIT funding initiative.

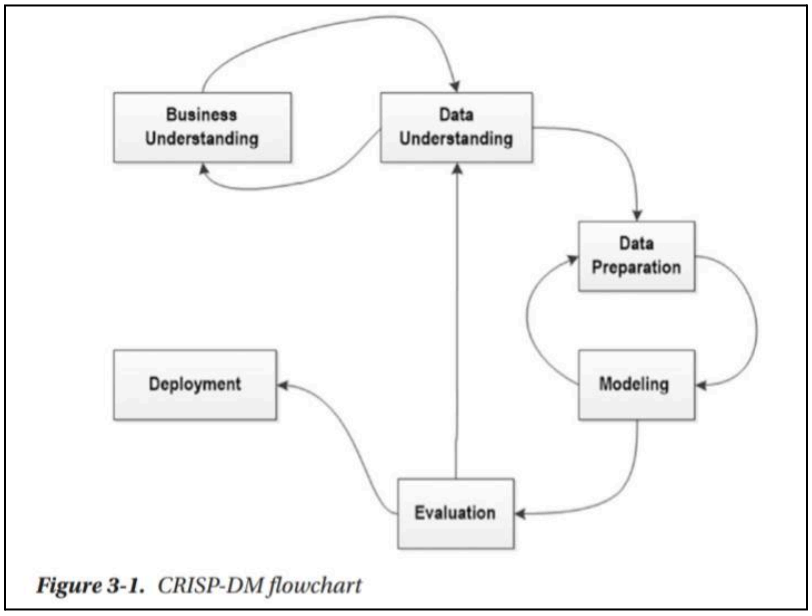


Figure 3-1. CRISP-DM flowchart

1. **Business Understanding**
    1. **This initial phase indicated concentrates on discovery of the data science goals and requests from a business perspective.**
    2. **Many businesses use a decision model or process mapping tool that is based on the Decision Model and Notation (DMN) standard**

2. **Data Understanding**
    1. **The data understanding phase starts with an initial data collection and continues with actions to discover the characteristics of the data.**
    2. **This phase identifies data quality complications and insights into the data.**

3 . **Data Preparation**
    1. **The data preparation phase covers all activities to construct the final data set for modeling tools.**
    2. **This phase is used in a cyclical order with the modeling phase, to achieve a complete model.**

4. **Modeling**
    1. **In this phase different data science modeling techniques are nominated and evaluated for accomplishing the prerequisite outcomes, as per the business requirements.**

5. **Evaluation**
    1. **If this fails, the process returns to the data understanding phase, to improve the delivery.**

6. **Deployment**
    1. **Creation of data science is generally not the end of the project.**
    2. **Once the data science is past the development and pilot phases, it has to go to production**

Q. Homogeneous Ontology for Recursive Uniform Schema

1. **The Homogeneous Ontology for Recursive Uniform Schema (HORUS) is used as an internal data format structure that enables the framework to reduce the permutations of transformations required by the framework.**
2. **The use of HORUS methodology results in a hub-and-spoke data transformation approach.**
3. **External data formats are converted to HORUS format, and then a HORUS format is transformed into any other external format.**
4. **The basic concept is to take native raw data and then transform it first to a single format.**
5. **That means that there is only one format for text files, one format for JSON or XML, one format for images and video.**
6. **Therefore, to achieve any-to-any transformation coverage, the framework's only requirements are a data-format-to-HORUS and HORUS-to-data-format converter.**
7. **By using the hub-and-spoke methodology on text files to reduce the amount of convertor scripts you must achieve an effective data science solution.**
8. **You can at only 100 text data sets, you can save 98% on a non-hub-and-spoke framework.**

6640_Shivam

Q. Top Layered Framework

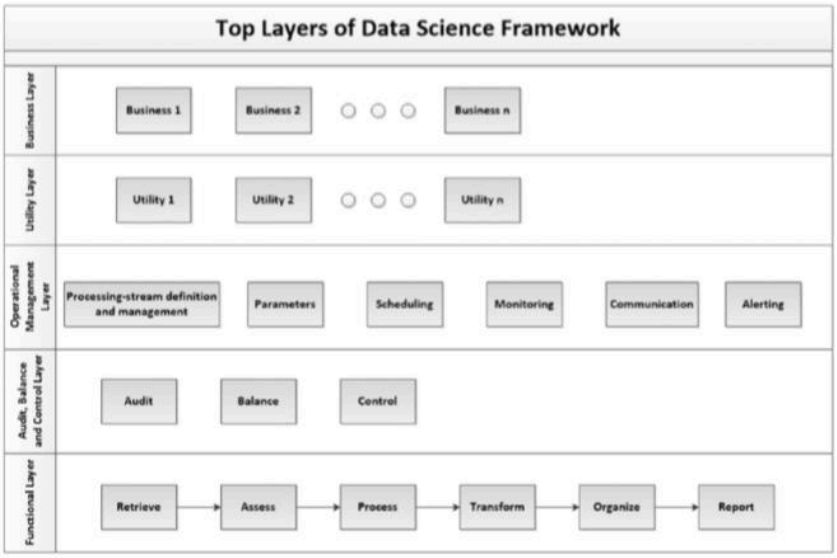


Figure 3-2. *Top layers of the data science framework*

The top layers are to support a long-term strategy of creating a Center of Excellence for your data science work.

Q. What is a functional layer? What are the steps of the processing algorithm?

1. **The functional layer of the data science ecosystem is the main layer of programming required.**
2. **The functional layer is the part of the ecosystem that executes the comprehensive data science. It consists of several structures.**
    - **Data models**
    - **Processing algorithms**
    - **Provisioning of infrastructure**

The processing algorithm is spread across six supersteps of processing, as follows:

1. Retrieve: This super step contains all the processing chains for retrieving data from the raw data lake via a more structured format.

2. Assess: This superstep contains all the processing chains for quality assurance and additional data enhancements.

3. Process: This superstep contains all the processing chains for building the data vault.

4. Transform: This superstep contains all the processing chains for building the data warehouse.

5. Organize: This superstep contains all the processing chains for building the data marts.

6. Report: This superstep contains all the processing chains for building virtualization and reporting the actionable knowledge.

6640_Shivam

Q. Audit , Balance and Control

1. **The audit, balance, and control layer is the area from which you can observe what is currently running within your data science environment.**
2. **It records**
   - **Process-execution statistics**
   - **Balancing and controls**
   - **Rejects and error-handling**
   - **Codes management**
3. **Audit**
   a. **The audit sublayer records any process that runs within the environment.**
   b. **This information is used by data scientists and engineers to understand and plan improvements to the processing.**
   c. **Make sure your algorithms and processing generate a good and complete audit trail.**
4. **Balance**
   a. **The balance sublayer ensures that the ecosystem is balanced across the available processing capability or has the capability to top-up capability during periods of extreme processing.**
   b. **The processing on demand capability of a cloud ecosystem is highly desirable for this purpose.**
   c. **Plan your capability as a combination of always-on and top-up processing. That way, you get maximum productivity from your processing cycle.**

5. **Control**
   a. **The control sublayer, controls the execution of the current active data science processes in a production ecosystem.**
   b. **The control elements are a combination of the control element within the Data Science Technology Stack's individual tools plus a custom interface to control the primary workflow.**
   c. **The control also ensures that when processing experiences an error, it can attempt a recovery, as per your requirements, or schedule a clean-up utility to undo the error.**

6640_Shivam

Ch 3 Business Layer

Short Tips: The business layer is where we record the interactions with the business. This is where we convert business requirements into data science requirements.

Q.What is meant by dimension? Discuss SCD Type I and SCD Type II dimensions in brief.

1. **A dimension is a structure that categorizes facts and measures, to enable you to respond to business questions.**
2. **A slowly changing dimension is a data structure that stores the complete history of the data loads in the dimension structure over the life cycle of the data lake.**
3. **There are several types of Slowly Changing Dimensions (SCDs) in the data warehousing design toolkit that enable different recording rules of the history of the dimension.**
4. **Types of SCD**
   a. **SCD Type 1—Only Update**
   b. **SCD Type 2—Keeps Complete History**
   c. **SCD Type 3—Transition Dimension**
   d. **SCD Type 4—Fast-Growing Dimension.**
5. **SCD Type 1—Only Update**

This dimension contains only the latest value for specific business information.

Example: Dr Jacob Roggeveen lived on Rapa Nui, Easter Island, but now lives in Middelburg, Netherlands.

Load	First Name	Last Name	Address
1	Jacob	Roggeveen	Rapa Nui, Easter Island
2	Jacob	Roggeveen	Middelburg, Netherlands

The SCD Type 1 records only the first place Dr Jacob Roggeveen lived and ignores the second load's changes.

Person		Location
First Name	Last Name	Address
Jacob	Roggeveen	Rapa Nui, Easter Island

I find this useful for storing the first address you register for a user, or the first product he or she has bought.

6. **SCD Type 2—Keeps Complete History**
   a. **This dimension contains the complete history of the specific business information. There are three ways to construct a SCD Type 2.**
   1. Versioning: The SCD keeps a log of the version number, allowing youto read the data back in the order in which the changes were made.

2. Flagging: The SCD flags the latest version of the business data rows a 1 and the rest as 0. This way, you can always get the latest row.
3. Effective Date: The SCD keeps records of the period in which the specific value was valid for the business data row.

Q. What are capacity, current and forecast requirements of the data science ecosystem?

1. **Capacity is the ability to load, process, and store a specific quantity of data by the data science processing solution.**
2. **You must track the current and forecast the future requirements, because as a data scientist, you will design and deploy many complex models that will require additional capacity to complete the processing pipelines you create during your processing cycles.**
3. **Capacity is measured per the component's ability to consistently maintain specific levels of performance as data load demands vary in the solution.**
4. **The correct way to record the requirement is Component C will provide P% capacity for U users, each with M MB of data during a time frame of T seconds.**
5. **Example:**
   The data hard drive will provide 95% capacity for 1000 users, each with 10MB of data during a time frame of 10 minutes.
6. **Concurrency is the measure of a component to maintain a specific level of performance when under multiple simultaneous loads conditions.**
7. **The correct way to record the requirement is**
   Component C will support a concurrent group of U users running predefined acceptance script S simultaneously.
8. **Example:**
   The memory will support a concurrent group of 100 users running a sort algorithm of 1000 records simultaneously.

Q. Sun model 1 , 2 , and 3.

Important questions from bharti ma'am for Unit 1

Q. Explain the following terms in brief:
What is Homogeneous Ontology for Recursive Uniform Schema (HORUS) and hub-and-spoke methodology?
What are the top layers of the data science framework?
What is a functional layer? What are the steps of the processing algorithm?
Discuss audit, balance and control sublayers in brief.
What is meant by dimension? Discuss SCD Type I and SCD Type II dimensions in brief.
What are capacity, current and forecast requirements of the data science ecosystem?

Unit 2 : Ch 4 Utility Layer

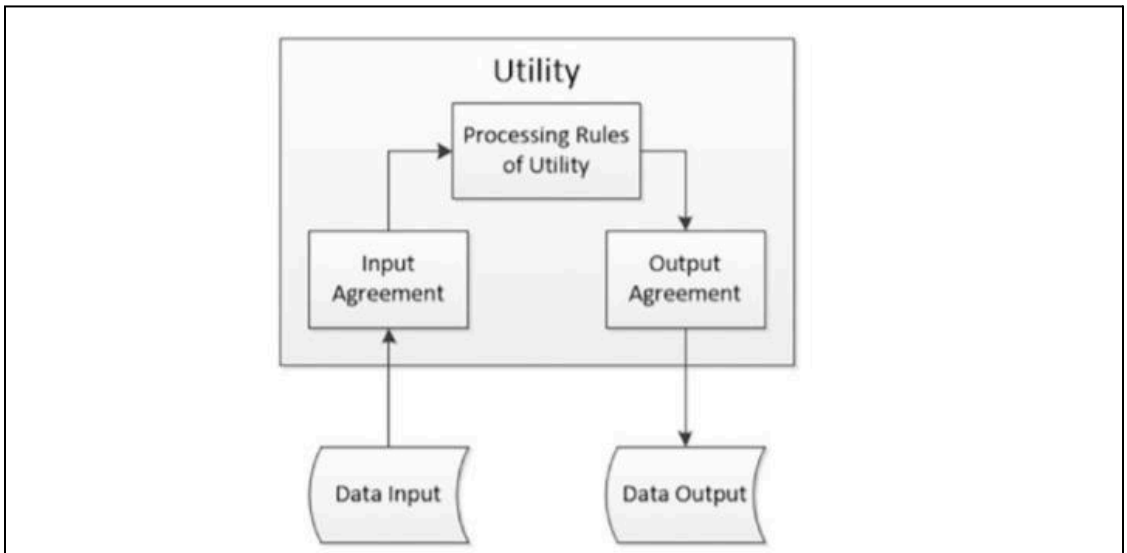
Short Tips: The utility layer is used to store repeatable practical methods of data science

Utilities are the common and verified workhorses of the data science ecosystem. The

utility layer is a central storehouse for keeping all one's solutions utilities in one place.

Q. What are the stages of basic utility design? Discuss each in brief.

1. **The basic utility must have a common layout to enable future reuse and enhancements.**
2. **This standard makes the utilities more flexible and effective to deploy in a large-scale ecosystem.**
3. **Three-stage process of Basic Utility.**
    1. **Load data as per input agreement.**
    2. **Apply processing rules of utility.**
    3. **Save data as per output agreement.**



gure 5-1. Basic utility design

4. **The main advantage of this methodology in the data science ecosystem is that you can build a rich set of utilities that all your data science algorithms require.**
5. **You have a basic pre-validated set of tools to use to perform the common processing and**
6. **then spend time only on the custom portions of the project.**

7. **You can also enhance the processing capability of your entire project collection with one single new utility update.**
8. **There are three types of utilities**
   - **Data processing utilities**
   - **Maintenance utilities**
   - **Processing utilities**

Q. Write a note on retrieving utilities. Or Discuss the various retrieve utilities of data collection in brief.

1. **Retrieve utilitie comes under Data Processing Utility.**
2. **Utilities for this superstep contain the processing chains for retrieving data out of the raw data lake into a new structured format.**
3. **You build all your retrieve utilities to transform the external raw data lake format into the Homogeneous Ontology for Recursive Uniform Schema (HORUS) data format.**
4. **Text-Delimited to HORUS**
   These utilities enable your solution to import text-based data from your raw data
5. **XML to HORUS**
   These utilities enable your solution to import XML-based data from your raw data sources
6. **JSON to HORUS**
   These utilities enable your solution to import XML-based data from your raw data sources
7. **Database to HORUS**
   These utilities enable your solution to import data from existing database sources
8. **Picture to HORUS**
   These expert utilities enable your solution to convert a picture into extra data.
9. **Video to HORUS**
   These expert utilities enable your solution to convert a video into extra data.

Q. Explain different fixer utilities with an example for each. (Examples imp)

1. **It comes under Access Utilities.**
2. **Fixers enable your solution to take your existing data and fix a specific quality issue.**
3. **Examples include**
   Removing leading or lagging spaces from a data entry
   Example in Python:

```
baddata = " Data Science with too many spaces is bad!!! "
print('>',baddata,'<')
cleandata=baddata.strip()
print('>',cleandata,'<')
```

4. **Removing nonprintable characters from a data entry**
   Example in Python:

```
import string
printable = set(string.printable)
baddata = "Data\x00Science with\x02 funny characters is
\x10bad!!!"
cleandata=''.join(filter(lambda x: x in string.printable,
baddata))
print(cleandata)
```

5. **Reformatting data entry to match specific formatting criteria. Convert 2017/01/31 to 31 January 2017**
   Example in Python:

```
import datetime as dt
baddate = dt.date(2017, 1, 31)
baddata=format(baddate,'%Y-%m-%d')
print(baddata)
gooddate = dt.datetime.strptime(baddata,'%Y-%m-%d')
gooddata=format(gooddate,'%d %B %Y')
print(gooddata)
```

Q. Explain various data vault utilities.

1. **The data vault is a highly specialist data storage technique that was designed by Dan Linstedt.**
2. **The data vault is a detail-oriented, historical-tracking, and uniquely linked set of normalized tables that support one or more functional areas of business.**
3. **It is a hybrid approach encompassing the best of breed between 3rd normal form (3NF) and star schema.**

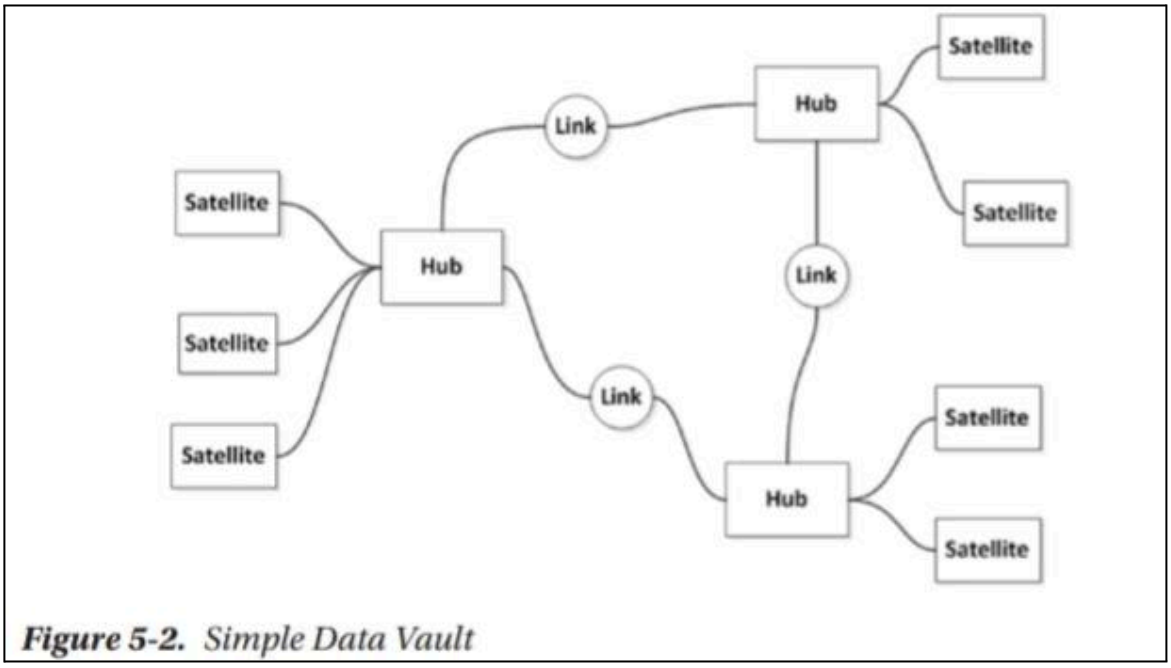


Figure 5-2. Simple Data Vault

4. **Hub Utilities**
   Hub utilities ensure that the integrity of the data vault's (Time, Person, Object, Location, Event) hubs is 100% correct, to verify that the vault is working as designed.
5. **Satellite Utilities**
   Satellite utilities ensure the integrity of the specific satellite and its associated hub.
6. **Link Utilities**
   Link utilities ensure the integrity of the specific link and its associated hubs.

6640_Shivam

Q. What are the objectives of organizing, reporting and maintaining utilities.

1. **Organize Utilities**

   1. Utilities for this superstep contain all the processing chains for building the data marts.

   2. The organize utilities are mostly used to create data marts against the data science results stored in the data warehouse dimensions and facts.

2. **Report Utilities**
   a. **Utilities for this superstep contain all the processing chains for building virtualization and reporting of the actionable knowledge**
   b. **The report utilities are mostly used to create data virtualization against the data science results stored in the data marts.**

3. **Maintenance Utilities**
   a. **The data science solutions you are building are a standard data system and, consequently, require maintenance utilities, as with any other system.**
   b. **Data engineers and data scientists must work together to ensure that the ecosystem works at its most efficient level at all times.**

6640_Shivam

Q. Discuss various scheduling utilities in brief.

1. **Scheduling Utilities**
   The scheduling utilities I use are based on the basic agile scheduling principles.

2. **Backlog Utilities**
   Backlog utilities accept new processing requests into the system and are ready to be processed in future processing cycles.

3. **To-Do Utilities**
   The to-do utilities take a subset of backlog requests for processing during the next processing cycle.
   They use classification labels, such as priority and parent-child relationships, to decide what process runs during the next cycle.

4. **Doing Utilities**
   The doing utilities execute the current cycle's requests.

5. **Done Utilities**
   The done utilities confirm that the completed requests performed the expected processing.

6. **Monitoring Utilities**
   The monitoring utilities ensure that the complete system is working as expected.

Ch. 5 Three Management Layers

Q. What are operational management layers? (Imp)

1. **The operational management layer is the core store for the data science ecosystem's complete processing capability.**
2. **The layer stores every processing schedule and workflow for the all-inclusive ecosystem.**
3. **This area enables you to see a singular view of the entire ecosystem. It reports the status of the processing.**
4. **The operations management layer is the layer where you record the following.**
5. **Processing-Stream Definition and Management**
    a. **The processing-stream definitions are the building block of the data science ecosystem.**
    b. **You can store all your current active processing scripts in this section.**
6. **Parameters**
   The parameters for the processing are stored in this section, to ensure a single location for all the system parameters
7. **Scheduling**
   The scheduling plan is stored in this section, to enable central control and visibility of the complete scheduling plan for the system
8. **Monitoring**
   The central monitoring process is in this section to ensure that there is a single view of the complete system.
9. **Communication**
   All communication from the system is handled in this one section, to ensure that the system can communicate any activities that are happening.
10. **.Alerting**
    The alerting section uses communications to inform the correct person, at the correct time, about the correct status of the complete system.

6640_Shivam

Q. Explain the concept of scheduling with the help of the Drum-Buffer-Rope method.

1. **The scheduling plan is stored in this section, to enable central control and visibility of the complete scheduling plan for the system.**

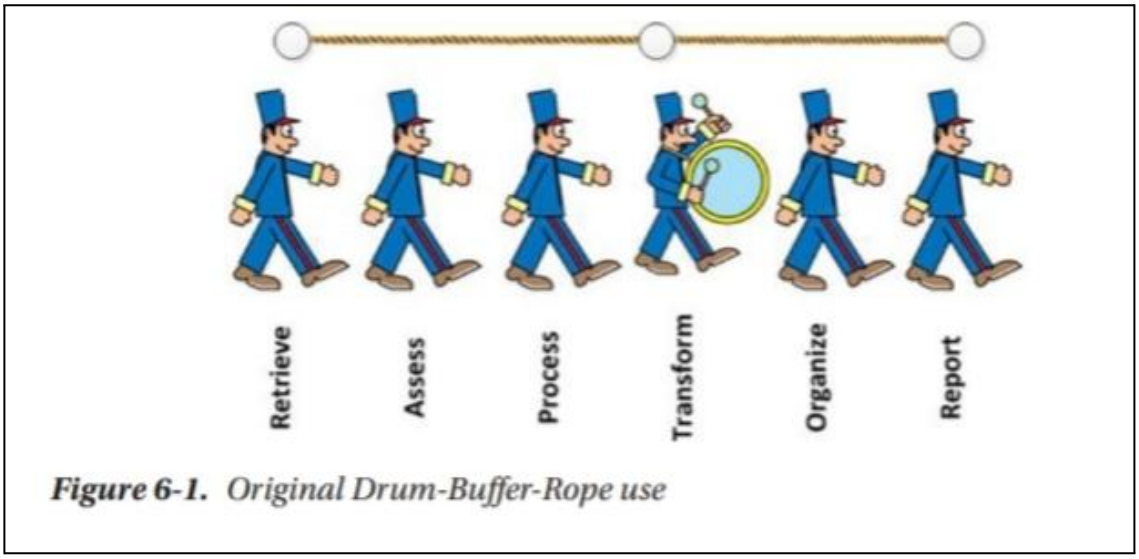


Figure 6-1. Original Drum-Buffer-Rope use

2. **The Drum-Buffer-Rope methodology is a standard practice to identify the slowest process and then use this process to pace the complete pipeline.**
3. **You then tie the rest of the pipeline to this process to control the eco-system's speed.**
4. **So, you place the "drum" at the slow part of the pipeline, to give the processing pace, and attach the "rope" to the beginning of the pipeline,**
5. **And the end by ensuring that no processing is done that is not attached to this drum.**
6. **This ensures that your processes complete more efficiently, as nothing is entering or leaving the process pipe without been recorded by the drum's beat.**

6640_Shivam

Q. What is built-in logging? Discuss debug, information, warning, error and fatal watcher in brief.

1. **Design your logging into an organized preapproved location, to ensure that you capture every relevant log entry**
2. **Debug Watcher**
    a. **This is the maximum verbose logging level.**
    b. **If you discover any debug logs in my ecosystem, you can normally raise an alarm.**
3. **Information Watcher**
    a. **The information level is normally utilized to output information that is beneficial to the running and management of a system.**
    b. **You can pipe these logs to the central Audit, Balance, and Control data store.**
4. **Warning Watcher**
    a. **Warning is often used for handled “exceptions” or other important log events.**
    b. **Usually this means that the tool handled the issue and took corrective action for recovery.**
5. **Error Watcher**
    a. **Error is used to log all unhandled exceptions in the tool.**
    b. **It means that a specific step in the planned processing did not complete as expected.**
6. **Fatal Watcher**
    a. **Fatal is reserved for special exceptions/conditions for which it is imperative that you quickly identify these events**
    b. **It means a specific step in the planned processing has not completed as expected.**

Ch. 6. Retrieve Superstep

Q. What is a data swamp? Explain the basic steps of avoiding the data swamp.

1. **Data swamps are simply data lakes that are not managed. They are not to be feared. They need to be tamed.**
2. **Following are my four critical steps to avoid a data swamp.**
3. **Start with Concrete Business Questions**
    a. **Perform a comprehensive analysis of the entire set of data you have and then apply a metadata classification for the data, stating full data lineage for allowing it into the data lake.**
    b. **The data lake must be enabled to collect the data required to answer your business questions.**
4. **Data Governance**
    a. **The role of data governance, data access, and data security does not go away with the volume of data in the data lake.**
    b. **It simply collects together into a worse problem, if not managed.**
5. **Data Source Catalog**
    a. **Metadata that link ingested data-to-data sources are a must-have for any data lake.**
    b. **Following as general rules for the data you process.**
    c. **Unique data catalog number,Short description (keep it under 100 characters),Long description (keep it as complete as possible),Contact information for external data source,Expected frequency, Internal business purpose.**
6. **Business Glossary**
    a. **The business glossary maps the data-source fields and classifies them into respective lines of business.**
    b. **This glossary is a must-have for any good data lake.**

6640_Shivam

Important questions from bharti ma'am for Unit 2

Ch. Utility Layer
What are the stages of basic utility design? Discuss each in brief.
Write a note on retrieving utilities. Or Discuss the various retrieve utilities of data collection in brief.
Explain different fixer utilities with an example for each.
Explain various data vault utilities.
Explain following data science utilities with an example for each:
1. Data binning and bucketing
2. Averaging of data
3. Outlier detection
What are the objectives of organizing, reporting and maintaining utilities.
Discuss various scheduling utilities in brief.

Ch. Three Management Layers
What are operational management layers?
Explain the concept of scheduling with the help of the Drum-Buffer-Rope method.
What is built-in logging? Discuss debug, information, warning, error and fatal watcher in brief.

Ch. Retrieve Superstep
What is a data swamp? Explain the basic steps of avoiding the data swamp.
Explain the data analytical model that should be executed on every data set in the data lake.
Explain the concept of training the trainer model.
Explain the following shipping terms in brief:
1. Seller and Buyer
2. Carrier
3. Port and Named Place
4. Ship
5. Terminal
Explain the following shipping terms with example:
1. Ex Works
2. Free Carrier
3. Carriage Paid To
4. Carriage and Insurance Paid To
5. Delivered at Terminal
6. Delivered at Place
7. Delivery Duty Paid
List and explain the different data stores used in data science.

6640_Shivam

MCQ From Unit1 Chapter 1

1. **Which of the following is a primary component of the Data Science Technology Stack?
- a) Data Collection
- b) Data Visualization
- c) Data Storage
- d) All of the above

Answer: d) All of the above

2. **Which programming language is widely used for data analysis and has extensive libraries for data science?
- a) Java
- b) R
- c) C++
- d) Ruby

Answer: b) R

3. **What is the role of a Data Warehouse in the Data Science Technology Stack?
- a) To store raw, unstructured data
- b) To integrate and store data from different sources for analysis
- c) To perform data cleaning
- d) To visualize data insights

Answer: b) To integrate and store data from different sources for analysis

4. **Which tool is commonly used for data visualization in the Data Science Technology Stack?
- a) Hadoop
- b) Tableau
- c) Spark
- d) MySQL

Answer: b) Tableau

5. **Which of the following frameworks is used for big data processing?
- a) Pandas
- b) TensorFlow
- c) Apache Spark
- d) Scikit-learn

Answer: c) Apache Spark

6. **Which SQL-based system is known for its scalability and ability to handle large datasets?
- a) PostgreSQL
- b) Oracle DB
- c) MySQL
- d) Amazon Redshift

Answer: d) Amazon Redshift

7. **In the context of machine learning, which library is commonly used in Python for building neural networks?
- a) Matplotlib
- b) NumPy
- c) TensorFlow
- d) SciPy

Answer: c) TensorFlow

8. **What is the purpose of ETL in the Data Science Technology Stack?
- a) Extract, Transform, Load data for analysis
- b) Encrypt, Transfer, Log data for security
- c) Extract, Test, Load data for validation
- d) Evaluate, Transform, Link data for integration

Answer: a) Extract, Transform, Load data for analysis

9. **Which of the following is NOT typically a part of a data science technology stack?
- a) Data Mining
- b) Data Cleansing
- c) Data Visualization
- d) Network Security

Answer: d) Network Security

10. **Which tool is best known for its ability to handle and process large-scale data across distributed systems?
- a) Excel
- b) Hadoop
- c) SAS
- d) SPSS

Answer: b) Hadoop

Certainly! Here are some multiple-choice questions (MCQs) related to the topics you provided:

Chapter 1

MCQs on Data Science Technology Stack Topics

1. **What is the primary function of a Data Lake in the data science ecosystem?
- a) To store structured data in a relational database
- b) To integrate and analyze real-time data streams
- c) To store raw and unstructured data for later processing and analysis
- d) To provide a user interface for data visualization

Answer: c) To store raw and unstructured data for later processing and analysis

2. **Which of the following is a key feature of the Data Vault model in data warehousing?
- a) Real-time data processing
- b) Data security and encryption
- c) Separation of business logic from data storage
- d) Historical data tracking with auditable history

Answer: d) Historical data tracking with auditable history

3. **What does a Data Warehouse Bus Matrix typically represent?
- a) A mapping of data sources to data targets
- b) A graphical representation of data flows between ETL processes
- c) A framework for organizing and managing data warehouse dimensions and facts
- d) A schedule for data backup operations

Answer: c) A framework for organizing and managing data warehouse dimensions and facts

4. **Which tool is primarily used for distributed data processing and can handle large-scale data operations across multiple nodes?
- a) Elasticsearch
- b) Apache Spark
- c) Apache Kafka
- d) Apache Cassandra

Answer: b) Apache Spark

5. **Which of the following technologies is designed for managing distributed data and ensuring high availability with fault tolerance?
- a) Apache Kafka
- b) Apache Mesos
- c) Elasticsearch

- d) Apache Cassandra

Answer: d) Apache Cassandra

6. What is the main purpose of Apache Kafka in the data science technology stack?
- **a) To perform real-time data analytics**
- **b) To store large volumes of structured data**
- **c) To provide a distributed streaming platform for real-time data feeds**
- **d) To visualize data with interactive dashboards**

Answer: c) To provide a distributed streaming platform for real-time data feeds

7. Which of the following is NOT a programming language commonly used for data science?
- **a) Python**
- **b) Scala**
- **c) R**
- **d) JavaScript**

Answer: d) JavaScript

8. Which of the following is a tool used for full-text search and analysis, known for its distributed nature and scalability?
- **a) Apache Mesos**
- **b) Apache Spark**
- **c) Elasticsearch**
- **d) R**

Answer: c) Elasticsearch

9. Which technology provides a framework for managing cluster resources and job scheduling in a distributed environment?
- **a) Apache Akka**
- **b) Apache Spark**
- **c) Apache Mesos**
- **d) Apache Kafka**

Answer: c) Apache Mesos

10. Which programming language is known for its ease of use in statistical analysis and has a strong ecosystem of libraries for data science?
- **a) Python**
- **b) Scala**
- **c) Java**
- **d) C#**

Answer: a) Python

Chapter 2

Certainly! Here are some multiple-choice questions (MCQs) related to the topics you've mentioned from the chapter on "Layered Framework" in the context of data science:

1. **What does a Data Science Framework typically describe?**
   - a) A set of algorithms for data processing
   - b) A structured approach to organizing and managing data science tasks and processes
   - c) A method for visualizing data insights
   - d) A programming language for statistical analysis

   Answer: b) A structured approach to organizing and managing data science tasks and processes

2. **Which of the following is the full form of CRISP-DM?**
   - a) Cross-Industry Standard Process for Data Management
   - b) Critical Information Standard Process for Data Mining
   - c) Cross-Industry Standard Process for Data Mining
   - d) Core Integrated Standard Process for Data Modeling

   Answer: c) Cross-Industry Standard Process for Data Mining

3. **What is the primary purpose of the CRISP-DM methodology?**
   - a) To define a standard data warehousing architecture
   - b) To provide a structured approach for data mining and analytics projects
   - c) To develop real-time data processing pipelines
   - d) To create data visualizations

   Answer: b) To provide a structured approach for data mining and analytics projects

4. **The Homogeneous Ontology for Recursive Uniform Schema is primarily used for:**
   - a) Integrating heterogeneous data sources
   - b) Standardizing data storage in data lakes
   - c) Creating consistent data schemas across different systems
   - d) Visualizing data relationships

   Answer: c) Creating consistent data schemas across different systems

5. **Which of the following best describes the Top Layers of a Layered Framework in data science?**
   - a) The physical storage and retrieval of data

- b) The data processing and transformation stages
- c) The user interfaces and visualization components
- d) The data collection and preprocessing stages
Answer: c) The user interfaces and visualization components

6. **In the context of a Layered Framework for High-Level Data Science and Engineering, which layer is typically responsible for handling data storage and retrieval?**
- a) Application Layer
- b) Data Layer
- c) Presentation Layer
- d) Business Logic Layer
Answer: b) Data Layer

7. **Which of the following is a key component of the Layered Framework for High-Level Data Science and Engineering?**
- a) Data Preprocessing
- b) Data Visualization
- c) Data Storage and Management
- d) Data Collection
Answer: c) Data Storage and Management

8. **What is the main function of the Application Layer in a Layered Framework?**
- a) To provide tools for data storage
- b) To implement business logic and data processing
- c) To create visual representations of data
- d) To manage and retrieve data from databases
Answer: b) To implement business logic and data processing

9. **In the context of data science frameworks, which layer focuses on presenting data insights to users?**
- a) Data Layer
- b) Application Layer
- c) Presentation Layer
- d) Integration Layer
Answer: c) Presentation Layer

10. **Which layer in a Layered Framework is responsible for the integration and processing of data from various sources?**
- a) Data Layer
- b) Application Layer
- c) Integration Layer
- d) Presentation Layer
Answer: c) Integration Layer

Chapter 3

Certainly! Here are some multiple-choice questions (MCQs) related to the "Business Layer" and "Engineering a Practical Business Layer" topics:

1. **What is the primary role of the Business Layer in a data science or data engineering architecture?**
   - a) To manage and store raw data
   - b) To implement and execute data processing algorithms
   - c) To provide a bridge between business requirements and technical implementation
   - d) To visualize and present data insights

   Answer: c) To provide a bridge between business requirements and technical implementation

2. **Which component is typically included in the Business Layer to ensure that data solutions align with organizational goals?**
   - a) Data Warehousing
   - b) Business Intelligence Tools
   - c) Business Rules and Logic
   - d) Data Integration Tools

   Answer: c) Business Rules and Logic

3. **In engineering a practical Business Layer, which of the following is essential for translating business needs into technical specifications?**
   - a) Data Warehousing
   - b) Requirements Gathering and Analysis
   - c) Data Visualization
   - d) Data Storage Solutions

   Answer: b) Requirements Gathering and Analysis

4. **What is a key challenge when engineering a practical Business Layer in a data system?**
   - a) Ensuring data is stored in a normalized format
   - b) Managing real-time data streaming
   - c) Aligning technical solutions with changing business requirements
   - d) Developing data models for historical data analysis

   Answer: c) Aligning technical solutions with changing business requirements

5. **Which of the following practices is crucial when designing a Business Layer to support decision-making processes?**
   - a) Implementing advanced data encryption techniques
   - b) Developing detailed data models
   - c) Incorporating user feedback and business insights
   - d) Setting up distributed data storage systems

   Answer: c) Incorporating user feedback and business insights

6. **Which of the following best describes the relationship between the Business Layer and the Data Layer in a data architecture?**
   - a) The Business Layer provides data storage capabilities to the Data Layer.
   - b) The Business Layer implements data processing algorithms that the Data Layer manages.
   - c) The Business Layer translates business needs into data requirements that the Data Layer fulfills.
   - d) The Business Layer handles data visualization, while the Data Layer focuses on business logic.

   Answer: c) The Business Layer translates business needs into data requirements that the Data Layer fulfills.

7. **What is a common tool or technique used in the Business Layer to support business decision-making?**
   - a) SQL Databases
   - b) Data Warehouses
   - c) Business Intelligence (BI) Tools
   - d) Data Lakes

   Answer: c) Business Intelligence (BI) Tools

8. **In the context of engineering a Business Layer, which approach is often used to ensure that the layer remains effective as business needs evolve?**
   - a) Implementing rigid data models
   - b) Regularly updating and revising business rules and logic
   - c) Focusing solely on technical specifications
   - d) Limiting user access to data

   Answer: b) Regularly updating and revising business rules and logic

9. **When creating a Business Layer, which aspect is crucial for effective communication between technical teams and business stakeholders?**
   - a) Developing complex data algorithms

- b) Creating clear and comprehensive documentation of business requirements
- c) Implementing real-time data processing systems
- d) Designing intricate data storage solutions

Answer: b) Creating clear and comprehensive documentation of business requirements

10. **Which of the following is NOT typically a responsibility of the Business Layer?**
- a) Mapping business processes to data solutions
- b) Defining data integration strategies
- c) Designing the user interface for data visualization
- d) Implementing low-level data storage technologies

Answer: d) Implementing low-level data storage technologies

6640_Shivam

U2 - Chapter 1

Certainly! Here are some multiple-choice questions (MCQs) focused on the "Utility Layer" and its practical engineering aspects in the context of data science:

1. **What is the main purpose of the Utility Layer in a data science architecture?**
   - a) To store and manage raw data
   - b) To provide essential services that support the functionality of other layers
   - c) To handle user authentication and access control
   - d) To create data visualizations and reports

   Answer: b) To provide essential services that support the functionality of other layers

2. **Which of the following is a common function of the Utility Layer?**
   - a) Data ingestion and processing
   - b) Data integration and ETL (Extract, Transform, Load)
   - c) User interface design
   - d) Business rule application

   Answer: b) Data integration and ETL (Extract, Transform, Load)

3. **What is a key consideration when designing the Utility Layer for data science projects?**
   - a) Ensuring high performance of data visualization tools
   - b) Providing robust data integration and transformation capabilities
   - c) Developing complex machine learning models
   - d) Creating interactive dashboards

   Answer: b) Providing robust data integration and transformation capabilities

4. **In the context of the Utility Layer, what does "data quality management" involve?**
   - a) Creating data visualizations
   - b) Implementing security measures
   - c) Monitoring and improving the accuracy and consistency of data
   - d) Designing data storage solutions

   Answer: c) Monitoring and improving the accuracy and consistency of data

5. **Which technology is commonly used in the Utility Layer to facilitate communication between different data systems?**
   - a) Data Lakes
   - b) Middleware and APIs
   - c) Data Warehouses
   - d) Data Visualization Tools

Answer: b) Middleware and APIs

6. **What is a critical component to include in the Utility Layer to support data processing and transformation?**
   - a) Data visualization libraries
   - b) Data integration platforms
   - c) User interface frameworks
   - d) Real-time analytics tools

   Answer: b) Data integration platforms

7. **When engineering a practical Utility Layer, which aspect is crucial for ensuring that it meets business needs?**
   - a) High-speed data storage solutions
   - b) Seamless integration with other data systems and tools
   - c) Complex algorithm development
   - d) Advanced data encryption

   Answer: b) Seamless integration with other data systems and tools

8. **Which of the following is NOT typically a responsibility of the Utility Layer?**
   - a) Providing data transformation services
   - b) Managing user access controls
   - c) Implementing data validation and cleansing processes
   - d) Generating business intelligence reports

   Answer: d) Generating business intelligence reports

9. **How does the Utility Layer support data science workflows in practical scenarios?**
   - a) By providing tools for real-time data streaming
   - b) By offering a platform for developing predictive models
   - c) By ensuring smooth data integration and processing pipelines
   - d) By designing interactive visualizations

   Answer: c) By ensuring smooth data integration and processing pipelines

10. **Which approach is commonly used in the Utility Layer to handle data transformation and enrichment?**
    - a) Data warehousing
    - b) ETL (Extract, Transform, Load) processes
    - c) Real-time data analytics
    - d) Data mining

    Answer: b) ETL (Extract, Transform, Load) processes

6640_Shivam

Unit 2 Chapter 2

Certainly! Here are some multiple-choice questions (MCQs) related to the "Three Management Layers" chapter from a data science perspective:

1. **What is the primary focus of the Operational Management Layer in data science?**
   - a) Strategic planning and decision-making
   - b) Day-to-day operations and monitoring of data systems
   - c) Data transformation and processing
   - d) Long-term data storage and archiving

   Answer: b) Day-to-day operations and monitoring of data systems

2. **Which component is crucial for defining and managing data streams in the Processing-Stream Definition and Management layer?**
   - a) Data Integration Tools
   - b) Data Quality Management
   - c) Data Visualization Tools
   - d) Real-Time Data Processing Frameworks

   Answer: d) Real-Time Data Processing Frameworks

3. **What is the main purpose of the Audit, Balance, and Control Layer?**
   - a) To oversee and ensure data quality and compliance
   - b) To manage data transformation processes
   - c) To design and implement data storage solutions
   - d) To visualize data insights for end-users

   Answer: a) To oversee and ensure data quality and compliance

4. **Which technique is used to measure and maintain balance and control in the context of data management?**
   - a) Predictive Analytics
   - b) Cause-and-Effect Analysis
   - c) Statistical Modeling
   - d) Data Mining

   Answer: b) Cause-and-Effect Analysis

5. **What is the role of the Balance, Control, Yoke Solution in data management?**
   - a) To balance workload across data systems
   - b) To implement data governance and control mechanisms
   - c) To integrate various data sources

- d) To design data storage and retrieval systems

Answer: b) To implement data governance and control mechanisms

6. **In the Functional Layer, what is the primary focus?**
- a) Managing operational tasks and monitoring
- b) Providing a framework for data processing and analytics
- c) Implementing data storage solutions
- d) Ensuring compliance with data regulations

Answer: b) Providing a framework for data processing and analytics

7. **Which system or approach is commonly used to ensure compliance and control in data management?**
- a) Data Warehousing
- b) Data Governance Frameworks
- c) Real-Time Data Processing Systems
- d) Data Mining Algorithms

Answer: b) Data Governance Frameworks

8. **What type of analysis is often used in the Audit, Balance, and Control Layer to identify data anomalies and ensure integrity?**
- a) Predictive Analytics
- b) Descriptive Statistics
- c) Cause-and-Effect Analysis
- d) Prescriptive Analytics

Answer: c) Cause-and-Effect Analysis

9. **Which of the following best describes the purpose of a Data Science Process within the management layers?**
- a) To provide a methodology for managing data operations
- b) To facilitate data integration and storage
- c) To outline steps for conducting data analysis and generating insights
- d) To implement real-time data streaming solutions

Answer: c) To outline steps for conducting data analysis and generating insights

10. **How does the Processing-Stream Definition and Management layer contribute to data science operations?**
- a) By providing tools for data visualization
- b) By ensuring data streams are properly defined, managed, and optimized for processing

6640_Shivam

- c) By creating data storage and retrieval solutions
- d) By implementing data security measures

Answer: b) By ensuring data streams are properly defined, managed, and optimized for processing

Chapter 3

Certainly! Here are some multiple-choice questions (MCQs) based on the topics from the chapter on "Retrieve Superstep" in data science:

1. **What is the primary purpose of a Data Lake?**
   - a) To store structured data in relational databases
   - b) To provide a centralized repository for storing raw, unstructured, and structured data at scale
   - c) To create real-time data visualizations
   - d) To implement complex machine learning algorithms

   Answer: b) To provide a centralized repository for storing raw, unstructured, and structured data at scale

2. **What is a Data Swamp?**
   - a) A well-organized and easily accessible data repository
   - b) A data lake that has become chaotic and disorganized, making it difficult to find and use data
   - c) A system for real-time data processing
   - d) A type of data visualization tool

   Answer: b) A data lake that has become chaotic and disorganized, making it difficult to find and use data

3. **What does the Training the Trainer Model refer to in the context of data science?**
   - a) A model for training machine learning algorithms
   - b) A framework for training internal stakeholders and data scientists on data lake usage and analytics
   - c) A method for training new data scientists on data engineering tools
   - d) A strategy for training end-users on data visualization tools

   Answer: b) A framework for training internal stakeholders and data scientists on data lake usage and analytics

4. **Why is understanding the business dynamics of the Data Lake important?**
   - a) To ensure that the data lake has sufficient storage capacity
   - b) To align data lake usage with business goals and needs, ensuring that the data can drive actionable insights
   - c) To implement real-time data processing
   - d) To design complex machine learning models

   **Answer: b) To align data lake usage with business goals and needs, ensuring that the data

can drive actionable insights**

5. **What is a key benefit of extracting actionable business knowledge from Data Lakes?**
   - a) Improved data encryption
   - b) Enhanced ability to generate business insights and drive decision-making based on comprehensive data analysis
   - c) Increased storage capacity
   - d) Simplified data visualization

   Answer: b) Enhanced ability to generate business insights and drive decision-making based on comprehensive data analysis

6. **When engineering a practical Retrieve Superstep, which aspect is critical for success?**
   - a) Implementing real-time data streaming solutions
   - b) Ensuring that data can be efficiently retrieved and processed from the data lake
   - c) Developing complex machine learning algorithms
   - d) Creating interactive dashboards

   Answer: b) Ensuring that data can be efficiently retrieved and processed from the data lake

7. **How can connecting to other data sources enhance the functionality of a Data Lake?**
   - a) By increasing data storage capacity
   - b) By enabling integration and consolidation of data from diverse sources, leading to more comprehensive insights
   - c) By improving real-time data processing speeds
   - d) By simplifying data encryption

   Answer: b) By enabling integration and consolidation of data from diverse sources, leading to more comprehensive insights

8. **Which of the following is NOT typically a challenge associated with Data Lakes?**
   - a) Difficulty in managing and organizing large volumes of diverse data
   - b) Ensuring data quality and consistency across sources
   - c) Simplifying data integration and transformation
   - d) Avoiding data redundancy and duplication

   Answer: c) Simplifying data integration and transformation

9. **What approach is commonly used to prevent a Data Lake from becoming a Data Swamp?**
   - a) Regularly updating data security protocols
   - b) Implementing robust data governance practices and metadata management
   - c) Increasing data storage capacity
   - d) Using advanced data visualization tools

Answer: b) Implementing robust data governance practices and metadata management

10. **Which of the following strategies helps in deriving actionable insights from Data Lakes?**
    - a) Focusing solely on historical data storage
    - b) Applying advanced analytics and data processing techniques to extract meaningful patterns and trends
    - c) Implementing data encryption techniques
    - d) Designing user-friendly interfaces for data access

Answer: b) Applying advanced analytics and data processing techniques to extract meaningful patterns and trends