(Big) Data Engineering In Depth

From Beginner to Professional

Moustafa Alaa
Senior Big Data Engineer
MoustafaAlaa Moustafa Alaa @Moustafa_alaa22
Garage Education
mustafa.alaa.mohamed@gmail.com

The Definitive Guide to Big Data Engineering Tasks

Videos classification

Watching Method / Audience	Computer	Mobile/Tablet	Just listening
Developer		●	
DevOps		●	
Business		●	

Table: Video classification
The green circle ● means short video.
The blue circle ● means medium video.
The red circle ● means long video

Section: Hot vs Cold Storage

Hot vs Cold Storage

What is multi-temperature Storage?

- (Most of) DWH solution design has multi-temperature data management model.
- What is the multi-temperature data management model?
  - It is a data classification design which allows us to have the following characteristics

What is multi-temperature Storage?

- (Most of) DWH solution design has multi-temperature data management model.
- What is the multi-temperature data management model?
  - It is a data classification design which allows us to have the following characteristics
    - (high performance) access on the frequent data (Hot data).

What is multi-temperature Storage?

- (Most of) DWH solution design has multi-temperature data management model.
- What is the multi-temperature data management model?
  - It is a data classification design which allows us to have the following characteristics
    - (high performance) access on the frequent data (Hot data).
    - Good (average performance) access to less-frequently data (warm data).

What is multi-temperature Storage?

- (Most of) DWH solution design has multi-temperature data management model.
- What is the multi-temperature data management model?
  - It is a data classification design which allows us to have the following characteristics
    - (high performance) access on the frequent data (Hot data).
    - Good (average performance) access to less-frequently data (warm data).
    - Availability to access rarely accessed data (cold data).

What is multi-temperature Storage?

- (Most of) DWH solution design has multi-temperature data management model.
- What is the multi-temperature data management model?
  - It is a data classification design which allows us to have the following characteristics
    - (high performance) access on the frequent data (Hot data).
    - Good (average performance) access to less-frequently data (warm data).
    - Availability to access rarely accessed data (cold data).
  - Who is responsible for data temperature classifications?

What is multi-temperature Storage?

- (Most of) DWH solution design has multi-temperature data management model.
- What is the multi-temperature data management model?
  - It is a data classification design which allows us to have the following characteristics
    - (high performance) access on the frequent data (Hot data).
    - Good (average performance) access to less-frequently data (warm data).
    - Availability to access rarely accessed data (cold data).
  - Who is responsible for data temperature classifications?
    - Demand team, product owner, or data architect (Based on the business needs).

Why do we need it?

- Why do we need the multi-temperature data management model?
  - Cost reduction **$$$$**

Why do we need it?

- Why do we need the multi-temperature data management model?
  - Cost reduction **$$$$**
  - Performance.

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:
    - Frequency of access

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:
    - Frequency of access
    - Data change rate.

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:
    - Frequency of access
    - Data change rate.
  - Identify which storage type is suitable for the project

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:
    - Frequency of access
    - Data change rate.
  - Identify which storage type is suitable for the project
    - We store the hot data on the fast storage system.

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:
    - Frequency of access
    - Data change rate.
  - Identify which storage type is suitable for the project
    - We store the hot data on the fast storage system.
    - Warm data (usual) stored on slightly slower storage.

How to implement a multi-temperature storage system?

- How to implement the multi-temperature data management model?
  - Before implementation, we need to know the following:
    - Frequency of access
    - Data change rate.
  - Identify which storage type is suitable for the project
    - We store the hot data on the fast storage system.
    - Warm data (usual) stored on slightly slower storage.
    - We store the cold data on the slowest storage.

How to implement a multi-temperature storage system? Cont.

- Design consideration to make the retention easily.
  - Table partitions need to be split based on the retention policy plan (date).

How to implement a multi-temperature storage system? Cont.

- Design consideration to make the retention easily.
  - Table partitions need to be split based on the retention policy plan (date).
  - Summary tables (agg) need to be maintained to reduce the need for access the cold storage.

How to implement a multi-temperature storage system? Cont.

- Design consideration to make the retention easily.
  - Table partitions need to be split based on the retention policy plan (date).
  - Summary tables (agg) need to be maintained to reduce the need for access the cold storage.
  - Backup, Recovery, and Rollback plans need to be automated and prepared/tested before moving the data.

How to implement a multi-temperature storage system? Cont.

- Implementation (summary):
  - There are lots of tools for this purpose and categorized as follows:

How to implement a multi-temperature storage system? Cont.

- Implementation (summary):
  - There are lots of tools for this purpose and categorized as follows:
    - Enterprise.

How to implement a multi-temperature storage system? Cont.

- Implementation (summary):
  - There are lots of tools for this purpose and categorized as follows:
    - Enterprise.
    - Open source.

How to implement a multi-temperature storage system? Cont.

- Implementation (summary):
  - There are lots of tools for this purpose and categorized as follows:
    - Enterprise.
    - Open source.
    - Cloud tools.

How to implement a multi-temperature storage system? Cont.

- Enterprise
  :
  - IBM InfoSphere
  - Informatica PowerCenter
  - Oracle Data Service Integrator
  - Talend Data Integration
  - Microsoft SQL

How to implement a multi-temperature storage system? Cont.

- Open source
  :
  - Apache NiFi*
  - CloverETL
  - Pentaho
  - Talend Open Studio*

How to implement a multi-temperature storage system? Cont.

- Cloud tools
  :
  - AWS Migration Services.
  - Azure Migration Tools.
  - Google Migration Services/Velostrata.
- Some cloud providers offer physical data movement services.
- How to choose the most suitable storage type for your project/organization?