**Reading Modes:**

3 type of reading mode into pyspark:

**1.Permissive Mode** :

* its a default reading mode.
* if spark is not able to parse it due to data type mismatch then make it as null without impacting other result

**2.DROPMALFORMED** :

* whatever record having parsing issue, i.e in any column there is issues it will totally ignore that record, it only show correct record.

**3.FailFast Mode** :

* while reading data into this mode it will fail as soon as it encounters any record have parsing issues.

**Writing Modes:**

**1.Append:**

* it is used to add the new rows or data to existing table or dataframe with out affecting the existing one.

**2.overwrite:**

* It is used to delete the existed data and insert the new data to a file in a given path

**3.Merge:**

* It is an advanced method in writing where we can do multiple operations in a single operation.
* We can update the data when there is a condition match, insert new data when condition matches and delete the data.

**Storage Account:**

* It is a high-level container for data storage in the cloud.
* We can create multiple containers in a sigle storage account
* It provides a centralized, scalable, and secure way to store various types of data like files, blobs, queues, tables.
* We have two types.
  + Blob storage account
  + ADLS gen2(Azure delta lake storage)
* In blob storage accounts we dont have hierarchical name space.
* In ADLS gen2 we have hierarchical name space.

**Blob Storage**

* This type is optimized for **unstructured data** and supports **blobs** like text and binary data(eg:- images, videos …).
* It has three access tiers for Blob Storage: **Hot**, **Cool**, and **Archive**.

**File Storage**

* For Azure File Shares, these are used to store and manage files that can be shared across Azure VMs and on-premises machines.

**Hierarchical Namespace**

* A Hierarchical Namespace is a feature in Azure Data Lake Storage Gen2
* It provides a folder-like structure for organizing files and directories, similar to the file system of an operating system.

**Delta Lake:**

* Delta Lake is an open-source storage layer developed by databricks.
* It is built on top of Parquet.
* It has features like ACID transactions, schema enforcement, versioning, and faster query performance.

**Write Modes in Delta Lake:**

1. **Overwrite**: Replaces existing data with new data.
2. **Append**: Adds new data to the existing data.
3. **Merge (Upsert)**: Performs insert, update, and delete operations in one step.

**Performance Improvements in Delta Lake:**

Delta Lake uses several techniques to optimize performance:

* **Z-Ordering**: Improves query performance by organizing data based on specific columns.
* **Data Skipping**: Skips irrelevant data blocks during queries.
* **Caching**: Stores frequently accessed data in memory for faster access.
* **Compaction (Optimize Command)**: Reduces small files and speeds up queries.