# Activity 6.5 – Parquet files and Delta Lake

**Summary.** To gain full advantage of modern data management frameworks like polars and pyspark, we need to rethink data storage. In this activitiy, you will watch a couple of videos that introduce the parquet file format, as well as delta lake. In the process of watching each video, you complete a number of tasks.

**Part 1 – parquet files.** Please watch <https://www.youtube.com/watch?v=1j8SdS7s_NY&t=8s> and record your answers to each of the following questions or discussion items

* What are advantages and disadvantages of TXT and CSV files?

Advantages – simple and widely supported, human-readable, for small datasets / sharing

Disadvantages – Not meant for large scale data, no compression, no schema / data type support, slow read / write performance

* What are the differences between Row-wise, columnar, and hybrid models for physical storage layout?

Row-wise storage is where data is stored row by row. Columnar storage is where data is stored column by column, hybrid models combine features of the other 2 models for optimized mixed workloads.

* What is projections pushdown?

It is a technique where only the required columns are retrieved from storage > all col

* What are the advantages of the hybrid model?

It combines the benefits of row / columnar storage models, improves performance, allows flexibility for operations

* How is the data organized in a parquet file?

In a columnar format with the file split into row groups, then organized into column chunks

* Discuss how parquet compression deals with repeated values.

Reduces data size by storing repeated values more efficiently

* When do you gain a benefit from page compression?

Returns storage saving and improved query performance (less data to read)

* What is predicate pushdown? Discuss various ways to optimize the predicate pushdowns.

It is a way to filter conditions at the storage level, reducing amount of data read into the memory. They can be optimized by creating efficient conditions, leveraging index, and data pruning

**Part 2 – Delta Lake.** Please watch <https://www.youtube.com/watch?v=LJtShrQqYZY> and record your answers to each of the following questions or discussion items. Note that you can stop at the start of the **Getting Started** section (about 23 minutes in).

* What is a data lake and why are they advantageous?

Data lakes are centralized repositories for storing un / semi / structured data at scale, advantages include diverse data type flexibility, large-scale data storage, and cost-effectiveness

* Why is it difficult to quality and reliability out of a data lake?

Usually lack schema enforcement = inconsistent / messy data, reliability issues come from potential data corruption, missing data, and complex data pipelines w/out strong governance

* Discuss each of the challenges faced when working with streaming data.

Dealing with high velocity / ensuring low latency processing, dealing with out-of-order / late data, ensuring consistency / correctness

* Summarize the data lake distractions.

Data lakes can become data swamps if not properly managed, lack of governance / data quality enforcement means unusable data / unreliable analytics

* Give an overview of the advantages of delta lake.
* Brings reliability, quality, performance improvements to data lakes, provides scalable metadata handling, schema enforcement
* What do we mean by data quality levels?

Bronze = raw data, silver = cleaned data, gold = aggregated / curated for analytics, helps to ensure data governance / improve data quality

* Give a short summary of the various patterns of use (streaming, batch, re-computation).

Streaming is supporting real-time data ingestion / processing with high reliability. Batch processing allows efficient handing of large-scale data transformations, re-computation is allowed based on historical data versions for consistency / recovery purposes