# Exercise #1: Research the parquet format (15 marks)

In the lab tutorial of module #1 we loaded a parquet file. Research this type of file, and provide a one-page summary describing the following:

## A bit of history of when it was introduced and what are the main benefits of this format.

Apache Parquet is an open-source, column-oriented data storage format introduced in July 2013 through a collaboration between Twitter and Cloudera to improve existing columnar storage solutions like Trevni (Wikipedia, 2024). Parquet offers significant benefits, including efficient data compression and encoding schemes, optimized storage utilization, and enhanced performance for handling complex, bulk data—making it ideal for big data analytics and interactive workloads (Databricks, 2024).

## The main structure and technologies behind it.

Parquet uses a columnar layout, meaning data is stored by columns rather than rows, which allows for efficient compression and query performance. Parquet employs several key technologies, such as dictionary encoding, bit-packing, and run-length encoding (RLE), to reduce storage space and enhance performance. The hybrid use of bit-packing and RLE ensures optimal compression for integer data, while the column-wise structure allows for more efficient querying and compression compared to row-based formats (Wikipedia, 2024).

## Which software/platform (s) support it.

Parquet is supported by various big-data processing frameworks within the Hadoop ecosystem, including Apache Hive, Apache Drill, and Apache Spark (Wikipedia, 2024). Additionally, it is compatible with serverless technologies like AWS Athena and Google BigQuery (Databricks, 2024). Parquet also supports several programming languages, such as Java, Python, and C++, making it highly versatile for developers (Wikipedia, 2024).

## A comparison between csv and parquet.

Apache Parquet and CSV differ significantly in structure and performance. CSV is a simple, row-based format commonly used by tools like Excel and Google Sheets, but it has limitations in terms of efficiency, especially for large datasets. CSV files require scanning entire rows of data, even when only specific columns are needed, leading to slower query times and higher costs in cloud environments. In contrast, Parquet is a columnar format, which allows it to only read the necessary columns, significantly reducing the amount of data scanned. This column-based approach, combined with efficient compression, results in faster query performance, lower storage requirements, and reduced costs (Databricks, 2024).

## List two other file types comparable to parquet.

Two file formats comparable to Apache Parquet are Avro and ORC. Avro is a row-based format that excels in write-heavy operations and schema evolution, making it ideal for use cases like data serialization in Hadoop. ORC, a columnar format like Parquet, is optimized for read-heavy operations, particularly in Hive environments, and offers superior compression and query performance (Upsolver, 2022).

# Exercise #2: setup your workspace (20 marks)

## Log into your azure portal and expand your data lake resource group created during lab exercise #1 and take a screen shot showing both the storage account and the analytics workspace. Make it contains all the naming explained in lab #1.

A computer screen shot of a computer screen

Description automatically generated

## Expand the storage account view and take a screenshot showing the first page and add it to your analysis report.

A computer screen shot of a computer

Description automatically generated

## Click on the “JSON view” tab, examine the output. Copy the content of the JSON file and add it to your analysis report.

{

    "sku": {

        "name": "Standard\_RAGRS",

        "tier": "Standard"

    },

    "kind": "StorageV2",

    "id": "/subscriptions/21e5f8a4-35c5-4c29-8c86-883650bd3f0e/resourceGroups/datalake2024Matheus04/providers/Microsoft.Storage/storageAccounts/contosolake2024matheus04",

    "name": "contosolake2024matheus04",

    "type": "Microsoft.Storage/storageAccounts",

    "location": "canadacentral",

    "tags": {},

    "properties": {

        "keyCreationTime": {

            "key1": "2024-09-19T00:05:17.8305902Z",

            "key2": "2024-09-19T00:05:17.8305902Z"

        },

        "allowCrossTenantReplication": false,

        "privateEndpointConnections": [],

        "minimumTlsVersion": "TLS1\_2",

        "allowBlobPublicAccess": false,

        "isHnsEnabled": true,

        "networkAcls": {

            "bypass": "AzureServices",

            "virtualNetworkRules": [],

            "ipRules": [],

            "defaultAction": "Allow"

        },

        "supportsHttpsTrafficOnly": true,

        "encryption": {

            "services": {

                "file": {

                    "keyType": "Account",

                    "enabled": true,

                    "lastEnabledTime": "2024-09-19T00:05:17.8462164Z"

                },

                "blob": {

                    "keyType": "Account",

                    "enabled": true,

                    "lastEnabledTime": "2024-09-19T00:05:17.8462164Z"

                }

            },

            "keySource": "Microsoft.Storage"

        },

        "accessTier": "Hot",

        "provisioningState": "Succeeded",

        "creationTime": "2024-09-19T00:05:17.7368324Z",

        "primaryEndpoints": {

            "dfs": "https://contosolake2024matheus04.dfs.core.windows.net/",

            "web": "https://contosolake2024matheus04.z9.web.core.windows.net/",

            "blob": "https://contosolake2024matheus04.blob.core.windows.net/",

            "queue": "https://contosolake2024matheus04.queue.core.windows.net/",

            "table": "https://contosolake2024matheus04.table.core.windows.net/",

            "file": "https://contosolake2024matheus04.file.core.windows.net/"

        },

        "primaryLocation": "canadacentral",

        "statusOfPrimary": "available",

        "secondaryLocation": "canadaeast",

        "statusOfSecondary": "available",

        "secondaryEndpoints": {

            "dfs": "https://contosolake2024matheus04-secondary.dfs.core.windows.net/",

            "web": "https://contosolake2024matheus04-secondary.z9.web.core.windows.net/",

            "blob": "https://contosolake2024matheus04-secondary.blob.core.windows.net/",

            "queue": "https://contosolake2024matheus04-secondary.queue.core.windows.net/",

            "table": "https://contosolake2024matheus04-secondary.table.core.windows.net/"

        }

    }

}

## On the left side click on the “Containers” tab under “Data storage”, take a screenshot and add it to your analysis report.

A computer screen with a white box

Description automatically generated

## Navigate back to the resource group and expand the “Synapse workspace”, take two screenshots showing all the configurations and add that to your analysis report.

A computer screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Exercise #3: Retrieve data using a synapse serverless SQL pool (35 mark)

## Take three screenshots and add them to your analysis report, the first showing the query, the second showing the results and the third showing the message.

A computer screen shot of a computer screen

Description automatically generated

A computer screen shot of a computer

Description automatically generated A computer screen shot of a computer

Description automatically generated

## In your workspace navigate to the monitor hub, click on the SQL requests, take a full screenshot and add it to your report.

A computer screen shot of a computer screen

Description automatically generated

**References**

Databricks. (2024). *What is Parquet?* Retrieved from <https://www.databricks.com/glossary/what-is-parquet>

Upsolver. (2022). *Parquet, ORC, and Avro: The file format fundamentals of big data*. Retrieved from <https://www.upsolver.com/blog/the-file-format-fundamentals-of-big-data>

Wikipedia. (2024). *Apache Parquet*. In *Wikipedia*. Retrieved from <https://en.wikipedia.org/wiki/Apache_Parquet>