**Which file format to choose**

**Avro / Parquet**

***Key Factors***

[*Apache*](https://nixondata.com/knowledge/kafka-fundamentals/apache-kafka-and-apache-flink-comparision/)[*Avro*](https://nixondata.com/knowledge/big-data-fundamentals/big-data-file-formats/)***and Apache Parque****t are both open-source file formats that are designed for storing and processing large datasets in a distributed environment.*

However, there are some key differences between the two formats:

1. [Data](https://nixondata.com/knowledge/apache-spark-for-experts/apache-spark-what-is-data-skew-how-to-identify-data-skew-from-data-skew-impact-fix/)**storage***: Avro is a data serialization system that is designed for efficient, language-independent data interchange***. It stores data in a binary format that includes both the data itself and a schema that describes the data. Parquet, on the other hand, is a columnar data storage format that is optimized for storing and processing large datasets*.****It stores data in a columnar format, which can be more efficient in terms of space and time compared to row-based storage formats.*
2. **Use cases**: *Avro is often used in*[*big data*](https://nixondata.com/knowledge/apache-spark-fundamentals/what-is-apache-spark/)*environments to store large datasets, as it is efficient and easy to use*. **It is also often used as the data serialization format for message-based systems, such as Apache**[**Kafka**](https://nixondata.com/knowledge/kafka-fundamentals/what-is-apache-kafka/).**Parquet, on the other hand, is primarily used for storing and processing large datasets,** and ***is often used in conjunction with tools like Apache***[***Spark***](https://nixondata.com/knowledge/apache-spark-fundamentals/apache-spark-jobs-stage-task/)***and Apache Impala for big data analysis****.*
3. **Performance**: ***Both Avro and Parquet are optimized for storing and processing large datasets****, and they can both offer good performance in terms of space and time efficiency*. However,**Parquet may offer slightly better performance due to its columnar storage format, which can be more efficient for certain types of queries.**

**Overall, Avro and Parquet are both useful file formats for storing and processing large datasets, and the specific format that is best suited for a given application will depend on the specific needs and requirements of the organization.**

***CSV DATA 🡪 comma file***

**1,John,Jon@mail.com**

**2,Rom,Rom@gmail.com**

**3,Kyara,Kyara104@gmail.com**

***Structured Format 🡪 RDBMS + Typical rows and columns***

|  |  |  |
| --- | --- | --- |
| **Roll No** | **Name** | **Email** |
| **1** | **John** | [john@mail.com](mailto:john@mail.com) |
| **2** | **Rom** | [Rom@gmail.com](mailto:Rom@gmail.com) |
| **3** | **Kyara** | **Kyara104@gmail.com** |

***Avro 🡪 Binary Encoding + JSON encoding + Row based***

**Roll No Name Email**

**1 John** [john@mail.com](mailto:john@mail.com)

**2 Rom** [Rom@gmail.com](mailto:Rom@gmail.com)

**3 Kyara** [Kyara104@gmail.com](mailto:Kyara104@gmail.com)

***Parquet 🡪 Columnar + Compression + Optimized storage***

Roll No 1 2 3

Name John Rom Kyara

Email [john@mail.com](mailto:john@mail.com) [Rom@gmail.com](mailto:Rom@gmail.com) Kyara104@gmail.com