Good questions and explanation

<http://data-flair.training/forums/topic/list-the-advantage-of-parquet-file-in-apache-spark>

<http://data-flair.training/forums/topic/explain-transformation-and-action-in-rdd-in-apache-spark>

Roll of Driver in Spark

* Driver program is responsible for launching various parallel operations on the cluster.
* **Driver program contains application's *main()* function.**
* **It is the process which is running the user code which in turn create the SparkContext object,**[**create RDDs**](http://data-flair.training/blogs/how-to-create-rdds-in-apache-spark/)**and performs**[**transformation and action operation on RDD**](http://data-flair.training/blogs/rdd-transformations-actions-apis-apache-spark/)**.**
* Driver program access [**Apache Spark**](http://data-flair.training/blogs/apache-spark-introduction-spark-comprehensive-tutorial/)through a [**SparkContext**](http://data-flair.training/blogs/sparkcontext-in-apache-spark-tutorial/) object which represents a connection to computing cluster (From Spark 2.0 onwards we can access SparkContext object through SparkSession).
* **Driver program is responsible for converting user program into the unit of physical execution called task.**
* It also defines distributed datasets on the cluster and we can apply different operations on Dataset (transformation and action).
* **Spark program creates a logical plan called**[**Directed Acyclic graph**](http://data-flair.training/blogs/directed-acyclic-graph-dag-in-apache-spark/)**which is converted to physical execution plan by the driver when driver program runs.**

What are the benefits of using parquet file-format in Apache Spark?

Parquet is a columnar format supported by many data processing systems. The benifits of having a columnar storage are -

1- Columnar storage limits IO operations.

2- Columnar storage can fetch specific columns that you need to access.

3-Columnar storage consumes less space.

4- Columnar storage gives better-summarized data and follows type-specific encoding.

Parquet is an open source file format for [**Hadoop**](http://data-flair.training/blogs/hadoop-introduction-tutorial-quick-guide/). Parquet stores nested data structures in a flat columnar format compared to a traditional approach where data is stored in row-oriented approach, parquet is more efficient in terms of storage and performance.

There are several advantages to columnar formats:

1)Organizing by column allows for better compression, as data is more homogeneous. The space savings are very noticeable at the scale of a Hadoop cluster.  
2)I/O will be reduced as we can efficiently scan only a subset of the columns while reading the data. Better compression also reduces the bandwidth required to read the input.  
3)As we store data of the same type in each column, we can use encoding better suited to the modern processors’ pipeline by making instruction branching more predictable.