**WHAT SQOOP DOES**

YARN coordinates data ingest from Apache Sqoop and other services that deliver data into the Enterprise Hadoop cluster.

* Apache Sqoop does the following to integrate bulk data movement between Hadoop and structured datastores:

| **Function** | **Benefit** |
| --- | --- |
| Import sequential datasets from mainframe | Satisfies the growing need to move data from mainframe to HDFS​ |
| Import direct to ORC (**Optimized Row Columnar**) Files | ​Improved compression and light-weight indexing for improved query performance |
| Data imports | Moves certain data from external stores and EDWs into Hadoop to optimize cost-effectiveness  of combined data storage and processing |
| Parallel data transfer | For faster performance and optimal system utilization |
| Copies data faster | From external systems into Hadoop |
| Efficient data analysis | Improves efficiency of data analysis by combining structured data with unstructured data in a  schema-on-read data lake |
| Load balancing | Mitigates(making something easier) excessive storage and processing loads to other systems |

**ORC Format:**

* RC (Record Columnar) File, the previous Hadoop Big Data storage format on Hive, is being challenged by the smart ORC (Optimized Row Columnar) format.
* RC (Record Columnar) File already improves the storage requirements significantly. ORC files are even better at storing the same information without compression.

**Working of Sqoop:**

* Sqoop provides a pluggable mechanism for optimal connectivity to external systems.
* Apache Sqoop efficiently transfers bulk data between Apache Hadoop and structured data stores such as relational databases.
* Sqoop can also be used to extract data from Hadoop and export it into external structured datastores.
* Sqoop works with relational databases such as Teradata, Netezza, Oracle, MySQL, Postgres, and HSQLDB.

