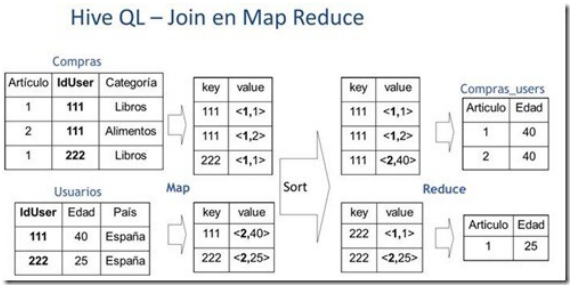
1.5 Pig y Hive

Hive proporciona un mecanismo para abstraer la estructura de los datos y consultar los datos mediamente un lenguaje parecido a SQL, llamado HiveQL. Luego estás sentencias se descomponen en tareas MapReduce y se ejecutan.



En nuestro caso concreto contamos con dos ficheros, por un lado uno que contiene el título de la película con su id y sus géneros y por otro uno que contiene los ratings dados por cada usuario a cada película. Hive sería muy útil ya que nos permitiría hacer un join de ambos ficheros de una forma muy sencilla y obtener como resultado el título de la película con su rating, que es justo lo que queremos.

Pig ofrece una plataforma con un lenguaje de alto nivel que permite analizar grandes volúmenes de datos. Como característica importante está el paralelismo de datos, por lo que se pueden analizar todos los datos a la vez.

En nuestro caso concreto tenemos archivos bastante grandes. Además, al tratarse de archivos en texto plano y relacionados con ratings de usuarios, cada archivo cuenta con muchas filas, aunque su tamaño en bytes no sea muy grande. Por tanto poder usar Pig para paralelizar el análisis de datos, haría que éste se produjera de una forma mucho más rápida.

Como conclusión, pig y hive nos ayudarían a unir los datos y a tratarlos mucho más rápido.