HBase desde MapReduce/Hadoop

- Las clases y utilitarios de HBase en el paquete org.apache.hadoop.hbase.mapreduce facilitan el uso de HBase como fuente o destino de datos en tareas MapReduce.
- La clase TableInputFormat crea splits en límites de región para que los maps trabajen con datos en una única región.
- La clase TableOutputFormat escribe los resultados del reduce en Hbase.
- Ejemplo: clase SimpleRowCounter (version simple de RowCounter disponible en el paquete HBase de mapreduce) para contar filas de una table usando TableInputFormat.

```
public class SimpleRowCounter extends Configured implements Tool {
    static class RowCounterMapper extends TableMapper<ImmutableBytesWritable, Result> {
        public static enum Counters { ROWS }
        @Override
        public void map(ImmutableBytesWritable row, Result value, Context context) {
            context.getCounter(Counters.ROWS).increment(1);
        }
}
```

- RowCounterMapper es una subclase de la clase abstracta TableMapper de HBase, una especialización de org.apache.hadoop.mapreduce.Mapper que setea los tipos de entrada pasados por TableInputFormat.
- Las claves de entrada son objetos ImmutableBytesWritable (claves de las filas) y los valores son objetos Result (resultados de un scan de las filas).
- El job no emite salida de map, solo incrementa un Contador (Counters.ROWS) por cada fila.

```
@Override
public int run(String[] args) throws Exception {
    if (args.length != 1) {
        System.err.println("Usage: SimpleRowCounter <tablename>");
        return -1;
    }
 String tableName = args[0];
    Scan scan = new Scan();
    scan.setFilter(new FirstKeyOnlyFilter());
    Job job = new Job(getConf(), getClass().getSimpleName());
    job.setJarByClass(getClass());
    TableMapReduceUtil.initTableMapperJob(tableName, scan,
    RowCounterMapper.class, ImmutableBytesWritable.class, Result.class, job);
    job.setNumReduceTasks(0);
    job.setOutputFormatClass(NullOutputFormat.class);
    return job.waitForCompletion(true) ? 0 : 1;
}
public static void main(String[] args) throws Exception {
    int exitCode = ToolRunner.run(HBaseConfiguration.create(), new SimpleRowCounter(), args);
    System.exit(exitCode);
}
}
```

- En el método run() se crea un objeto scan que configure el trabajo invocando al método TableMapReduceUtil.initTableMapJob() que setea la clase map y el input format a TableInputFormat.
- Se setea un filtro (instancia de FirstKeyOnlyFilter) en el scan para que se cargue el objeto Result solamente con el primer campo en cada fila (optimización útil porque el mapper no toma en cuenta los valores).
- El número de filas de una tabla puede calcularse en el shell de Hbase con el commando '*tablename*', pero no es distribuido. La version Mapreduce puede ser más eficiente para tablas de gran dimension.