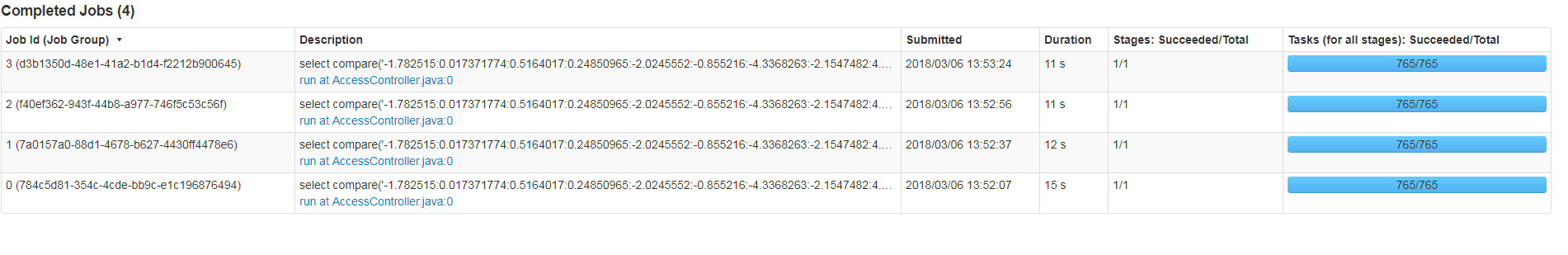
数据量为10030068： 6个节点，256G\*5，48 核\*5， + 125G，12核，spark.parquet.block 大小为32M

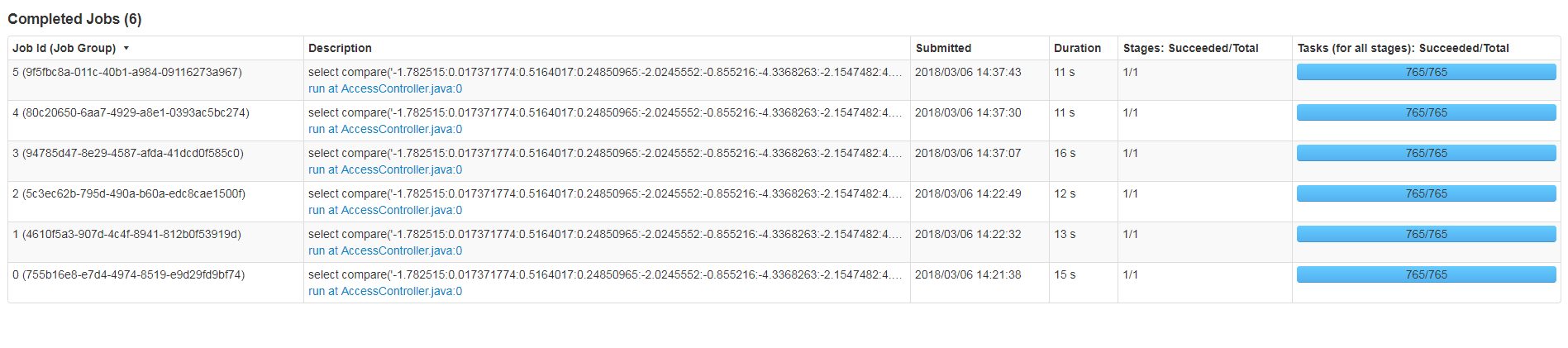
约为1000万： 没有加对比的时候，时间为12到15秒，

SQL 语句：

select compare(“\*\*\*\*\*\*”, feature) as related from person\_table\_v1 where date='2018-02-32' limit 100;



加上比对的时间：



总结：从上面两个对比结果看，加上对比和不加上对比，时间上基本一致。（即是对比基本不花费时间）

时间主要耗费在数据的读取和对结果的排序上面

当Block 减少到16M每块，即生成parquet 文件的时候，设置spark.parquet.block 为16 M，单位Byte，的时候，情况比较好。

