# Purpose and example

These are the rows I want to insert into my Tables A and B.

| Rows for table A: | Rows for table B: |
|---|---|
| ID | ID |
| 0 | 0 |
| 0 | 0 |
| 1 | 0 |
| 2 | 1 |
| 2 | 1 |
| 3 | 2 |
| | 2 |
| | 2 |
| | 3 |
| | 3 |
| | 3 |

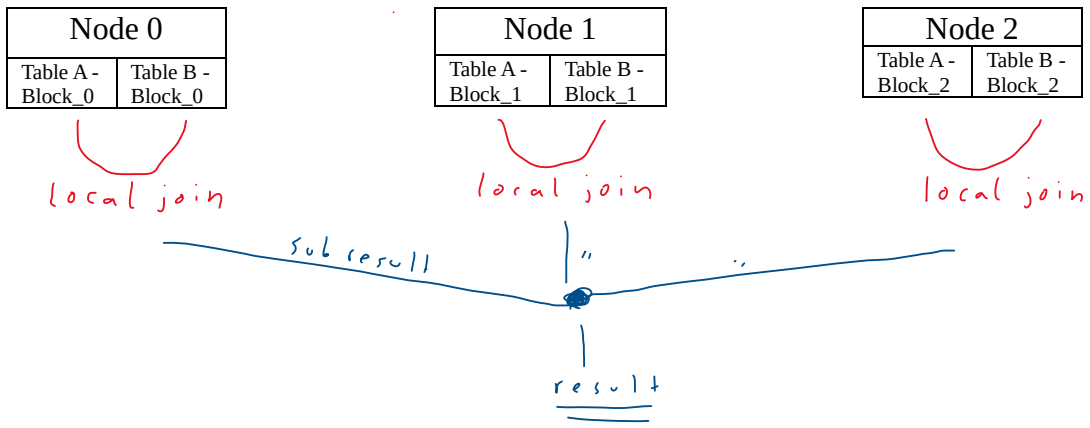
Let's assume, we have 3 data nodes, so the modulo function is : ID % 3. The creation of the parquet files is done manually (outside of Impala) and results in following parquet files:

| Table A - Parquet_0 | Table A - Parquet_1 | Table A - Parquet_2 |
|---|---|---|
| 0 | 1 | 2 |
| 0 | | 2 |
| 3 | | |

| Table B - Parquet_0 | Table B - Parquet_1 | Table B - Parquet_2 |
|---|---|---|
| 0 | 1 | 2 |
| 0 | 1 | 2 |
| 0 | | 2 |
| 3 | | |
| 3 | | |
| 3 | | |

The idea behind this data distribution is to distribute blocks and its data equally across the cluster.

Afterwards, the parquet files are transferred to the corresponding tables and blocks are created + distributed by HDFS. Since the distribution is done randomly, I move corresponding blocks to the same node:




This is the starting position for this project. Red and blue describe our purpose.