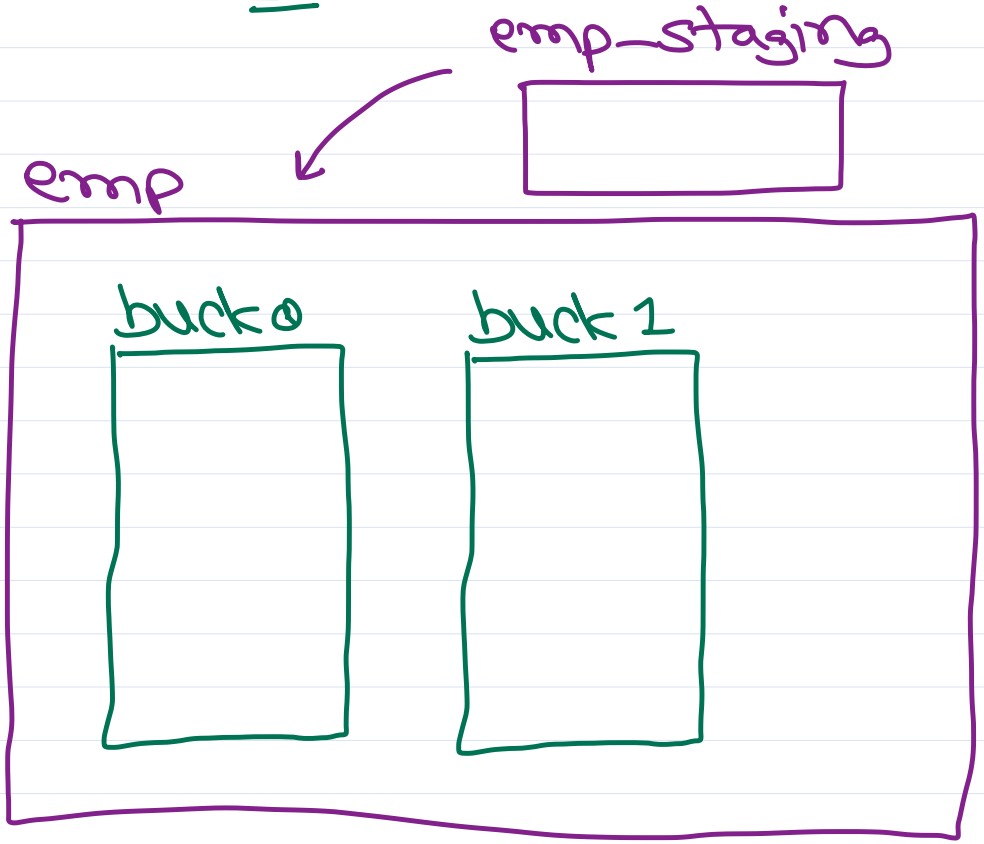
Big Data Technologies

Trainer: Mr. Nilesh Ghule.

Bucketing

emp table → clustered by (ename) into 2 buckets.



ename

FORD → hashCode() % 2 → 0 → bucket0
                      → 1 → bucket1

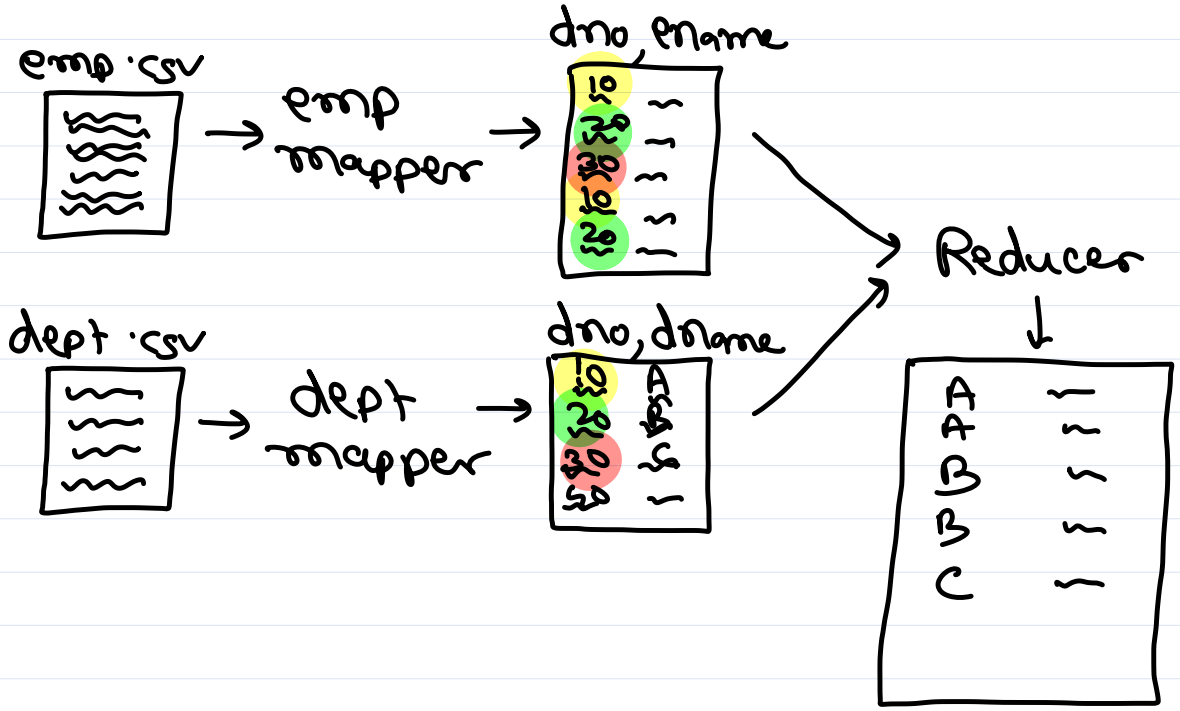
JAMES → hashCode() % 2 → 0 → bucket0
                       → 1 → bucket1

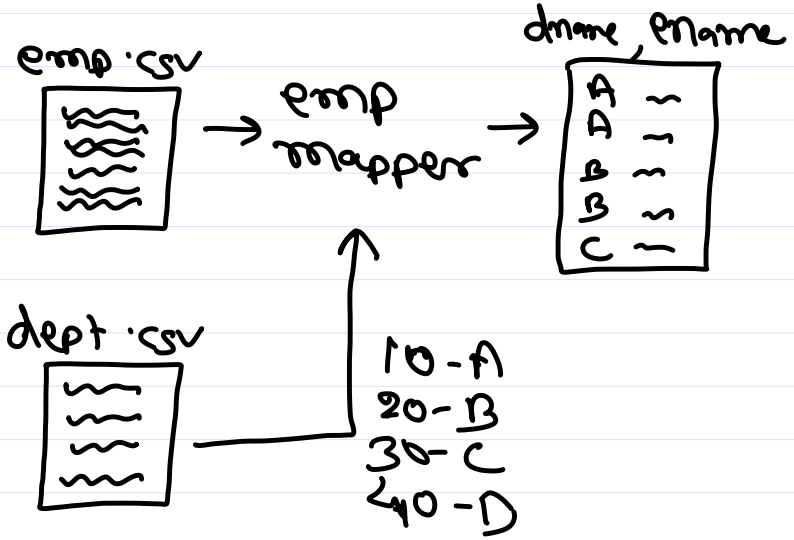
KING → hashCode() % 2 → 0 → bucket0
                      → 1 → bucket1

When DML ops are performed on bucketed tables, num of reducers are set to num of buckets.

In Hive 3.x, when DML ops are perform on non-bucketed table, one bucket is considered for the table (one reducer is used).

Reduce side joins vs Map side joins





Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>