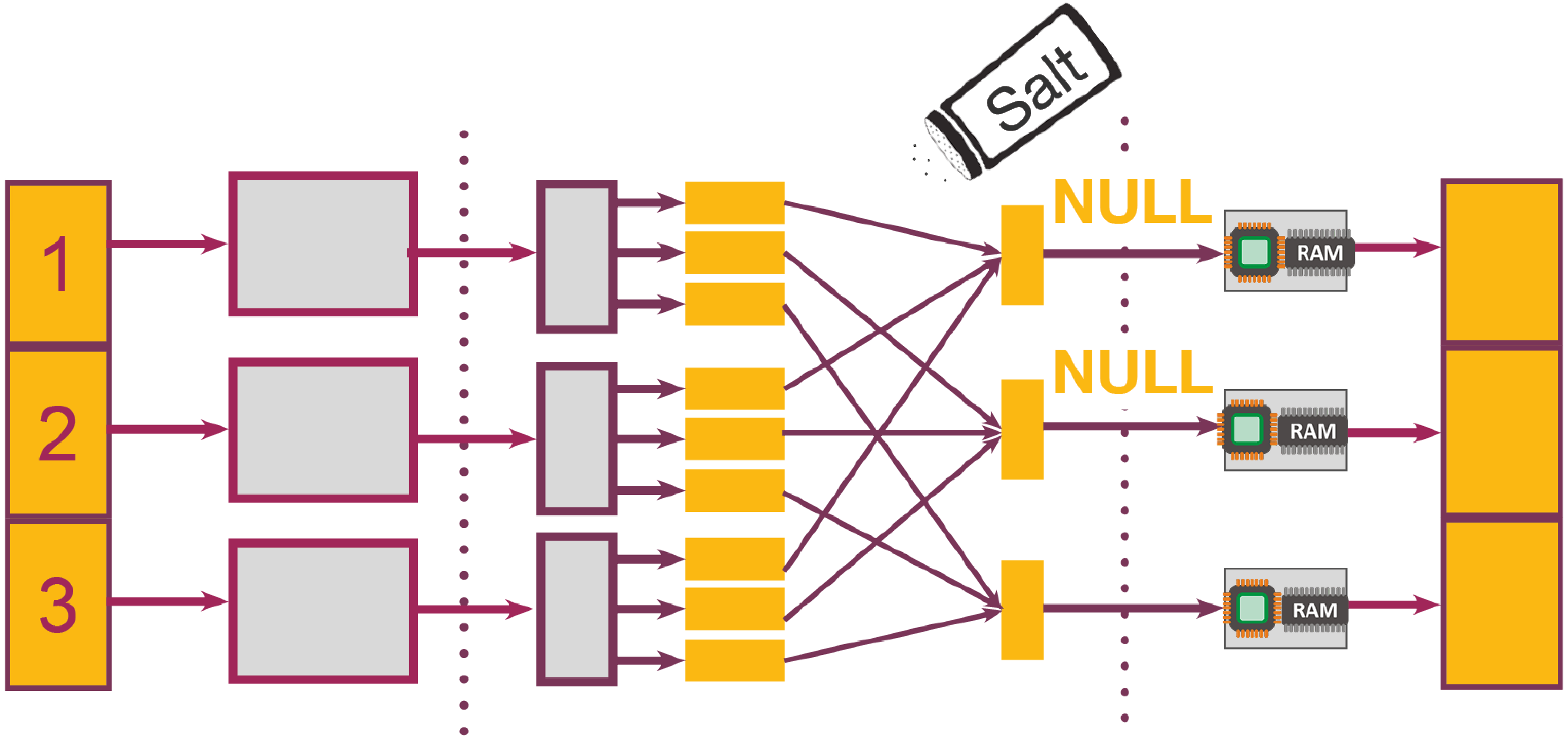
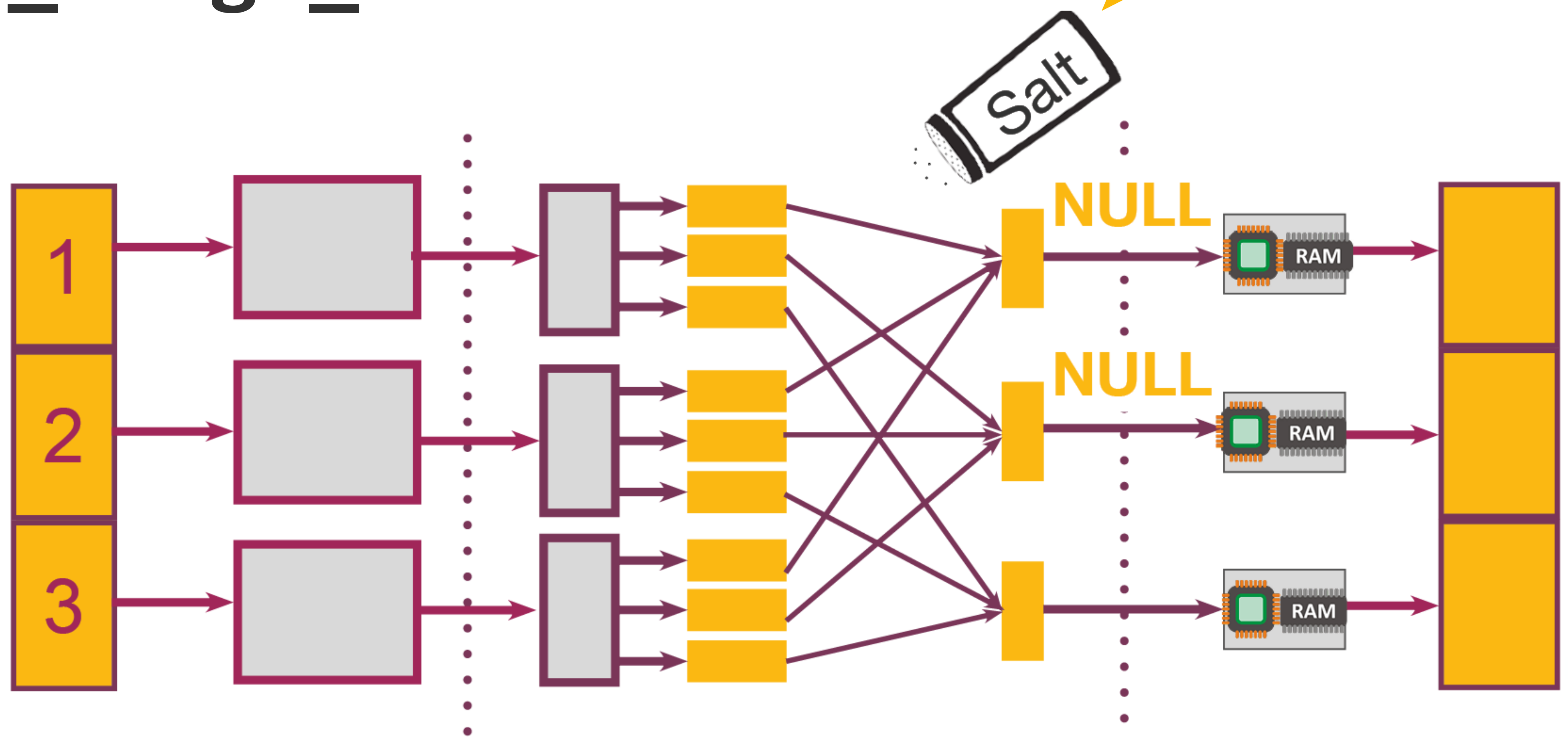
Hive Optimisation

Data Skew

Data Skew



```
SET mapred.reduce.tasks = 128;
SELECT TRANSFORM(user_id, …)
    USING "./count.sh" AS user_id, some_stat
FROM (
    SELECT *
    FROM access_log
    DISTRIBUTE BY (
      hash(user_id)
      + IF(user_id IS NULL, my_salt_UDF(), 0)
      )
) table_stage_0
```



```
…
FROM (
    SELECT *
    FROM access_log
    DISTRIBUTE BY (
      hash(user_id)
      + IF(user_id IS NULL, my_salt_UDF(), 0)
      )
) table_stage_0
```

example

```
SELECT CONCAT("none-", SUBSTR(
        reflect("java.util.UUID", "randomUUID"), 0, 8))
FROM some_table …;
```

```
…
none-0a1a15ac
none-29e78368
none-3daa8e36
…
```

```
CREATE TABLE skewed_access_log (
    ip STRING,
    …
    user_id STRING,
    …
)
SKEWED BY (user_id) ON ("unknown", "1")
…
```

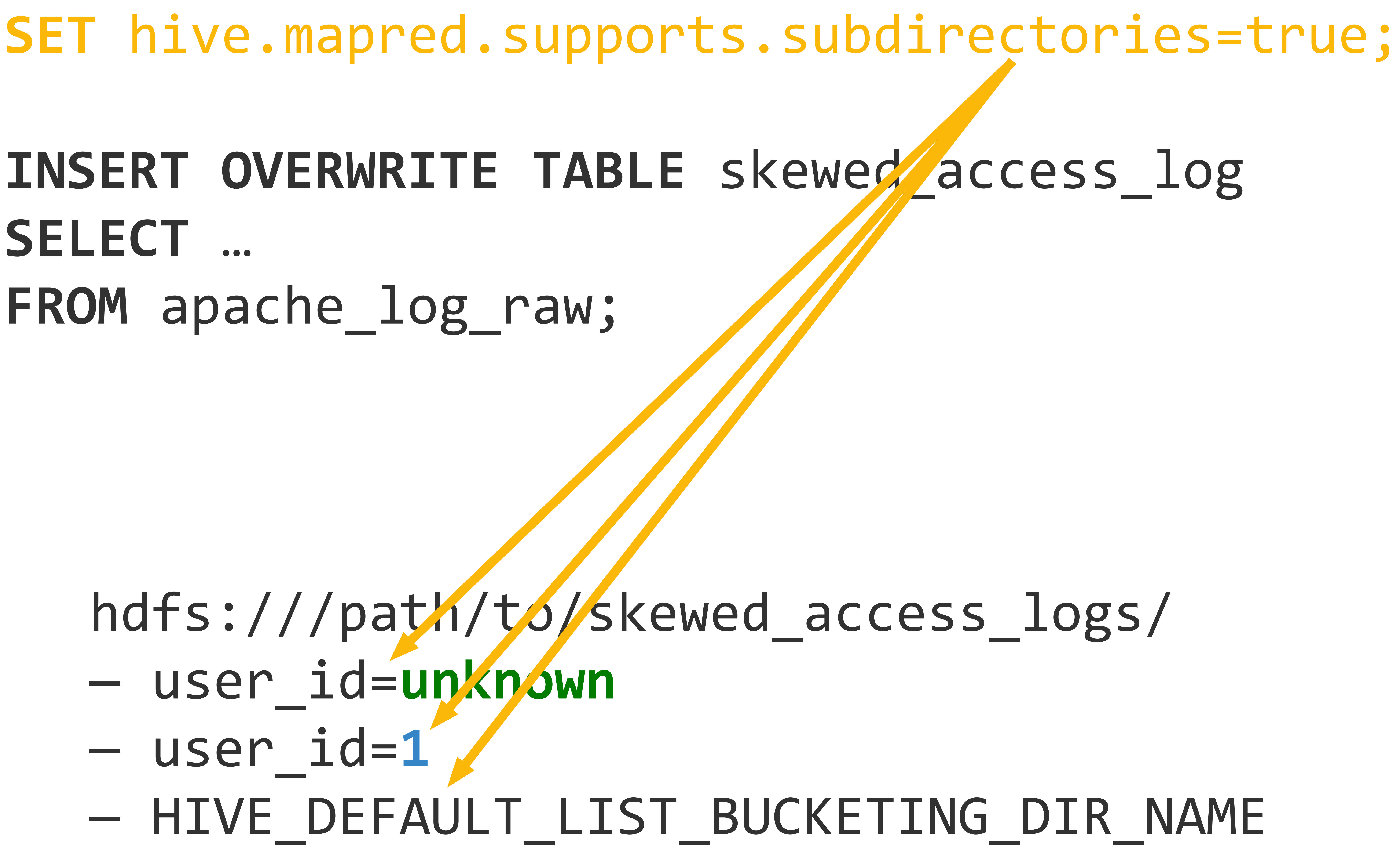
```
CREATE TABLE skewed_access_log (
    ip STRING,
    …
    ~~request_date STRING~~,
    user_id STRING,
    …
)
PARTITIONED BY (request_date STRING)
SKEWED BY (user_id) ON ("unknown", "1")
…
```

1. List Bucketing

```
CREATE TABLE skewed_access_log (
    ip STRING,
    …
    user_id STRING,
    …
)
SKEWED BY (user_id) ON ("unknown", "1")
    STORED AS DIRECTORIES
…

hdfs:///path/to/skewed_access_logs/
– user_id=unknown
– user_id=1
– HIVE_DEFAULT_LIST_BUCKETING_DIR_NAME
```

1. List Bucketing



```
SET hive.mapred.supports.subdirectories=true;

INSERT OVERWRITE TABLE skewed_access_log
SELECT …
FROM apache_log_raw;
```

```
hdfs:///path/to/skewed_access_logs/
– user_id=unknown
– user_id=1
– HIVE_DEFAULT_LIST_BUCKETING_DIR_NAME
```

2. Skewed Table

```
CREATE TABLE skewed_access_log (
    ip STRING,
    …
    user_id STRING,
    …
)
SKEWED BY (user_id) ON ("unknown", "1")
    ~~STORED AS DIRECTORIES~~
…
```

```
hdfs:///path/to/skewed_access_logs/
~~- user_id=unknown~~
~~- user_id=1~~
~~- HIVE_DEFAULT_LIST_BUCKETING_DIR_NAME~~
```

Summary

Summary

- You can **create** and **insert** data into **skewed** Hive tables

Summary

- You can **create** and **insert** data into **skewed** Hive tables
- You can **explain** usage scenarios of “**List Bucketing**” and “**Skewed Table**”

Summary

- You can **create** and **insert** data into **skewed** Hive tables
- You can **explain** usage scenarios of “**List Bucketing**” and “**Skewed Table**”

see: https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-SkewedTables
see: https://cwiki.apache.org/confluence/display/Hive/ListBucketing
see: https://cwiki.apache.org/confluence/display/Hive/Skewed+Join+Optimization