Hive at Last.fm

March 2010

What is Last.fm?

A music community **website**, powered by **scrobbling** that provides personalised **radio**.

We aggregate scrobbles. A single scrobble is the smallest unit of music attention data.

1 scrobble = (track, artist, timestamp).

In numbers

- 40 million users visit the site every month
- 39 billion scrobbles (600 per second)
- 400k personalised radio stations per day

enter hadoop...

last.fm

Hadoop cluster

- 44 nodes
- 8 cores per node
- 16 gig ram per node
- 4x 1TB 7200rpm disks per node

last.fm

Hadoop what is it good for?

- Charts
- Reporting
- Corrections
- Site stats / metrics
- Neighbours
- Recommendations

last.fm

But wait, can you tell us about <stuff/>?

- How many?
- When?
- Where?
- Who?
- Why? Why not?

last.fm

Ad hoc questions

- We get them all the time.
- Questions are good things, but answers take up time.
- We would typically write programs once, run once.

enter Hive...

last.fm

What is Hive?

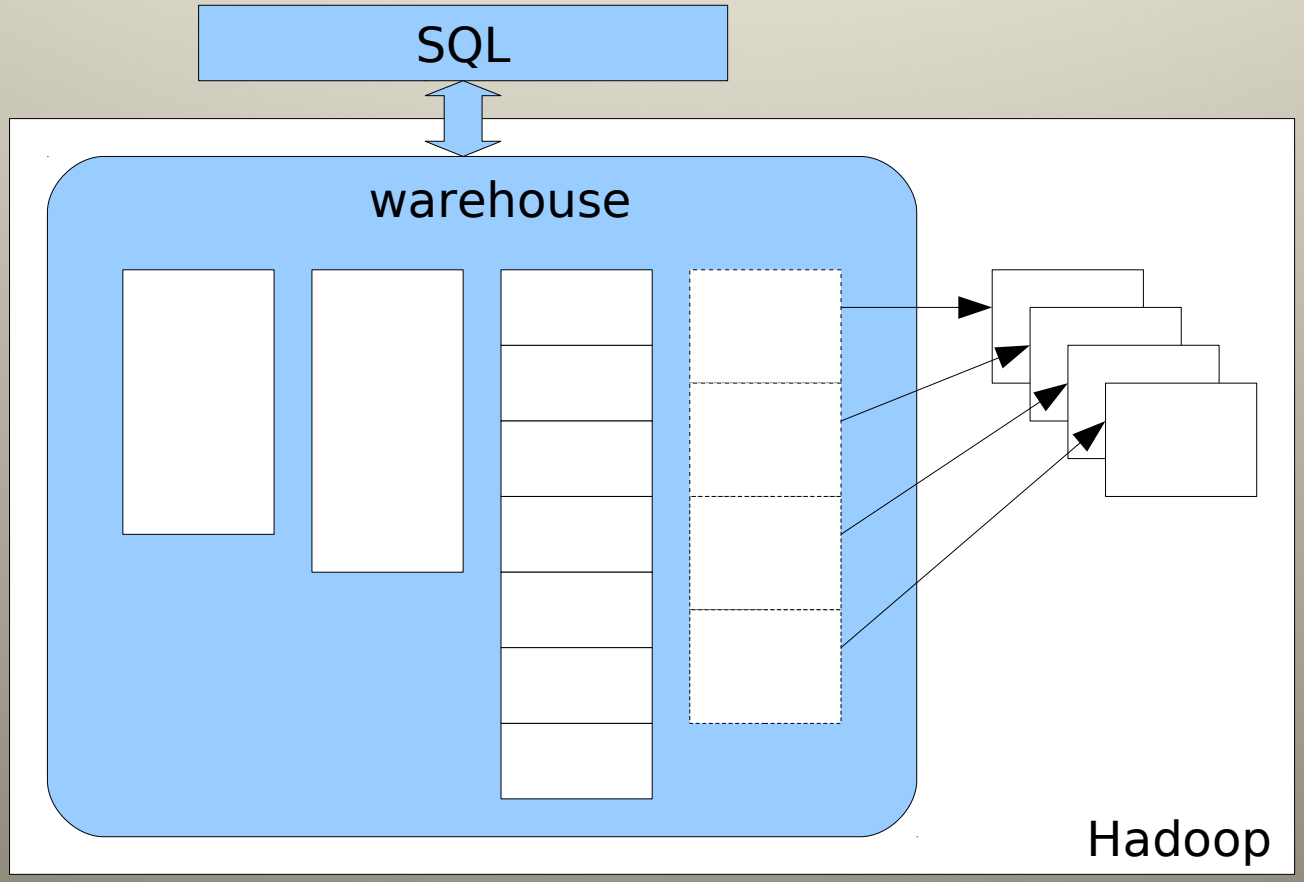
"Hive is a data warehouse infrastructure built on top of Hadoop"

You get an SQL-like language for queries.

Start queries from a shell, file, jdbc, thrift.

last.fm

Hive:



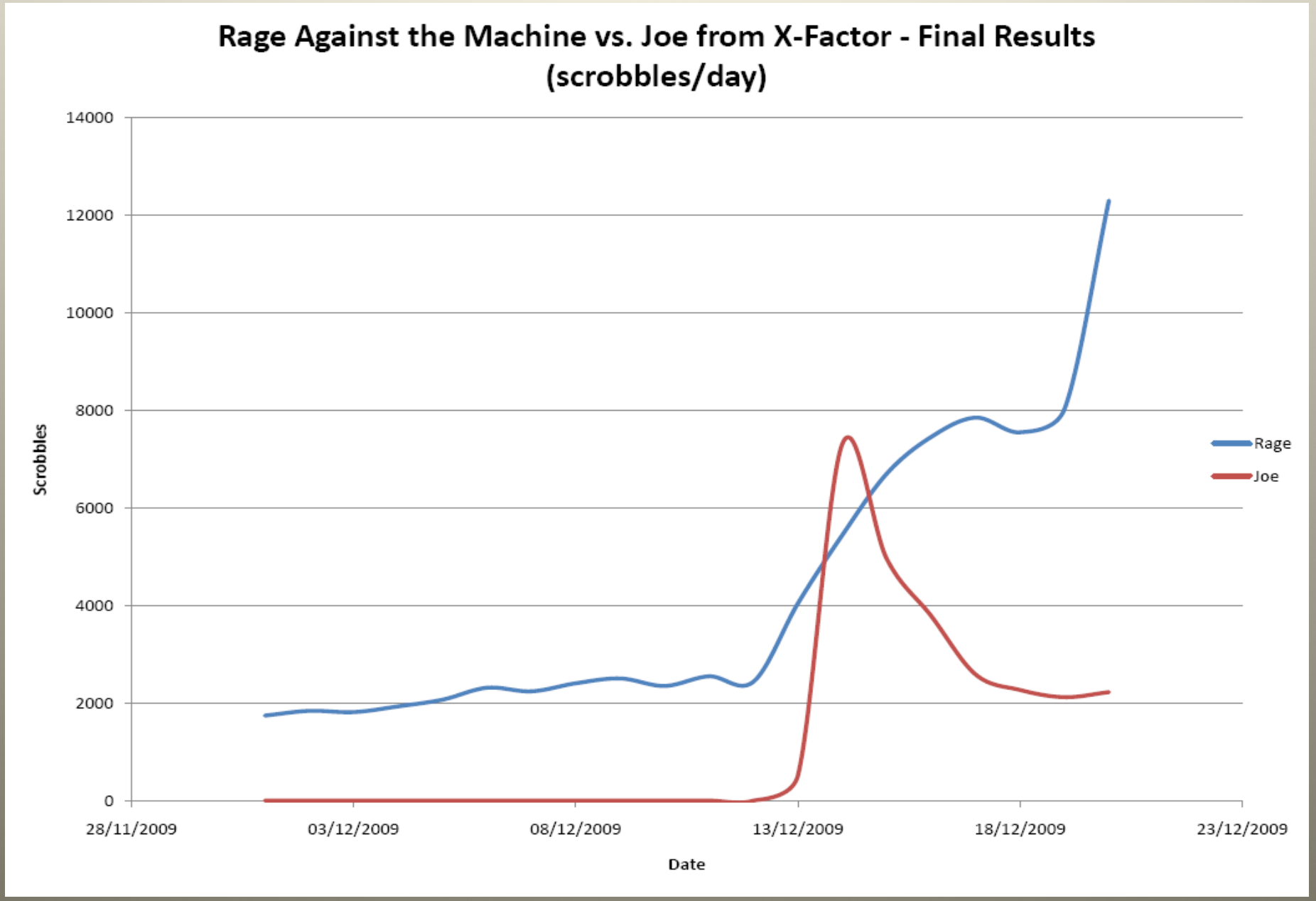
last.fm

Why we chose Hive?

- SQL familiarity suits non data engineers.
- It integrates well with existing data sets.
- It worked.

Johan set it up..

eg: http://www.flickr.com/photos/lozzd/4203345000/

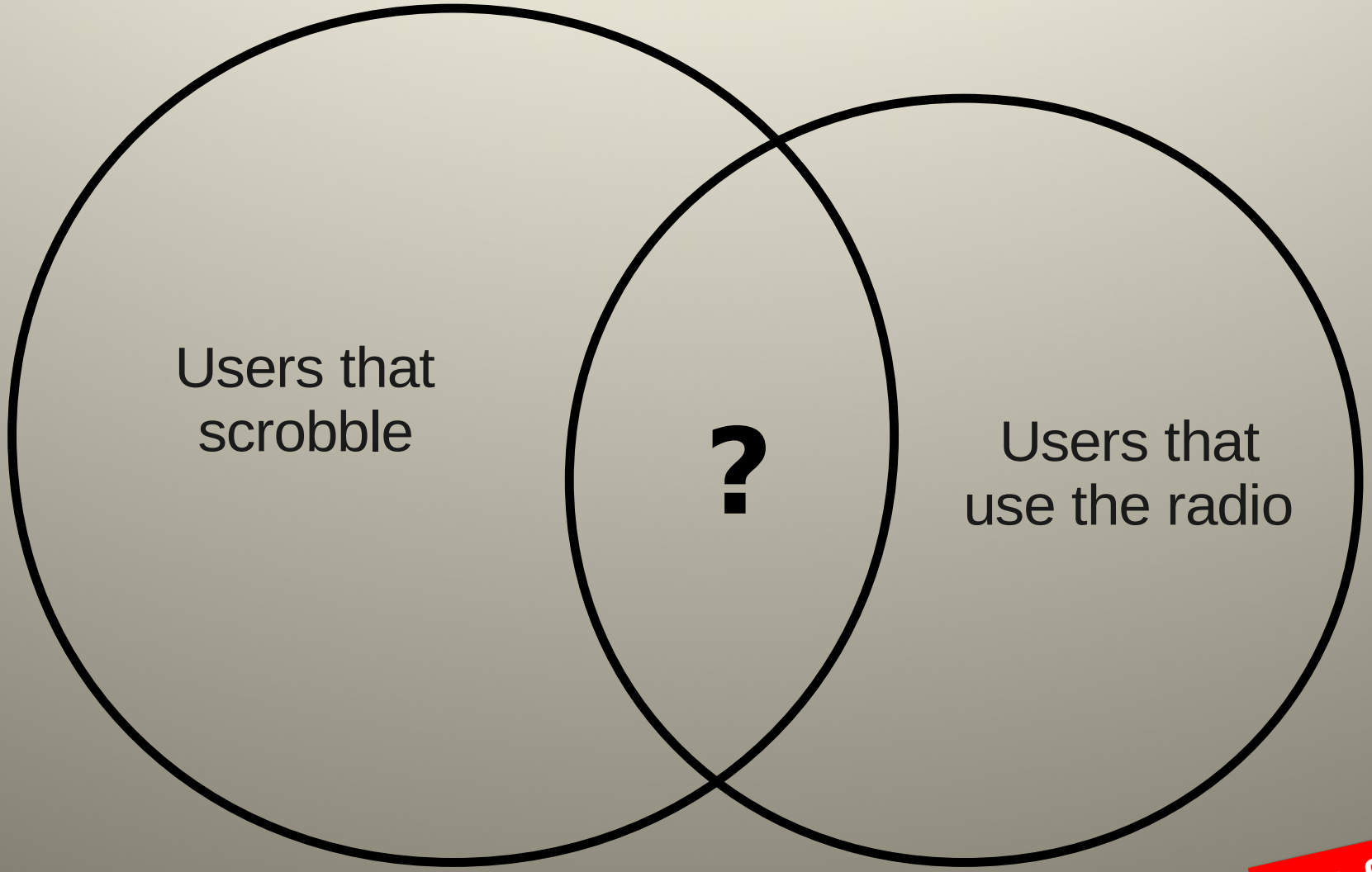


Example:

```
SELECT artistid, insertdate, count(1)
FROM scrobbles
WHERE (trackid = 10019 OR trackid = 368575614)
  AND insertdate >= '2009-12-01'
  AND insertdate <= '2009-12-31'
GROUP BY artistid, insertdate
ORDER BY artistid, insertdate;
```

last.fm

Example:



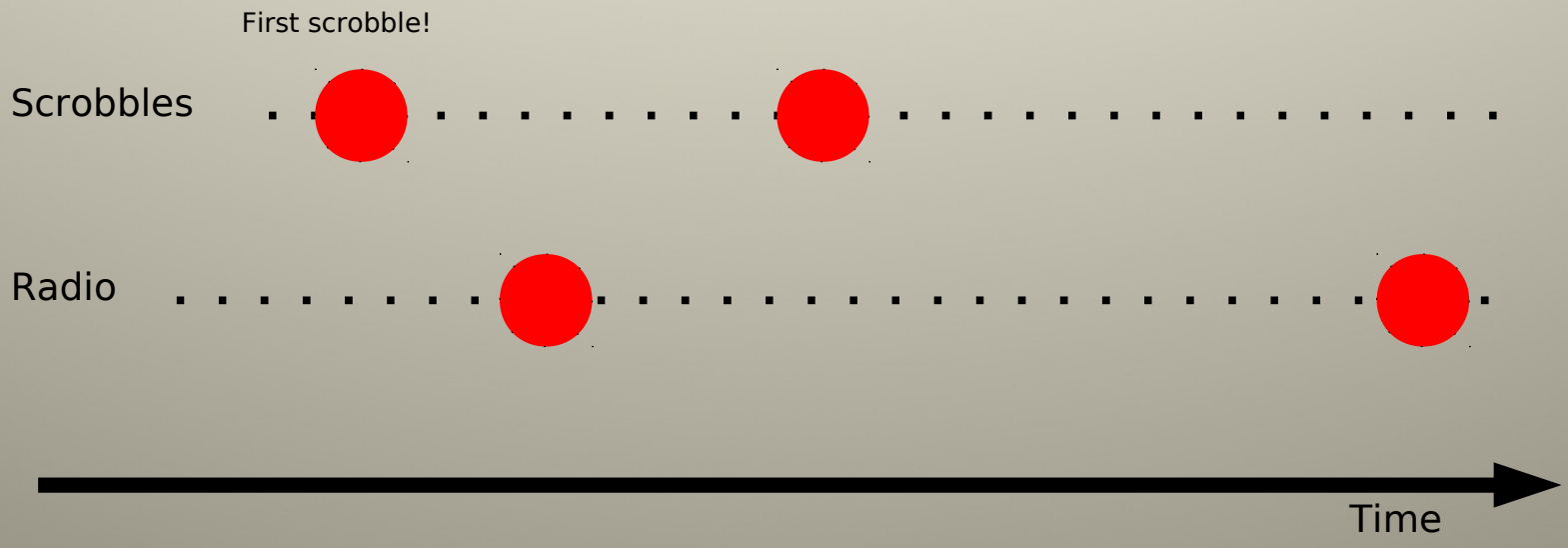
Example:

```
SELECT count(1) FROM scrobbles GROUP BY userid;

SELECT count(1) FROM radiologs GROUP BY userid;

SELECT count(1) FROM
  radiologs r JOIN scrobbles s
  ON r.userid = s.userid
GROUP BY r.userid;
```

last.fm

Example:

Consider a user's scrobbles and radio listens for just one track



Example:

```
SELECT r.userid, r.trackid, count(1)
FROM
 (
  SELECT userid, trackid, min(unixtime) as unixtime
  FROM scrobbles GROUP BY userid, trackid
 ) s
 JOIN
 radiologs r
 ON r.userid = s.userid AND r.trackid = r.trackid
WHERE s.unixtime < r.unixtime
GROUP BY r.userid, r.trackid
```

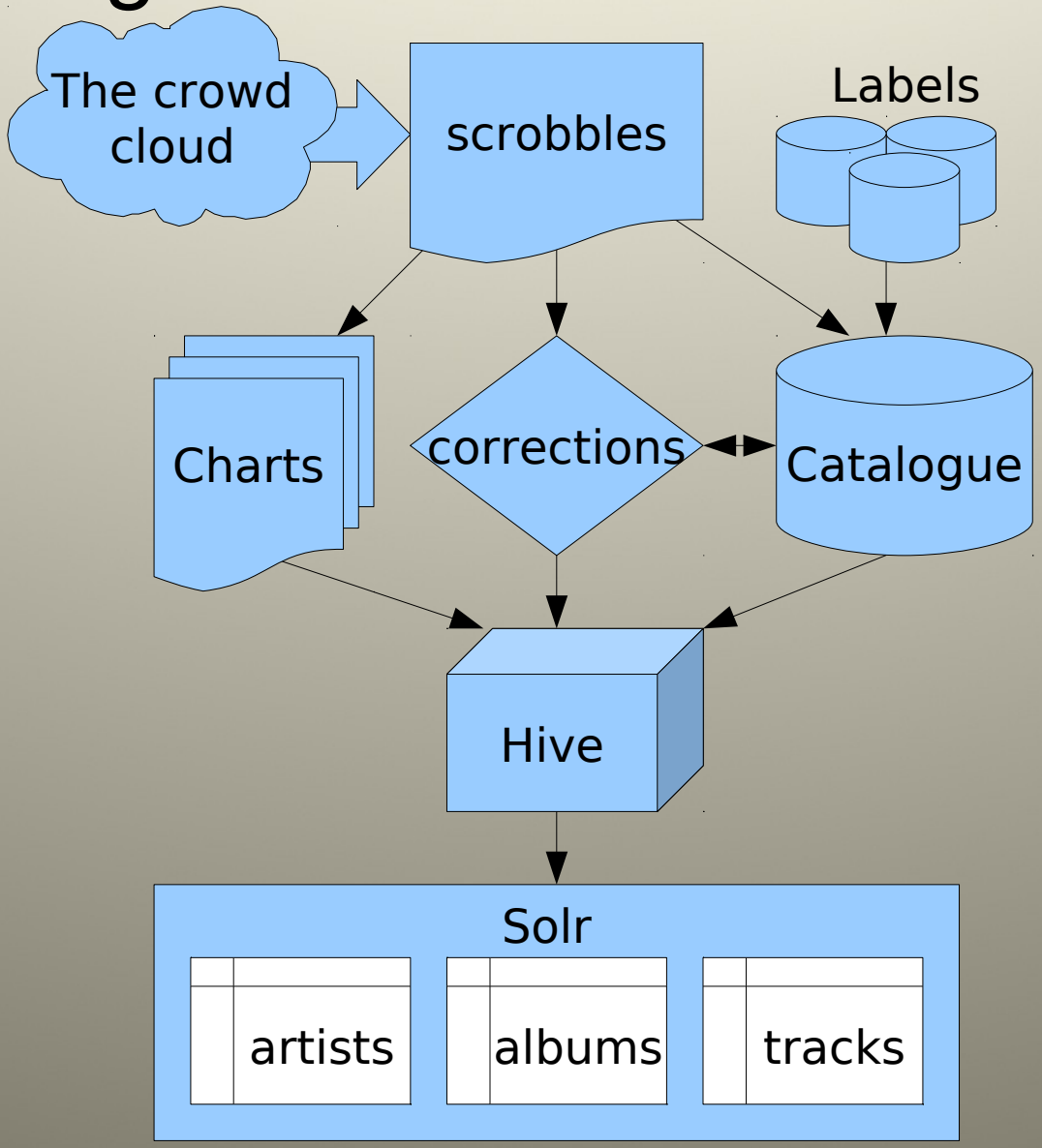
last.fm

Other nice things about hive

- Joins are really really easy (most of the time).

last.fm

Preparing a search index



last.fm

Not so great

- No recordio.
- Really huge joins can cause out of memory exceptions.

last.fm