Use Hadoop, One Week, No Coding

About Me

Noam Hasson
Team Leader for Big Data

- 12 years experience in web development
- Hadoop enthusiast since 2011
- noam.hasson@kenshoo.com

What is Kenshoo?

Closed-loop Targeting. Universal Integration. Dynamic Attribution.

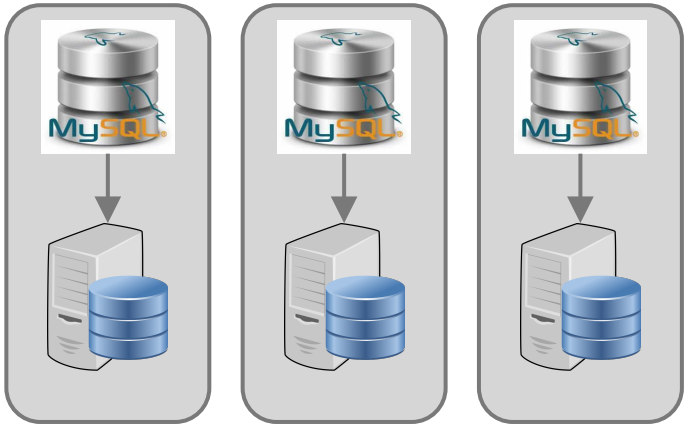
Infinite Optimization™.

Agenda

- Benefits of Hadoop:
  - Solve big data challenges
  - See actual results in less than a week
  - Use current RDMBS infrastructure
  - No coding experience necessary

- See how with an actual case study

Kenshoo Data Infrastructure

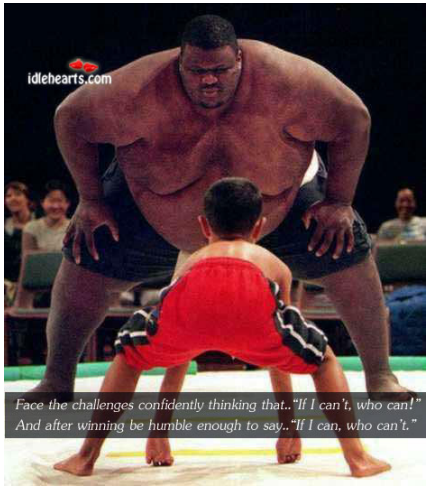


...X 250

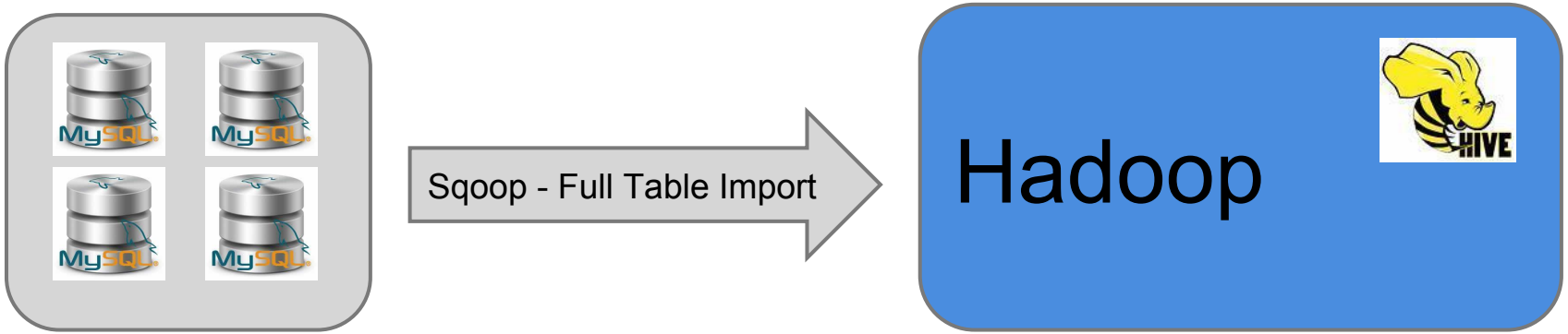
- 250 MySQL & application server pairs, each pair dedicated to a single client
- Identical schema on all MySQL servers
- Different table size & capacities on each MySQL server
- Total size of more than 100 TB

The Challenge

- Avoid load-intensive queries on production clusters
- Reduce run-time for data analysis queries
- Run cross-server queries for clients deployed on more than one DB server
- Compare statistical information across several DB servers



The Solution



- Use Sqoop to migrate tables to Hive
- Running SQL queries on Hive
- That's it!

Example Code

- **Oneline Import**

```
sqoop import --hive-import --connect jdbc:mysql://ServerAddress/
Database -m 10 --table DBTable --hive-database DBTable --username
dbUsername --password dbPassword --hive-overwrite -z
```

- **Query**

```
hive
use myDatabase
Select * from myTable where field = 'value'
```

Performance

- **Single-node Hardware**
  - 12 Hard drives each 4TB
  - 12 cores, 2 x CPU Sockets
  - 32GB memory

- **Performance**
  - Import table of 300M rows took 3 hours
  - Select count on 5.5 Billion rows took 90 minutes
  - Group By on 5.5 Billion rows, with 1.1 Billion rows took 18 hours

What We've Learned

- Use **partitions**
- Import directly to **compressed** files
- **Compare** the row count in the source and destination tables
- Import **only the columns you need** to query
- Use a **full table import** for easy & quick results
- Refine the number of map tasks used for import
- Adopt Hive for your Map-Reduce jobs
- Increase query speed by **avoiding "order by"**