Introduction to Hive

For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com

Agenda

Limitation of MR
Hive Overview
Data Vs Metadata
Pros of Hive
Cons of Hive
Hive applications
Hive usages @ Facebook

limitation of MR

Map-Reduce code is not reusable.

Need to write a new MR job for every new requirement.

Error prone

For complex jobs:

Multiple stage of Map/Reduce functions

Hive overview

A data warehouse for Hadoop.

It makes structured data looks like tables.

SQL based query can be used on these tables.

Queries are automatically converted into Map-Reduce jobs which are executed on Hadoop.

Complex queries may be converted into more than 1 Map-Reduce jobs.

Data Vs Metadata

Hive table's data is stored on HDFS.

Big tables can be stored on multiple machine distributively.

All the metadata is stored in **Derby** by default

Derby doesn’t allow more than 1 concurrent connections.

Any database with JDBC (**MYSQL**) can be used which provides more than 1 concurrent connections.

Pros

No need to write Map-Reduce programs.

An easy way to process large scale data

Support SQL based queries

Support features like Indexes, Joins, Group by etc.

Provides more user defined interfaces to extend

Efficient execution plans for performance

Cons

You can't edit data in Hive.

You can't delete particular row from table.

Default database (Derby) doesn't allow more than 1 concurrent connections.

Can't be used for transactions.

You can't access Hive tables if Hadoop is down.

More operator required (In/Exists/Views etc)

Hive Applications

Business intelligence
Advertising Delivery
Spam Detection
Log processing
Daily Report
User Activity Measurement
Data/Text mining
Machine learning

Hive Usage @ Facebook

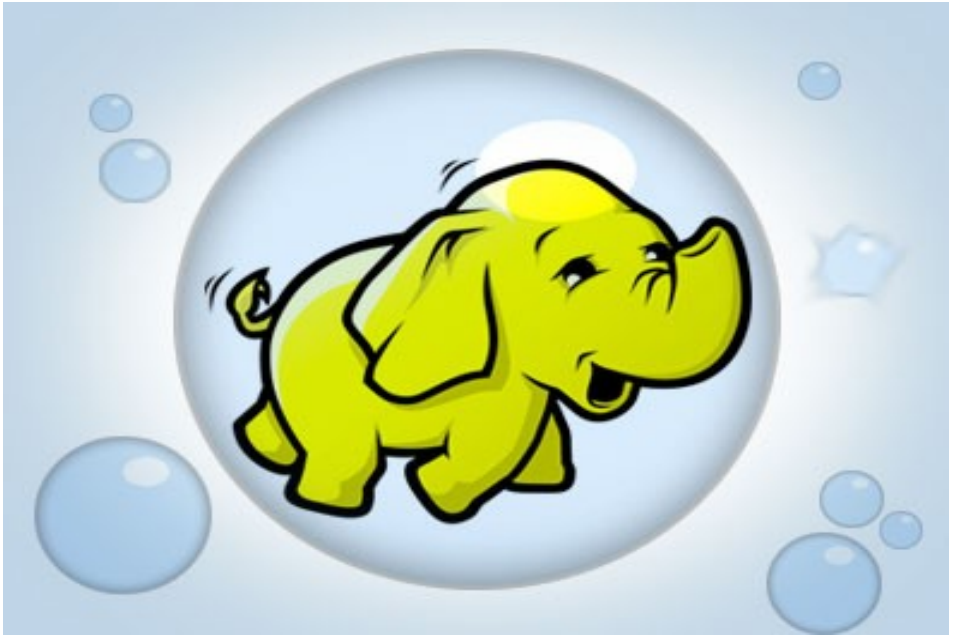
Statistics per day:

- 4 TB of compressed new data added per day
- 135 TB of compressed data scanned per day
- 7500+ Hive jobs on per day

Hive simplifies Hadoop:

- ~200 people/month run jobs on Hadoop/Hive
- Analysts use Hadoop through Hive
- 95% of jobs are Hive Jobs

...Thanks...



For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com