Hadoop/Hive

Keeyong Han

Table of Content

What is Apache Hadoop?

Quick Introduction

What is Hadoop for and isn't for?

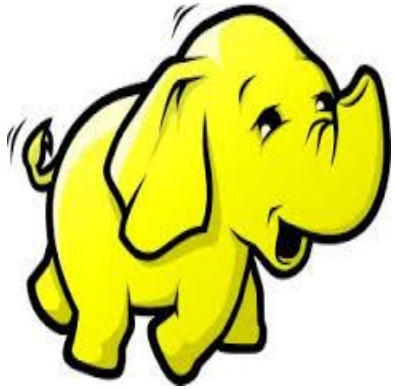
- Apache Open Source
- Processing large scale data (TB to PB) in batch fashion
- Fault-tolerant
- Linear scalability (scale-out)

Mainly for offline batch processing. Not for realtime

Certain computation pattern isn't a good fit.

In MapReduce framework, network bandwidth is a common bottleneck

Apache Hadoop



Written in Java

Hadoop 1.0

- HDFS (Hadoop Distributed File System): Distributed File System
- Map Reduce Framework: Distributed Computation Framework

Hadoop 2.0

- HDFS remains the same (with some enhancements)
- Mapreduce Framework is now replaced by more generic distributed computation framework called YARN
  - Now Mapreduce Framework is an application layer on top of YARN
  - Other application layers are Spark, Storm and so on

HDFS

- Distributed File System
  - On top of regular file systems (Linux)
    - Replication Factor
- Name Node (Master)
  - Stores directory structure, filename + data block mapping, …
    - Single point of failure
- Data Node(s)
  - Store data block itself (blindly).
    - Data block is 64MB by default

Map Reduce Framework

- Computation is done in a series of map/reduce steps
  - It is very powerful in certain tasks (like log handling)
  - Continuous (Key, Value) pair transformation
- Start-up overhead is in minutes
  - Not a good fit for realtime processing
- Uses HDFS as input and output storage
- Job Tracker
  - Gets job requests and distribute them to Task Trackers
- Task Tracker(s)
  - Real Workers. Essentially JVMs

Hadoop Distribution

- Apache
- CDH (Cloudera Distribution Hadoop)
  - De facto standard
- HDP (Hortonworks Data Platform)
- MapR
  - Not open source
- AWS EMR (Elastic MapReduce)
- Azure HDInsight (Microsoft)
- Hadoop on Google Compute Engine

What is Apache Hive?

Quick Introduction

SQL on Hadoop

Support most SQL concepts:

- database
- schema
- table/view

Load/Store data in HDFS/S3

Convert SQL query into MapReduce jobs

Support UDF for any missing features

- Scalar function (UDF)
- Aggregate function (UDAF)

Data Types

Numeric type

tinyint, smallint, int, bigint, float, double, decimal

Date/Time type

timestamp, date

String type

string, varchar (0.12+), char (0.13+)

Misc type

boolean, binary

Complex type

array (0.14+), map, struct, union

Metastore

- Store metadata for Hive tables and partitions
- Also provide this schema info to Hive client
- Have to install HiveServer or HiveServer2
  - HiveServer will be installed by default which uses Derby as an underlying information (no concurrency though)
  - We set up HiveServer2 with MySQL (Sungju did)
    - Now you can use JDBC/ODBC to access this

Concept of External Table

- What if you already have data in HDFS or S3?
  - Do I have to load it into Hive? Or can I just use it as if it is a table?
- You don’t have to load it again into Hive
- Use “CREATE **EXTERNAL** TABLE”
  - This will be READONLY

Example - create external table

```
USE redshift;

DROP TABLE IF EXISTS s3_course;
CREATE EXTERNAL TABLE s3_course (
  id bigint,
  userId bigint,
  title string,
  titleCleaned string,
  …
) row format delimited fields terminated by '\t'
LOCATION 's3n://udemy-bigdata-east/course/';
```

Example - Create a Hive table

```
DROP TABLE IF EXISTS user_course;
CREATE TABLE user_course (
    userid bigint,
    courseid bigint
) row format delimited fields terminated by '\t'
lines terminated by '\n'
STORED AS TEXTFILE LOCATION '/user/root/user_course/';
-- you can use sequencefile instead of textfile if you want
compression
```

Example - Insert into the Hive table

```
INSERT OVERWRITE TABLE user_course
SELECT chu.userId, chu.courseId
FROM s3_course_has_user chu, s3_course c
WHERE c.id = chu.courseId
    AND c.adminRating >= ${hiveconf:ADMIN_RATING_THRESHOLD}
    AND c.isPublished = 'Yes'
    AND c.isPrivate = 'No'
    AND c.sourceOrganizationId is null
    AND chu.id > 0
    AND chu.userId > 0;
```

Example - How to run a Hive query

Command mode vs. Interactive mode
Command mode:

```
hive -hiveconf
ADMIN_RATING_THRESHOLD=7 -
hiveconf
OUTPUT=/user/root/user_course_meta_d
ata/ -f build_user_course.sql
```

Q & A