Hive - A Petabyte Scale Data WareHouse Using Hadoop

Facebook Data Infrastructe Team. Hive - A Petabye Scale Data Warehouse Using Hadoop. Facebook. Web. 19 October 2016

A comparison of Approaches to Large-Scale Data Analysis

Pavlo, Andrew, Erik Paulson, Alex Rasin, Daniel Abadi, David DeWitt, Samuel Madden, Michael Stonebraker. A Comparison of Approaches to Large-Scale Data Analysis Web. 19 October 2016

Nicholas Barranco
October 19th, 2016

What is Hive?

- Due to the rapid growth in which data sets are measured in, the process for warehouse solutions is impractical as well as expensive.
- Hadoop is a popular map-reduce implementation, and is open-source, the downside to is that it requires a lot of upkeep ultimately requiring more time by programmers to perform maintenance.
- Hive is another open source data warehouse implementation that uses Hadoop.
- Hive uses HiveQL, which is similar to other SQL languages.
- It compiles into map-reduce jobs, which then run on Hadoop.
- Hive uses an implementation of Serialization/Deserialization (SerDe) java interface when the user associates the provided one to a table.
- It contains: Metastore, a system catalog.
  Driver, manages the lifestyle.
  Query Compiler, the component that compiles the HiveQL
  Execution Engine, the part that runs tasks created by the compiler;
  Hive Server, that has a thrift interface and a JBDC/ODBC server and helps integrate Hive amongst other applications.

Implementation of Hive

- Facebook uses, with 5TB (15 after replication) of compressed data added daily.
- Runs on Hadoop, so it is not fully distinctive.
- Hive can take implementation of SerDe java interface.
- Due to Hadoop’s questionable efficiency, Hive was well-accepted after implementation.
- Uses HiveQL, which is similar to any other SQL languages. Very similar and almost identical in some ways.
- Open-Source project that was easily adapted to and is a work in progress.

Analysis of Hive

- Much more efficient than previous process. Cheaper and uses less man power.
- Lack of inserting into pre-existing tables to be a problem, but has apparently not caused an error yet.
- Great support amongst SerDe, File Formats, and Data Storage.
- Its also a great effort towards moving towards more productive and cost efficient methods in the future.

A Comparison of Approaches to Large-Scale Data Analysis Summary

stsea

Advantages and Disadvantages

Advantages:

Disadvantages

XX%

Use this slide to show a major stat. It can help enforce the presentation’s main message or argument.

Analysis of Hive

- Much more efficient than previous process. Cheaper and uses less man power.
- Lack of inserting into pre-existing tables to be a problem, but has apparently not caused an error yet.
- Great support amongst SerDe, File Formats, and Data Storage.
- Its also a great effort towards moving towards more productive and cost efficient methods in the future.

Section Title

Final point

A one-line description of it

“This is a super-important quote”

- From an expert

This is the most important takeaway that everyone has to remember.

Thanks!

Contact us:

Your Company
123 Your Street
Your City, ST 12345

no_reply@example.com
www.example.com