FIU

FLORIDA INTERNATIONAL UNIVERSITY

COP 5725 Principles of Database Management Systems

Final Project Proposal

Student Name: Yezehao Huai

PID: 5965823

• Title

Research of Hive data warehouse

• Abstract

With the amount of access to the websites and electronic business platforms become more and more, the related information like Web logs is wildly increasing and the traditional database is incapable for the requirements. Hive is Data Warehouse infrastructure developed by Facebook. The HiveQL of Hive is a SQL-like declarative language to compile queries expression into map-reduce jobs executed on Hadoop. This technology made big data processing efficiency significantly improved in Hive data warehouse compared with the general database. It is very important for the big data storage and analysis.

• Motivations of the research

For the class, data warehouse is a significant idea of database. This research is related to the Part 10(chapter 23, 24, 25) of course textbook. It is really helpful for understanding database deeply by learning data warehouse and distributed database.

For myself, I was doing some work for a P2P financing platform in China. Mining the useful information in database is the inevitable way to create the report. I could learn plenty of knowledges to solve big data problems by doing this research

For general, Hadoop is a popular big data technologies and the Hive is a warehousing solution over Hadoop Map-Reduce framework. It also support other big data distributed programming technologies like Tez and Spark. My research could be helpful for other researchers who need further study in big data technologies.

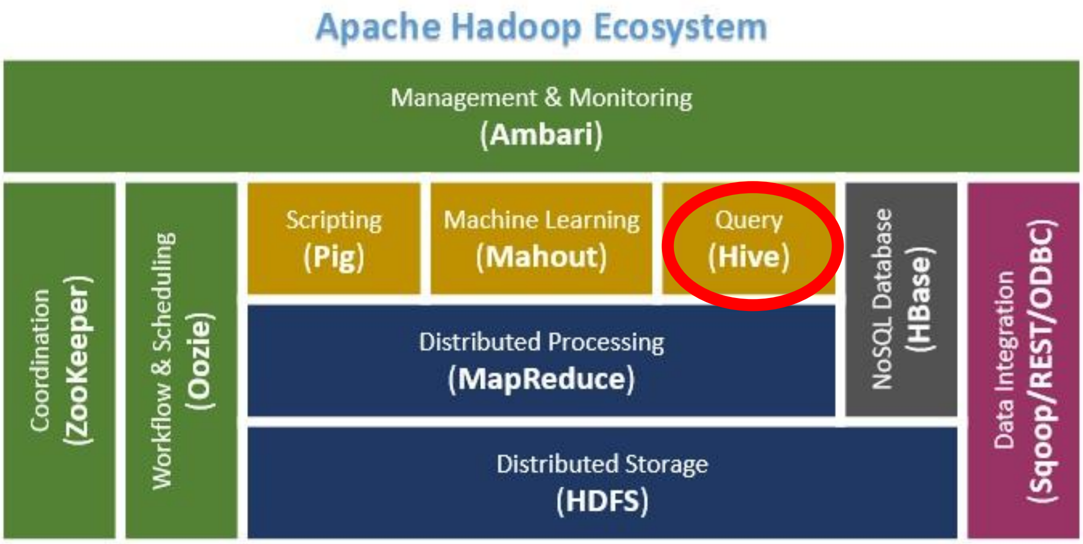
- **Main methods and content**

To achieve the aims, the research will focus on the principles of data warehouse and Hive based on the study of HDFS and MapReduce. The research on the features and applications of Hive and HiveQL is also a major part. Then, I will do some research on a case about establishing a Web logs data warehouse based on Hive. The further study is something about Hive on Spark and its relationship with HBase.

- **Result and conclusion**

The deliverable of the project will be a paper research and a simple Web logs data warehouse based on Hive.

Nowadays, the Hadoop Ecosystem is unavoidable and essential for the people from every fields doing work with big data. The research on Hive not only provide the knowledges of database, but also the comprehensions of the basic Hadoop Ecosystem. I am focusing on the financial big data analysis, and this is really interesting and cooperative with my study and the course.


Apache Hadoop Ecosystem
Management & Monitoring
(Ambari)
Coordination
(ZooKeeper)
Workflow & Scheduling
(Oozie)
Scripting
(Pig)
Machine Learning
(Mahout)
Query
(Hive)
NoSQL Database
(HBase)
Data Integration
(Sqoop/REST/ODBC)
Distributed Processing
(MapReduce)
Distributed Storage
(HDFS)