

EXAM SCRIPT COVER PAGE



J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS

SEMESTER 2 EXAMINATION

Module Name and Code

MS5106

Student ID

21230412

In submitting this script, I am aware that it is my responsibility to adhere to the examination guidelines. Please tick (**Yes/No**) for the following:

| | Yes | No |
|---|-------------------------------------|--------------------------|
| I have read the module owners description/expectations for this exam and the submission process. | <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| I have read the examination guidelines. | <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| I have read the examination checklist. | <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| I am aware of what the NUI Galway plagiarism policy entails. | <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| I have submitted a Learning and Educational Needs Summary (LENS) report to the School office (if applicable) | <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| I have saved the script file (MS Word, Excel or PDF) to contain only my student ID and the module code (e.g., 1187404-MS2101.doc) for submission. | <input checked="" type="checkbox"/> | <input type="checkbox"/> |

Declaration for this Exam Submission:

In submitting this work I confirm that it is entirely my own. I acknowledge that I may be invited to online interview if there is any concern in relation to the integrity of my exam, and I am aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations.

Big data

An IT industry term, refers to a collection of data that cannot be captured, managed and processed within a certain time frame with conventional software tools (such as mysql, ssm, etc.), and is a massive, high growth rate and diverse (image, voice, etc.) information asset that requires new processing model to have stronger decision-making power, insight discovery and process optimization capabilities. It mainly addresses the storage of massive data and the analysis and calculation of massive data





Clustering

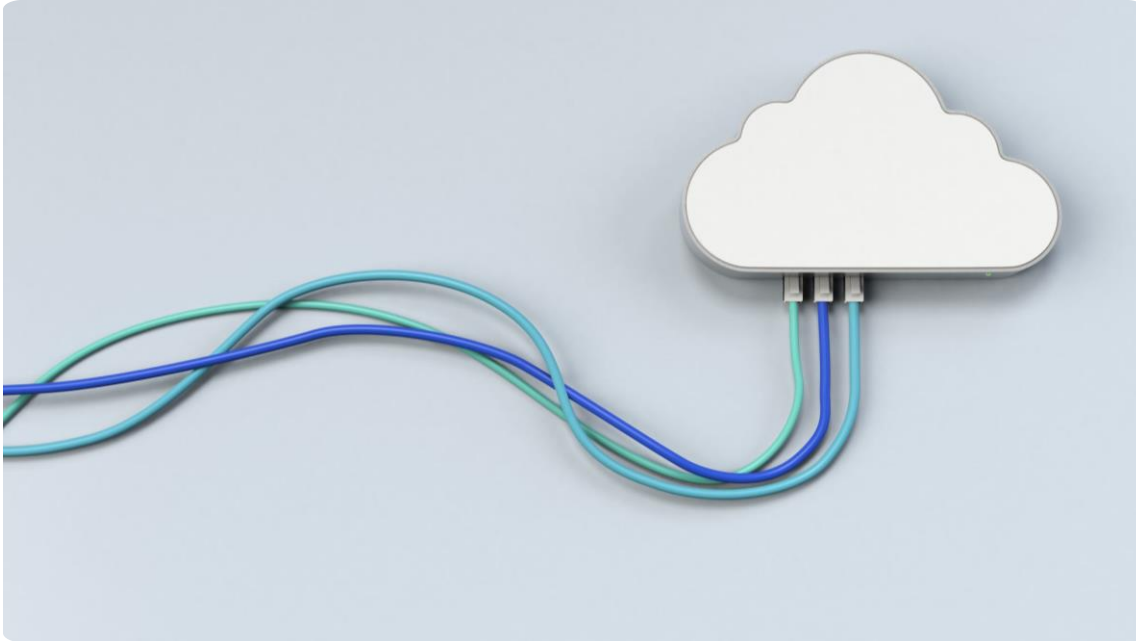
- Clustering means that a number of servers are grouped together to perform the same service, which appears to the client as if there is only one server. Clustering can use multiple computers for parallel computing to achieve high computing speed, and it can also use multiple computers for backup, so that if any one machine goes down the whole system can still run normally.

A large group of Doraemon characters, yellow cats with white faces and red noses, are standing in rows. They are all looking forward with expressions of surprise or excitement. The background is a bright blue sky with some clouds. The characters are arranged in a way that suggests a large-scale operation or a team effort.

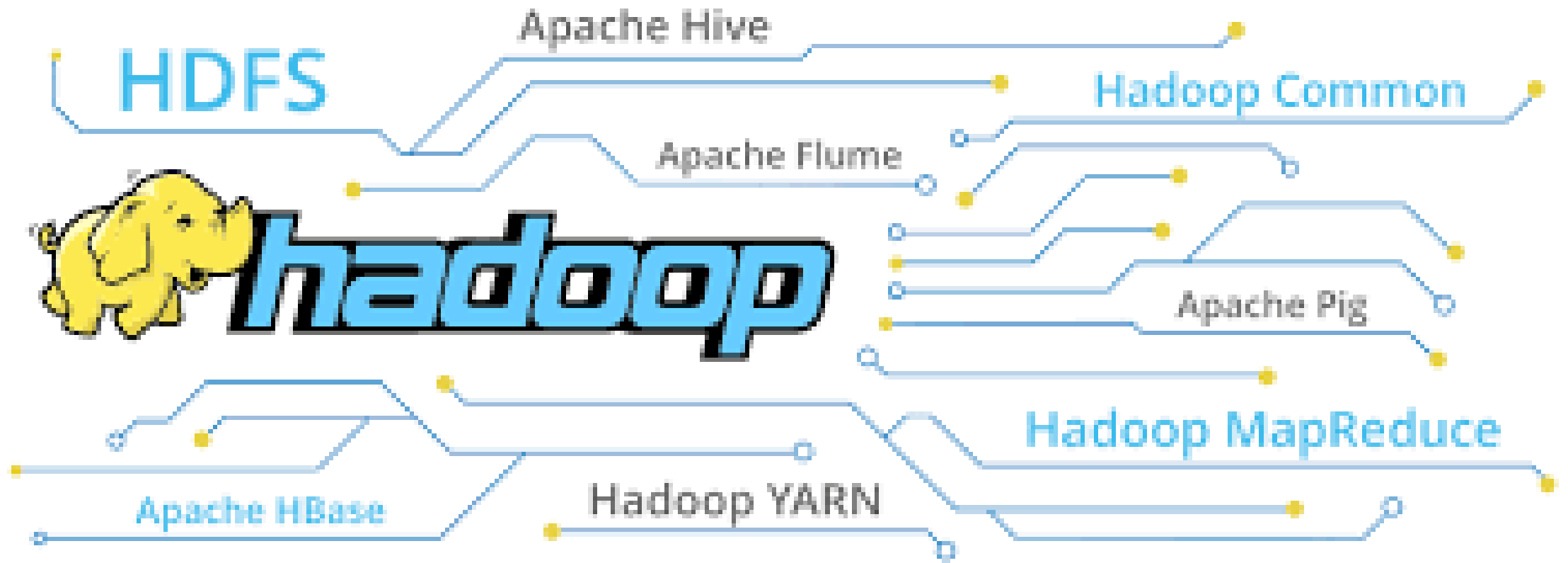
Imagine a lot of computers performing a task at the same time, or a lot of people working on a project at the same time, the big task is divided into different small tasks to calculate, for example, you open an excel sheet with a lot of data, one computer loading time is very slow, but if you use the computing resources of multiple computers to open the data at the same time, the speed will become much faster than a single computer.

These Doraemon are like one server at a time, they are completing a task at the same time, and the failure of a single node does not affect the completion of the whole task

Cloud



- You can rent a server, deploy your services to the network to execute, so you do not have to spend a lot of equipment and venues to build computing clusters, you just need to rent a server such as AWS to deploy these services can easily use it, a simple example is that you do not have to buy these Doraemon and venues to build your server cluster, you just need to tell (deploy) them through the network You just need to tell (deploy) them the services you want through the network, let them help you execute, and return the results to you.

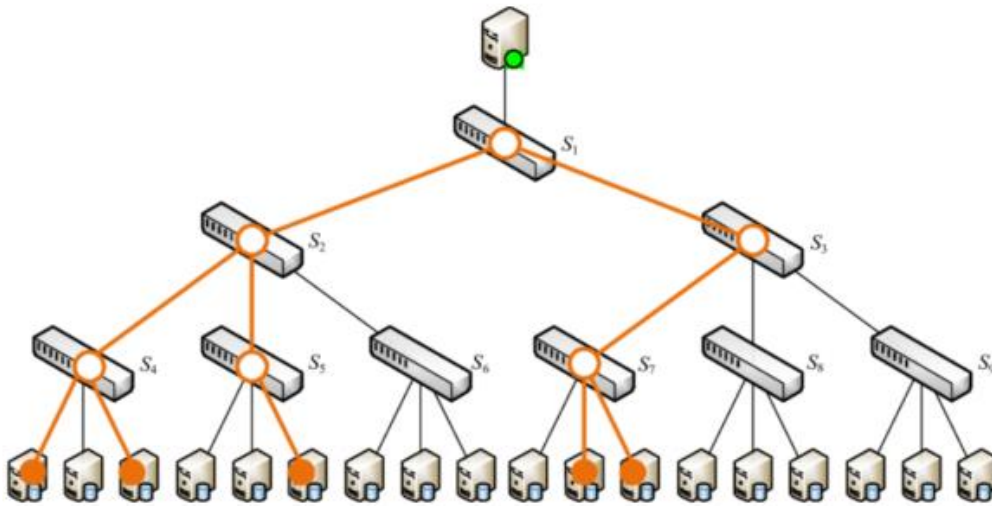


Hadoop is written in java language, open source framework for storing massive data and running distributed analytic applications on distributed server clusters, whose core components are HDFS and MapReduce.

The core design of Hadoop framework is: **HDFS** and **MapReduce**. HDFS provides storage for large amounts of data, MapReduce provides computation for large amounts of data.

HDFS

(Hadoop Distributed FileSystem)



HDFS (Hadoop Distributed FileSystem) is essentially designed for large amounts of data to span hundreds or thousands of machines, but users see one file system instead of many file systems. For example, if a user wants to get data from `/hdfs/tmp/file1`, the reference is to a file path, but the actual data is stored on many different machines. As a user, do not need to know these .

How to use Hadoop

To put it simply, you put the Hadoop installation package on each server, modify and adjust its configuration, and start it to finish building the Hadoop cluster.

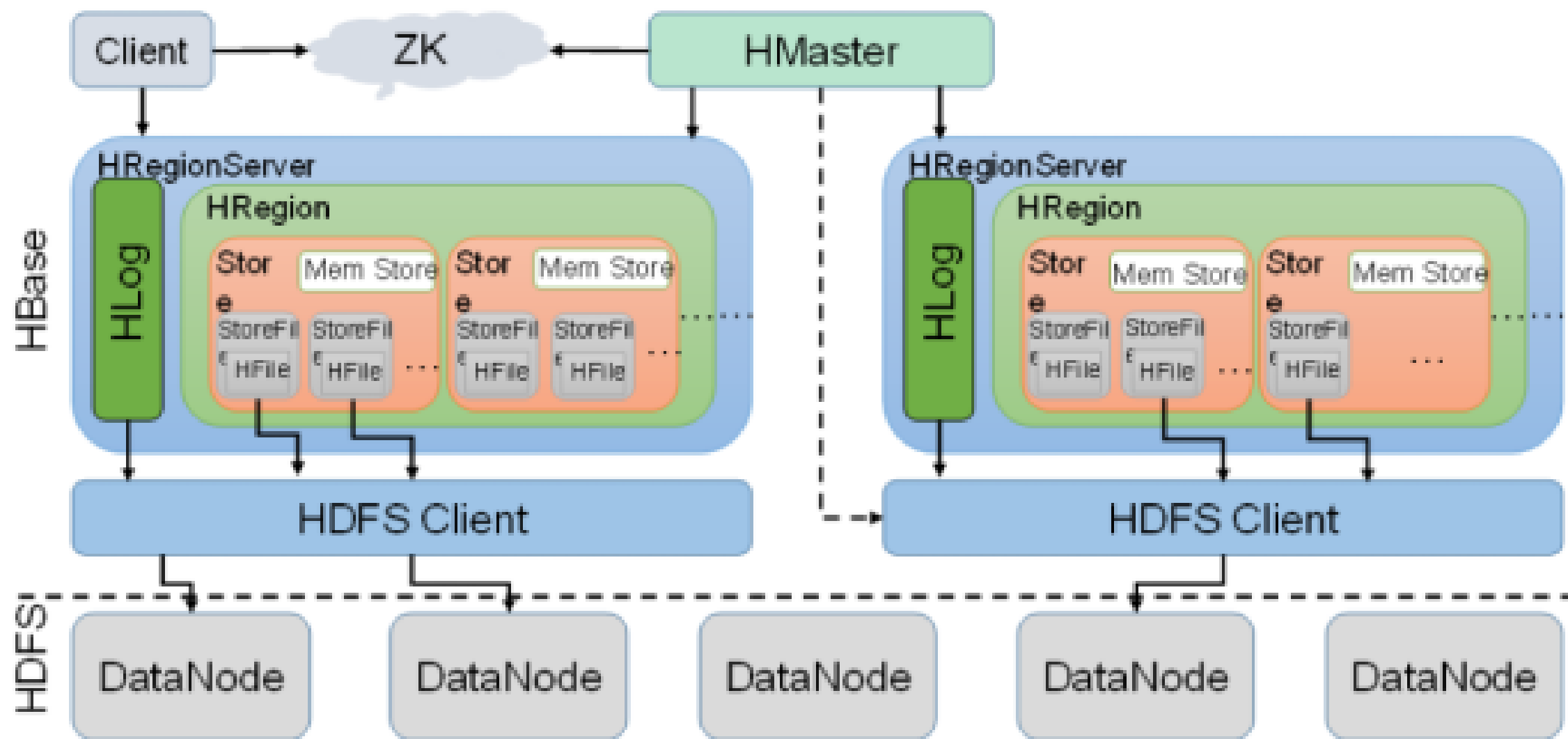
Uploading files to the Hadoop cluster for file storage

Imagine, is when you are operating a computer, you store files to your desktop, only here replaced by you through the HDFS to store your files in the server cluster above, Hadoop cluster set up, you can view the cluster through the web page, but also through the Hadoop command to upload files to the hdfs cluster, through the Hadoop command to create a directory on the hdfs cluster, through the Hadoop command to delete the files on the cluster and so on.

HBase

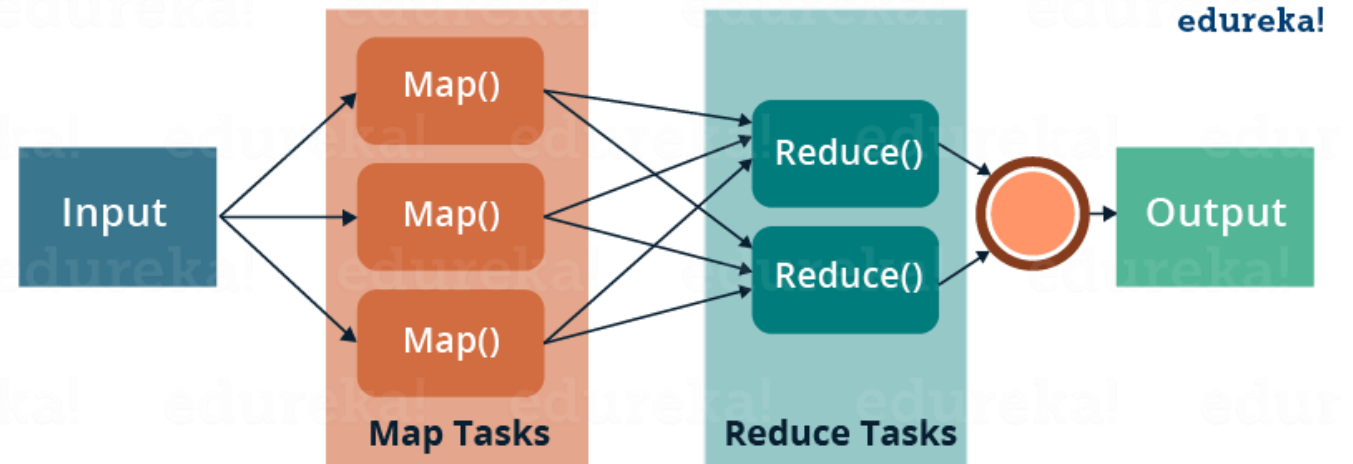
- In addition, HDFS with Hbase can be used in the form of key value to organize the content of the file, when the search first find the key can easily narrow down the search content, similar to the library you want to find a book, you can first classify the book into different categories, and then according to the category to find a book, so you do not have to search all the books at once.





MapReduce

- Write map/reduce program to complete the computation task
- Import Hadoop related jar packages through integrated development tools (eclipse, for example), write the map/reduce program, and throw the program as a jar package on the cluster to execute it, and then run it to get the computation results. You can also give the task of writing map/reduce to Hive, and use Sql-like language to do map/reduce to get the relevant data.



Yarn framework

Write map/reduce program to complete the computation task

Import Hadoop related jar packages through integrated development tools (eclipse, for example), write the map/reduce program, and throw the program as a jar package on the cluster to execute it, and then run it to get the computation results. You can also give the task of writing map/reduce to Hive, and use Sql-like language to do map/reduce to get the relevant data.

For companies, they can choose a Hadoop and hive solution

Or they can choose Hadoop and spark.

hive and spark they are both built on the yarn framework

The Yarn framework is composed of four parts

ResourceManager

Compute resource management for the entire compute cluster (equivalent to the administrator of all Doraemon)

NodeManager

Computational resource management for a single node server (equivalent to a single Doraemon)

Container

A container for encapsulating individual tasks, allocating computational resources to each task (these tasks include AppicaionMaster , map, reduce)

AppicaionMaster

Manage the computational resources of a set of map,reduce tasks



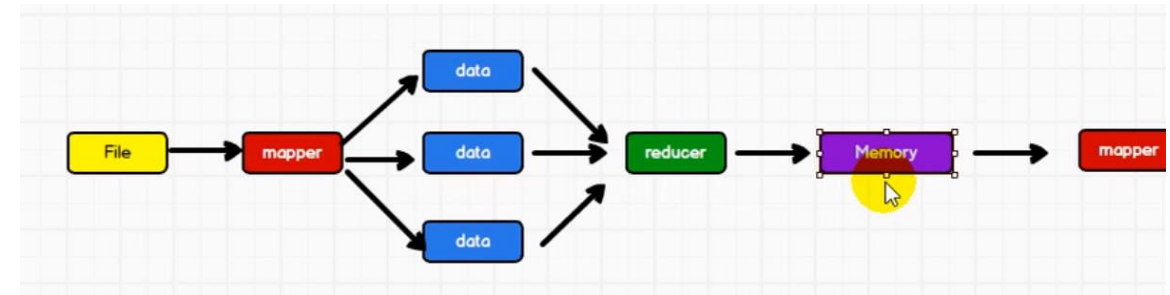
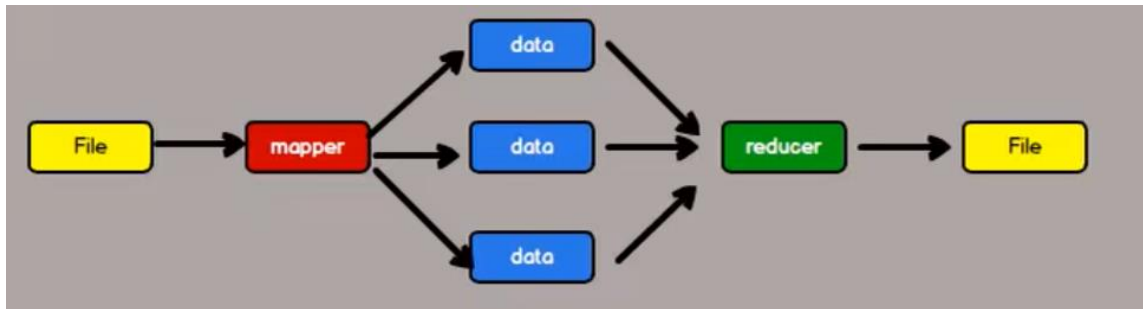
Hive and Spark SQL are both tools used to help people with big data queries, They are similar to mapreduce program steps to query data

The difference is that Hive is used with Hadoop's MapReduce to perform a single calculation, the advantage is that it requires less memory, the disadvantage is that it can not calculate data online in real time, the calculation latency is higher than spark, but the memory performance required to build a cluster is not high.

And spark is the use of spark core to replace MapReduce, the latter can facilitate the rapid implementation of multiple iterative computing, more suitable for machine learning and real-time computing, but the cost is that it requires more memory, which requires more money to build or rent a cluster server.

spark saves data in memory after MapReduce to facilitate continued iterative computation, which is the biggest difference between spark and MapReduce.

Spark and MapReduce



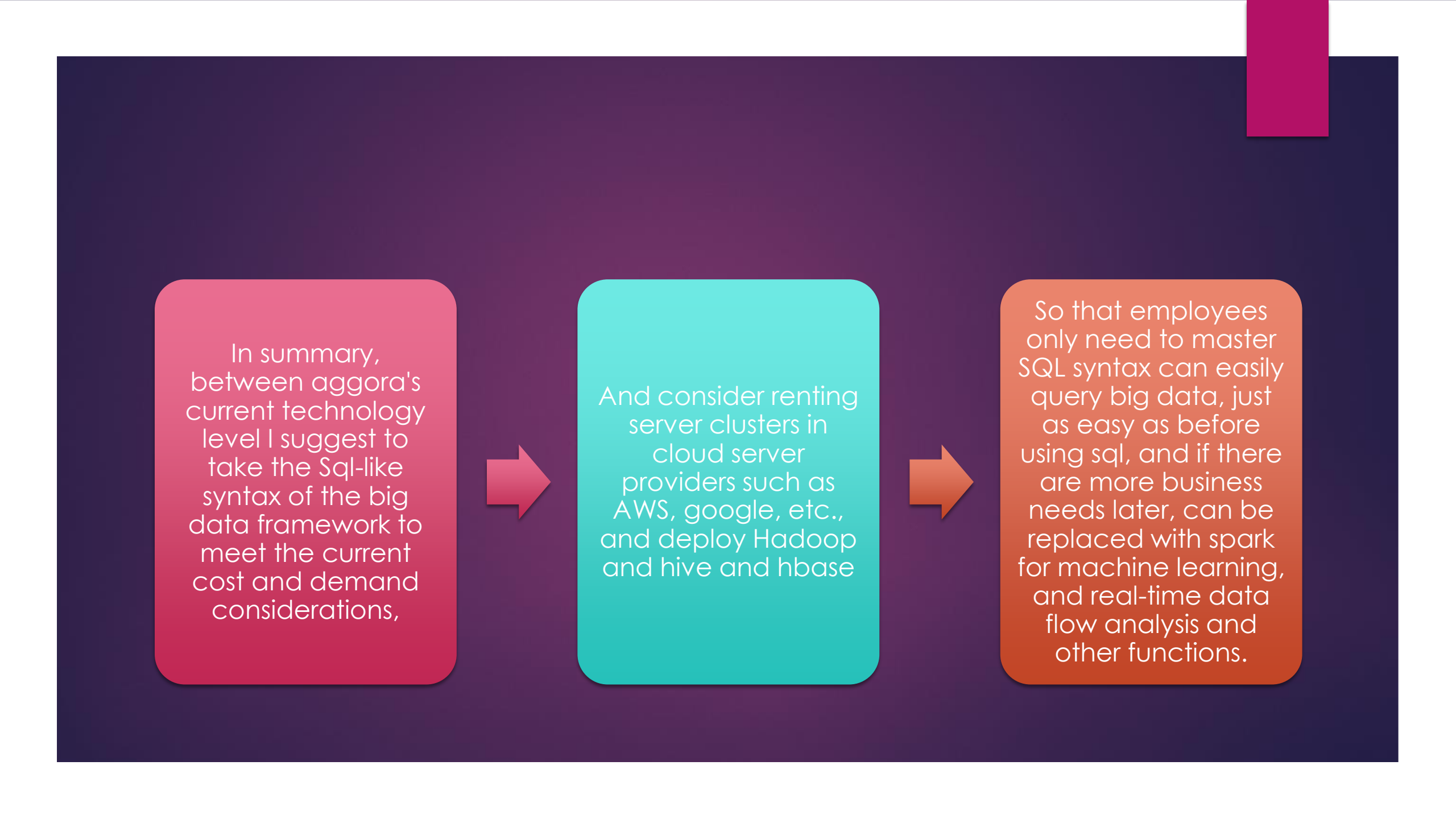
PIG

- In addition, you can also take **pig**, which works very similar to hive but is more suitable for semi-structured data
- Pig is a large-scale data analysis platform based on Hadoop, which provides a SQL-LIKE language called Pig Latin, whose compiler converts SQL-like data analysis requests into a series of optimized MapReduce operations. pig provides a simple operational and programming interface for complex parallel computation of large amounts of data.
- Hive is more suited to data warehousing tasks, and Hive is primarily used for static structures and jobs that require frequent analysis. hive's similarity to SQL makes it an ideal intersection for combining Hadoop with other BI tools. pig gives developers more flexibility in the area of large data sets and allows the development of clean scripts for transforming data streams for embedding into larger applications. Pig is relatively lightweight compared to Hive, and its main advantage is that it cuts the amount of code significantly compared to using the Hadoop Java APIs directly.





| Hive | Spark Sql | Pig |
|---|---|---|
| Lower server cost overhead | Higher server cost overhead | Lower server cost overhead |
| Slightly slower calculation speed | Fast computation speed | Basically the same as Hive |
| SQL-like query language | SQL-like query language | Pig Latin |
| Not suitable for iterative computing and machine learning | Supports iterative computing and machine learning | Not suitable for iterative computing and machine learning |



In summary,
between aggora's
current technology
level I suggest to
take the Sql-like
syntax of the big
data framework to
meet the current
cost and demand
considerations,



And consider renting
server clusters in
cloud server
providers such as
AWS, google, etc.,
and deploy Hadoop
and hive and hbase



So that employees
only need to master
SQL syntax can easily
query big data, just
as easy as before
using sql, and if there
are more business
needs later, can be
replaced with spark
for machine learning,
and real-time data
flow analysis and
other functions.