DATA WAREHOUSE & MINING

TA COURSE #01

INTRO

TA COURSE #01

INTRO

CONTACT ANOTHER TA

CONTACT ANOTHER TA

- Email: dmdw_ustc2019@163.com
  - Nothing except for mails sent to the mail addr shown above will be responded.
- QQ group: 966512038
- All code works must be kept as git repos for examinations

TA COURSE #01

INTRO

PURPOSE

GLOBAL PURPOSE

- To learn by participating
- To improve by practicing
- To create by experimenting

PURPOSE OF LESSON TODAY

- To have an initial knowledge of the stuffs to learn to use
-

TA COURSE #01

INTRO

SYSTEM ENV

LINUX SYSTEM

- Nothing for rookie, all for geek
- Env for almost all professional applications
- Various distinctive distributions —— Important to choose a suitable one for oneself
- Access [our LUG repo] for more info





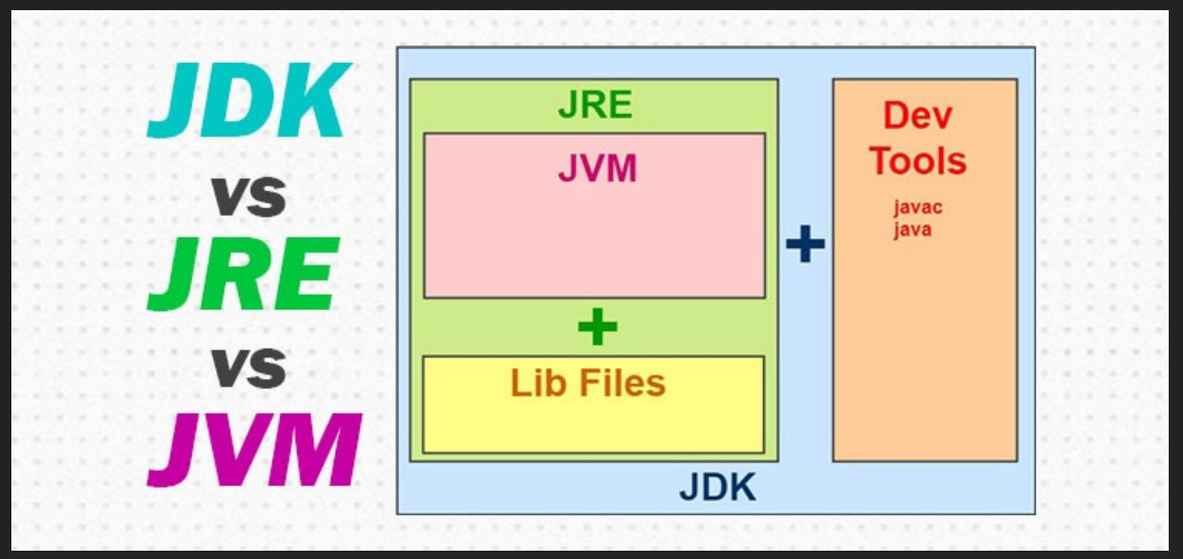
GNU COMPILER COLLECTION (GCC)

- GNU is Not a Unix
- Make basic compilations for everything you will need
- HPC development
- Cross-language embedding with other frameworks
- Difference between C99 and C11~: compiler version matters
- More info

(I'm not a clang gang)

JAVA DEVELOPMENT KIT (JDK)

- GC makes me healthy
- Java & scala-based development
- JVM–>JRE–>JDK
- More info



SCALA

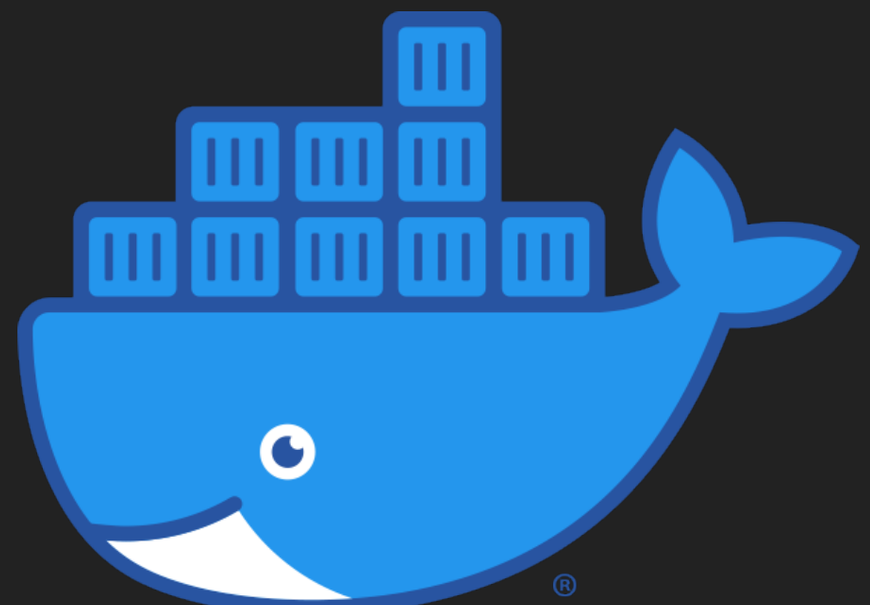
- God knows...the charms of functional language
- A JVM-based functional language
- Some special language characteristics
- Will be introduced at next TA lesson talked by me
- More info about scala and functional programming

DEPENDENCY & COMPILATION MANAGEMENT

- Tomorrow is in your hands
  - ——Hideo Kojima
- Key for large scale project management
- C/C++ developments: CMake
- Java developments: maven
- Scala developments: sbt
- ...

VIRTUALIZATION

- Pretend to be rich enough
- To simulate a customized system env
  - Virtual machine
    - Virtualbox
    - VMWare(not recommended)
  - Kernel virtualization
    - docker(recommended)
  - ...



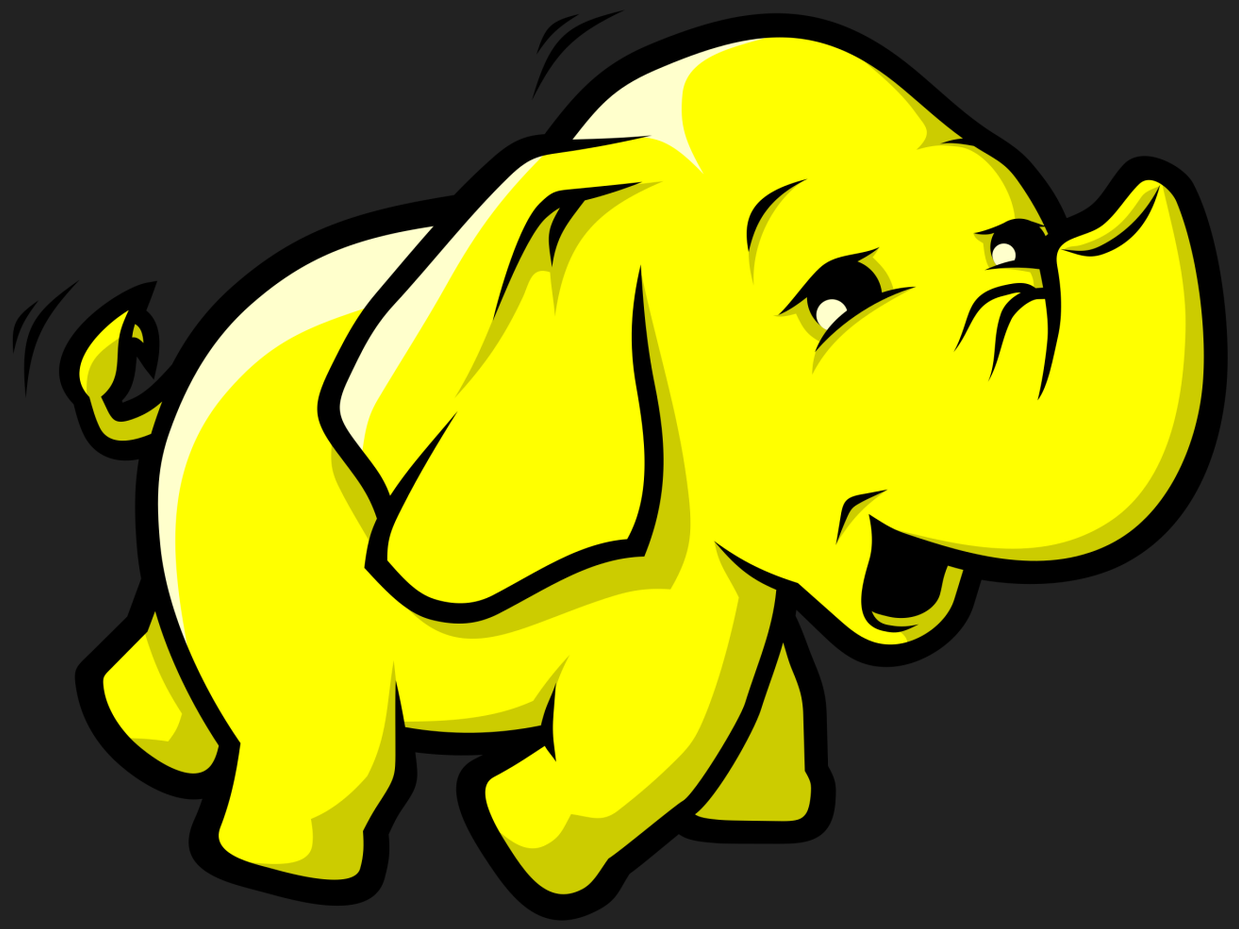
TA COURSE #01

INTRO

APACHE FAMILY

APACHE HADOOP

- You will certainly fail if you can not build this elephant env
- A framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models
- The base of apache data management toolkits
- More info

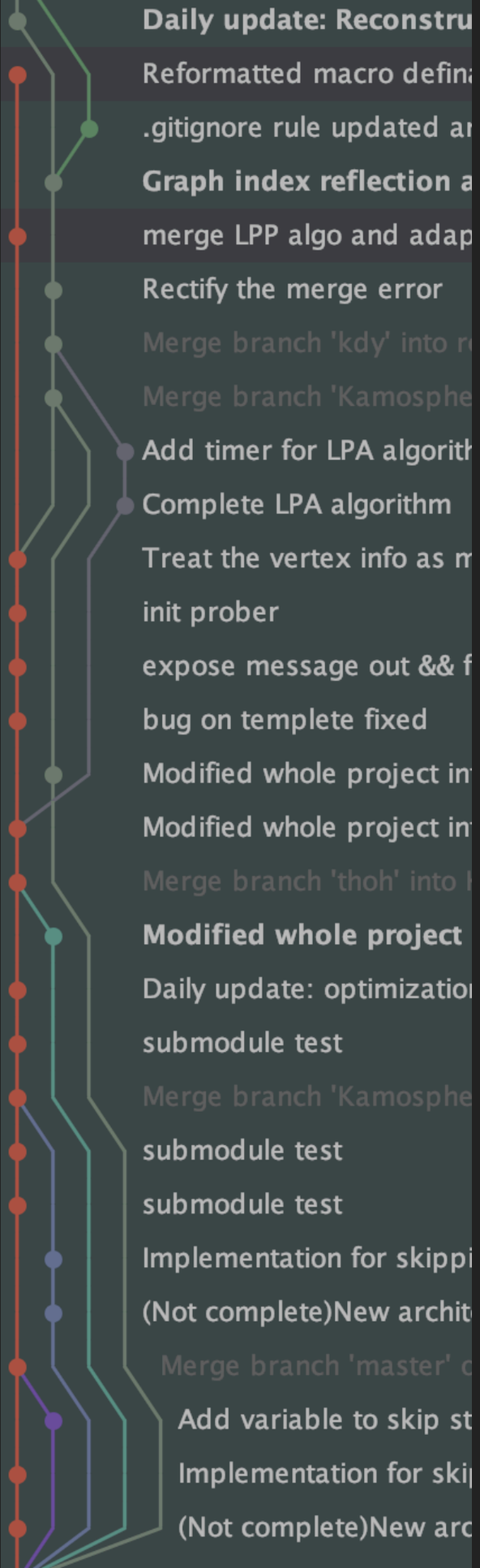


APACHE HIVE

- A heaven feeling like databases
- A data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL
- Built on Hadoop
- (Maybe replaced by Spark SQL sometimes)
- More info

APACHE SPARK

- Just like mathematics theories, easy to use, and (maybe) not difficult to build
- A unified analytics engine for large-scale data processing on distributed system framework
- Also built on Hadoop
- More info



TA COURSE #01

INTRO

VCS

GIT

- git, a British slang
- I'm an egotistical bastard, and I name all my projects after myself. First Linux, now git.
  - ——Linus Benedict Torvalds
- Useful for large scale project with multiple developers.
- More info

REPOS

- A place we make friend?s
- Easy to hold, easy to manage
- Use existing repo:
  - github click here
- Or build a customized one
  - gitlab click here
-

TA COURSE #01

INTRO

SOME RECOMMEND ATION

RECOMMENDATION

- A better IDE
- A better text editor
- A better ssh shell
- A better clarketech to cross a specific thing
- …

RECOMMENDATION

- A better spirit to keep trying

Q&A AND THANKS