* 1. **Database creation**

git log –pretty="%H,%aN, %s" > db.csv

git log –pretty=”%H, %aN, %aI, %at” > timestamp.csv

**1.2 Database system installation**

• What are the most common database systems?

MySQL, SQLite, Oracle, DB2…

• Briefly list the pros and cons of the three most common ones.

Oracle:

pros: robust database management tools

cons: high cost & significant resources required

MySQL:

pros: free & able to work with other databases

cons: no built-in support for XML or OLAP

DB2:

pros: multiple jobs can be run at once

cons: third party tools or additional software is required to make some functions work

• Create an empty SQLite database.

sqlite3 ve482.sqlite3

• Use the SQLite shell to prepare two empty tables for each of your .csv file.

create table db

(

HC TEXT NOT NULL,

aN TEXT NOT NULL,

s TEXT NOT NULL

);

create table timestamp

(

HC TEXT NOT NULL,

aN TEXT NOT NULL,

aI TEXT NOT NULL,

at TEXT NOT NULL

);

• Import each .csv file in its corresponding SQLite table.

.separator “,”

.import db.csv db

.import timestamp.csv timestamp

1. **Database queries**

• Who are the top five contributors to the Linux kernel since the beginning?

select aN, cnt from (select aN, COUNT(aN) AS cnt from timestamp group by aN order by cnt DESC) limit 5;

The top five contributors to the Linux kernel since the beginning are: Linus Torvalds, David S. Miller, Mark Brown, Takashi Iwai, Arnd Bergmann.

• Who are the top five contributors to the Linux kernel for each year over the past five years?

date time stamp

2019/1/1 1546272000

2018/1/1 1514736000

2017/1/1 1483200000

2016/1/1 1451577600

2015/1/1 1420041600

2014/1/1 1388505600

select aN, cnt from (select aN, COUNT(aN) AS cnt from timestamp where at>1388505600 and at<1420041600group by aN order by cnt DESC) limit 5;

The top five contributors to the Linux kernel for each year over the past five years is:

2014-2015: Linus Torvalds, Hartley Sweeten, David S. Miller, Jes Sorensen, Malcolm

2015-2016: Linus Torvalds, David S. Miller, H Hartley Sweeten, Arnd Bergmann, Geert Uytterhoeven

2016-2017: Linus Torvalds, Arnd Bermann, David S. Miller, Chris Wilson, Mauro Carvalho Chehab

2017-2018: Linus Torvalds, Arnd Bermann, David S. Miller, Chris Wilson, Arvind Yadav

2018-2019: Arnd Bermann, David S. Miller, Chris Wilson, Christoph Hellwig, Coin Ian King

• What is the most common “commit subject”?

select s, MAX(cnt) from (select s, COUNT(s) as cnt from db group by s);

The most common “commit subject” is x86.

• On which day is the number of commits the highest?

select aI, MAX(cnt) from (select aI, COUNT(aI) AS cnt from timestamp group by at);

The day that the number of commits the highest is 2017/11/1.

• Determine the average time between two commits for the five main contributor.

First use:

select COUNT(\*) from timestamp where aN=”Linus Torvalds”

to get the total number of commit of Linus Torvalds, which is 28312.

Then:

select avg(aa) from (select (julianday(a)) as aa from (select aI as a from (select aI from timestamp where An=”Linus Torvalds”) limit 28311));

select avg(bb) from (select (julianday(b)) as bb from (select aI as b from (select aI from timestamp where An=”Linus Torvalds”) limit 28311 offset 1));

The first result is 2456368.6036361, and the second result is 2456268.4179278. Average time between two commits for Linus Torvald = the first result - the second result = 0.185703 day.

The average time between two commits for the other four main contributor can be obtained similarly.