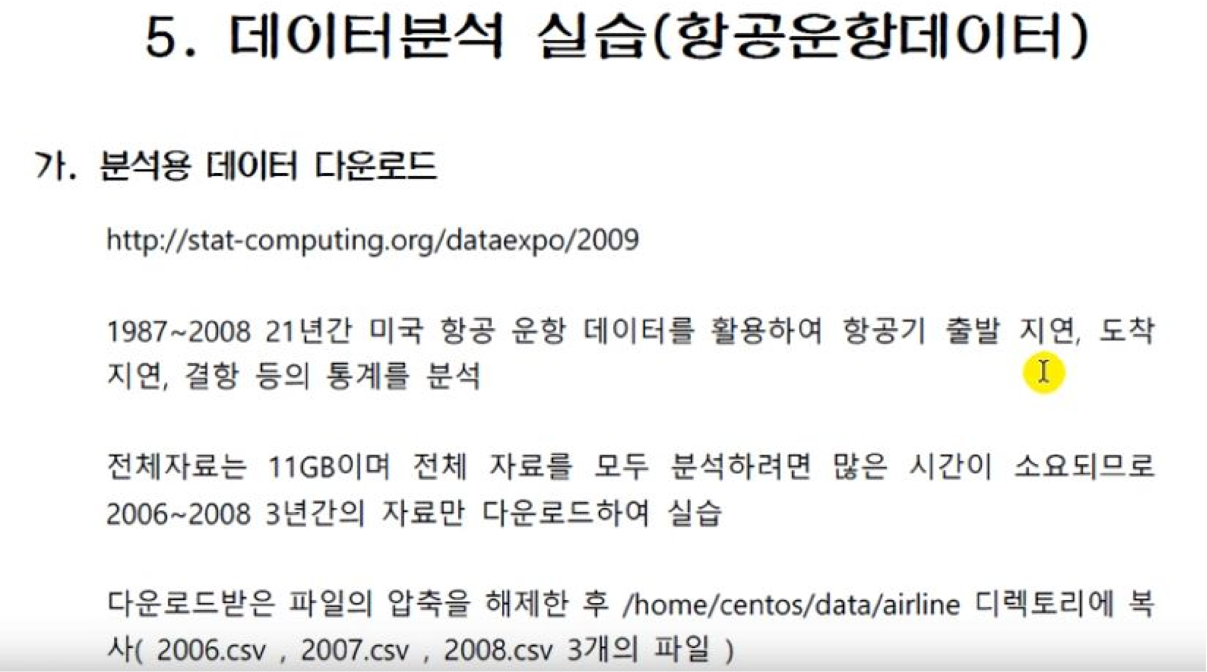


5. 데이터분석 실습(항공운항데이터)

가. 분석용 데이터 다운로드

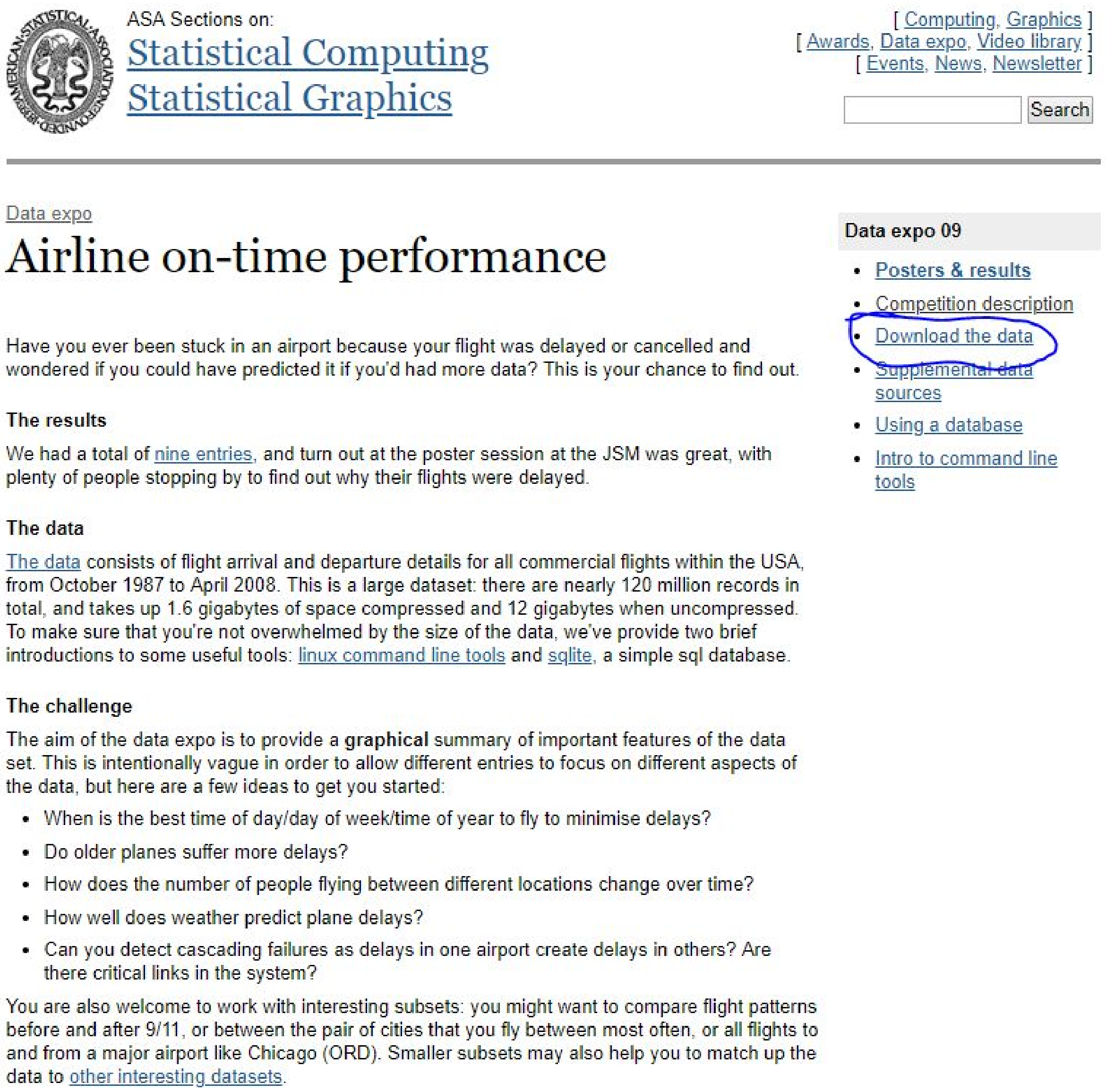
http://stat-computing.org/dataexpo/2009

1987~2008 21년간 미국 항공 운항 데이터를 활용하여 항공기 출발 지연, 도착 지연, 결항 등의 통계를 분석

전체자료는 11GB이며 전체 자료를 모두 분석하려면 많은 시간이 소요되므로 2006~2008 3년간의 자료만 다운로드하여 실습

다운로드받은 파일의 압축을 해제한 후 /home/centos/data/airline 디렉토리에 복사(2006.csv , 2007.csv , 2008.csv 3개의 파일)

http://stat-computing.org/dataexpo/2009/


ASA Sections on:
Statistical Computing
Statistical Graphics

[Computing, Graphics]
[Awards, Data expo, Video library]
[Events, News, Newsletter]

Search

Data expo

Airline on-time performance

Have you ever been stuck in an airport because your flight was delayed or cancelled and wondered if you could have predicted it if you'd had more data? This is your chance to find out.

The results

We had a total of nine entries, and turn out at the poster session at the JSM was great, with plenty of people stopping by to find out why their flights were delayed.

The data

The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed. To make sure that you're not overwhelmed by the size of the data, we've provide two brief introductions to some useful tools: linux command line tools and sqlite, a simple sql database.

The challenge

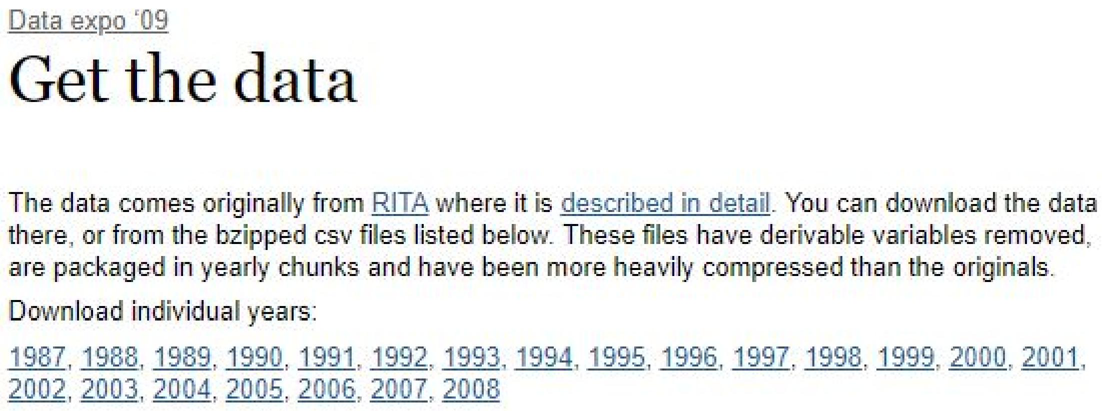
The aim of the data expo is to provide a **graphical** summary of important features of the data set. This is intentionally vague in order to allow different entries to focus on different aspects of the data, but here are a few ideas to get you started:

- When is the best time of day/day of week/time of year to fly to minimise delays?
- Do older planes suffer more delays?
- How does the number of people flying between different locations change over time?
- How well does weather predict plane delays?
- Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?

You are also welcome to work with interesting subsets: you might want to compare flight patterns before and after 9/11, or between the pair of cities that you fly between most often, or all flights to and from a major airport like Chicago (ORD). Smaller subsets may also help you to match up the data to other interesting datasets.

Data expo 09

- **Posters & results**
- Competition description
- Download the data
- Supplemental data sources
- Using a database
- Intro to command line tools


Data expo '09

Get the data

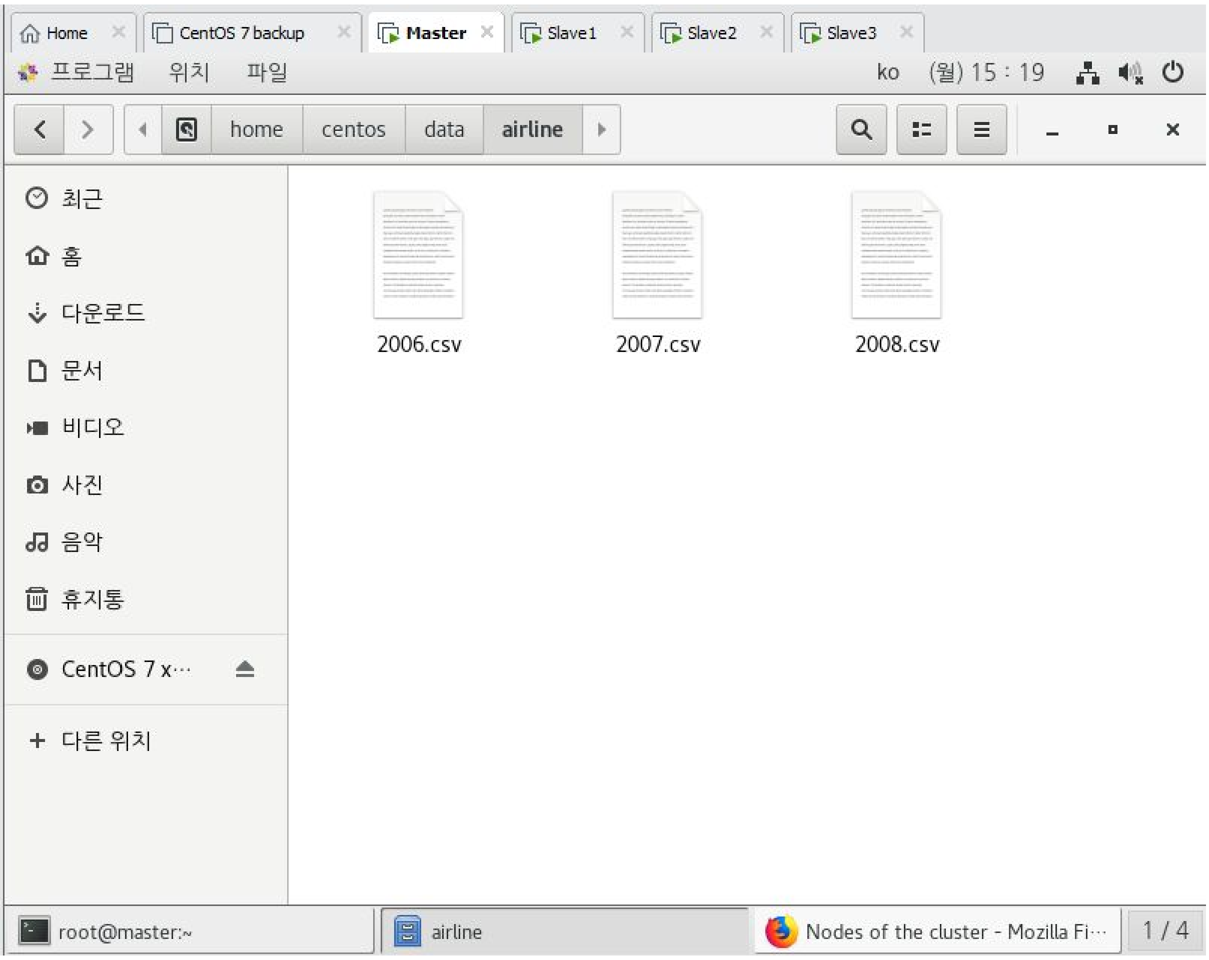
The data comes originally from RITA where it is described in detail. You can download the data there, or from the bzipped csv files listed below. These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals.

Download individual years:

1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008

1987~2008까지 출발시간 및 도착시간이 얼마나 지연되었는지에 대한 데이터이다.

총 합치면 11GB이기 때문에, 2006~2008까지의 데이터로만 실습하도록 한다.



하둡분석시간과 mysql을 통한 DBMS의 분석시간의 차이를 분석해보려고한다.

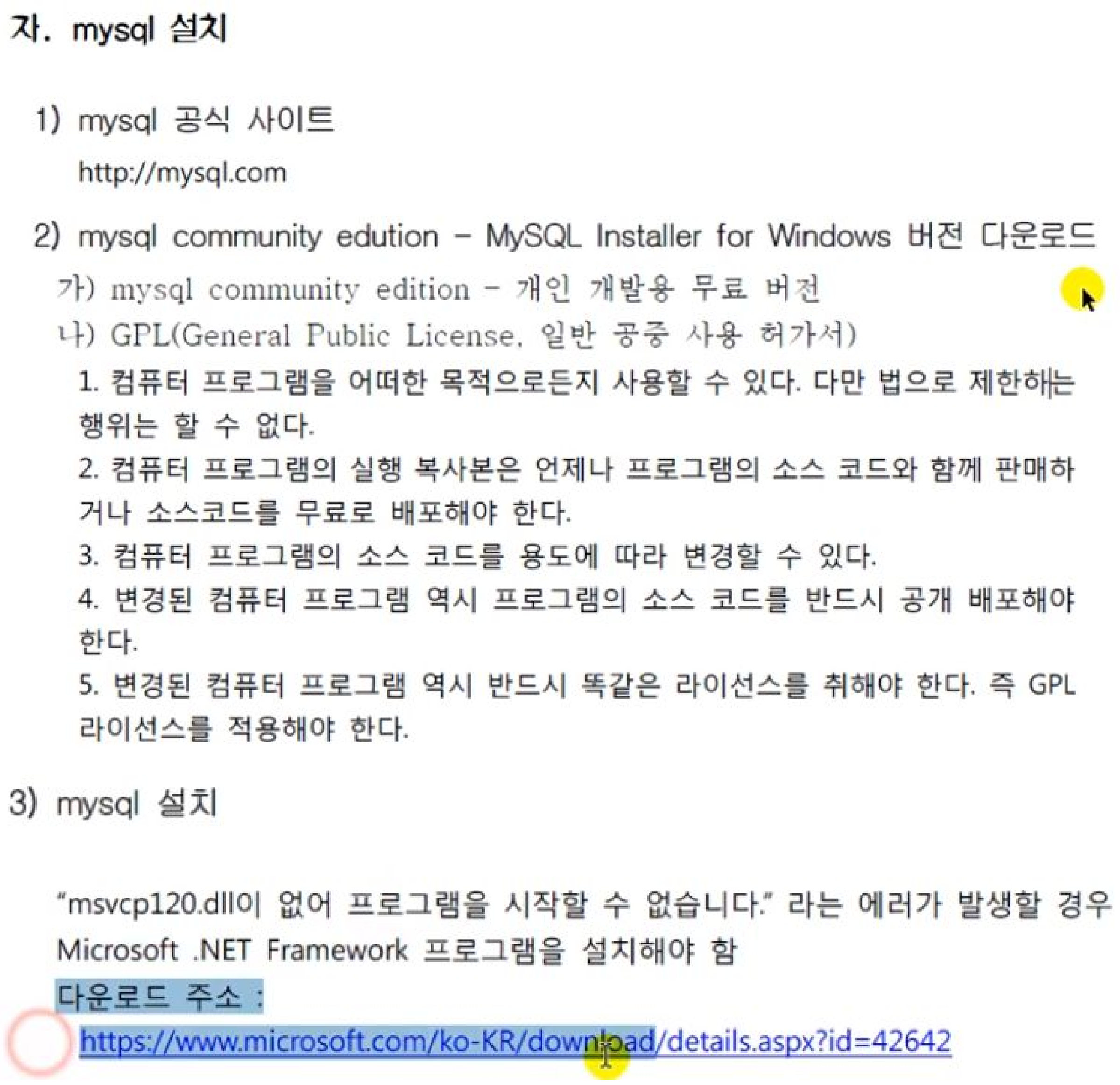
나. mysql 설치(윈도우즈에서 작업)

HDFS에서 분석하는 방법과 비교하기 위하여 다운로드받은 csv 파일을 mysql 데이터베이스 테이블로 import

mysql 설치 방법은 Java 21강 - 데이터베이스프로그래밍 참조

csv 파일을 mysql 테이블로 import하기 위하여 cmd에서 아래와 같이 mysql에 접속

```
mysql --local-infile -u root -p
```



자. mysql 설치

1) mysql 공식 사이트

http://mysql.com

2) mysql community edution – MySQL Installer for Windows 버전 다운로드

가) mysql community edition – 개인 개발용 무료 버전

나) GPL(General Public License, 일반 공중 사용 허가서)

1. 컴퓨터 프로그램을 어떠한 목적으로든지 사용할 수 있다. 다만 법으로 제한하는 행위는 할 수 없다.
2. 컴퓨터 프로그램의 실행 복사본은 언제나 프로그램의 소스 코드와 함께 판매하거나 소스코드를 무료로 배포해야 한다.
3. 컴퓨터 프로그램의 소스 코드를 용도에 따라 변경할 수 있다.
4. 변경된 컴퓨터 프로그램 역시 프로그램의 소스 코드를 반드시 공개 배포해야 한다.
5. 변경된 컴퓨터 프로그램 역시 반드시 똑같은 라이선스를 취해야 한다. 즉 GPL 라이선스를 적용해야 한다.

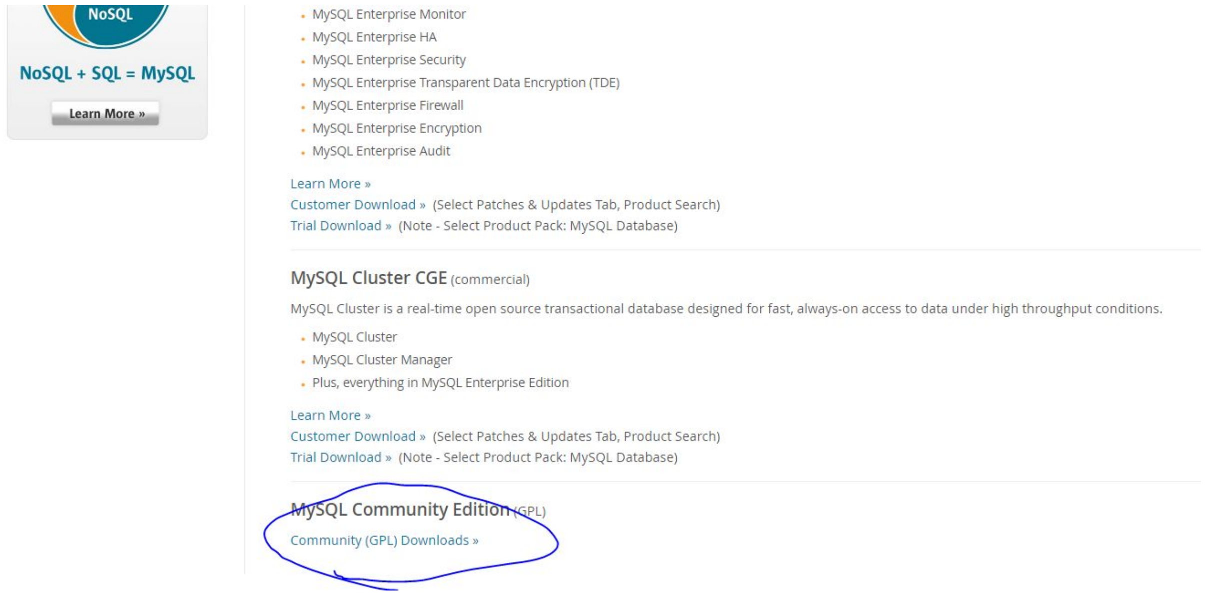
3) mysql 설치

"msvcp120.dll이 없어 프로그램을 시작할 수 없습니다." 라는 에러가 발생할 경우 Microsoft .NET Framework 프로그램을 설치해야 함

다운로드 주소 :

https://www.microsoft.com/ko-KR/download/details.aspx?id=42642

https://www.mysql.com/



NoSQL

NoSQL + SQL = MySQL

Learn More »

- MySQL Enterprise Monitor
- MySQL Enterprise HA
- MySQL Enterprise Security
- MySQL Enterprise Transparent Data Encryption (TDE)
- MySQL Enterprise Firewall
- MySQL Enterprise Encryption
- MySQL Enterprise Audit

Learn More »
Customer Download » (Select Patches & Updates Tab, Product Search)
Trial Download » (Note - Select Product Pack: MySQL Database)

MySQL Cluster CGE (commercial)

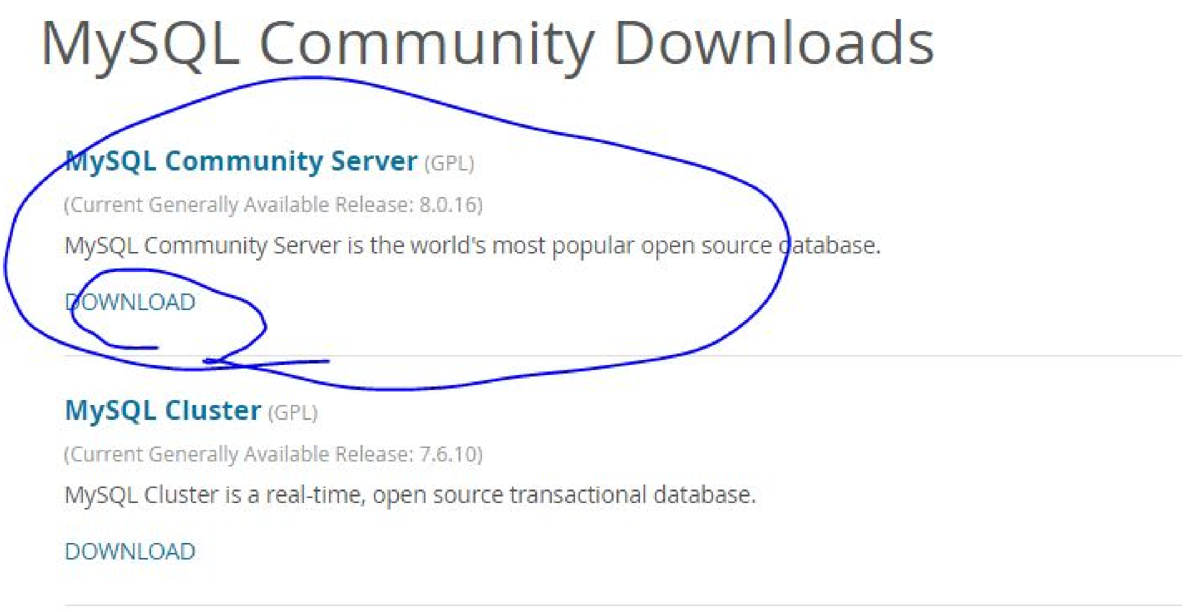
MySQL Cluster is a real-time open source transactional database designed for fast, always-on access to data under high throughput conditions.

- MySQL Cluster
- MySQL Cluster Manager
- Plus, everything in MySQL Enterprise Edition

Learn More »
Customer Download » (Select Patches & Updates Tab, Product Search)
Trial Download » (Note - Select Product Pack: MySQL Database)

MySQL Community Edition (GPL)

Community (GPL) Downloads »



MySQL Community Downloads

MySQL Community Server (GPL)

(Current Generally Available Release: 8.0.16)

MySQL Community Server is the world's most popular open source database.

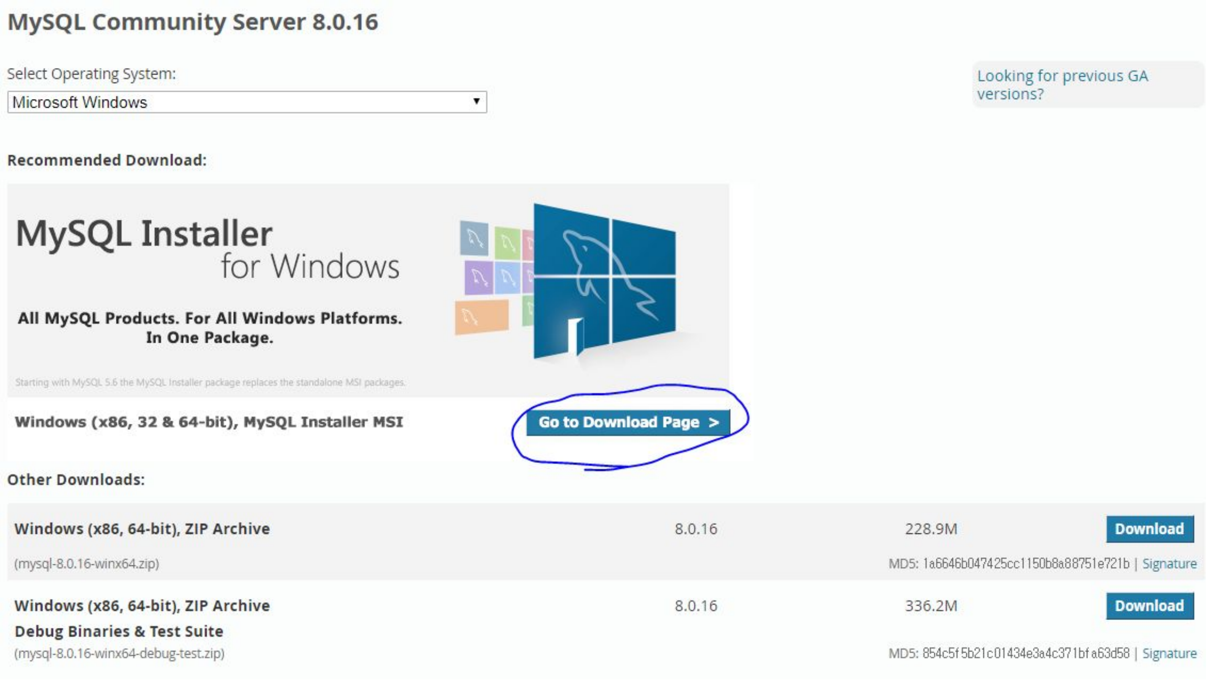
DOWNLOAD

MySQL Cluster (GPL)

(Current Generally Available Release: 7.6.10)

MySQL Cluster is a real-time, open source transactional database.

DOWNLOAD



MySQL Community Server 8.0.16

Select Operating System:
Microsoft Windows ▾

Looking for previous GA versions?

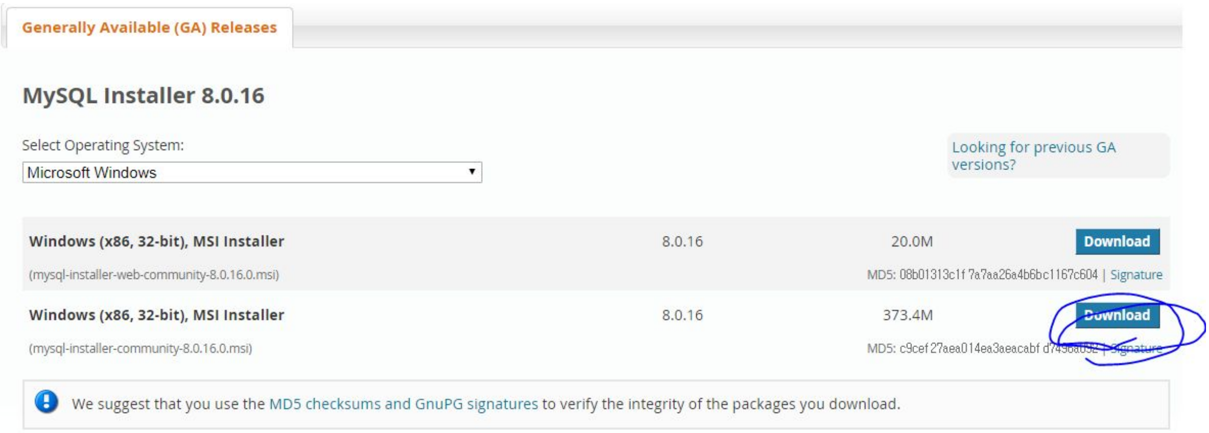
Recommended Download:

MySQL Installer
for Windows
All MySQL Products. For All Windows Platforms.
In One Package.
Starting with MySQL 5.6 the MySQL Installer package replaces the standalone MSI packages.

Windows (x86, 32 & 64-bit), MySQL Installer MSI Go to Download Page >

Other Downloads:

Windows (x86, 64-bit), ZIP Archive (mysql-8.0.16-winx64.zip)	8.0.16	228.9M MD5: 1a6646b047425cc1150b8a88751e721b | Signature	Download
Windows (x86, 64-bit), ZIP Archive **Debug Binaries & Test Suite** (mysql-8.0.16-winx64-debug-test.zip)	8.0.16	336.2M MD5: 854c5f5b21c01434e3a4c371bfa63d58 | Signature	Download



Generally Available (GA) Releases

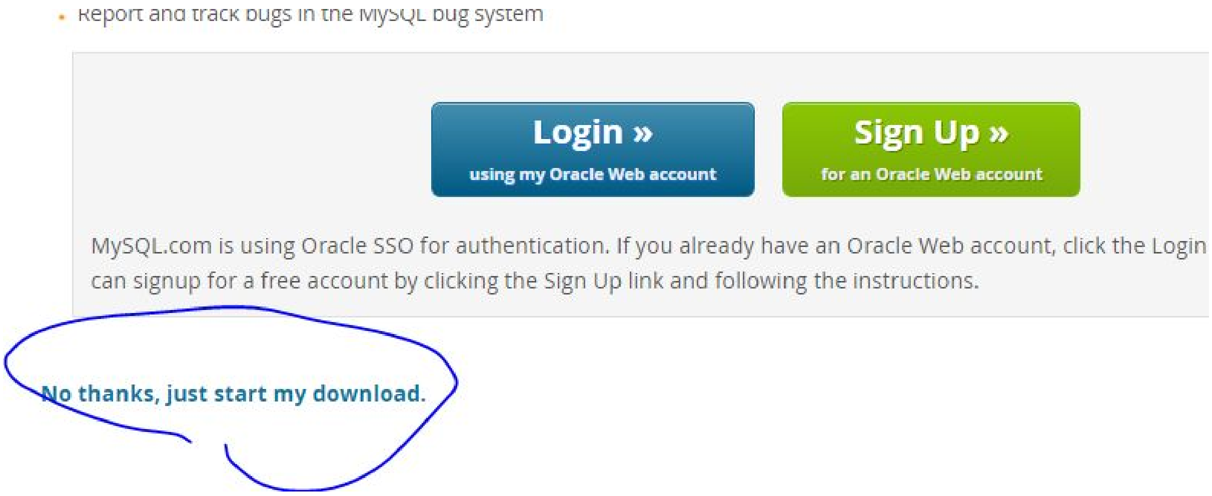
MySQL Installer 8.0.16

Select Operating System:
Microsoft Windows

Looking for previous GA versions?

Windows (x86, 32-bit), MSI Installer (mysql-installer-web-community-8.0.16.0.msi)	8.0.16	20.0M	Download MD5: 08b01313c1f7a7aa26a4b6bc1167c604 | Signature
Windows (x86, 32-bit), MSI Installer (mysql-installer-community-8.0.16.0.msi)	8.0.16	373.4M	Download MD5: c9cef27aea014ea3aeacabfd7496a092 | Signature

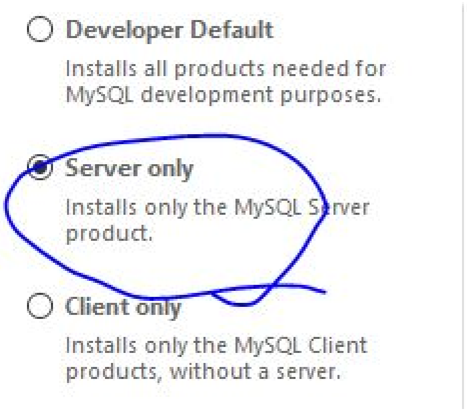
We suggest that you use the MD5 checksums and GnuPG signatures to verify the integrity of the packages you download.



- Report and track bugs in the MySQL bug system

MySQL.com is using Oracle SSO for authentication. If you already have an Oracle Web account, click the Login
can signup for a free account by clicking the Sign Up link and following the instructions.

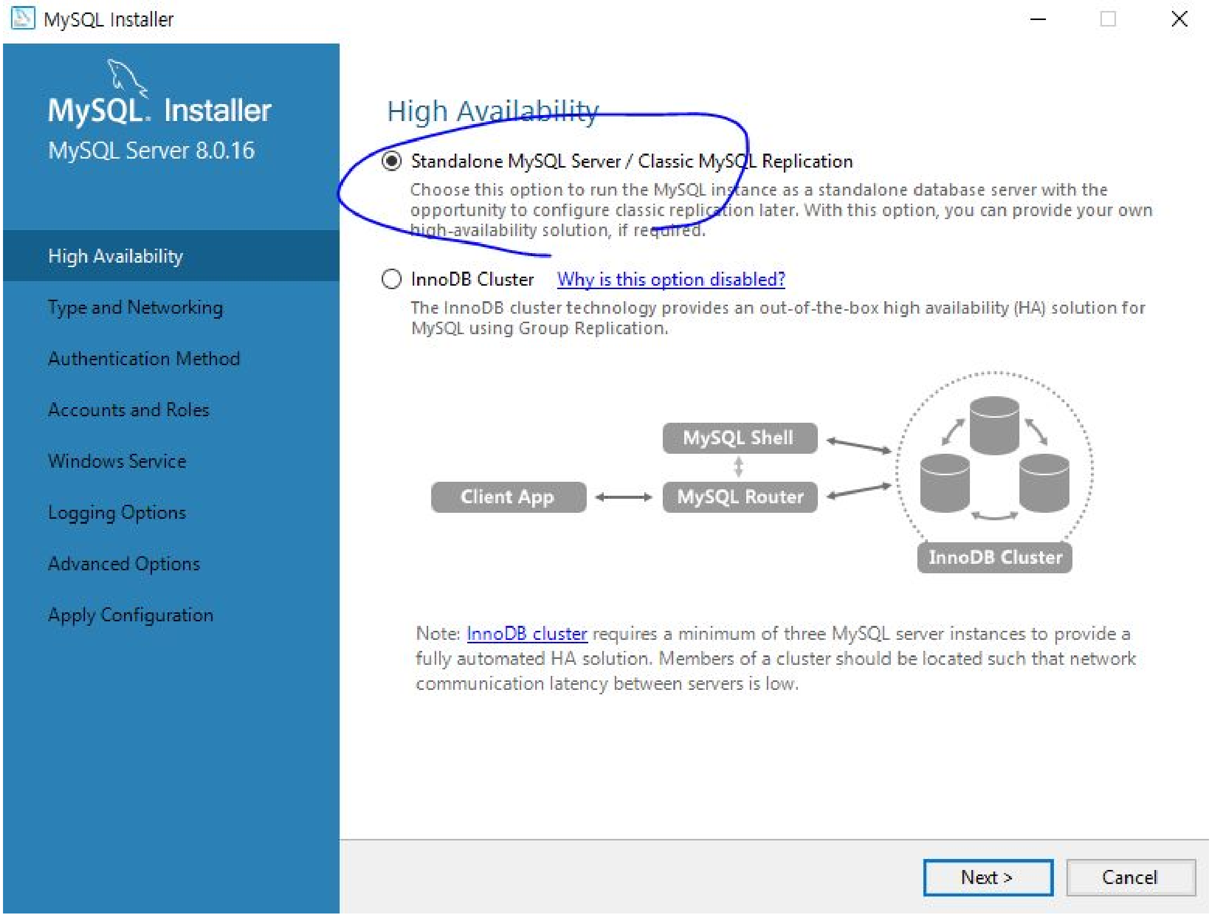
No thanks, just start my download.



○ **Developer Default**
Installs all products needed for MySQL development purposes.

● **Server only**
Installs only the MySQL Server product.

○ **Client only**
Installs only the MySQL Client products, without a server.



High Availability

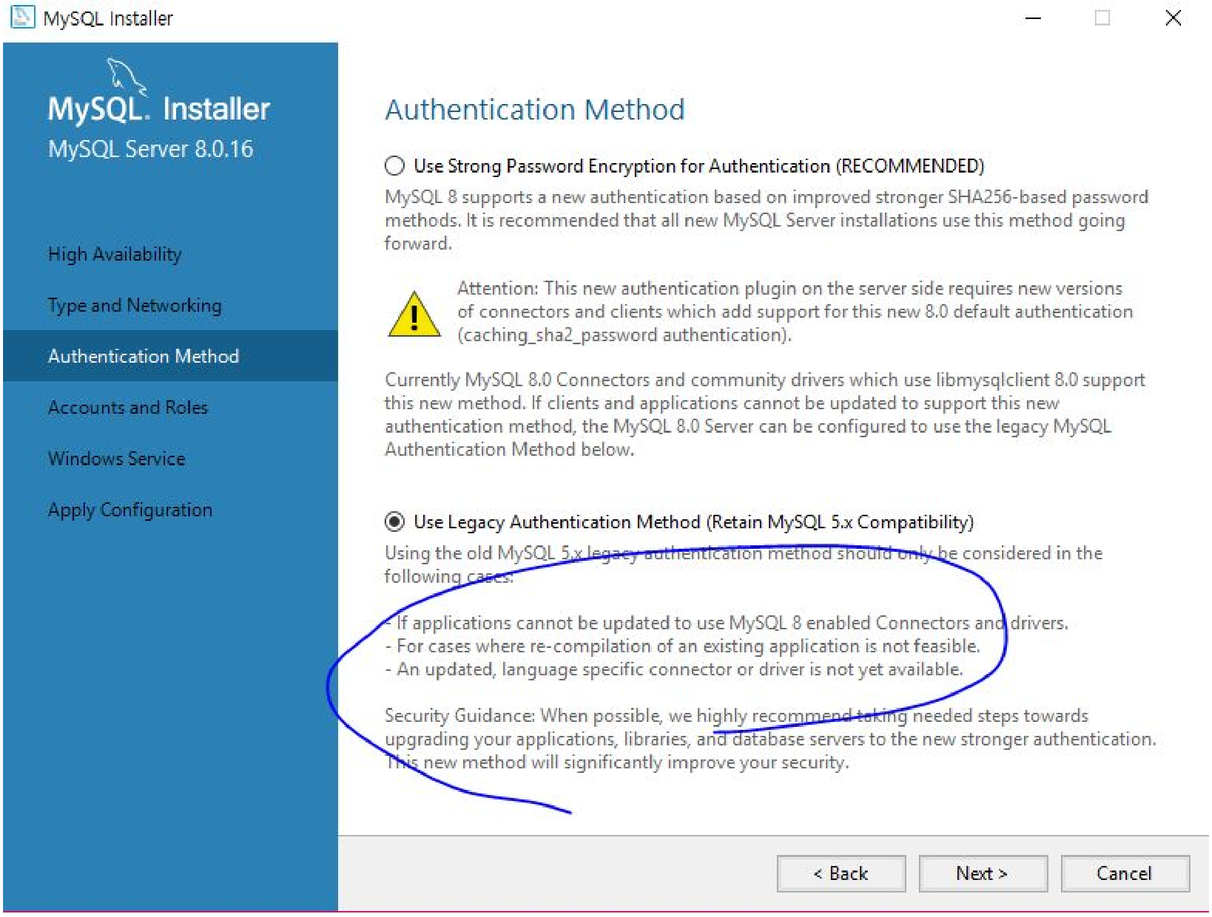
● Standalone MySQL Server / Classic MySQL Replication

Choose this option to run the MySQL instance as a standalone database server with the opportunity to configure classic replication later. With this option, you can provide your own high-availability solution, if required.

○ InnoDB Cluster Why is this option disabled?

The InnoDB cluster technology provides an out-of-the-box high availability (HA) solution for MySQL using Group Replication.

Note: InnoDB cluster requires a minimum of three MySQL server instances to provide a fully automated HA solution. Members of a cluster should be located such that network communication latency between servers is low.



Authentication Method

○ Use Strong Password Encryption for Authentication (RECOMMENDED)

MySQL 8 supports a new authentication based on improved stronger SHA256-based password methods. It is recommended that all new MySQL Server installations use this method going forward.

Attention: This new authentication plugin on the server side requires new versions of connectors and clients which add support for this new 8.0 default authentication (caching_sha2_password authentication).

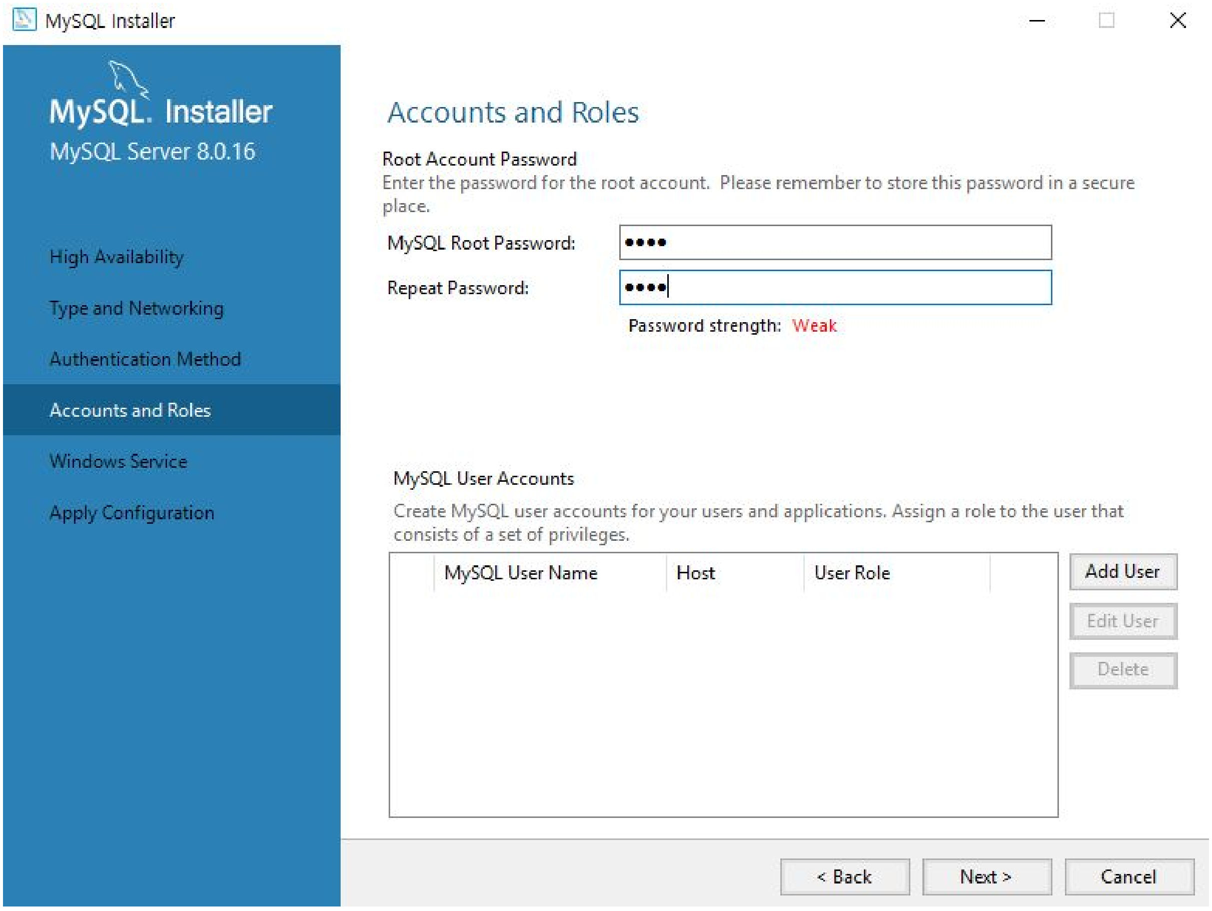
Currently MySQL 8.0 Connectors and community drivers which use libmysqlclient 8.0 support this new method. If clients and applications cannot be updated to support this new authentication method, the MySQL 8.0 Server can be configured to use the legacy MySQL Authentication Method below.

● Use Legacy Authentication Method (Retain MySQL 5.x Compatibility)

Using the old MySQL 5.x legacy authentication method should only be considered in the following cases:

- If applications cannot be updated to use MySQL 8 enabled Connectors and drivers.
- For cases where re-compilation of an existing application is not feasible.
- An updated, language specific connector or driver is not yet available.

Security Guidance: When possible, we highly recommend taking needed steps towards upgrading your applications, libraries, and database servers to the new stronger authentication. This new method will significantly improve your security.



Accounts and Roles

Root Account Password
Enter the password for the root account. Please remember to store this password in a secure place.

MySQL Root Password: ••••

Repeat Password: ••••

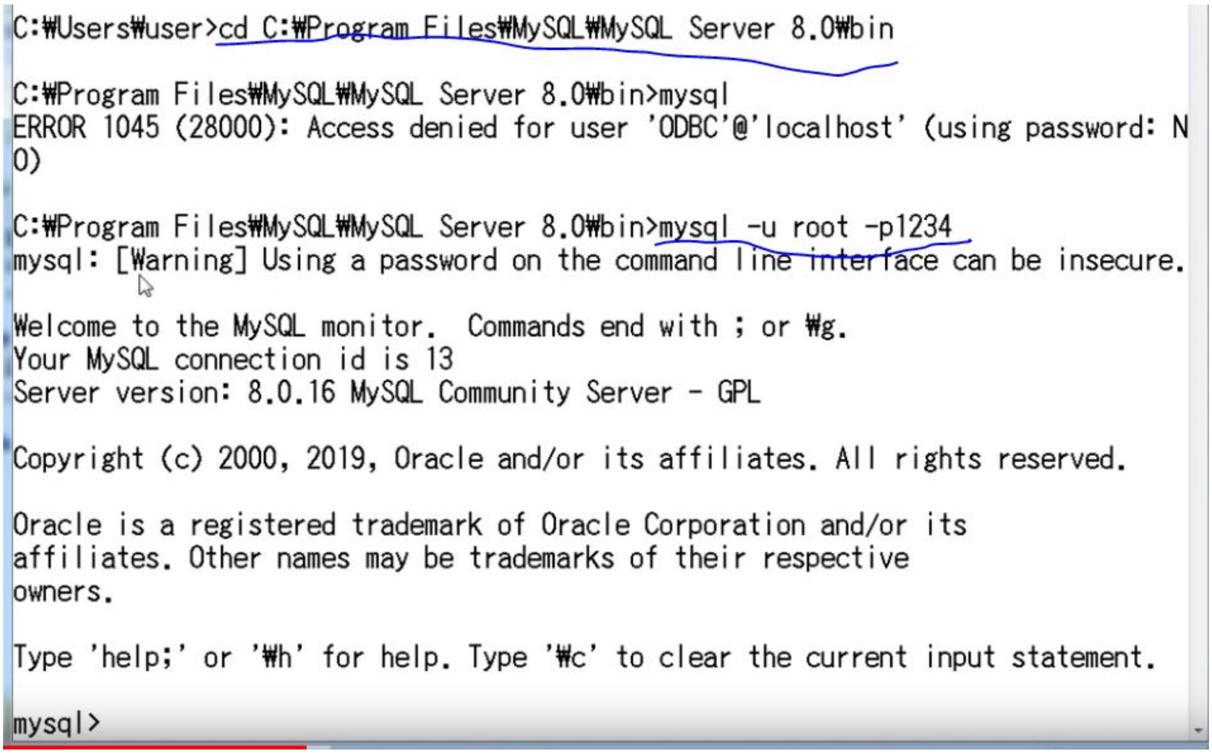
Password strength: Weak

MySQL User Accounts
Create MySQL user accounts for your users and applications. Assign a role to the user that consists of a set of privileges.

MySQL User Name	Host	User Role

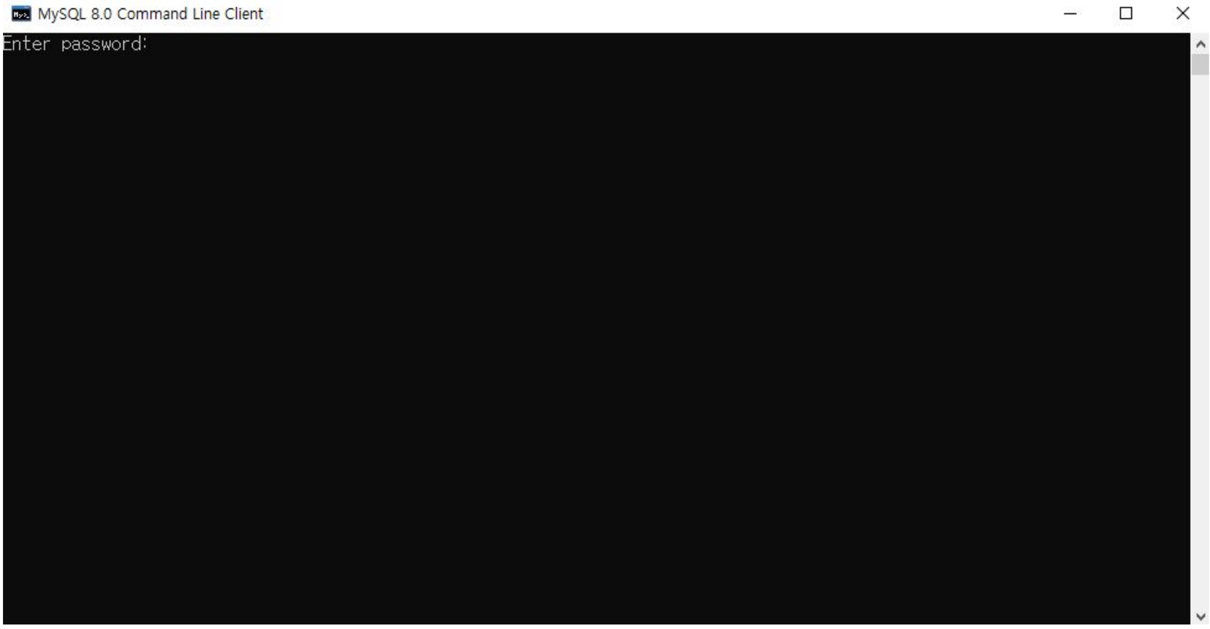
1234로 설정하자.

(cmd 접속방법)



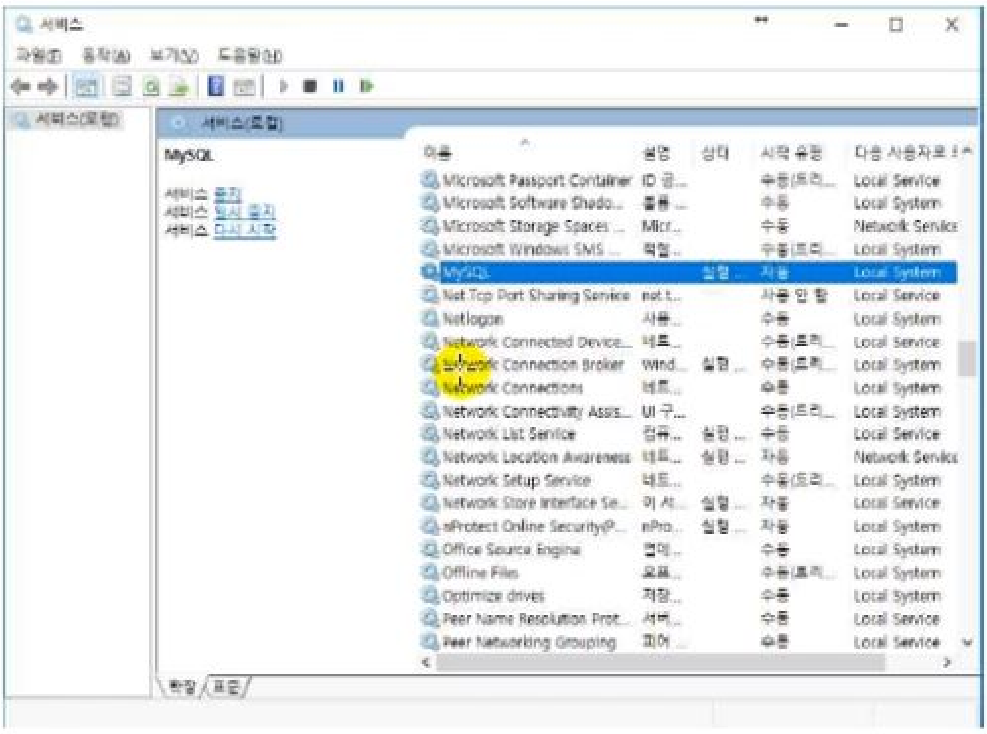
```
C:₩Users₩user>cd C:₩Program Files₩MySQL₩MySQL Server 8.0₩bin

C:₩Program Files₩MySQL₩MySQL Server 8.0₩bin>mysql
ERROR 1045 (28000): Access denied for user 'ODBC'@'localhost' (using password: N
O)

C:₩Program Files₩MySQL₩MySQL Server 8.0₩bin>mysql -u root -p1234
mysql: [Warning] Using a password on the command line interface can be insecure.

Welcome to the MySQL monitor.  Commands end with ; or ₩g.
Your MySQL connection id is 13
Server version: 8.0.16 MySQL Community Server - GPL

Copyright (c) 2000, 2019, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '₩h' for help. Type '₩c' to clear the current input statement.

mysql>
```

(프로그램으로 DB접속)



```
Enter password:
```

4) mysql 서비스 확인



MySQL

서비스 중지
서비스 일시 중지
서비스 다시 시작

이름	설명	상태	시작 유형	다음 사용자로 로그온
Microsoft Passport Container	ID 공...		수동(트리...	Local Service
Microsoft Software Shado...	볼륨 ...		수동	Local System
Microsoft Storage Spaces ...	Micr...		수동	Network Service
Microsoft Windows SMS ...	적절...		수동(트리...	Local System
MySQL		실행 ...	자동	Local System
Net.Tcp Port Sharing Service	net.t...		사용 안 함	Local Service
Netlogon	사용...		수동	Local System
Network Connected Device...	네트...		수동(트리...	Local Service
Network Connection Broker	Wind...	실행 ...	수동(트리...	Local System
Network Connections	네트...		수동	Local System
Network Connectivity Assis...	UI 구...		수동(트리...	Local System
Network List Service	컴퓨...	실행 ...	수동	Local Service
Network Location Awareness	네트...	실행 ...	자동	Network Service
Network Setup Service	네트...		수동(트리...	Local System
Network Store Interface Se...	이 서...	실행 ...	자동	Local Service
nProtect Online Security(P...	nPro...	실행 ...	자동	Local System
Office Source Engine	업데...		수동	Local System
Offline Files	오프...		수동(트리...	Local System
Optimize drives	저장...		수동	Local System
Peer Name Resolution Prot...	서버...		수동	Local Service
Peer Networking Grouping	피어 ...		수동	Local Service

5) HeidiSQL 설치

http://heidisql.com

6) Heidi SQL에서 사용자 계정 생성

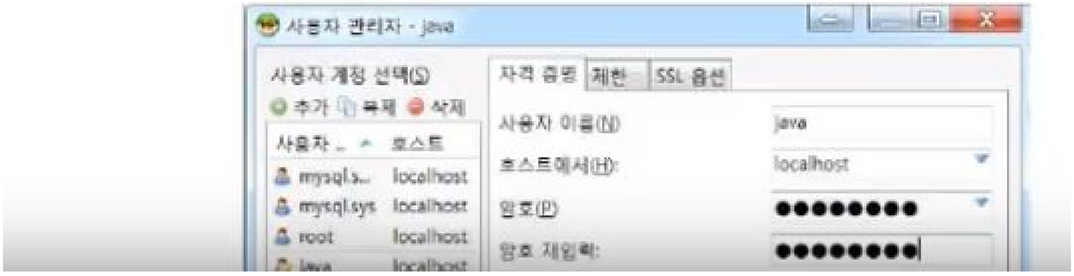
도구 - 사용자관리자 메뉴에서 사용자 추가

아이디 : java

비번 : java1234

호스트 : localhost

전체권한에 체크



사용자 관리자 - java

사용자 계정 선택(S)

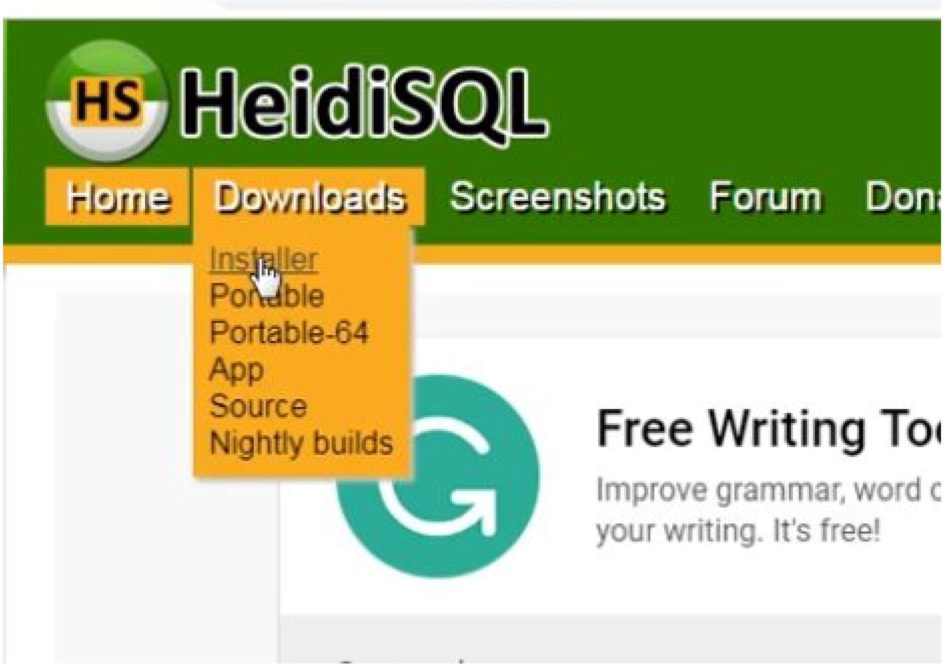
사용자	호스트
mysql.s...	localhost
mysql.sys	localhost
root	localhost
java	localhost

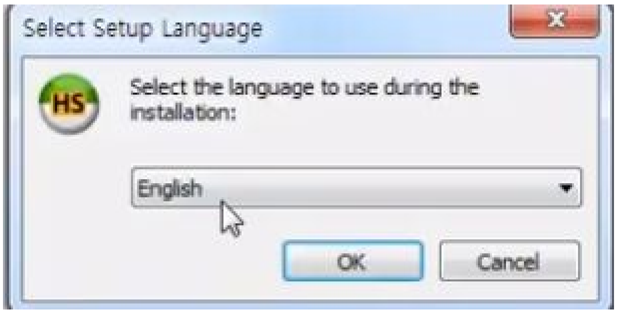
사용자 이름(N): java

호스트에서(H): localhost

암호(P): ●●●●●●●●●

암호 재입력: ●●●●●●●●●





Select Setup Language

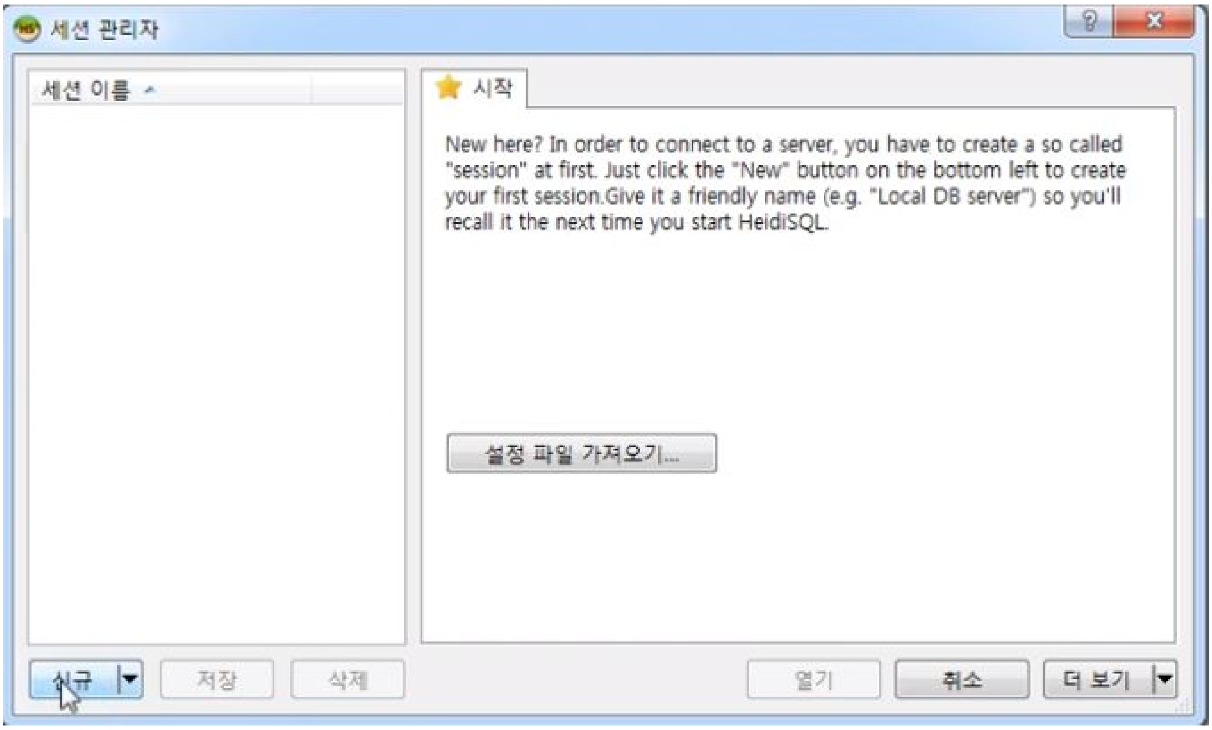
Select the language to use during the installation:

English

기본값으로 설치

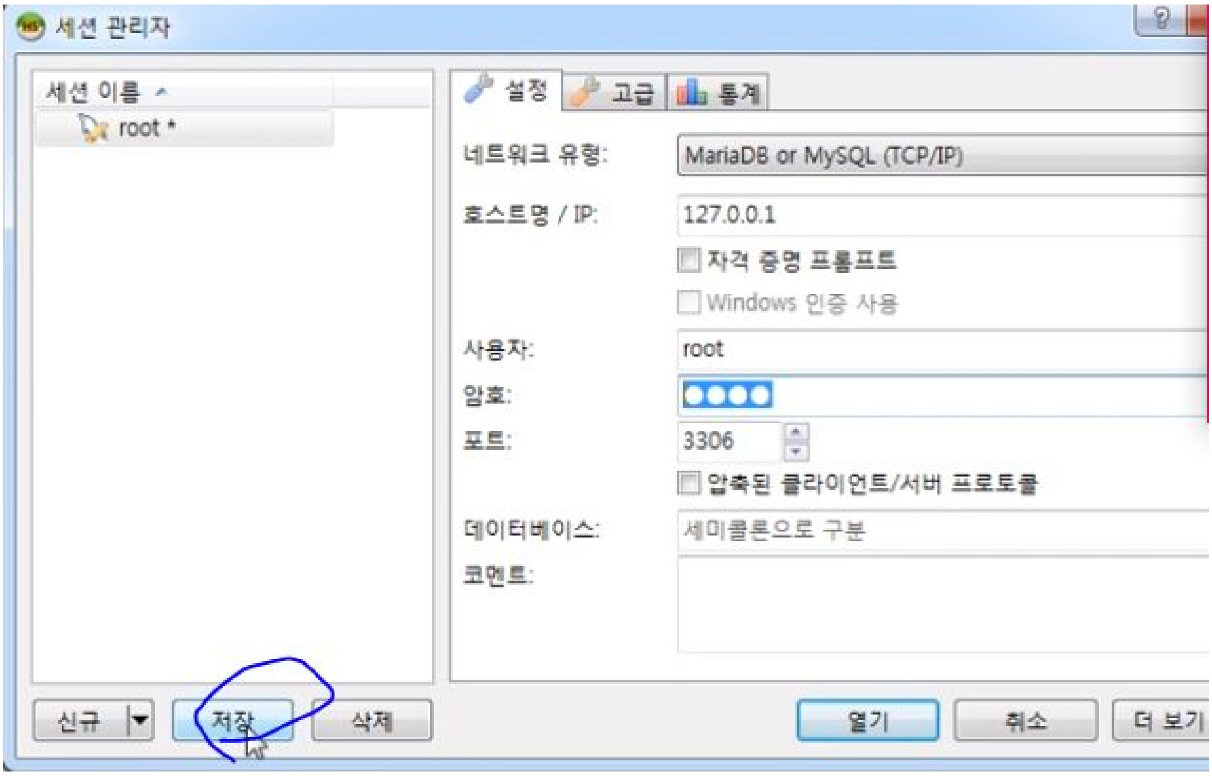
=>실행

mysql에 접속해야한다.



세션 관리자

New here? In order to connect to a server, you have to create a so called "session" at first. Just click the "New" button on the bottom left to create your first session.Give it a friendly name (e.g. "Local DB server") so you'll recall it the next time you start HeidiSQL.



네트워크 유형: MariaDB or MySQL (TCP/IP)

호스트명 / IP: 127.0.0.1

□ 자격 증명 프롬프트

□ Windows 인증 사용

사용자: root

암호: ●●●●

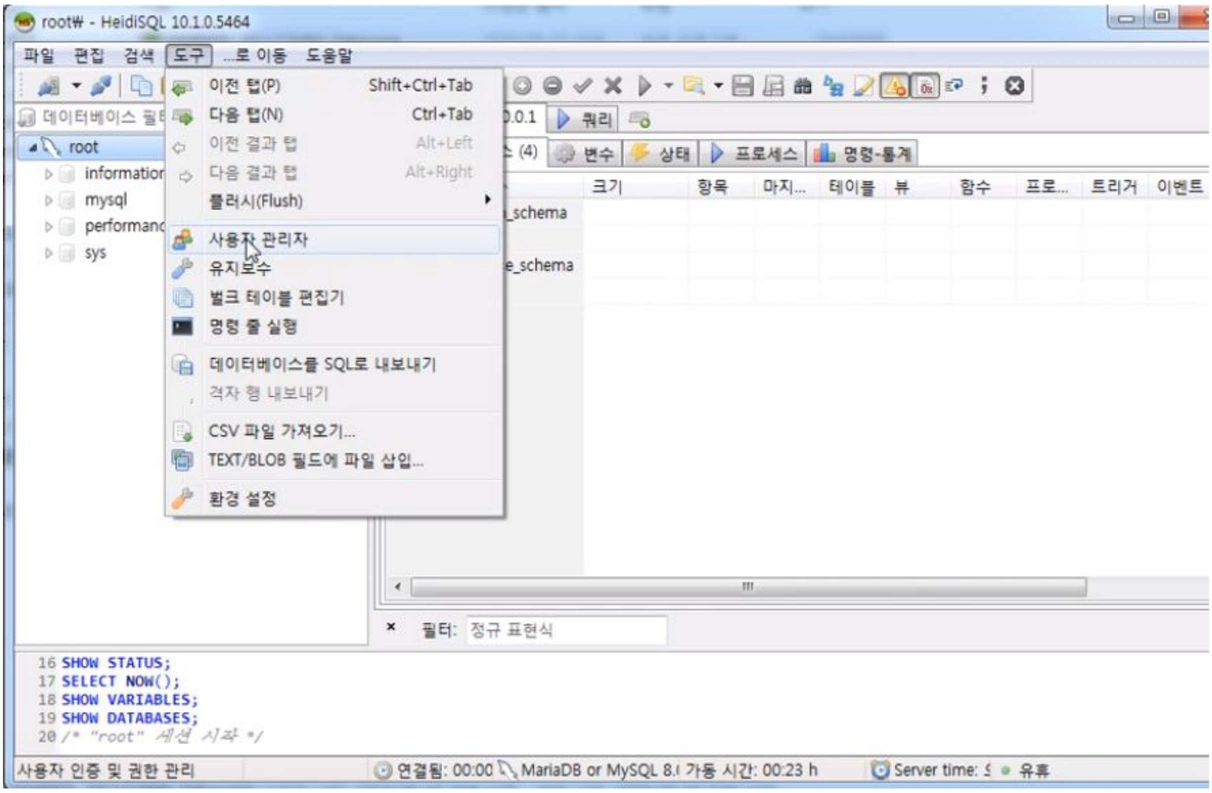
포트: 3306

□ 압축된 클라이언트/서버 프로토콜

데이터베이스: 세미콜론으로 구분

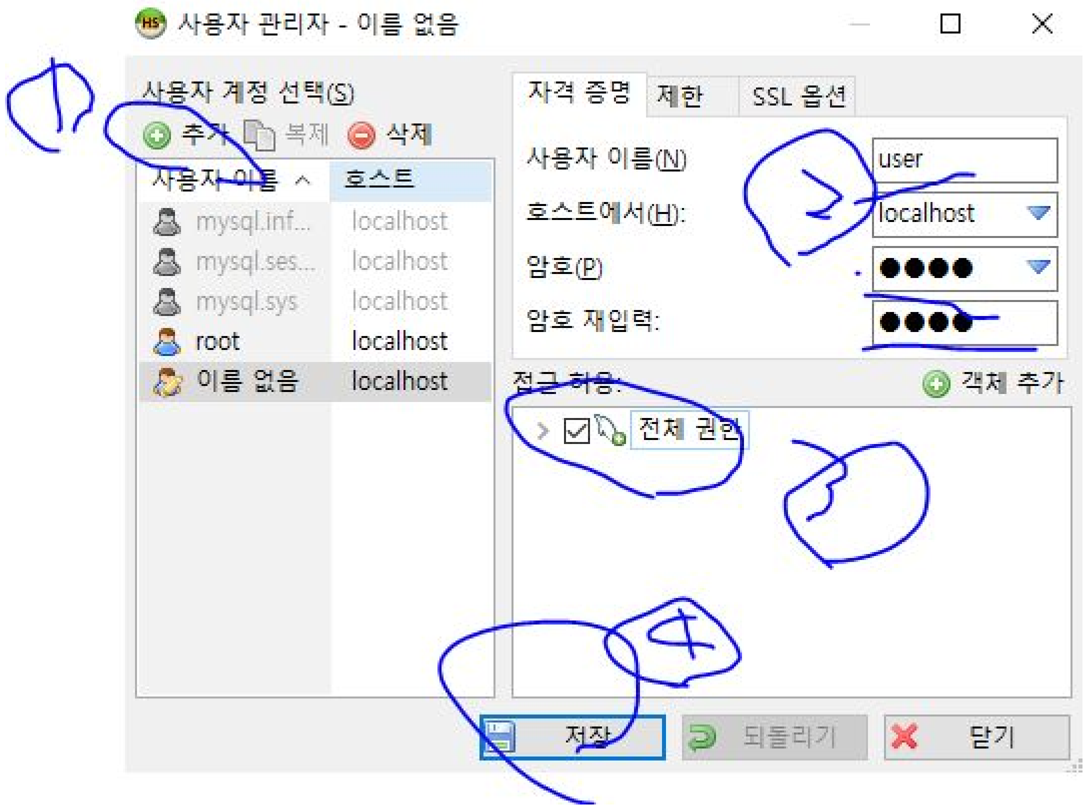
코멘트:

저장을 누르면 다시 비밀번호를 묻지 않는다.



```
SHOW STATUS;
SELECT NOW();
SHOW VARIABLES;
SHOW DATABASES;
/* "root" 세션 시작 */
```

root이외에 사용자 계정을 하나 만든다.



사용자 이름	호스트
mysql.inf...	localhost
mysql.ses...	localhost
mysql.sys	localhost
root	localhost
이름 없음	localhost

사용자 이름(N): user

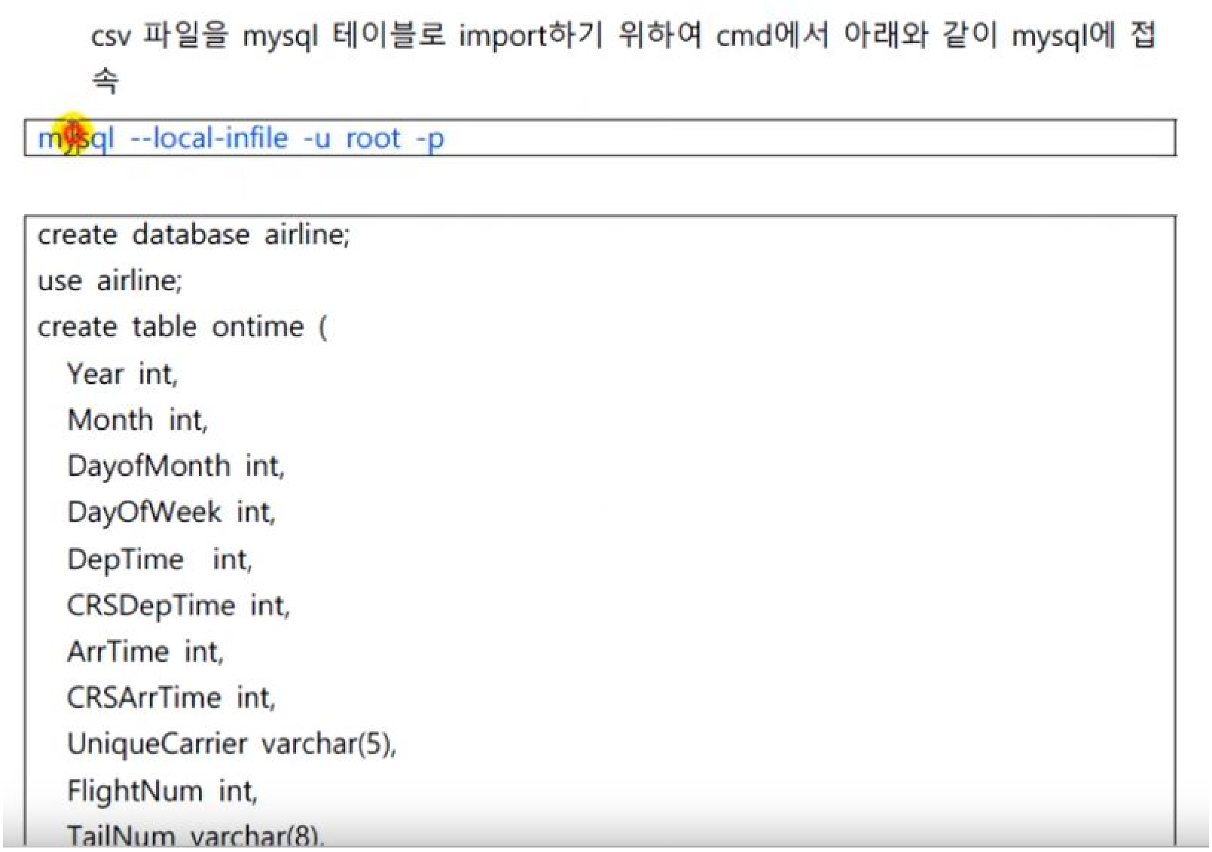
호스트에서(H): localhost

암호(P): ●●●●

암호 재입력: ●●●●

id:user
pw:user

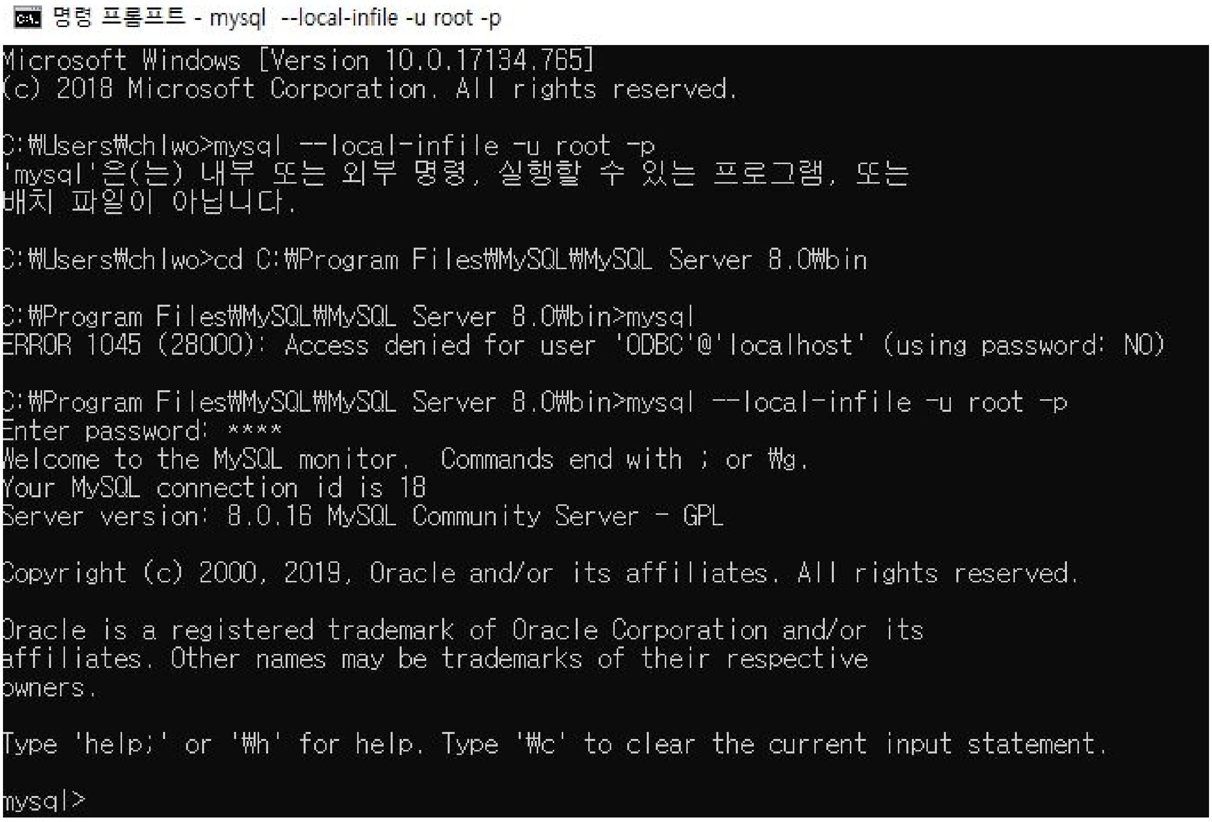
, id:java
pw:java1234



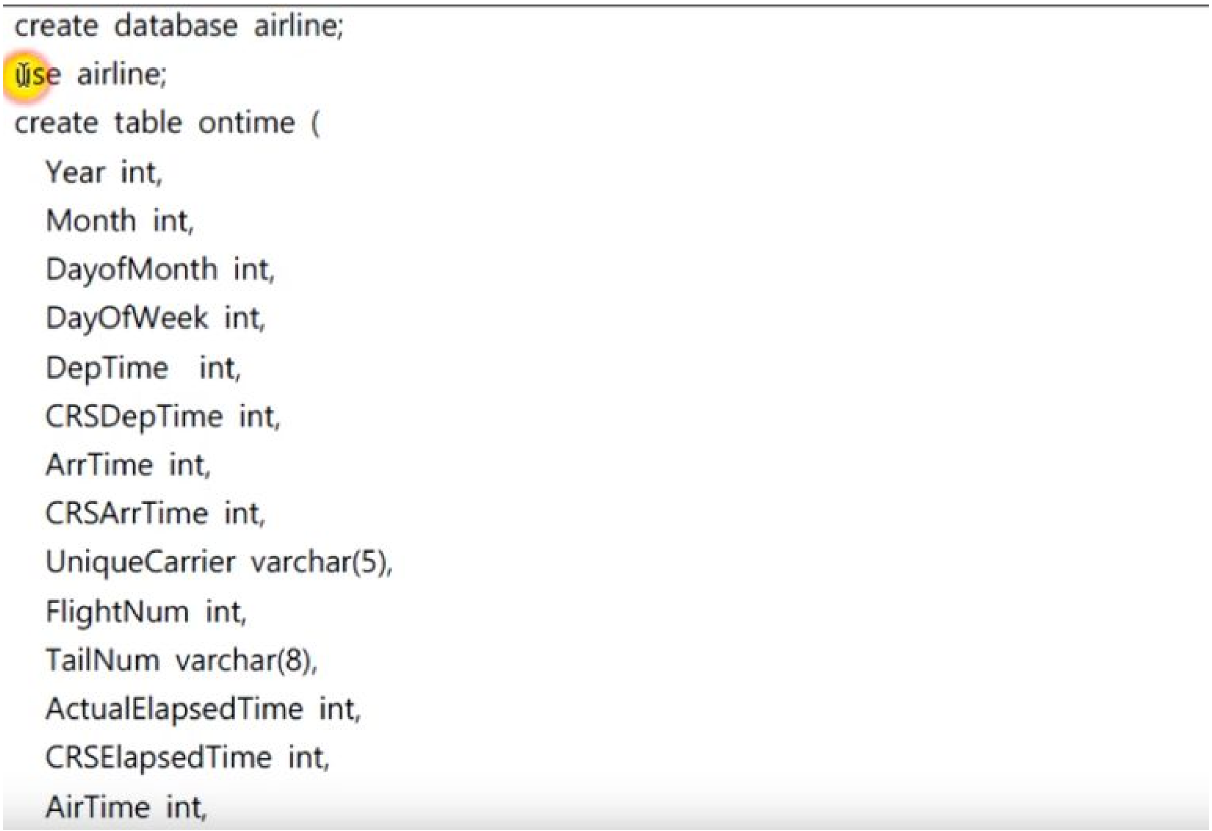
csv 파일을 mysql 테이블로 import하기 위하여 cmd에서 아래와 같이 mysql에 접속

```
mysql --local-infile -u root -p
```

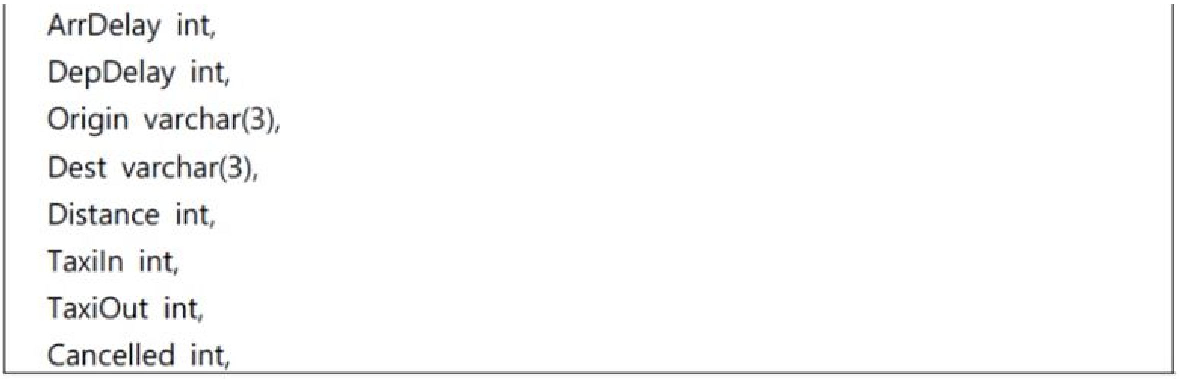
```
create database airline;
use airline;
create table ontime (
  Year int,
  Month int,
  DayofMonth int,
  DayOfWeek int,
  DepTime  int,
  CRSDepTime int,
  ArrTime int,
  CRSArrTime int,
  UniqueCarrier varchar(5),
  FlightNum int,
  TailNum varchar(8),
```



```
Microsoft Windows [Version 10.0.17134.765]
(c) 2018 Microsoft Corporation. All rights reserved.

C:₩Users₩chlwo>mysql --local-infile -u root -p
'mysql'은(는) 내부 또는 외부 명령, 실행할 수 있는 프로그램, 또는
배치 파일이 아닙니다.

C:₩Users₩chlwo>cd C:₩Program Files₩MySQL₩MySQL Server 8.0₩bin

C:₩Program Files₩MySQL₩MySQL Server 8.0₩bin>mysql
ERROR 1045 (28000): Access denied for user 'ODBC'@'localhost' (using password: NO)

C:₩Program Files₩MySQL₩MySQL Server 8.0₩bin>mysql --local-infile -u root -p
Enter password: ****
Welcome to the MySQL monitor.  Commands end with ; or ₩g.
Your MySQL connection id is 18
Server version: 8.0.16 MySQL Community Server - GPL

Copyright (c) 2000, 2019, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '₩h' for help. Type '₩c' to clear the current input statement.

mysql>
```

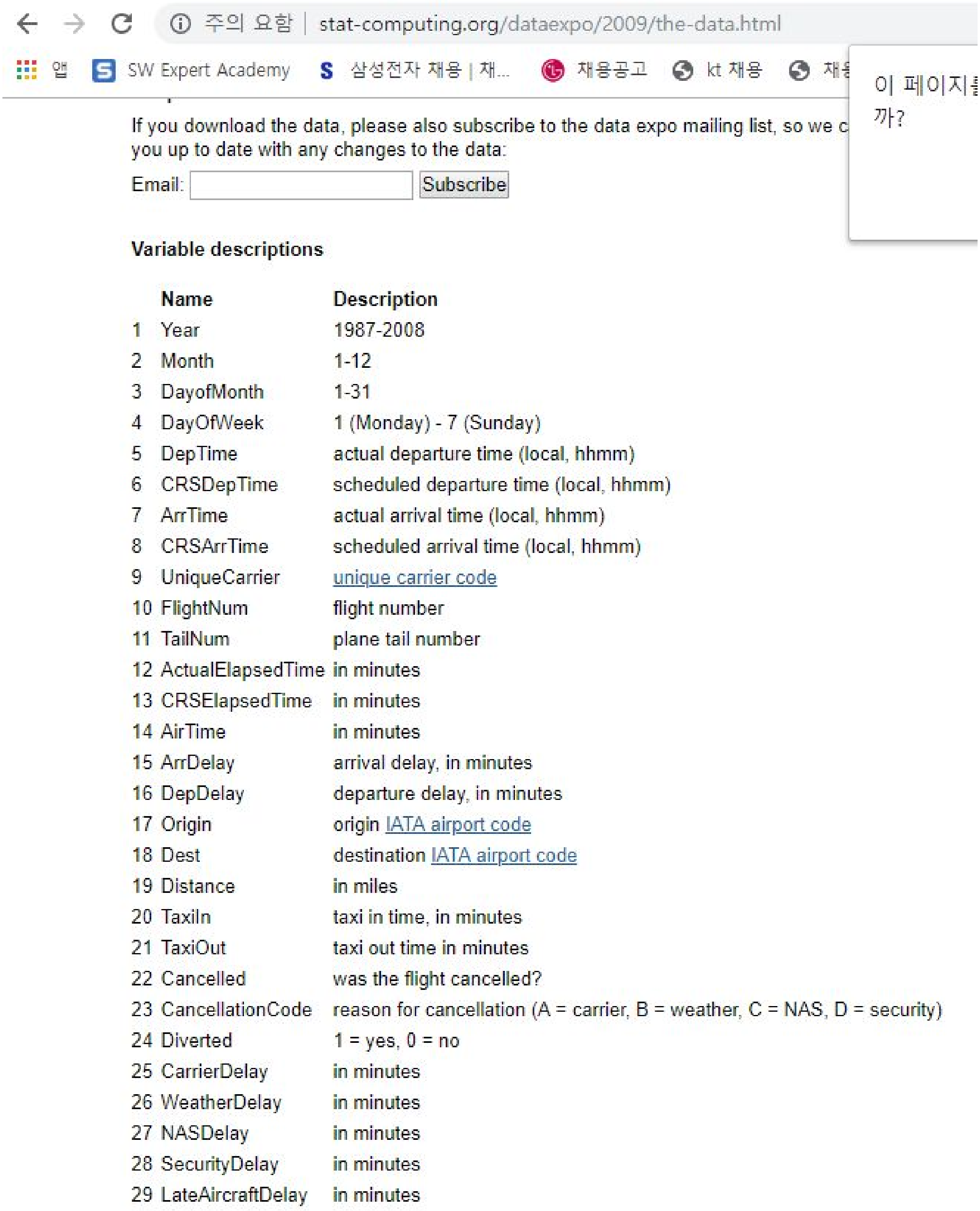


```
create database airline;
use airline;
create table ontime (
  Year int,
  Month int,
  DayofMonth int,
  DayOfWeek int,
  DepTime  int,
  CRSDepTime int,
  ArrTime int,
  CRSArrTime int,
  UniqueCarrier varchar(5),
  FlightNum int,
  TailNum varchar(8),
  ActualElapsedTime int,
  CRSElapsedTime int,
  AirTime int,
```



```
  ArrDelay int,
  DepDelay int,
  Origin varchar(3),
  Dest varchar(3),
  Distance int,
  TaxiIn int,
  TaxiOut int,
  Cancelled int,
```

```
create table ontime(
Year int,
Month int,
DayofMonth int,
DayOfWeek int,
DepTime int,
CRSDepTime int,
ArrTime int,
CRSArrTime int,
UniqueCarrier varchar(5),
FlightNum int,
TailNum varchar(8),
ActualElapsedTime int,
CRSElapsedTime int,
AirTime int,
ArrDelay int,
DepDelay int,
Origin varchar(3),
Dest varchar(3),
Distance int,
TaxiIn int,
TaxiOut int,
Cancelled int,
CancellationCode varchar(1),
Diverted varchar(1),
CarrierDelay int,
WeatherDelay int,
NASDelay int,
SecurityDelay int,
LateAircraftDelay int
);
```



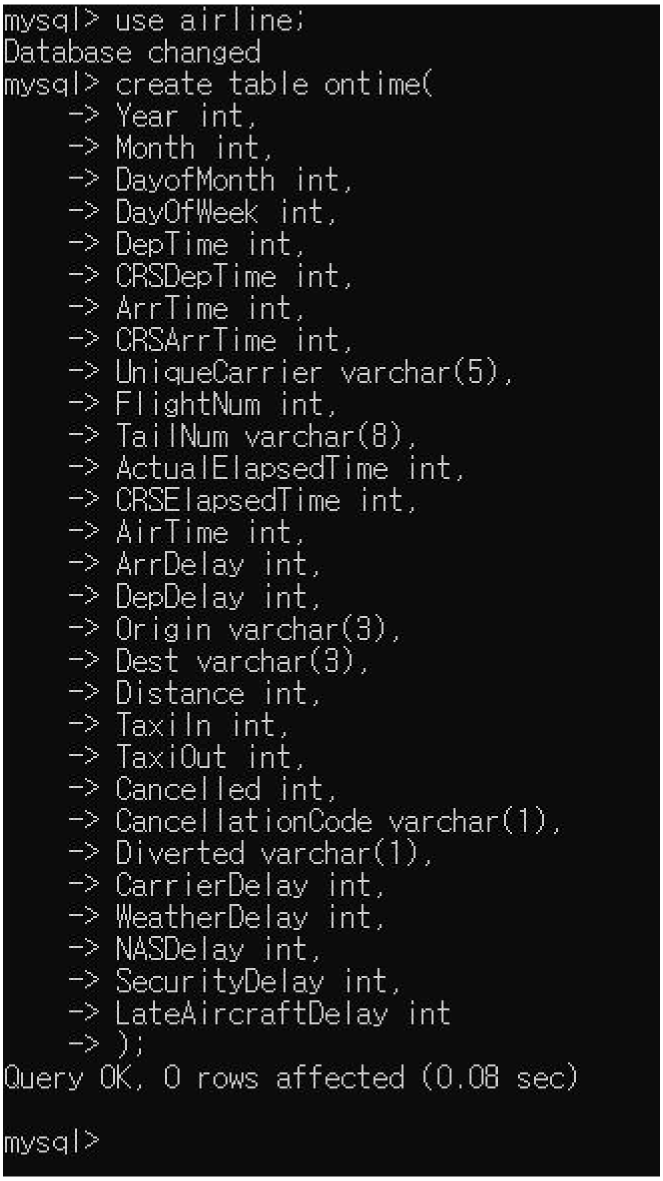
If you download the data, please also subscribe to the data expo mailing list, so we c you up to date with any changes to the data:

Email: Subscribe

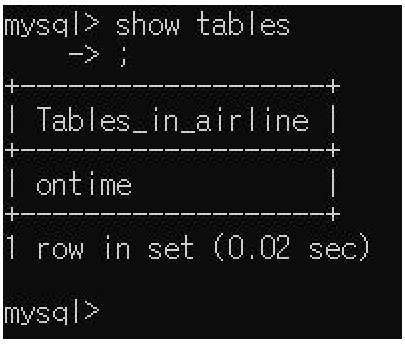
Variable descriptions

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
17	Origin	origin IATA airport code
18	Dest	destination IATA airport code
19	Distance	in miles
20	TaxiIn	taxi in time, in minutes
21	TaxiOut	taxi out time in minutes
22	Cancelled	was the flight cancelled?
23	CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24	Diverted	1 = yes, 0 = no
25	CarrierDelay	in minutes
26	WeatherDelay	in minutes
27	NASDelay	in minutes
28	SecurityDelay	in minutes
29	LateAircraftDelay	in minutes

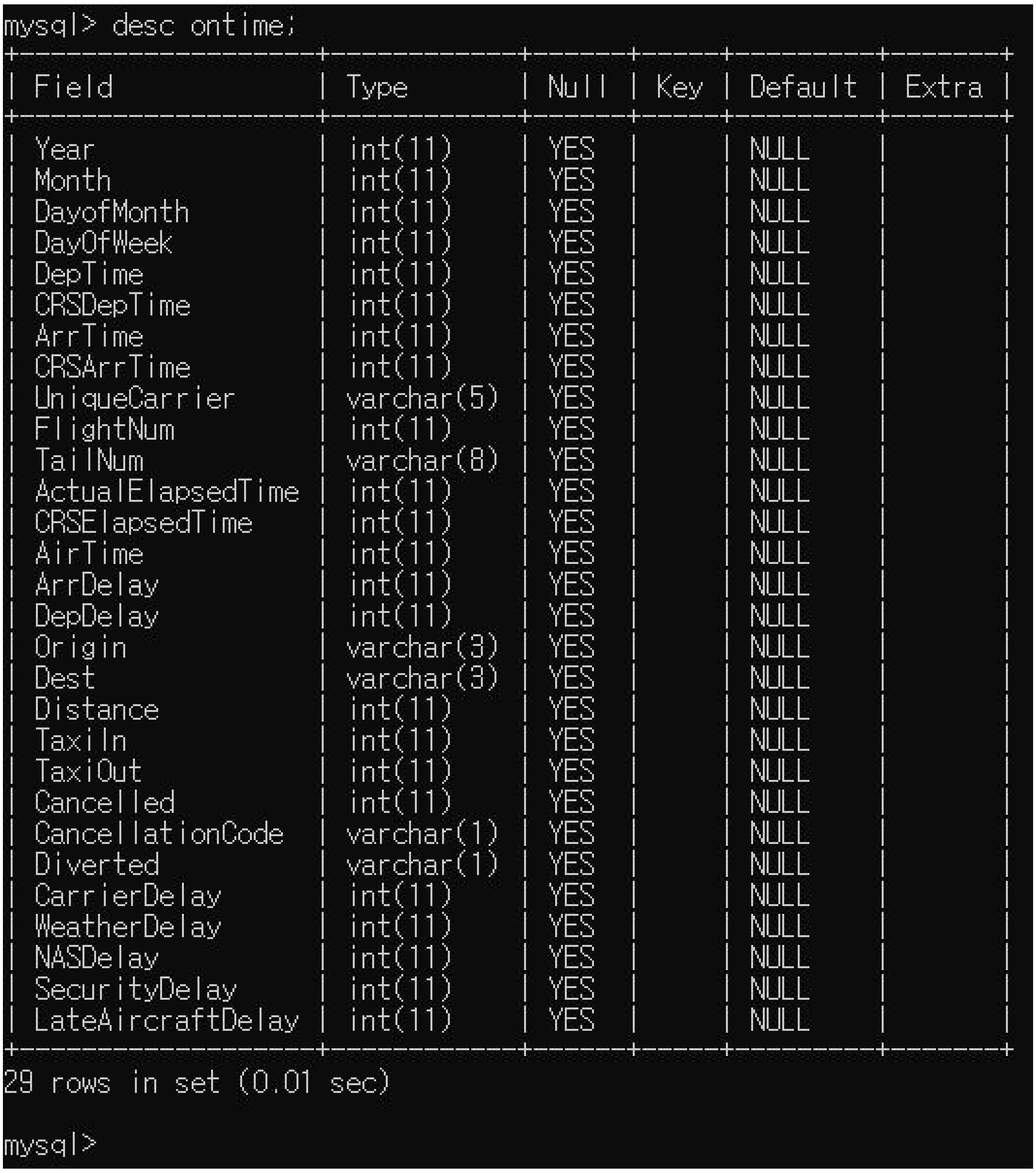
사이트에 나와있는 다음 필드명들을 가져온것이다.



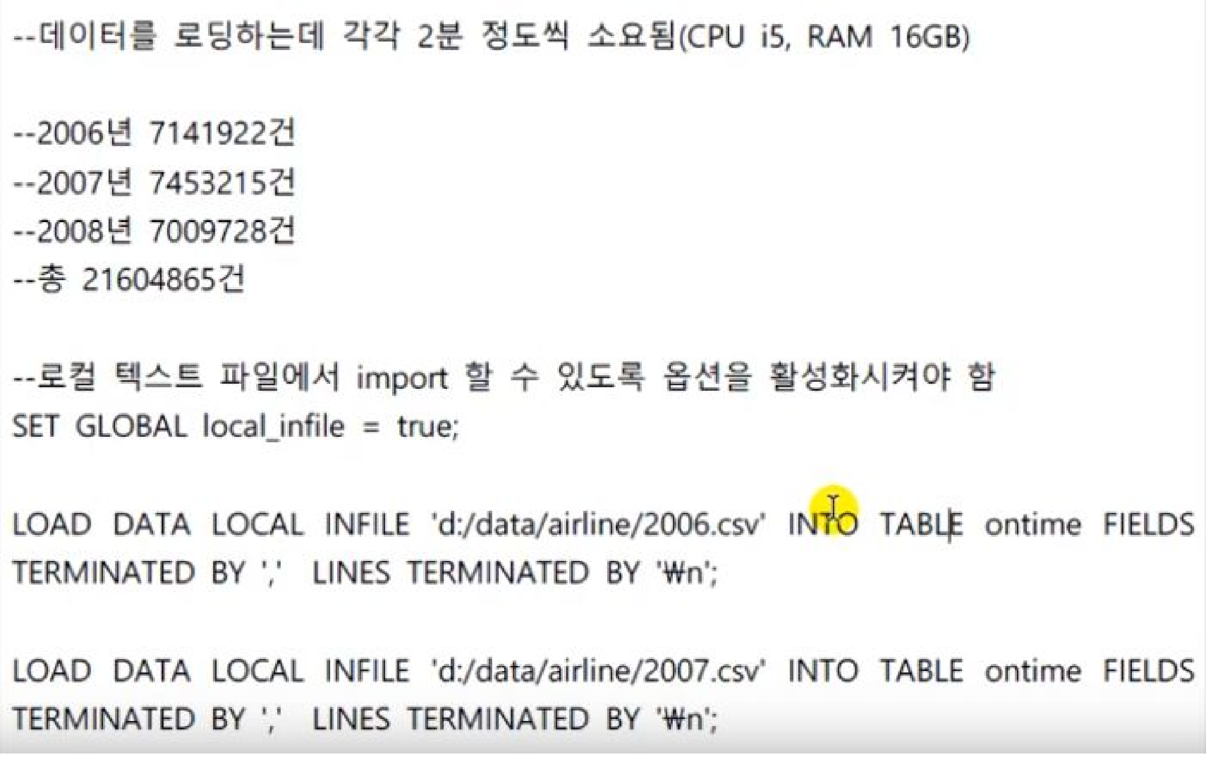
```
mysql> use airline;
Database changed
mysql> create table ontime(
    -> Year int,
    -> Month int,
    -> DayofMonth int,
    -> DayOfWeek int,
    -> DepTime int,
    -> CRSDepTime int,
    -> ArrTime int,
    -> CRSArrTime int,
    -> UniqueCarrier varchar(5),
    -> FlightNum int,
    -> TailNum varchar(8),
    -> ActualElapsedTime int,
    -> CRSElapsedTime int,
    -> AirTime int,
    -> ArrDelay int,
    -> DepDelay int,
    -> Origin varchar(3),
    -> Dest varchar(3),
    -> Distance int,
    -> TaxiIn int,
    -> TaxiOut int,
    -> Cancelled int,
    -> CancellationCode varchar(1),
    -> Diverted varchar(1),
    -> CarrierDelay int,
    -> WeatherDelay int,
    -> NASDelay int,
    -> SecurityDelay int,
    -> LateAircraftDelay int
    -> );
Query OK, 0 rows affected (0.08 sec)

mysql>
```



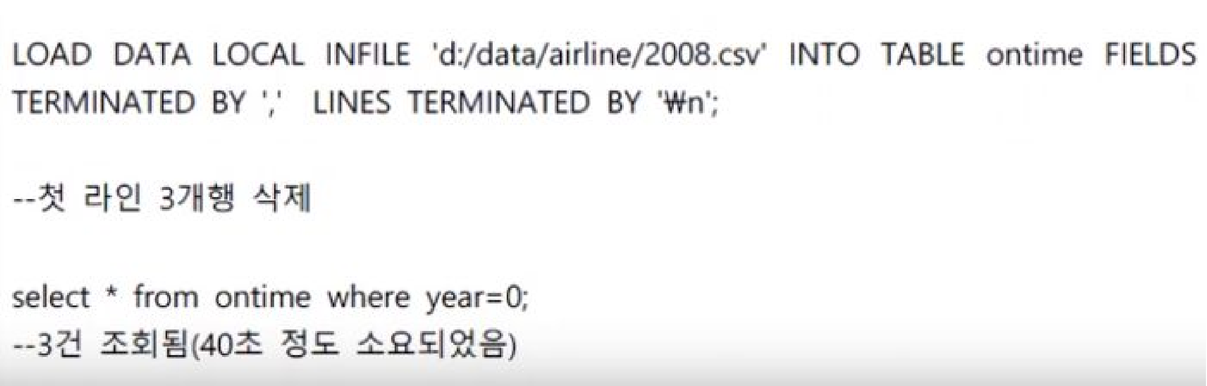
```
mysql> show tables
    -> ;
+--------------------+
| Tables_in_airline |
+--------------------+
| ontime             |
+--------------------+
1 row in set (0.02 sec)

mysql>
```



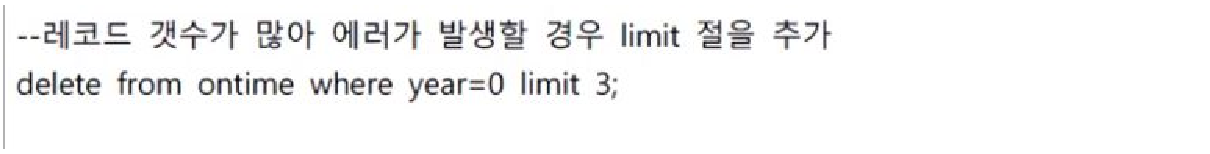
```
mysql> desc ontime;
+-------------------+------------+------+-----+---------+-------+
| Field             | Type       | Null | Key | Default | Extra |
+-------------------+------------+------+-----+---------+-------+
| Year              | int(11)    | YES  |     | NULL    |       |
| Month             | int(11)    | YES  |     | NULL    |       |
| DayofMonth        | int(11)    | YES  |     | NULL    |       |
| DayOfWeek         | int(11)    | YES  |     | NULL    |       |
| DepTime           | int(11)    | YES  |     | NULL    |       |
| CRSDepTime        | int(11)    | YES  |     | NULL    |       |
| ArrTime           | int(11)    | YES  |     | NULL    |       |
| CRSArrTime        | int(11)    | YES  |     | NULL    |       |
| UniqueCarrier     | varchar(5) | YES  |     | NULL    |       |
| FlightNum         | int(11)    | YES  |     | NULL    |       |
| TailNum           | varchar(8) | YES  |     | NULL    |       |
| ActualElapsedTime | int(11)    | YES  |     | NULL    |       |
| CRSElapsedTime    | int(11)    | YES  |     | NULL    |       |
| AirTime           | int(11)    | YES  |     | NULL    |       |
| ArrDelay          | int(11)    | YES  |     | NULL    |       |
| DepDelay          | int(11)    | YES  |     | NULL    |       |
| Origin            | varchar(3) | YES  |     | NULL    |       |
| Dest              | varchar(3) | YES  |     | NULL    |       |
| Distance          | int(11)    | YES  |     | NULL    |       |
| TaxiIn            | int(11)    | YES  |     | NULL    |       |
| TaxiOut           | int(11)    | YES  |     | NULL    |       |
| Cancelled         | int(11)    | YES  |     | NULL    |       |
| CancellationCode  | varchar(1) | YES  |     | NULL    |       |
| Diverted          | varchar(1) | YES  |     | NULL    |       |
| CarrierDelay      | int(11)    | YES  |     | NULL    |       |
| WeatherDelay      | int(11)    | YES  |     | NULL    |       |
| NASDelay          | int(11)    | YES  |     | NULL    |       |
| SecurityDelay     | int(11)    | YES  |     | NULL    |       |
| LateAircraftDelay | int(11)    | YES  |     | NULL    |       |
+-------------------+------------+------+-----+---------+-------+
29 rows in set (0.01 sec)

mysql>
```



```
--데이터를 로딩하는데 각각 2분 정도씩 소요됨(CPU i5, RAM 16GB)

--2006년 7141922건
--2007년 7453215건
--2008년 7009728건
--총 21604865건

--로컬 텍스트 파일에서 import 할 수 있도록 옵션을 활성화시켜야 함
SET GLOBAL local_infile = true;

LOAD DATA LOCAL INFILE 'd:/data/airline/2006.csv' INTO TABLE ontime FIELDS
TERMINATED BY ','  LINES TERMINATED BY '₩n';

LOAD DATA LOCAL INFILE 'd:/data/airline/2007.csv' INTO TABLE ontime FIELDS
TERMINATED BY ','  LINES TERMINATED BY '₩n';
```

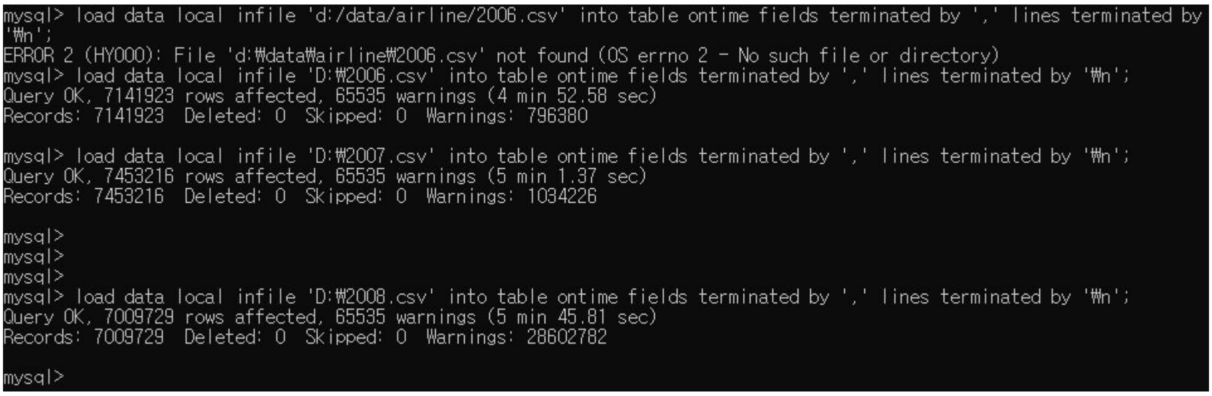


```
LOAD DATA LOCAL INFILE 'd:/data/airline/2008.csv' INTO TABLE ontime FIELDS
TERMINATED BY ','  LINES TERMINATED BY '₩n';

--첫 라인 3개행 삭제

select * from ontime where year=0;
--3건 조회됨(40초 정도 소요되었음)
```

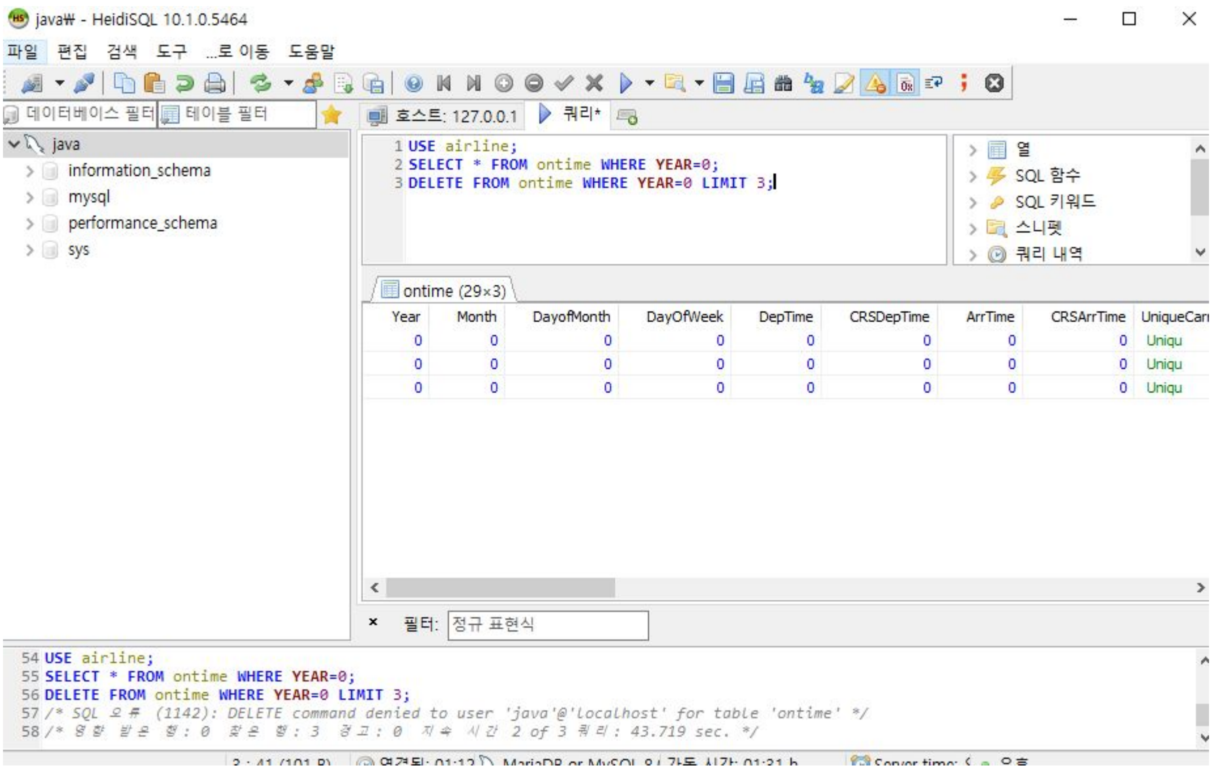


```
--레코드 갯수가 많아 에러가 발생할 경우 limit 절을 추가
delete from ontime where year=0 limit 3;
```

```
set global local_infile =true;

load data local infile 'D:\2006.csv' into table ontime fields terminated by ',' lines terminated by '\n';

load data local infile 'D:\2007.csv' into table ontime fields terminated by ',' lines terminated by '\n';

load data local infile 'D:\2008.csv' into table ontime fields terminated by ',' lines terminated by '\n';
```



```
mysql> load data local infile 'd:/data/airline/2006.csv' into table ontime fields terminated by ',' lines terminated by
'₩n';
ERROR 2 (HY000): File 'd:₩data₩airline₩2006.csv' not found (OS errno 2 - No such file or directory)
mysql> load data local infile 'D:₩2006.csv' into table ontime fields terminated by ',' lines terminated by '₩n';
Query OK, 7141923 rows affected, 65535 warnings (4 min 52.58 sec)
Records: 7141923  Deleted: 0  Skipped: 0  Warnings: 796380

mysql> load data local infile 'D:₩2007.csv' into table ontime fields terminated by ',' lines terminated by '₩n';
Query OK, 7453216 rows affected, 65535 warnings (5 min 1.37 sec)
Records: 7453216  Deleted: 0  Skipped: 0  Warnings: 1034226

mysql>
mysql>
mysql>
mysql> load data local infile 'D:₩2008.csv' into table ontime fields terminated by ',' lines terminated by '₩n';
Query OK, 7009729 rows affected, 65535 warnings (5 min 45.81 sec)
Records: 7009729  Deleted: 0  Skipped: 0  Warnings: 28602782

mysql>
```



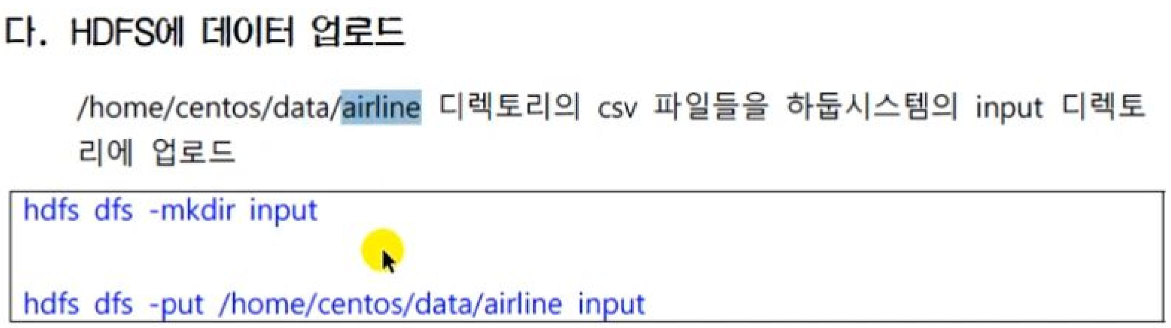
```
USE airline;
SELECT * FROM ontime WHERE YEAR=0;
DELETE FROM ontime WHERE YEAR=0 LIMIT 3;
```

Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarr
0	0	0	0	0	0	0	0	Uniqu
0	0	0	0	0	0	0	0	Uniqu
0	0	0	0	0	0	0	0	Uniqu

```
USE airline;
SELECT * FROM ontime WHERE YEAR=0;
DELETE FROM ontime WHERE YEAR=0 LIMIT 3;
/* SQL 오류 (1142): DELETE command denied to user 'java'@'localhost' for table 'ontime' */
/* 영향 받은 행: 0  찾은 행: 3  경고: 0  지속 시간 2 of 3 쿼리: 43.719 sec. */
```

delete를 할때 limit 3을 넣어줘도 메모리 초과가 뜬다.
따라서 너무 데이터가 크면, 연산이 불가한 상황이 발생할 수 있는 것이다.

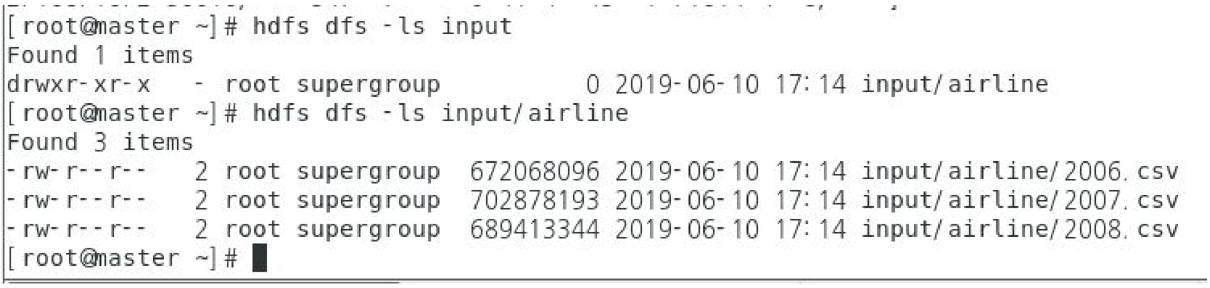
따라서 hadoop을 이용해서 처리를 해보려고한다.



다. HDFS에 데이터 업로드

/home/centos/data/airline 디렉토리의 csv 파일들을 하둡시스템의 input 디렉토리에 업로드

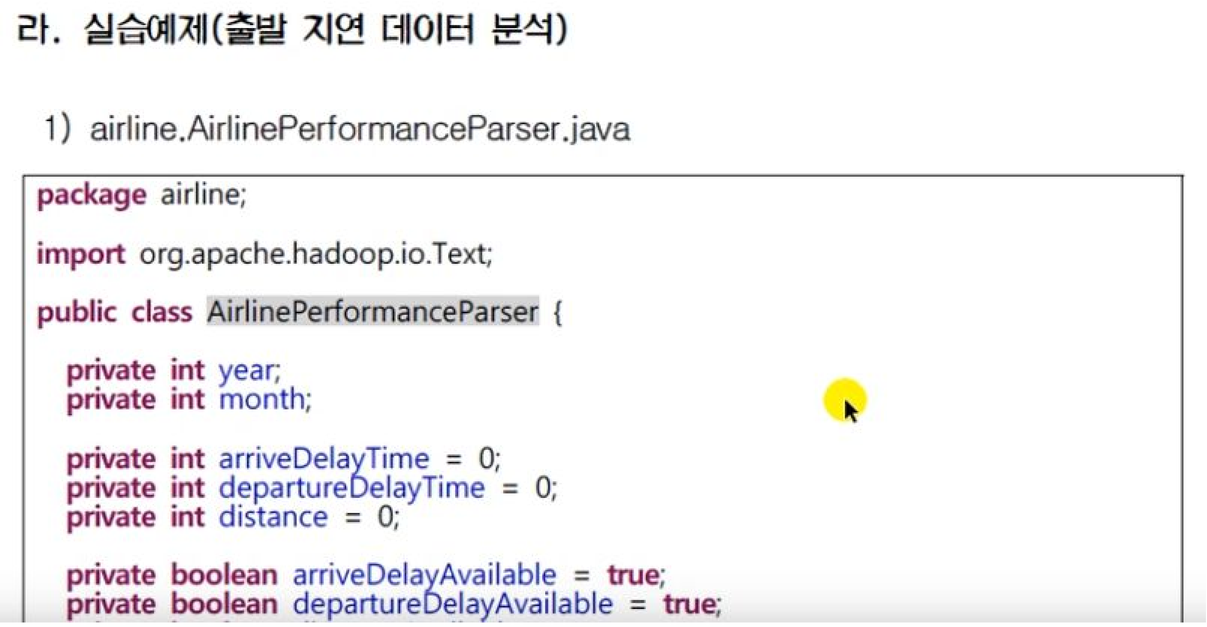
```
hdfs dfs -mkdir input

hdfs dfs -put /home/centos/data/airline input
```



```
[root@master ~]# hdfs dfs -ls input
Found 1 items
drwxr-xr-x   - root supergroup          0 2019-06-10 17:14 input/airline
[root@master ~]# hdfs dfs -ls input/airline
Found 3 items
-rw-r--r--   2 root supergroup  672068096 2019-06-10 17:14 input/airline/2006.csv
-rw-r--r--   2 root supergroup  702878193 2019-06-10 17:14 input/airline/2007.csv
-rw-r--r--   2 root supergroup  689413344 2019-06-10 17:14 input/airline/2008.csv
[root@master ~]#
```

올라간것 확인

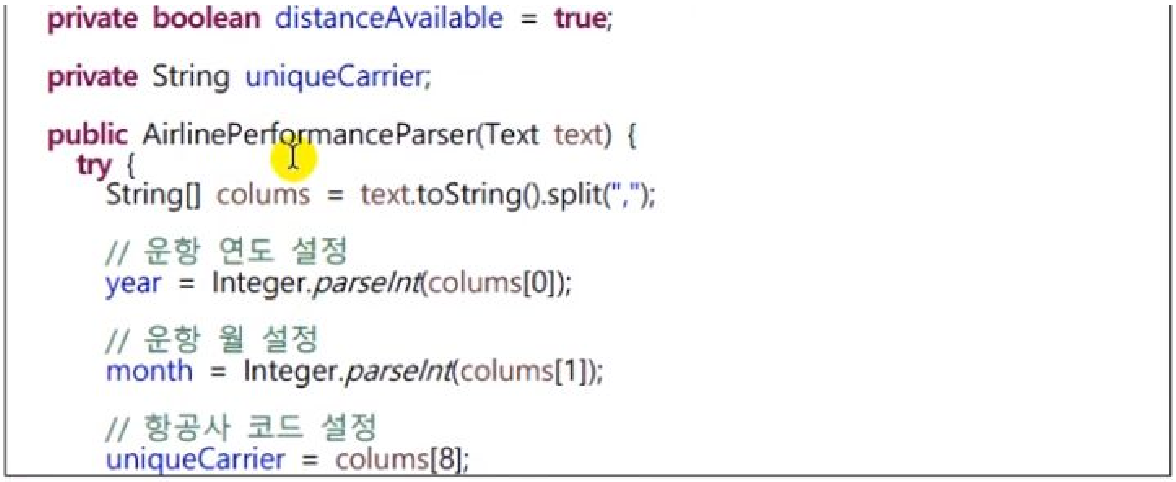
월별로 지연된 건수를 연도별 또는 월별로 값을 알고 싶다.
만약에 DBMS로 분석을 한다면,
select year, month, count(*)
from ontime
where depdelay >0
group by year, month
order by year, month;



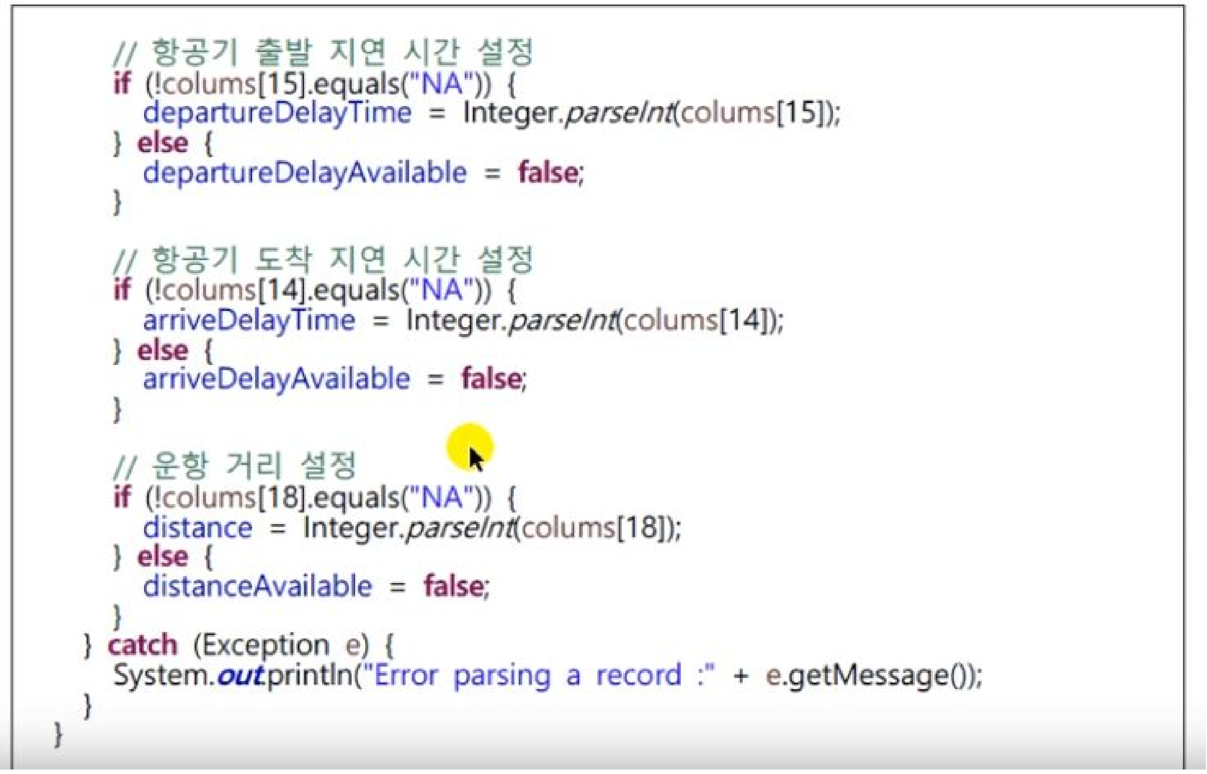
라. 실습예제(출발 지연 데이터 분석)

1) airline.AirlinePerformanceParser.java

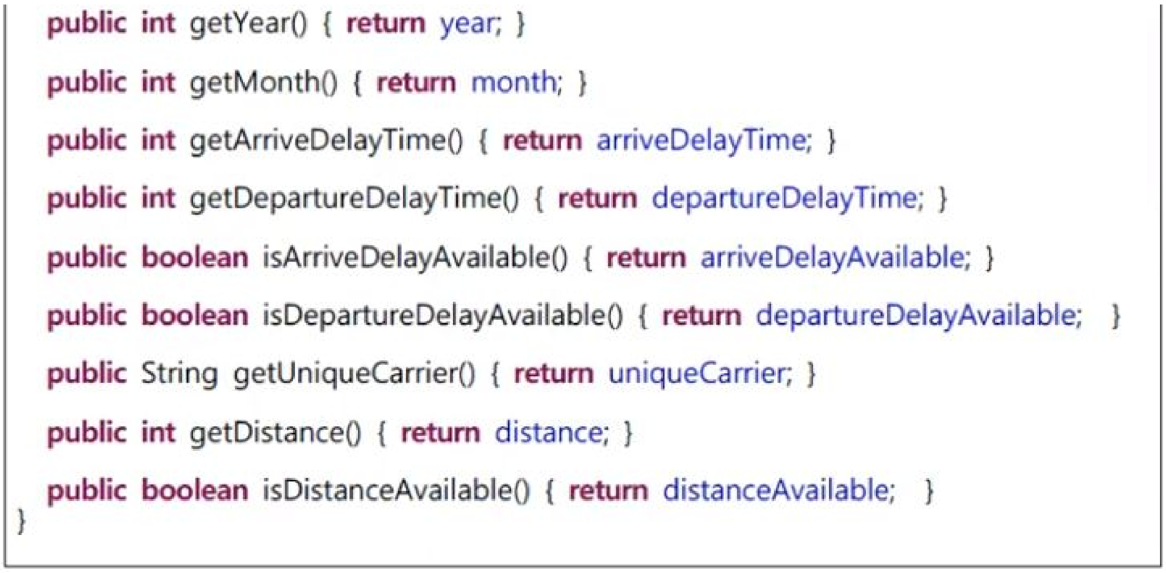
```
package airline;

import org.apache.hadoop.io.Text;

public class AirlinePerformanceParser {

  private int year;
  private int month;

  private int arriveDelayTime = 0;
  private int departureDelayTime = 0;
  private int distance = 0;

  private boolean arriveDelayAvailable = true;
  private boolean departureDelayAvailable = true;
```



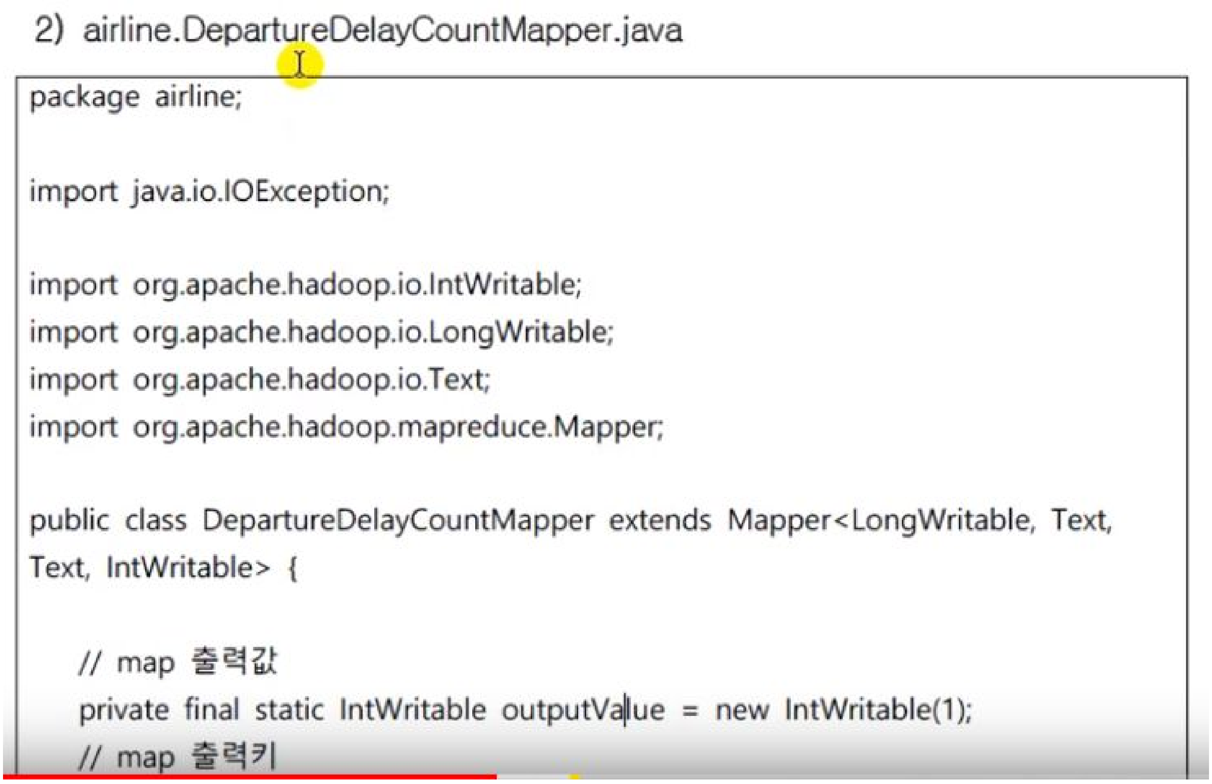
```
  private boolean distanceAvailable = true;

  private String uniqueCarrier;

  public AirlinePerformanceParser(Text text) {
    try {
      String[] colums = text.toString().split(",");

      // 운항 연도 설정
      year = Integer.parseInt(colums[0]);

      // 운항 월 설정
      month = Integer.parseInt(colums[1]);

      // 항공사 코드 설정
      uniqueCarrier = colums[8];
```



```
        // 항공기 출발 지연 시간 설정
        if (!colums[15].equals("NA")) {
          departureDelayTime = Integer.parseInt(colums[15]);
        } else {
          departureDelayAvailable = false;
        }

        // 항공기 도착 지연 시간 설정
        if (!colums[14].equals("NA")) {
          arriveDelayTime = Integer.parseInt(colums[14]);
        } else {
          arriveDelayAvailable = false;
        }

        // 운항 거리 설정
        if (!colums[18].equals("NA")) {
          distance = Integer.parseInt(colums[18]);
        } else {
          distanceAvailable = false;
        }
      } catch (Exception e) {
        System.out.println("Error parsing a record :" + e.getMessage());
      }
    }
```

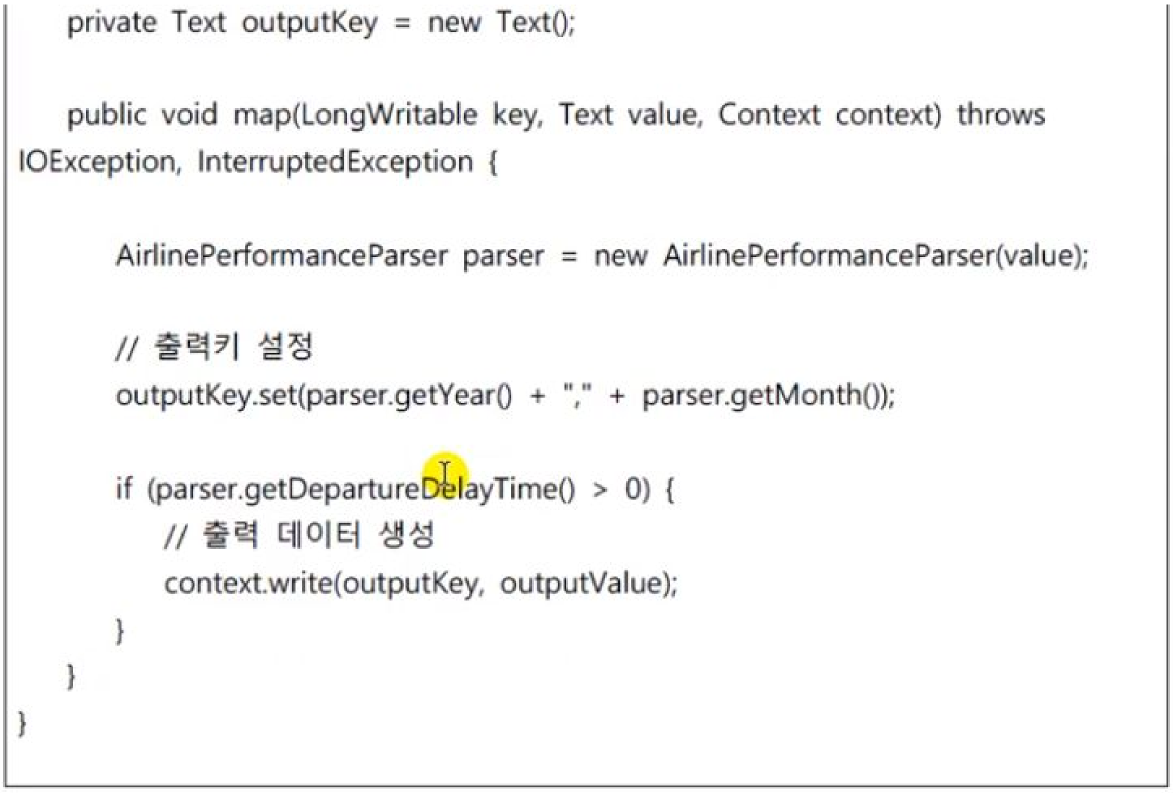


```
  public int getYear() { return year; }

  public int getMonth() { return month; }

  public int getArriveDelayTime() { return arriveDelayTime; }

  public int getDepartureDelayTime() { return departureDelayTime; }

  public boolean isArriveDelayAvailable() { return arriveDelayAvailable; }

  public boolean isDepartureDelayAvailable() { return departureDelayAvailable;  }

  public String getUniqueCarrier() { return uniqueCarrier; }

  public int getDistance() { return distance; }

  public boolean isDistanceAvailable() { return distanceAvailable;  }
}
```

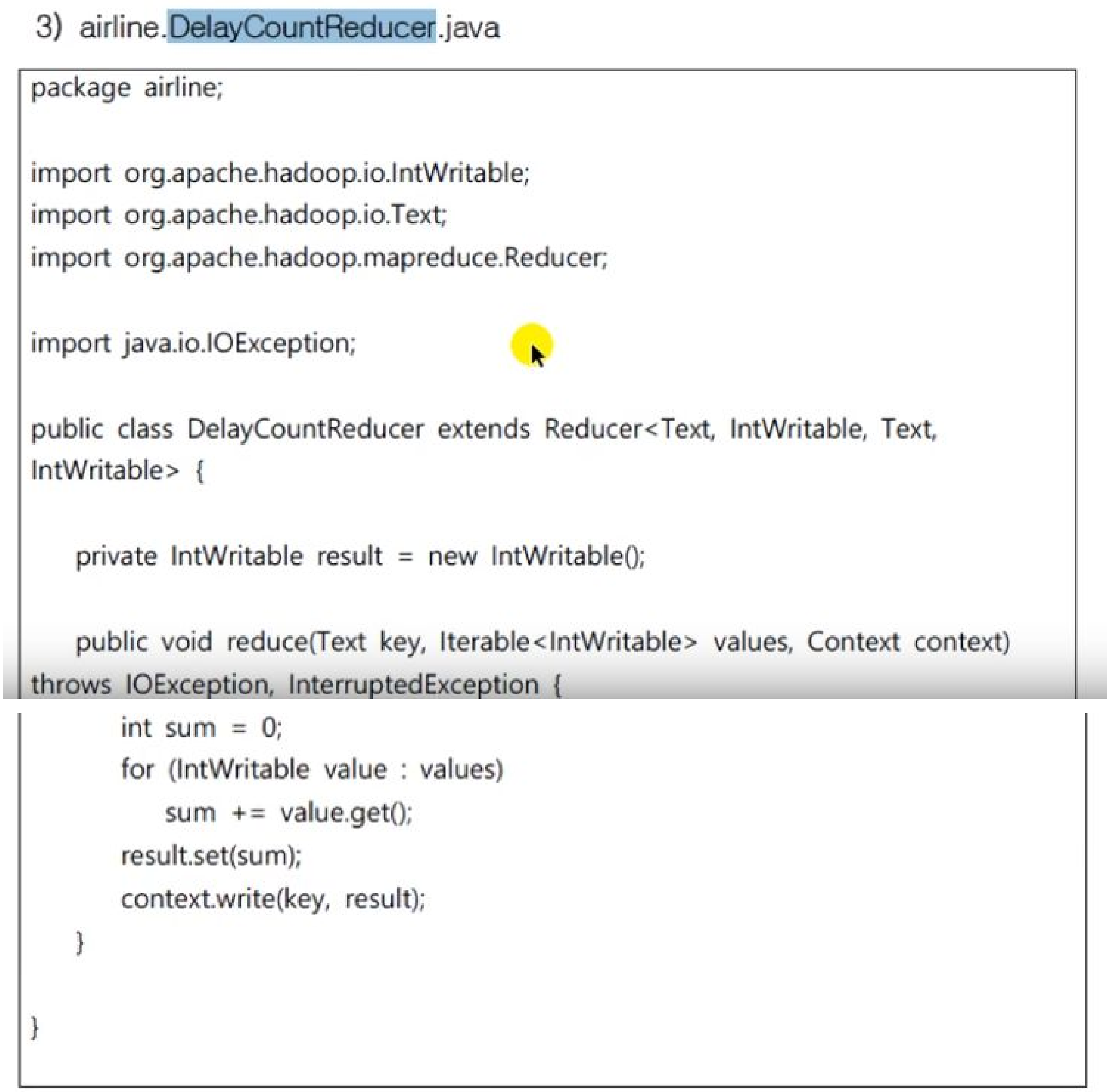


2) airline.DepartureDelayCountMapper.java

```
package airline;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class DepartureDelayCountMapper extends Mapper<LongWritable, Text,
Text, IntWritable> {

    // map 출력값
    private final static IntWritable outputValue = new IntWritable(1);
    // map 출력키
```



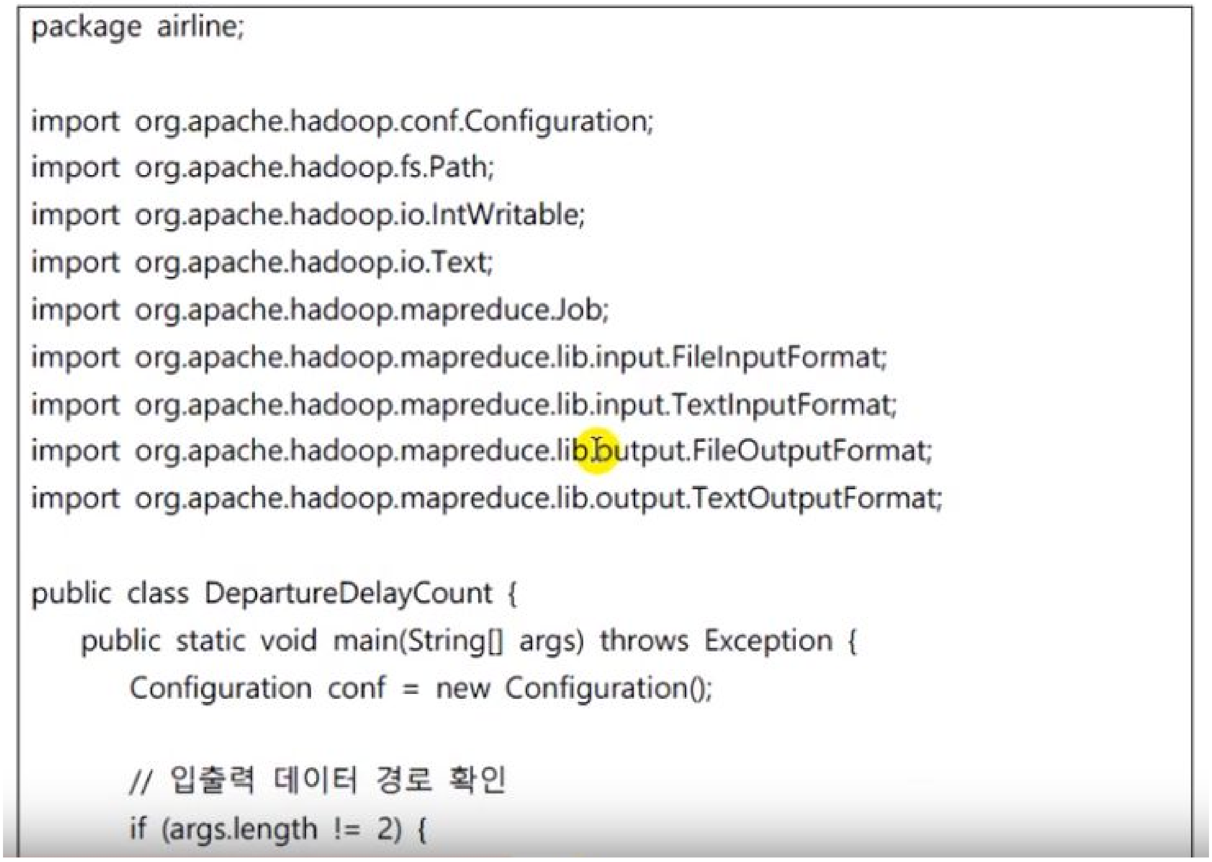
```
    private Text outputKey = new Text();

    public void map(LongWritable key, Text value, Context context) throws
IOException, InterruptedException {

        AirlinePerformanceParser parser = new AirlinePerformanceParser(value);

        // 출력키 설정
        outputKey.set(parser.getYear() + "," + parser.getMonth());

        if (parser.getDepartureDelayTime() > 0) {
            // 출력 데이터 생성
            context.write(outputKey, outputValue);
        }
    }
}
```



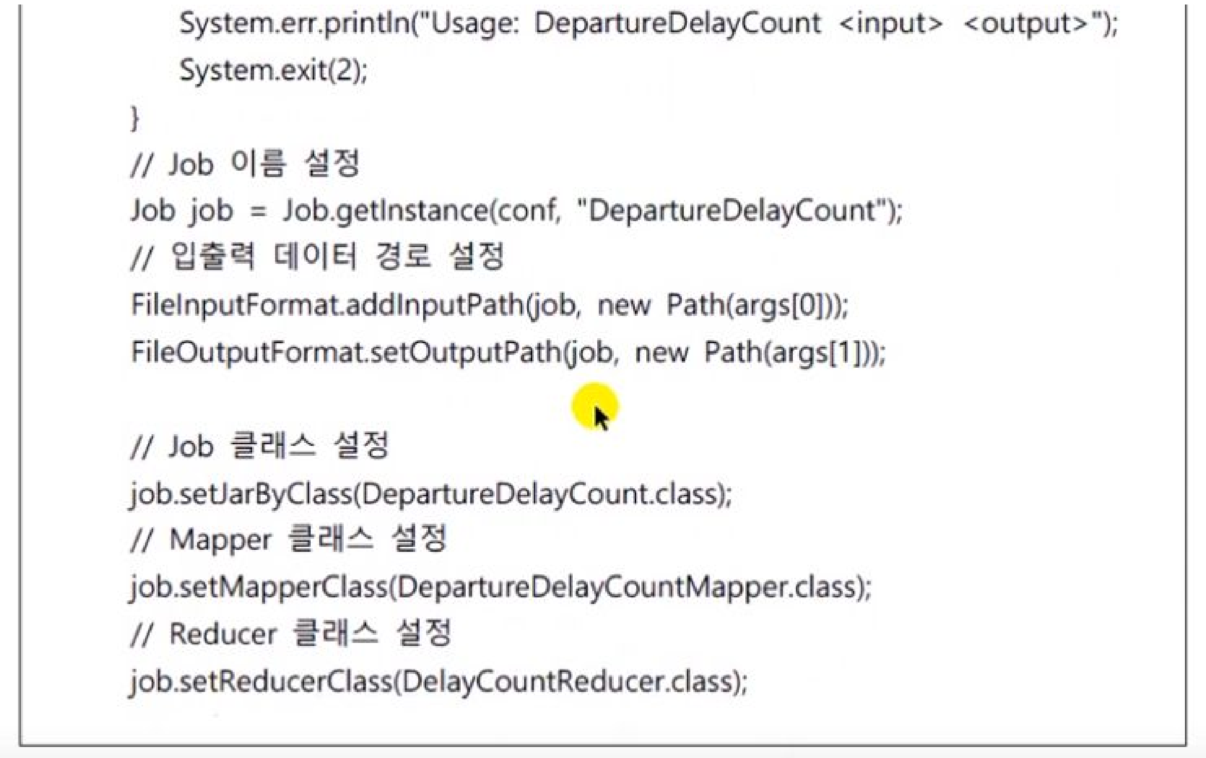
3) airline.DelayCountReducer.java

```
package airline;

import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

import java.io.IOException;

public class DelayCountReducer extends Reducer<Text, IntWritable, Text,
IntWritable> {

    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values, Context context)
throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable value : values)
            sum += value.get();
        result.set(sum);
        context.write(key, result);
    }

}
```

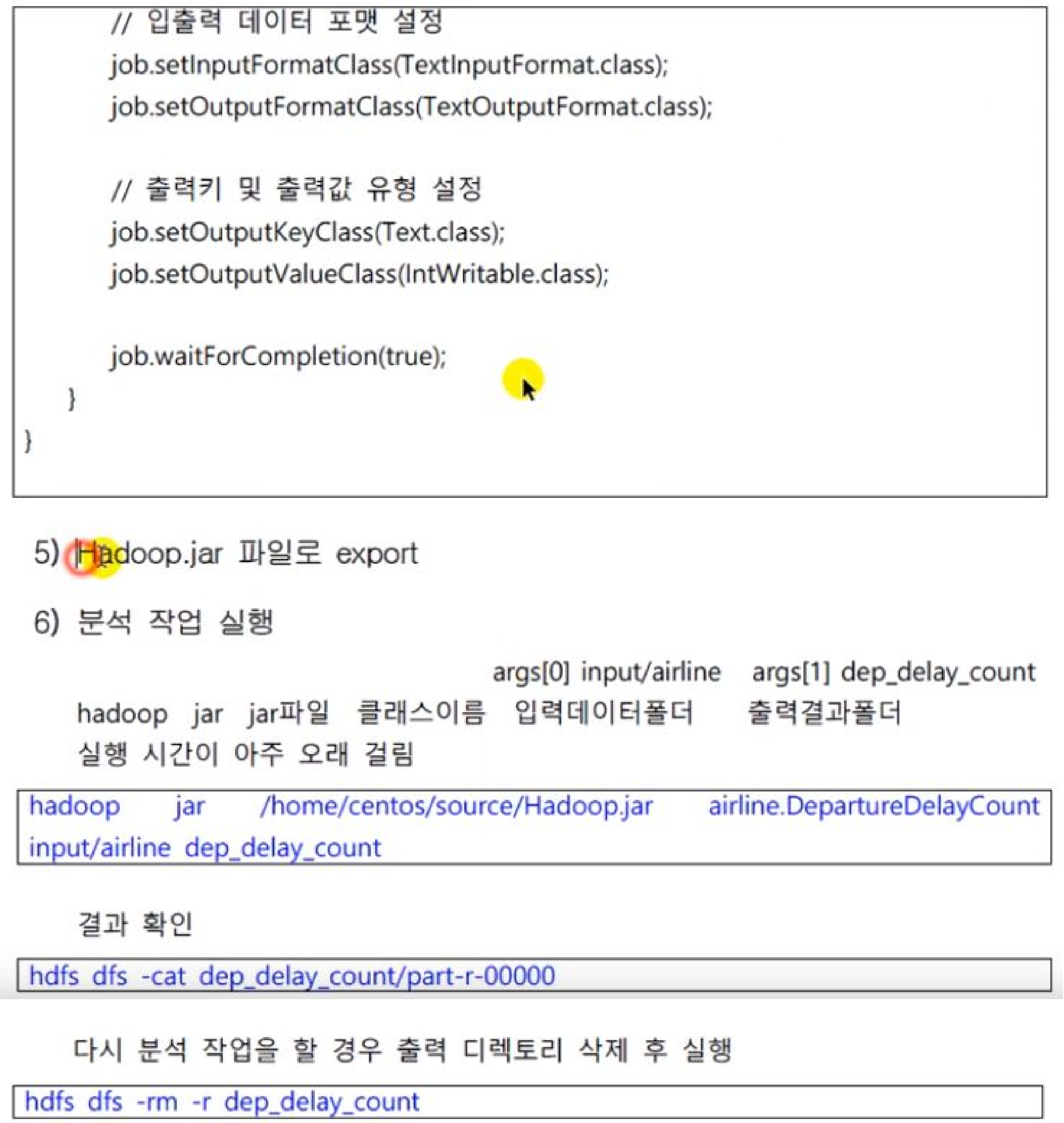
4) DepartureDelayCount.java



```
package airline;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class DepartureDelayCount {
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();

        // 입출력 데이터 경로 확인
        if (args.length != 2) {
```



```
            System.err.println("Usage: DepartureDelayCount <input> <output>");
            System.exit(2);
        }
        // Job 이름 설정
        Job job = Job.getInstance(conf, "DepartureDelayCount");
        // 입출력 데이터 경로 설정
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        // Job 클래스 설정
        job.setJarByClass(DepartureDelayCount.class);
        // Mapper 클래스 설정
        job.setMapperClass(DepartureDelayCountMapper.class);
        // Reducer 클래스 설정
        job.setReducerClass(DelayCountReducer.class);
```



```
        // 입출력 데이터 포맷 설정
        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

        // 출력키 및 출력값 유형 설정
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);

        job.waitForCompletion(true);
    }
}
```

5) Hadoop.jar 파일로 export

6) 분석 작업 실행

```
                                  args[0] input/airline   args[1] dep_delay_count
hadoop  jar  jar파일  클래스이름  입력데이터폴더          출력결과폴더
```

실행 시간이 아주 오래 걸림

```
hadoop   jar   /home/centos/source/Hadoop.jar   airline.DepartureDelayCount
input/airline  dep_delay_count
```

결과 확인

```
hdfs dfs -cat dep_delay_count/part-r-00000
```

다시 분석 작업을 할 경우 출력 디렉토리 삭제 후 실행

```
hdfs dfs -rm -r dep_delay_count
```

하지만, sql을 통해서 한줄이면 되는 것을 자바 코딩으로 가져온다는 것이 굉장히 비효율적이라는 생각이 든다.
따라서, hive라는 기술을 사용하면, 분산처리시스템을 sql문을 사용하여 결과값을 갖고올 수 있다.

22분까지 들었음.