Class 1. 데이터베이스에서 데이터 조회하기

데이터베이스란?

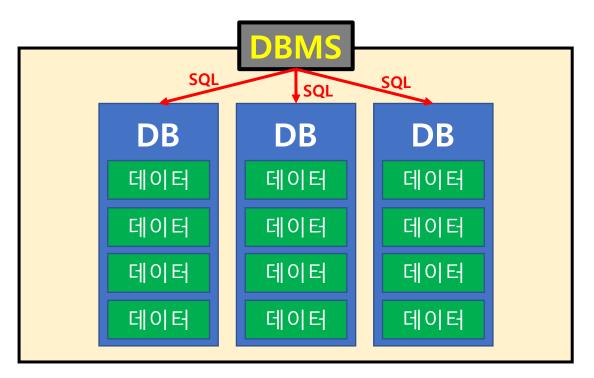
데이터(data, 자료): 숫자, 문자, 소리, 이미지, 영상 등의 형태로 된 의미 단위.

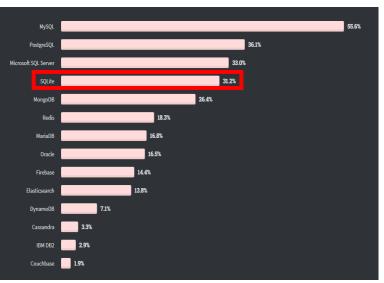
데이터베이스(DataBase, DB): 데이터가 저장되어 있는 공간. 데이터를 모아둔 집합.

DBMS(DataBase Management System): 데이터베이 스를 관리하고 운영하기 위한 시스템.

SQL(Structured Query Language, 구조화된 질의 언어): 데이터베이스를 구축하고 관리하고 활용하기 위해서 사용되는 언어.

세상에는 많은 종류의 DBMS가 존재. 하지만 SQL을 알면 Oracle, MySQL, MariaDB, SQLite, Google Cloud SQL, Amazon RDS, MS SQL Server, PostgreSQL 등의 DBMS를 다룰 수 있음.





출처: Stack Overflow Developer Survey, 2020

SQLite3 설치하기

우리는 DBMS 중에 하나인 SQLite3로 실습 진행

SQLite: "에스큐엘라이트" 또는 "시퀄라이트"라고 부름

터미널에 다음과 같이 입력

sudo apt-get install sqlite3 libsqlite3-dev

- 1) sudo: super user do의 약자. 가장 강력한 권한을 가진 존재가 처리한다는 뜻.
- 2) apt-get: 우분투(Ubuntu)와 같은 데비안 계열의 리눅스 운영체제에서 사용되는 패키지 관리 도구.

터미널에 sqlite3 --version 를 입력했을 때 다음과 같은 결과가 나온다면 제대로 설치된 것

root@goorm:/workspace/KUSF_data# sqlite3 --version 3.22.0 2018-01-22 18:45:57 0c55d179733b46d8d0ba4d88e01a25e10677046ee3da1d5b1581e86726f2alt1

레먼 데이터베이스 다운로드 받기

DBMS를 설치했으니 테스트 해볼 DB가 필요. MLB DB 중 하나인 레먼 데이터베이스로 실습 진행.

다운로드 링크: http://www.seanlahman.com/baseball-archive/statistics/

2019 - MS Access version

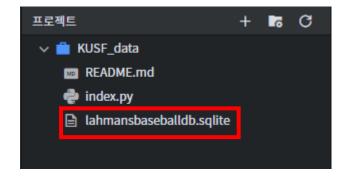
2019 - comma-delimited version

2019 - R Package

2019 - MySQL version

2019 - SQL Lite

2019-SQL Lite 클릭!



다운로드 받은 파일을 현재 작업 디렉토리에 넣어주세요

레먼 데이터베이스:

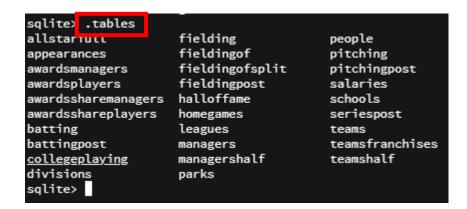
1871년부터 2020년까지의 MLB 데이터 보유.

숀 레먼 만듦.

매 시즌 후 갱신.

레먼 데이터베이스 열기

```
root@goorm:/workspace/KUSF_data# sqlite3 lahmansbaseballdb.sqlite
SQLite version 3.22.0 2018-01-22 18:45:5/
Enter ".help" for usage hints.
sqlite>
```



1. 데이터베이스 열기

sqlite3 lahmansbaseballdb.sqlite

2. 데이터베이스 내 테이블 확인

.tables

레먼 데이터베이스는 people, pitching, batting 등 다양한 테이블로 구성되어 있음

레먼 데이터베이스 설명서:

http://www.seanlahman.com/files/database/readme2017.txt

people 테이블: 선수 데이터		레먼 데이터베이스
	pitching 테이블: 투수 데이터	
		salaries 테이블: 연봉 데이터
fielding 테이블: 수비 데이터	batting 테이블: 타자 데이터	

1. 연봉 테이블 구조 확인

.schema 테이블명

```
SQLITE SCHEMA SALARIES
CREATE TABLE JE NOT EXISTS "calaries" (
"ID" INTEGER NOT NULL,
"yearID" SMALLINT NOT NULL,
"teamID" CHARACTER(3) NOT NULL,
"team_ID" INTEGER NULL,
"lgID" CHARACTER(2) NOT NULL,
"lgID" CHARACTER(2) NOT NULL,
"playerID" VARCHAR(9) NOT NULL,
"salary" DOUBLE NULL,
"salary" DOUBLE NULL,
FRIMARY KEY ( ID ),
FOREIGN KEY("lgID") REFERENCES "leagues" ("lgID") ON UPDATE NO ACTION ON DELETE NO ACTION,
FOREIGN KEY("team_ID") REFERENCES "teams" ("ID") ON UPDATE NO ACTION ON DELETE NO ACTION,
FOREIGN KEY("playerID") REFERENCES "people" ("playerID") ON UPDATE NO ACTION ON DELETE NO ACTION
);
CREATE INDEX "salaries_lgID" ON "salaries" ("lgID");
CREATE INDEX "salaries_playerID" ON "salaries" ("playerID");
CREATE INDEX "salaries_team_ID" ON "salaries" ("team_ID");
CREATE UNIQUE INDEX "salaries_team_ID" ON "salaries" ("team_ID");
CREATE UNIQUE INDEX "salaries_yearID" ON "salaries" ("yearID", "teamID", "lgID", "playerID");
```

필드명 또는 컬럼명



salaries 테이블 구조

ID	yearID	teamID	team_ID	lgID	playerID	salary

2. 연봉 데이터 조회

SELECT * FROM salaries;

```
|2016|WAS|2835|NL|werthja01|21733615.0
```

DB에 누적되어 있는 연봉 데이터가 모두 조회됨.

3. 2016년 연봉 데이터 조회

SELECT * FROM salaries WHERE yearID = 2016;

```
2016|TOR|2834|AL|smoakju01|3900000.0
2016|TOR|2834|AL|storedr01|8375000.0
2016|WAS|2835|NL|werthja01|21733615.0
```

2016년 연봉 데이터만 조회됨.

4. 2016년 연봉이 높은 사람부터 낮은 사람 순으로 정렬

SELECT * FROM salaries WHERE yearID = 2016 ORDER BY salary DESC;

5. 2016년 연봉 TOP 10 조회

조회한 결과 중 처음 10개 행만

SELECT * FROM salaries WHERE yearID = 2016 ORDER BY salary DESC LIMIT 10;

6. 컬럼별로 정리되어 보여지게 하고, 컬럼명도 함께 보여지게 하기

.mode column

.header on

sqlite>	.mode co	lumn					
sqlite>	SELECT *	FROM salaries	WHERE yearID	= 2016 ORDE	R BY	salary DESC	LIMIT 10;
25966	2016	LAN	2819	NL		kershcl01	33000000.0
25589	2016	ARI	2806	NL		greinza01	31799030.0
25674	2016	BOS	2809	AL		priceda01	30000000.0
25833	2016	DET	2815	AL		cabremi01	28000000.0
25859	2016	DET	2815	AL		verlaju01	28000000.0
26067	2016	NYN	2824	NL		cespeyo01	27328046.0
26242	2016	SEA	2829	AL		hernafe02	25857143.0
25702	2016	CHN	2811	NL		lestejo01	25000000.0
25933	2016	LAA	2818	AL		pujolal01	25000000.0
26111	2016	NYA	2823	AL		sabatcc01	25000000.0
							*

	.header on SELECT * FROM	salaries W	HERE vearID =	2016 ORDE	R BY salarv DESC	LIMIT 10:
ID	yearID	teamID	team_ID	lgID	playerID	salary
25966	2016	LAN	2819	NL	kershcl01	33000000.0
25589	2016	ARI	2806	NL	greinza01	31799030.0
25674	2016	BOS	2809	AL	priceda01	30000000.0
25833	2016	DET	2815	AL	cabremi01	28000000.0
25859	2016	DET	2815	AL	verlaju01	28000000.0
26067	2016	NYN	2824	NL	cespeyo01	27328046.0
26242	2016	SEA	2829	AL	hernafe02	25857143.0
25702	2016	CHN	2811	NL	lestejo01	25000000.0
25933	2016	LAA	2818	AL	pujolal01	25000000.0
26111	2016	NYA	2823	AL	sabatcc01	25000000.0

컬럼별로 정리됨

컬럼명도 보임

7. people 테이블의 구조 확인

.schema people

```
sqlite> .schema people
CREATE TABLE IF NOT EXISTS "people" (
        "playerID" VARCHAR(9) NOT NULL,
        "birthYear" INTEGER NULL,
        "birthMonth" INTEGER NULL,
        "birthDay" INTEGER NULL,
        "birthCountry" VARCHAR(255) NULL,
        "birthState" VARCHAR(255) NULL,
        "birthCity" VARCHAR(255) NULL,
        "deathYear" INTEGER NULL,
        "deathMonth" INTEGER NULL,
        "deathDay" INTEGER NULL,
        "deathCountry" VARCHAR(255) NULL,
        "deathState" VARCHAR(255) NULL,
        "deathCity" VARCHAR(255) NULL,
        "nameFirst" VARCHAR(255) NULL,
        "nameLast" \ARCHAR(255) NULL,
        "nameGiven" VARCHAR(255) NULL,
        "weight" INTEGER NULL,
        "height" INTEGER NULL,
        "bats" VARCHAR(255) NULL,
        "throws" VARCHAR(255) NULL,
        "debut" VARCHAR(255) NULL,
        "finalGame" VARCHAR(255) NULL,
        "retroID" VARCHAR(255) NULL,
        "bbrefID" VARCHAR(255) NULL,
        "birth_date" DATE NULL,
        "debut_date" DATE NULL,
        "finalgame_date" DATE NULL,
        "death_date" DATE NULL,
        PRIMARY KEY ("playerID")
```

8. playerID로 어떤 선수인지 확인하기

SELECT * FROM people WHERE playerID = 'kershcl01';

		OM people WHE ar birthMon				tate birthC	ity deathY	ear deathM	lonth deathDay	deathCountry	deathState	deathCity	nameFirst	nameLast	nameGiven	weight	hei
ght	bats	throws	debut	finalGame	retroID	bbrefID	birth_date	debut_date	finalgame_date	death_date							
kershcl01	1 1988	3	19	USA	TX	Dallas							Clayton	Kershaw	Clayton Edward	226	76
_	L	L	2008-05-25	2019-09-29	kersc001	kershcl01	1988-03-19	2008-05-25	2019-09-29								

1D yearID teamID team_ID lgID playerID salary	sqlite> SEL	.ECT * FROM s	alaries WHER	E yearID = 2	016 ORDER BY	salary DESC	LIMIT 10;	
25589 2016 ARI 2806 NL greinza01 31799030.0 잭 그레인키 25674 2016 BOS 2809 AL priceda01 30000000.0 데이빗 프라이스 25833 2016 DET 2815 AL cabremi01 28000000.0 미겔 카브레라 326067 2016 NYN 2824 NL cespeyo01 27328046.0 요에니스 세스페데스 26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 잘 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	ID	yearID	teamID	team_ID	lgID	playerID	salary	
25589 2016 ARI 2806 NL greinza01 31799030.0 잭 그레인키 25674 2016 BOS 2809 AL priceda01 30000000.0 데이빗 프라이스 25833 2016 DET 2815 AL cabremi01 28000000.0 미겔 카브레라 25859 2016 DET 2815 AL verlaju01 28000000.0 저스틴 벌랜더 26067 2016 NYN 2824 NL cespeyo01 27328046.0 요에니스 세스페데스 26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스								
25674 2016 BOS 2809 AL priceda01 30000000.0 데이빗프라이스 25833 2016 DET 2815 AL cabremi01 28000000.0 미겔 카브레라 25859 2016 DET 2815 AL verlaju01 28000000.0 저스틴 벌랜더 26067 2016 NYN 2824 NL cespeyo01 27328046.0 요에니스 세스페데스 26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	25966	2016	LAN	2819	NL	kershcl01	33000000.0	클레이튼 커쇼 (약 395억)
25833 2016 DET 2815 AL cabremi01 28000000.0 미겔 카브레라 25859 2016 DET 2815 AL verlaju01 28000000.0 저스틴 벌랜더 26067 2016 NYN 2824 NL cespeyo01 27328046.0 요에니스 세스페데스 26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	25589	2016	ARI	2806	NL	greinza01	31799030.0	잭 그레인키
25859 2016 DET 2815 AL verlaju01 28000000.0 저스틴 벌랜더 26067 2016 NYN 2824 NL cespeyo01 27328046.0 요에니스 세스페데스 26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	25674	2016	BOS	2809	AL	priceda01	30000000.0	데이빗 프라이스
26067 2016 NYN 2824 NL cespeyo01 27328046.0 요에니스 세스페데스 26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	25833	2016	DET	2815	AL	cabremi01	28000000.0	미겔 카브레라
26242 2016 SEA 2829 AL hernafe02 25857143.0 펠릭스 에르난데스 25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	25859	2016	DET	2815	AL	verlaju01	28000000.0	저스틴 벌랜더
25702 2016 CHN 2811 NL lestejo01 25000000.0 존 레스터 25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	26067	2016	NYN	2824	NL	cespeyo01	27328046.0	
25933 2016 LAA 2818 AL pujolal01 25000000.0 알버트 푸홀스	26242	2016	SEA	2829	AL	hernafe02	25857143.0	펠릭스 에르난데스
	25702	2016	CHN	2811	NL	lestejo01	25000000.0	
26111 _ 2016 NYA 2823 AL sabatcc01 25000000.0 CC 사바시아	25933	2016	LAA	2818	AL	pujolal01	25000000.0	
	26111	2016	NYA	2823	AL	sabatcc01	25000000.0	CC 사바시아

실습2: 류현진 연봉 확인하기

1. 류현진의 성을 이용해서 류현진의 playerID 찾기

```
SELECT playerID, nameFirst, nameLast FROM people WHERE nameLast = 'Ryu';
특정 컬럼들의 데이터만 보고 싶을 때
*은 모든 컬럼을 보고 싶을 때
```

```
sqlite> SELECT playerID, nameFirst, nameLast FROM people WHERE nameLast = 'Ryu';
playerID nameFirst nameLast
------
ryuhy01 Hyun-Jin Ryu
ryuja01 Jae Kuk Ryu
```

실습2: 류현진 연봉 확인하기

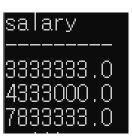
2. 류현진의 playerID를 활용하여 류현진의 연봉 조회

SELECT * FROM salaries WHERE playerID = 'ryuhy01';



3. 다른 데이터 말고 연봉만 보고 싶다면?

SELECT salary FROM salaries WHERE playerID = 'ryuhy01';

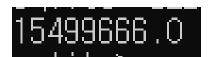


집계함수(aggregate)

집계함수	설명
AVG	평균
COUNT	개수
MAX	최대값
MIN	최소값
SUM	합계(정수들의 합계라면 정수로)
TOTAL	합계(항상 실수로)
GROUP_CONCAT	문자열들 연결

4. 류현진의 연봉을 모두 합하면?

SELECT SUM(salary) FROM salaries WHERE playerID = 'ryuhy01';



약 185억

실습3: MLB 역대 한국 선수들 조회하기

1. people 테이블에서 birthCountry가 South Korea인 선수 조회

SELECT * FROM people WHERE birthCountry = 'South Korea';

2. 데뷔순으로 조회

SELECT * FROM people WHERE birthCountry = 'South Korea' ORDER BY debut;

	deathCountry	deathStat	te deathC	ity nameFirst	nameLast	birthCit nameGiven finalgame_date 	weight	height		
parkch01	1973 6 8 2010-10-01			 South Korea Chan Ho 1973-06-30	Park	 	210	 74	 R	R
choji01 1998-07-0	1975 8 4 1999-07-30	cho-j001	16 choji01	South Korea Jin Ho 1975-08-16	North Jeolla Cho 1998-07-04	a Jeonju Jin Ho 1999-07-30	175	72	R	R
1999-05-2	9 2007-09-28	kim-b001	kimby01	Byung-Hyui 1979-01-19	n Kim 1999-05-29	Gwangju Byung-Hyun 2007-09-28 Seoul Sang-Hoon			R	R
2000-06-2	9 2000-10-01	lee-s001	leesa01 4	1971-03-11 South Korea	2000-06-29 Incheon	Sang-Hoon 2000-10-01 Incheon Sun-Woo			L R	L R
bongju01	1980 7		kimsu01 15	1977-09-04 South Korea	2001-06-15 Seoul Bong	2006-09-17 Seoul Jung Keun			L	L
seoja01 2002-07-2	1977 5 1 2007-05-29	seo-j001	24 seoja01	South Korea Jae Weong 1977-05-24	Gwangju Seo 2002-07-21	Gwangju Jae Weong 2007-05-29	215	73	R	R
choihe01	1979 3		16	South Korea Hee-Seop	South Jeolla Choi	a Hwasun Hee-Seop	235	77	L	L

3. 총 몇 명인지?

SELECT COUNT(*) FROM people WHERE birthCountry = 'South Korea';



롭 레프스나이더(김정태): 한국에서 태어났지만 생후 5개월만에 미국으로 입양됨 토미 펠프스: 아버지가 주한미군으로 근무할 때 출생.

실습4: MLB 2019시즌 홈런왕은 누구?

1. batting 테이블 구조 확인

.schema batting

2. 2019년도 타자기록 홈런 개수 역순으로 정렬한 후 첫번째 행만 조회

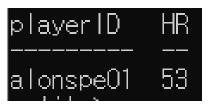
SELECT * FROM batting WHERE yearID = 2019 ORDER BY HR DESC LIMIT 1;



피트 알론소 2021 시즌에도 홈런 37개로 3위

3. 다른 데이터 말고 playerID와 홈런 개수만 보고 싶다면?

SELECT playerID, HR FROM batting WHERE yearID = 2019 ORDER BY HR DESC;



실습5: MLB 2019시즌 탈삼진 TOP 5는 누구?

1. pitching 테이블 구조 확인

.schema pitching

2. 2019년도 투수 기록 탈삼진 개수 내림차순으로 정렬한 후 처음 5개 행만 조회

SELECT playerID, SO FROM pitching WHERE yearID = 2019 ORDER BY SO DESC LIMIT 5;

playerID	SO
colege01	326
verlaju01	300
biebesh01	259
degroja01	255
strasst01	251

모두 굉장히 유명한 선수들

게릿 콜 저스틴 벌랜더 셰인 비버 제이콥 디그롬 스티븐 스트라스버그

* 삼진을 잘 잡는 투수의 몸값이 대체로 높은 편. 그 이유는? 이후 강의에서 설명 예정^^

Sqlite3 쿼리 문법 요약

목적	쿼리 문법	예시
데이터베이스 열기	.open 데이터베이스 파일명	open lahmansbaseballdb.sqlite.
존재 테이블 확인	.tables	
컬럼명과 함께 데이터 조회	.mode column .header on	
테이블 구조 확인	.schema 테이블명	.schema batting
데이터 조회	SELECT * FROM 테이블명; (모든 필드 조회시) SELECT 필드명1, 필드명2, FROM 테이블명; (특정 필드 조회시)	SELECT * FROM batting; SELECT playerID, HR FROM batting;
조건을 가지고 데이터 조회	SELECT * FROM 테이블명 WHERE 조건;	SELECT * FROM batting WHERE yearID = 2019;
지정 컬럼을 기준으로 데이터 정렬	SELECT * FROM 테이블명 ORDER BY 필드명;	SELECT * FROM batting ORDER BY HR; SELECT * FROM batting ORDER BY HR DESC; (내림차순 정렬)
설정한 개수의 데이터만 조회	SELECT * FROM 테이블명 LIMIT 개수;	SELECT * FROM batting LIMIT 5;
조회한 행 개수 확인	SELECT COUNT(*) FROM 테이블명;	SELECT COUNT(*) FROM batting;
조회한 데이터 합계	SELECT SUM(필드명) FROM 테이블명;	SELECT SUM(HR) FROM batting;

중요한 것은 필요에 따라 이 문법들을 잘 조합해서 사용해야 한다는 점!

과제#1

레먼 데이터베이스에서 흥미로운 데이터 5건 조회하기

예시1. 2018 정규시즌에 3루타를 가장 많이 친 타자 3명은?

예시2. 추신수 선수의 연봉이 가장 높았던 연도는?

예시3. 2019 정규시즌에 방어율이 가장 좋았던 투수 10명은?

예시4. 2019 포스트시즌에 가장 많은 홈런을 친 타자는?

예시5. 2019 정규시즌에 가장 많은 실책을 범한 선수 5명은?

예시에 있는 것을 해도 좋지만, 예시에 없는 것을 하는 것을 추천!

쿼리와 조회 결과 캡쳐 화면을 word로 정리해서 kyohoonsim@gmail.com 으로 보내주세요~!

문서 제목 양식:

KUSF데이터분석_과제1_이름.docx

ex) KUSF데이터분석_과제1_심교훈.docx