Big Data Analysis

Lecture 3

2019/11/29

A Real Big Data Case in Finance

- Quantitative Trading
- Cover stocks (3600+), ETFs(250), futures(55*120), bonds(different types), options(tenors*strikes), funds(7000+), etc.
- Since 2011 or the first listed dates, around 2160 business days
- Various data types: prices (OHLC), volume, amount, news, …
- Derived data information such as dividend adjusted prices, returns, consensus, etc.

Actually Available Data

lev2_lz_201712130400.zip	2017/12/23 11:06	WinRAR ZIP archive	5,813,602 KB	2018/7/22 21:40
lev2_lz_201712131510.zip	2017/12/23 11:07	WinRAR ZIP archive	5,530,698 KB	2018/7/22 21:47
lev2_lz_201712140400.zip	2017/12/23 11:14	WinRAR ZIP archive	6,633,712 KB	2018/7/22 21:54
lev2_lz_201712141510.zip	2017/12/23 11:13	WinRAR ZIP archive	5,271,076 KB	2018/7/22 22:03
lev2_lz_201712150400.zip	2017/12/23 11:21	WinRAR ZIP archive	6,381,791 KB	2018/7/22 20:02
lev2_lz_201712151510.zip	2017/12/23 11:21	WinRAR ZIP archive	5,278,234 KB	2018/7/22 20:11
lev2_lz_201712180400.zip	2017/12/23 11:29	WinRAR ZIP archive	5,922,013 KB	2018/7/22 20:19
lev2_lz_201712181510.zip	2017/12/23 11:28	WinRAR ZIP archive	5,352,670 KB	2018/7/22 20:27
lev2_lz_201712190400.zip	2017/12/23 11:36	WinRAR ZIP archive	6,291,980 KB	2018/7/22 20:35
lev2_lz_201712191510.zip	2017/12/23 11:35	WinRAR ZIP archive	5,448,475 KB	2018/7/22 20:44
lev2_lz_201712200400.zip	2017/12/23 11:43	WinRAR ZIP archive	5,714,756 KB	2018/7/22 20:52
lev2_lz_201712201510.zip	2017/12/23 11:43	WinRAR ZIP archive	5,272,442 KB	2018/7/22 21:01
lev2_lz_201712210400.zip	2017/12/23 11:50	WinRAR ZIP archive	6,289,756 KB	2018/7/22 21:08
lev2_lz_201712211510.zip	2017/12/23 11:50	WinRAR ZIP archive	5,328,823 KB	2018/7/22 21:16
lev2_lz_201712220400.zip	2017/12/23 11:58	WinRAR ZIP archive	6,205,395 KB	2018/7/22 21:24
lev2_lz_201712221510.zip	2017/12/23 11:57	WinRAR ZIP archive	5,161,314 KB	2018/7/22 21:32

lev2_lz_201712221510.zip\logs - ZIP64 archive, unpacked size 42,661,117,642 bytes

Name	Size	Packed	Type	Modified	CRC32
..			File folder		
shfe_lev2_debug_20171222_0.log	201,429,106	25,223,208	Text Document	2017/12/22 9:00	EAC67E61
shfe_lev2_debug_20171222_1.log	201,427,028	25,175,292	Text Document	2017/12/22 9:01	DEAD3460
shfe_lev2_debug_20171222_2.log	201,427,255	25,155,006	Text Document	2017/12/22 9:02	E1388095
shfe_lev2_debug_20171222_3.log	201,427,023	25,164,127	Text Document	2017/12/22 9:03	7FB7C096
shfe_lev2_debug_20171222_4.log	201,426,901	25,044,202	Text Document	2017/12/22 9:04	7F022DB1
shfe_lev2_debug_20171222_5.log	201,427,271	24,994,527	Text Document	2017/12/22 9:06	4757B480
shfe_lev2_debug_20171222_6.log	201,427,373	25,159,058	Text Document	2017/12/22 9:07	9159F55F
shfe_lev2_debug_20171222_7.log	201,427,377	25,133,149	Text Document	2017/12/22 9:08	15B91140
shfe_lev2_debug_20171222_8.log	201,427,159	25,141,481	Text Document	2017/12/22 9:09	2DF94861
shfe_lev2_debug_20171222_9.log	201,427,541	25,151,749	Text Document	2017/12/22 9:10	666C0D1D
shfe_lev2_debug_20171222_10.log	201,427,412	25,192,096	Text Document	2017/12/22 9:11	F77BCE69
shfe_lev2_debug_20171222_11.log	201,427,449	25,276,414	Text Document	2017/12/22 9:12	9C3FBEA8
shfe_lev2_debug_20171222_12.log	201,426,736	25,114,907	Text Document	2017/12/22 9:13	48DA96F0
shfe_lev2_debug_20171222_13.log	201,426,014	25,167,170	Text Document	2017/12/22 9:15	2E3E32E8
shfe_lev2_debug_20171222_14.log	201,425,855	25,275,992	Text Document	2017/12/22 9:16	822A0B0A
shfe_lev2_debug_20171222_15.log	201,426,049	25,224,382	Text Document	2017/12/22 9:17	8FA7BEA1
shfe_lev2_debug_20171222_16.log	201,425,762	25,241,692	Text Document	2017/12/22 9:18	49C5838A
shfe_lev2_debug_20171222_17.log	201,425,431	25,205,182	Text Document	2017/12/22 9:19	A986679D
shfe_lev2_debug_20171222_18.log	201,425,748	25,219,757	Text Document	2017/12/22 9:20	0553771E
shfe_lev2_debug_20171222_19.log	201,425,581	25,244,401	Text Document	2017/12/22 9:21	37975977
shfe_lev2_debug_20171222_20.log	201,425,831	25,198,292	Text Document	2017/12/22 9:22	DE21DFA0
shfe_lev2_debug_20171222_21.log	201,425,926	25,225,519	Text Document	2017/12/22 9:23	FC17BEBD
shfe_lev2_debug_20171222_22.log	201,426,400	25,200,522	Text Document	2017/12/22 9:24	94D8DD0D
shfe_lev2_debug_20171222_23.log	201,425,759	25,222,424	Text Document	2017/12/22 9:25	FDAC59D7
shfe_lev2_debug_20171222_24.log	201,425,853	25,230,254	Text Document	2017/12/22 9:26	A7440F42
shfe_lev2_debug_20171222_25.log	201,426,033	25,212,660	Text Document	2017/12/22 9:27	FE28BE12
shfe_lev2_debug_20171222_26.log	201,425,739	25,196,459	Text Document	2017/12/22 9:28	5CEB8A69
shfe_lev2_debug_20171222_27.log	201,425,995	25,219,766	Text Document	2017/12/22 9:29	A1F800A5
shfe_lev2_debug_20171222_28.log	201,425,872	25,205,341	Text Document	2017/12/22 9:30	B9A1D8F0
shfe_lev2_debug_20171222_29.log	201,425,692	25,172,221	Text Document	2017/12/22 9:31	B248E5E7
shfe_lev2_debug_20171222_30.log	201,426,509	25,086,412	Text Document	2017/12/22 9:32	68E7C629
shfe_lev2_debug_20171222_31.log	201,425,951	25,143,929	Text Document	2017/12/22 9:33	26EB299F
shfe_lev2_debug_20171222_32.log	201,425,762	25,180,425	Text Document	2017/12/22 9:34	53E460D3
shfe_lev2_debug_20171222_33.log	[illegible]	[illegible]	Text Document	[illegible]	[illegible]

Selected 201,423,834 bytes in 1 file

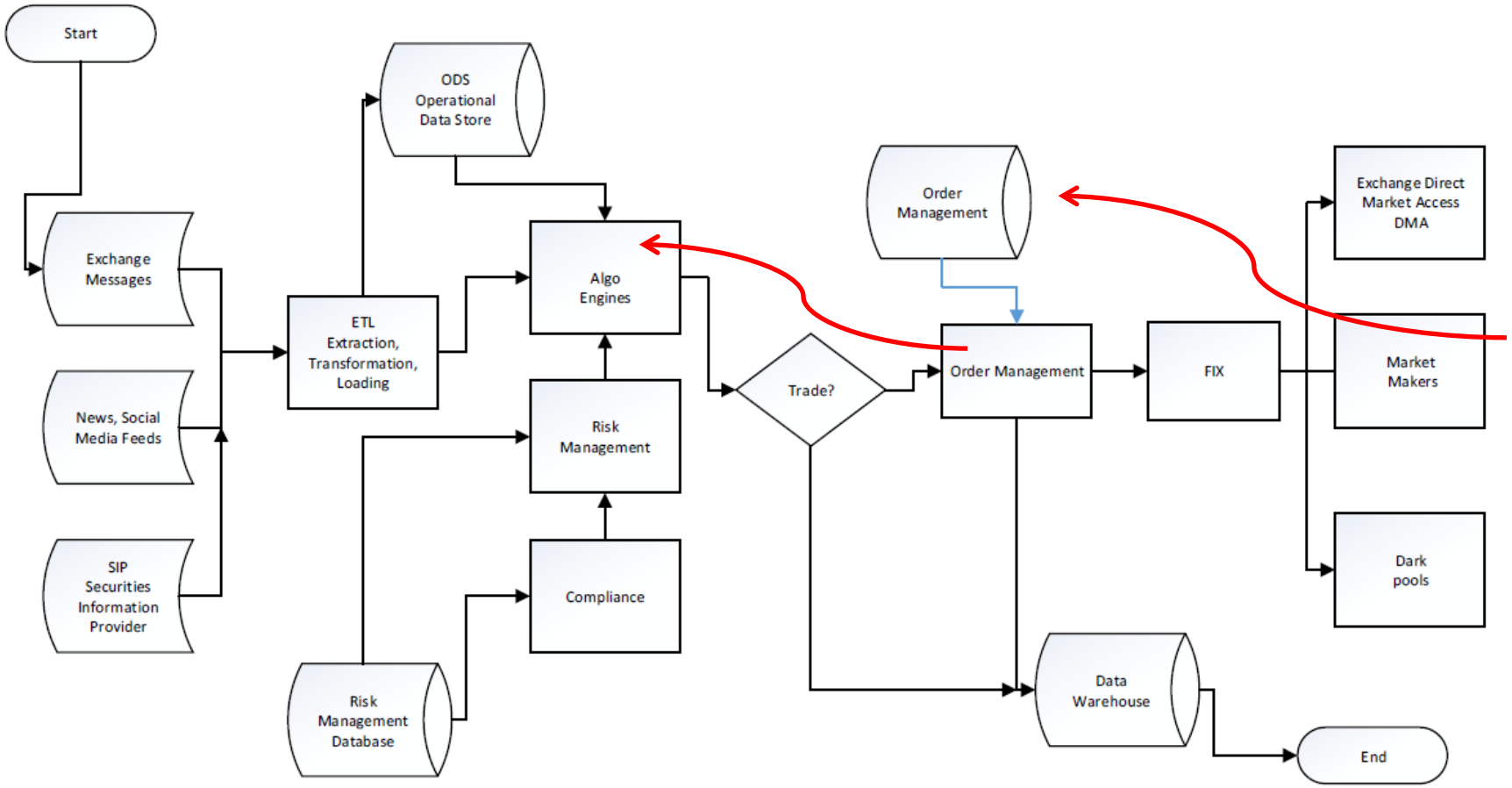
Total 42,661,117,642 bytes in 216 files

```
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[15035.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14285.0000],数量[4]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[15030.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14280.0000],数量[2]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[15025.0000],数量[15]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14270.0000],数量[3]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[15020.0000],数量[2]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14250.0000],数量[10]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[15010.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14240.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[15000.0000],数量[7]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14235.0000],数量[10]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14995.0000],数量[16]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14225.0000],数量[48]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14990.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14215.0000],数量[7]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14985.0000],数量[36]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14200.0000],数量[49]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14980.0000],数量[10]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14195.0000],数量[2]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14975.0000],数量[13]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14185.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14965.0000],数量[2]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14180.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14960.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14170.0000],数量[4]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14950.0000],数量[3]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14165.0000],数量[1]
[2017-12-22 14:39:10.537] lev2 node ask msg: [003,15723567],rb1812,14:39:46,500,5,10,[00:000:3803.0000:10],[01:001:3797.0000:4],[02:002:3733.000
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14945.0000],数量[9]
[2017-12-22 14:39:10.537] lev2 node time [ask:3:15723567] | 18472:22226:157909854:157887628 | 322123:529314:207191 | 159414072:159395600:1593733
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14150.0000],数量[26]
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14930.0000],数量[1]
[2017-12-22 14:39:10.537] lev2 node ask msg: [003,15723568],ru1801,14:39:46,500,5,5,[00:000:14915.0000:10],[01:001:14910.0000:6],[02:002:14905.0
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14145.0000],数量[3]
[2017-12-22 14:39:10.537] lev2 node time [ask:3:15723568] | 18472:22226:157909854:157887628 | 529853:603497:73644 | 159464579:159446107:15942388
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14925.0000],数量[1]
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14140.0000],数量[4]
[2017-12-22 14:39:10.537] lev2 node ask msg: [002,15723569],al1805,14:39:46,500,5,24,[00:003:15830.0000:1],[01:004:15820.0000:1],[02:005:15800.0
[2017-12-22 14:39:10.537] shfe_trade[2]: session idx:2, MBL:合约[al1805],方向[1],价格[14920.0000],数量[13]
[2017-12-22 14:39:10.537] lev2 node time [ask:2:15723569] | 21571:23090:158009279:157986189 | 32175:560283:528108 | 159558426:159536855:15951376
[2017-12-22 14:39:10.537] shfe_trade[3]: session idx:3, MBL:合约[ru1801],方向[1],价格[14130.0000],数量[2]
```

Objectives

- Predict the price movements of securities
- Construct a portfolio of selected securities
- Manage the risks of the portfolio

Workflow



Big Data, Fast: Database Requirements for High Frequency Trading, Ramesh K. Karne, 2017

Big Data is Perfect for Quant Strategies

- Central Limit Theorem
- https://nypost.com/2014/04/18/schneiderman-to-probe-virtus-htf-practices/

Strategy Categories

- Constraints
  - AUM
  - Research capabilities
  - Trading system, especially IT infrastructure

Frequency	High	Middle	Low
Upper limit of AUM	small	between	large
Expected Alpha Return	high	between	low
System Requirement	high	between	low

Multi Strategies

- No strategy works forever
- Each strategy has its good and bad times
- Most institutional investors diversify their investment portfolios with multi strategies
- Strategies with portable alphas can work together

Raw Data Treatment

- Slice and group data from raw data into different frequencies
- Typical example is the daily stock price time series
- Choices of how to summarize the discarded information between sample points
- Irregular and regular information handling
- Data cleaning

Data Storage Choices

- Relational databases
- NoSQL databases
- Time Series Databases

Relational databases

- Database engine supports concurrent transactions
- Relational model and schema
- Structured Query Languages (SQL)

Concurrent Transactions

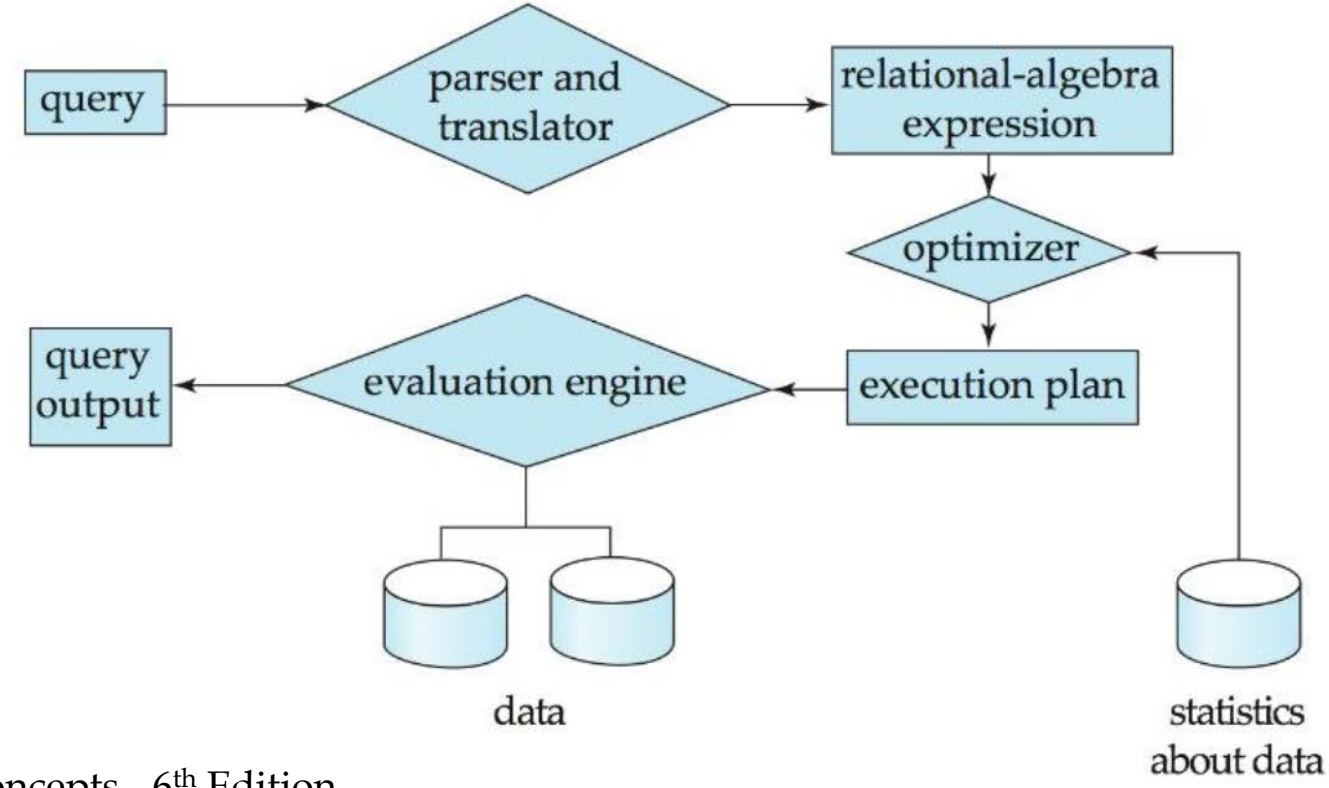
- Properties known as ACID:
  - **Atomicity**: No inconsistent state with partial updates of transactions.
  - **Consistency**: No illegal transactions.
  - **Isolation**: Concurrently executed transactions need to be isolated.
  - **Durability**: Committed transactions will stay.

Database Engine Components

- Storage manager
  - the interface between the database and the operating system
  - responsible for authorization, interaction with the OS file system (accessing storage and organizing files), and efficient data storage/modification (indexing, hashing, buffer management).
- Transaction manager
  - ensures the database is consistent (if a failure occurs) and ACID properties

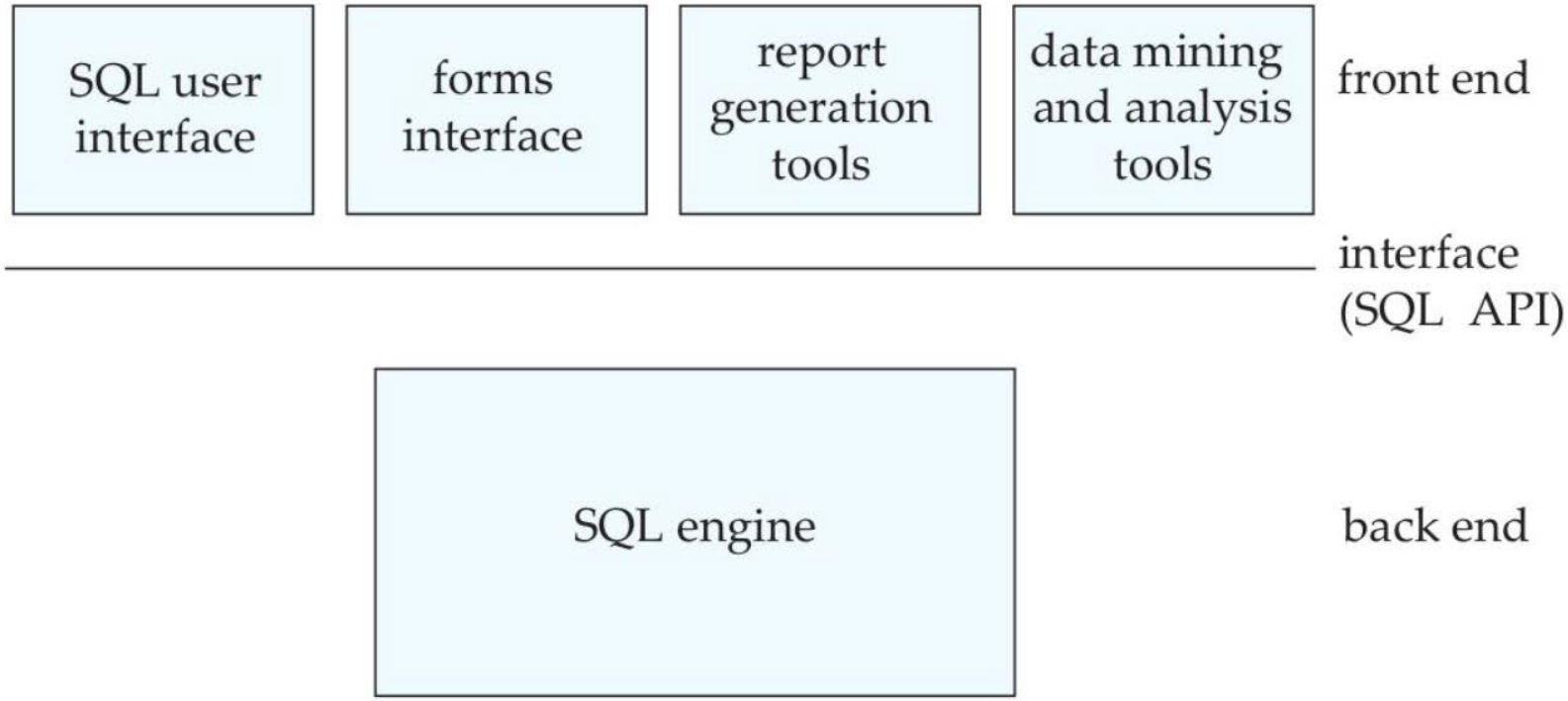
Query Processor

- Three major jobs: parsing and translation, optimization, and evaluation



Database System Concepts - 6th Edition

SQL API as Interface



Database System Concepts - 6th Edition

Relation Example: Instructor

attributes
(or columns)

ID	*name*	*dept_name*	*salary*
10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	60000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	62000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

tuples
(or rows)

Schema Example: Instructor

- Database schema -- is the logical structure of the database.
- Database instance -- is a snapshot of the data in the database at a given instant in time.
- Example:
  - schema: *instructor* (*ID, name, dept_name, salary*)
  - Instance:

ID	*name*	*dept_name*	*salary*
22222	Einstein	Physics	95000
12121	Wu	Finance	90000
32343	El Said	History	60000
45565	Katz	Comp. Sci.	75000
98345	Kim	Elec. Eng.	80000
76766	Crick	Biology	72000
10101	Srinivasan	Comp. Sci.	65000
58583	Califieri	History	62000
83821	Brandt	Comp. Sci.	92000
15151	Mozart	Music	40000
33456	Gold	Physics	87000
76543	Singh	Finance	80000

SQL Query Example

- The **where** clause specifies conditions that the result must satisfy
  - Corresponds to the selection predicate of the relational algebra.
- To find all instructors in Comp. Sci. dept

```
select name
from instructor
where dept_name = 'Comp. Sci.'
```

- SQL allows the use of the logical connectives **and, or,** and **not**
- The operands of the logical connectives can be expressions involving the comparison operators <, <=, >, >=, =, and <>.
- Comparisons can be applied to results of arithmetic expressions
- To find all instructors in Comp. Sci. dept with salary > 80000

```
select name
from instructor
where dept_name = 'Comp. Sci.'  and salary > 80000
```