Data Engineering Week2 Homework

17기 신예진

RDBMS와 Hadoop의 특징, 공통점과 차이점

RDBMS

- Relational DataBase Management System 의 약자로 관계형 데이터베이스를 생성하고 관리할 수 있는 소프트웨어이다.
- RDBMS 내의 모든 데이터는 2 차원 테이블 형태로 표현되며, 테이블은 다른 테이블과 관계를 맺고 있다.
- 관계를 나타내기 위해 foreign key 를 사용하며, 이 foreign key 를 이용하여 join 이 가능하다.

Hadoop

- 분산 환경에서 빅데이터를 저장하고 처리할 수 있는 자바 기반의 오픈소스 프레임워크이다.
- 여러 대의 컴퓨터 클러스터에 대규모 데이터셋을 분산 처리할 수 있게 해주는 프레임워크이다.
- hdfs 와 mapreduce 를 시작으로 하둡생태계가 확장, 발전되었다.

공통점

- 데이터를 저장하는 용도로 사용되며, 정형 데이터를 처리할 수 있다는 공통점이 있다.

차이점

	RDBMS	Hadoop
데이터 다양성	구조화된 데이터 저장	여러가지 형태의 데이터 저장(정형, 반정형, 비정형)
데이터 저장고	평균 데이터 양을 저장	훨씬 데이터 양을 저장
속도	읽기가 빠름	읽기 및 쓰기가 빠름
확장성	수직확장성	수평확장성
하드웨어	고급 서버 사용	상용 하드웨어 사용
처리량	처리량이 높음	처리량이 낮음

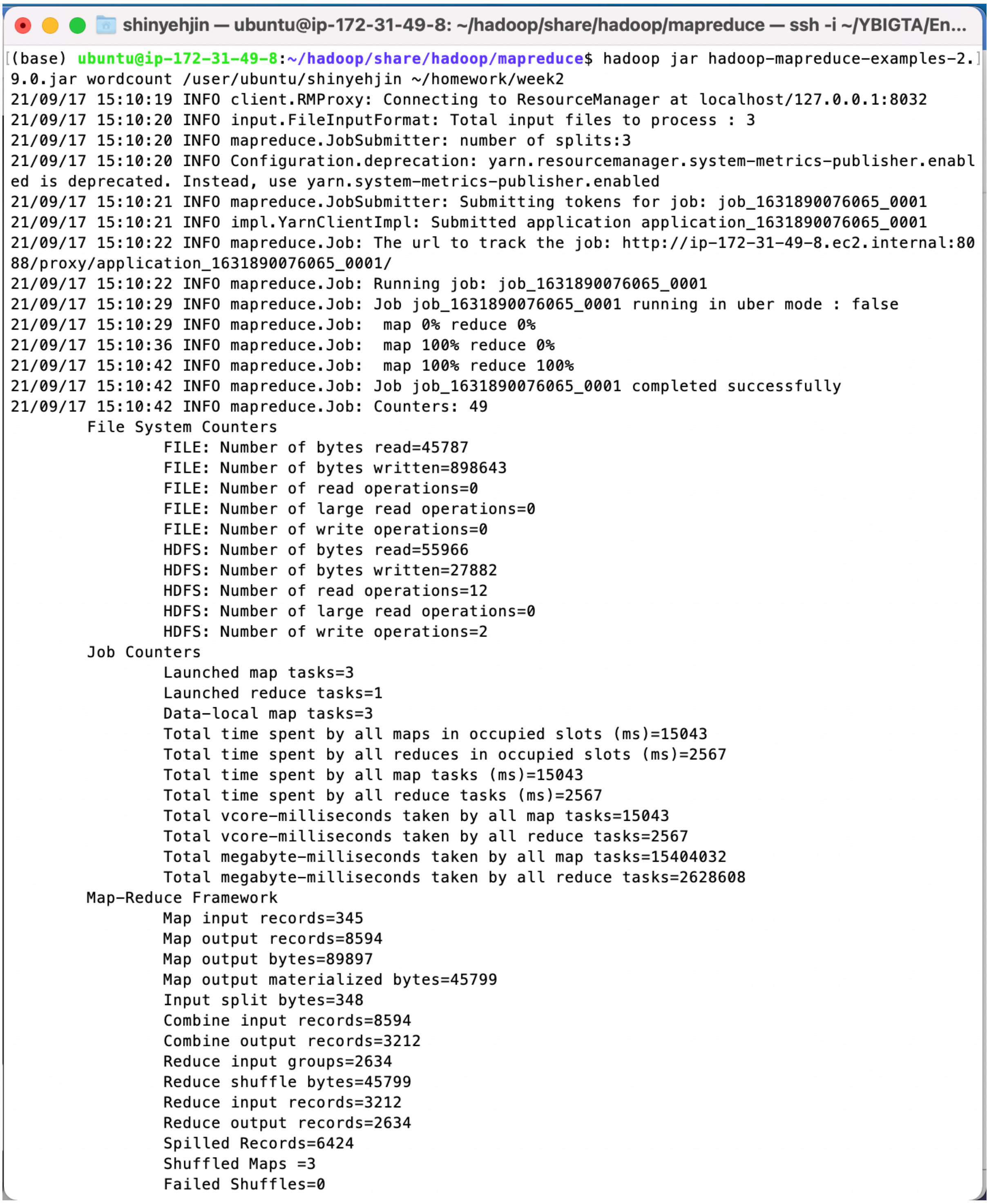
출처

https://khj93.tistory.com/entry/Database-RDBMS 와-NOSQL-차이점

https://ko.strephonsays.com/rdbms-and-hadoop-8502

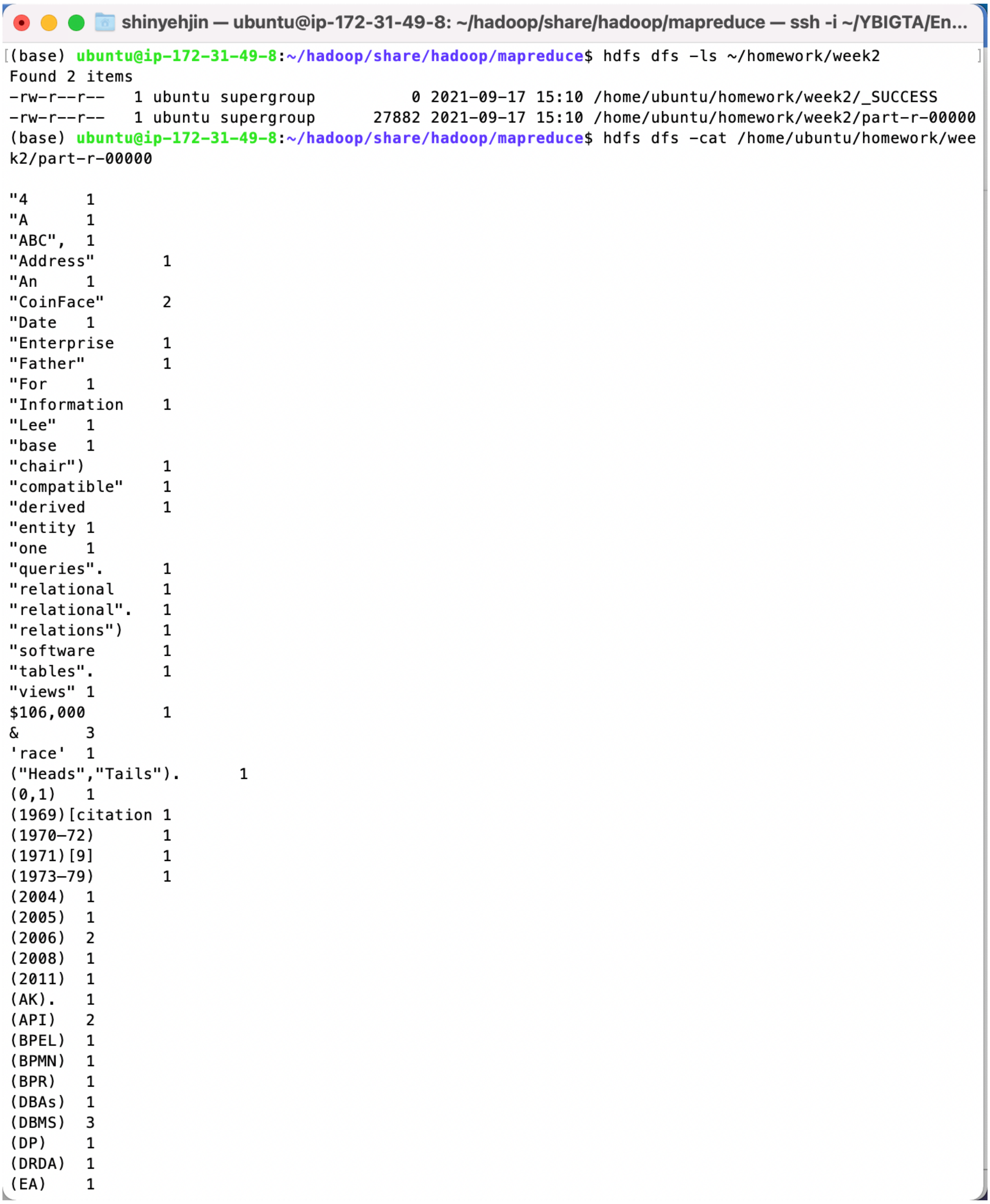
Wordcounter 실행

wordcounter


shinyehjin — ubuntu@ip-172-31-49-8: ~/hadoop/share/hadoop/mapreduce — ssh -i ~/YBIGTA/En...

```
(base) ubuntu@ip-172-31-49-8:~/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-2.
9.0.jar wordcount /user/ubuntu/shinyehjin ~/homework/week2
21/09/17 15:10:19 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
21/09/17 15:10:20 INFO input.FileInputFormat: Total input files to process : 3
21/09/17 15:10:20 INFO mapreduce.JobSubmitter: number of splits:3
21/09/17 15:10:20 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabl
ed is deprecated. Instead, use yarn.system-metrics-publisher.enabled
21/09/17 15:10:21 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1631890076065_0001
21/09/17 15:10:21 INFO impl.YarnClientImpl: Submitted application application_1631890076065_0001
21/09/17 15:10:22 INFO mapreduce.Job: The url to track the job: http://ip-172-31-49-8.ec2.internal:80
88/proxy/application_1631890076065_0001/
21/09/17 15:10:22 INFO mapreduce.Job: Running job: job_1631890076065_0001
21/09/17 15:10:29 INFO mapreduce.Job: Job job_1631890076065_0001 running in uber mode : false
21/09/17 15:10:29 INFO mapreduce.Job:  map 0% reduce 0%
21/09/17 15:10:36 INFO mapreduce.Job:  map 100% reduce 0%
21/09/17 15:10:42 INFO mapreduce.Job:  map 100% reduce 100%
21/09/17 15:10:42 INFO mapreduce.Job: Job job_1631890076065_0001 completed successfully
21/09/17 15:10:42 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=45787
                FILE: Number of bytes written=898643
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=55966
                HDFS: Number of bytes written=27882
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=15043
                Total time spent by all reduces in occupied slots (ms)=2567
                Total time spent by all map tasks (ms)=15043
                Total time spent by all reduce tasks (ms)=2567
                Total vcore-milliseconds taken by all map tasks=15043
                Total vcore-milliseconds taken by all reduce tasks=2567
                Total megabyte-milliseconds taken by all map tasks=15404032
                Total megabyte-milliseconds taken by all reduce tasks=2628608
        Map-Reduce Framework
                Map input records=345
                Map output records=8594
                Map output bytes=89897
                Map output materialized bytes=45799
                Input split bytes=348
                Combine input records=8594
                Combine output records=3212
                Reduce input groups=2634
                Reduce shuffle bytes=45799
                Reduce input records=3212
                Reduce output records=2634
                Spilled Records=6424
                Shuffled Maps =3
                Failed Shuffles=0
```

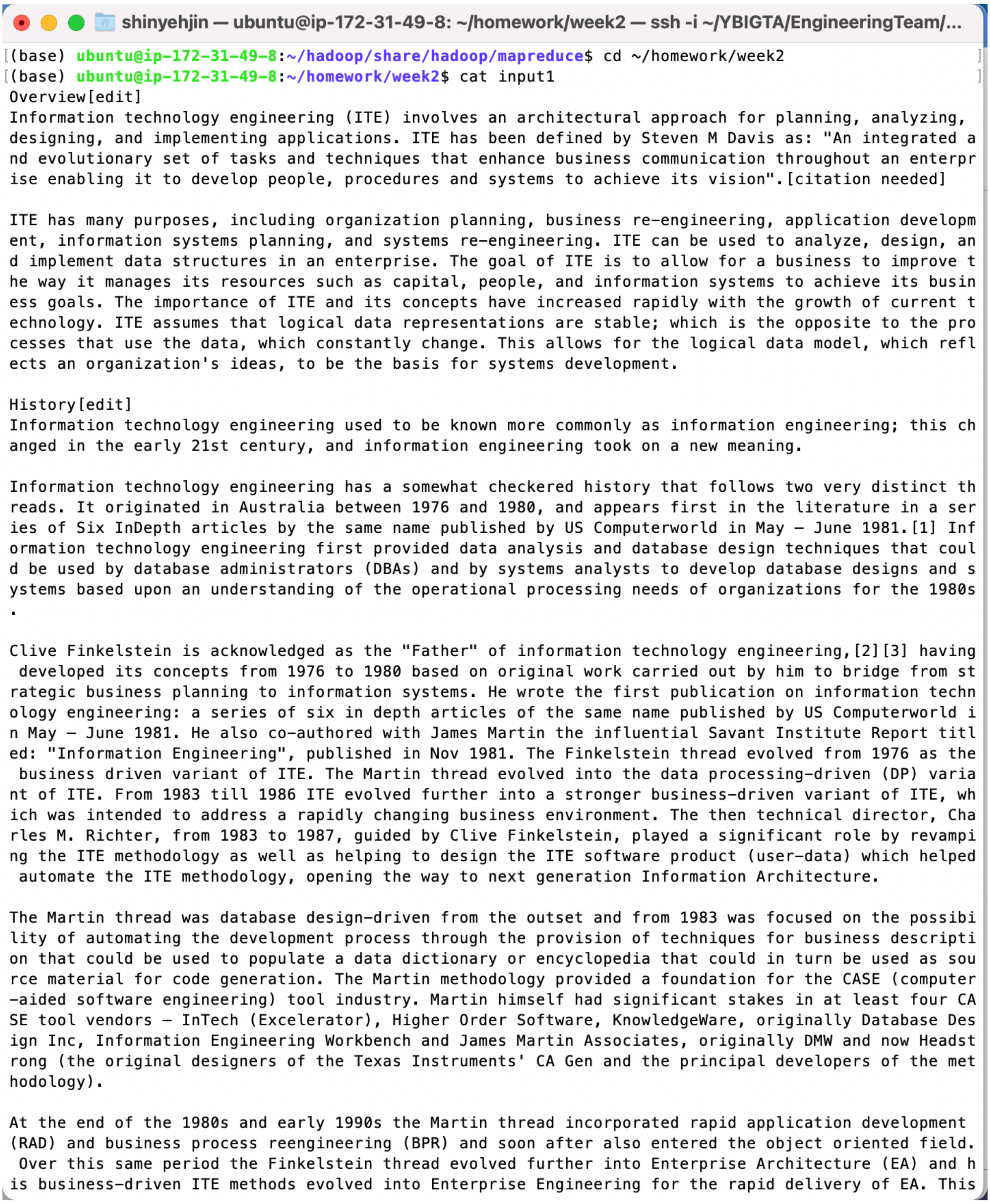
결과파일 스크린샷 (폴더명은 제맘대로 했어요..ㅎ)


shinyehjin — ubuntu@ip-172-31-49-8: ~/hadoop/share/hadoop/mapreduce — ssh -i ~/YBIGTA/En...

```
(base) ubuntu@ip-172-31-49-8:~/hadoop/share/hadoop/mapreduce$ hdfs dfs -ls ~/homework/week2
Found 2 items
-rw-r--r--   1 ubuntu supergroup          0 2021-09-17 15:10 /home/ubuntu/homework/week2/_SUCCESS
-rw-r--r--   1 ubuntu supergroup      27882 2021-09-17 15:10 /home/ubuntu/homework/week2/part-r-00000
(base) ubuntu@ip-172-31-49-8:~/hadoop/share/hadoop/mapreduce$ hdfs dfs -cat /home/ubuntu/homework/wee
k2/part-r-00000

"4      1
"A      1
"ABC",  1
"Address"       1
"An     1
"CoinFace"      2
"Date   1
"Enterprise     1
"Father"        1
"For    1
"Information     1
"Lee"   1
"base   1
"chair")        1
"compatible"    1
"derived        1
"entity 1
"one    1
"queries".      1
"relational     1
"relational".   1
"relations")    1
"software       1
"tables".       1
"views" 1
$106,000        1
&       3
'race'  1
("Heads","Tails").      1
(0,1)   1
(1969)[citation 1
(1970–72)       1
(1971)[9]       1
(1973–79)       1
(2004)  1
(2005)  1
(2006)  2
(2008)  1
(2011)  1
(AK).   1
(API)   2
(BPEL)  1
(BPMN)  1
(BPR)   1
(DBAs)  1
(DBMS)  3
(DP)    1
(DRDA)  1
(EA)    1
```

input 파일 일부 (wiki 아무거나 3 개 긁어서 사용했습니다)


shinyehjin — ubuntu@ip-172-31-49-8: ~/homework/week2 — ssh -i ~/YBIGTA/EngineeringTeam/...

```
(base) ubuntu@ip-172-31-49-8:~/hadoop/share/hadoop/mapreduce$ cd ~/homework/week2
(base) ubuntu@ip-172-31-49-8:~/homework/week2$ cat input1
Overview[edit]
Information technology engineering (ITE) involves an architectural approach for planning, analyzing,
designing, and implementing applications. ITE has been defined by Steven M Davis as: "An integrated a
nd evolutionary set of tasks and techniques that enhance business communication throughout an enterpr
ise enabling it to develop people, procedures and systems to achieve its vision".[citation needed]

ITE has many purposes, including organization planning, business re-engineering, application developm
ent, information systems planning, and systems re-engineering. ITE can be used to analyze, design, an
d implement data structures in an enterprise. The goal of ITE is to allow for a business to improve t
he way it manages its resources such as capital, people, and information systems to achieve its busin
ess goals. The importance of ITE and its concepts have increased rapidly with the growth of current t
echnology. ITE assumes that logical data representations are stable; which is the opposite to the pro
cesses that use the data, which constantly change. This allows for the logical data model, which refl
ects an organization's ideas, to be the basis for systems development.

History[edit]
Information technology engineering used to be known more commonly as information engineering; this ch
anged in the early 21st century, and information engineering took on a new meaning.

Information technology engineering has a somewhat checkered history that follows two very distinct th
reads. It originated in Australia between 1976 and 1980, and appears first in the literature in a ser
ies of Six InDepth articles by the same name published by US Computerworld in May – June 1981.[1] Inf
ormation technology engineering first provided data analysis and database design techniques that coul
d be used by database administrators (DBAs) and by systems analysts to develop database designs and s
ystems based upon an understanding of the operational processing needs of organizations for the 1980s
.

Clive Finkelstein is acknowledged as the "Father" of information technology engineering,[2][3] having
 developed its concepts from 1976 to 1980 based on original work carried out by him to bridge from st
rategic business planning to information systems. He wrote the first publication on information techn
ology engineering: a series of six in depth articles of the same name published by US Computerworld i
n May – June 1981. He also co-authored with James Martin the influential Savant Institute Report titl
ed: "Information Engineering", published in Nov 1981. The Finkelstein thread evolved from 1976 as the
 business driven variant of ITE. The Martin thread evolved into the data processing-driven (DP) varia
nt of ITE. From 1983 till 1986 ITE evolved further into a stronger business-driven variant of ITE, wh
ich was intended to address a rapidly changing business environment. The then technical director, Cha
rles M. Richter, from 1983 to 1987, guided by Clive Finkelstein, played a significant role by revampi
ng the ITE methodology as well as helping to design the ITE software product (user-data) which helped
 automate the ITE methodology, opening the way to next generation Information Architecture.

The Martin thread was database design-driven from the outset and from 1983 was focused on the possibi
lity of automating the development process through the provision of techniques for business descripti
on that could be used to populate a data dictionary or encyclopedia that could in turn be used as sou
rce material for code generation. The Martin methodology provided a foundation for the CASE (computer
-aided software engineering) tool industry. Martin himself had significant stakes in at least four CA
SE tool vendors – InTech (Excelerator), Higher Order Software, KnowledgeWare, originally Database Des
ign Inc, Information Engineering Workbench and James Martin Associates, originally DMW and now Headst
rong (the original designers of the Texas Instruments' CA Gen and the principal developers of the met
hodology).

At the end of the 1980s and early 1990s the Martin thread incorporated rapid application development
(RAD) and business process reengineering (BPR) and soon after also entered the object oriented field.
 Over this same period the Finkelstein thread evolved further into Enterprise Architecture (EA) and h
is business-driven ITE methods evolved into Enterprise Engineering for the rapid delivery of EA. This
```