Big Data Management: Past, Present, and Future

Kelly Magruder

What is Big Data?

- Extremely **large data sets** that may be **analyzed computationally** to reveal patterns, trends, and associations

Need for Management Systems

- "Every two days now we create as much information as we did from the dawn of civilization up until 2003. That's something like five exabytes of data."
  - Google CEO Erik Schmidt in 2010 (even larger now)

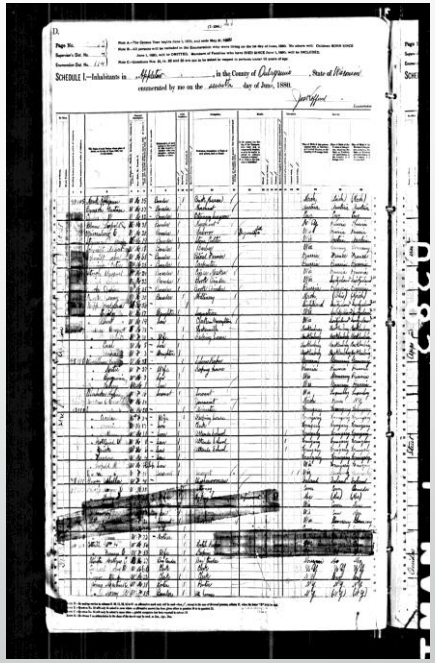
- Gigabyte = 10^9 = 1,000,000,000
  Terabyte = 10^12 = 1,000,000,000,000
  Petabyte = 10^15 = 1,000,000,000,000,000
  Exabyte = 10^18 = 1,000,000,000,000,000,000

Management System Goals

- The Four V's of Big Data:
  - Volume – Scale of data
  - Variety – Different forms of data
  - Velocity – Analysis of streaming data
  - Veracity – Uncertainty and accuracy of data

First Big Data Problem and Solution

- 1880 Census took eight years to tabulate 50 million people
- The Hollerith Tabulating Machine punched cards with 80 variables and completed the 1890 Census in 6 weeks

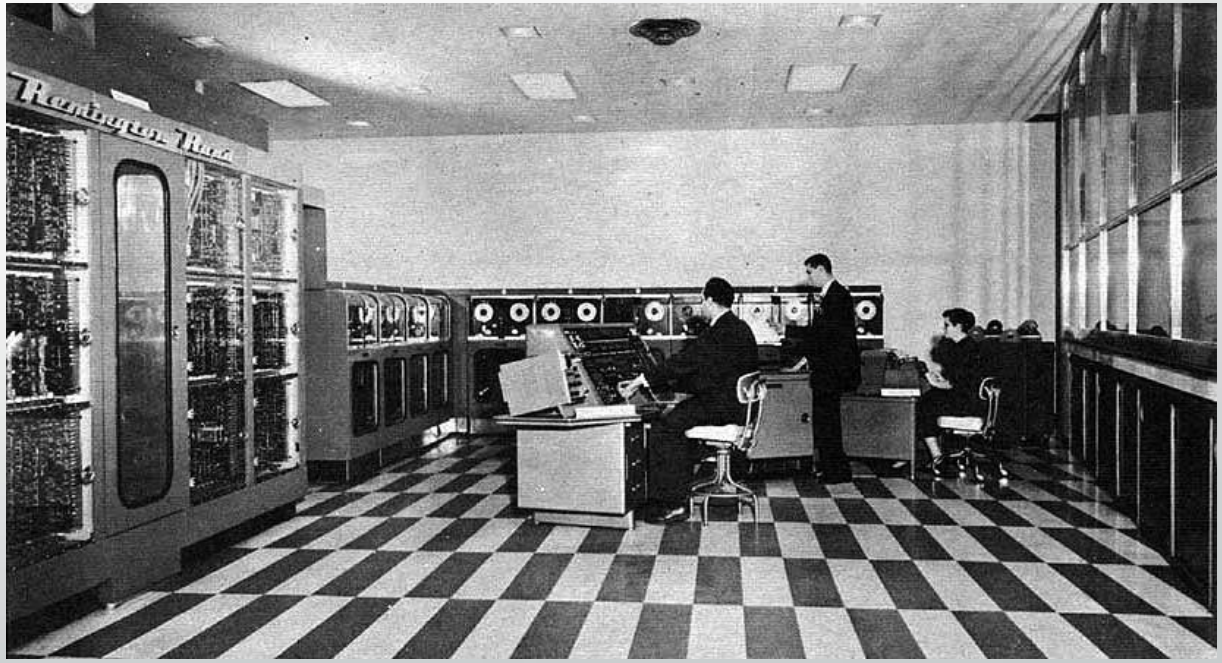


Information Overload

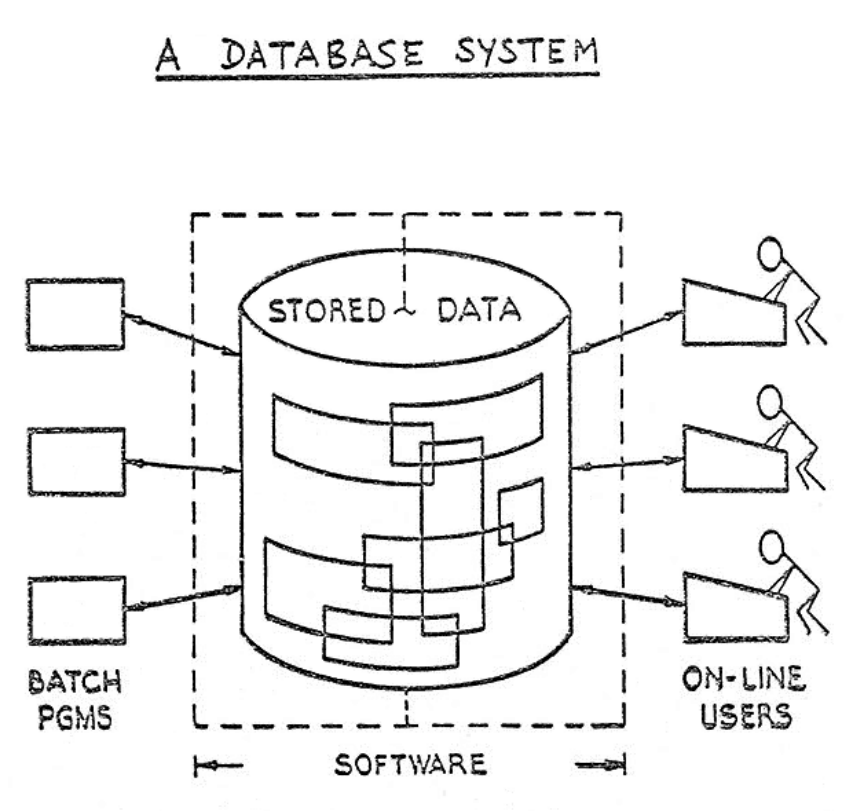
- Population boom in the 1930's
  - Census, social security numbers, new research
- Libraries adapted new storage methods to meet demand of publications
- In 1944, libraries were predicted to double every sixteen years

Centralized Computing Systems

- 1960s: Information and research had reached the point of critical magnitude (Information Scientist Derek Price)
- Organizations developed electronic automation of their inventory systems

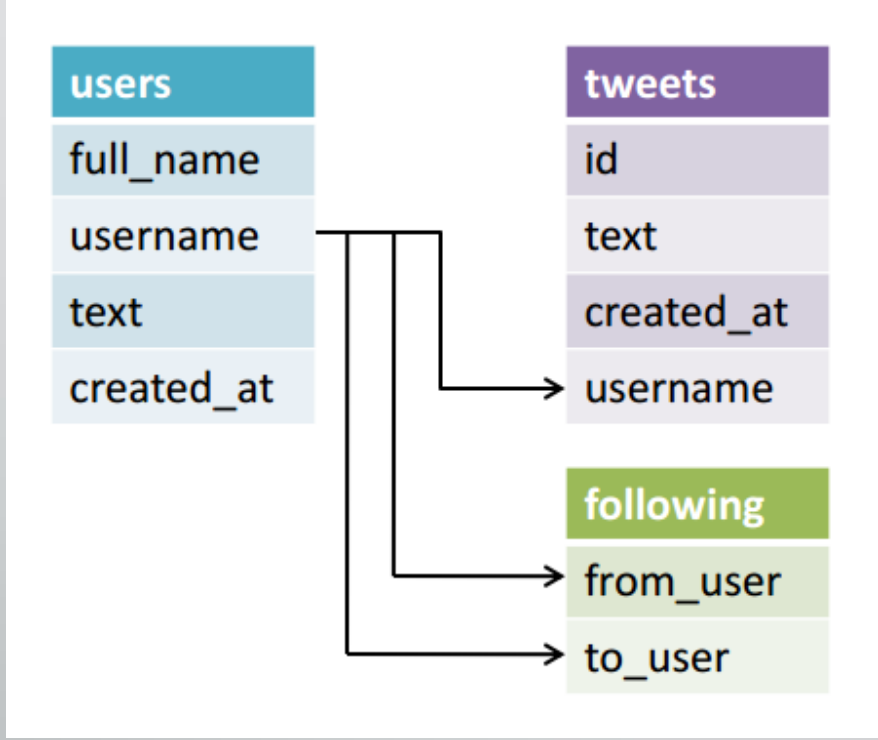


The Relational Database

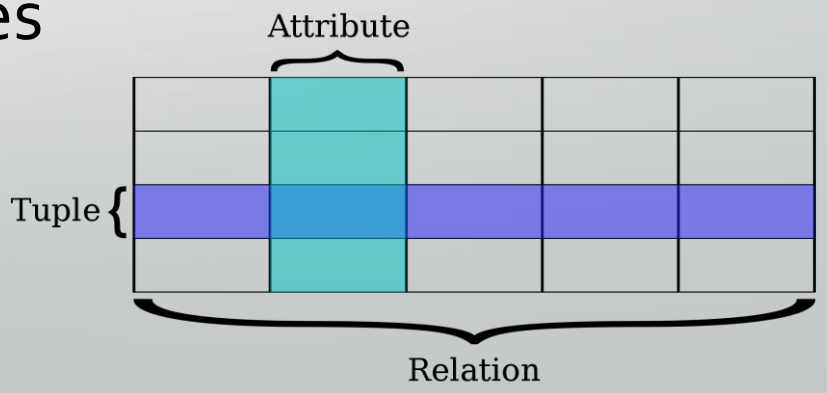


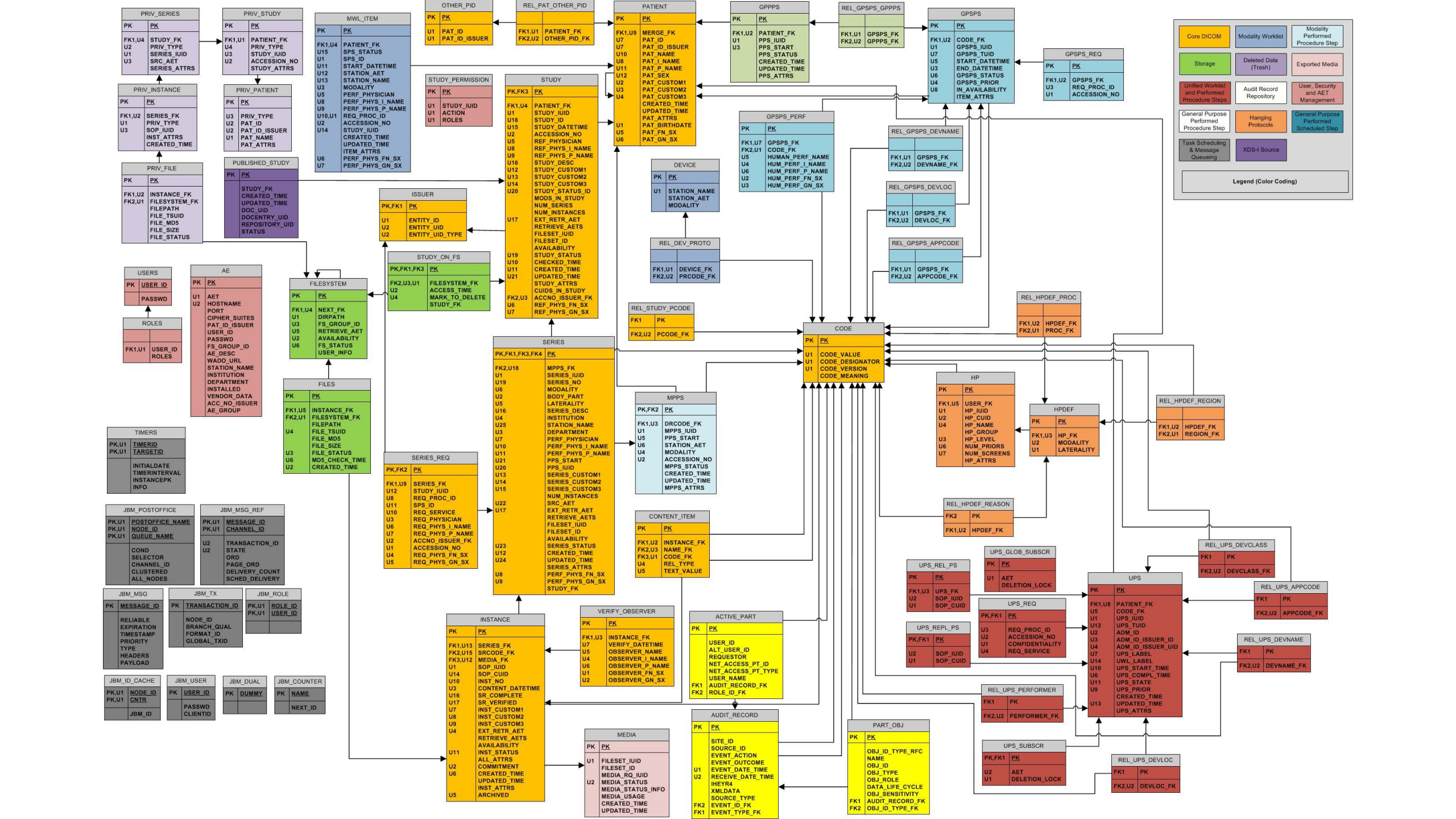
- 1970: IBM Research Lab published a paper showing how information in large databases could be easily accessed
- Modern data transactions all use structures based on relational database theory

The Relational Database



- Possible schema of Twitter's database
- Three tables: users, tweets, following
- Each table has several attributes



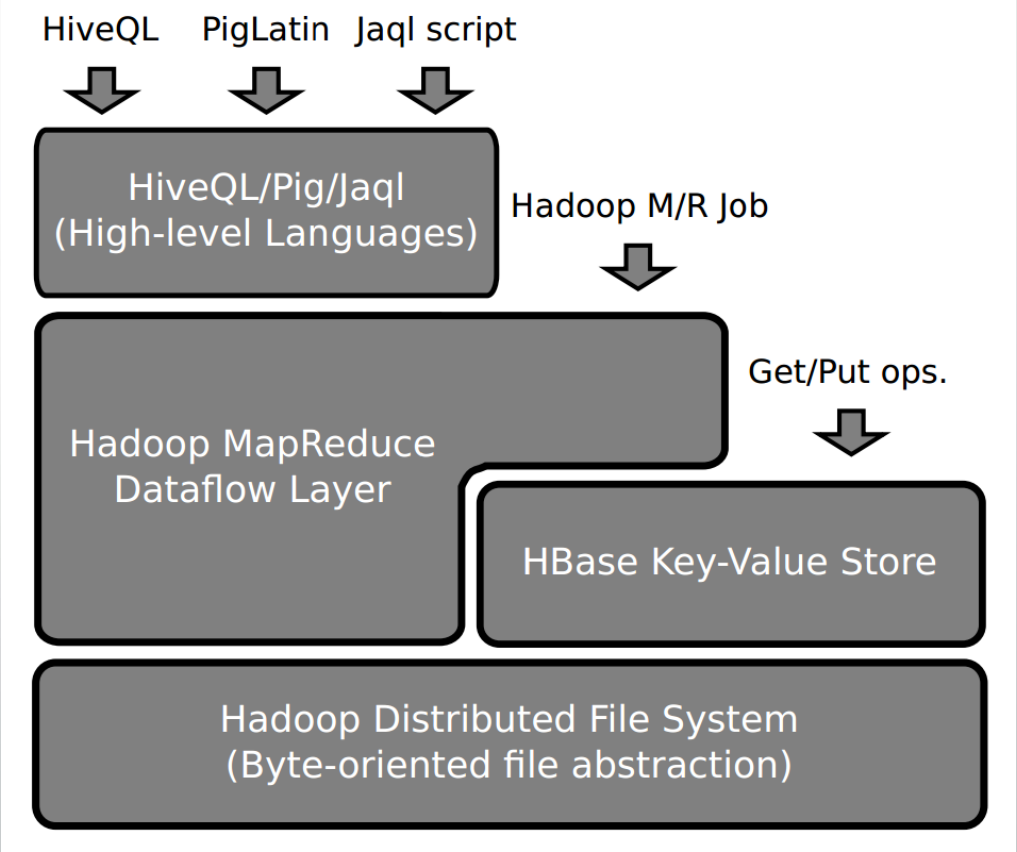

PRIV_SERIES
PK PK
FK1,U4 STUDY_FK
U2 PRIV_TYPE
U1 SERIES_IUID
U3 SRC_AET
SERIES_ATTRS
PRIV_STUDY
PK PK
FK1,U1 PATIENT_FK
U4 PRIV_TYPE
U3 STUDY_IUID
U2 ACCESSION_NO
STUDY_ATTRS
PRIV_INSTANCE
PK PK
FK1,U2 SERIES_FK
U1 PRIV_TYPE
U3 SOP_IUID
INST_ATTRS
CREATED_TIME
PRIV_PATIENT
PK PK
U3 PRIV_TYPE
U2 PAT_ID
U2 PAT_ID_ISSUER
U1 PAT_NAME
PAT_ATTRS
PRIV_FILE
PK PK
FK1,U2 INSTANCE_FK
FK2,U1 FILESYSTEM_FK
FILEPATH
FILE_TSUID
FILE_MD5
FILE_SIZE
FILE_STATUS
MWL_ITEM
PK PK
FK1,U4 PATIENT_FK
U15 SPS_STATUS
U1 SPS_ID
U11 START_DATETIME
U12 STATION_AET
U13 STATION_NAME
U3 MODALITY
U5 PERF_PHYSICIAN
U8 PERF_PHYS_I_NAME
U9 PERF_PHYS_P_NAME
U10,U1 REQ_PROC_ID
U2 ACCESSION_NO
U14 STUDY_IUID
CREATED_TIME
UPDATED_TIME
ITEM_ATTRS
U6 PERF_PHYS_FN_SX
U7 PERF_PHYS_GN_SX
OTHER_PID
PK PK
U1 PAT_ID
U1 PAT_ID_ISSUER
REL_PAT_OTHER_PID
FK1,U1 PATIENT_FK
FK2,U2 OTHER_PID_FK
PATIENT
PK PK
FK1,U9 MERGE_FK
U7 PAT_ID
U7 PAT_ID_ISSUER
U10 PAT_NAME
U8 PAT_I_NAME
U11 PAT_P_NAME
U12 PAT_SEX
U2 PAT_CUSTOM1
U3 PAT_CUSTOM2
U4 PAT_CUSTOM3
CREATED_TIME
UPDATED_TIME
PAT_ATTRS
U1 PAT_BIRTHDATE
U5 PAT_FN_SX
U6 PAT_GN_SX
GPPPS
PK PK
FK1,U2 PATIENT_FK
U1 PPS_IUID
U3 PPS_START
PPS_STATUS
CREATED_TIME
UPDATED_TIME
PPS_ATTRS
REL_GPSPS_GPPPS
FK1,U1 GPSPS_FK
FK2,U2 GPPPS_FK
GPSPS
PK PK
FK1,U2 CODE_FK
U1 GPSPS_IUID
U7 GPSPS_TUID
U5 START_DATETIME
U3 END_DATETIME
U6 GPSPS_STATUS
U4 GPSPS_PRIOR
IN_AVAILABILITY
U8 ITEM_ATTRS
GPSPS_REQ
PK PK
FK1,U2 GPSPS_FK
U3 REQ_PROC_ID
U1 ACCESSION_NO
Core DICOM
Modality Worklist
Modality Performed Procedure Step
Storage
Deleted Data (Trash)
Exported Media
Unified Worklist and Performed Procedure Steps
Audit Record Repository
User, Security and AET Management
General Purpose Performed Procedure Step
Hanging Protocols
General Purpose Performed Scheduled Step
Task Scheduling & Message Queueing
XDS-I Source
Legend (Color Coding)
STUDY_PERMISSION
PK PK
U1 STUDY_IUID
U1 ACTION
U1 ROLES
STUDY
PK,FK3 PK
FK1,U4 PATIENT_FK
U1 STUDY_IUID
U18 STUDY_ID
U15 STUDY_DATETIME
U2 ACCESSION_NO
U5 REF_PHYSICIAN
U8 REF_PHYS_I_NAME
U9 REF_PHYS_P_NAME
U16 STUDY_DESC
U12 STUDY_CUSTOM1
U13 STUDY_CUSTOM2
U14 STUDY_CUSTOM3
U20 STUDY_STATUS_ID
MODS_IN_STUDY
NUM_SERIES
NUM_INSTANCES
U17 EXT_RETR_AET
RETRIEVE_AETS
FILESET_IUID
FILESET_ID
AVAILABILITY
U19 STUDY_STATUS
U10 CHECKED_TIME
U11 CREATED_TIME
U21 UPDATED_TIME
STUDY_ATTRS
CUIDS_IN_STUDY
FK2,U3 ACCNO_ISSUER_FK
U6 REF_PHYS_FN_SX
U7 REF_PHYS_GN_SX
GPSPS_PERF
PK PK
FK1,U7 GPSPS_FK
FK2,U1 CODE_FK
U5 HUMAN_PERF_NAME
U4 HUM_PERF_I_NAME
U6 HUM_PERF_P_NAME
U2 HUM_PERF_FN_SX
U3 HUM_PERF_GN_SX
REL_GPSPS_DEVNAME
FK1,U1 GPSPS_FK
FK2,U2 DEVNAME_FK
REL_GPSPS_DEVLOC
FK1,U1 GPSPS_FK
FK2,U2 DEVLOC_FK
REL_GPSPS_APPCODE
FK1,U1 GPSPS_FK
FK2,U2 APPCODE_FK
DEVICE
PK PK
U1 STATION_NAME
STATION_AET
MODALITY
REL_DEV_PROTO
FK1,U1 DEVICE_FK
FK2,U2 PRCODE_FK
ISSUER
PK,FK1 PK
U1 ENTITY_ID
U2 ENTITY_UID
U2 ENTITY_UID_TYPE
STUDY_ON_FS
PK,FK1,FK3 PK
FK2,U3,U1 FILESYSTEM_FK
U2 ACCESS_TIME
U4 MARK_TO_DELETE
STUDY_FK
USERS
PK USER_ID
PASSWD
ROLES
FK1,U1 USER_ID
ROLES
AE
PK PK
U1 AET
U2 HOSTNAME
PORT
CIPHER_SUITES
PAT_ID_ISSUER
USER_ID
PASSWD
FS_GROUP_ID
AE_DESC
WADO_URL
STATION_NAME
INSTITUTION
DEPARTMENT
INSTALLED
VENDOR_DATA
ACC_NO_ISSUER
AE_GROUP
FILESYSTEM
PK PK
FK1,U4 NEXT_FK
U1 DIRPATH
U3 FS_GROUP_ID
U5 RETRIEVE_AET
U2 AVAILABILITY
U6 FS_STATUS
USER_INFO
FILES
PK PK
FK1,U5 INSTANCE_FK
FK2,U1 FILESYSTEM_FK
FILEPATH
U4 FILE_TSUID
FILE_MD5
FILE_SIZE
U3 FILE_STATUS
U6 MD5_CHECK_TIME
U2 CREATED_TIME
REL_STUDY_PCODE
FK1 PK
FK2,U2 PCODE_FK
CODE
PK PK
U1 CODE_VALUE
U1 CODE_DESIGNATOR
U1 CODE_VERSION
CODE_MEANING
REL_HPDEF_PROC
FK1,U2 HPDEF_FK
FK2,U1 PROC_FK
SERIES
PK,FK1,FK3,FK4 PK
FK2,U18 MPPS_FK
U1 SERIES_IUID
U19 SERIES_NO
U6 MODALITY
U2 BODY_PART
U5 LATERALITY
U16 SERIES_DESC
U4 INSTITUTION
U25 STATION_NAME
U3 DEPARTMENT
U7 PERF_PHYSICIAN
U10 PERF_PHYS_I_NAME
U11 PERF_PHYS_P_NAME
U21 PPS_START
U20 PPS_IUID
U13 SERIES_CUSTOM1
U14 SERIES_CUSTOM2
U15 SERIES_CUSTOM3
NUM_INSTANCES
U22 SRC_AET
U17 EXT_RETR_AET
RETRIEVE_AETS
FILESET_IUID
FILESET_ID
AVAILABILITY
U23 SERIES_STATUS
U12 CREATED_TIME
U24 UPDATED_TIME
SERIES_ATTRS
U8 PERF_PHYS_FN_SX
U9 PERF_PHYS_GN_SX
STUDY_FK
MPPS
PK,FK2 PK
FK1,U3 DRCODE_FK
U1 MPPS_IUID
U5 PPS_START
U6 STATION_AET
U4 MODALITY
U2 ACCESSION_NO
MPPS_STATUS
CREATED_TIME
UPDATED_TIME
MPPS_ATTRS
HP
PK PK
FK1,U5 USER_FK
U1 HP_IUID
U2 HP_CUID
U4 HP_NAME
HP_GROUP
U3 HP_LEVEL
U6 NUM_PRIORS
U7 NUM_SCREENS
HP_ATTRS
HPDEF
PK PK
FK1,U3 HP_FK
U2 MODALITY
U1 LATERALITY
REL_HPDEF_REGION
FK1,U2 HPDEF_FK
FK2,U1 REGION_FK
TIMERS
PK,U1 TIMERID
PK,U1 TARGETID
INITIALDATE
TIMERINTERVAL
INSTANCEPK
INFO
SERIES_REQ
PK,FK2 PK
FK1,U9 SERIES_FK
U12 STUDY_IUID
U8 REQ_PROC_ID
U11 SPS_ID
U10 REQ_SERVICE
U3 REQ_PHYSICIAN
U6 REQ_PHYS_I_NAME
U7 REQ_PHYS_P_NAME
U2 ACCNO_ISSUER_FK
U1 ACCESSION_NO
U4 REQ_PHYS_FN_SX
U5 REQ_PHYS_GN_SX
CONTENT_ITEM
PK PK
FK1,U2 INSTANCE_FK
FK2,U3 NAME_FK
FK3,U1 CODE_FK
U4 REL_TYPE
U5 TEXT_VALUE
REL_HPDEF_REASON
FK2 PK
FK1,U2 HPDEF_FK
JBM_POSTOFFICE
PK,U1 POSTOFFICE_NAME
PK,U1 NODE_ID
PK,U1 QUEUE_NAME
COND
SELECTOR
CHANNEL_ID
CLUSTERED
ALL_NODES
JBM_MSG_REF
PK,U1 MESSAGE_ID
PK,U1 CHANNEL_ID
U2 TRANSACTION_ID
U2 STATE
ORD
PAGE_ORD
DELIVERY_COUNT
SCHED_DELIVERY
JBM_MSG
PK MESSAGE_ID
RELIABLE
EXPIRATION
TIMESTAMP
PRIORITY
TYPE
HEADERS
PAYLOAD
JBM_TX
PK TRANSACTION_ID
NODE_ID
BRANCH_QUAL
FORMAT_ID
GLOBAL_TXID
JBM_ROLE
PK,U1 ROLE_ID
PK,U1 USER_ID
JBM_ID_CACHE
PK,U1 NODE_ID
PK,U1 CNTR
JBM_ID
JBM_USER
PK USER_ID
PASSWD
CLIENTID
JBM_DUAL
PK DUMMY
JBM_COUNTER
PK NAME
NEXT_ID
INSTANCE
PK PK
FK1,U13 SERIES_FK
FK2,U15 SRCODE_FK
FK3,U12 MEDIA_FK
U1 SOP_IUID
U14 SOP_CUID
U10 INST_NO
U3 CONTENT_DATETIME
U16 SR_COMPLETE
U17 SR_VERIFIED
U7 INST_CUSTOM1
U8 INST_CUSTOM2
U9 INST_CUSTOM3
U4 EXT_RETR_AET
RETRIEVE_AETS
AVAILABILITY
U11 INST_STATUS
ALL_ATTRS
U2 COMMITMENT
U6 CREATED_TIME
UPDATED_TIME
INST_ATTRS
U5 ARCHIVED
VERIFY_OBSERVER
PK PK
FK1,U3 INSTANCE_FK
U7 VERIFY_DATETIME
U5 OBSERVER_NAME
U4 OBSERVER_I_NAME
U6 OBSERVER_P_NAME
U1 OBSERVER_FN_SX
U2 OBSERVER_GN_SX
ACTIVE_PART
PK PK
USER_ID
ALT_USER_ID
REQUESTOR
NET_ACCESS_PT_ID
NET_ACCESS_PT_TYPE
USER_NAME
FK1 AUDIT_RECORD_FK
FK2 ROLE_ID_FK
MEDIA
PK PK
U1 FILESET_IUID
FILESET_ID
MEDIA_RQ_IUID
U2 MEDIA_STATUS
MEDIA_STATUS_INFO
MEDIA_USAGE
CREATED_TIME
UPDATED_TIME
AUDIT_RECORD
PK PK
SITE_ID
SOURCE_ID
EVENT_ACTION
EVENT_OUTCOME
U1 EVENT_DATE_TIME
U2 RECEIVE_DATE_TIME
IHEYR4
XMLDATA
SOURCE_TYPE
FK2 EVENT_ID_FK
FK1 EVENT_TYPE_FK
PART_OBJ
PK PK
OBJ_ID_TYPE_RFC
NAME
OBJ_ID
OBJ_TYPE
OBJ_ROLE
DATA_LIFE_CYCLE
OBJ_SENSITIVITY
FK1 AUDIT_RECORD_FK
FK2 OBJ_ID_TYPE_FK
UPS_REL_PS
PK PK
FK1,U3 UPS_FK
U2 SOP_IUID
U1 SOP_CUID
UPS_GLOB_SUBSCR
PK PK
U1 AET
DELETION_LOCK
UPS_REQ
PK,FK1 PK
U3 REQ_PROC_ID
U2 ACCESSION_NO
U1 CONFIDENTIALITY
U4 REQ_SERVICE
UPS_REPL_PS
PK,FK1 PK
U2 SOP_IUID
U1 SOP_CUID
REL_UPS_DEVCLASS
FK1 PK
FK2,U2 DEVCLASS_FK
REL_UPS_APPCODE
FK1 PK
FK2,U2 APPCODE_FK
UPS
PK PK
FK1,U8 PATIENT_FK
U5 CODE_FK
U1 UPS_IUID
U12 UPS_TUID
U2 ADM_ID
U3 ADM_ID_ISSUER_ID
U4 ADM_ID_ISSUER_UID
U7 UPS_LABEL
U14 UWL_LABEL
U10 UPS_START_TIME
U6 UPS_COMPL_TIME
U11 UPS_STATE
U9 UPS_PRIOR
CREATED_TIME
U13 UPDATED_TIME
UPS_ATTRS
REL_UPS_DEVNAME
FK1 PK
FK2,U2 DEVNAME_FK
REL_UPS_PERFORMER
FK1 PK
FK2,U2 PERFORMER_FK
UPS_SUBSCR
PK,FK1 PK
U2 AET
U1 DELETION_LOCK
REL_UPS_DEVLOC
FK1 PK
FK2,U2 DEVLOC_FK

Transitioning to Modern Data Management

- Mid 1990s: Content rapidly generated in the wake of the world-wide Web's impact
- Business Intelligence storage transitioned from paper to digital to reduce cost

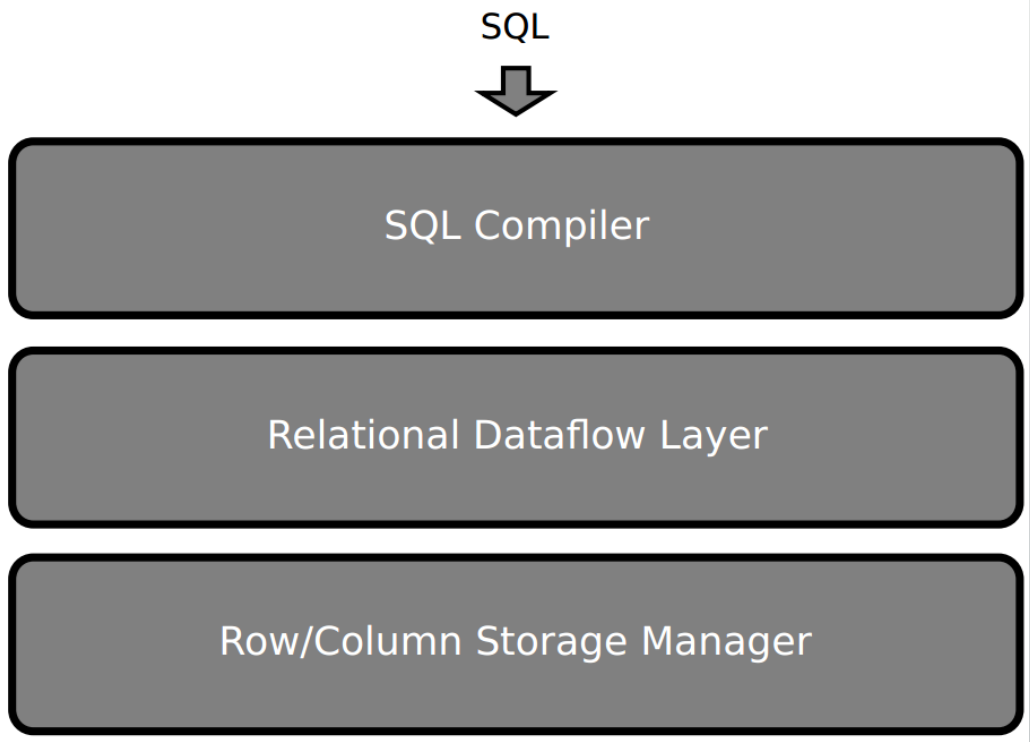
MapReduce Programming Models

- Google pioneered divide-and-conquer parallelism based on hashing contiguous bytes
- Apache Hadoop is open source, cheap, and scalable, and remains popular today
- Community developed high-level declarative languages for queries and data analysis (Pig, Jaql, Hive)



Parallel Database Systems

- Instead of hashed abstractions, the core file system is relational tables and is more object-oriented
- Users only able to access the outermost layer of SQL as most commercial options are closed-source

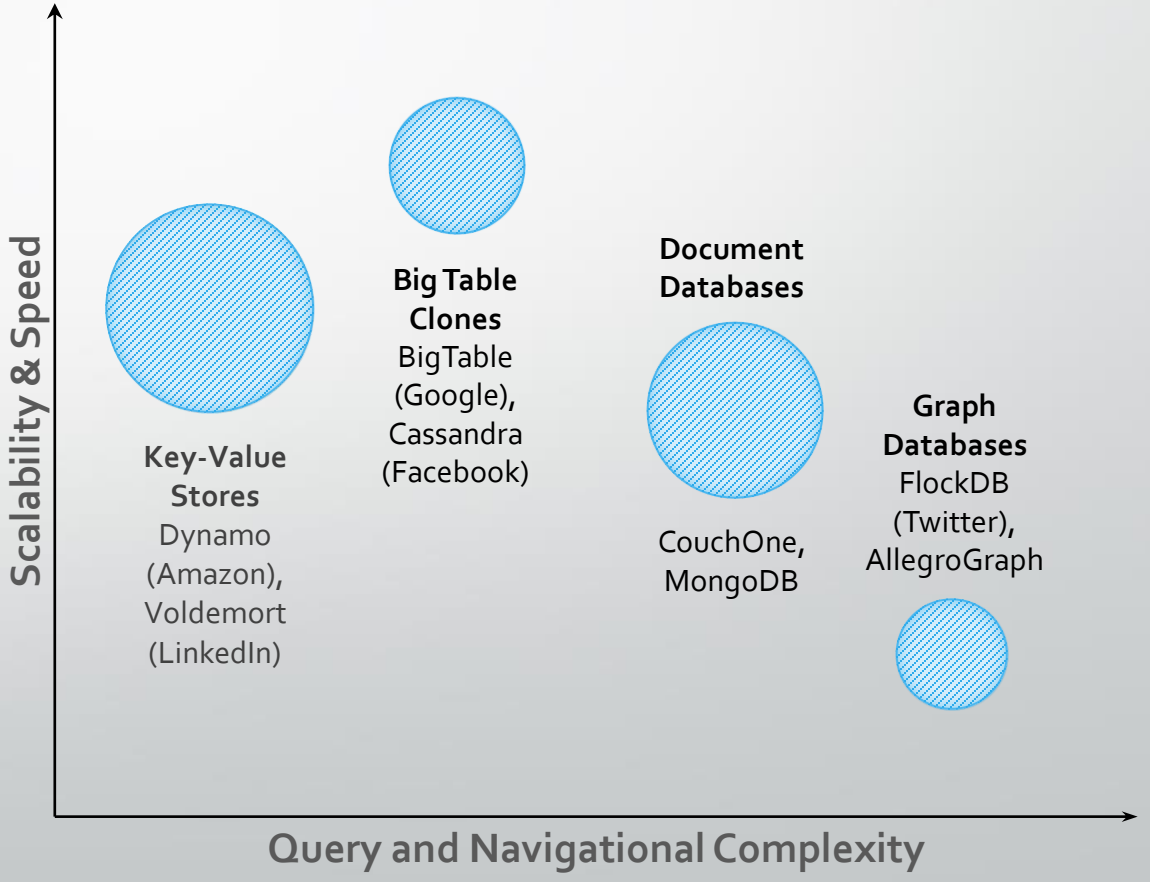


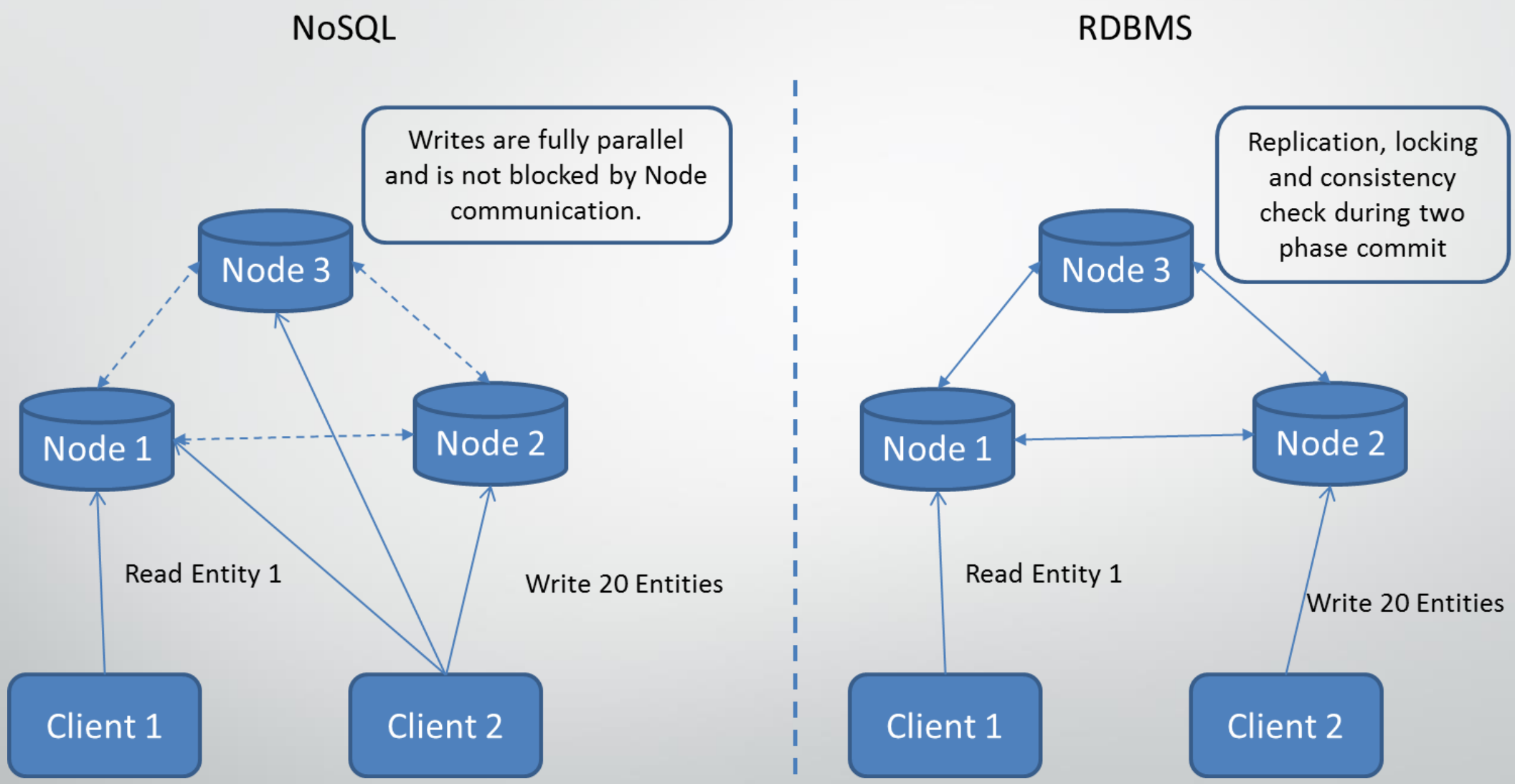
Problems in Big Data Management

- Low-latency
- High scalability grow / shrink elastically with business demand
- Support push, pull, transactional, and analytic based workloads
- Cost-effective
- Fault and failure tolerant

The NoSQL Movement

- **N**ot **O**nly **SQL** departs from the relational model
- Key-value, graph, or document data structures
- Open source, more versatile, scales better in some cases




NoSQL
Writes are fully parallel and is not blocked by Node communication.
Node 3
Node 1
Node 2
Read Entity 1
Write 20 Entities
Client 1
Client 2
RDBMS
Replication, locking and consistency check during two phase commit
Node 3
Node 1
Node 2
Read Entity 1
Write 20 Entities
Client 1
Client 2

Future of Big Data Management

Ultimately, depends on business preference, capabilities for implementation, and specific data management needs

- NoSQL
- Open Source
- High-level query languages

Thank you!

Questions?