Lecture 19: NoSQL I

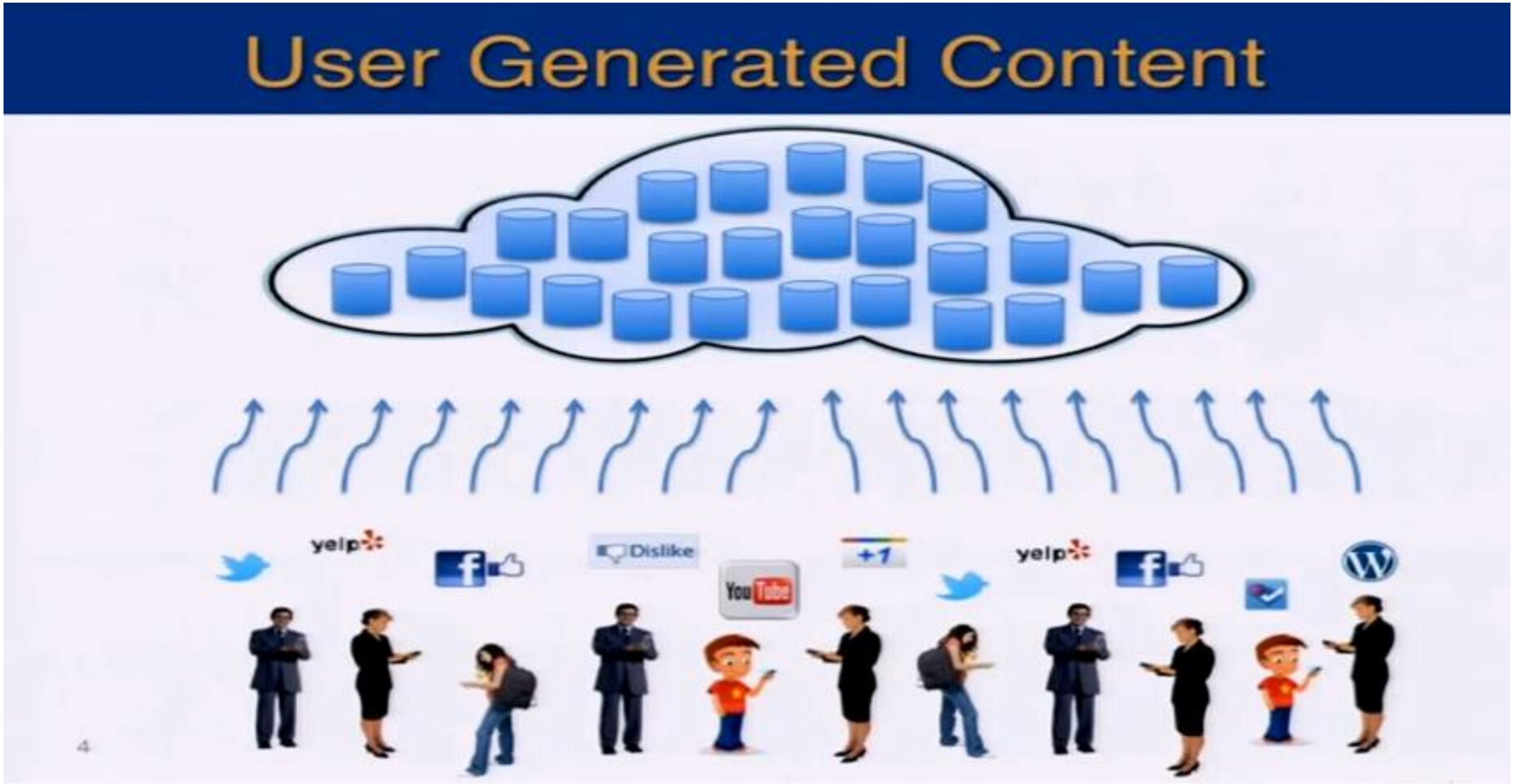
Wednesday, April 8, 2015

Where We Are

- Mostly done with class project (phase 2 is optional)
- Today: Big Data
- Next class: MapReduce & Pig
- Next Wed: Cloud platforms
- In 2 weeks: MongoDB & other Data Stores
- In 3 weeks: Prep for Final

Very important: Keep up with readings and tutorials:

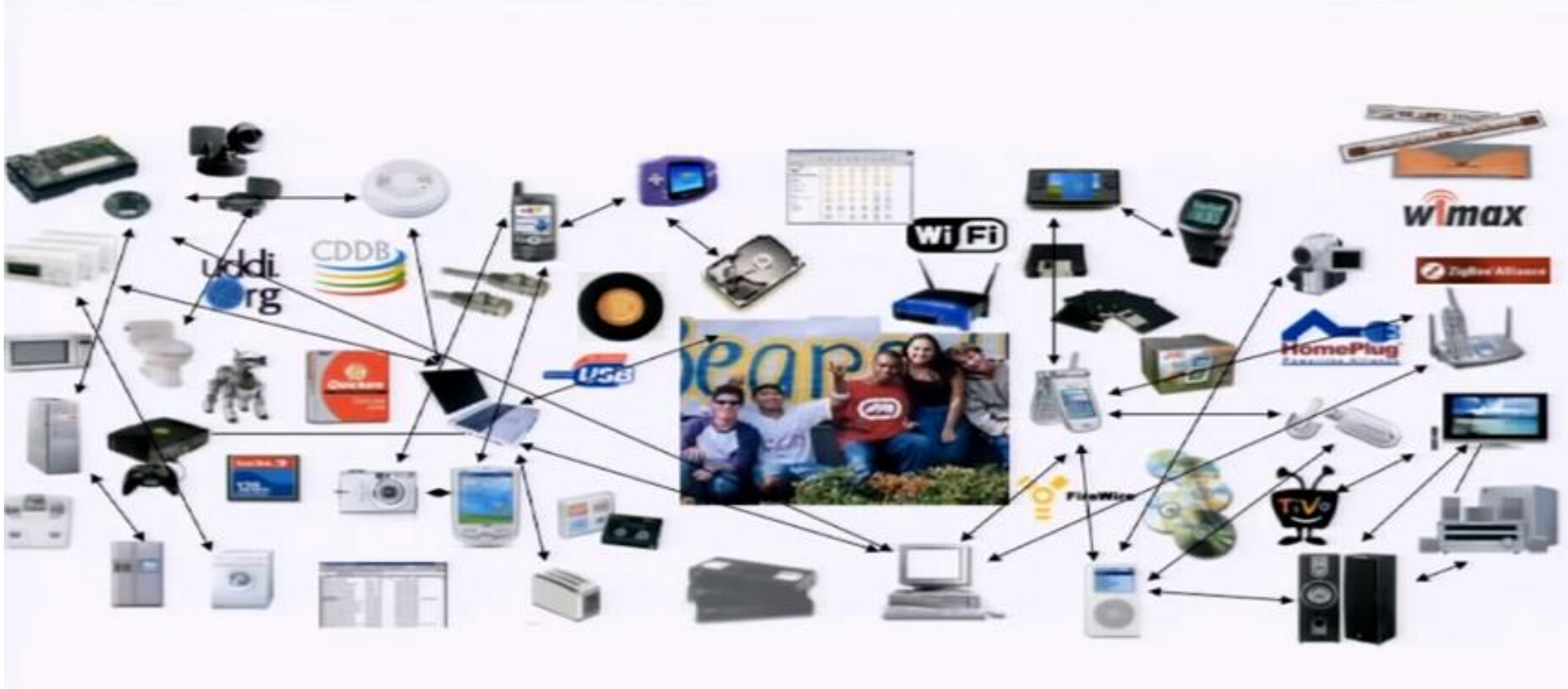
- Sadalage and Fowler, *NoSQL Distilled* (Addison-Wesley, 2013)
- MongoDB video tutorials (links on course web site)



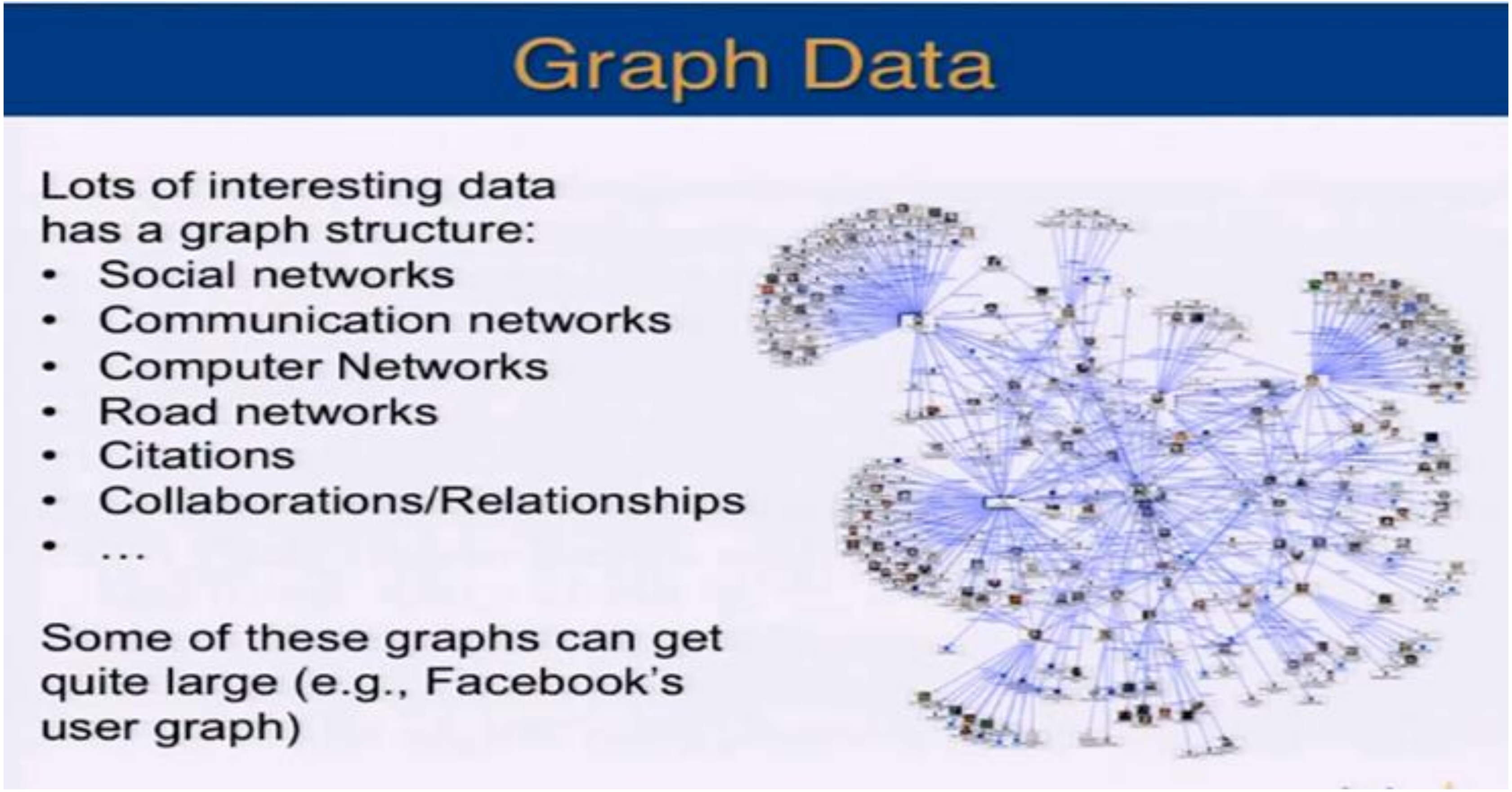
User Generated Content

Source: UC Berkeley AMP Lab

M2M - Internet of things



Source: UC Berkeley AMP Lab



Graph Data

Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

Some of these graphs can get quite large (e.g., Facebook's user graph)

Source: UC Berkeley AMP Lab

“Big Data”

- Just a buzzword?

- Gartner 2011 report*:
  - High volume
  - High variety
  - High velocity

Question: what do **you** think about “Big Data”?

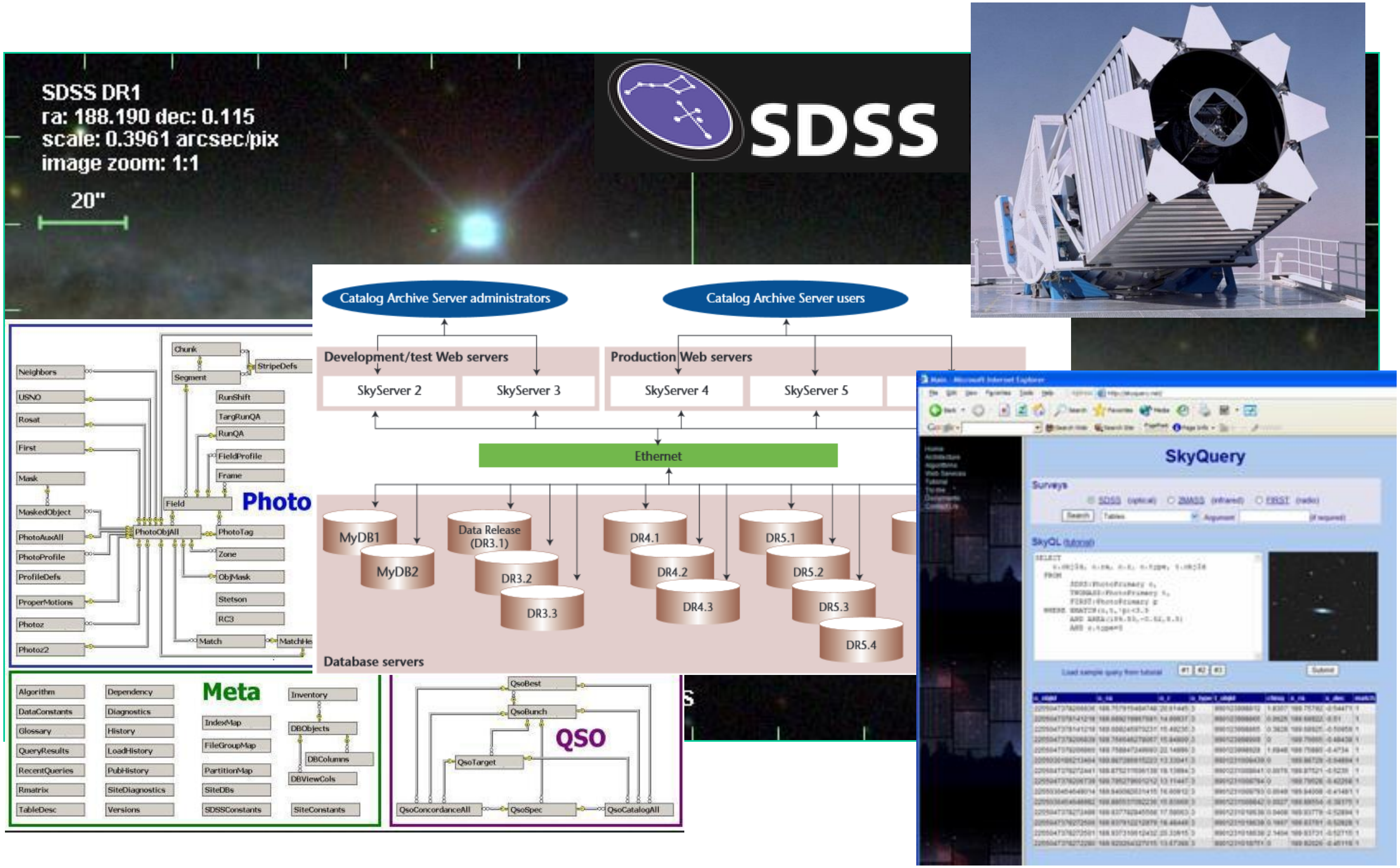
* http://www.gartner.com/newsroom/id/1731916

“Big Data” is really two problems

- The **analysis** problem:
  - How to extract useful info, using aggregate queries, machine learning and statistics

- The **storage** problem:
  - How to organize and partition huge amounts of data to support interactive queries

“Big Data” Meets RDBMS



```
SELECT
   o.objId, o.ra, o.r, o.type, t.objId
 FROM
      SDSS:PhotoPrimary o,
      TWOMASS:PhotoPrimary t,
      FIRST:PhotoPrimary p
 WHERE XMATCH(o,t,!p)<3.5
      AND AREA(185.33,-0.32,4.5)
      AND o.type=3
```

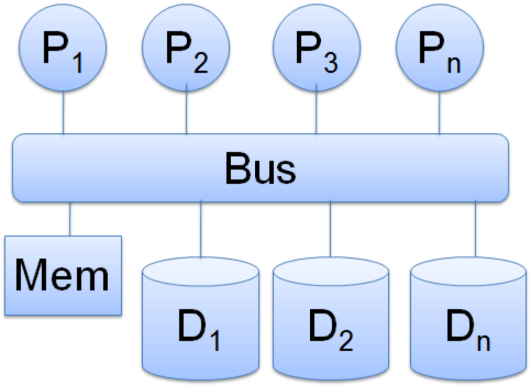
Source: Sloan Digital Sky Survey images obtained from http://skyserver.sdss.org

Classical DBMS (“Elephant” systems)

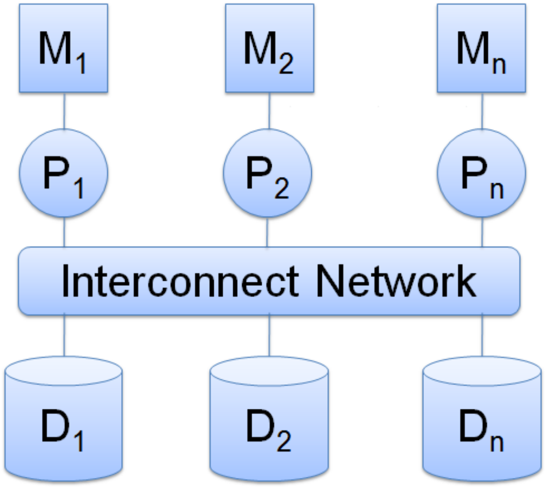
- Fixed schema (but alterations are possible)
- High-level query language (i.e. SQL)
- Limited analytics
- Structured & persistent data (e.g. inventory, banking, payroll, etc.)
- ACID properties
- Query optimization for consistent workloads
- Complex install & configurations
- Consumes time to load data
- Limited clustering and fault tolerance
- Primitive data partitioning technology
- Prohibitively expensive at web scale

Parallel Architectures

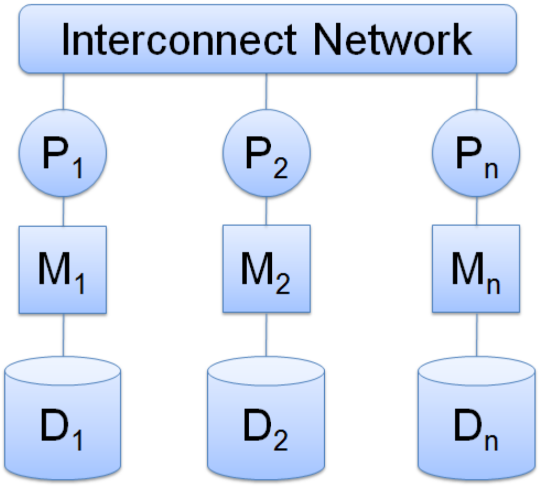
Shared Memory



Shared Disk



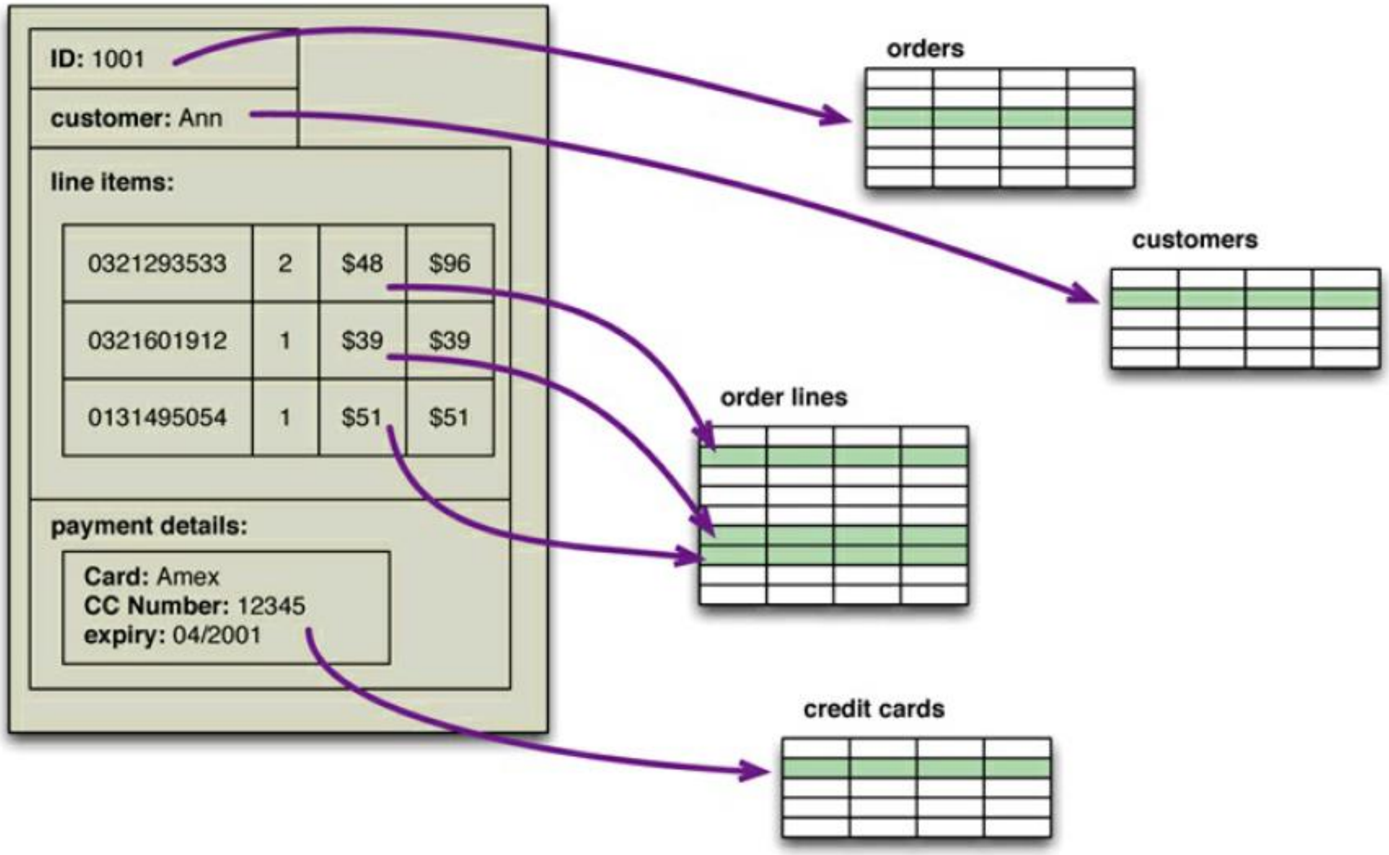
Shared Nothing



Performance metrics: speedup v.s. scaleup
Challenges: communication, resource contention,
data skew

Discussion of Readings

What is the “impedance mismatch” problem?

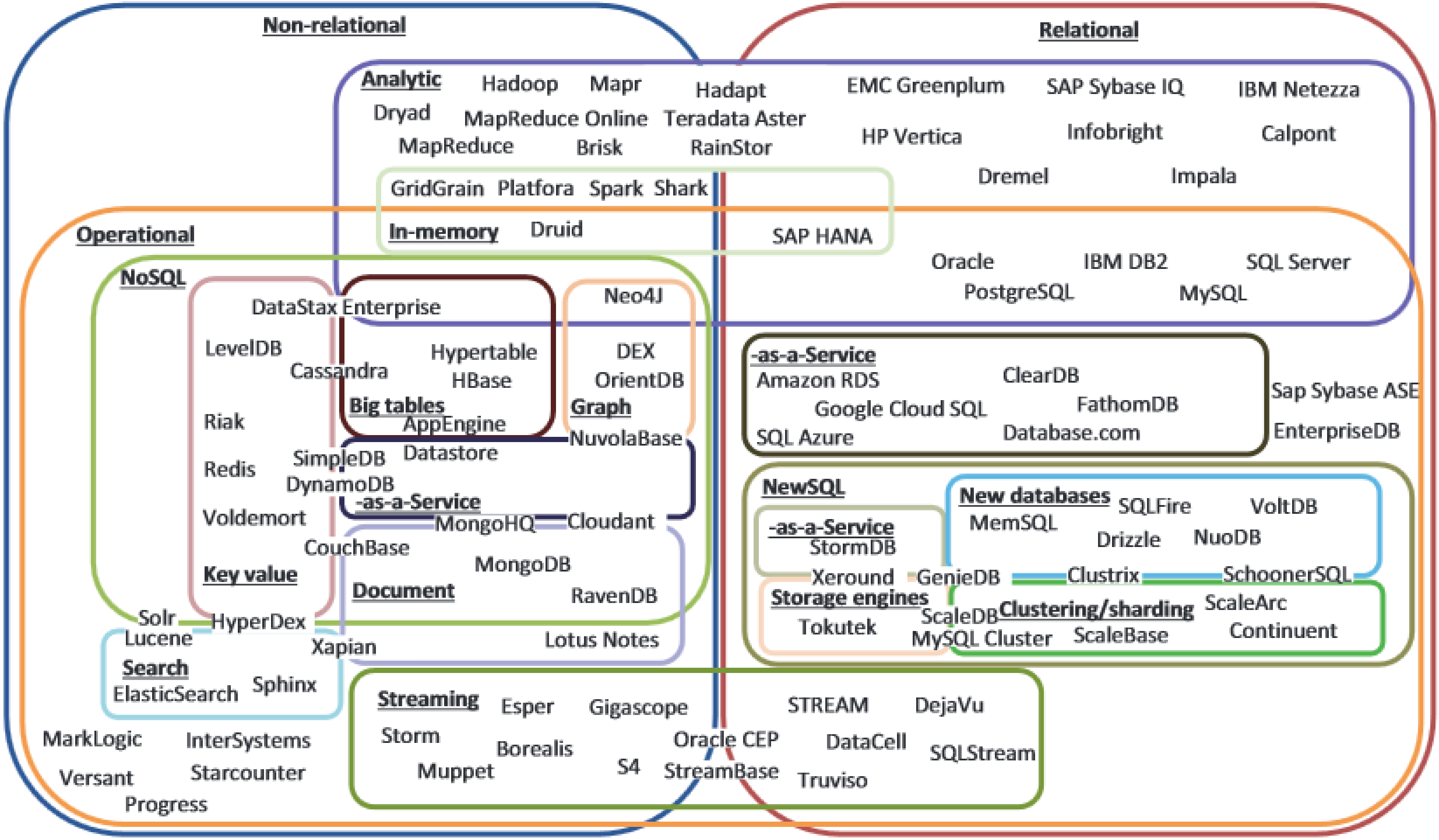


Source: Sadalage and Fowler, *NoSQL Distilled* (Addison-Wesley, 2013).

NoSQL Systems

- Name “NoSQL” = “Not SQL” or “Not Only SQL”
- Typical characteristics:
  - don't use relational model
  - “flexible” schema => implicit schema
  - unstructured and semi-structured data
  - simple APIs (no joins)
  - eventual consistency (=> immature consistency)
  - mostly open-source systems
  - easy to prototype and deploy
  - designed for use on clusters
  - support for data partitioning and replication
- Major forces driving NoSQL systems:
  - cloud platforms (will come back to this topic)
  - web 2.0 apps

“Data Systems” Landscape



Source: Lim et al, “How to Fit when No One Size Fits”, CIDR 2013.

DBMS Market Shares

- From 2011 Gartner report*:
  - Oracle: 48% market with $11.7BN in sales
  - IBM: 20% market with $4.8BN in sales
  - Microsoft: 17% market with $4.0BN in sales
  - Other vendors (i.e. NoSQL): 5.8% market with $1.3BN in sales

* http://www.gartner.com/newsroom/id/1731916

Discussion of Readings

- NoSQL taxonomy proposed by Sadalage and Fowler:
  - Analytics: MapReduce, Pig, Hive, Spark, Dremel
  - Key/Value: Redis, Memcached, Voldemort
  - Column: BigTable, DynamoDB, HBase, Cassandra
  - Document: CouchDB, MongoDB, SimpleDB
  - Graph: GraphDB, Neo4j

- “NewSQL” or Hybrid Systems:
  - Megastore, Spanner, F1, VoltDB, NuoDB

Optional References

The Unreasonable Effectiveness of Data [Alon Halevy et. al., IEEE Intelligent Systems 24(2): 8-12, 2009]

Challenges and Opportunities with Big Data – A community white paper developed by leading researchers across the United States. [D. Agrawal et al., http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf, Mar 2012]

The elephant in the room: getting value from Big Data [ACM Sigmod Blog. http://wp.sigmod.org/?p=1519, Feb 2015]

Next Class

- MapReduce and Pig
- HW 4