CC5212-1
PROCESAMIENTO MASIVO DE DATOS
OTOÑO 2015

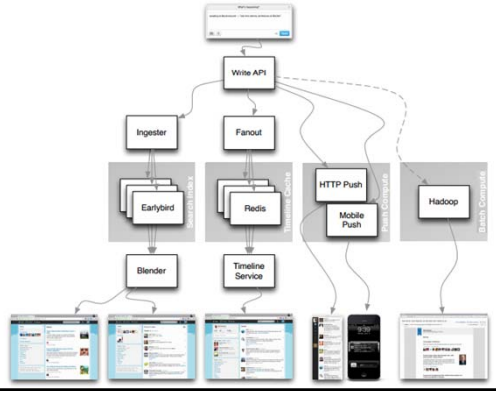
Lecture 11: Conclusion

Aidan Hogan
aidhog@gmail.com

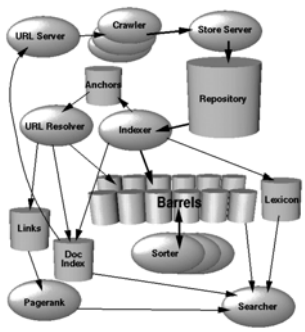
FULL-CIRCLE

The value of data …

Twitter architecture



Google architecture



Generalise concepts to …

Working with large datasets

Value/danger of distribution

Frameworks

- For Distrib. Processing
- For Distrib. Storage

The Big Data Buzz-word

Distributed Systems

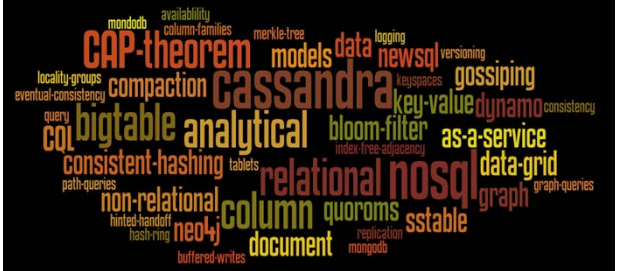
external sorts replication consistency
consensus protocols cap theorem
availability two phase commit
fault tolerance
distributed hash table partitions
client server synchronous
paxos java rmi
distributed systems
peer to peer asynchronous fallacies
three phase commit
transparency three tier architecture

GFS (HDFS) / MapReduce (Hadoop)

output map
shuffle transparency
pipelined-writes
hdfs replication gfs
rack-awareness
direct-reads
schema-on-read hadoop
sort reduce mapreduce
availability
consistency
rebalancing
fault-tolerance
partition
combiner

Information Retrieval

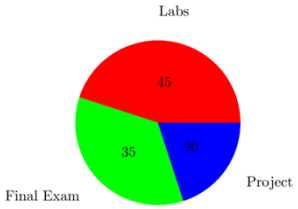
NoSQL and Querying



EXAM …

Course Marking

- 45% for Weekly Labs (~3% a lab!) [Easy]
- 20% for Small Class Project [Medium]
- 35% for Final Exam [Hard]



Final Exam (35%)

- Goal: test your understanding of *concepts*
  - Coding covered by labs/project
  - No syntax *writing* questions!
    - but there will be design and syntax reading questions
- Max. three hours
- Not marking you on English
  - If stuck, write in Spanish!
- Four questions (marked on best three) …

The following is not a legally abiding agreement.
It is just a helpful guide for what's important.

Q1: Distributed Systems

Q1: Distributed Systems (Slides)

Slides:

Lecture 2: Distributed Systems I

Lecture 3: Distributed Systems II

Names per the homepage:
http://aidanhogan.com/teaching/cc5212-1/

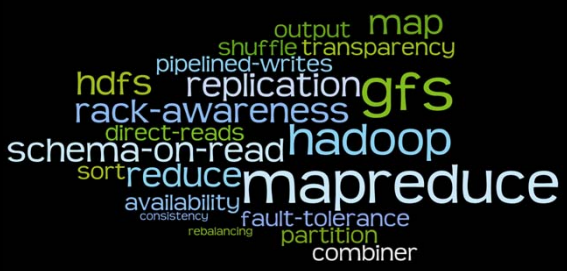
Q1: Distributed Systems (Topics)

Possible Topics:

- Advantages/disadvantages of a distributed system
- Five distributed system design goals
- Distributed architectures (P2P vs. C–S, Fat/Thin, n-Tier, etc.)
- Java RMI (high-level)
- Eight fallacies of distributed computing
- Consensus basics (fail-stop vs. Byzantine, synchronous vs. asynchronous, goals)
- Consensus protocols (2PC, 3PC, Paxos)

- CAP theorem may appear in Q4, but will not appear in Q1

GFS (HDFS) / MapReduce (Hadoop)



Q2: GFS (HDFS) / MapReduce (Hadoop)

Slides:

Lecture 4: DFS & MapReduce I

Lecture 6: DFS & MapReduce III

(Lecture 5 was whiteboard only, practicing MapReduce design)

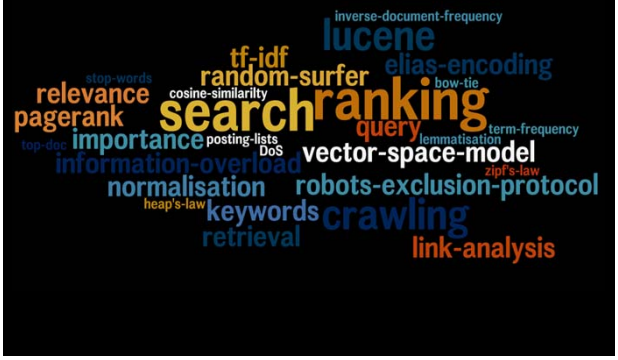
Names per the homepage:
http://aidanhogan.com/teaching/cc5212-1/

Q2: GFS (HDFS) / MapReduce (Hadoop)

Possible Topics:

- Google File System (reads, writes, fault-tolerance)
- MapReduce (incl. design question)
- HDFS/Hadoop (architecture)
- Pig (high-level, e.g., explain what a given script does)

Information Retrieval



Q3: Information Retrieval (Slides)

Slides:

Lecture 7: Information Retrieval I

Lecture 8: Information Retrieval II

Names per the homepage:
http://aidanhogan.com/teaching/cc5212-1/

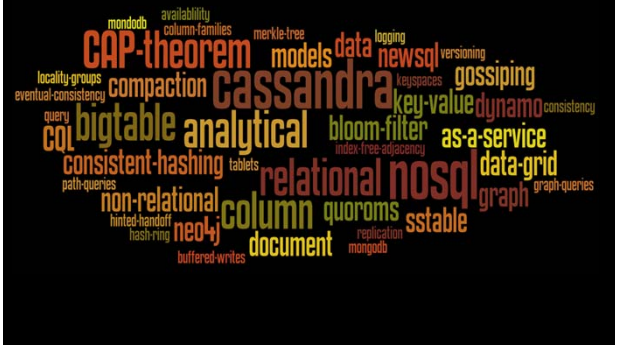
Q3: Information Retrieval (Topics)

Possible Topics:

- Crawling (high-level multi-threading, (D)DoS, robots.txt, sitemap, distribution, bow-tie)
- Inverted indexes (data structure, normalisation, Heap's law, Ziph's law, Elias encoding, etc.)
- Ranking (relevance vs. importance, TF-IDF, Vector Space Model, etc.)
- PageRank (concept, random surfer, calculation)

Bring a Calculator!

NoSQL and Querying



Q4: NoSQL and Querying (Slides)

Slides:

Lecture 9: NoSQL I

Lecture 10: NoSQL II

Lecture 3: Distributed Systems II (slides on CAP)

Names per the homepage:
http://aidanhogan.com/teaching/cc5212-1/

Q4: NoSQL and Querying (Topics)

Possible Topics:

- CAP theorem (<- note out of order)
- The Database Landscape
- Key–Value stores (data model, operations, distribution, consistent hashing, replication, Dynamo, Merkle trees)
- Document stores (high-level)
- Tabular/column-families (data model, Bigtable, sorting, tablets, column families, SSTables, writes, reads, compactions, hierarchy, bloom filters)
- Graph databases (high-level)
- Cassandra (high-level)

Final Exam (35%)

- Goal: test your understanding of *concepts*
  - Coding covered by labs/project
  - No syntax *writing* questions!
    - but there will be design and syntax reading questions
- Max. three hours
  - Not marking you on English
  - If stuck, write in Span(gl)ish!
- Four questions (marked on best three) …

SOME ADVERTISING

Poll … lab on Wednesday?

APACHE HBASE

vs.

Next semester … La Web de Datos

- Web of Data / Semantic Web / Linked Data

November … ayudante wanted

- "Diplomado" version of this course
- 8 x 3 hour lectures over 3 weeks
- Help with:
  - Translating slides
  - Attending class
  - Laboratories
  - Marking
- 1 UF per hour! ($25,000/hour)