Big Data Management and NoSQL Databases

Lecture 14. Exam Preparation

PD Dr. Andreas Behrend
Relevant Topics

• Lecture 1: Foundation
  • What Big Data: (Volume, Veracity, Velocity, Veracity)
  • Motivation NoSQL (distributed computation, fault tolerance, eventual consistency, schemaless data, ..)
  • scale-up vs scale-out
  • types of NoSQL stores (key-value, column family, document, graph)

• Lecture 2+3: Map Reduce
  • Map (k1,v1) -> list(k2,v2)
  • Reduce (k2,list(v2))->(k2, possibly smaller list(v2))
  • Keywords: chunks, shuffle, partition function, combine, counters,
  • Examples (sort, counting words, grep)
  • Multiple MR jobs (DAG)
  • Bloom Filter + Probability of False Positives
Relevant Topics

• Lecture 4: Application of MR
  • N-Grams (example)
  • MR approach: naive vs. a priori vs suffix based

• Lecture 5: Similarity of Documents
  • Jaccard Distance+Similarity
  • Shingling, Min-Hashing, Signatures
  • LSH

• Lecture 6: Foundations of NoSQL stores (1)
  • Fault tolerance,
  • pessimistic vs. optimistic replication
  • one-copy serializability, ROWA, PrimCopy
  • failures classes: byzantine vs Fail-Stop
  • consensus algorithm, Paxos (proposer, acceptor)
Relevant Topics

• Lecture 7/8: Foundations of NoSQL stores (2)
  • Ordering events - Lamport Timestamps
  • Last write wins, multiversion concurrency control
  • ACID vs. Base, CAP theorem
  • eventual consistency, quorums N,W,R
  • vector clocks

• Lecture 9: Foundations of NoSQL stores (3)
  • casual consistency
  • placement of data/nodes in network: consistent hashing
  • lookup table for logarithmic routing, gossip

• Lecture 10: Column-Family Stores
  • motivation
  • explanation of a provided sample instance
Relevant Topics

• Lecture 11: Document Stores
  • motivation (what is it?)
  • replica management (primary vs secondary, elections)

• Lecture 12: Graph Databases
  • data model
  • Cypher/Gremlin examples (only understand provided code snippets)

• Lecture 13: Data Streams
  • motivation
  • probabilistic counting
  • types of sliding windows (sliding/tumbling, range/row)
  • ISTREAM vs. DSTREAM vs. RSTREAM
  • Push vs pull
  • Explain provided CQL code samples