CSE 132C
Database System Implementation

Arun Kumar

Topic 8: Other “Big Data” Systems

Optional; NOT included in course syllabus or Final Exam
Other “Big Data” Systems

❖ Key-Value/NoSQL Systems
❖ Graph Processing Systems
❖ Machine Learning Systems
Key-Value/NoSQL Systems

- **Simple API**: `get` and `put` unique records very quickly!
  - Records usually uniquely identified by a “key”; information in record is the “value” (could be general JSON object)
- Used extensively by Web companies, e.g., get product record quickly and update stock count, update Facebook status, etc.
- Need high availability, high scalability, “eventual” consistency
- **Idea**: Discard ACID and 30+ years of DB lessons; use “BASE” (Basically Available, Soft state, and Eventually consistent)
- The new RDBMS-hating “movement” was christened “NoSQL”
Key-Value/NoSQL Systems

Also called transactional NoSQL (read-write)

Hadoop / Spark aka analytical NoSQL (read mostly)
Recent work on relaxed consistency models with guarantees in between full ACID and fuzzy best-effort BASE/Eventual

5 consistency levels of Microsoft Azure CosmosDB (a geodistributed cloudnative DBMS)

My take: Key area of research at the intersection of DB & distributed systems

Ad: Take CSE 124 or 223B to learn more
Other “Big Data” Systems

❖ Key-Value/NoSQL Systems
❖ Graph Processing Systems
❖ Machine Learning Systems
Graph Processing Systems

- Not a workload DB folks used to care much about but a hotter area of R&D in the last few years
- Specialized graph systems have been around for many years (Neo4j) but more popular now (Facebook, LinkedIn, etc.)
- **Data Model**: set of nodes, and set of (multi-)edges
- **Ops/queries**: nearest neighbors, shortest path, connectivity, density, cliques, etc.
Graph Processing Systems

Can be handled as an application on an RDBMS, but might be inefficient: transitive closure, repeated self-joins, etc.
Graph Processing Systems

My take:
Hot area of R&D in DB + algorithms + systems intersection

Ad:
Take HDSI’s DSC 104 or “Graph Analytics” course on UCSD Coursera to learn more about graph databases and systems

https://graphaware.com/graphaware/2019/02/01/graph-technology-landscape.html
Other “Big Data” Systems

- Key-Value/NoSQL Systems
- Graph Processing Systems
- Machine Learning Systems
Machine Learning Systems

- Systems for mathematically advanced data analysis and prediction computations, not (just) SQL aggregates:
  - Statistics, data mining, machine learning, deep learning

- Two Orthogonal Dimensions of Categorization:
  - Packages of Algorithms vs. Linear Algebra Systems
  - Layered on Existing Platforms vs. Customized Systems
Packages of Algorithms Layered on Existing Platforms:

**In-RDBMS**: use RDBMS’s UDFs/UDAs
- Apache MADlib, Oracle Data Mining, etc.

**On Dataflow Systems**: use their APIs
- Spark MLlib, Apache Mahout, etc.

*Key challenge*: Rewrite statistical and ML algorithms to use the extensibility abstractions of the data platforms

*My take*: Was a hot R&D topic in DB + ML intersection
Customized Systems/Frameworks:
- TensorFlow (Google): tailored for deep learning
- PyTorch (Facebook): also for deep learning
- XGBoost (UWash): popular tree-learning system

Cloud-Native and Other Packages of Algorithms:
- AWS SageMaker, Microsoft AzureML, etc.
- AutoML: DataRobot, H2O.AI, SalesForce Einstein

Each system has its own set of challenges and ideas

My take: Hot R&D topic in ML + DB + systems intersection
Machine Learning Systems

❖ **Linear Algebra Systems** (mostly, R-based or R-like):
   R is popular for statistical analysis on structured data

❖ **Layered on Existing Platforms:**
   - **In-RDBMS:** Oracle R Enterprise, SAP HANA R
   - **Others:** Apache SystemML on Spark, SparkR

❖ **Customized Platforms:**
   - ScaLAPACK, Microsoft Revolution R

*My take:* A lot of industrial R&D last decade; less active now
Machine Learning Systems

Summary: Scalable and efficient advanced data analytics using ML is crucial for unlocking the value of “Big Data”

If you are interested in learning more about this topic, read my book (the first on ML systems!) at:

PDF: https://www.morganclaypool.com/doi/10.2200/S00895ED1V01Y201901DTM057

(PDF is free on most university networks; use UCSD VPN)

Ad: Fall’21: CSE 234: Data Systems for Machine Learning
This Topic (Other “Big Data Systems) is NOT included for the final exam.

Thank you for taking CSE 132C!