CSE 190A
Database System Implementation

Arun Kumar

Topic 7.2: Advanced “Big Data” Systems

Not included for final exam

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”

"Big Data" Systems

- Parallel RDBMSs
- Beyond RDBMSs: A Brief History
- "Big Data" Systems
  - The MapReduce/Hadoop Craze
  - Spark and Other Dataflow Systems
  - **Key-Value NoSQL Systems**
  - Graph Processing Systems
  - Advanced Analytics/ML Systems

Key-Value NoSQL Systems

Also called **transactional NoSQL** (read-write)

Hadoop / Spark aka **analytical NoSQL** (read mostly)
**Key-Value NoSQL Systems**

- 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 geo-distributed cloud-native DBMS)

*My bias: Key area of research at the intersection of DB & distributed systems!*

**Advertisement:** Take CSE 223B to learn more!

---

**“Big Data” Systems**

- Parallel RDBMSs
- Beyond RDBMSs: A Brief History
- “Big Data” Systems
  - The MapReduce/Hadoop Craze
  - Spark and Other Dataflow Systems
  - Key-Value NoSQL Systems
  - Graph Processing Systems
  - Advanced Analytics/ML Systems

---

**Graph Processing Systems**

- Not a workload DB folks used to care much about
- Specialized graph systems have been around for 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.*
Advanced Analytics/ML Systems

- Systems for mathematically advanced data analysis ops (not just SQL aggregates):
  - Statistics, machine learning, data mining, …
- Two Orthogonal Dimensions of Categorization:
  - Packages of Algorithms vs. Linear Algebra Systems
  - Layered on Existing Platforms vs. Customized Systems

Key challenge: Rewrite statistical and ML algorithms to use the extensibility abstractions of these systems

My bias: Key area of research in DB & ML intersection

Parallel RDBMSs
- Beyond RDBMSs: A Brief History
- "Big Data" Systems
  - The MapReduce/Hadoop Craze
  - Spark and Other Dataflow Systems
  - Key-Value NoSQL Systems
  - Graph Processing Systems
  - Advanced Analytics/ML Systems

Key challenge: Rewrite statistical and ML algorithms to use the extensibility abstractions of these systems

My bias: Key area of research in DB & ML intersection
Advanced Analytics/ML Systems

- **Customized Systems/Frameworks:**
  - TensorFlow (Google): esp. good for "deep learning"
  - GraphLab (UWash): graph-parallel analytics; uses MPI
  - DeepDive (Stanford): for statistical relational learning
    (uses an RDBMS for some parts)

  *Each system has its own set of challenges and ideas*

  *My bias: Key area of research in DB & ML & systems intersection*

Advanced Analytics/ML Systems

- **Linear Algebra Systems** (mostly, R-based or R-like):
  - R is incredibly popular for statistical analysis

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

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

  *My bias: Is (was?) a focus area for industrial R&D*

Advanced Analytics/ML Systems

- **Summary:** Scalable and efficient advanced analytics/ML is crucial for unlocking the value of “Big Data”

  If you are interested in learning more about this topic, read my survey paper at:


  *Advertisement: This topic is also the focus of my research*
  - Fall’18: CSE 190 “Seminar on Adv. Data Science Applications”

Advanced Analytics/ML Systems

- **NoSQL, Graphs, and Adv. Analytics/ML systems not included for the final exam!**

  Thank you for taking CSE 190A!