CSE 190D
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”
Key-Value NoSQL Systems

- Recent work on relaxed consistency models with guarantees in between full ACID and fuzzy best-effort BASE/Eventual Consistency

5 consistency levels of Microsoft Azure CosmosDB (a geodistributed cloud-native DBMS)

My bias: One of the most important and challenging open research problems in DB/distributed systems intersection!

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 Systems

Graph Processing Systems

- Not a workload DB folks used to really care 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.

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

“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 Systems**

Advanced Analytics 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

Advanced Analytics Systems

- **Packages of Algorithms Layered on Existing Platforms:**
  - **In-RDBMS:** use RDBMS’s UDFs/UDAs
    - Oracle DM, MADlib, Bismarck (Wisconsin), etc.
  - **On-Hadoop/Spark/etc.:** use their APIs
    - Apache Mahout, Spark MLib, AzureML, etc.

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

  **My bias:** Hot area of research in DB / ML intersection
Advanced Analytics Systems

- **Customized Systems/Frameworks:**
  - TensorFlow (Google): especially good for deep learning
  - GraphLab (Wash): graph-parallel analytics/ML; 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: Hot area of research in DB / ML / systems intersection*

Advanced Analytics Systems

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

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

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

  *My bias: Hot area of research and industrial R&D*

Advanced Analytics Systems

**Summary**: Scalable and efficient advanced data analytics / data science is key to unlocking the value of “Big Data”

If you are interested in learning more about advanced analytics systems, read my survey paper at:

http://cseweb.ucsd.edu/~arunkk/vision/

Click “Associated Survey of ML Systems” under Downloads

*Advertisement: Advanced analytics is the focus of my research*

Fall’17: CSE 190 Seminar on Advanced Data Science
Winter’18: CSE 291 Topics in Advanced Analytics and ML Systems

NoSQL, Graphs, and Advanced Analytics are **not included for the final exam!**

Thank you for taking CSE 190D!