Database Management Systems - I, CS 157A

Ahmed K. Ezzat,

Scalability Models and Big Data Issues

Outline

- Scalability Models:
  - Strong vs. Eventually Consistent Models
- Big Data Issues:
  - Introduction
  - Top Ten Challenges
  - Security, Compliance, Auditing, and Protection
    - Steps to securing Big Data
    - Classifying Data
    - Protecting Big Data Analytics
    - Big Data and Compliance
    - The Intellectual Property Challenge
- Summary

Scalability Models

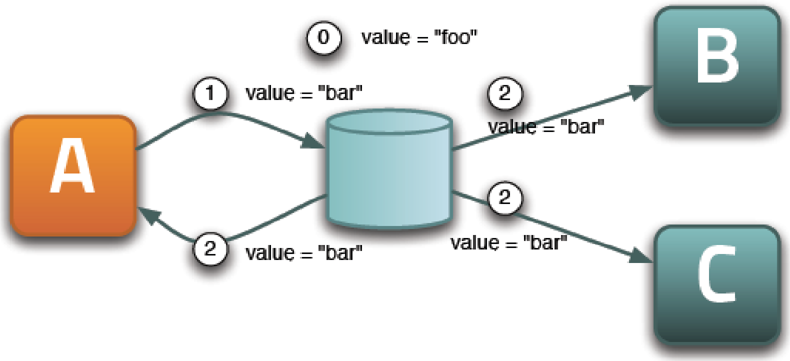
1. Scalability Models

- Overview
  - RDBMS supports ACID properties with Strong Consistency
  - The CAP theorem (**C**onsistency, **A**vailability, tolerance to network **P**artitioning) presented by Eric Brewer states that one can achieve only 2 out of these 3 attributes. Given for Big Data large scalable systems, network partitioning is a given, either consistency or availability have to be relaxed; hence the eventual consistency model.
  - Consistency Levels depends on the application needs:
    - Strong Consistency: after an update completes and subsequent access from any copy should be the same.
    - Weak Consistency: no guarantee that after an update an access will be the same during the inconsistency window.
    - Eventual Consistency: after an update and assuming no more updates, eventually all copies will be the same.

1. Scalability Models

- Overview
  - □ Strong Consistency

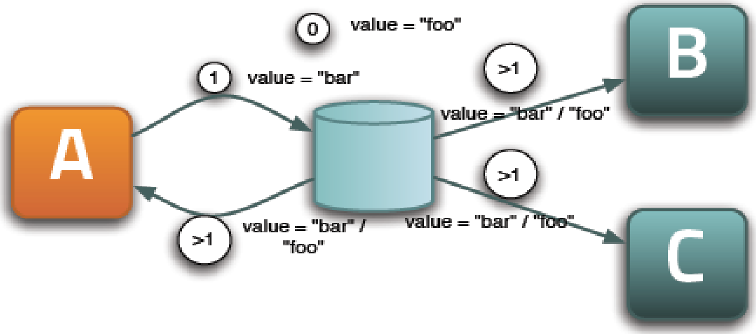
Strong Consistency



After the update, any subsequent access will return the updated value.

  - □ Weak Consistency

Weak Consistency

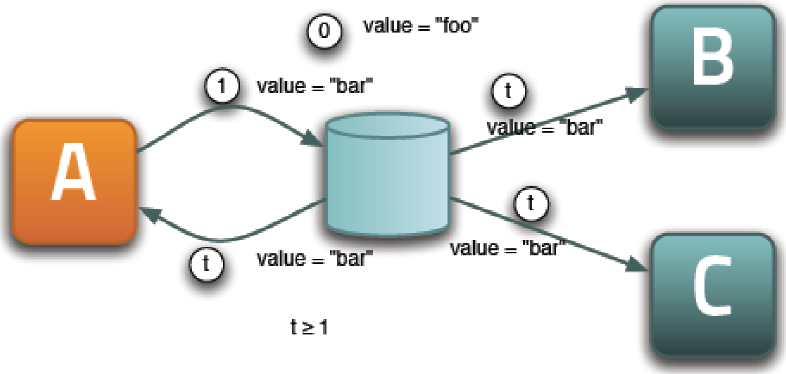


The system does not guarantee that at any given point in the future subsequent access will return the updated value

1. Scalability Models

- Overview
  - Eventual Consistency

Eventual Consistency



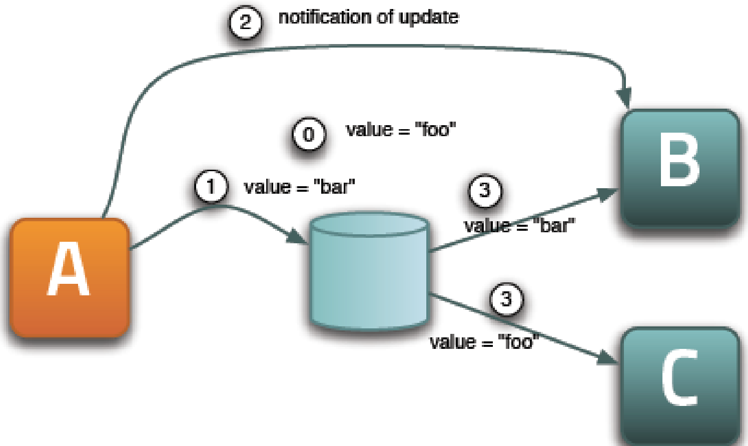
If no updates are made to the object, eventually all accesses will return the last updated value.

  - Inconsistent Window
    - Can be computed based on metrics like communication delay and workload on the system
    - If reading from asynchronous replica, inconsistency window = length of log shipment.

1. Scalability Models

- Eventual Consistency Flavors:
  - □ Causal consistency: If process A communicated to process B that it has updated an item, subsequent access by process B will return the updated value, and write is guaranteed to supersede the earlier write. Access by process C that has no causal relationship with A is subject to normal eventual consistency rules

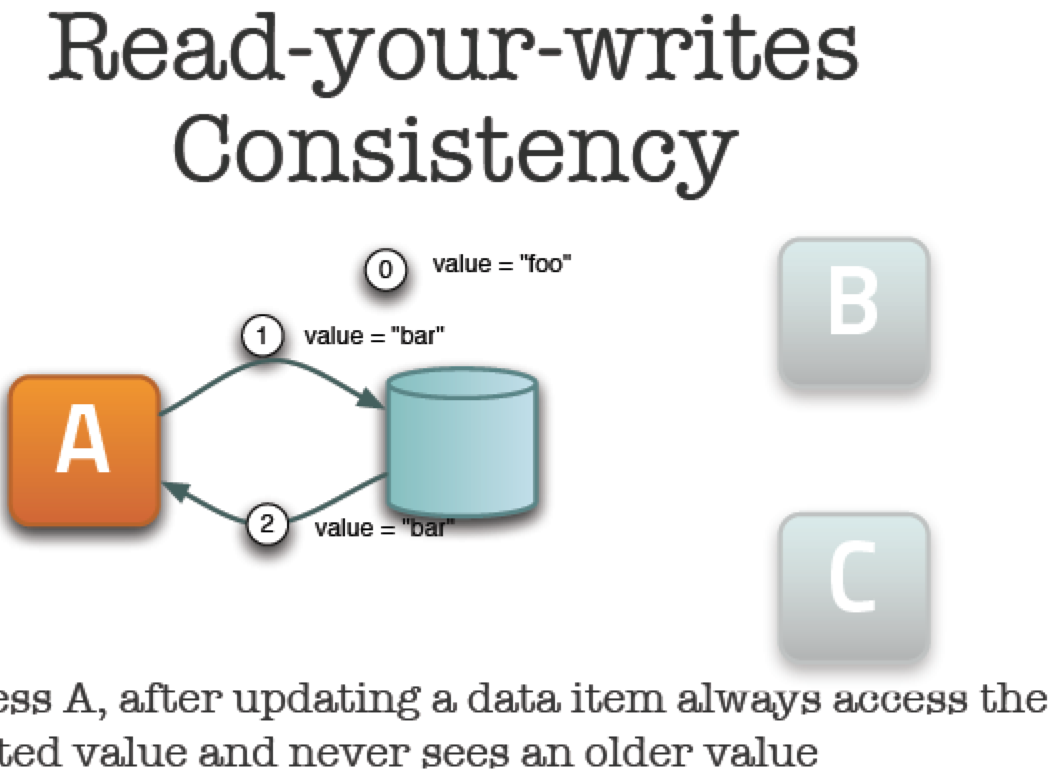
Causal Consistency



Subsequent access by process B will return the updated value, and a write is guaranteed to supersede the earlier write.

1. Scalability Models

- Eventual Consistency Flavors:
  - Read-your-writes consistency: Process A after updating an item always access the updated value; special case of causal model.



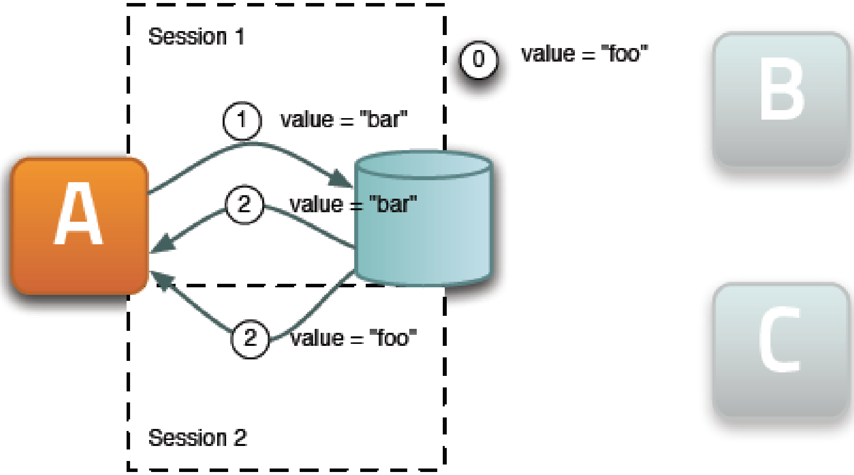
Read-your-writes Consistency

Process A, after updating a data item always access the updated value and never sees an older value

1. Scalability Models

- Eventual Consistency Flavors:
  - □ Session consistency: Process access with a session context. Within a session the system guarantee R/W consistency, and no guarantee with overlapping sessions.

Session Consistency

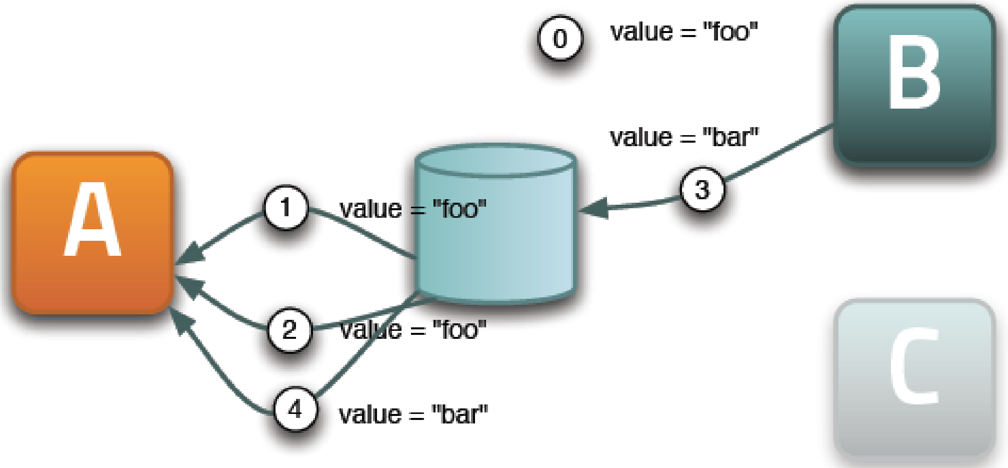


Within the "session", the system guarantees read-your-writes consistency

1. Scalability Models

- Eventual Consistency Flavors:
  - □ Monotonic read consistency: Once a process read a value of an object it will never access previous values.

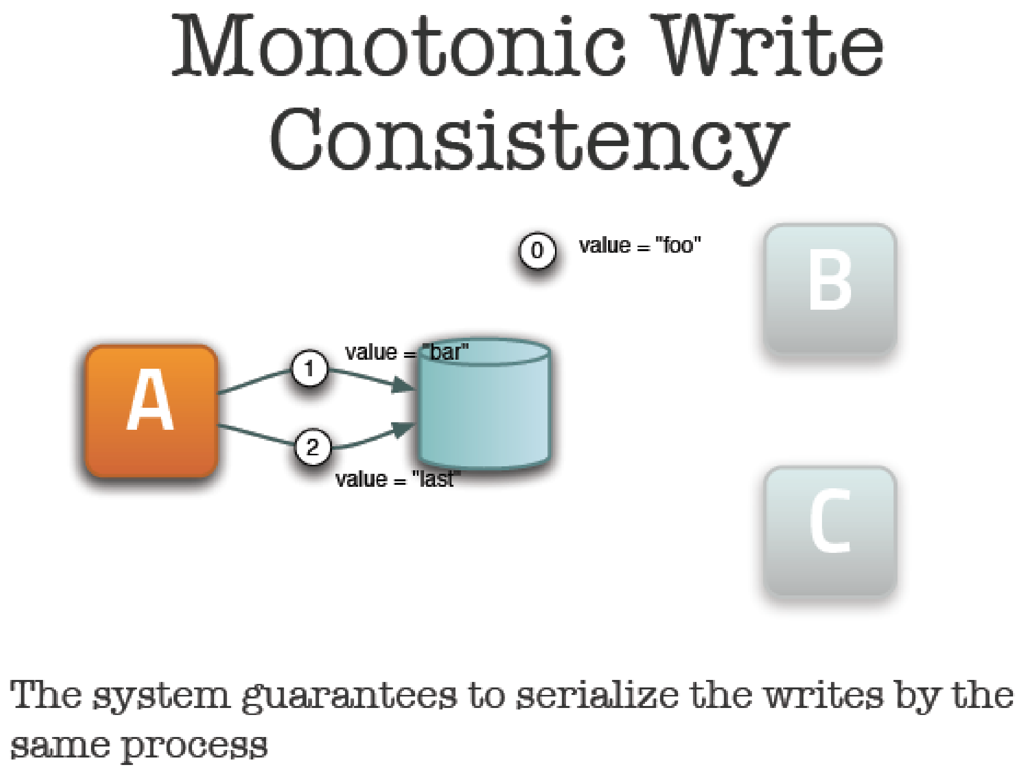
Monotonic Read Consistency



If a process has seen a particular value for the object, any subsequent access will never return any previous values

1. Scalability Models

- Eventual Consistency Flavors:
  - □ Monotonic write consistency: System guarantee serializing writes by the same process.



Monotonic Write Consistency

The system guarantees to serialize the writes by the same process

Big Data Issues

2. Big Data Issues:

Introduction

- Security and privacy issues are magnified by velocity, volume, and variety of Big data, such as large scale cloud infrastructure, diversity of data sources and formats, streaming nature of data acquisition, and high volume inter-cloud migration.
- Traditional security mechanisms are in-adequate
- Streaming data demands ultra-fast response time from any security and privacy solution.
- Below is highlights of the top ten big data security and privacy challenges:

2. Big Data Issues:

Introduction

- Secure computations in distributed programming framework
- Security best practices for non-relational data stores
- Secure data storage and transaction logs
- End-point input validation/filtering
- Real-time security/compliance monitoring
- Scalable and privacy-preserving data mining & analytics
- Cryptographically enforced access control and secure communication
- Granular access control
- Granular audit
- Data provenance

2. Big Data Issues:

Top Ten Challenges

- (1) Secure computations in distributed programming framework: typically we have parallel computation and storage access.
- One model is to use M/R framework, which splits an input file into chunks. Mapper would read a chunk and performs some computation and output K/V pairs. Reducer combines values belong to each distinct key and output the result.
- There are 2 potential issues with the above model: securing the mappers (can return wrong results to reducers) and securing the data in the presence of untrusted mapper!

2. Big Data Issues:

Top Ten Challenges

- (2) Security best practices for non-relational data stores: these stores are still evolving w.r.t. security. For example, secure NoSQL injection is still not mature.
- Each NoSQL DB was built to tackle different challenge posed by analytics and hence security was not part of the model in the design phase
- Typically NoSQL developers embed security in the middleware rather then being embedded in the DB itself
- Clustering in NoSQL poses additional challenge to the robustness of such security practices

2. Big Data Issues:

Top Ten Challenges

- (3) Secure data storage and transaction logs: data and transactional logs are stored in multi-tiered storage media. Manually moving data gives the IT manager direct control over exactly what data is moved and when.

- When data is very large (Big Data) scalability necessitated auto-tiering for big data storage management. Auto tiering poses new challenges to secure data storage and new mechanisms are needed.

2. Big Data Issues:

Top Ten Challenges

- (4) End Point Input validation/Filtering: many Big Data use cases in an enterprise setting require data/event collection from many sources, such as end-point devices.

- A key challenge in the data collection is data validation: how can we trust the data? How we can validate that the source is not malicious and how do we filter malicious input from our collection? Input validation and filtering is a daunting challenge posed by untrusted sources

2. Big Data Issues:
Top Ten Challenges

- (5) Real-time security/compliance monitoring: real-time security monitoring is always a challenge (i.e., given the # of alerts that may lead to false positives). This problem increases with big data given the volume and velocity of data streams.
- However, big data technology might also provide an opportunity with fast processing and analytics which in turn can be used to provide real-time anomaly detection based on scalable security analytics.
- Examples include: who is accessing which data from which resource at what time; do we have a breach of compliance standard C because of action A?

2. Big Data Issues:

Top Ten Challenges

- (6) Scalable and Composable Privacy-preserving Data Mining and Analytics: this is manifested in enabling invasion of privacy, invasive marketing, decreased civil freedom, and increase state and corporate control!
- User data collected by companies and government agencies are constantly mined and analyzed by inside analysts and possibly outside contractors. A malicious insider or untrusted partner can abuse these datasets and extract private information about customers.
- It is important to establish guidelines and recommendations for preventing inadvertent privacy disclosures.

2. Big Data Issues:

Top Ten Challenges

- (7) Cryptographically enforce access control and secure communication: to ensure that the sensitive private data is end-to-end secure and is only accessible by authorized entities, data has to be encrypted based on access control policies.

- To ensure authentication and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented

2. Big Data Issues:

Top Ten Challenges

- (8) Granular access control: the problem with course-grained access mechanism is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security.

- Granular access control gives data managers ability to share data as much as possible without compromising security.

2. Big Data Issues:
Top Ten Challenges

- (9) Granular audits: with real-time security monitoring, we try to be notified at the moment an attack takes place. In reality, this will not always be the case.

- To get to the bottom of any attack, we need audit information. This is not only helpful to understand what happened but also is important from compliance and regulations point of view.

- Auditing is not new, but the scope and granularity might be different, e.g., we might need to deal with a large number of distributed objects!

2. Big Data Issues:

Top Ten Challenges

- (10) Data provenance: provenance metadata will grow in complexity due to large provenance graphs generated from provenance-enabled programming environments in big data applications.

- Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- The sheer size of Big Data brings with it a major security challenge. Proper security entails more than keeping the bad guys out; it also means backing up data and protecting data from corruption.
- Data access: data can be protected if you eliminate access to the data! Not pragmatic so we opt to control access.
- Data availability: controlling where the data are stored and how it is distributed; more control position you better to protect the data.
- Performance: encryption and other measures can improve security but they carry a processing burden that can severely affect the system performance!

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Liability: accessible data carry with them liability, such as the sensitivity of the data. The legal requirements connected to the data privacy issues, and IP concerns.
- Adequate security becomes a strategic balancing act among the above concerns. With planning, logic, and observations, security becomes manageable. Effectively protecting data while allowing access to the authorized users and systems.

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Pragmatic Steps to Securing Big Data:
  - First get rid of data that are no longer needed. If not possible to destroy then the information should be securely archived and kept offline
  - A real challenge is to decide which data is needed? As value can be found in unexpected places. For example, activity logs represent a risk but logs can be used to determine scale, use, and efficiency of big data analytics
  - There is no easy answer to the above question, and it becomes a case of choosing the lesser of two evils.

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Classifying Data:
  - Protecting data is much easier if data is classified into categories, e.g., internal email between colleagues is different from financial report, etc.
  - Simple classification can be: financial, HR, sales, inventory, and communications.
  - Once organizations better understand their data, they can take important steps to segregate the information and that makes it easier to employ security measures like encryption and monitoring more manageable

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Protecting Big Data Analytics:
  - A real concern with Big Data is the fact that Big Data contains all of the things you don’t want to see when are trying to protect data, very unique sample set, etc.
  - Such uniqueness also means that you can’t leverage time-saving backup and security technologies such as deduplication.
  - Significant issue is the large size and number of files involved in Big Data Analytics environment. Backup bandwidth and/or the backup appliance must be large and the receiving devices must be able to ingest data at the delivery rate of data.

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Big Data and Compliance:
  - Compliance has major effect on how Big Data is protected, stored, accessed, and archived.
  - Big Data is not easily handled by RDBMS; this means it is harder to understand how compliance affects the data.
  - Big Data is transforming the storage and access paradigm to a new world of horizontally scaling, unstructured databases, which are more suited to solve old business problems with analytics.
  - New data types and methodologies are still expected to meet the legislative requirements expected by compliance laws

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Big Data and Compliance:
  - Preventing compliance from becoming the next Big Data nightmare is going to be the job of security professionals.
  - Health care is a good example of Big Data compliance challenge, i.e., different data types and vast rate of data from different devices, etc.
  - NoSQL is evolving as the new data management approach to unstructured data. No need for federating multiple RDBMS. Clustered single NoSQL database and being deployed in the cloud.

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Big Data and Compliance:
  - Unfortunately, most data stores in the NoSQL world (i.e., Hadoop, Cassandra and MongoDB) do not incorporate sufficient data security tools to provide what is needed.
  - Big Data changed few things: For example network security developers spent a great deal of time and money on perimeter-based security mechanisms (e.g., firewalls) but that cannot prevent unauthorized access to data once a criminal/hacker has entered the network!
  - Lessons learned:
    - Control access by process, not job function

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Secure the data at the data store level
- Protect the cryptographic keys and store them separately from the data
- Create trusted applications and stacks to protect data from rogue users

- Once you begin to map and understand the data, opportunities will be evident that will lead to automating and monitoring compliance and security compliance.
- Of course automation does not solve every problem; there are still basic rules to be used to enable security while not derailing the value of Big Data:

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Ensure that security does not impede performance or availability
- Pick the right encryption scheme, i.e., file, document, column, etc.
- Ensure that the security solution can evolve with your changing requirements

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- The Intellectual Property (IP) Challenge:
  - One of the biggest issues with Big Data is the concept of IP.
  - IP refers to creations of the human mind, such as inventions, literary and artistic works, and symbols, names, images used in commerce.
  - Some basic rules are:
    - Understand what IP is and know what you have to protect
    - Prioritize protection
    - Label (confidential information should be labeled

2. Big Data Issues:

Security, Compliance, Auditing and Protection

- Educate employees
- Know your tools: tools that can be used to track IP stores
- Use a holistic approach: includes internal risks as well as external ones.
- Use a counterintelligence mind-set: think as if you are spying on your company and ask how would you do it?

- The above guidelines can be applied to almost any information security paradigm that is geared toward protecting IP.

4. Summary

- CAP is fundamental in distributed scalable systems design.
- Network partitioning is a given with distributed scalable systems and hence compromising on consistency
- Integrating Big Data applications and analysis into an existing data security infrastructure rather than relying on homegrown scripts and monitors.
- We covered the top ten challenges in Big Data.
- We covered the Big Data issue related to Security, Compliance, Auditing, and Protection.
- We concluded with “Best Practices” when dealing with Big Data.

END