BiG Data Security

CISC 6640 PRIVACY AND SECURITY IN BIG DATA

Instructor:
Md Zakirul Alam Bhuiyan
Assistant Professor
Department of Computer and Information Sciences
Fordham University

We Have Learned …

- Database Security
- Relational Databases
  - Database security models
- No SQL Databases
- Object Based vs. Object Oriented
- Overview of Database Vulnerabilities
  - Common DBMS vulnerabilities
- Overview of Database topics/issues (indexing, inference, aggregation, polyinstantiation)
  - Security issues of inference and aggregation
- Hashing and Encryption
- Database access controls (DAC, MAC, RBAC, Clark-Wilson)
- Information flow between databases/servers & applications

What We Are Going to Learn…

- Data Administration
  - Big Data Security
  - Data Policies, Data Quality
  - Data Ownership
  - Data Warehousing
  - Long Term Archival
  - Data Validation
  - Data Security (access control, encryption)

Implications of Big Data Key Aspects

- Security impact of big data dephasing 5Vs
  - Volume, Velocity, Variety, Veracity, Value
- The key aspects increase challenge for security
  - Science and foundations
  - Infrastructure
  - Management
  - Searching and mining
  - Applications.

Big Data Security

- Data security through Hadoop
  - Challenges centers around the fragmented data security issues, though new tools and technologies are surfacing.
    - The Kerberos (a network authentication protocol) is a great step toward making Hadoop environments secure.
      - It is still immature
- Hadoop File Permissions
  - Added in HADOOP-1298
    - Hadoop 0.16, Early 2008
  - Authorization without authentication

Big Data Security

- MapReduce ACLs
  - **Added in HADOOP-3698**
    - Hadoop 0.19, Late 2008
  - **ACLs per job queue**
  - **Set a list of allowed users or groups per operation**
    - Job submission
    - Job administration
  - **No authentication**

Big Data Security

- Securing a Cluster Through a Gateway
  - **Hadoop cluster runs on a private network**
  - **Gateway server dual-homed (Hadoop network and public network)**
  - **Users SSH onto gateway**
    - Optionally can create an SSH proxy for jobs to be submitted from the client machine
  - **Provides minimum level of protection**

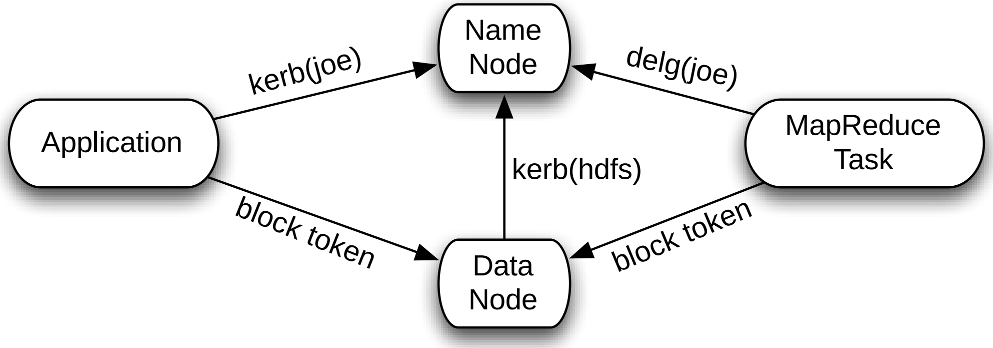
SSH: Secure Shell, a cryptographic network protocol, used to log into a remote machine and execute commands

Big Data Security

- Why security matters
  - Prevent Accidental Access
    - Doesn't require strong authentication
  - Stop Malicious Users
    - Security has to get real
    - Hadoop runs arbitrary code
    - Implicit trust doesn't prevent the insider threat
  - Big data means getting rid of stovepipes
    - Scalability and flexibility are only 50% of the problem
    - Trust your data in a multi-tenant environment

Big Data Security

- An evolving story
  - Files
  - Service-level authorization
    - Whitelists and blacklists of hosts and users
- Authentication
  - HADOOP-4487, Hadoop 0.22 and 0.20.205, Late 2010
  - Based on Kerberos and internal delegation tokens
    - Provides strong user authentication
    - Also used for service-to-service authentication



Big Data Security

- Encryption with Hadoop
  - Over the wire encryption for some socket connections
  - RPC (Remote Proc. Call) encryption added after Kerberos
  - Shuffle encryption (HTTPS) added in Hadoop 2.0.2-alpha
  - HDFS block streamer encryption added in Hadoop 2.0.2-alpha
  - Volume-level encryption for data at rest
- Pluggable Authentication
  - Currently supports username/password authentication backed by ZooKeeper
  - Authentication info replaced with generic tokens
  - Supports multiple security level implementations (e.g. Kerberos)

Big Data Security

Application Level

- **Apache Accumulo is a computer software project that developed a sorted, distributed key/value store based on the BigTable technology from Google**
  - Accumulo often paired with application level authentication/authorization
  - Accumulo users created secure access level per application
  - Each application granted access level of most permitted user
  - Application authenticates users, grabs user authorizations, passes user labels with requests

Big Data Security

Apache Hbase

- **HBase is the Hadoop database, a distributed, scalable, big data store**
  - Also based on Google's BigTable
  - Started as a Hadoop contrib project
  - Supports column-level ACLs
  - Kerberos for authentication
  - Discussion and early prototypes of cell-level security ongoing

Future: Encryption for Data at Rest

- **Need multiple levels of granularity**
- **Encryption keys tied to authorization labels (like Accumulo labels or HBase ACLs)**
- **APIs for file-level, block-level, or record-level encryption**

HBase

- HBase is an open-source, distributed, column-oriented database built on top of HDFS based on BigTable!
  - Designed to operate on top of the Hadoop
  - Distributed file system (HDFS) or Kosmos File System (KFS, aka Cloudstore) for scalability, fault tolerance, and high availability.
    - No real indexes
    - Automatic partitioning
    - Scale linearly and automatically with new nodes
    - Commodity hardware
    - Fault tolerance
    - Batch processing

HBase Security

- Cell Tags
  - All values in HBase are now written in cells, can also carry arbitrary no. of tags such as metadata
- Cell ACLs
  - Enables the checking of (R)ead, (W)rite, E(X)excute, (A)dmin & (C)reate
- Cell Labels
  - Visibility expression support via new security coprocessor

HBase Security

- Transparent Encryption
  - Data is encrypted on disk – HFiles are encrypted when written and decrypted when read
- RBAC
  - Uses Hadoop Group Mapping Service and ACL's to implement

Encryption

- Over the wire encryption for some socket connections
  - RPC encryption added soon after Kerberos
  - Shuffle encryption (HTTPS) added in Hadoop 2.0.2-alpha, back ported to CDH4 MR1
  - HDFS block streamer encryption added in Hadoop 2.0.2-alpha

Encryption

- Over the wire encryption for some socket connections
  - Need multiple levels of granularity
  - Encryption keys tied to authorization labels (like HBase ACLs)
  - APIs for file-level, block-level, or record-level encryption
  - Volume-level encryption for data at rest

Cell Level Security

- Cell level ACL means explicit 'RW' access can be set on individual cells when the cell data is put into HBase.
- Visibility labels allow administrators to associate secure access to cells with visibility labels

Cell Level Security

- Labels stored per cell
- Labels consist of Boolean expressions (AND, OR, nesting)
- Labels associated with each user
- Cell labels checked against user's labels with a built-in iterator

Cell Level Security

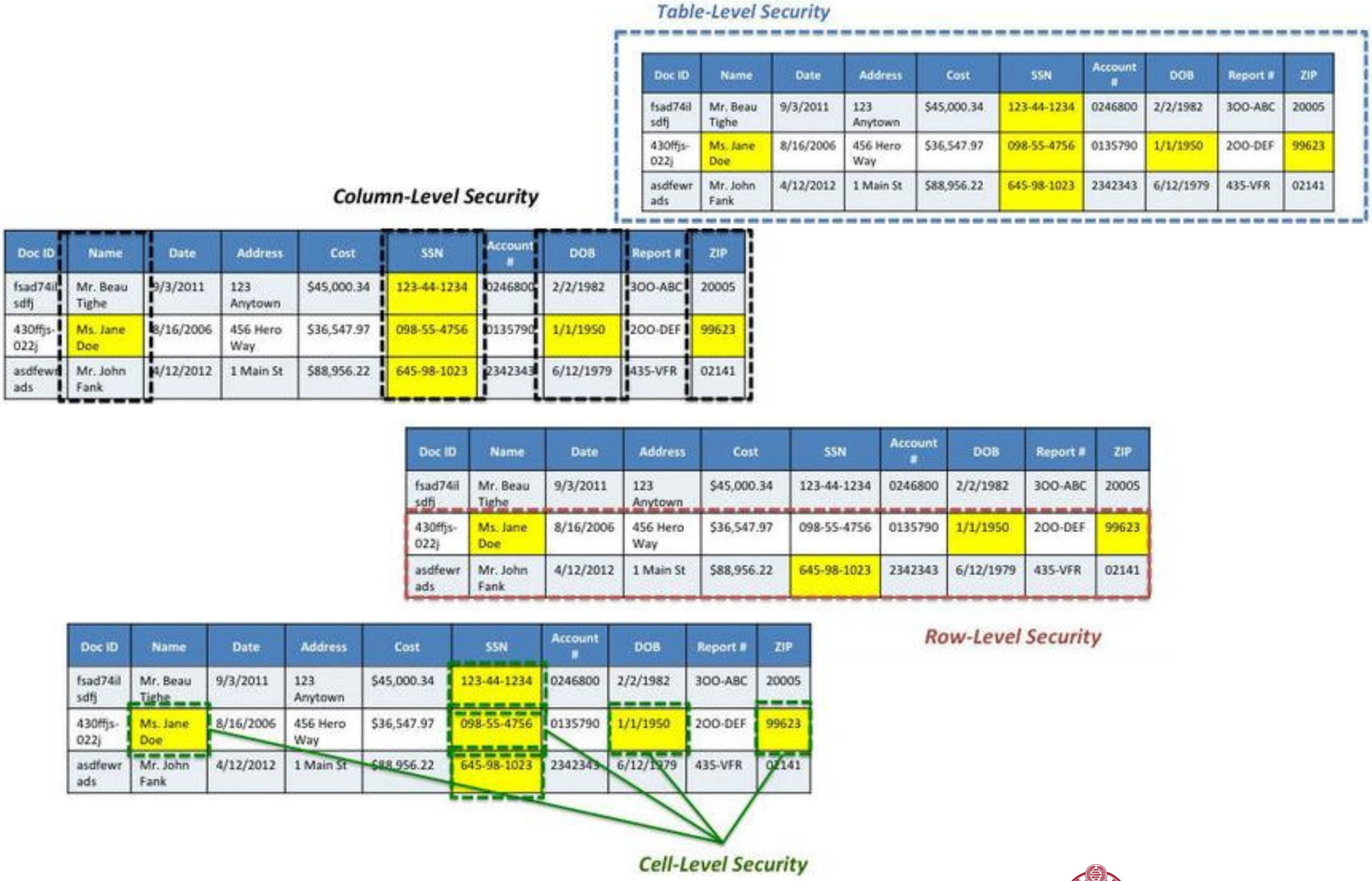


Table-Level Security

Doc ID	Name	Date	Address	Cost	SSN	Account #	DOB	Report #	ZIP
fsad74il sdfj	Mr. Beau Tighe	9/3/2011	123 Anytown	$45,000.34	123-44-1234	0246800	2/2/1982	300-ABC	20005
430ffjs-022j	Ms. Jane Doe	8/16/2006	456 Hero Way	$36,547.97	098-55-4756	0135790	1/1/1950	200-DEF	99623
asdfewr ads	Mr. John Fank	4/12/2012	1 Main St	$88,956.22	645-98-1023	2342343	6/12/1979	435-VFR	02141

Column-Level Security

Doc ID	Name	Date	Address	Cost	SSN	Account #	DOB	Report #	ZIP
fsad74il sdfj	Mr. Beau Tighe	9/3/2011	123 Anytown	$45,000.34	123-44-1234	0246800	2/2/1982	300-ABC	20005
430ffjs-022j	Ms. Jane Doe	8/16/2006	456 Hero Way	$36,547.97	098-55-4756	0135790	1/1/1950	200-DEF	99623
asdfewr ads	Mr. John Fank	4/12/2012	1 Main St	$88,956.22	645-98-1023	2342343	6/12/1979	435-VFR	02141

Doc ID	Name	Date	Address	Cost	SSN	Account #	DOB	Report #	ZIP
fsad74il sdfj	Mr. Beau Tighe	9/3/2011	123 Anytown	$45,000.34	123-44-1234	0246800	2/2/1982	300-ABC	20005
430ffjs-022j	Ms. Jane Doe	8/16/2006	456 Hero Way	$36,547.97	098-55-4756	0135790	1/1/1950	200-DEF	99623
asdfewr ads	Mr. John Fank	4/12/2012	1 Main St	$88,956.22	645-98-1023	2342343	6/12/1979	435-VFR	02141

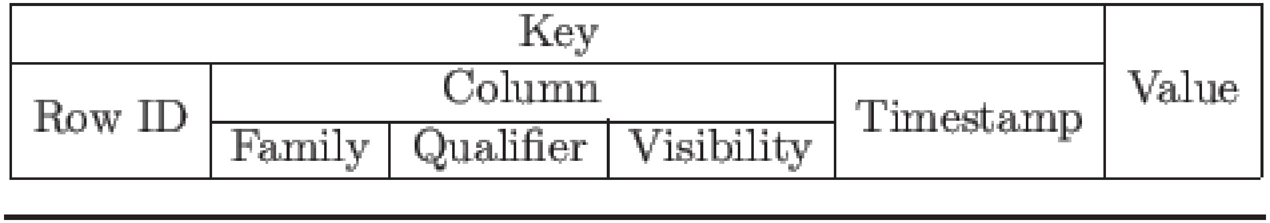
Row-Level Security

Doc ID	Name	Date	Address	Cost	SSN	Account #	DOB	Report #	ZIP
fsad74il sdfj	Mr. Beau Tighe	9/3/2011	123 Anytown	$45,000.34	123-44-1234	0246800	2/2/1982	300-ABC	20005
430ffjs-022j	Ms. Jane Doe	8/16/2006	456 Hero Way	$36,547.97	098-55-4756	0135790	1/1/1950	200-DEF	99623
asdfewr ads	Mr. John Fank	4/12/2012	1 Main St	$88,956.22	645-98-1023	2342343	6/12/1979	435-VFR	02141

Cell-Level Security

Cell Level Security

- Data Model
  - Multi-dimensional, persistent, sorted map
  - Key/Value store with a twist
  - A single primary key (Row ID)
  - Secondary key (Column) internal to a row
    - Family
    - Qualifier
  - Per-cell timestamp



Key					Value
Row ID	Column			Timestamp	
	Family	Qualifier	Visibility		

- Data Administration
  - Big Data: Hadoop / Mongo DB / HBASE
  - Data Policies, Data Quality
  - Data Ownership
  - Data Warehousing
  - Long Term Archival
  - Data Validation
  - Data Security (access control, encryption)

Data Policies

- Data is any type of stored digital information that is asset.
- Security is about the protection of assets.
- Prevention
  - Measures taken to protect your assets from being damaged.
- Detection
  - Measures taken to allow you to detect when an asset has been damaged, how it was damaged and who damaged it.
- Reaction
  - Measures that allow you to recover your assets.

Data Policies

- Data confidentiality, integrity, and availability.
- Accountability is audit information that is kept and protected so that security actions can be traced to the responsible party.

Data Policies

- Audit Standards
  - Data Security is subject to several types of audit standards and verification.
    - The most common are ISO 17799, ISO 27001-02, PCI, ITIL, SAS-70, HIPPA, SOX
  - Security Administrators are responsible for creating and enforcing a policy that forms to the standards that apply to their organizations business.
  - IT certification audits are generally carried out by 3rd party accounting firms.

Data Policies

- Security Policies
  - **A security policy is a comprehensive document**
    - That defines a companies' methods for prevention, detection, reaction, classification, accountability of data security practices and enforcement methods.
      - It generally follows industry best practices as defined by ISO 17799,27001-02, PCI, ITIL, SAS-70, HIPPA , SOX or a mix of them.

Data Policies

- Security Policies
  - **A security policy is a comprehensive document**
    - The security policy is the key document in effective security practices.
      - Once it has been defined it must be implemented and modified and include any exceptions that may need to be in place for business continuity.
    - All users need to be trained on these best practices with continuing education at regular intervals.

Data Policies

- Tools to Secure Data
  - Data needs to be classified in the security policy according to its sensitivity.
  - Once this has taken place, the most sensitive data has extra measures in place to safeguard and ensure its integrity and availability.
  - All access to this sensitive data must be logged.

Data Policies

- Tools to Secure Data
  - **Secure data is usually isolated from other stored data.**
    - Controlling physical access to the data center or area where the data is stored.
    - Active or Open Directory is a centralized authentication management system that is available to companies to control and log access to any data on the system.
  - **Encryption of the sensitive data is critical before transmission across public networks.**

Data Policies

- Tools to Monitor Secure Data
  - Walk around and look for passwords in the open.
  - Event Viewer / Log Files
  - Intrusion Detection/ Protection systems (IDS/IPS) such as SNORT.
  - These will alert Administrators of suspicious data flows.

Data Policies

- Tools to Document Data Security
  - **Microsoft Visio is the standard for drawing network maps.**
    - These maps allow a detailed overview of the system and how it is functions.
    - They also allow the spotting of weak points of security and flaws in design that can impact reliability or continuity of the data to the end user.

Data Policies

- End User Education
  - All relevant security polices must be clearly explained to the end users.
  - A clear explanation of the consequences for violating these polices must also be explained.
  - The end user needs to sign a document acknowledging that they understand the policies and consequences for violating these policies.

Data Policies

- Enforcement
  - Must obtain executive authority to enforce security policy.
  - Systematic approach of warnings and punishments.
  - Coordinate with HR to document continued issues with staff.

Data Administration

- Big Data: Hadoop / Mongo DB / HBASE
- Data Policies, Data Quality
- Data Ownership
- Data Warehousing
- Long Term Archival
- Data Validation
- Data Security (access control, encryption)

Data Quality

- Data in the real world is of low quality (dirty!)
  - Poor data across businesses and the government costs the U.S. economy $3.1 trillion dollars a year.
  - Poor data can cost businesses 20%–35% of their operating revenue.
  - Bad data or poor data quality costs US businesses $600 billion annually.

Data Quality

- Data in the real world is of low quality (dirty!)
- When data can be of high quality?
  - The data is accurate
  - The data is stored according to data type
  - The data has integrity
  - The data is consistent
  - The databases are well designed
  - The data is not redundant
  - The data follow business rules
  - The data corresponds to established domains
  - The data is timely
  - The data is well understood
  - etc.

Data Quality

- Data in the real world is of low quality (dirty!)
  - Inconsistent
    - Containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records
  - Incomplete
    - Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - Noisy
    - Containing errors or outliers
    - e.g., Salary="-10"

Data Quality

- Data in the real world is dirty
  - Illegal values
  - Violated attribute dependencies
  - Uniqueness violation
  - Referential integrity violation
  - Missing values
  - Misspellings
  - Cryptic values
  - Embedded values
  - Misfielded values

Data Quality

- Data in the real world is dirty
  - Word transpositions
  - Duplicate records
  - Contradicting records
  - Wrong references
  - Overlapping data/matching records
  - Name conflicts
  - Structural conflicts

Data Quality

- Assessment of existing data quality
  - Programs that abnormally terminate with data exceptions
  - Clients who experience errors/anomalies
  - Clients who do not know or are confused about what the data actually means
  - Data that cannot be shared due to lack of integration

Data Administration

- Big Data: Hadoop / Mongo DB / HBASE
- Data Policies, Data Quality
- Data Ownership
- Data Warehousing
- Long Term Archival
- Data Validation
- Data Security (access control, encryption)

Data Ownership

- Data Owner Definition:
  - A person with statutory or operational authority for specified information (e.g., supporting a specific business function) and responsibility for establishing the controls for its generation, collection, processing, access, dissemination, and disposal

Data Ownership

- Grants access to the Information System under his/her responsibility.
- Classifies Digital Data based on Data sensitivity and risk.
- Backs up Data under his/her responsibility in accordance with risk management decisions and secures back up media.

Data Administration

- Big Data: Hadoop / Mongo DB / HBASE
- Data Policies, Data Quality
- Data Ownership
- Data Warehousing
- Long Term Archival
- Data Validation
- Data Security (access control, encryption)

Data Warehouse

- Definition
  - An enterprise structured repository of subject-oriented, time-variant, historical data used for information retrieval and decision support. The data warehouse stores atomic and summary data. [Oracle Data Warehouse Method]

Why Do We Need Data Warehouses?

- Consolidation of information resources
  - Improved query performance
  - Separate research and decision support functions from the operational systems
  - Foundation for data mining, data visualization, advanced reporting and OLAP tools

Why Do We Need Data Warehouses?

Which are our lowest/highest margin customers ?

What is the most effective distribution channel?

Who are my customers and what products are they buying?

What product prom--otions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

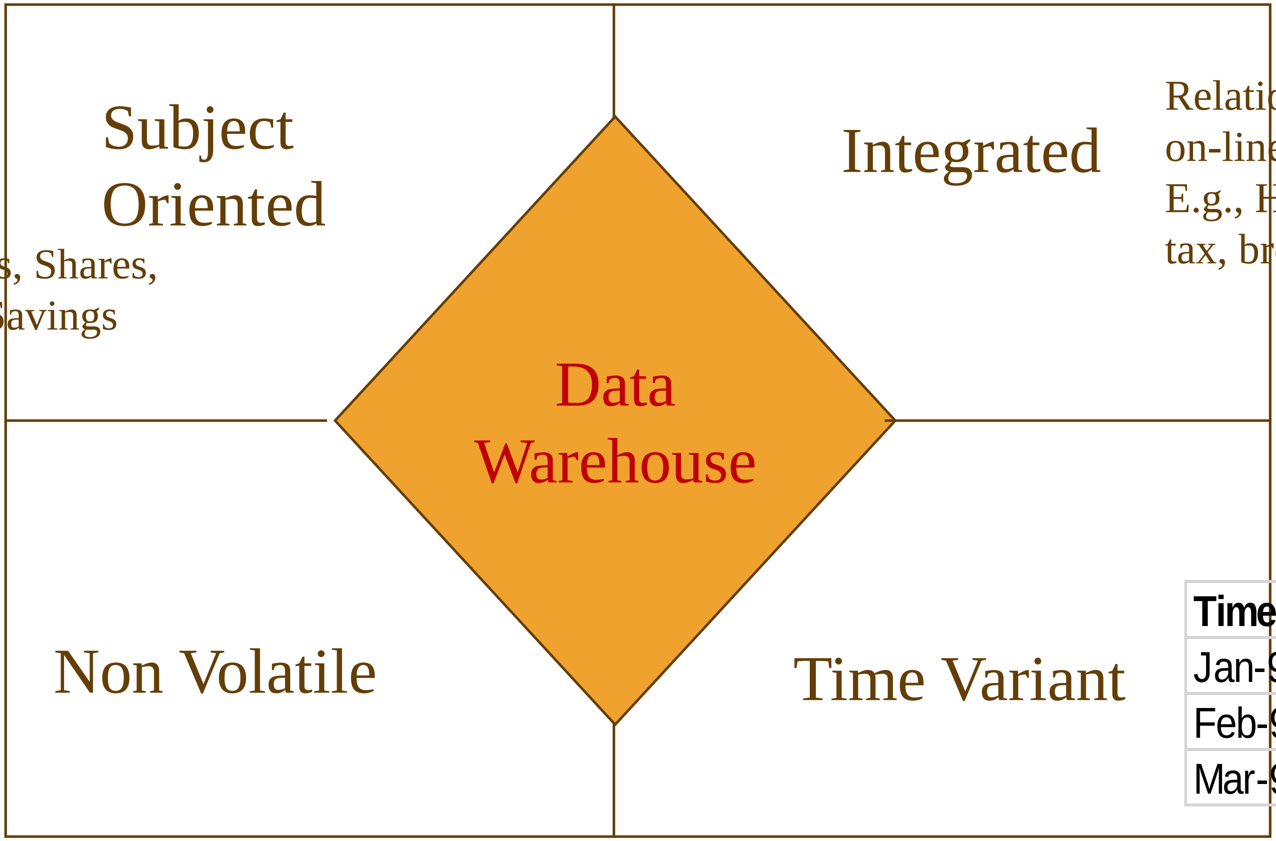
What impact will new products/services have on revenue and margins?

What Is a Data Warehouse Used for?

- Knowledge discovery
  - Making consolidated reports
  - Finding relationships and correlations
  - Examples
    - Banks identifying credit risks
    - Insurance companies searching for fraud
    - Medical research

Data Warehouse Properties

○



Subject Oriented

Integrated

Data Warehouse

Non Volatile

Time Variant

Equity Plans, Shares,
Insurance, Savings

Relational databases, flat
on-line transaction recor
E.g., Hotel price: currenc
tax, breakfast covered, et

Load, Read
No:
Modificati
on, update,

Time	Data
Jan-97	January
Feb-97	February
Mar-97	March

How Do Data Warehouses Differ From Operational Systems?

- Operational
  - Online transaction and query processing
  - Day-to-day operations
  - Users and system orientation: customer oriented
    - Online transaction processing (OLTP)
- Data warehouse
  - Serve users or knowledge workers in the role of data analysis and decision making
  - Decision support: operations as defined or needed
  - Market-oriented
    - Online analytical processing (OLAP)

How Do Data Warehouses Differ From Operational Systems?

Property	Operational DB	Data Warehouse
Response time	Sub seconds to seconds	Seconds to hours
Operations	Historic data	Primarily read only
Nature of Data	30-60 days	Snapshots over time
Data Organization	Applications	Subject, time
Size	Small to large	Large to very large
Data Source	Operational, Internal	Operational, Internal, External
Activities	Processes	Analysis

How Do Data Warehouses Differ From Operational Systems?

OLTP Examples	OLAP Examples
Which product is mostly purchased (or sold)?	Amazon analyzes purchases by its customers to come up with an individual screen with products of likely interest to the customer.
Answering queries from a Web interface, sales at cash registers, e.g., selling particular medicine.	Analysts at Wal-Mart look for items with increasing sales in some region.

Decision Support

- Used to manage and control business
- Data is historical or point-in-time
- Optimized for inquiry rather than update
- Use of the system is loosely defined and can be ad-hoc
- Used by managers and end-users to understand the business and make judgements

Data Warehouse Models

- Enterprise warehouse
  - Collects all of the information about subjects spanning the entire organization
- Data mart
  - A subset of corporate-wide data that is of value to a specific groups of users.
  - Its scope is confined to specific, selected groups, such as marketing data mart
- Virtual warehouse
  - A set of views over operational databases
  - Only some of the possible summary views may be materialized

Warehouse Products

- Computer Associates -- CA-Ingres
- Hewlett-Packard -- Allbase/SQL
- Informix -- Informix, Informix XPS
- Microsoft -- SQL Server
- Oracle – Oracle
- Red Brick -- Red Brick Warehouse
- SAS Institute -- SAS
- Software AG -- ADABAS
- Sybase -- SQL Server, IQ, MPP

Security in the Data Warehouse

- Unfortunately, most data warehouses are built with little or no consideration given to security during the development phase.
  - Basic security concepts
  - Physical security
  - Stand-alone or shared security
  - Remote access

Security in the Data Warehouse

- Achieving proactive security requirements of DW is a seven-phase process:
  1) identifying data
  2) classifying data
  3) quantifying the value of data
  4) identifying data security vulnerabilities
  5) identifying data protection measures and their costs
  6) selecting cost-effective security measures, and
  7) evaluating the effectiveness of security measures. These phases are part of an enterprise-wide vulnerability assessment and management program.

Security in the Data Warehouse

- Controlling Access to Warehouse Data
- Role-based Access Control
- Row-Level Security
- Virtual Private Database (VPD)
  - Column-Level Virtual Private Database
  - Policy Types for Added Performance

Data Administration

- Big Data: Hadoop / Mongo DB / HBASE
- Data Policies, Data Quality
- Data Ownership
- Data Warehousing
- Long Term Archival
- Data Validation
- Data Security (access control, encryption)

Long Term Archival

- Long Term Data that can be entered and/or checked over a longer time frame, using collaborative arrangements
- Archiving is for long-term storage of data that is not required immediately.
- More often than not it is never required again but it kept just in case.
- Data is often removed from a system and stored separately.

Long Term Archival

- Consider the following example:
  - A school records data about pupils’ performance every year. If they continued to collect data, even after pupils had left school, the system resources would soon diminish. Instead, records about pupils are removed from the system once they leave.
  - However, some data may be archived such as average test scores and achievement rates. The data is not needed immediately but may be useful to keep for the future.

Data Administration

- Big Data: Hadoop / Mongo DB / HBASE
- Data Policies, Data Quality
- Data Ownership
- Data Warehousing
- Long Term Archival
- Data Validation
- Data Security (access control, encryption)

Data Validation

- It allows us to provide data with confidence in its accuracy, and we can consistently provide this data by implementing thorough security.

- Data Integrity
  - Validity, consistency, and accuracy of the data in a database.
    - Table-level
    - Field-level
    - Relationship-level
    - Business Rules

Data Validation

- The process of determining if an update to a value in a **table's data cell** is within a preestablished range or is a member of a set of allowable values.

Data Validation

- Methods for checking for data validity
  - Visual/manual
  - Aggregation
  - Reviewers guide
  - Auto data checks
  - Record counts
  - Spell checks
  - Have data provider review

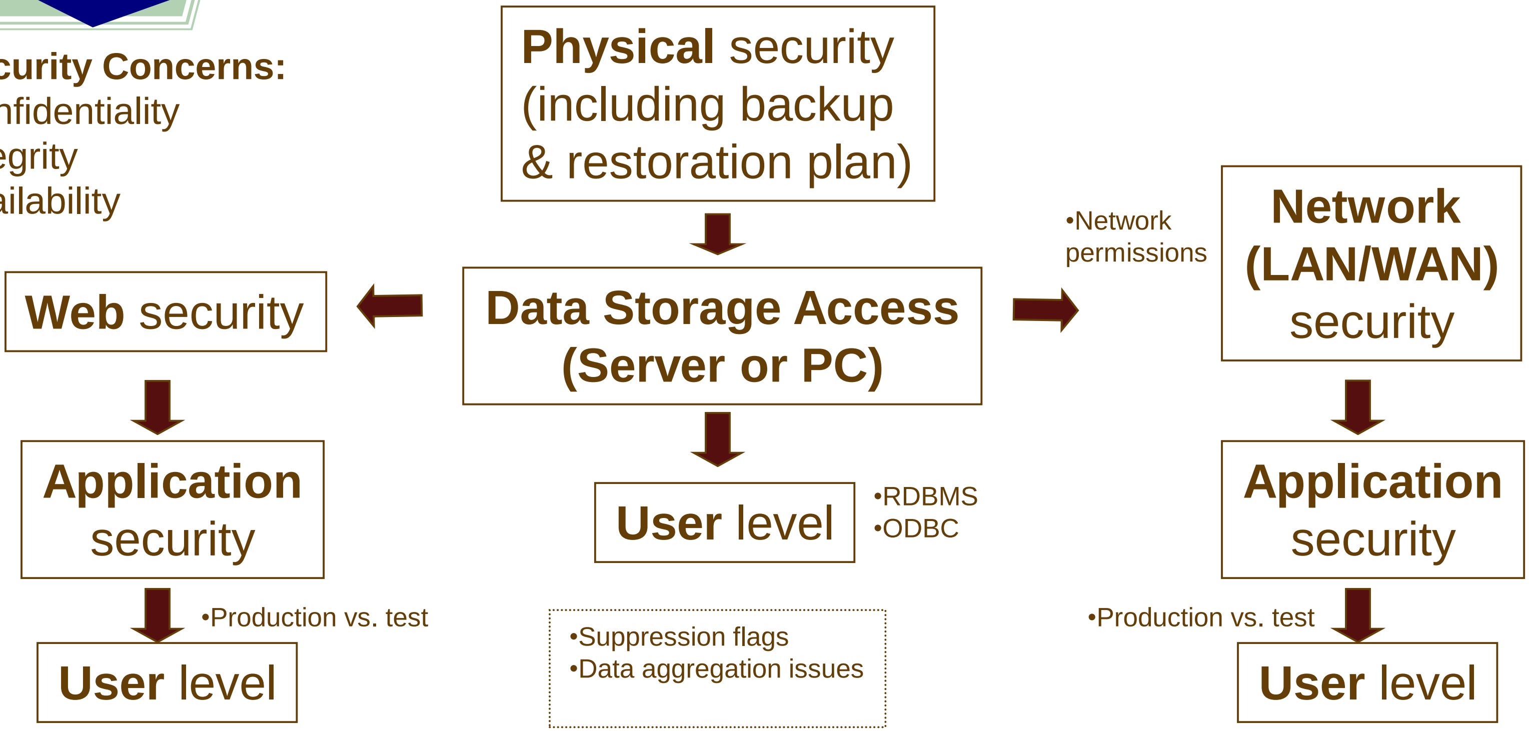
Data Validation

○ Secure Systems for Data Validation

Data Security Considerations

The contact in my state is:______________

Security Concerns:
Confidentiality
Integrity
Availability



Physical security (including backup & restoration plan)

Data Storage Access (Server or PC)

Web security

Network (LAN/WAN) security

- Network permissions

Application security

- Production vs. test

User level

User level

- RDBMS
- ODBC

- Suppression flags
- Data aggregation issues

Application security

- Production vs. test

User level

Database security prevents unauthorized person(s) from viewing, destroying or altering data within the database.

Coming Attraction…

- Reading: Slides for quiz
- Assignment 1
- Next class
  - Proposal Presentaiton