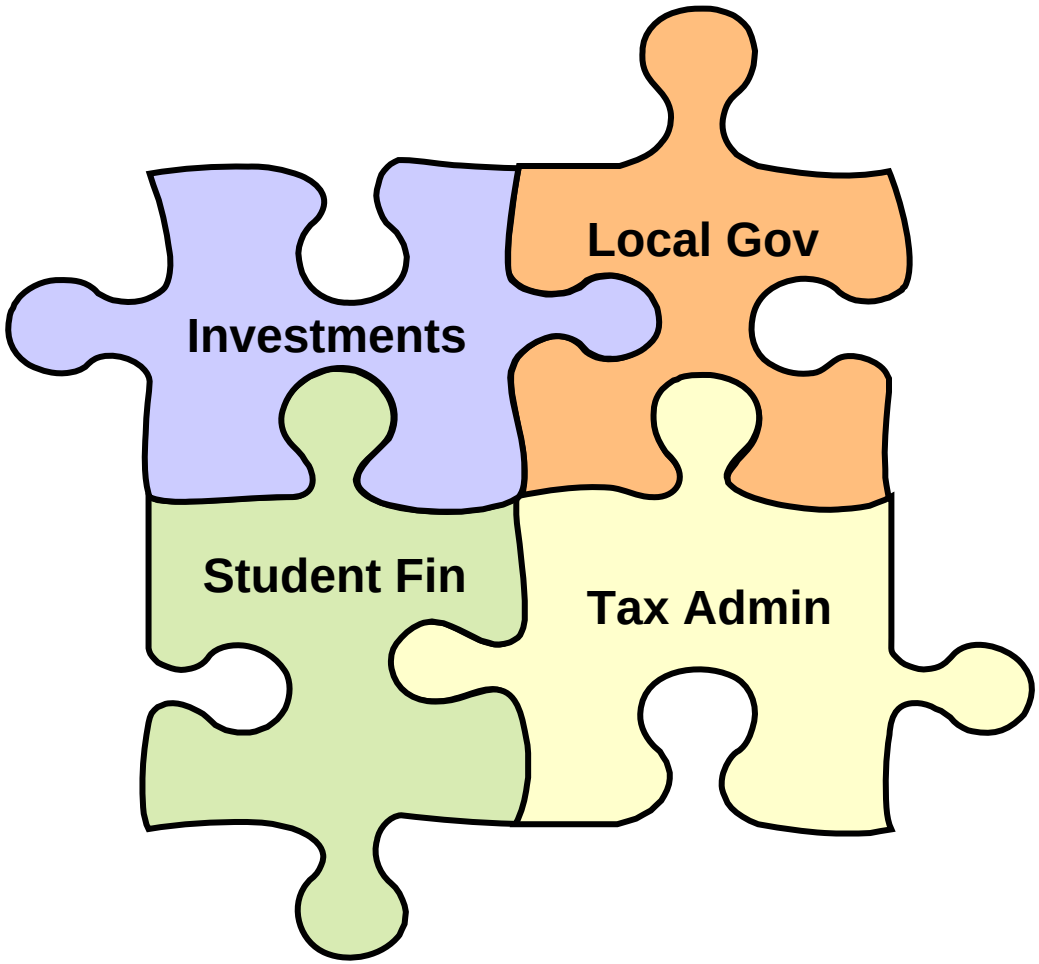
ETL: The Promise

Dawn Shattuck
Department of Treasury

ETL: The Promise

- Why the interest from Treasury
- What current challenges drive ETL use
- Why is a tool needed
- What are expected agency benefits
- Why is the timing so important
- Why should this be of interest to DW geeks / users

Treasury

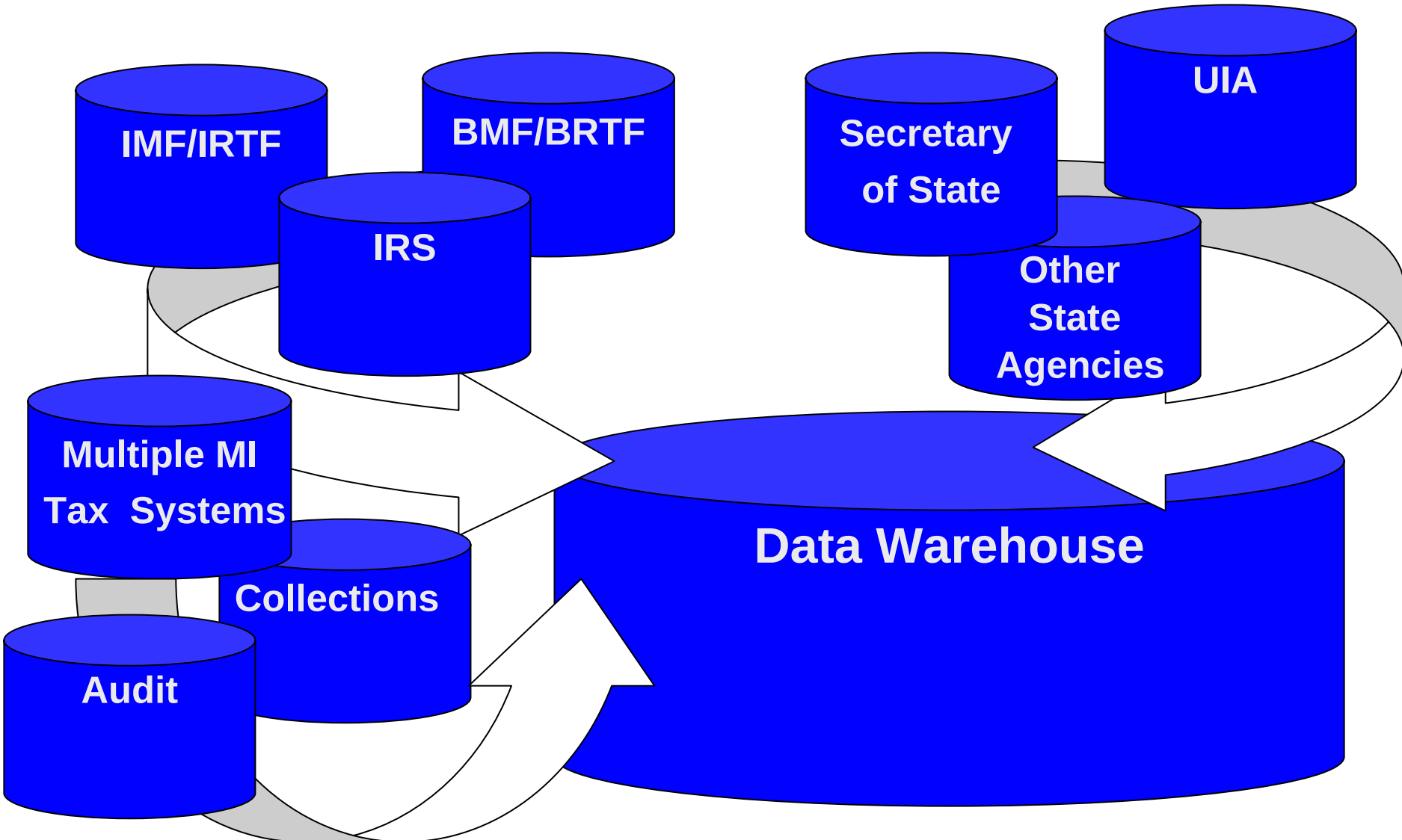


Challenge #1

- Tax Administration
  - $24.2 Billion
  - 350,000 businesses file multiple taxes
  - 5,000,000 individual taxpayers

  - Data resides in multiple aged systems
  - Limited data per tax type is in the DW
  - IRS Data tapes are manually loaded

Tax Data: Many Systems



Challenge #2

- Local Government / Revenue Sharing
  - Provides services to 3,023 local units of government, counties, schools, and other
  - Coordinates over $1 billion in bonds/notes and over $100M in loans to school districts

---

  - Analytics are needed across cities, counties, school districts, multiple bond issues, etc.
  - Better access to external data is desired

Challenge #3

- Student Financial Services
  - $240 M in loans and grants / 15 programs
  - 3 Million Student records / 1M annually
  - 1000 high schools and 100 post secondary

  - Cannot look at data across systems
  - Not able to push information to customers
  - One centralized “version of the truth” on each student does not exist across programs

Challenge #4

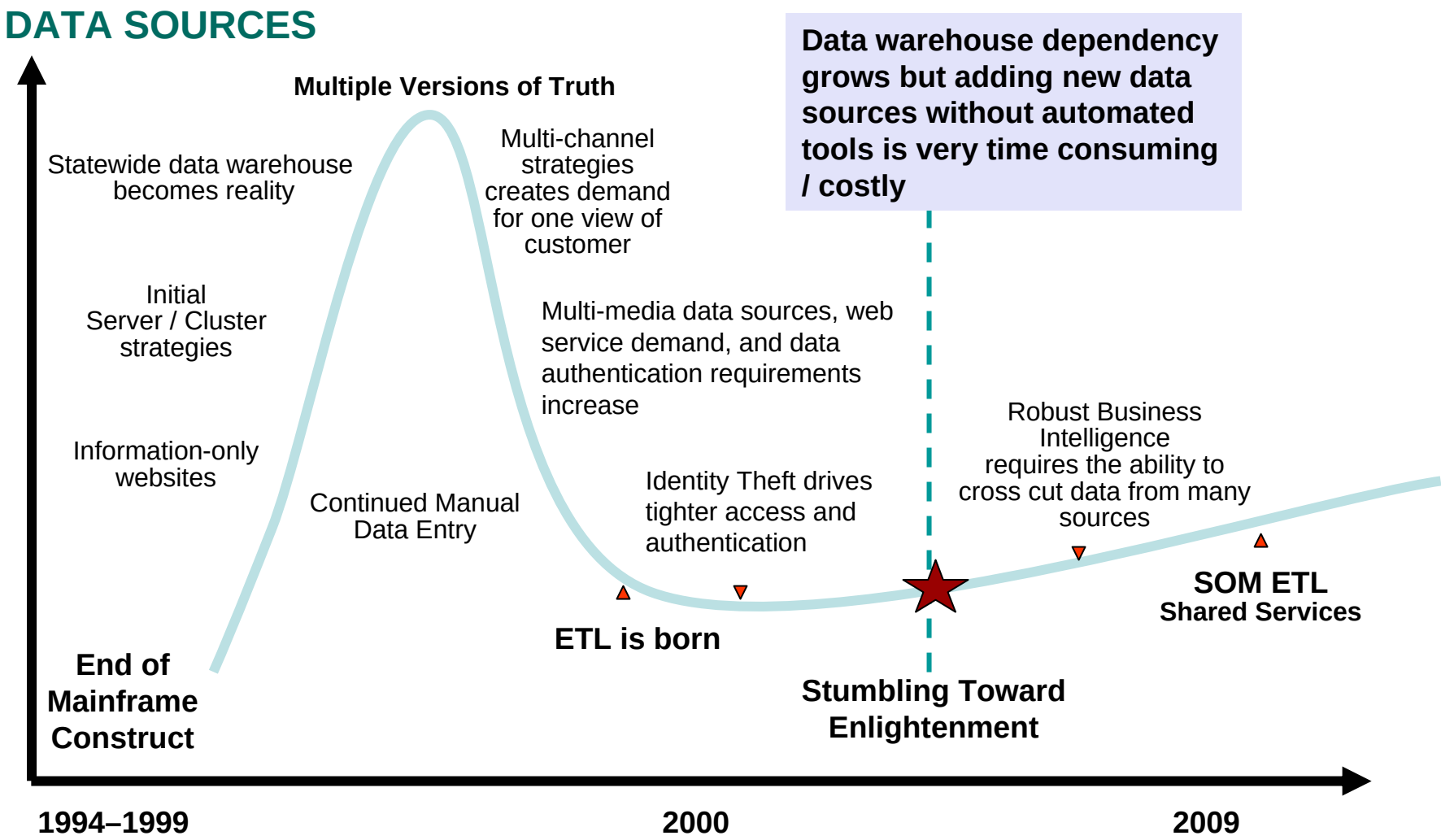
- Financial Management
  - Revenue Forecasting / critical reporting
  - Revenue Collection for taxes, state debt, etc
  - Investments manages $58B in pension funds

  - All functions require statistical analysis and forecasting across many data sources
  - Data is available but in different media and environments

Treasury Data Challenges

- Varied sources of data needed daily
- No tools available for ETL type functions
- Mature analytics needed to improve decision support, fraud detection, and critical reporting
- Users do not really know what data resides where with so many different systems
- Growing need to create one holistic view of a business or an individual
- Not able to create useful information and push out to clients in a structured method

Data Management Challenge



Complexity =Tool Required

- Until the late 1990's manipulation of data between multiple sources was very painful
- The first version of ETL (Extract, Transform, and Load) provided hope but not much more
- The toolset still required complex scripting and data mapping by hand and…
- The middleware package was expensive to purchase and deploy

ETL: The Big Easy *@?!!

New and Improved ETL

- Gui toolset facilitates rapid deployment
- Improved rule establishment for complex data manipulation and delivery
- Virtual data management and manipulation; does not require co-location of data sources
- External sources can be scheduled to load immediately and can be transformed on the fly
- Data / information can be delivered to other systems like web services or CRM
- Products now contain the ability to handle partitioning and parallel processing abilities

ETL V2: Agency Benefits

- Ability to standardize data from multiple environments into one useable format
- Ability to create rules for matching, cleansing, and transforming data on the fly
- Ability to create one view of business / individual across multiple systems, medias, and programs
- Ability to transform and merge data and deliver for use in various agency operations / services
- Ability to better manage our data universe in a more holistic manner

Why Now ?

- Cannot afford loss of revenue by data not being loaded in a timely manner
- Cannot take 100-200 manhours of effort to load tapes from IRS multiple times / year
- Can no longer afford to hand write scripts and still pull or push bad data into the DW
- Can no longer afford the one off approach to reoccuring work

Keeping the End in Mind

- Maximize decision support; fraud detection
- Ability to cross pollinate data from multiple sources to create the most accurate and up-to-date information
- Being able to deliver “improved” data to multiple systems like CRM, audit, collections, web self service,etc.
- Minimize data replication within our systems to alleviate cost overhead, etc.

Treasury Status

- Align technology with business needs; implement ETL tool set in pilot
- Assess current data model(s) in use in the data warehouse (and other key systems); determine what data do we have and how is it related or redundant
- Start small/ simple with a straight forward load process; load our IRS data loads from tape as they are time sensitive for various auditing processes

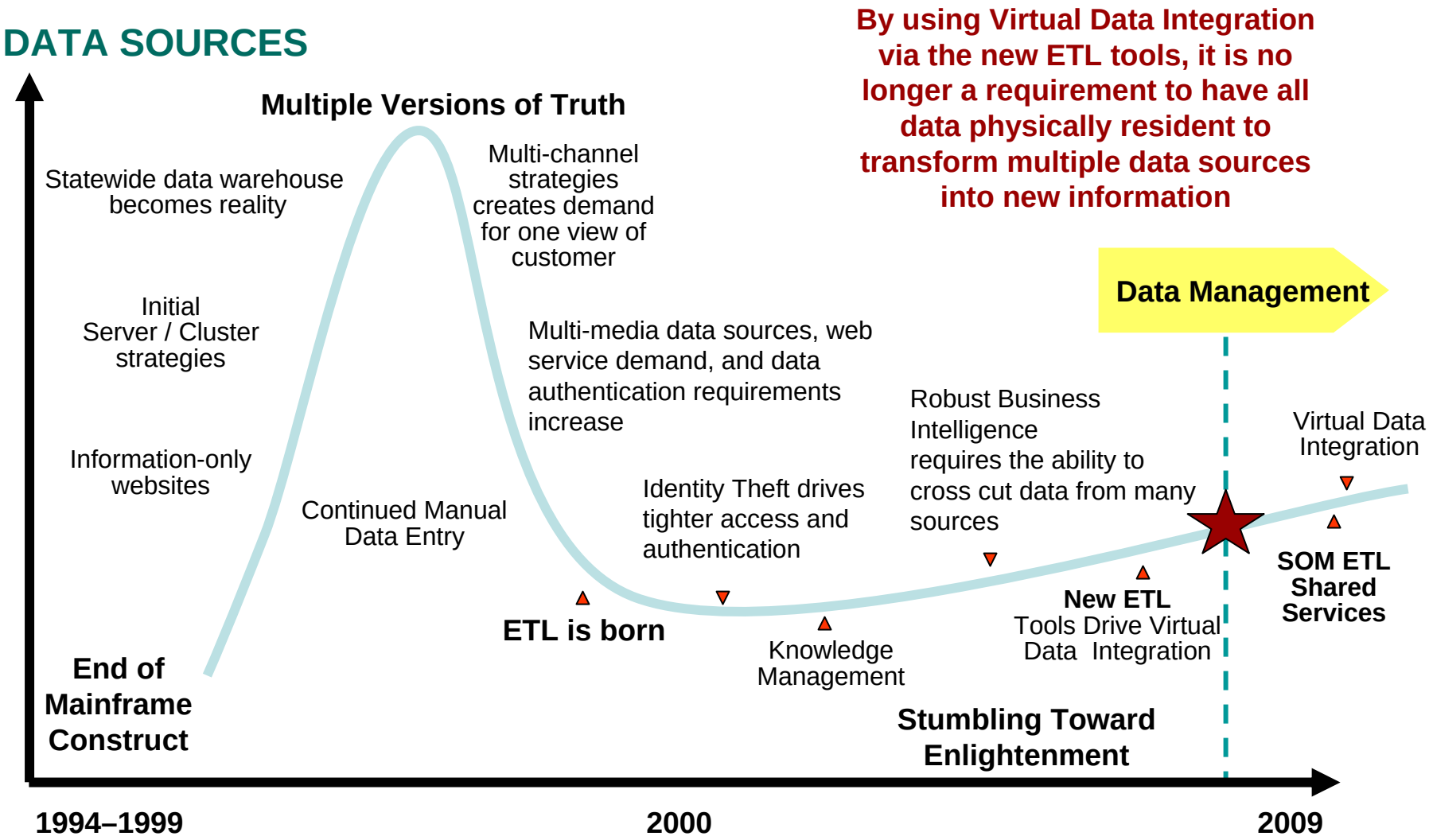
Next Steps

- Next step is transformation of data from disparate systems /media into real information
- Extraction, transformation, and/or Disbursement or delivery of “information” to other departments, other states, feds, customers.
- Virtual data integration to keep from redundant storage of data in the DW

So What ?

- Most departments run on data and access to multiple sources of data
- Data integrity / accuracy is vital to the quality of our operations and services
- Most departments with the use of ETL will be able to better utilize their data and the data warehouse
- No more Cobol pgms and load scripts needed; faster completion of work requests
- Data management that is truly holistic and seamless and less redundant storage of data in multiple places
- Virtual data integration is critical to advanced customer self service / web services

Data Management: NEXT



ETL: Promise Delivered

Michigan Data Warehouse Users Group
Agencies Sharing Knowledge

Michigan State Police

State Court
Administrators Office

Automated Data Matching Project

Chad Canfield
Michigan State Police

Overview of Automated Data Matching Project

The Problem?

- Completing Criminal History Records with Missing Court Disposition Data

The Systems

- Criminal History System
- Michigan Court Disposition reporting

The Players

- Michigan State Police-Criminal Records Division
- State Court Administrators Office-representing District and Circuit Courts

The Solution

- Automated Data Matching using the Data on the State Data Warehouse

Criminal History Background

Act 289 of 1925 –Gives MSP the role of CHR central repository

- Shall procure and file for purposes of criminal identification criminal history record information on all persons who have been convicted within this state of either a **felony** or a **misdemeanor**
- Agencies **may** submit arrests for misdemeanor offenses where the maximum sentence is 92 days or less
- It is permissible to send Local Ordinance violations in that “substantially correspond” to state statue
- Shall provide access to criminal history record information and juvenile history record information, as prescribed by the department and as authorized by law.
- A copy of an arrest card shall be forwarded to the federal bureau of investigation

Criminal History Background cont.

All Criminal History is Based on Fingerprint Impressions

There are three components of a complete criminal history:

- Arrest Segment-Consists of a Fingerprint Submission taken by the police
- Prosecutor Segment-Consists of the charges that will be brought against an individual
- Judicial Segment-Consists of the disposition data for the case (final charges and outcome)

This creates one line on a rapsheet and is considered one “Incident”

Criminal History Background cont.

Judicial Records are Reported to MSP Electronically or by Paper

- There are over 150 Circuit and District courts in Michigan
- There are two major software systems that cover a majority of courts in the state.
  - **JMS**: Judicial Management System
  - **JIS:** Judicial Information Systems
- There are at least 20 other homegrown court information systems that report disposition information to CHR electronically
- More than 30,000 paper dispositions are still submitted yearly

Criminal History Background cont.

Criminal History Records are permanent

- Criminal records do not “go away” after seven years like driving records
- Criminal History Records are removed from system in three ways:
  - Person on system reaches age 99
  - MSP is notified and can verify that individual is deceased
  - Court orders record expunged

Criminal History Records are considered public unless the law specifically makes them non-public:

- Juvenile case convictions under adult statue are public
- Certain cases have a public component until probation is completed
- (7411, Spouse Abuse and Parental Kidnapping)
- Conviction Set-Aside can be done after 5 years

The Problem

Finding Court Disposition Data on Incomplete Cases

- Overall the CHR has above 95% completeness on disposition
- Most data is now reported via electronic interface from court to MSP
- However 5% incomplete on a volume of over 100,000 dispositions per year adds up
  - 5% per year = 5,000 plus records per year
  - 20 years x 5,000 per year equals a candidate pool of over 100,000 possible incident records

The Problem, cont.

CHR Auditor Assigned Full-Time to Work the Problem

- Over 5 year period of time est. that < 10,000 records total were found.
- Open case reports sent monthly to courts
- CHR user community is active in completing records as they find them:
  - Law Enforcement
  - Prosecutors
  - Courts
  - Parole and Probation Officers
  - Employer
  - Citizens
  - National Instant Criminal Checks system (NICS) for gun buyers

The Light goes On!!

The Decision was made to set up a Special Project to resolve all the missing records

- Under the current process, finding Dispositions was time consuming, costly and only done on a case by case basis.
- The steps involved in adding a Disposition to the record:
  - An agency becomes aware of missing CHR information
  - Research is performed
  - Phone calls made
  - Information faxed
  - System updated
  - Follow up with request originator

The Light goes On!!

Updating a Record Takes Time!

- Seven to ten day turnaround is common
- Staff has to go through old records or log into obsolete systems to obtain necessary information
- It is estimated that the manual process in total takes an average of 35 minutes per transaction:
  - regardless of whether the information is found or not
  - 35 minutes x 100,000 possible records x $30.00 per hour*=$1,749,000

*Includes benefits

The Light goes On!!

There Has to be a Better Way!!!

- This is the computer age, much of this information is out there electronically.

- How can we get access to it and create a way to update the records automatically?

The Light goes On!!

Judicial Data Warehouse (JDW) to the rescue!!!

SCAO/JDW: Mark Dobek

- In 2004 the State Court Administrative Office (SCAO) developed a central repository (JDW) for better statistical information and for information sharing.
- 219 courts are submitting data to the JDW on a regular basis.
- JDW is run over the State Secured Network only (LGNET).
- SCAO shares court data with other state agencies, including MSP, Department of Corrections, and Department of State through formal Data Sharing Agreements (DSA).

Automated Data Matching The first steps

Brainstorming Session (Spring 2007)

- MSP
- SCAO
- Bull

Could This Work?

- Can we Identify candidate records from CHR to be matched?
- Can we find matches for those records in the JDW?
- Can we feel confident to 100% certainty that the correct records were being updated?
- Can we afford not to try?

Automated Data Matching Prototype phase

Prototyping Performed

1. Bull/MSP created extract process for CHR records
2. Bull/SCAO/MSP agreed to matching criteria:
   - Criminal Tracking Number
   - Name at Arrest or SID
3. Test Courts were involved in examination of records match and validation of matches
4. Test Courts asked to sign off on match process
5. Electronic update file created and sent from JIS
6. Reject messages analyzed
7. Process fine tuned
   - Court charge data did not conform to reporting standards
   - Those records are identified for courts in validation step

Automated Data Matching Putting it into production

Prototype to Production

- September 10th, 2007 first production batch file sent
- 1000 dispositions added to CHR

Viewed as a One-time Only or “Catch-up” Process

- Will not take the place of everyday data updates
- Process has now been completed over 150 times
- To date over 43,000 counts have been added to CHR

Lessons Learned

Think of Outcomes from the Beginning

- Reports you might want from the data, for example:
  - MSP never asked what types of cases were found. Was there a disproportionate number of a certain type that is under-reported?
  - That question might have lead to some training or suggested system improvements to the field

Contact Information

Chad Canfield
Acting Manager, Criminal Records Division Michigan State Police
Canfieldc@michigan.gov
517.202.2045

Mark Dobek
Director of Judicial Information Systems State Court Administrators Office
DobekM@courts.mi.gov
517.373.8978

Kristen Pawlowski
JDW Project Manager, Ingenix/Bull
KristenPawlowski@Bull.com
517.327.2206