Costs and Implications of Moving Large Data Sets

Initial Assessment of the SHRP2 NDS Data

Safety Data Oversight Committee
November 2, 2016
J. Spotts

TRB
TRANSPORTATION RESEARCH BOARD

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

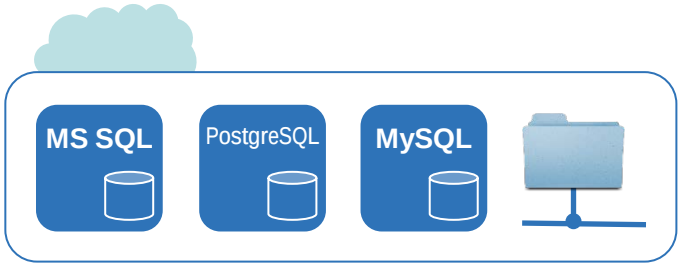
Some Metrics

- 1 petabyte (PB) = 1,000 terabytes (TB)
- 1 TB = 1,000 gigabytes (GB) – typical storage capacity for today's desktops
- 4.7 GB – storage capacity of a DVD
- 40 gigabits/sec (Gbps) – state of the art local network bandwidth

SHRP2 NDS Data Landscape

VTTI Collection Administration

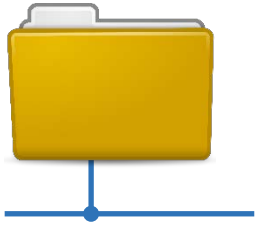
Insight Website



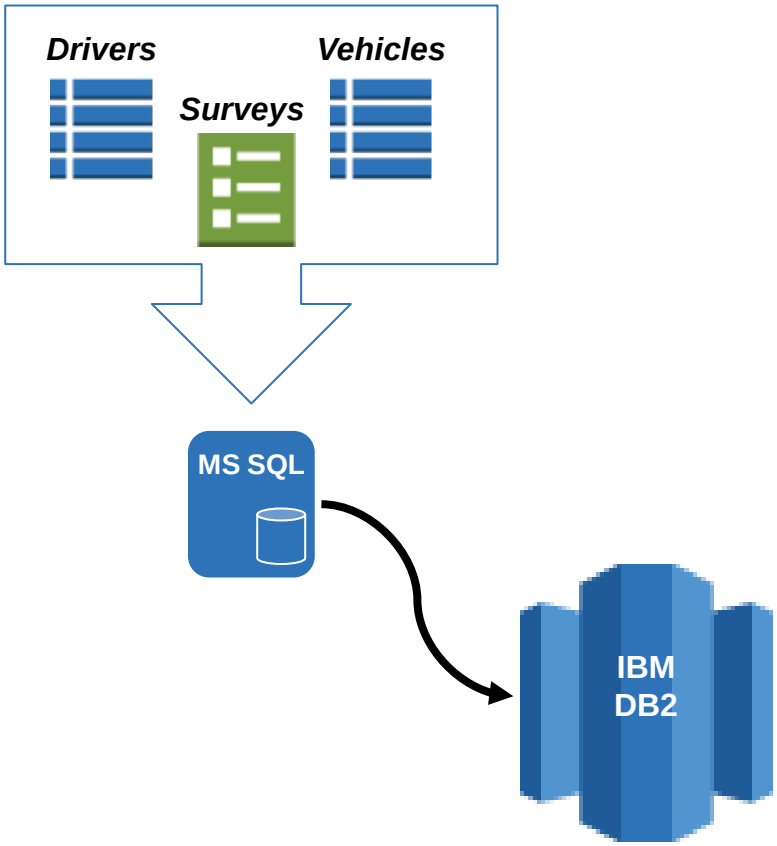
Archival Storage

Primary Database

File Store

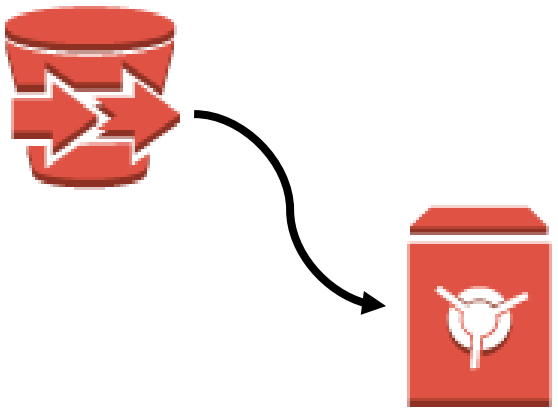


Collection Administration



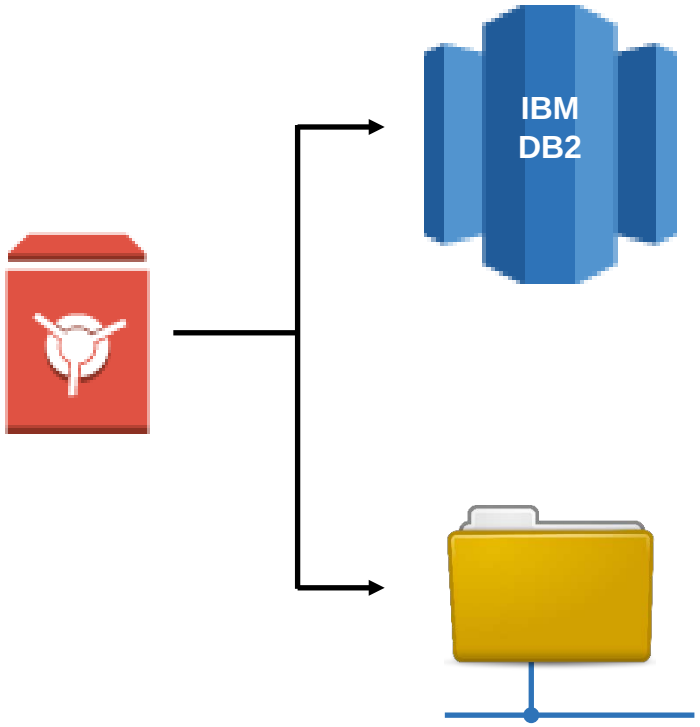
- VTTI-internal database contains sensitive information about driving studies
  - Each study represents a “Collection”
- These data include driver info, vehicle info, survey responses
- Some–but not all–of the appropriate data were loaded to primary DB2 database
- Occasionally, researchers request data elements that can only be found in this database, e.g., vehicle width, which is important in a lane keeping study
- Important non-PII would have to be identified and replicated to the primary database as part of any data migration effort

Data Collection



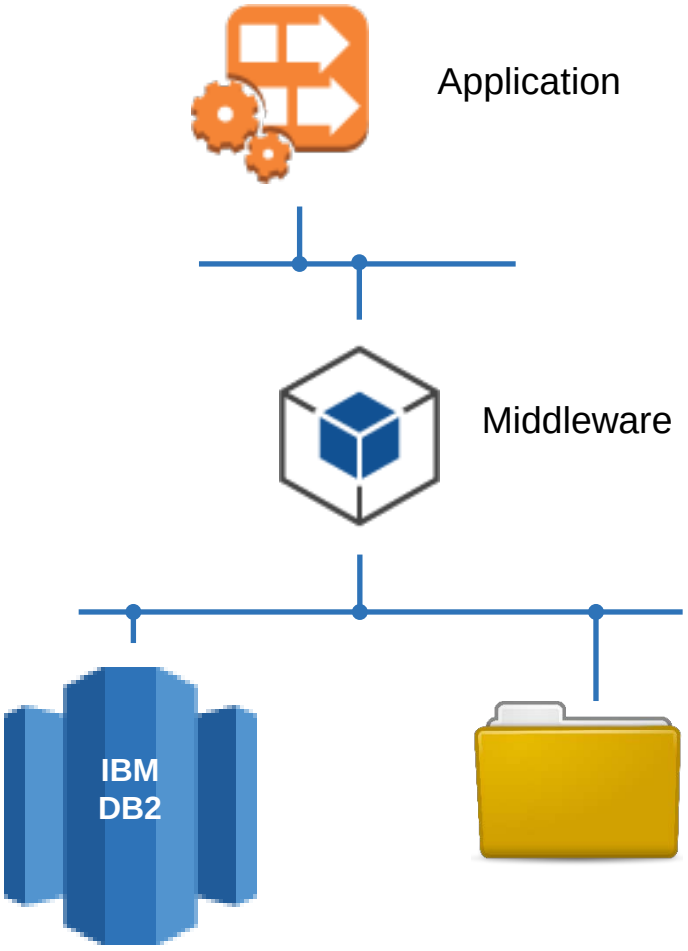
- Data from in-vehicle DAS offloaded periodically during study and uploaded to archival storage
- Approximately 1 PB of raw sensor data and video
- Derived data and research results should be traceable to these sources

Data Ingestion and Processing



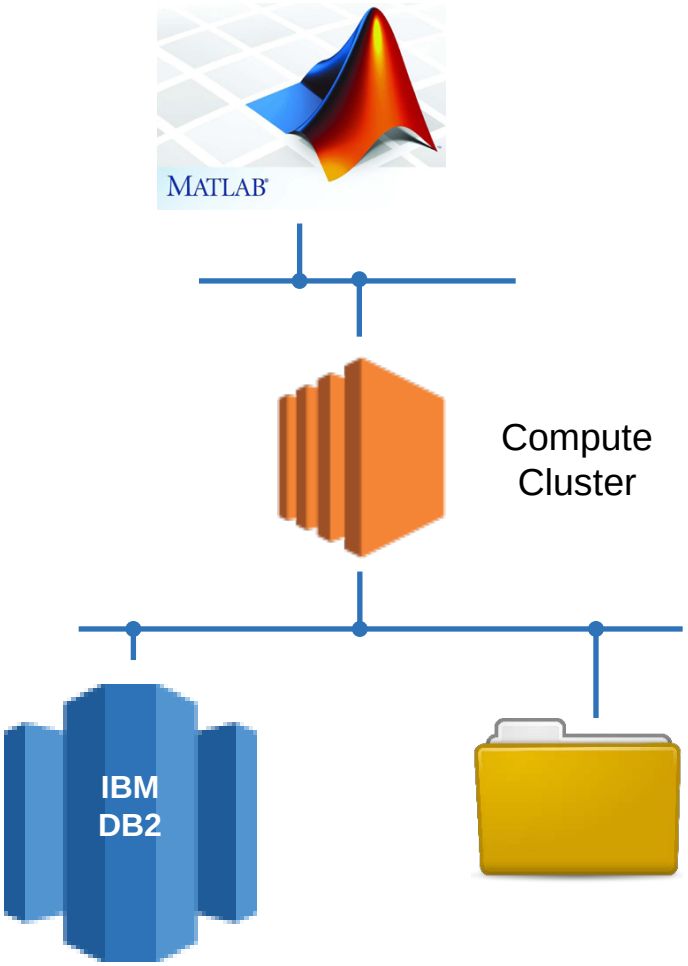
- Sensor data loaded into database tables
  - ~18 T rows
  - Raw data remains untouched
- Videos undergo format conversion and are loaded into hi-performance file store
  - Database contains linkage to video
- Events are cataloged by algorithmic and manual processes
- Annotations are created and associated with events
- Annotations accumulate as research continues, thus the database continues to grow
- ~ 1 PB of video
- ~1.5 PB of database tables, indexes, and metadata

Data Reduction and Analysis



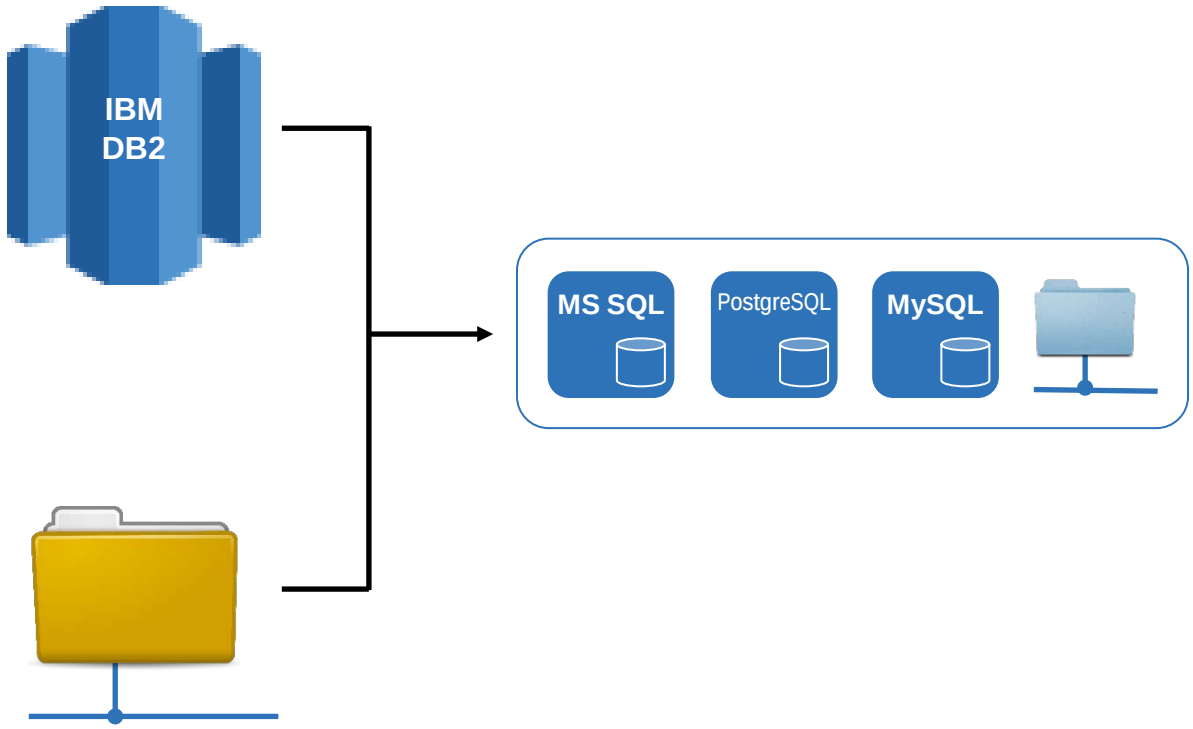
- “Hawkeye” application provides database query and synchronized video viewing, plus event annotation capabilities
- VTTI-developed application and supporting middleware
- Used by VTTI personnel as well as researchers whose DUL allows them to work in VTTI SDE

Data Analysis and Data Set Extraction



- Matlab is principal tool used by VTTI analysts to identify and extract data based on DUL “specs”
- Compute cluster provides parallelization of Matlab jobs
- Cluster also used for algorithmic DULs
- Significant set of reusable Matlab scripts have accumulated

Insight



- Web-facing repository of de-identified data and video files
- Periodically updated from primary sources
- ~ 100 GB of content, moving to cloud hosting
- Web application (not shown) is also part of the overall system

Challenge: Database Complexity and Scale

- While the DB2 database provides an essential framework for accessing the data, this structure and organization must be maintained when the data is exported
- Database mechanisms are also used to restrict access by DUL (Schema defines a “Shadow Collection”)
- Because of the database overhead involved, it could take as long as a year to export the data
- It could take 9 mo. to a year to import this data into a new database environment
- If DB2 is not the target database platform, a significant number of compatibility / functional equivalency issues must be addressed

Challenge: Data Movement

- It’s impractical to move 3.5 PB of data over a wide area network
- Physical movement of data would be accomplished by installing a new storage subsystem(s) in the VTTI data center, copying data to it via a high-speed local network connection, then shipping the storage subsystem to its final destination

Challenge: Intellectual Property

- A substantial set of software tools (applications, middleware, and scripts) are part of the overall SHRP2 NDS data “ecosystem”
- IP rights are unclear
- The data are of limited use without these tools or their equivalent

Summary

- Migrating the SHRP2 NDS data is not a simple task of copying data from point A to B
- The data span multiple platforms
- Data, tools, and processes are inextricably linked
- Any migration scenario will require professional services efforts and VTTI's involvement

Further Thoughts on Data Migration

March 2017

Migration Scope

- Migrating the NDS data to a new target environment would be a major undertaking
- To continue serving NDS data to researchers, the new environment must provide functional equivalence to the current VTTI-supported data ecosystem
- At a minimum this requires:
  - Replicating the current database and video file store
    - *Replicating the integrated functionality, not merely copying data*
  - Building software tools needed to support on-going data reduction and data set extraction
- If the DAS data will be maintained at VTTI on a long-term basis the archival storage would not need to be duplicated

Key Elements of the Target Environment

- Enterprise-class database software, servers, and related storage
- High-performance file store for video data
- High-performance compute cluster
- High-speed network backbone
- Could reside in an on-premises datacenter or in the cloud, which can be thought of as a virtual datacenter
- On-premises deployment implies hardware and software acquisition costs, on-going maintenance and support fees, and operational costs
- Cloud deployment is a pay-as-you-go model where the above costs are “baked in”
- May not be significant cost differences between these two deployment models when measured over time

Required Roles and Key Migration Tasks*

Role	Key Tasks
Systems architect	Define overall architecture of target environment, scope and spec of migration and development tasks
Project manager	Manage and track project execution
Database administrator	Create target database instances, schemas, manage database import/export
Database developer	Develop database queries, stored procedures
System administrator	Configure and manage compute, database, and file servers; manage file transfer activities
Network administrator	Support database and file transfer activities
Software architect	Develop architecture and specs for software tools
Software developer	Programming

*Some of these roles/tasks will be required from VTTI as well as from the target organization

Database Migration

- Most complex, and therefore most costly, aspect of any migration
- Simplest path is a DB2-to-DB2 “copy”
  - DB2 software licensing costs alone are approximately $1M
  - Still requires substantial time from highly skilled DBAs on both sides
  - Exporting/importing a petabyte of data presents numerous challenges and will take a significant length of time
- It’s possible to migrate from DB2 to another database technology
  - Target database would have to support various enterprise-class features
  - Licensing costs depend on target
  - Additional professional services costs to deal with data conversion issues, re-write stored procedures, etc.

Cloud Hosting is Not a Panacea

- Cloud services are most cost-effective for on-demand compute capacity and infrequently accessed storage
- NDS data serving requires dedicated compute, network, and storage resources
- Difficult to estimate compute costs without in-depth analysis
- But sizing storage costs is very straightforward:

Storage Type	Example Commercial Service	Price	Monthly Cost
Archive (1 PB)	Amazon Glacier	$0.004 per GB-month	$40,000
Files (1.5 PB)	Amazon Elastic File Store (EFS)	$0.30/GB-month	$450,000
Database (1 PB)	Amazon Elastic Block Store (EBS)	$0.10 per GB-month	$100,000
			$554,000

Concluding Thoughts

- Not feasible to do a bottom-up estimate of cost without:
  - Knowing exact details of the target environment
  - An extensive scoping effort
- Enough is known about the broad requirements and the diversity of skills/tasks to be accomplished to say that an absolute minimum cost estimate would be in the range of $2M