LSU

E. J. OURSO COLLEGE OF BUSINESS

Moving to a Data Warehouse

THE HIGHWAY SAFETY RESEARCH GROUP

What is the Highway Safety Research Group (HSRG)?

- A division of the Information Systems and Decision Sciences Department (ISDS) within the E. J. Ourso College of Business at Louisiana State University

- Website:

  http://hsrg.lsu.edu

What is the Highway Safety Research Group (HSRG)?

- Grant funded by the LA DOTD
- Responsible for collecting, maintaining, storing, and reporting crash data captured from law enforcement agencies throughout the state of Louisiana
- Analyzing crash data for LA since 1994

Collecting Data

- Have 180+ law enforcement agencies using our LACRASH software
- Collect electronic crash reports from 3rd party vendors using xml and ftp processes
- Receive paper crash reports and manually enter data

Maintaining Data

- Create yearly crash databases
- Offer back-up services
- Provide real-time fail over services

- Manually review reports for data quality
  - completeness and accuracy

Storing Data

- Data stored in SQL databases
  - Normalized
    - Organize fields and tables to minimize redundancy and dependency
    - Divide large tables into less redundant tables and define relationships between them

Reporting Data

- Crashes
  - – Aggressive driving
  - – Alcohol
  - – CMV
  - – Fatal
  - – Occupant protection
  - – Young drivers

Reporting Data

- Crashes
  - Driver characteristics
  - Roadway characteristics
  - Vehicle types
  - Weather conditions
  - When
  - Where

http://datareports.lsu.edu

ON-LINE TRANSACTIONAL PROCESSING (OLTP) SYSTEM

OLTP System at HSRG

- Capture and store data based on transactions of business process
  - Transaction = crash
- LA averages about 150,000 crashes a year

OLTP System at HSRG

- Normalized data
- Stored in yearly databases

Data at HSRG

- Tables in yearly database
  - Crash
    - (Crash_Num)
  - Vehicle
    - (Crash_Num, Veh_Num)
  - Occupant
    - (Crash_Num, Veh_Num, Occ_Num)
  - Pedestrian
    - (Crash_Num, Ped_Num)

Crash Example

- Crash occurs involving two cars:
  - Car 1
    - Driver
    - Occupant
  - Car 2
    - Driver
    - Occupant
    - Occupant
- How do we determine if the crash was a fatal crash?

How do we determine is the crash was a fatal crash?

- Join vehicle and occupant table
  – Evaluate injury for each person in 1st vehicle
    - Driver and occupant
  – Evaluate injury for each person in 2nd vehicle
    - Driver and 2 occupants

  – If any person was killed, the crash was a fatal crash

How is this calculation performed?

- Ad-hoc
  - – When needed
- Stored Procedure
  - – Scheduled process on new records

Ad-hoc

- Write SQL Statement
- Do all employees know correct SQL statement?
- Processing time
  - – Joining tables
  - – Same SQL statements executed multiple times to receive same data

Ad-hoc

- Write SQL Statement

```
select VEHIC_TB.CRASH_NUM
From VEHIC_TB, OCCUP_TB
Where VEHIC_TB.CRASH_NUM =
OCCUP_TB.CRASH_NUM
and VEHIC_TB.VEH_NUM = OCCUP_TB.VEH_NUM
and (VEHIC_TB.DR_INJ_CD = 'A'
or OCCUP_TB.OCC_INJ_CD = 'A')
```

Ad-hoc

- Do all employees know correct SQL statement?

- Processing time
  - Joining tables
  - Same SQL statements executed multiple times to receive same data

Stored Procedures

- Create computed field
  - – Fatal_Crash within Crash Table
- Create stored procedure to evaluate crash and update new field (Y/N)

- Efficient?
  - – Injury code changes
    - People can pass away days after crash

Ad-hoc and Stored Procedures

- Multiple processes
  - – Crash severity
  - – # people killed, # people injured
  - – Aggressive driving crash
  - – Alcohol crash
  - – CMV crash
  - – Young driver crash
  - – Etc…

Roadway Departure Definition

Prior_Movement_Cd IN ('E', 'G')

OR F_Harm_Ev_Cd In ('a','j','k','l''s','x','z','aa','bb','cc','dd','ee','ff','gg',
'hh','ii','jj','kk','ll','mm','nn','oo','pp','qq')

OR S_Harm_Ev_Cd In ('a','j','k','l''s','x','z','aa','bb','cc','dd','ee','ff','gg',
'hh','ii','jj','kk','ll','mm','nn','oo','pp','qq')

OR T_Harm_Ev_Cd In ('a','j','k','l''s','x','z','aa','bb','cc','dd','ee','ff','gg',
'hh','ii','jj','kk','ll','mm','nn','oo','pp','qq')

OR FO_Harm_Ev_Cd In ('a','j','k','l''s','x','z','aa','bb','cc','dd','ee','ff','gg',
'hh','ii','jj','kk','ll','mm','nn','oo','pp','qq')

OR M_Harm_Ev_Cd In ('a','j','k','l''s','x','z','aa','bb','cc','dd','ee','ff','gg',
'hh','ii','jj','kk','ll','mm','nn','oo','pp','qq'))

Ad-hoc and Stored Procedures

- Dynamic
  - Definition changes
  - Where is definition used
    - Have to know all reports to change
- Flexible
  - Add new process
  - Need age range 16 – 20, instead of 16-24

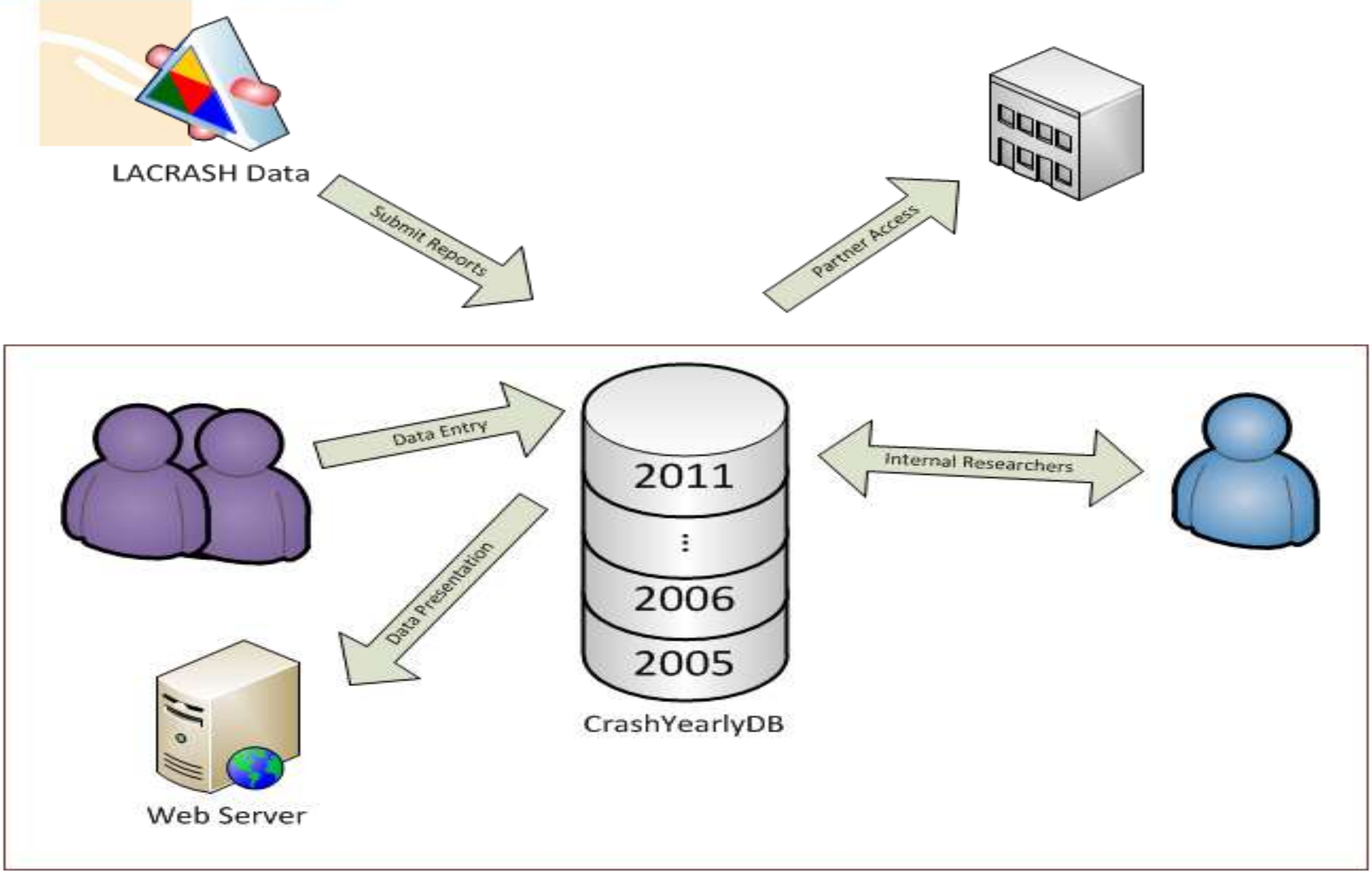
OLTP System at HSRG

- Works great for collecting, storing, and maintaining data
- However, it is not as efficient for reporting and analyzing data

Ad-hoc and Stored Procedures

- Basically, we are trying to pre-calculate aggregate values for reporting purposes.
  - Number of fatal crashes
  - Number of injury crashes
  - Number of fatalities
  - Number of injuries

Pre BI Database Overview



Challenges

- – Shift focus from data delivery to data analytics
- – Provide information to decision makers in a timely manner
- – Separate transactional and reporting operations
- – Provide single version of the “truth”
- – Leverage new technology and provide platform standardization in-line with our current competencies

How to move forward?

- In 2010, we began looking into Business Intelligence

BUSINESS INTELLIGENCE DEFINED

Business

- Encompasses all of the traditional functional activities in business:
  - Examples: marketing, manufacturing, accounting, finance, distribution, and support operations
- Provided by transactional processing systems and other basic technology

Intelligence

- Includes all mathematical and statistical tools developed to solve business “problems”
  - Examples: applied mathematics, statistical quality control, and operations research

- While business flow concentrates on **efficiency,** intelligence focuses on **effectiveness**

What is Business Intelligence (BI)?

- Broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions

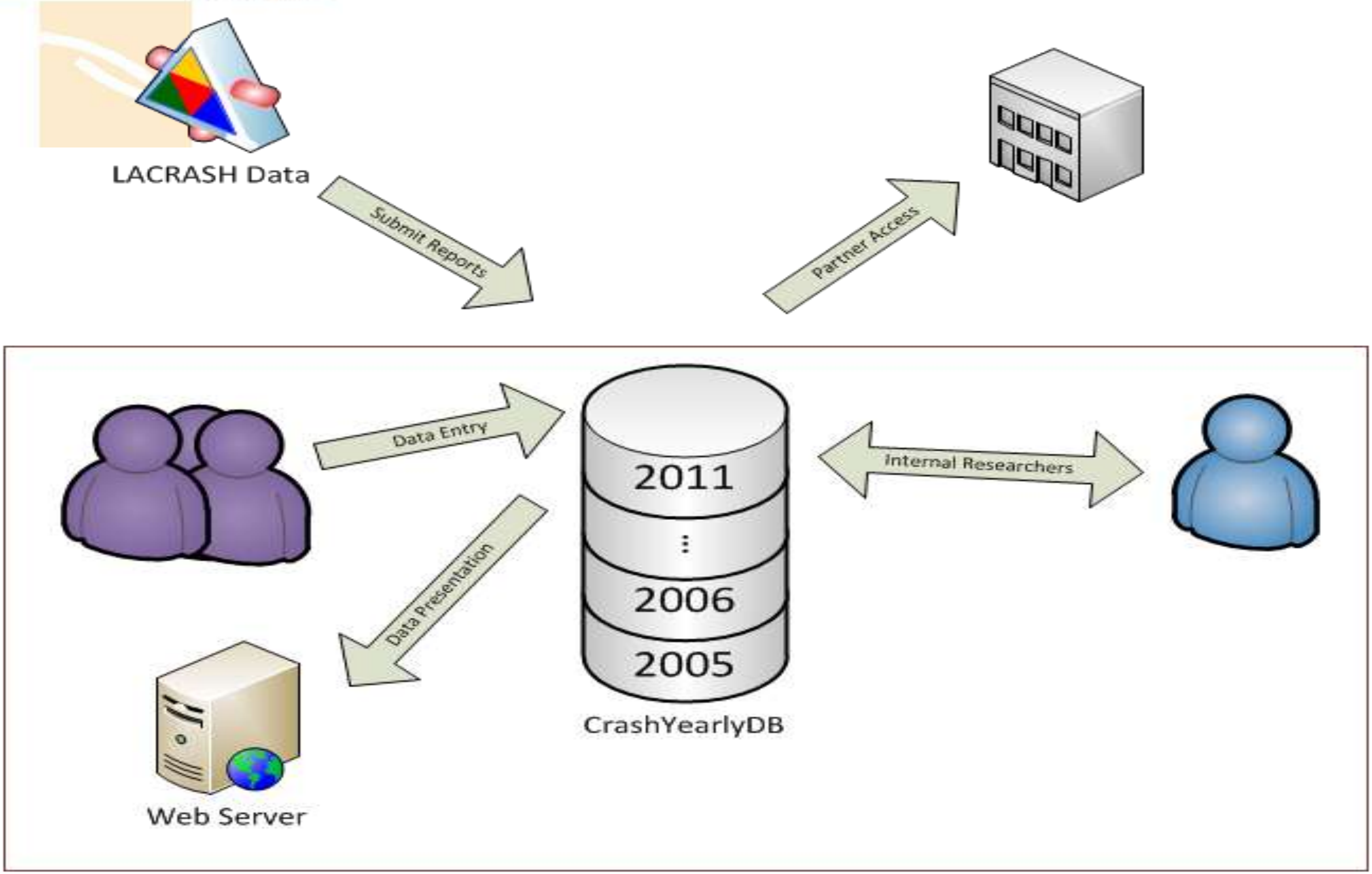
- Process of transforming data into information and making it available to users in a timely manner to make effective decisions

ON-LINE ANALYTIC PROCESSING (OLAP) SYSTEM

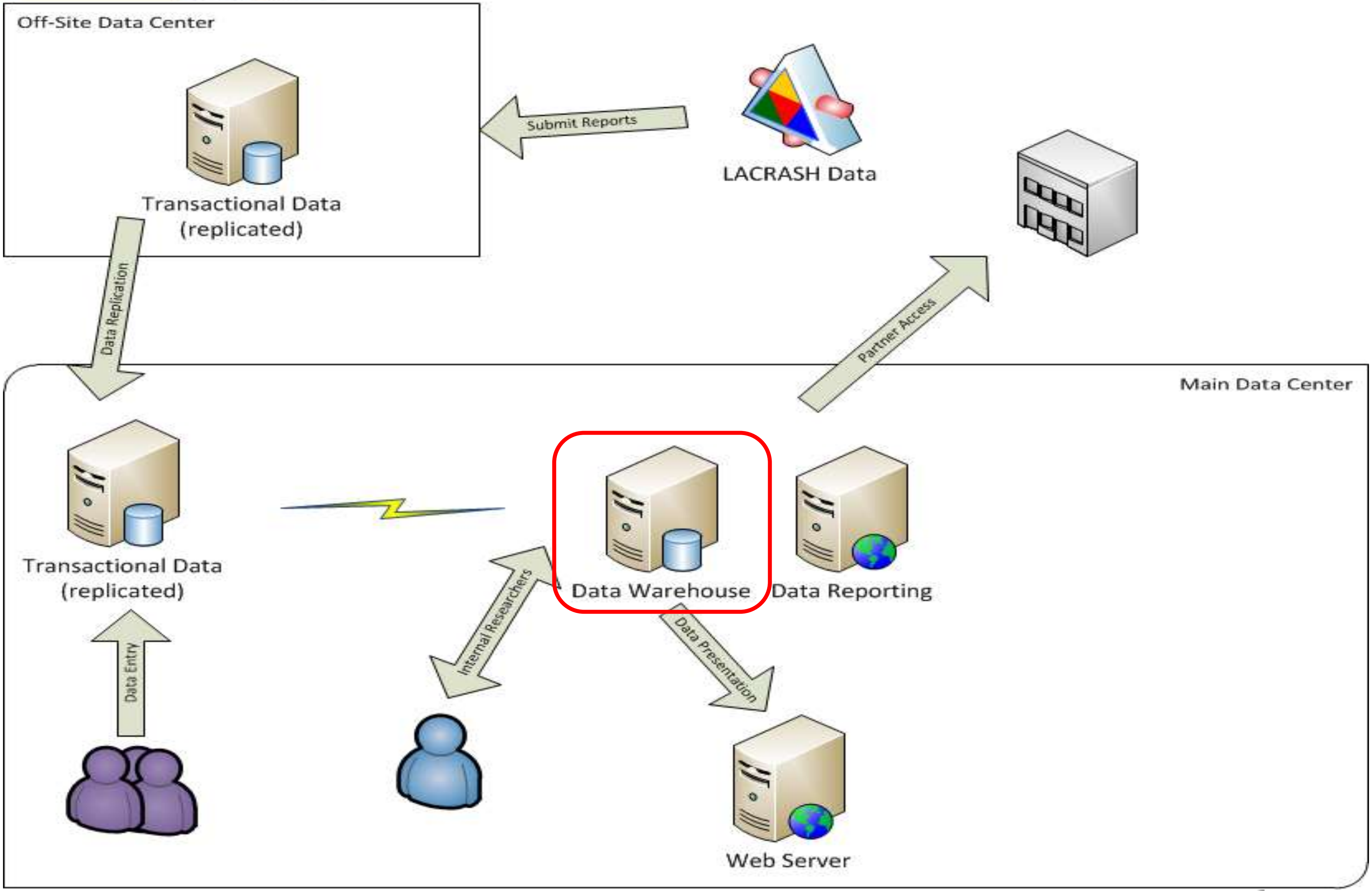
Data Warehouse

- Relational database used for reporting and analysis
- Stored in star or snowflake schema
- Contains cleaned and transformed data made available for use by managers and other business professionals

Pre BI Database Overview



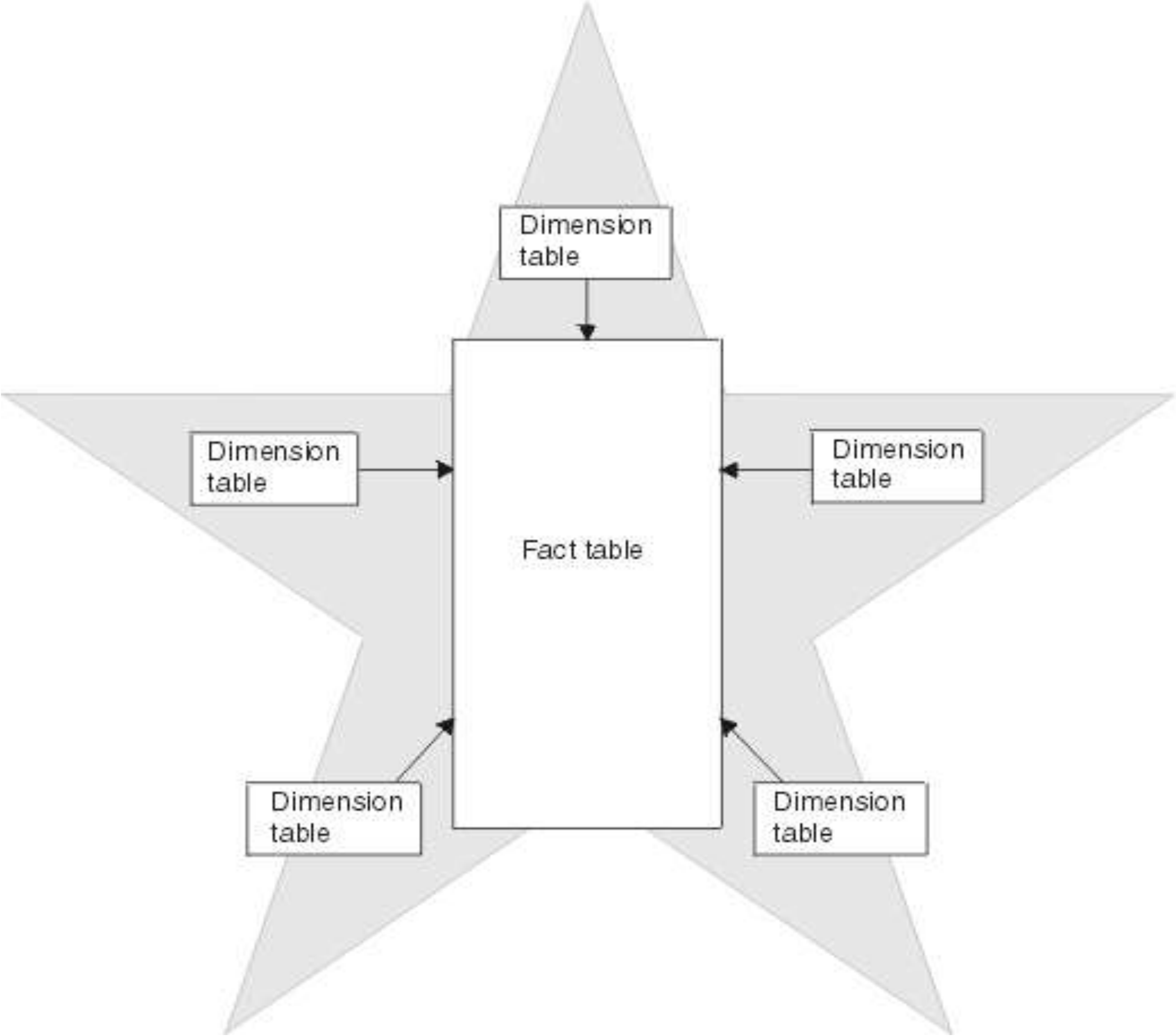
Post BI Database Overview



Data Warehouse

- Build with decision in mind
  - Automate repeated decision
    - Crashes
      - Severity
      - Type
      - When
      - Where
    - Driver
      - Age
      - Race
      - Sex

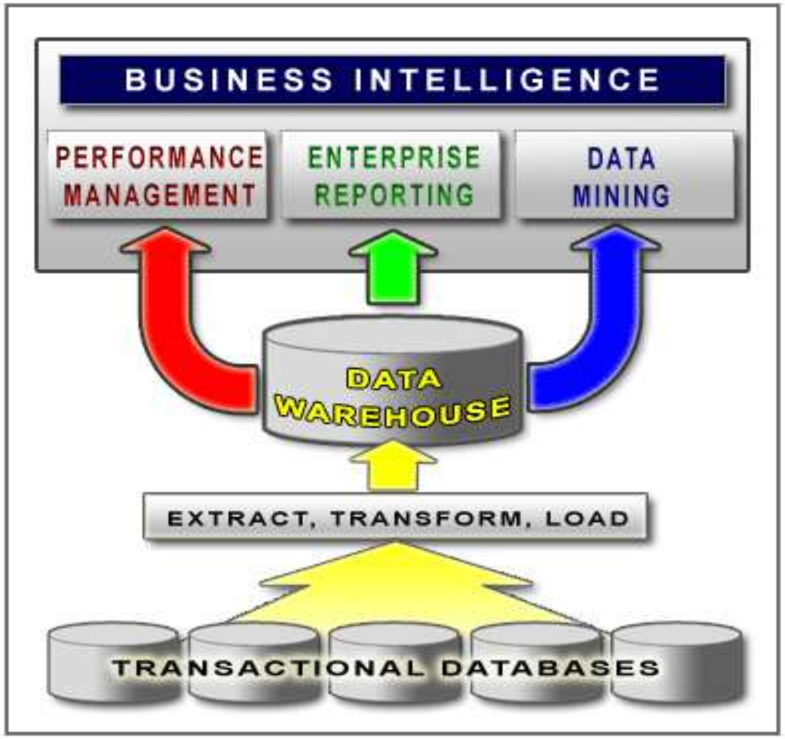
Star Schema


Dimension table
Dimension table
Dimension table
Fact table
Dimension table
Dimension table

Star Schema

- Fact
  - What do we want to measure
    - Driver
- Dimension
  - How to we want to ‘slice and dice’ the measure
    - Age
    - Race
    - Sex

BI Using Microsoft SQL 2008R2



Extract, Transform, Load (ETL)

- Extract data from OLTP system
  - – Normalized
- Transform the data
  - – Data quality
  - – Calculations (severity, cmv, alcohol)
- Load the data into data warehouse
  - – Star or snowflake schema

Extract, Transform, Load (ETL)

- Now, there is ONE place that contains all the definitions
  - Standardized
  - Easy to maintain
  - Flexible
  - Dynamic
  - Efficient
    - Can drop and reload DW from 2005 – present in less than 20 minutes (over 10 million records)
    - Perform on weekly basis

ETL and DW

- Most time is spent designing the DW, writing the ETL, and then cleaning & validating the process
- Once the DW is created, loaded, and validated, cubes can be built

What is a cube?

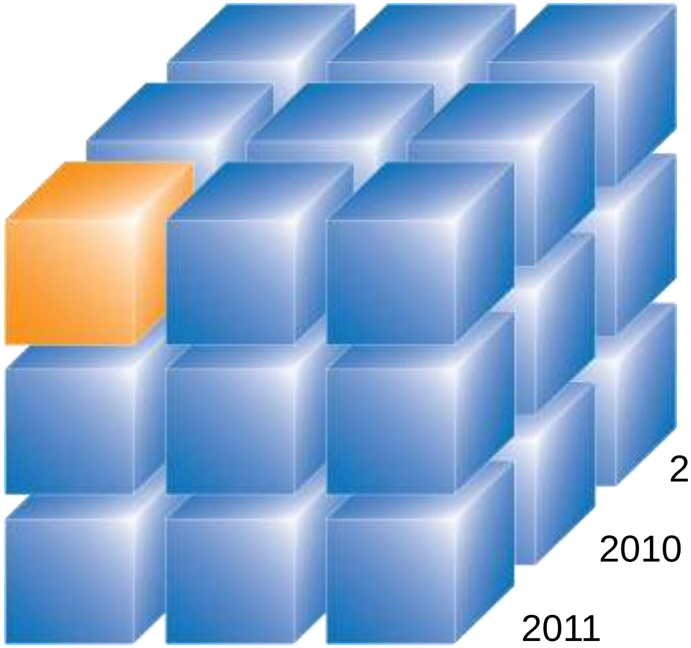
- *A* multidimensional dataset that can have an arbitrary number of dimensions
- Each cell of the cube holds a number that represents some *measure* of the business process

Cube Example

- *Fact*
  - – *Number of crashes*
- *Dimensions*
  - – Where (Parish)
  - – Severity (Fatal, Injury, PDO)
  - – When (Year)

Cube Structure

Number of Fatal crashes in Acadia parish in 2011



Acadia

Baton Rouge

Caddo

Fatal

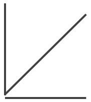
Injury

PDO

2009

2010

2011



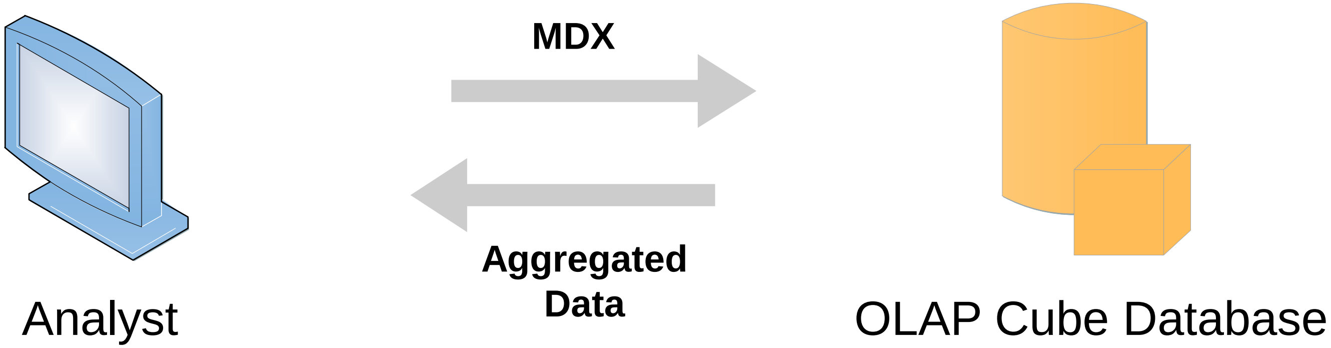
Where

When

Severity

Cubes

- OLAP databases are called ‘Cubes’
- The **Multi-Dimensional** Expression (MDX) language accesses cube data



Browsing a Cube

- BIDS
- Web
  - – http://datareportsdev.lsu.edu/
- Analysis Services Database

Reporting from a Cube

- Web
  - http://datareports.lsu.edu/
  - http://lashspdata.lsu.edu/#/Home

Next Steps

- Data Mining
- Forecasting
- Fraud Detection

Contact Information

- Cory Hutchinson
  - Associate Director
  - cory@lsu.edu
  - (225) 578-1433