Dat5 Presentation

Group d521a

Aalborg University

1^{st} December 2008

Presentation Outline

- Introduction
- Related Work
- Data Warehouse Design
- Extract-Transform-Load (ETL)
- Demonstration
- Conclusion

Introduction

Project Introduction

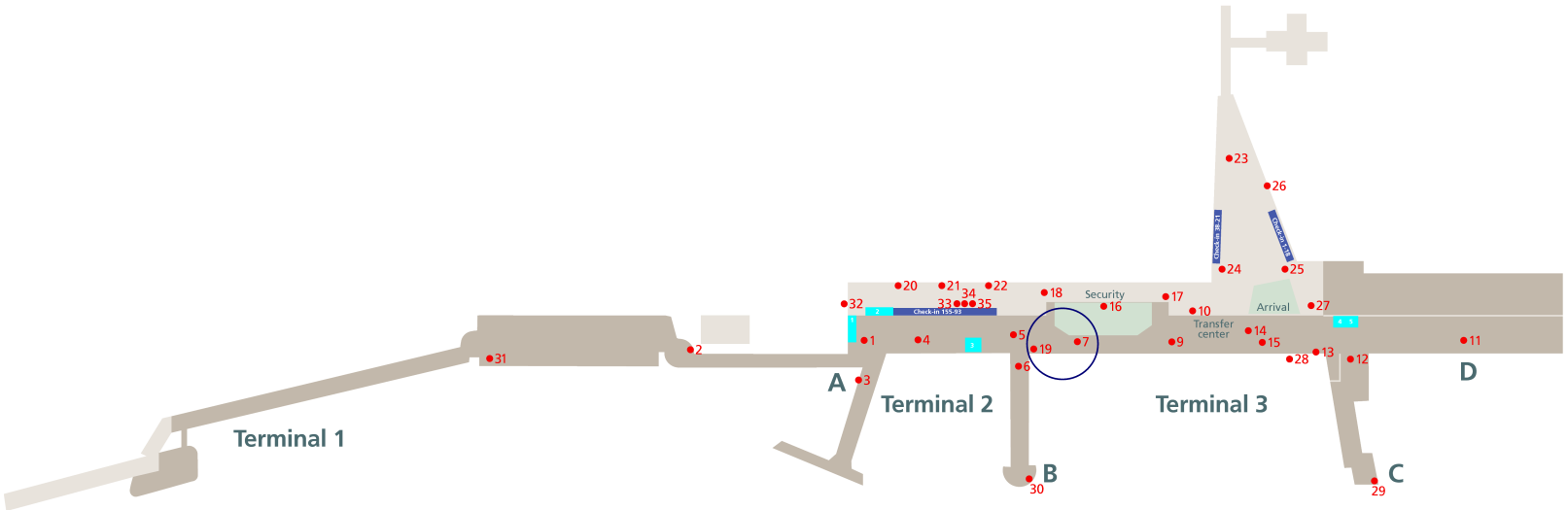
- Project done in collaboration with BLIP Systems A/S.
  - BLIP was founded in 2003 by former Ericsson employees.
  - Area of expertise includes LBS, Bluetooth marketing, Bluetooth networks in general, consultancy etc.
- Project is about tracking people in an airport.
  - Tracking data collected using Bluetooth access points.
  - Only devices with active Bluetooth modules are registered.
  - Enough people with Bluetooth devices to provide useful results.
- Involves modeling a multidimensional data warehouse.
  - Performing analysis on the data using OLAP tools.
  - Representing results in a meaningful and useful manner.

Motivation

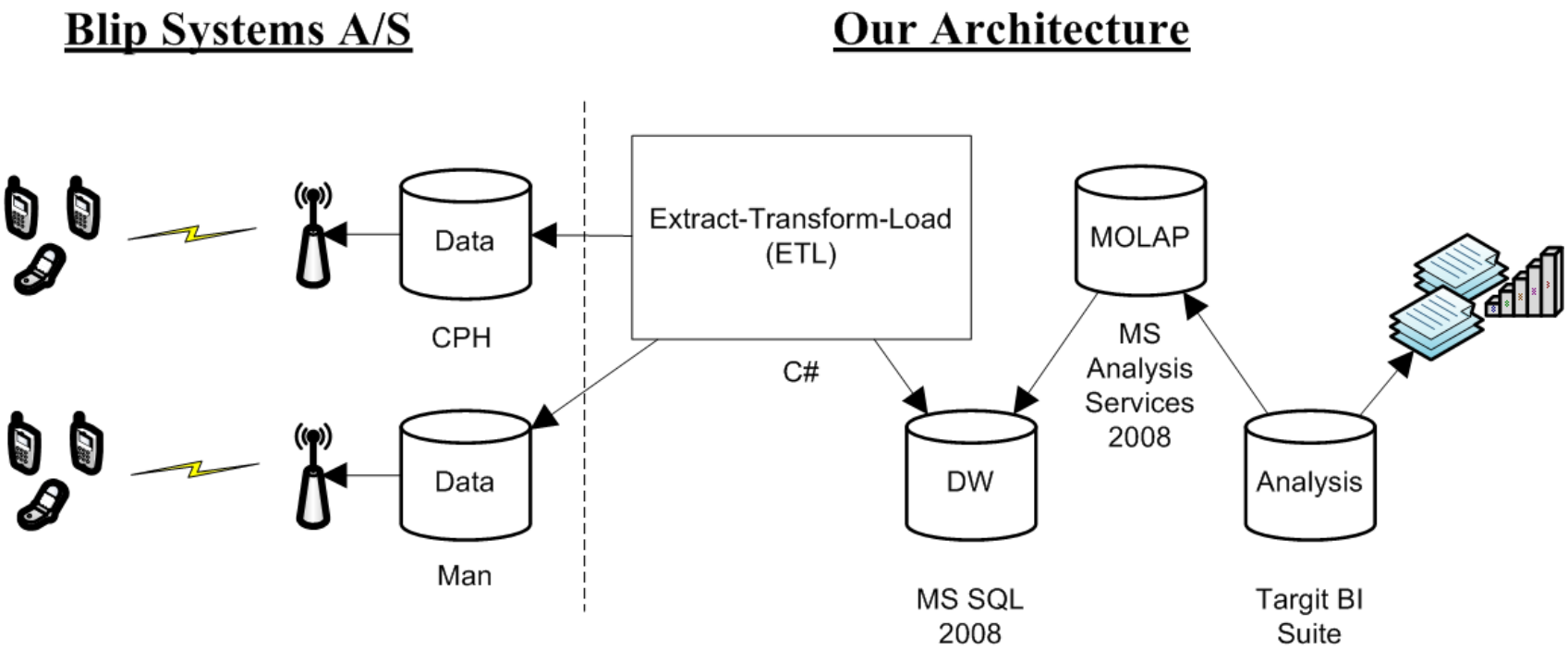
- What information can be discovered by tracking people in the airport?
  - What are the queue times for check-in, security and boarding?
  - How many users are there of the airport?
  - How many are local, transit and frequent flyers?
  - How many people follow the *forced* routes?
- Manual tracking disadvantages.
  - Tracking of up to 200 people per day using video surveillance.
  - Tracking criteria must be very specific due to the size of the area.
    - Must keep eyes on the subject.
- Push relevant information to the phones.

Physical Setup

- 26 access points tracking Bluetooth devices.
- More added in areas needing extra attention.



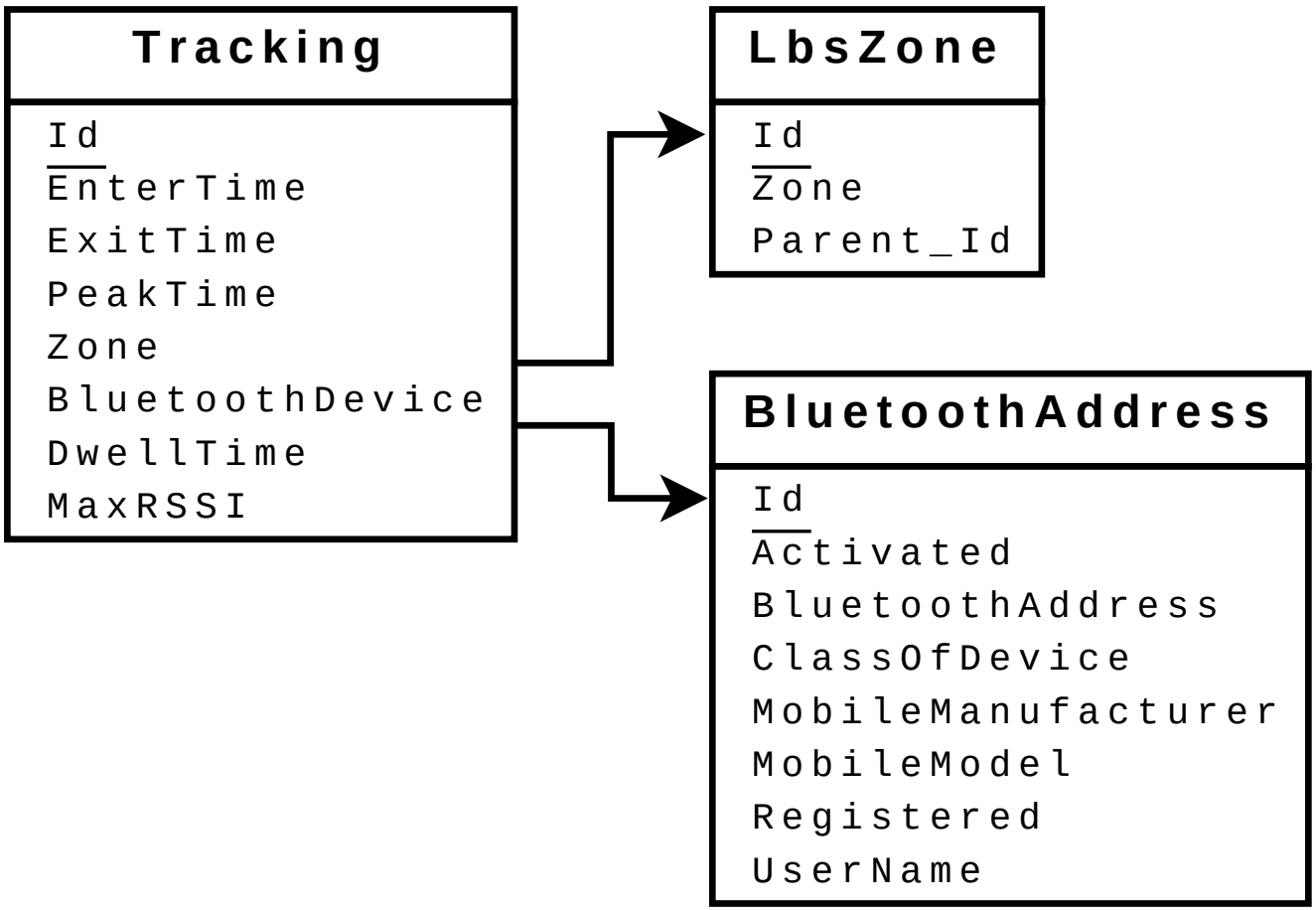
System Architecture



Data Collection

1. Server queries each access point for Bluetooth devices once per second.
2. When a device comes into the vicinity of an access point - the server creates an *open* record in a database, containing a timestamp (*enter time*) as well as zone and device information.
3. The server monitors the device, keeping track of the value and timestamp of the strongest signal as well as time last seen.
4. When the device enters another zone, or 1 minute has passed since the device was last seen, the server closes the record.
    - Note: A device can only be in 1 zone at any given time!

Data Format


Tracking
Id
EnterTime
ExitTime
PeakTime
Zone
BluetoothDevice
DwellTime
MaxRSSI
LbsZone
Id
Zone
Parent_Id
BluetoothAddress
Id
Activated
BluetoothAddress
ClassOfDevice
MobileManufacturer
MobileModel
Registered
UserName

Quantitative Measurements

Copenhagen Dataset

- Up to 6.500 unique passengers per day.
- Up to 500.000 tracking records per day.
- Data collected from 26 access points in the airport.
- 200 people tracked per day, using manual video surveillance.

Related Work

Temporary Mobile Subscriber Identity (TMSI)

- Path Intelligence Ltd. has develop a system tracking mobile phones using TMSI.
- Advantages.
  - Long range.
  - High precision - 1-2 meters accuracy.
  - High penetration - powered on mobile phones.
- Disadvantages.
  - Infrequent updates - minutes.

Radio Frequency Identification (RFID)

- IT University of Copenhagen and Lyngsoe Systems have developed a system that tracks users using RFID tags.
- Advantages/disadvantages.
  - Penetration dependent on whether the tags are handed out, attached to boarding cards, incorporated into baggage trolleys etc.
  - Limited maneuverability if tags are fixed onto equipment.
  - Complete penetration if tags are attached to boarding cards.
    - Airlines can tell if their passengers can make the flight and act accordingly.

Overview

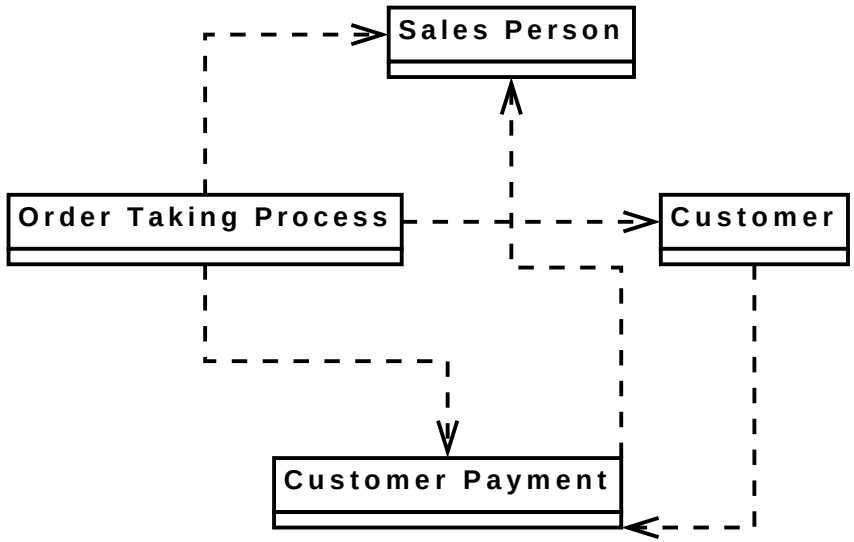
- Approach
- Modeling
- Dimensional Modeling
- Online Analytical Processing(OLAP)
- Dimension Design

Developing the Data Warehouse Design

- Limited knowledge.
- Iterative approach(3rd version).
- Simplest design.
- Problems occurred
  - Smart keys
  - Recurrence in zones.

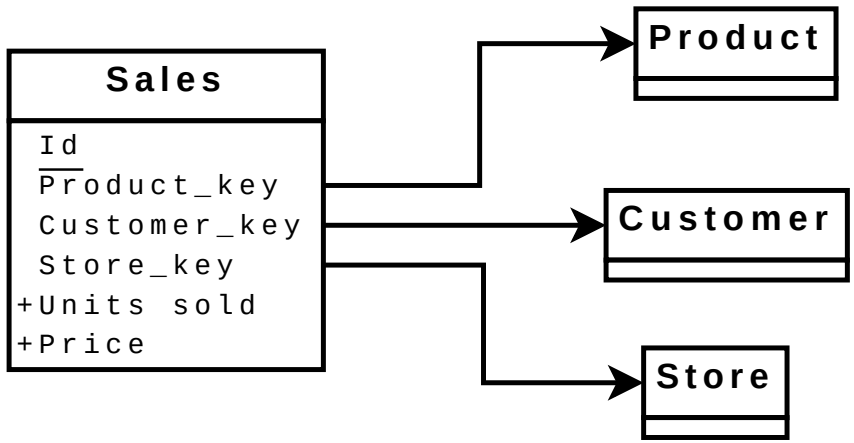
Modeling - ER Modeling

- Relationship based.
- Ultimate goal is to remove redundancy.



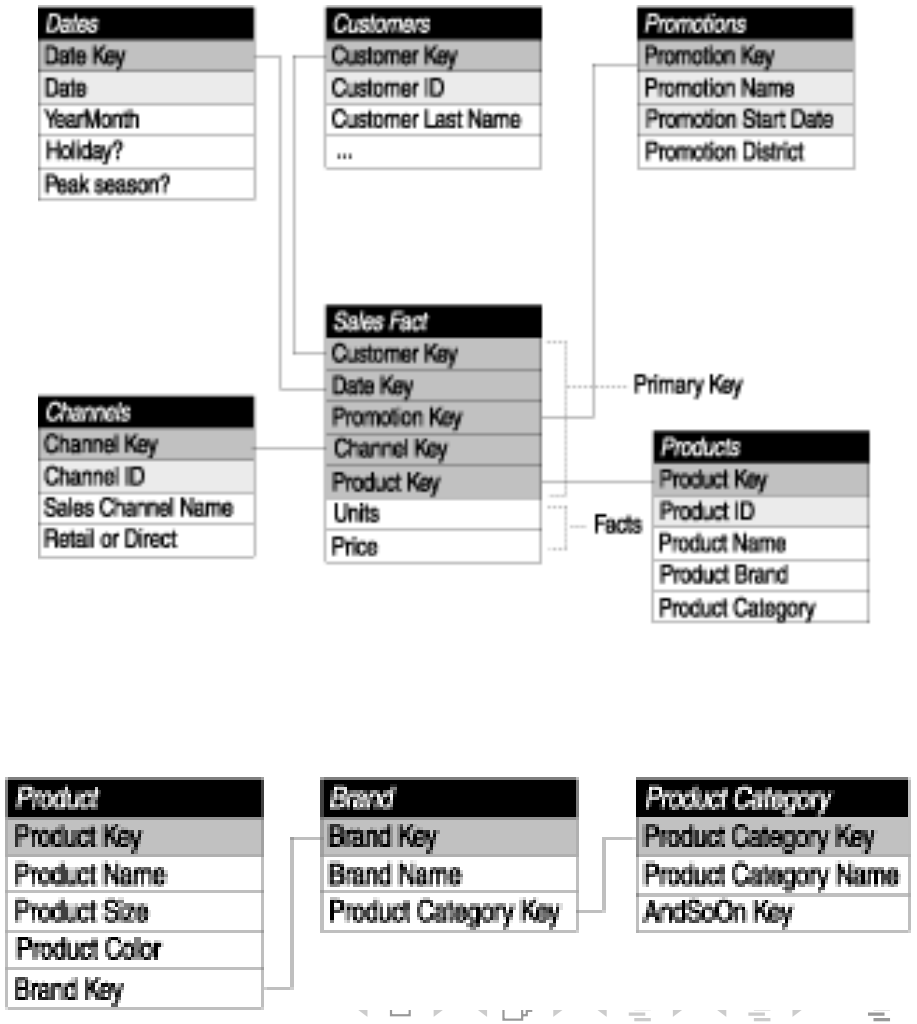
Modeling - Dimensional Modeling

- Fact based.
- Facts are numeric and additive.
- Textual information.

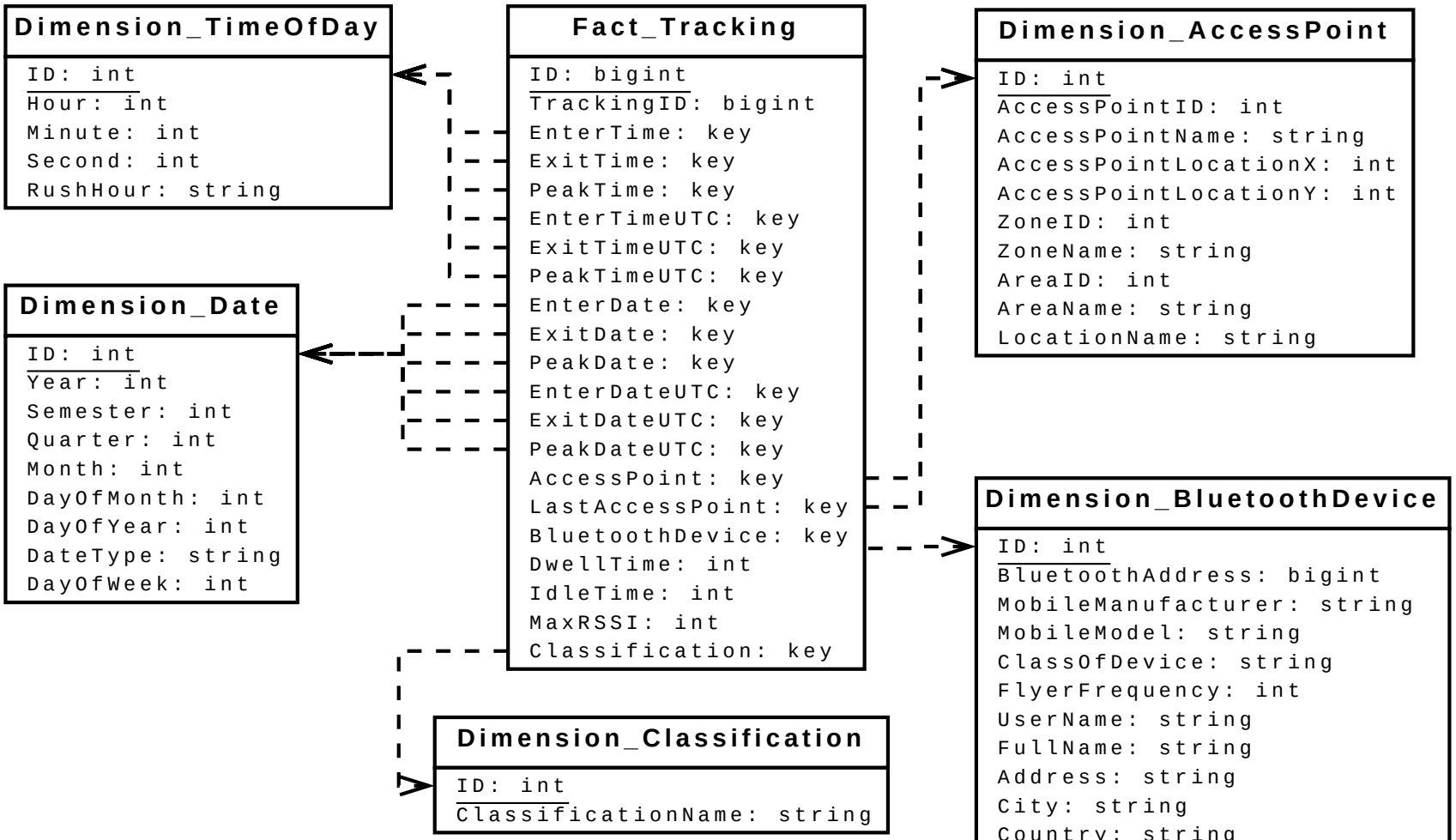


Dimensional Modeling - Star scheme vs Snowflake scheme

- Star scheme.
  - Simplicity.
  - Redundancy.
  - Single level joins.
- Snow flake.
  - Multi level star model.
  - Some degree of normalization.
  - Multi level joins.

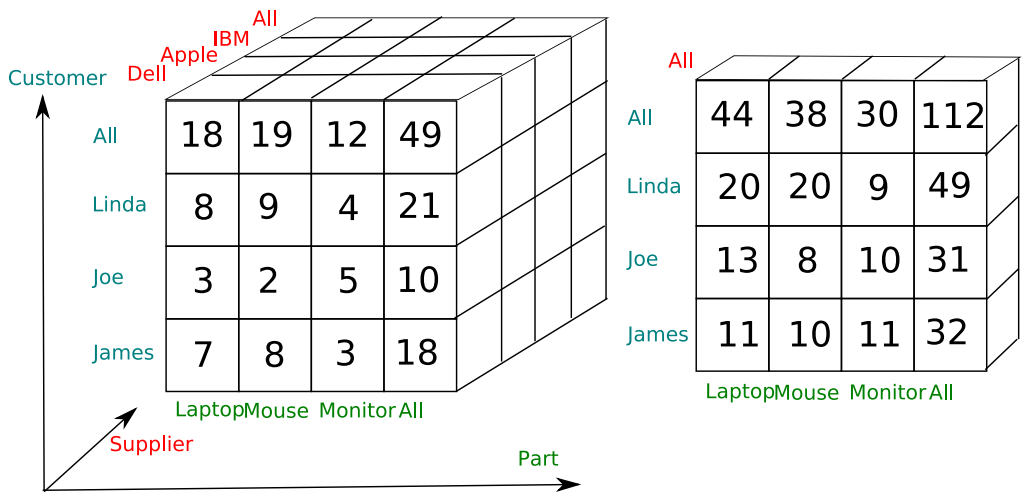


Dimensional Modeling - Our Design



Online Analytical Processing(OLAP)

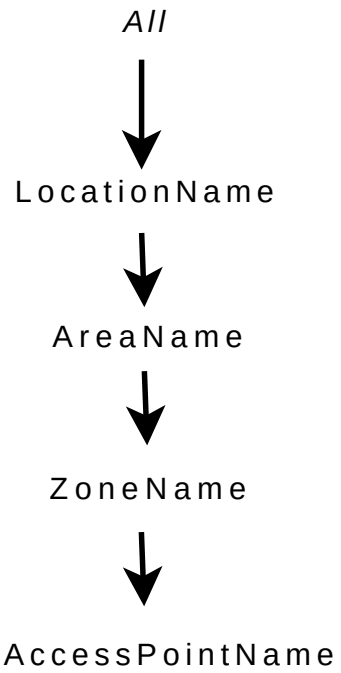
- OLAP cube.
  - Facts, called measures, derived from the records in the fact table.
  - Categorized by dimensions, derived from the dimension tables.



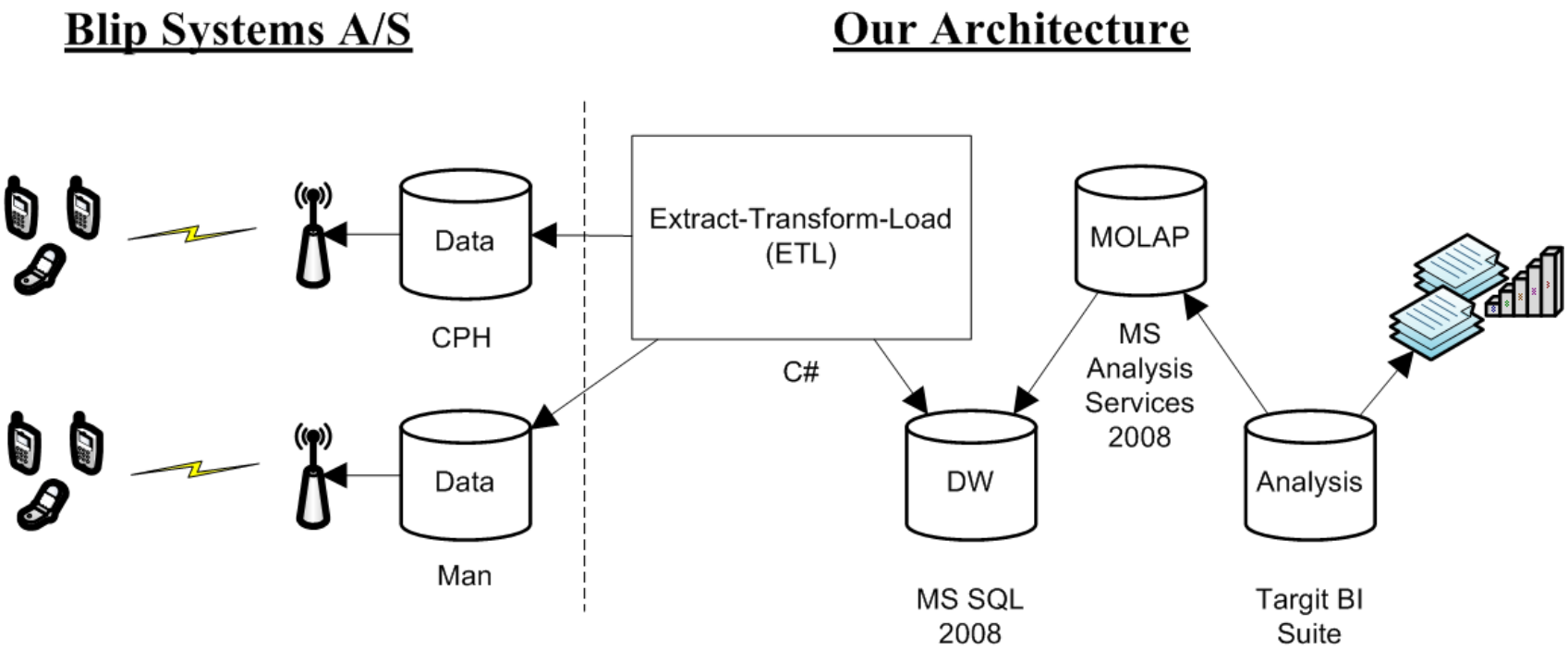
OLAP Taxonomy

- MOLAP
  - High query performance through Aggregations.
  - Introduces data redundancy.
- ROLAP
  - Good at handling non-aggregatable facts.
  - Slower at performing queries.
- HOLAP
  - Hybrid model of MOLAP and ROLAP.

Access Point Hierarchy



System Architecture



Extract-Transform-Load(ETL)

Extract-Transform-Load(ETL)

- The steps of our ETL
- Data Cleansing and Classification
- Bounce detection
- Frequent flyer

Extract-Transform-Load(ETL)

- Extractor
  1. Extract Min and Max timestamp
     - Pre-Populate Time and Date dimension in our DW
  2. Extract users
  3. Extract zones
  4. Extract tracking records
     - Data Cleansing and Classification
     - Discarding incomplete records
- Transformer
  1. Calculate DwellTime and IdleTime
  2. Data Cleansing and Classification
  3. Frequent Flyer Calculation
- Loader

Data Cleansing and Classification

- Types of source data issues.
  - Incomplete tracking records.
  - Exit timestamp $<$ Enter timestamp.
  - Enter timestamp $>$ Peak timestamp.
  - Peak timestamp $>$ Exit timestamp.
  - No Frequenct flyer attribute.
  - Bouncing problems (introduced later).
- The Loader is loading partitions of the source data.
  - No states.
  - Stream of data.

Bouncing

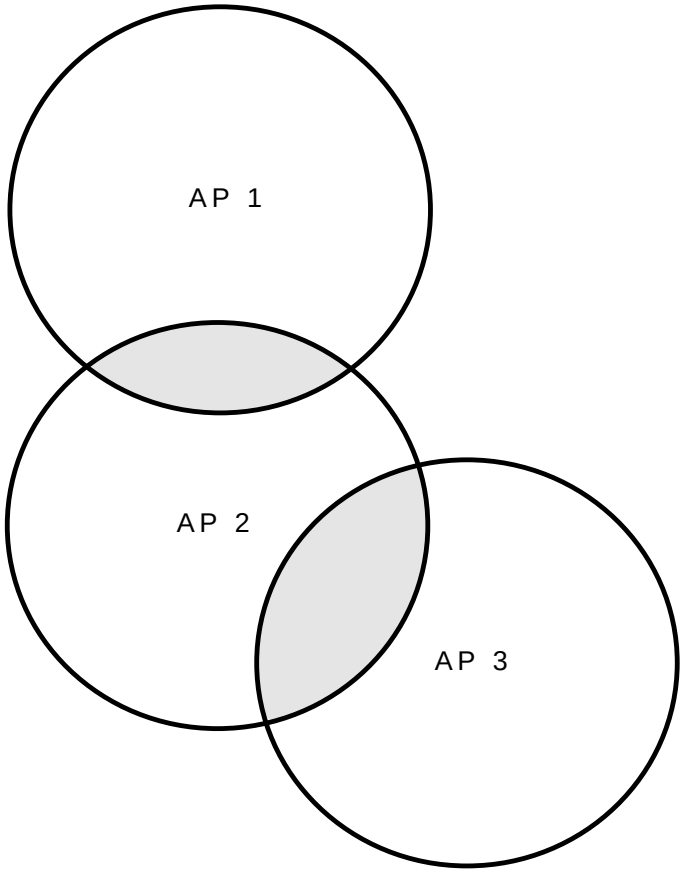


Figure: BT Device traversing AP 1, AP 2, and AP 3.

Bounce Detection

- Identify the timespan in which a BT device bounces.
- Identify which access points it bounces between.
- Split the total bounce time between the access points.
  - Weighted split.
  - Split in the same order as the access points are seen.
  - No collapsing of bouncing slices and earlier tracking records.
  - Amount of bouncing records equal to the amount of access points.

Bouncing Detection

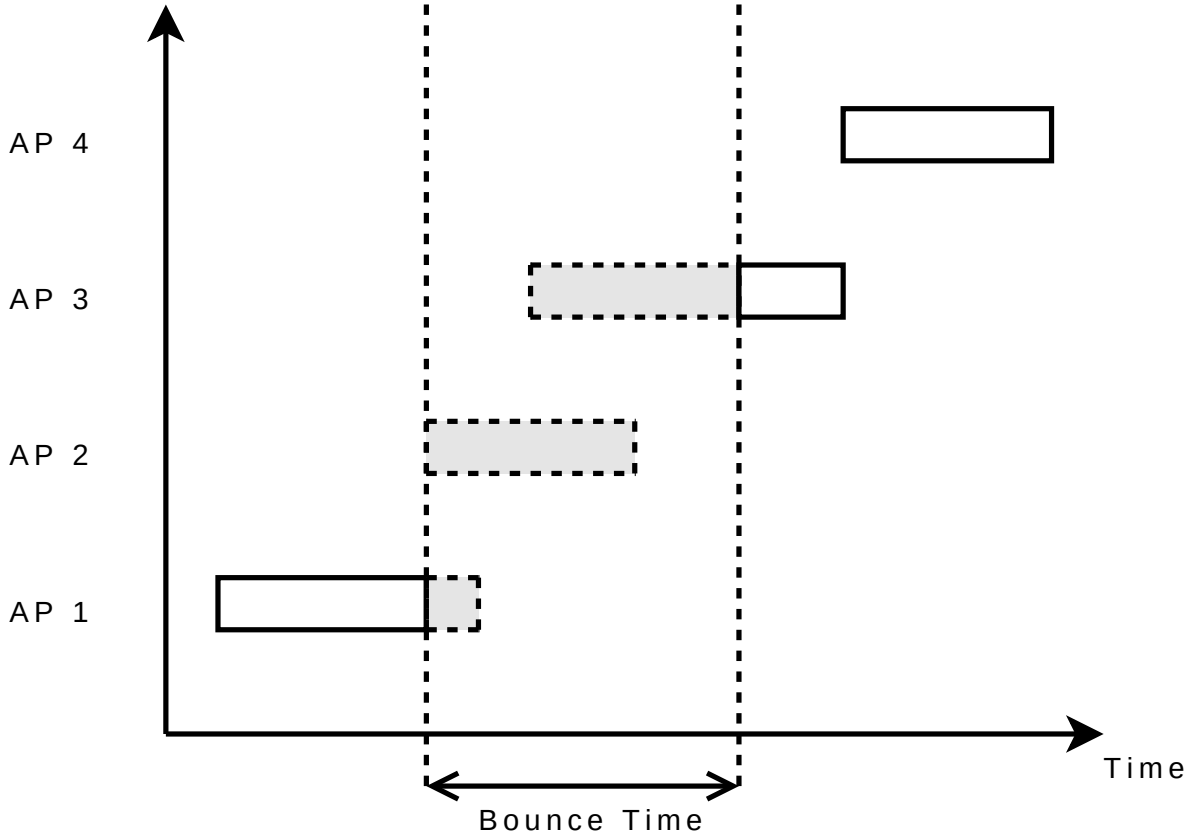


Figure: BT Device bouncing between AP 1, AP 2, and AP 3.

Bounce Detection Improvements

- Bounce Threshold set to 20 seconds.
- Source data consist of approximately 22 mil records.
- Bounce detection eliminated approximately 6 mil records.
- Bounce detection classify approximately 8 mil records as "Bounced".
- Analyze on all records of a given BT device.
  - Better bounce detection.
  - Possibility to collapse bounces with earlier and later tracking records.
  - Better classification.

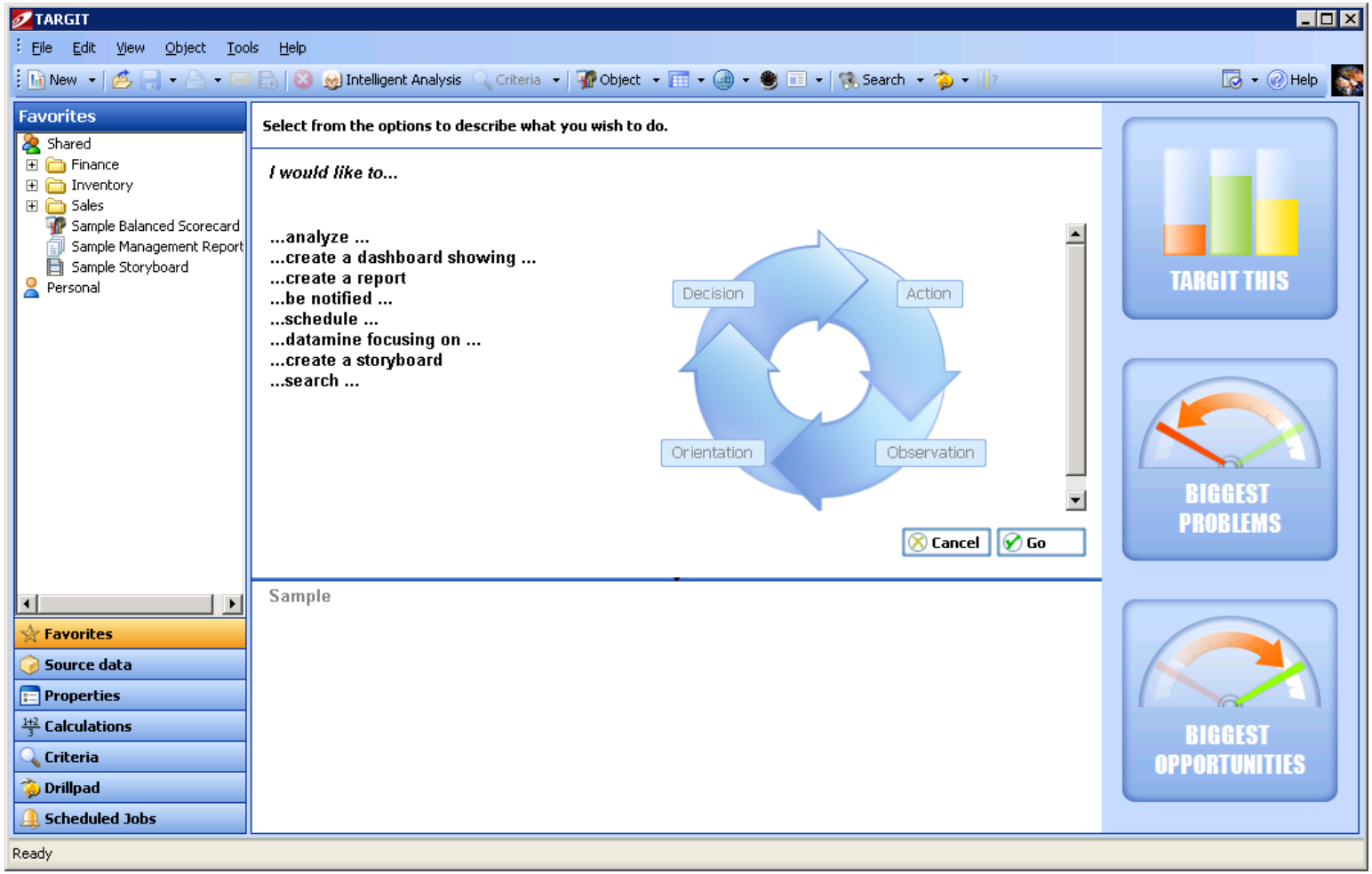
Frequent Flyer Calculation

- Frequent Flyer $\longrightarrow$ Flyer frequency count
- Updated when loading the date from source to DW.
- FrequentFlyerThreshold set to 43200sec (12Hours).

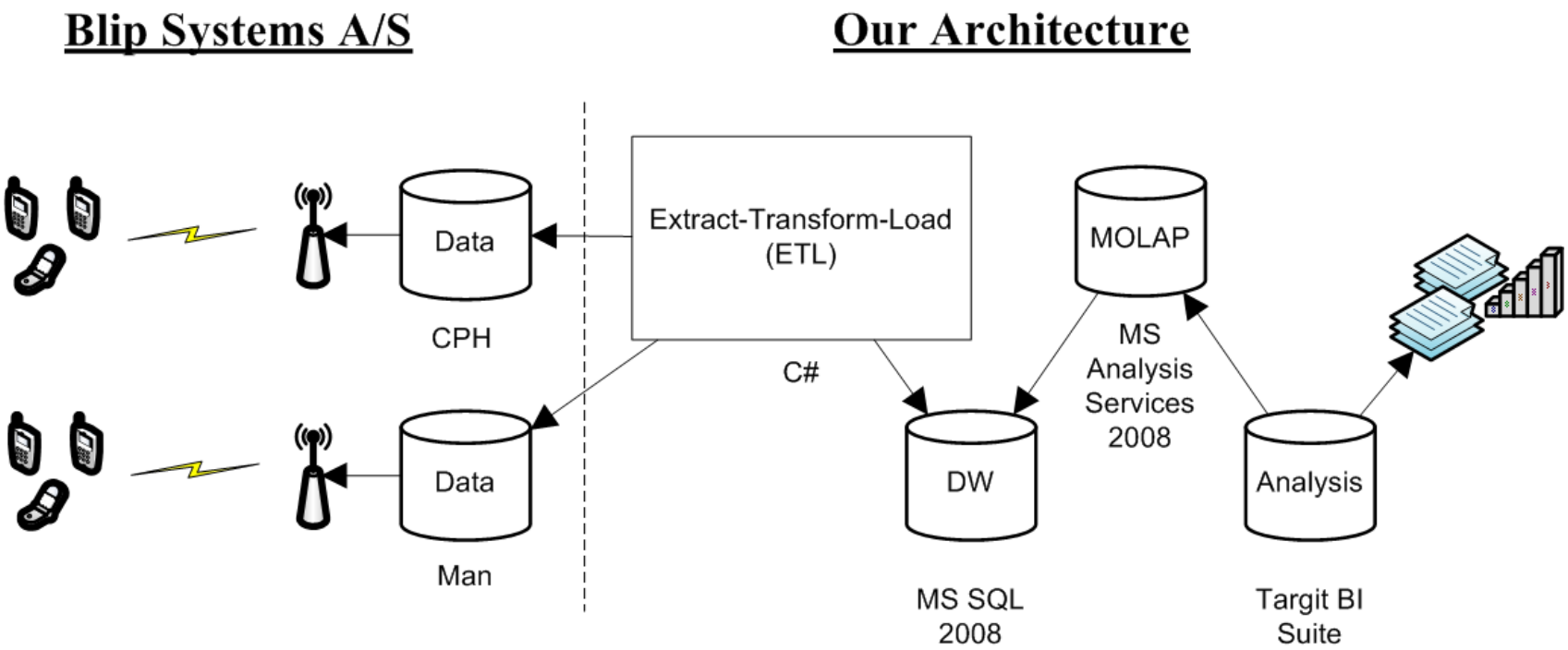
Frequent Flyer

$$BTDevice_{EnterTime} - BTDevice_{LastSeen} > FrequentFlyerThreshold \quad (1)$$

And now for something completely different...



System Architecture



Demonstration Questions

1. Historical analysis of the number of passengers.
2. How many passengers are frequent flyers?
3. How much time do passengers spend on average in the different zones?
4. How much time do passengers spend on average before entering different zones?
5. How is the distribution of time spent per passenger in a given zone?
6. Which day of week are the zones used the most?
7. When is there a risk of bottlenecks in specific zones?

Conclusion

- Status of the project.
- Main project contributions.
  - Data warehouse design and ETL.
  - Bounce detection.
- Future work
  - Flow analysis of passengers movement and trends.
  - Real-time monitoring of passengers in the airport.
  - Develop a mobile application that can deliver LBS to the passengers.

New unsolved problem

- Blip would like a graph that shows the percentage of new devices grouped by time.
- We should be able to show this with our current data.

Questions to the audience

- How can we solve the problem with the graph showing new passengers in percentage over time?
- Ideas on how to perform flow analysis?
- Ideas on how to improve bounce detection?
- If implemented, how can we utilize sessions?