Data Warehouse Projects:
A short course for IT executives

Bob Wakefield, Chancellor

Bob@MassStreetUniversity.com

@BobLovesData

Mass Street University

PER EDUCATIONEM PROGRESSUS

Bob’s Background

- IT professional, 17 years
- Entrepreneur
- Education
  - BS Business Administration (MIS) from Kansas State University
  - MBA (finance concentration) from University of Kansas
  - Coursework in Mathematics at Washburn University
  - Graduate Certificate Data Science from Rockhurst University
- Addicted to everything data

Follow Us!

- Twitter: @MassStreetU
- Website: www.MassStreetUniversity.com
- Facebook: @MassStreetUniversity

Motivations for the Course

- Give IT managers the tools and language necessary to successfully manage a data warehouse project.
- Set proper expectations for a data warehouse project.
- Help IT managers hire the right talent.
- Help IT managers navigate potential land mines.
- Help IT managers understand what they need from the rest of the organization.

This is an Opinionated Course!

- You can’t find this information on the internet
- Based on 17 years of building databases
- Material is based on Mass Street’s philosophy and practices
- Material does not invalidate any other opinions or methods

Course Outline

Data Warehouse Overview

Data Warehouse ETL

Organizational Considerations

Let’s answer the following question

Should I invest time and money learning about traditional databases?

The current reality of Big Data

- The job market isn’t as big as the data
- Most companies don’t have big data
- There is still a wide gap between the perception and the reality
- Having a lot of data and being able to do anything with it are two separate things

The current reality of Big Data

- The job market isn’t as big as the data
- Most companies don’t have big data
- There is still a wide gap between the perception and the reality
- Having a lot of data and being able to do anything with it are two separate things

What is Medium Data®?

- I’m just kidding. You can use it.
- Not big data but not small data either
- Far more ubiquitous than Big Data

Getting the Course Material

- The latest course material can be found on GitHub.

- https://github.com/MassStreetAnalytics/data-warehouse-projects

Data Warehouse Overview

What is some important terminology?

- Data model
- Dimensional modeling
- Dimension table
- Fact table
- Slowly changing dimension (SCD)
- Star schema
- Snowflake
- Online Analytical Processing (OLAP)
- Master Data Management (MDM)
- Data Contract

Data Warehouse Overview

Why do we need a data warehouse?

- Historical analysis in transactional systems isn’t efficient.
- Removes data silos.
- Provides a central source of truth.
- Enables 360 analysis.
- Bonus: Helps uncover errors and inefficiencies in existing business processes.

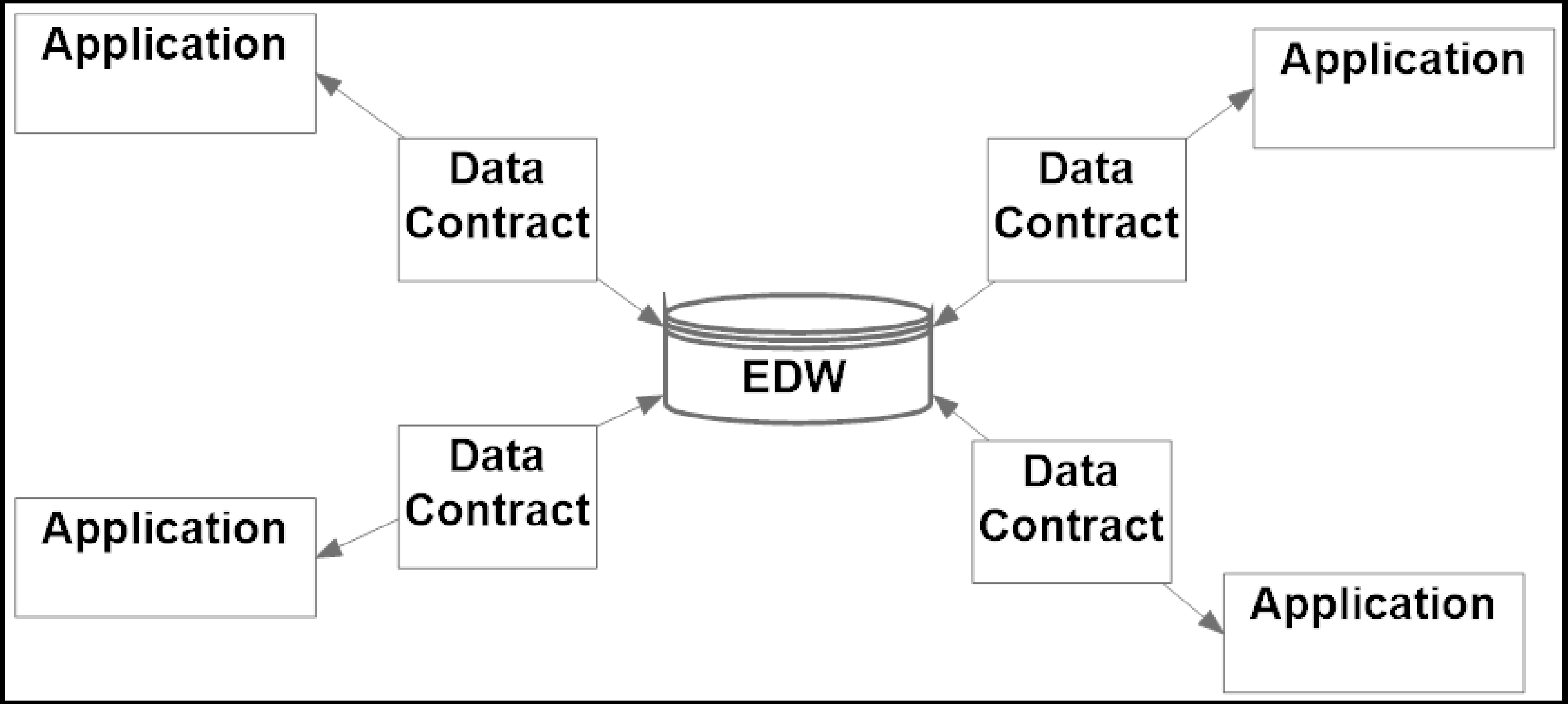
Data Warehouse Overview

What is a data warehouse?

- Enterprise Data Warehouse (EDW)
  - Central to any organization’s analytic efforts
  - Should be a central location of all historical data
  - Single source of truth
  - There are strict rules around processing and storage

Data Warehouse Overview

What is a data warehouse?



Data Warehouse Overview

What is NOT a data warehouse?

- Data Mart
  - Subset of data warehouse
  - Usually organized by business process
    - Finance, Marketing, HR, etc.

Data Warehouse Overview

What is NOT a data warehouse?

- Data Lake
  - A massive dumping ground for data.
  - There is no complex ETL.
  - Structure and relationships are determined after the load.

Data Warehouse Overview
What technology do you need?

- A server
  - Needs to be a dedicated box
  - Specs driven by selected DB
- A database
- Possibly a point-and-click ETL Tool
- A Master Data Management Tool
- Possibly an OLAP technology
- An analysis and visualization tool

Data Warehouse Overview

What technology do you need?

- Relational DBs
  - PostgreSQL, MySQL
  - SQL Server
  - Oracle
- NoSQL DBs
  - Druid
- Massively Parallel Processing DBs (MPP)
  - MySQL Cluster CGE
  - SQL Server Parallel Data Warehouse
  - Greenplum

Data Warehouse Overview

What technology do you need?

- An ETL Tool
  - Possibly
  - Usually used by junior people
  - More trouble than they are worth
- Popular ETL Tools
  - Informatica
  - SQL Server Integration Services (SSIS)
  - DataStage
- Less Popular Tools
  - Talend
  - Pentaho?

Data Warehouse Overview
What technology do you need?

- Master Data Management Tools (MDM)
  - SQL Server Master Data Services
    - It has some shortcomings
    - Some critical functions can't be automated
  - Talend
  - Informatica

Data Warehouse Overview

What technology do you need?

- OLAP software
- Can be used to create data marts
- OLAP isn’t that popular
- SQL Server Analysis Services
- There are others
  - Pyramid Analytics

Data Warehouse Overview

What technology do you need?

- Orchestration
- Called job scheduler in the old world
- Tools
  - SQL Server Agent
  - Control M
  - Apache Airflow

Data Warehouse Overview

What technology do you need?

- Analysis and visualization tools
- Traditional point-and-click tools
  - SQL Server Reporting Services
  - Power BI
  - Tableau
- New Tools
  - Jupyter Notebook
  - R Notebooks
  - Apache Zeppelin

Data Warehouse Overview

How are data warehouses built?

- There are two approaches to data warehouse design
  - Kimball
  - Inmon
- We build them as a destination for everything
  - Build a data mart from that

Data Warehouse Overview

How are data warehouses built?

- In general, there are two approaches to designing databases
  - Transactional
  - Dimensional
- Transactional
  - Highly normalized
  - Many tables
  - Many tables leads to many joins
  - Little, if any, data duplication

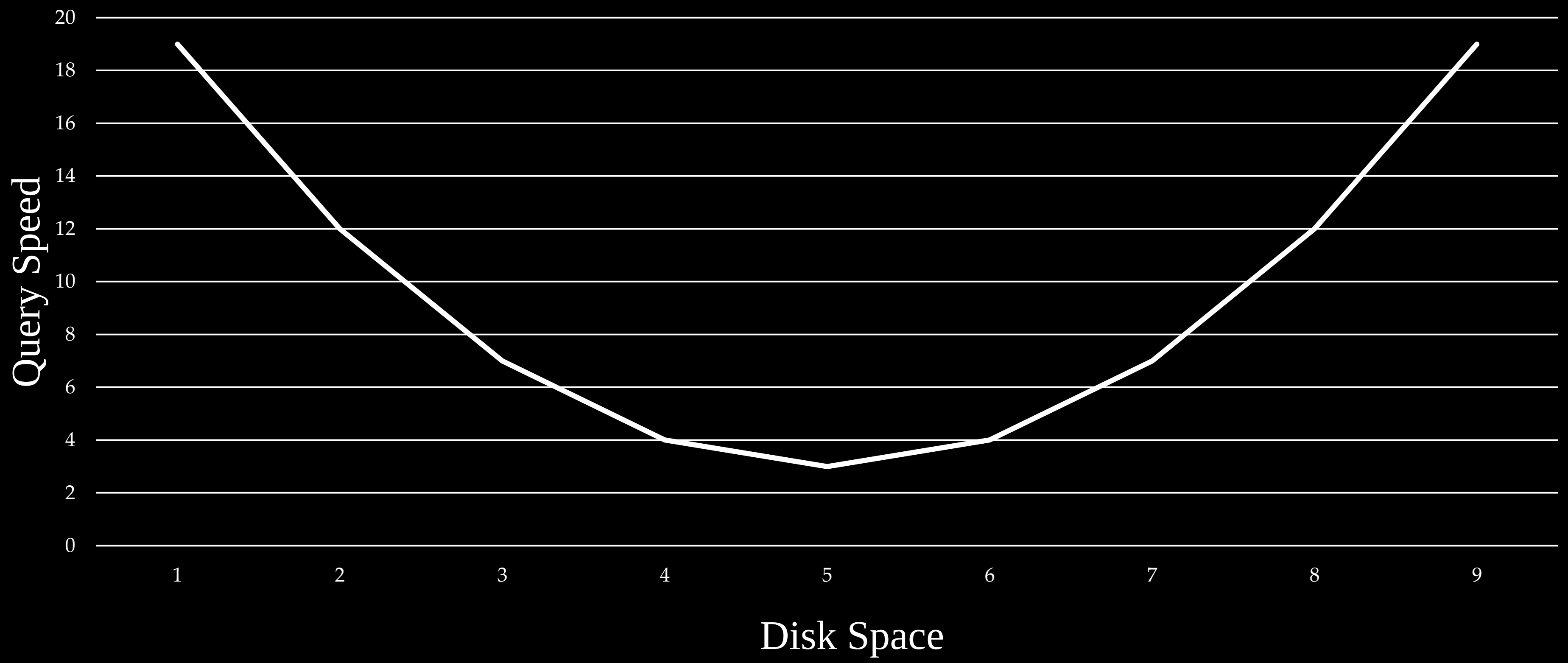
Data Warehouse Overview

How are data warehouses built?

- Dimensional
  - Denormalized
  - Far fewer tables
  - Fewer tables leads to fewer joins
  - Data duplication is ok

Data Warehouse Overview

How are data warehouses built?



Nobody likes a slow query.

Mass Street University
PER EDUCATIONEM PROGRESSUS

Data Warehouse Overview

How are data warehouses built?

- More than one way to skin the cat
- Convert the transactional database into a dimensional one.
  - Works if all the business rules are captured in the data model
  - Not resource-intensive
- Develop a two dimensional dataset to answer a specific question and model that.
  - This requires significant resources to support
  - Can be a much faster process

Data Warehouse Overview

How are data warehouses built?

- Recommended process flow for new models
    - Source data identified
    - Data model created
    - ETL created
    - Data validation/UAT
    - Deploy
- This should be a 1-3 week process
- If it takes longer, something is wrong

Data Warehouse Overview

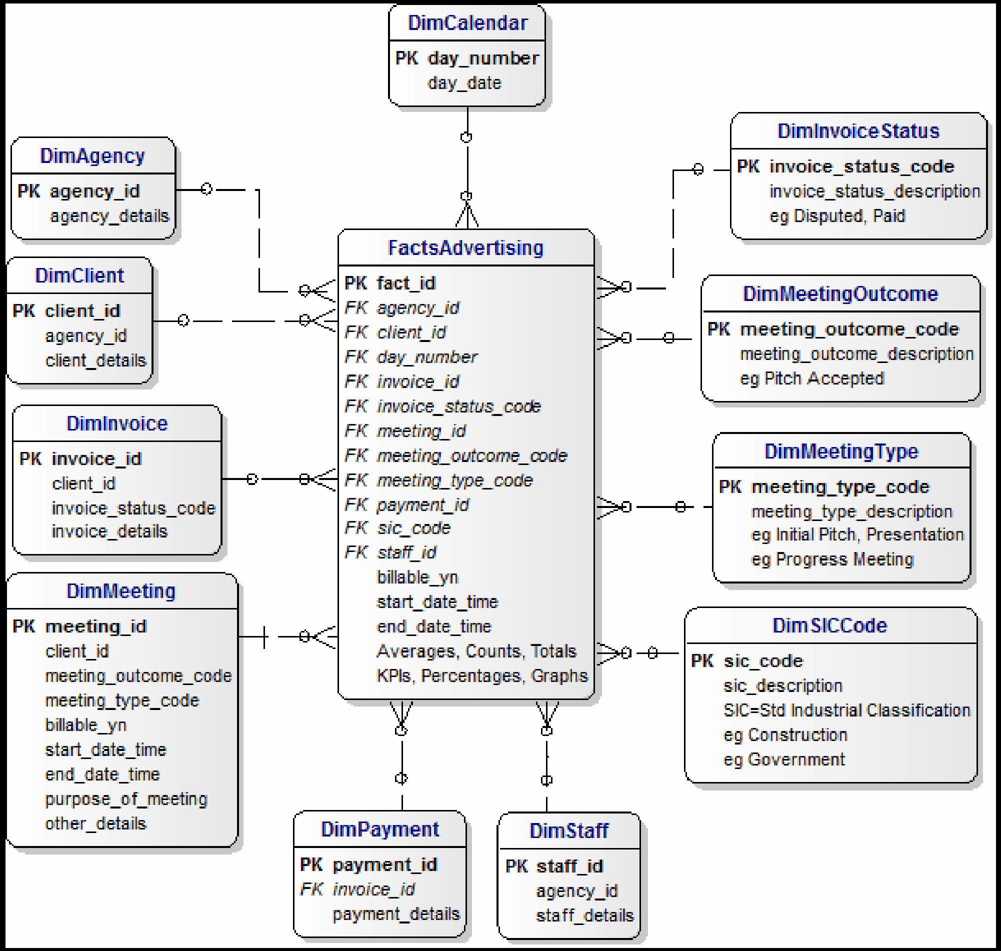
What is a data model?

- Abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world
- Models are built around fact tables
- Two dimensional representation of how data is stored
- Errors in data models will cause serious issues
  - Query performance suffers
  - Numbers won’t crunch correctly
  - Errors don’t get fixed because of priorities
  - Hacky workarounds worsen the situation
  - Fixing errors is expensive

Data Warehouse Overview

What is a data model?

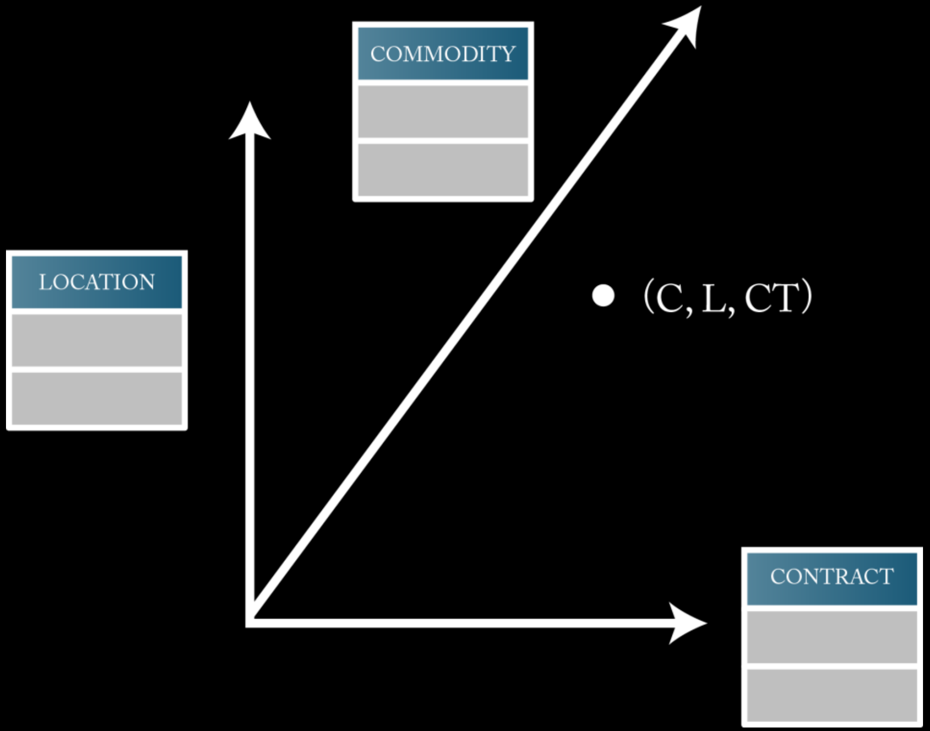
Data Model Example



Data Warehouse Overview

What is a dimension table?

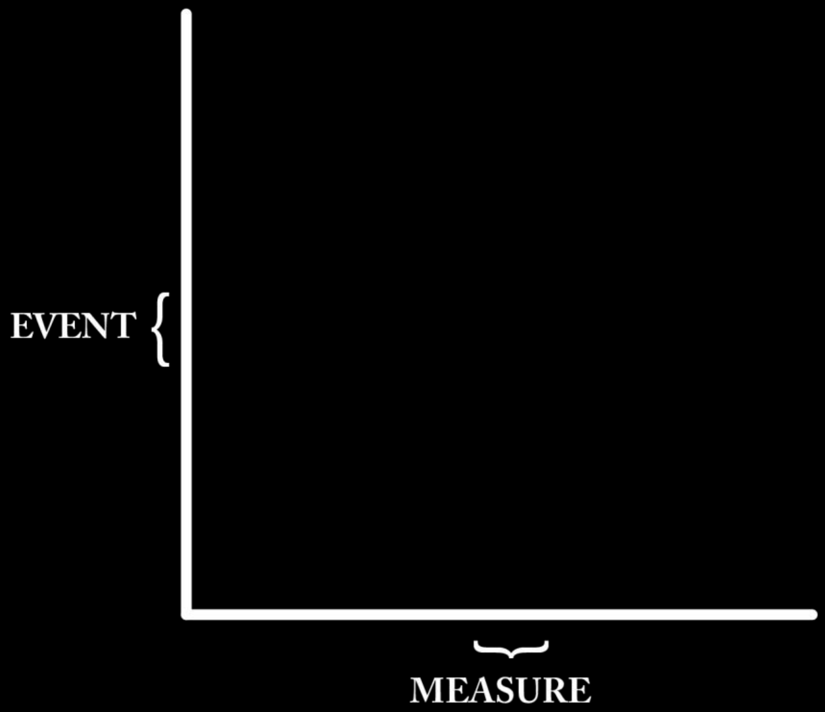
- Using the word dimension isn't arbitrary
- A dimensional model is a two dimensional representation of n dimensional space
- A dimension is like an axis on a chart
- Issuing a query is like issuing coordinates
- Facts (numbers) reside at the coordinates in n dimensional space



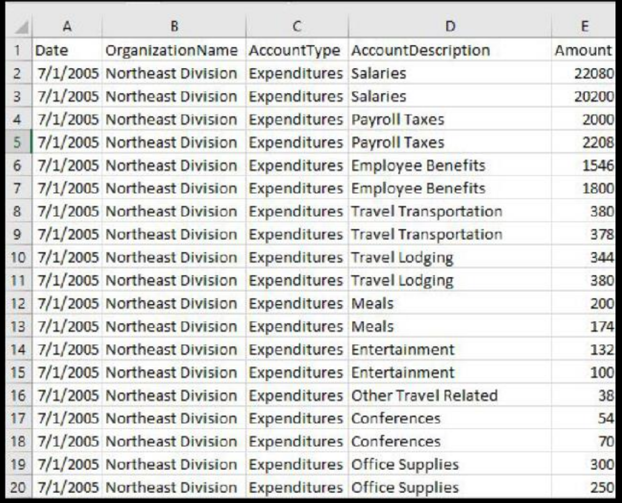
Data Warehouse Overview

What is a dimension table?

MENTAL MODEL



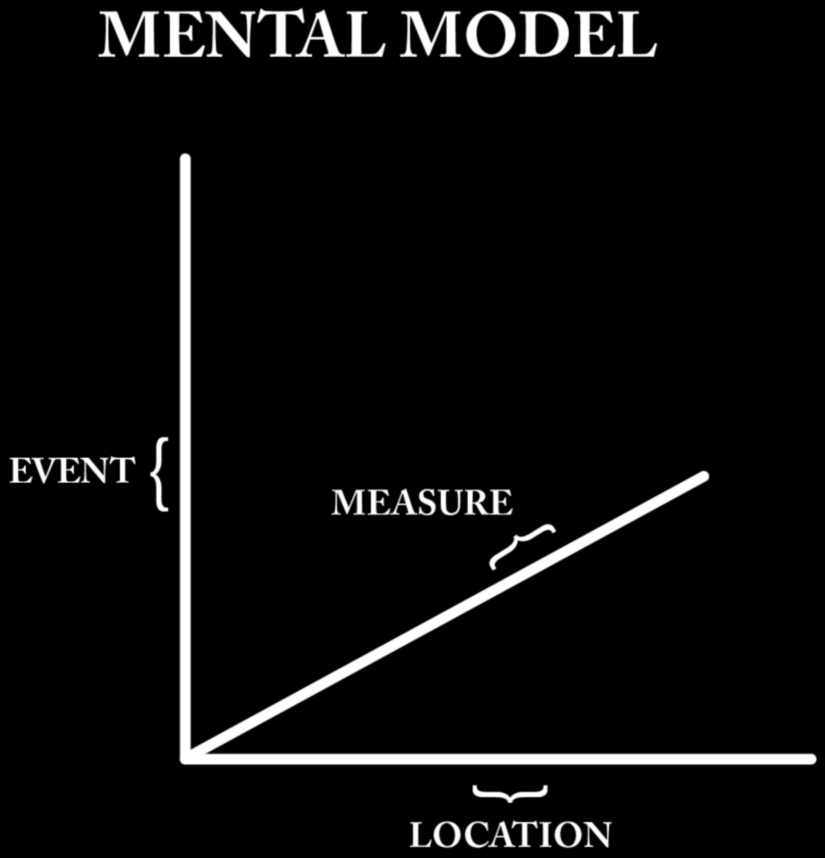
REAL WORLD OBJECT

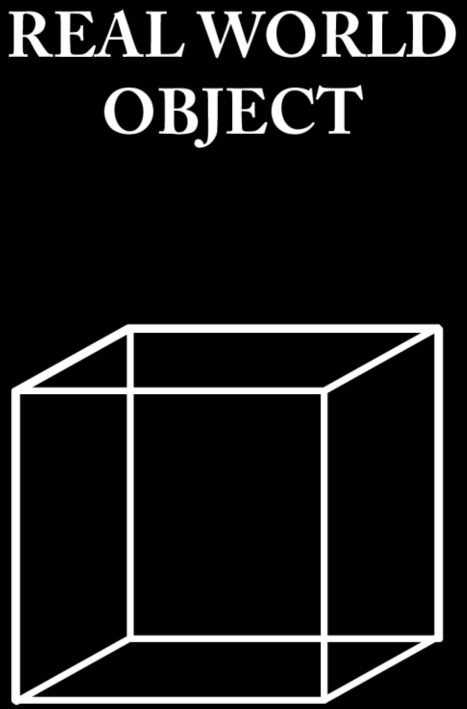


	A	B	C	D	E
1	Date	OrganizationName	AccountType	AccountDescription	Amount
2	7/1/2005	Northeast Division	Expenditures	Salaries	22080
3	7/1/2005	Northeast Division	Expenditures	Salaries	20200
4	7/1/2005	Northeast Division	Expenditures	Payroll Taxes	2000
5	7/1/2005	Northeast Division	Expenditures	Payroll Taxes	2208
6	7/1/2005	Northeast Division	Expenditures	Employee Benefits	1546
7	7/1/2005	Northeast Division	Expenditures	Employee Benefits	1800
8	7/1/2005	Northeast Division	Expenditures	Travel Transportation	380
9	7/1/2005	Northeast Division	Expenditures	Travel Transportation	378
10	7/1/2005	Northeast Division	Expenditures	Travel Lodging	344
11	7/1/2005	Northeast Division	Expenditures	Travel Lodging	380
12	7/1/2005	Northeast Division	Expenditures	Meals	200
13	7/1/2005	Northeast Division	Expenditures	Meals	174
14	7/1/2005	Northeast Division	Expenditures	Entertainment	132
15	7/1/2005	Northeast Division	Expenditures	Entertainment	100
16	7/1/2005	Northeast Division	Expenditures	Other Travel Related	38
17	7/1/2005	Northeast Division	Expenditures	Conferences	54
18	7/1/2005	Northeast Division	Expenditures	Conferences	70
19	7/1/2005	Northeast Division	Expenditures	Office Supplies	300
20	7/1/2005	Northeast Division	Expenditures	Office Supplies	250

Data Warehouse Overview

What is a dimension table?

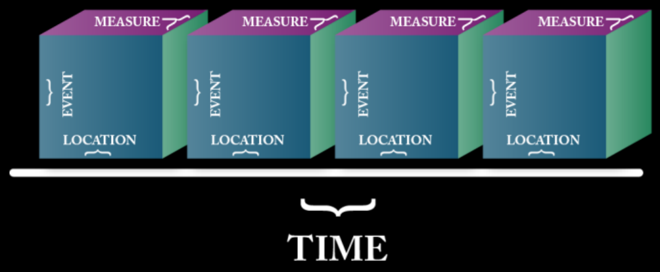




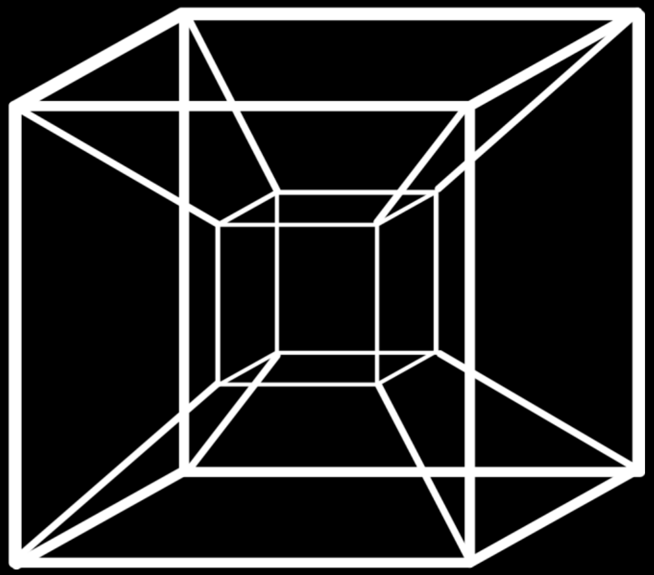
Data Warehouse Overview

What is a dimension table?

MENTAL MODEL

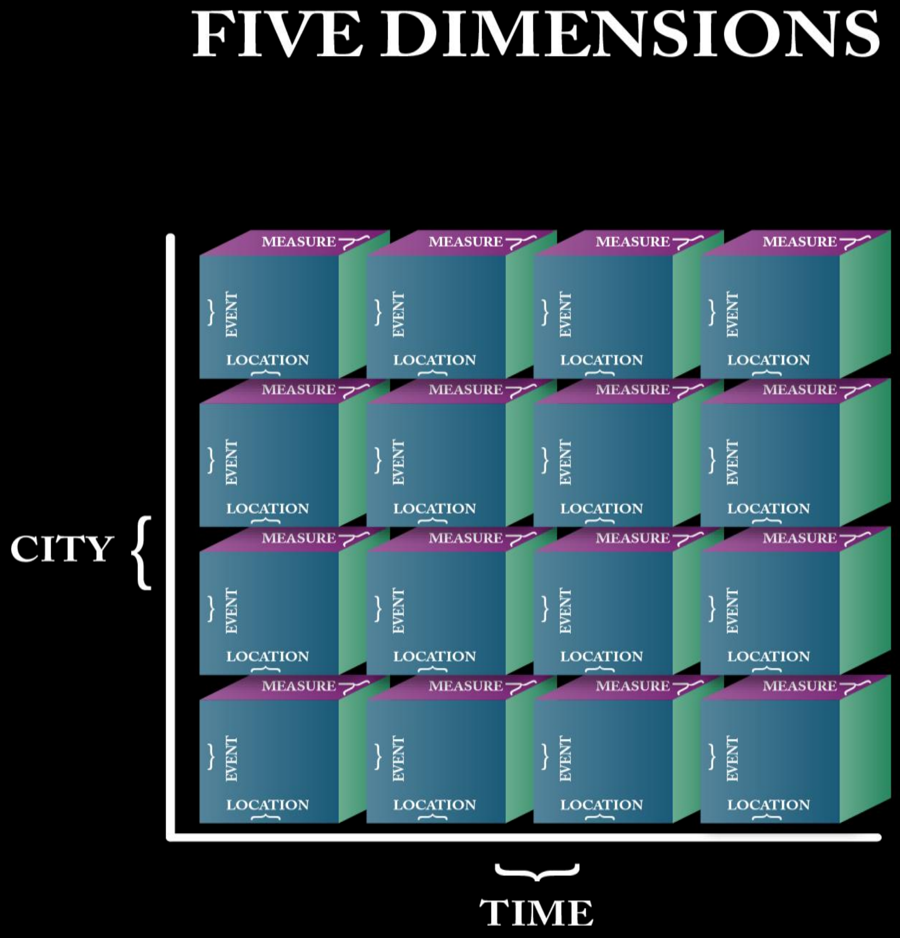


REAL WORLD OBJECT



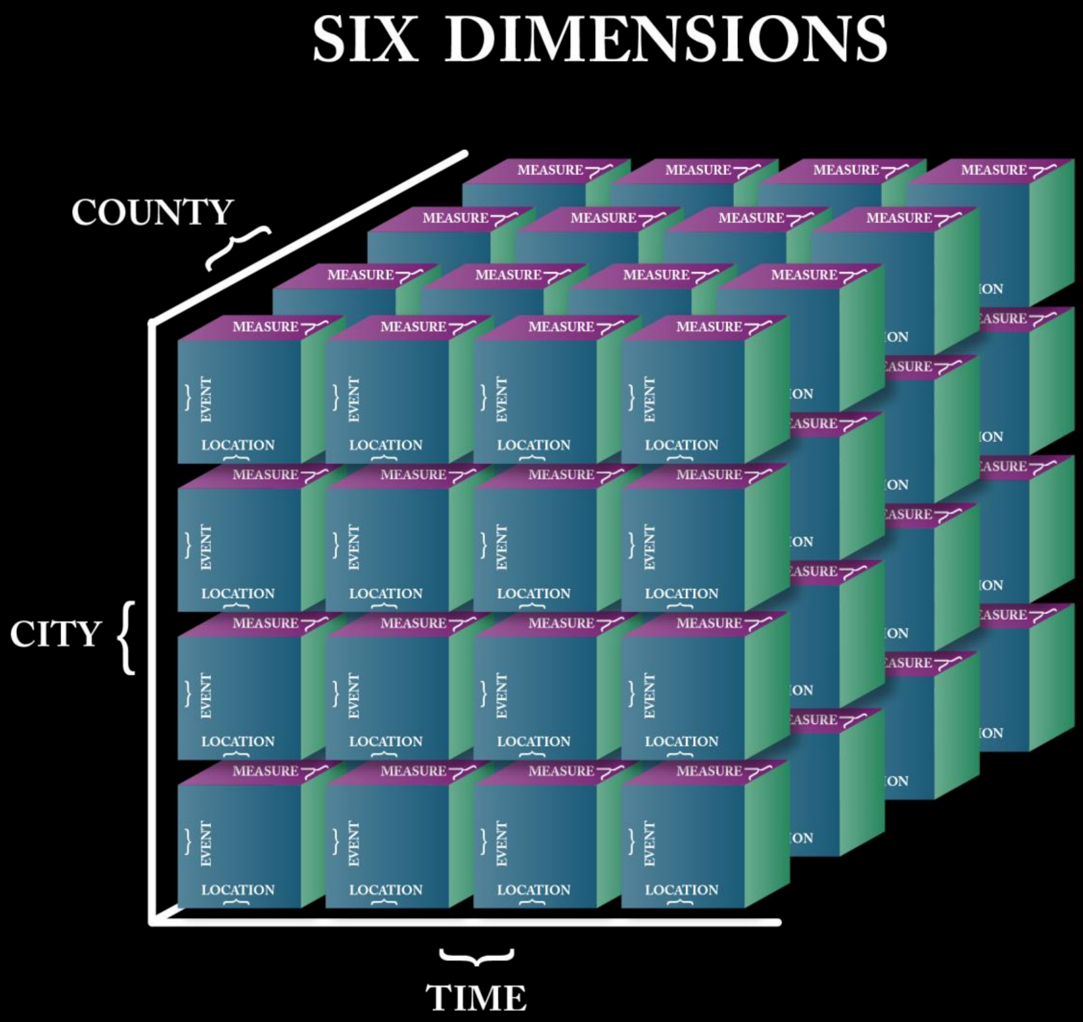
Data Warehouse Overview

What is a dimension table?



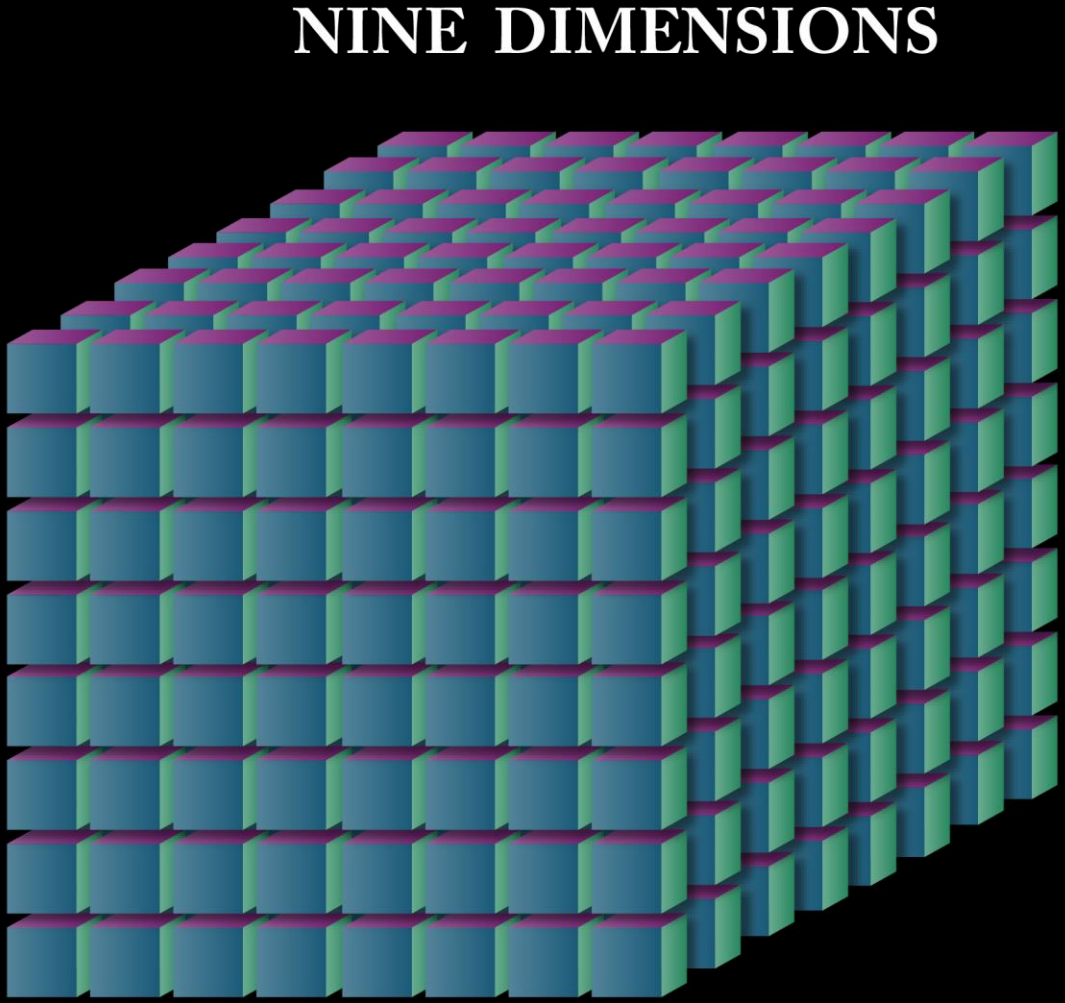
Data Warehouse Overview

What is a dimension table?



Data Warehouse Overview

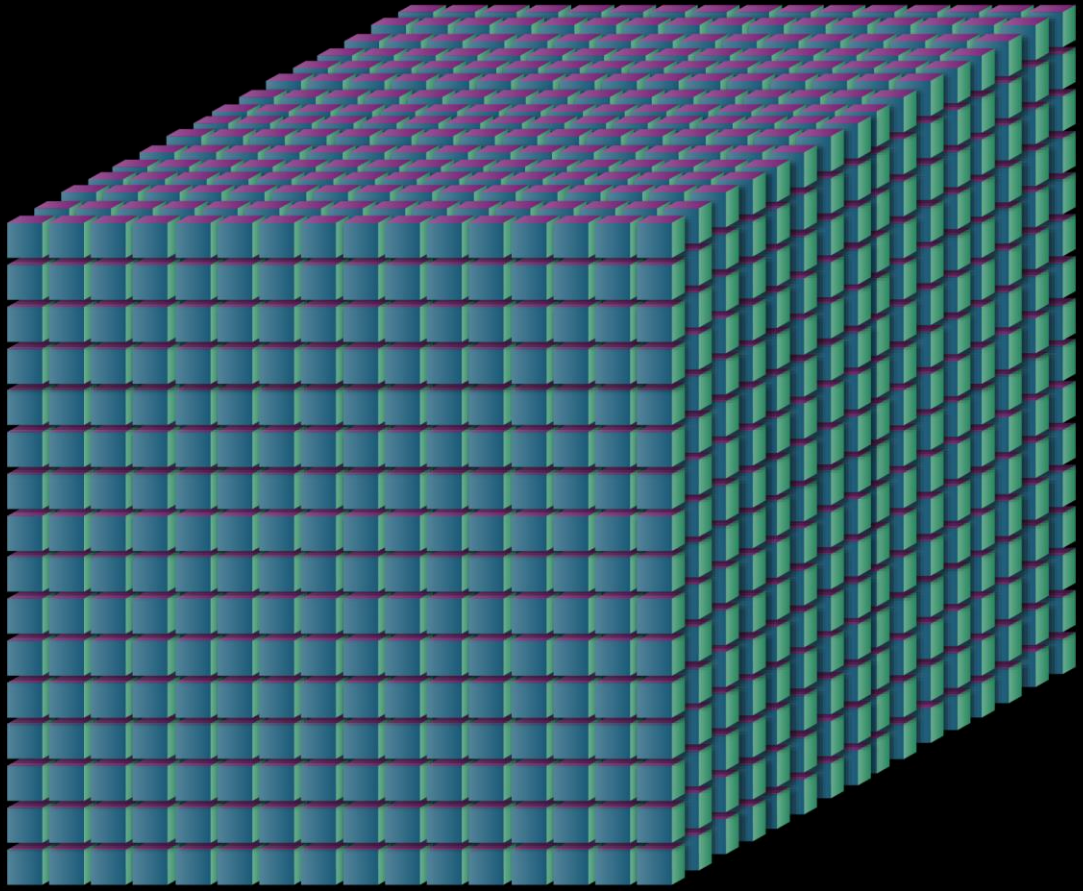
What is a dimension table?



Data Warehouse Overview

What is a dimension table?

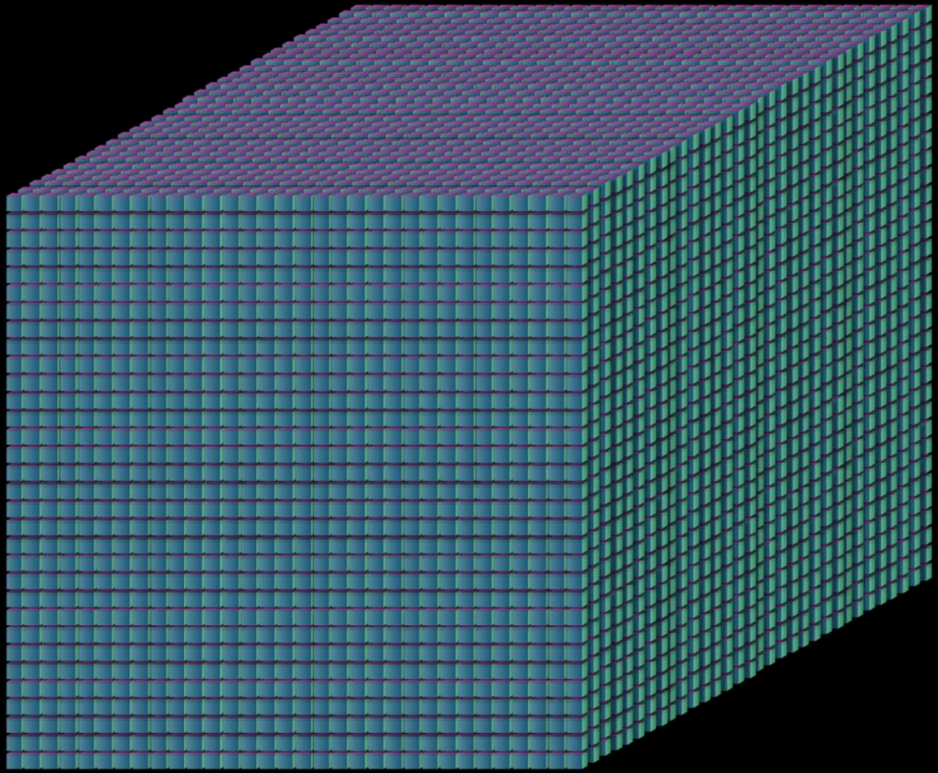
TWELVE DIMENSIONS



Data Warehouse Overview

What is a dimension table?

FIFTEEN DIMENSIONS



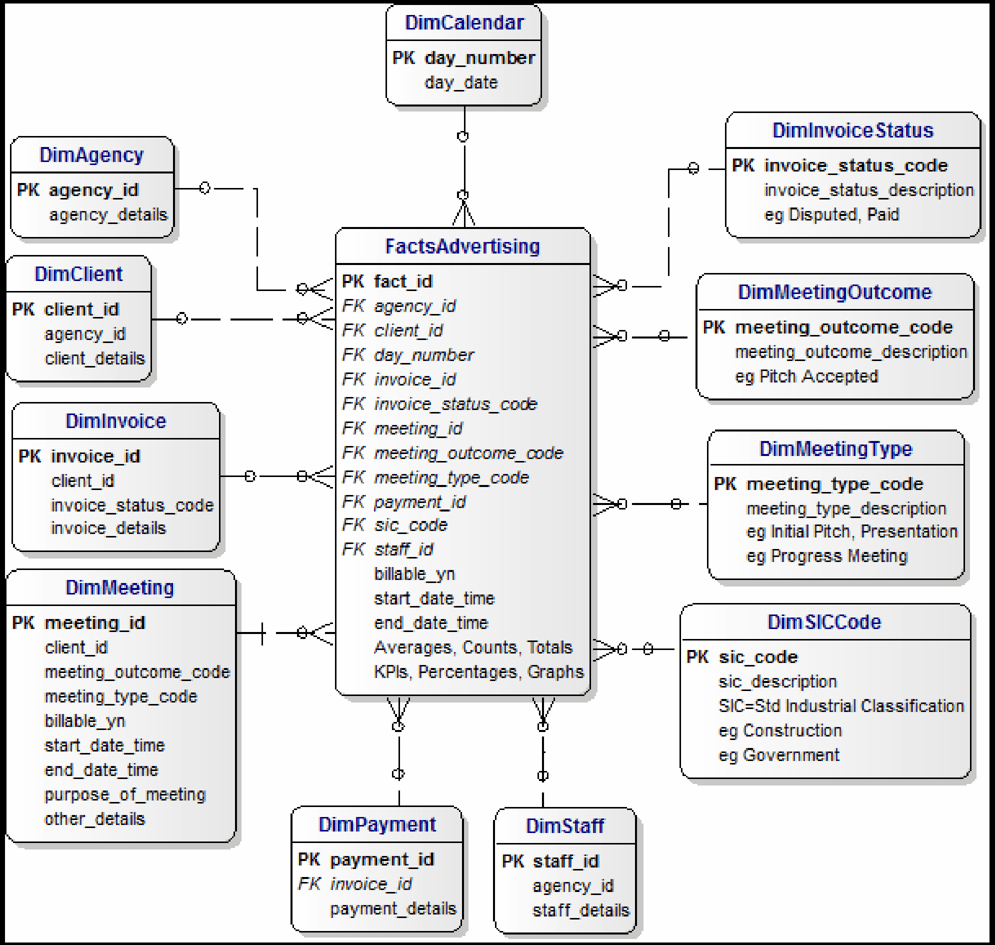
Data Warehouse Overview

What is a dimension table?

- Primary keys are meaningless (contrived key)
- Dimensions surround fact tables
- Dimensions contain textual filter data
- Dimensions can contain numbers as long as they aren’t being used to add
- There are various kinds of dimensions
- Data in dimensions can get duplicated
- Dimensional data changes slowly, if ever

Data Warehouse Overview

What is a dimension table?



DimCalendar
- PK day_number
- day_date

DimAgency
- PK agency_id
- agency_details

DimClient
- PK client_id
- agency_id
- client_details

DimInvoice
- PK invoice_id
- client_id
- invoice_status_code
- invoice_details

DimMeeting
- PK meeting_id
- client_id
- meeting_outcome_code
- meeting_type_code
- billable_yn
- start_date_time
- end_date_time
- purpose_of_meeting
- other_details

FactsAdvertising
- PK fact_id
- FK *agency_id*
- FK *client_id*
- FK *day_number*
- FK *invoice_id*
- FK *invoice_status_code*
- FK *meeting_id*
- FK *meeting_outcome_code*
- FK *meeting_type_code*
- FK *payment_id*
- FK *sic_code*
- FK *staff_id*
- billable_yn
- start_date_time
- end_date_time
- Averages, Counts, Totals
- KPIs, Percentages, Graphs

DimInvoiceStatus
- PK invoice_status_code
- invoice_status_description
- eg Disputed, Paid

DimMeetingOutcome
- PK meeting_outcome_code
- meeting_outcome_description
- eg Pitch Accepted

DimMeetingType
- PK meeting_type_code
- meeting_type_description
- eg Initial Pitch, Presentation
- eg Progress Meeting

DimSICCode
- PK sic_code
- sic_description
- SIC=Std Industrial Classification
- eg Construction
- eg Government

DimPayment
- PK payment_id
- FK *invoice_id*
- payment_details

DimStaff
- PK staff_id
- agency_id
- staff_details

Data Warehouse Overview

What is a dimension table?

- Most dimensions should change rarely
- Slowly Changing Dimension Types
  - Type I – Make the change. Don’t keep history.
  - Type II – Add a row. Retire the old row.
  - Type III – Put the change in a different column.
  - Type X – Use some combination of the above.

Data Warehouse Overview

What is a dimension table?

- Types of Dimensions
  - Date dimension – Used to make time-based queries
    - Not strictly a type of dimension
  - Junk – unique combinations of low cardinality values
  - Degenerate – Dimension that goes in fact table
  - Role Play – Copy of an existing dimension but renamed
  - Conformed – Dimension used to jump across models
  - Outrigger – Used to snowflake

Data Warehouse Overview

What is a dimension table?

- Date dimension
- Absolutely critical piece of any data warehouse
- Allows you to easily slice and dice data based on date parameters
- Can be customized for your particular organization
  - More columns can be added
  - Can be adjusted for your fiscal calendar
- Almost every fact table will connect to the date dimension

Data Warehouse Overview
What is a dimension table?

- Junk Dimension
  - Unique combination of low cardinality values
  - Usually a table filled with flags or binary values
  - You can use math to predict the number of rows in the table
    - $X_1 * X_2 * X_3 \ldots * X_n$
    - Where X = number of unique values

Data Warehouse Overview

What is a dimension table?

- Degenerate Dimension
  - A dimension that does not belong in its own table
  - Goes in the fact table
  - Invoice numbers are a good example
  - Resist the temptation to create inappropriate degenerate dimensions
    - If it repeats infinitely, it is NOT a degenerate dimension

Data Warehouse Overview

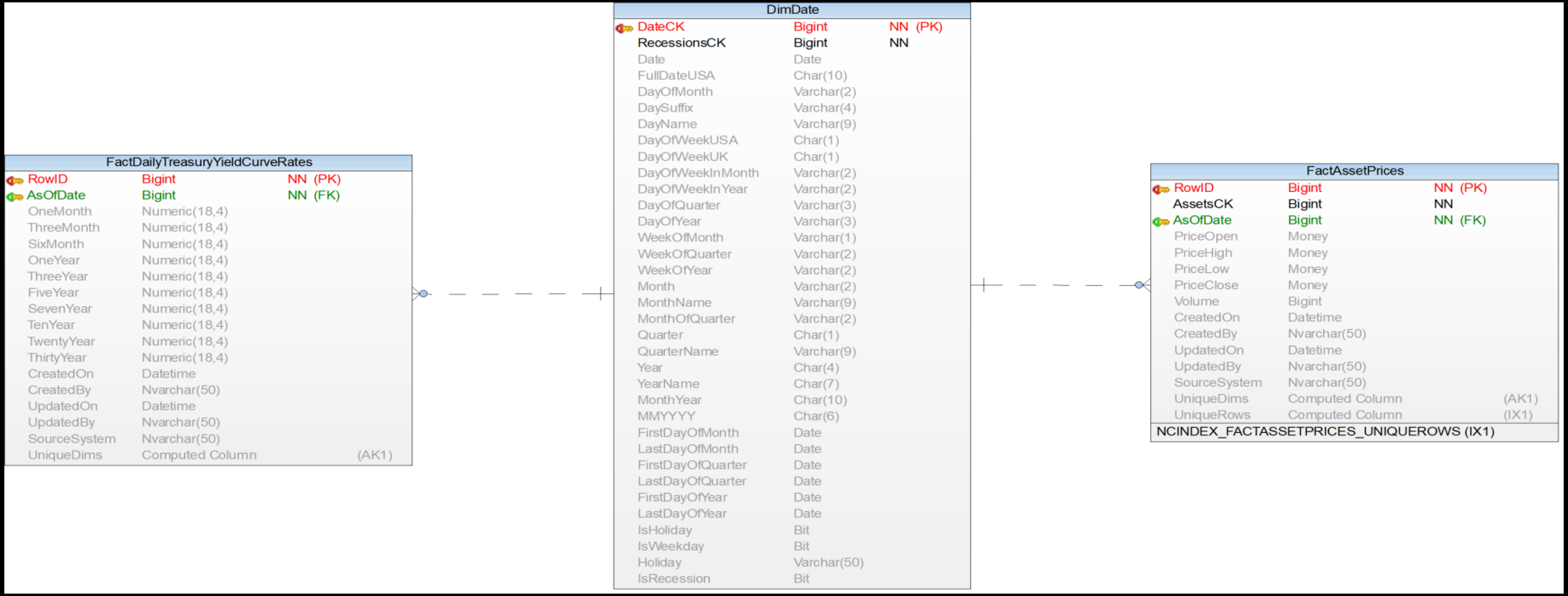
What is a dimension table?

- Role Play
  - Close copy of an existing dimension
    - Data is the same. Table and columns are named differently
  - Implemented with views
  - Used to make things clearer for users and creates cleaner code for data professionals
  - Most commonly implemented with the date dimension

Data Warehouse Overview

What is a dimension table?

Conformed Dimension Example



FactDailyTreasuryYieldCurveRates			
RowID	Bigint	NN (PK)	
AsOfDate	Bigint	NN (FK)	
OneMonth	Numeric(18,4)		
ThreeMonth	Numeric(18,4)		
SixMonth	Numeric(18,4)		
OneYear	Numeric(18,4)		
ThreeYear	Numeric(18,4)		
FiveYear	Numeric(18,4)		
SevenYear	Numeric(18,4)		
TenYear	Numeric(18,4)		
TwentyYear	Numeric(18,4)		
ThirtyYear	Numeric(18,4)		
CreatedOn	Datetime		
CreatedBy	Nvarchar(50)		
UpdatedOn	Datetime		
UpdatedBy	Nvarchar(50)		
SourceSystem	Nvarchar(50)		
UniqueDims	Computed Column		(AK1)

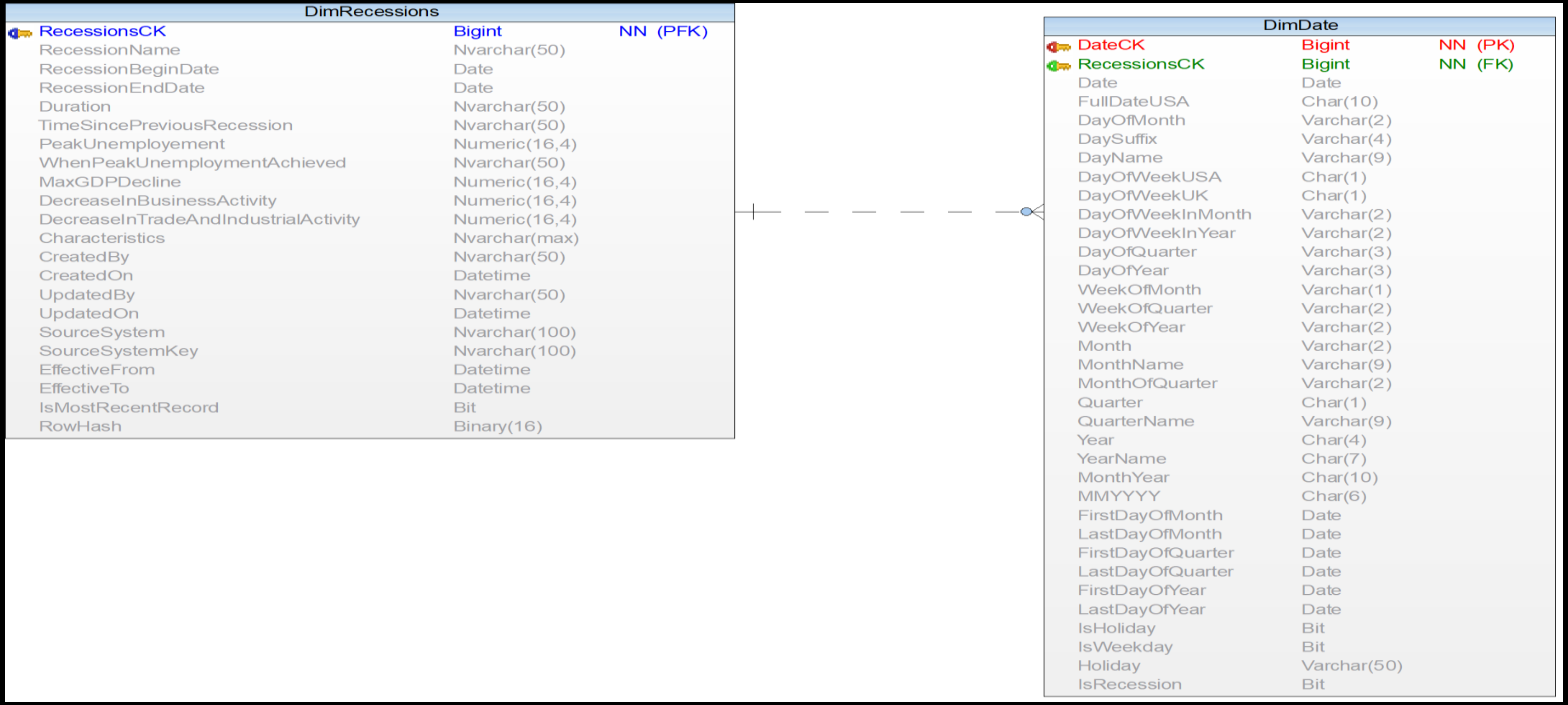
DimDate		
DateCK	Bigint	NN (PK)
RecessionsCK	Bigint	NN
Date	Date	
FullDateUSA	Char(10)	
DayOfMonth	Varchar(2)	
DaySuffix	Varchar(4)	
DayName	Varchar(9)	
DayOfWeekUSA	Char(1)	
DayOfWeekUK	Char(1)	
DayOfWeekInMonth	Varchar(2)	
DayOfWeekInYear	Varchar(2)	
DayOfQuarter	Varchar(3)	
DayOfYear	Varchar(3)	
WeekOfMonth	Varchar(1)	
WeekOfQuarter	Varchar(2)	
WeekOfYear	Varchar(2)	
Month	Varchar(2)	
MonthName	Varchar(9)	
MonthOfQuarter	Varchar(2)	
Quarter	Char(1)	
QuarterName	Varchar(9)	
Year	Char(4)	
YearName	Char(7)	
MonthYear	Char(10)	
MMYYYY	Char(6)	
FirstDayOfMonth	Date	
LastDayOfMonth	Date	
FirstDayOfQuarter	Date	
LastDayOfQuarter	Date	
FirstDayOfYear	Date	
LastDayOfYear	Date	
IsHoliday	Bit	
IsWeekday	Bit	
Holiday	Varchar(50)	
IsRecession	Bit	

FactAssetPrices			
RowID	Bigint	NN (PK)	
AssetsCK	Bigint	NN	
AsOfDate	Bigint	NN (FK)	
PriceOpen	Money		
PriceHigh	Money		
PriceLow	Money		
PriceClose	Money		
Volume	Bigint		
CreatedOn	Datetime		
CreatedBy	Nvarchar(50)		
UpdatedOn	Datetime		
UpdatedBy	Nvarchar(50)		
SourceSystem	Nvarchar(50)		
UniqueDims	Computed Column		(AK1)
UniqueRows	Computed Column		(IX1)
NCINDEX_FACTASSETPRICES_UNIQUEROWS (IX1)			

Data Warehouse Overview

What is a dimension table?

Outrigger Dimension Example



DimRecessions

RecessionsCK	Bigint	NN (PFK)
RecessionName	Nvarchar(50)	
RecessionBeginDate	Date	
RecessionEndDate	Date	
Duration	Nvarchar(50)	
TimeSincePreviousRecession	Nvarchar(50)	
PeakUnemployement	Numeric(16,4)	
WhenPeakUnemploymentAchieved	Nvarchar(50)	
MaxGDPDecline	Numeric(16,4)	
DecreaseInBusinessActivity	Numeric(16,4)	
DecreaseInTradeAndIndustrialActivity	Numeric(16,4)	
Characteristics	Nvarchar(max)	
CreatedBy	Nvarchar(50)	
CreatedOn	Datetime	
UpdatedBy	Nvarchar(50)	
UpdatedOn	Datetime	
SourceSystem	Nvarchar(100)	
SourceSystemKey	Nvarchar(100)	
EffectiveFrom	Datetime	
EffectiveTo	Datetime	
IsMostRecentRecord	Bit	
RowHash	Binary(16)	

DimDate

DateCK	Bigint	NN (PK)
RecessionsCK	Bigint	NN (FK)
Date	Date	
FullDateUSA	Char(10)	
DayOfMonth	Varchar(2)	
DaySuffix	Varchar(4)	
DayName	Varchar(9)	
DayOfWeekUSA	Char(1)	
DayOfWeekUK	Char(1)	
DayOfWeekInMonth	Varchar(2)	
DayOfWeekInYear	Varchar(2)	
DayOfQuarter	Varchar(3)	
DayOfYear	Varchar(3)	
WeekOfMonth	Varchar(1)	
WeekOfQuarter	Varchar(2)	
WeekOfYear	Varchar(2)	
Month	Varchar(2)	
MonthName	Varchar(9)	
MonthOfQuarter	Varchar(2)	
Quarter	Char(1)	
QuarterName	Varchar(9)	
Year	Char(4)	
YearName	Char(7)	
MonthYear	Char(10)	
MMYYYY	Char(6)	
FirstDayOfMonth	Date	
LastDayOfMonth	Date	
FirstDayOfQuarter	Date	
LastDayOfQuarter	Date	
FirstDayOfYear	Date	
LastDayOfYear	Date	
IsHoliday	Bit	
IsWeekday	Bit	
Holiday	Varchar(50)	
IsRecession	Bit	

Mass Street University

PER EDUCATIONEM PROGRESSUS

Data Warehouse Overview

What is a fact table?

- Two dimensional representation of an answer to a question
- Each record is a single event
- All records have to be at the same grain
- Records have to be additive (usually)
  - Corrections have to be complete or incremental
  - There are also semi-additive and non-additive records
- Types of fact tables
  - Snapshot
  - Cumulative

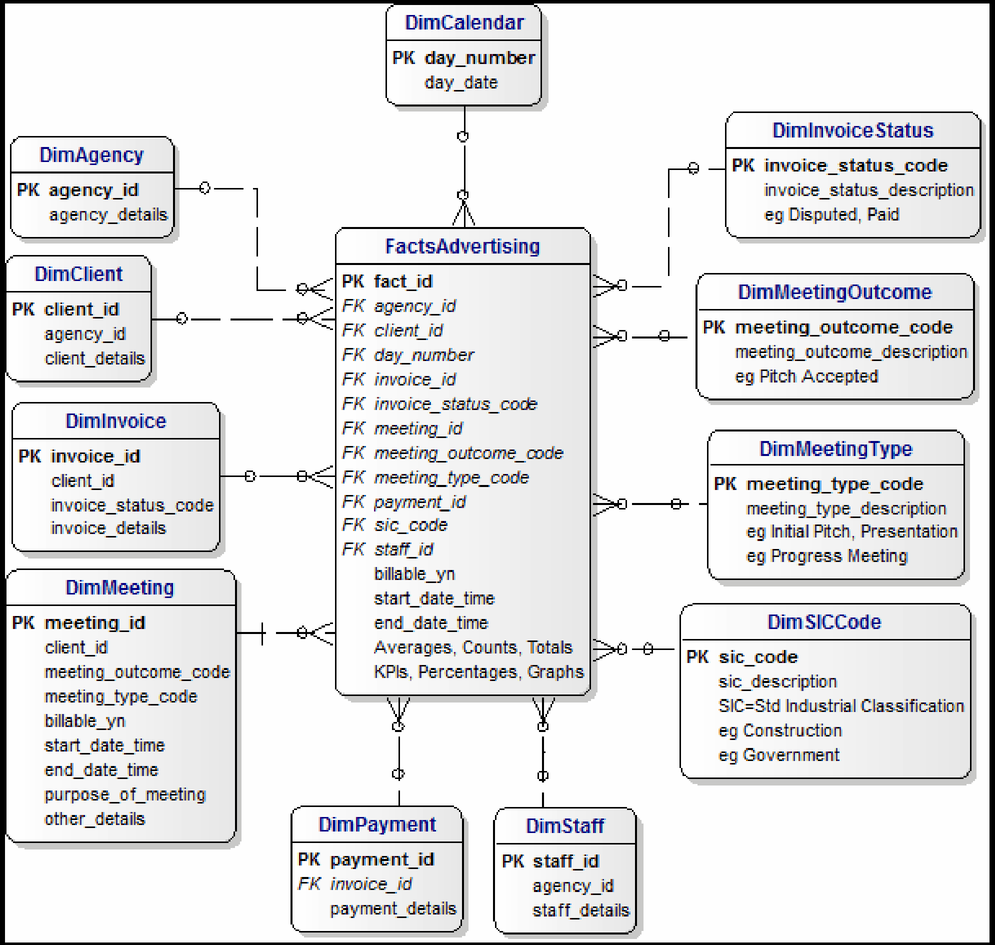
Data Warehouse Overview

What is a fact table?

- The primary key of a fact table is the unique combination of dimensions associated with that fact
- That key combination defines the record’s signature
- You can add an auto increment key but not for record identification
- Defining unique records is critical

Data Warehouse Overview

What is a fact table?



DimCalendar
PK day_number
day_date

DimAgency
PK agency_id
agency_details

DimClient
PK client_id
agency_id
client_details

DimInvoice
PK invoice_id
client_id
invoice_status_code
invoice_details

DimMeeting
PK meeting_id
client_id
meeting_outcome_code
meeting_type_code
billable_yn
start_date_time
end_date_time
purpose_of_meeting
other_details

FactsAdvertising
PK fact_id
FK agency_id
FK client_id
FK day_number
FK invoice_id
FK invoice_status_code
FK meeting_id
FK meeting_outcome_code
FK meeting_type_code
FK payment_id
FK sic_code
FK staff_id
billable_yn
start_date_time
end_date_time
Averages, Counts, Totals
KPIs, Percentages, Graphs

DimInvoiceStatus
PK invoice_status_code
invoice_status_description
eg Disputed, Paid

DimMeetingOutcome
PK meeting_outcome_code
meeting_outcome_description
eg Pitch Accepted

DimMeetingType
PK meeting_type_code
meeting_type_description
eg Initial Pitch, Presentation
eg Progress Meeting

Dim SICCode
PK sic_code
sic_description
SIC=Std Industrial Classification
eg Construction
eg Government

DimPayment
PK payment_id
FK invoice_id
payment_details

DimStaff
PK staff_id
agency_id
staff_details

Data Warehouse Overview

What is Master Data?

Plan to make MDM part of your data architecture from Day 1!

Data Warehouse Overview

What is Master Data?

- Master data is the set of data objects that are at the center of business activities (Customers, Products, Cost Centers, Locations, Assets, Task)
- Central concept necessary for any customer data integration project.
- MDM is a human in the loop process.

Data Warehouse Overview

What is Master Data?

- The disciplines, technologies, and solutions that are used to create and maintain consistent and accurate master data for all stakeholders across and beyond the enterprise
- The fundamental purpose of an MDM System is to serve as the authoritative source for master data; an MDM System is a system that provides clean, consistent master data to the enterprise.

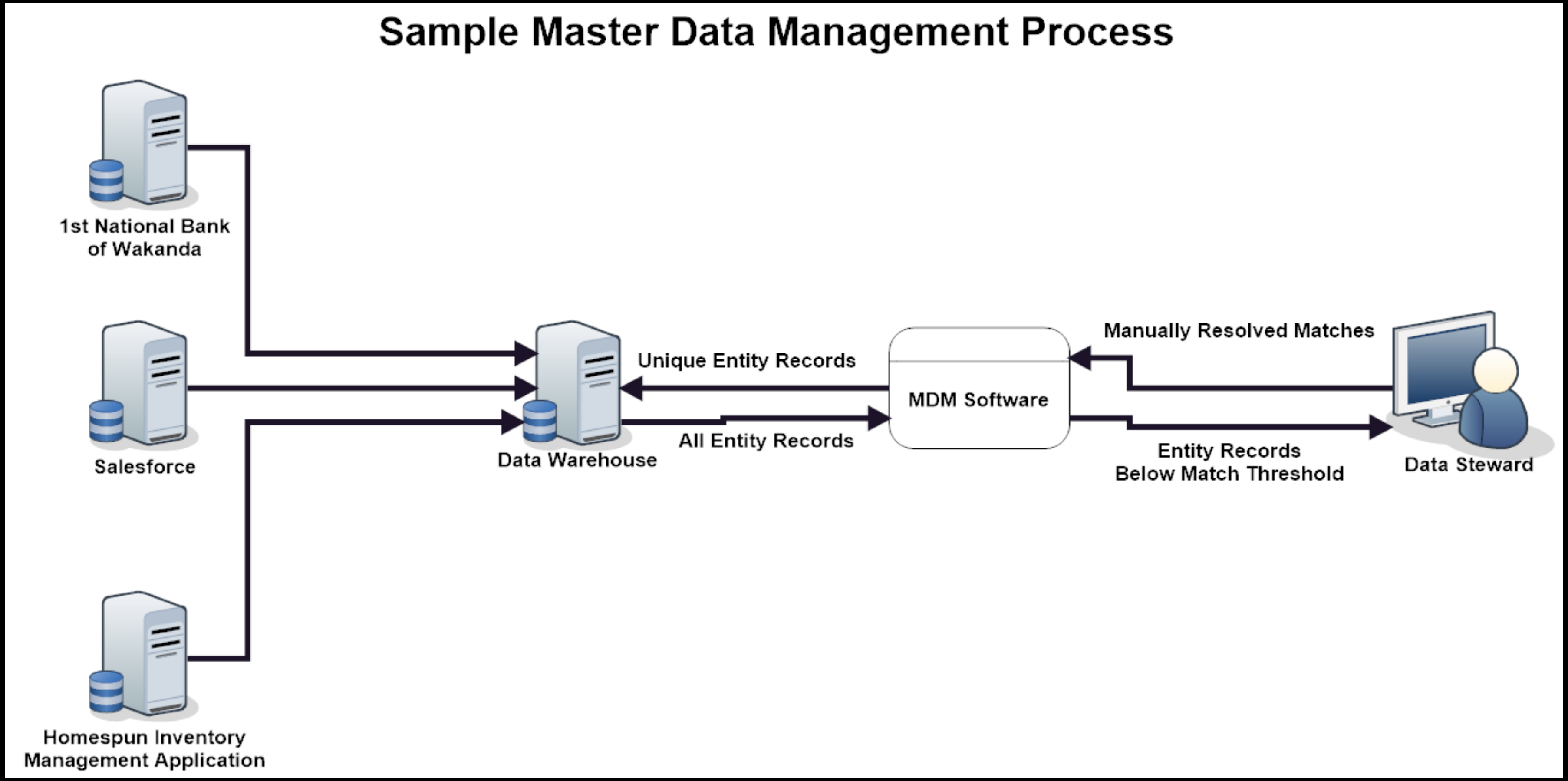
Data Warehouse Overview

What is Master Data?

- The problem is actually technical in nature
- De-duplicate identical entities from different systems
- De-duplicate identical entities from the same systems
- Manage all those keys so fact records flow through properly

Data Warehouse Overview

What is Master Data?


Sample Master Data Management Process
1st National Bank of Wakanda
Salesforce
Homespun Inventory Management Application
Data Warehouse
Unique Entity Records
All Entity Records
MDM Software
Manually Resolved Matches
Entity Records Below Match Threshold
Data Steward

Data Warehouse Overview

What is Master Data?

- How to do MDM from Day 1
  - Identify entities in your organization
  - Identify the source systems
  - Identify a data steward
  - Begin the process of selecting MDM software

Data Warehouse Overview
Section Conclusion

Data Warehouse ETL

What is ETL?

E = Extract

T = Transform

L = Load

Data Warehouse ETL

What is ETL?

- The process of moving data from one location to the next
- Usually reserved for talking about loading historical data
- This is different from the process of moving data in and out of a transactional system
- Usually involves moving large amounts of data in a batch process

Data Warehouse ETL

What is ETL?

- Also involves combining data sources
- Usually a rigorous process with many business rules
- Being able to write high performing code is important
- Can easily turn into a hot mess
- Might be referred to as ELT

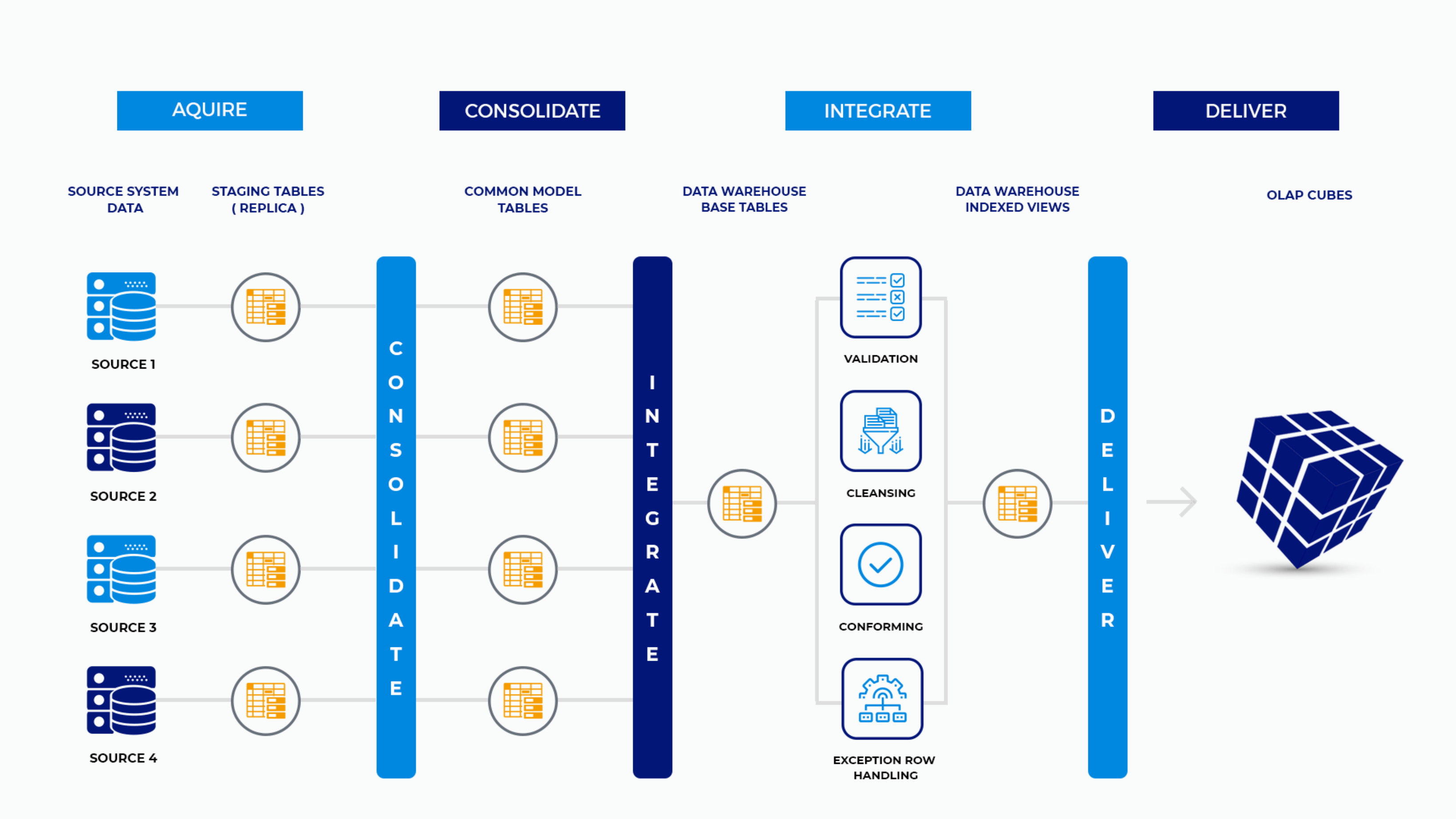
Data Warehouse ETL

What is ETL?

- Two methods for ETL
  - Supply push
  - Demand pull
- Demand pull is the most common and recommended approach
- Supply push is often necessary for third party integration

Data Warehouse ETL

What is ETL?

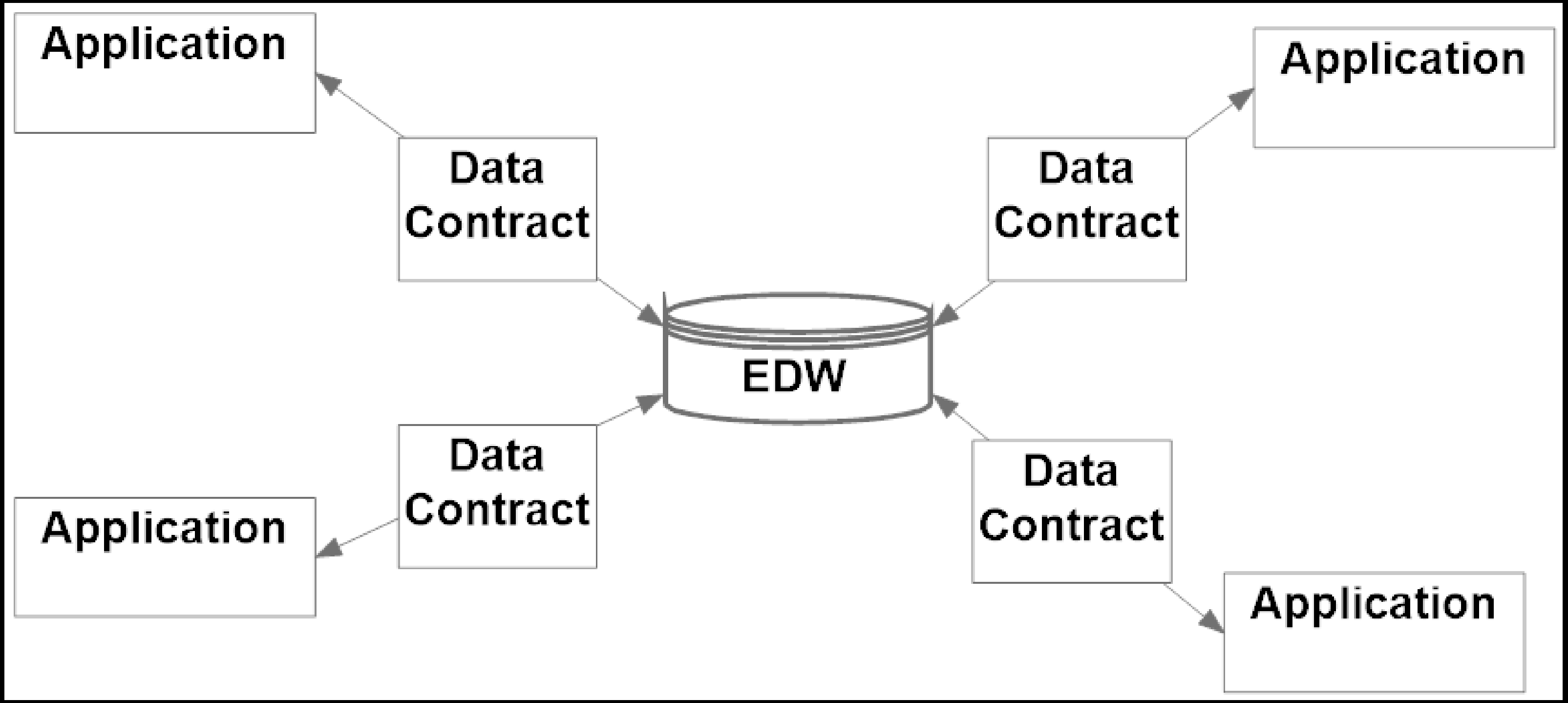


Mass Street University

PER EDUCATIONEM PROGRESSUS

Data Warehouse ETL

What is a Data Contract?



Data Warehouse ETL

What is a Data Contract?

- An agreement between systems that outlines the exchange of data
- Engineers on both ends of a pipe need to agree on what is flowing through that pipe
- Data contracts have two general implementations
  - As actual software via an interface
  - A written agreement between two departments

Data Warehouse ETL

What is a Data Contract?

- Data contracts as interfaces
  - An API is a type of data contract
  - Not appropriate for data warehouse work
  - Specialized APIs can be written but not recommended

Data Warehouse ETL

What is a Data Contract?

- Data contracts as written agreements
  - This is the recommended approach
  - It’s critical for supply push processes that might change frequently
  - Change control needs to be a part of the contract

Data Warehouse ETL

What is a Data Contract?

- Data contracts are messy in the real world
- A mixed approach may be necessary
- Be prepared to adapt and overcome

Data Warehouse ETL

What is an ETL Framework?

- A standardized methodology to move data into and out of the data warehouse
- Can be generalized to moving data around the entire organization
- Enables management of all data pipelines
- Helps significantly reduce the cost of ETL maintenance

Data Warehouse ETL

What is an ETL Framework?

- Why do we need an ETL Framework?
  - Helps prevent the Rube Goldberg machine
- Maintenance is not a data engineer’s highest or best use
- We see maintenance as a waste of cash

Data Warehouse ETL

What is the Mass Street ETL Framework?

- An open source ETL Framework for SQL Server
- Technically for any RDBMS
- Six years in the making
- Its use is taught in implementing data warehouses in the real world
- Freely available on the MIT license
- Pull requests are more than welcome

Data Warehouse ETL

What is the Mass Street ETL Framework?

- Mass Street ETL Framework components
  - Full documentation on usage
  - Sample model creation template
  - SQL Scripts for creating all necessary objects
  - Sample scripts for creating new objects
  - Sample scripts for data processing
  - Helper scripts to tackle common task

Data Warehouse ETL
Section Conclusion

Mass Street University
PER EDUCATIONEM PROGRESSUS

Organizational Considerations

Getting Organizational Buy-In

- Organizational buy-in is critical for project success
- This requires leadership from the top down
- You’re going to change the status quo
  - People will resist
  - People will be ambivalent
- You need to involve everybody at every level
  - Marketing is a great place to start

Organizational Considerations

Getting Organizational Buy-In

- The fastest way to get buy-in is quick wins!
  - The lure of big problems is seductive
- Sometimes quick wins won’t do it.
  - Involve business users and get them invested
- Don’t hide your tech resources
- Implementation should be a big deal
  - Anything having to do with the warehouse should be a newsworthy corporate event

Organizational Considerations

Getting Organizational Buy-In

- The data warehouse is exciting!
- It reduces cost!
- It saves time!
- It allows you to do analysis you could never do before!
- It's a necessary step before any data science efforts!
- It cures cancer! (Not really.)

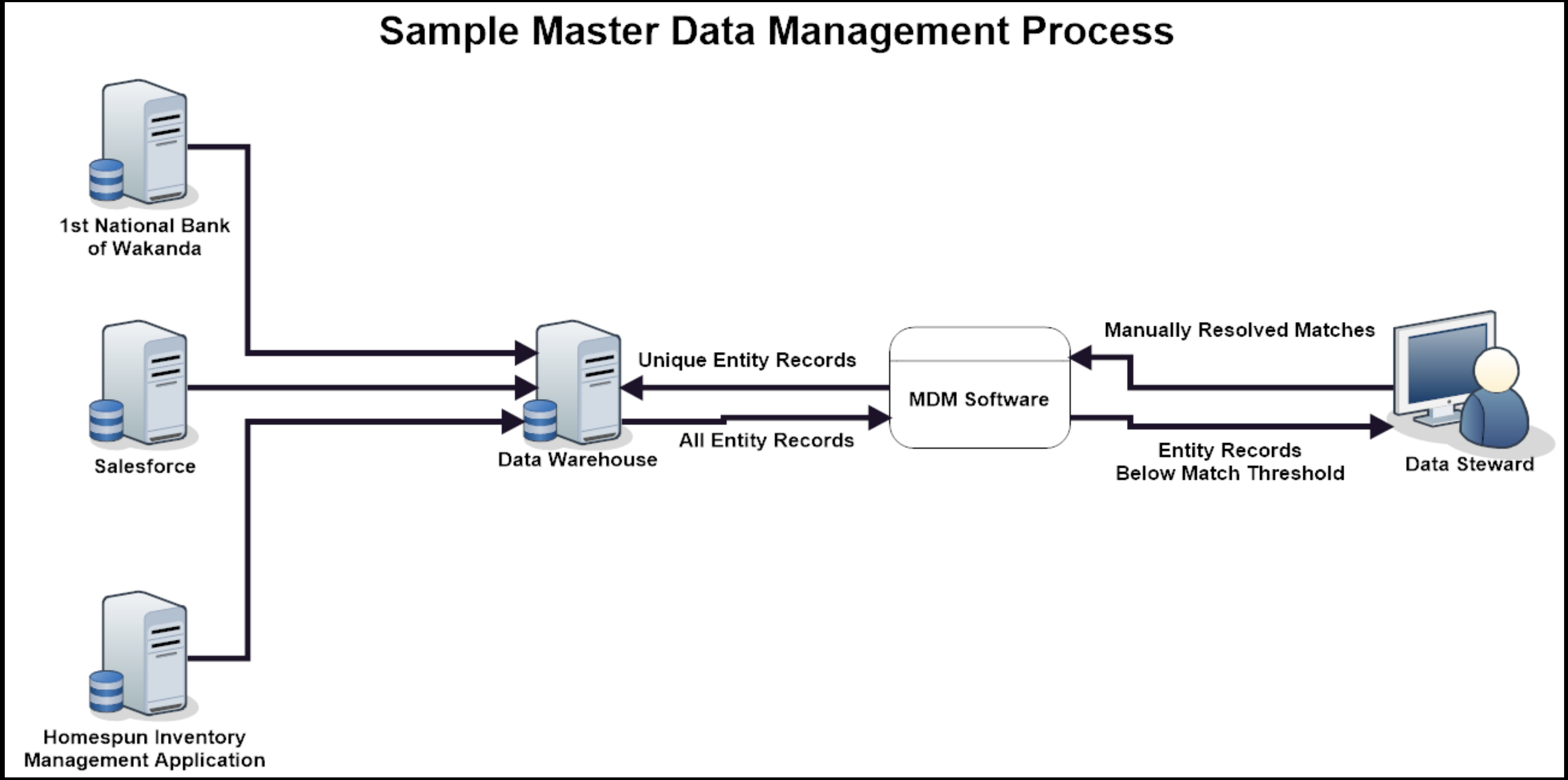
Organizational Considerations

Data Governance

- Data governance is a necessity
- This is how you keep junk data out of the warehouse
- Data governance looks at:
  - Who can access data
  - What they are accessing
  - Keeping data clean
  - Keeping data secure
- You can scale the team
  - You need at least one business user defining what does and doesn’t go into the warehouse

Organizational Considerations

MDM Again


Sample Master Data Management Process
1st National Bank of Wakanda
Salesforce
Homespun Inventory Management Application
Data Warehouse
Unique Entity Records
All Entity Records
MDM Software
Manually Resolved Matches
Entity Records Below Match Threshold
Data Steward

Organizational Considerations

MDM Again

- MDM is not done in isolation
- Every master object needs a human business owner
- You’ll have to convince a manager to let go of someone’s time for a bit
- It needs to be someone who is super familiar with the data
- After an initial clean-up, this should not take up a lot of time

Organizational Considerations
Documentation

- Users of the data warehouse
  - Highly technical data professionals
  - Somewhat technical business analysts and report writers
  - Non-technical business analysts
  - Non-technical report consumers (managers, executives, etc.)
- All these people will need documentation

Organizational Considerations

Documentation

- Types of documents
  - Source to target mapping
    - For engineers
  - Entity Relationship Diagram
    - For engineers, technical analysts, report writers
  - Data dictionary
    - For everybody

Organizational Considerations

How to Hire a Data Engineer

- Options to get this work done
  - Hire a specialized services firm like Mass Street
  - Engage a recruiter
    - You'll get mixed results with this
  - Try to hire them yourself (not recommended)
- There are not a lot of "right" answers in data warehousing
- Candidates should be selected on depth of experience
- I've provided a list of recommended interview questions

Organizational Considerations

How to Hire a Data Engineer

- You’re looking for an engineer, not a DBA
- Having an understanding of the inner workings of a specific RDBMS is not necessary
- Understanding dimensional modeling techniques and efficient ETL is what is required
- They should have a good understanding of all the material presented in this course

Organizational Considerations

Data Engineer Interview Questions

- Caution! Opinions ahead!
- The questions are ok. The thought processes behind the questions is more important.

Organizational Considerations

Data Engineer Interview Questions

Explain the difference between Kimball and Inmon methodologies.

Organizational Considerations

Data Engineer Interview Questions

Discuss the different types of slowly changing dimensions.

Organizational Considerations

Data Engineer Interview Questions

When you are pulling large amounts of data from a transactional source system, how do you keep from locking users out?

Organizational Considerations

Data Engineer Interview Questions

There is a number we look at that is the result of a calculation. Where in the ETL process does the math for that calculation belong?

Organizational Considerations

Data Engineer Interview Questions

We are getting started and this is what we are planning. [List off the things you are planning.] What else do you recommend we do?

Organizational Considerations

Data Engineer Interview Questions

What are the challenges surrounding managing the history of a Type II SCD?

Organizational Considerations

Data Engineer Interview Questions

We have customers stored in three different systems. Because they are entered three different ways, we often cannot tell when a customer is entered twice. We don’t want duplicates in the warehouse. How would you solve this problem?

Organizational Considerations

Data Engineer Interview Questions

How do you decide where to put indexes on tables?

Organizational Considerations

Data Engineer Interview Questions

Describe how you go about transforming a transactional system into a dimensional model.

Organizational Considerations

Data Engineer Interview Questions

We get bank reconciliation data in a feed. The data comes in at the transaction level on a daily basis but it also gives us a month-to-date rollup. How would you go about storing this?

Organizational Considerations Section Conclusion

Course Wrap Up