IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

January 12/14, 2016

Week 1 Presentation

Data Analytics 527
Course Outline

ITM - 527

This course will introduce the student to fundamental concepts in data analysis and implementation methodologies. The course assumes knowledge of SQL and data modeling to run queries and process data sets. The course will introduce/use Excel, SAS, and R to apply analysis techniques and concepts for various use cases. The course will also cover data management concepts e.g., data warehousing, to set the foundation for analytic reporting. Application of analytic techniques will span several industries. Lastly, the course will introduce concepts in Big Data and its applications.

Upon completion of the course, the student will be able to:

- Understand data analysis and data management concepts, theories, and implementation methodologies
- Source, process, and model data sets for analysis
- Use tools and technologies available for data analysis e.g., Excel, SAS, R
- Perform analysis and summarize findings in a presentation
- Understand Big Data concepts

Data Analytics 527
Class Exercises & Readings

Class exercises, assigned as needed, are due ***one week*** from assigned date. The assignment will be posted in Blackboard and submissions will be collected also via Blackboard. **There will be approximately 10 class assignments worth 10 points in the first 10 weeks.**

You will not be able to submit late. No exceptions.

Class readings will be assigned on a weekly basis, as needed. This is an important part of class preparation and augments required and optional book assignments. The expectation is that all required readings and review of weekly presentations are done prior to class.

Data Analytics 527
Midterm and Grading

Midterm grades will be assigned according to grading criteria below after completion of Week 7 class assignments. Midterm grades will be posted by March 9th.

Midterm Grading criteria for ITMD 527 students will be as follows:

- **A** *Outstanding work reflecting substantial effort:* 90-100%
- **B** *Adequate work fully meeting that expected of a graduate student:* 80-89.99%
- **C** *Weak but marginally satisfactory work not fully meeting expectations:* 65-79.99%
- **E** *Unsatisfactory work:* 0-64.99%
- *No Exceptions!*

The Midterm grade for the class will be calculated as follows:

- Class Exercises & Participation **100%**

Data Analytics 527
Final Project and Grading

Finals will consist of a presentation and supporting analysis work. Details for Finals will be distributed March 22nd. Deadline for submission is May 3rd. No late submissions will be accepted.

Grading criteria for ITMD 527 students will be as follows:

- **A** *Outstanding work reflecting substantial effort:* 90-100%
- **B** *Adequate work fully meeting that expected of a graduate student:* 80-89.99%
- **C** *Weak but marginally satisfactory work not fully meeting expectations:* 65-79.99%
- **E** *Unsatisfactory work:* 0-64.99%
- *No Exceptions!*

The final grade for the class will be calculated as follows:

- Final Project **40%**
- Class Exercises & Participation **60%**

Data Analytics 527
Weekly Schedule

Session	Date	Topic	Reading
1	January 12/14		Week 1 topics: Course Overview
2	January 19/21		Week 2 topics: Analysis in Excel I
3	January 26/28		Week 3 topics: Analysis in Excel II
4	February 2/4		Week 4 topics: Analysis in Excel III
5	February 9/11		Week 5 topics: Cluster Analysis/Optimization (in Excel) I
6	February 16/18		Week 6 topics: Cluster Analysis/Optimization (in Excel) II
7	February 23/25		Week 8 topics: Cluster Analysis/Optimization (in Excel) III
8	March 1/3		Week 8 topics: Analysis in SAS I
9	March 8/10		Week 9 topics: Analysis in SAS II
10	March 22/24		Week 10 topics: Analysis in SAS III
11	March 31		Week 11 topics: Big Data Analytics I
12	**April 5/7**		**Week 12 topics: Special Topic: kdb+**
13	April 12/14		Week 13 topics: Big Data Analytics II
14	April 19/21		Week 14 topics: Final Project Workshops
15	April 26/28		Week 15 topics: Final Project Presentations
Finals	May3		Finals Submission

ITM - 527

Data Analytics 527
Project Logistics

- Class meets in Stuart Building Room 111 on Thursdays and Thursdays from January 12th through April 28th except for March 15, 17, and 29th for Spring Break. Total of 14 weeks of classes followed by final project presentations during the last week of class.
- Class time is 10:00 PM ~ 11:15 PM. There will be no time for questions after class.

- Please use office hours or reach me via contact information below. Office hours are by appointment only. I will confirm a room in Perlstein Hall or Stuart Building for office hours.
- Course material communication will be done through Blackboard e.g., class materials, assignment submission, etc.
- Other means of contact:
  - Email: Best means of communication is email. I will respond within 24 hours. Email: sshin17@iit.edu
  - Phone: Use phone calls for emergencies and short requests. Please don't leave long winded questions on the phone. Phone: Phone: (773) 492-1321
- There will be weekly presentations for the class.
- Notifications will be sent for postings in Blackboard but please check regularly for updates.

Data Analytics 527
Course Books

- There is one required book for the class:
  1. Data Smart: Using Data Science to Transform Information into Insight 1st Edition by John W. Foreman. ISBN-13: 978-1118661468. ISBN-10: 111866146X

- However, I list a few *optional* references:
  1. (Optional Reference) SAS and R by Ken Kleinman, Nicholas J Horton. ISBN-13: 978-1420070576 ISBN-10: 1420070576
  2. (Optional Reference) Data Mining: The Textbook by Charu Aggarwal. ISBN-13: 978-3319141411. ISBN-10: 3319141414
  3. (Optional Reference) Microsoft Excel 2013 Data Analysis and Business Modeling 1st Edition by Wayne Winston. ISBN-13: 978-0735669130. ISBN-10: 0735669139
  4. (Optional Reference) Data Science for Business: What you need to know about data mining and data-analytic thinking 1st Edition by Foster Provost and Tom Tawcett. ISBN-13: 978-1449361327. ISBN-10: 1449361323

Week 1 Topic: Conditioning our minds to think data

For Week 1, we will cover some history leading up to present day data analytics environment, challenges and technologies available:

- Defining Data Analytics, Data Mining, and Business Intelligence
- Data Management Concepts
- Definition of a Data Warehouse
- Definition of a Data Marts
- Next for this course

Week 1 Topic: What do we mean by Data Analytics?

For this course, we discuss analysis of data in 3 overlapping disciplines; data analytics, data mining, and business intelligence. You will see in the market that most vendor products choose a discipline to focus their marketing efforts. However, dependent on the solution offering, it can have all or one functional offering.

Theoretically, it's all analysis of data to gather information but using these terms adds *purpose* to the activity. We loosely define the 3 disciplines further but as we progress, we will speak more about *why we use certain techniques for what purpose* rather than noodle over what discipline it belongs to.

- Data Analytics: Further than just visualization of data, the goal of analytics is to support a decision or prove a hypothesis by using quantitative techniques. This confirmatory approach will validate or summarize information to provide the answer.
- Data Mining: Describes examining data sets to identify undiscovered patterns and uncover hidden relationships. This exploratory approach may use sophisticated models to determine its patterns.
- Business Intelligence: Describes data analysis efforts that is focused on answering analytical questions about the business, supports business management processes, or provides views of the business whether it be enterprise or departmental.

Week 1 Topic: More on Data Analytics

We describe more confirmatory, inference driven data analysis activities to data analytics:

- **Descriptive Statistics:** Use of descriptive statistics like means, medians, and standard deviations
- **Graph Distributions**: Using distributions of data to summarize groupings and seek out anomalies in data using visual means e.g., histograms, pie charts, bar charts, etc.
- **String Operations:** Manipulating, parsing, or translating text values to otherwise interpret information that may be coded or concatenated.
- **Math Functions:** Using counts, sums, percentages, and other math functions to validate or summarize data. Same techniques used in string operations can also be applied to numeric values.
- Use of filters and sorts to understand data.
- Logical operations on data e.g., applying finite ranges, <, =, >.
- Calculating derived values and applying conditions on data.
- In general, we answer questions or confirm hypothesis through quantitative means that is not tied to sophisticated models. We tie simple math and view manipulations to this activity.

Week 1 Topic: More on Data Mining

We describe more model driven exploratory data analysis activities with data mining:

- **Outlier Analysis:** Identification of unusual data records, that might be interesting or data errors that require further investigation.
- **Correlations:** Association rule learning or dependency modelling – Searches for relationships between variables. This is sometimes referred to as market basket analysis.
- **Clustering (Segmentation/Summarization):** The act of grouping similar cases together. Discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Classification:** Predicting a discrete categorical value. The task of generalizing known structure to apply to defined categories.
  - **Forecasting/Regression:** Discovering patterns in data that can lead to reasonable predictions about the future. Attempts to find a function which models the data with the least error.
  - **Bayesian:** Use of conditional probabilities to infer an outcome.
  - **Others include use of Neural Networks and Decision Trees**

Week 1 Topic: More on Business Intelligence

We tie Business Intelligence (BI) with a Data Warehouse (DW) solution. Hence, we focus on OLAP functions as BI analysis techniques. We define OLAP as the ability to join discrete sets of data into a dimensional cube structure that is optimized/aggregated for analysis/reporting. With an OLAP cube, you can:

- **Slice**: Act of taking a subset of a cube by choosing a single value for one of its dimensions, creating a new cube with one fewer dimension.
- **Dice**: Creating a subset of a cube for analysis by choosing values from dimensions.
- **Drill Up/Down**: Allowing a user to navigate through the levels of a dimensional hierarchy up (summarized) and down (more detailed).
- **Roll Up**: Summarization of data along a dimension e.g., totals or derived.
- However, BI also encompasses reporting functions similar to the way we defined Data Analytics albeit focused on answering business questions and hypothesis.

Week 1 Topic: Data Analysis Vendor Landscape

One can perform data analytics in spreadsheets as long as one can acquire the data in the right format which is why Excel is so popular amongst the business users. For more sophisticated work, however, we categorize the following top vendors:

Data Mining (*More Business Use*)
SAS Enterprise Miner
SPSS (IBM)
Stata
Tibco Spotfire
Dell StatSoft

Statistical (More *Academic Use*)
MatLab
Mathematica
Minitab
R
SAS JMP

Business Intelligence – OLAP Included
Cognos (IBM)
BusinessObjects (SAP)
MicroStrategy
Hyperion (Oracle)
Informatica

Business Intelligence - Visualization
Tableau
QlikView
Domo
Sisense
birst

Week 1 Topic: Roles in Data Analysis

Business Analyst

- Supports business users
- Serves as subject matter expert for the business domain
- Usually owns a business application, process and/or function
- Performs business analysis mostly in business application or Excel
- Some super users can model and code as necessary

Reporting Analyst

- Supports business users and analysts by building and supporting reports, dashboards, or BI solutions
- Serves as subject matter expert for the reporting domain
- Usually owns the reporting business process or function and in some cases data steward to the data within reports
- Performs data analysis and validations for data processing and management

Quantitative Analyst

- Supports business users and analysts
- Serves as subject matter expert for statistical solutions domain
- PhD types that will use mathematical and statistical programming applications to build and execute model based analysis
- Consumes high volumes of raw data for model data feeds

Data Analyst

- Supports business users and analysts for ad hoc queries or other data related analysis projects
- Serves as subject matter expert for the data domain and typically serves as data stewards
- Performs data analysis, builds reports as necessary, and supports data processing and management tasks

Week 1 Topic: Database Vendor Landscape

Database Software	Language Support
RDBMS: • IBM (Mainframe, DB2, IMS, Cloudant, Informix) • Oracle (Latest 12c) • Microsoft SQL Server (Latest 2014) • SAP (ASE, IQ-columnar) • Teradata (columnar)	• Query: SQL, PL/SQL (Oracle) • API exists for various application build languages • Some recently extended to support JSON, XML
Open Source: PostgreSQL, MySQL, MariaDB Enterprise, Firebird	• Query: SQL and other scripting languages
NoSQL (non-relational database systems with no pre-defined structure): Cassandra, MongoDB, Dynamo	• Query: Cassandra uses CQL and others use various scripting languages • Open distributed file systems. Cassandra resembles RDBMS table structure while MongoDB uses JSON like file structures
In-Memory: KDB, EXtremeDB, MemSQL, SAP HANA	• Query: KDB uses Q and others use direct queries in C/C++, HANA supports Javascripts
Hadoop (a distributed file system) – *More on this when we cover Big Data*	
Specialty Appliances: *Netezza, XtremeData, Greenplum are* optimized for analysis using multi processors, lot of memory, faster networks, large disk space, etc.	

Week 1 Topic: Data Management Components

The following is a view into data management concepts and components for an enterprise. Dependent on the role of the data analyst, one can work in any one of the following areas:

Data Governance			
Organizational Model	Enablement	Standards & Policies	Processes & Procedures

Data Quality			
Profiling / Analysis	Cleansing	Controls	Enrichment / Enhancement

Data Usage		
Reporting	Analytics (OLAP)	Data Mining
Quantitative Analysis	Scorecard / Dashboards	Alerts/ Notifications

Data Management			
System of records	Operational data stores	Data warehouse/data marts	Data movement (ETL/EAI/EII)
Data protection	Metadata management	Reference data management	Master data management

Architecture		
Conceptual	Logical	Physical / technical
Design patterns	Services	Standards

Week 1 Topic:
Data Types – Transaction vs Snapshot

It's important to understand data requirements for analysis as, most likely, you will be defining it for the developers and act as a conduit for the business users' needs. Translating the needs into requirements is not an easy task and in some cases require most time and resources during an implementation. We'll cover data requirements gathering methodologies in the next class. For today, we define what we mean by transaction versus snapshot data:

Transaction Data:

- Records business events e.g., retail purchases, call detail records, bank deposits/withdrawals, insurance claims, stock trades/quotes, etc.
- Usually recorded along with date and timestamp for each transaction.
- Considered raw data or detailed view of data as analysis may be done at a point in time versus looking at each transaction.

Snapshot Data:

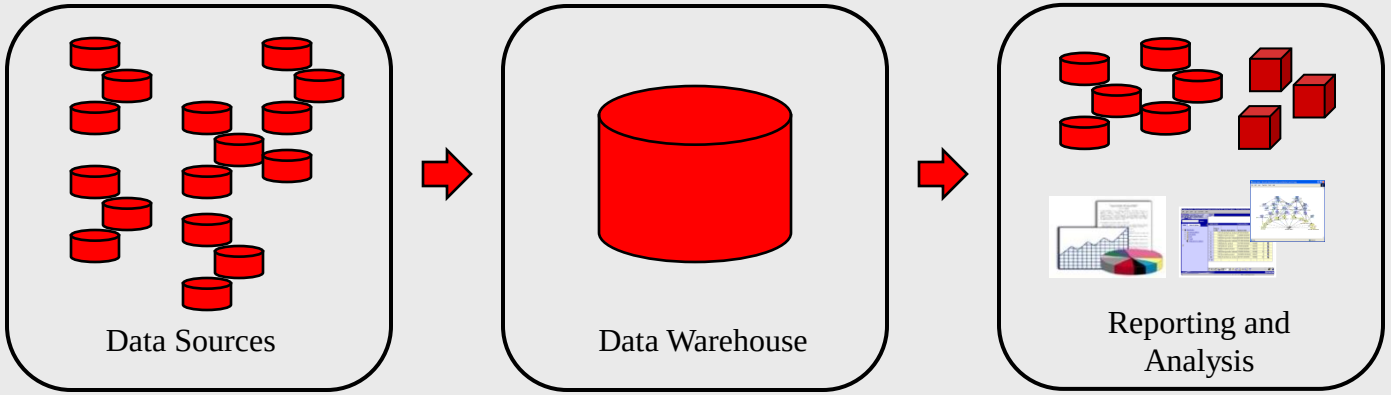
- Records current or past 'state' of a business entity or relationship e.g., customer, account or measures of metric values at a certain point in time.
- Unlike transaction data, a query will typically want to access only one "time instance" of the snapshot data e.g., balance on the account for month end close.
- Multiple snapshots can be used for trending or constructing averages over time.

Week 1 Topic: Data Warehouse: Definition

Most of all, you need data to do data analysis. Most business requirements will require data from multiple data sources with differing data types and varying degrees of data quality. This is where a data warehouse comes in.

Simply put, *a data warehouse is a data repository that collects data from multiple sources into a uniform structure*. Architecture of a data warehouse and its use varies when you start to consider what that uniform structure looks like.

There were two different philosophies on implementing data warehouses in the beginning. Here, we are talking back in the 90s, many years after the concept was first discussed in the 60s.



Week 1 Topic:
Data Warehouse: Inmon versus Kimball

When you start to think about building a data warehouse, one of the first things to consider is how you will model this uniform structure. Of course, this is after you have gathered sufficient business requirements and identified applicable data sources to understand the purpose of the data warehouse in the first place. When we say purpose here, we also mean analytics and its data needs as well.

The two philosophies were:

1. Bill Inmon claimed that this uniform structure is in a 3NF*. Analysis is done off of dimensionally structured data marts** that is produced from this 3NF data warehouse. More of a "top down" approach.

2. Ralph Kimball claimed that this uniform structure is a conglomeration of departmental data marts that may share data through an information bus. Hence, a data warehouse itself may also have a dimensional structure. More of a "bottom up" approach.

*3NF was originally defined by E.F. Codd in 1971. This class assumes that you know basics of data modeling. You will need some data modeling skills to prepare your data sets.

**We will discuss data marts and dimensional data modeling in subsequent slides.

Week 1 Topic:
Data Warehouse: Adoption and Evolution

There was more adoption of Kimball's method as it was considered a "lighter" more practical approach. Investment was easier to justify. This also meant creation of departmental data silos which is another topic all together. He wrote *The Data Warehouse Toolkit* in 1996 which was considered a must read for anyone building a data warehouse at the time.

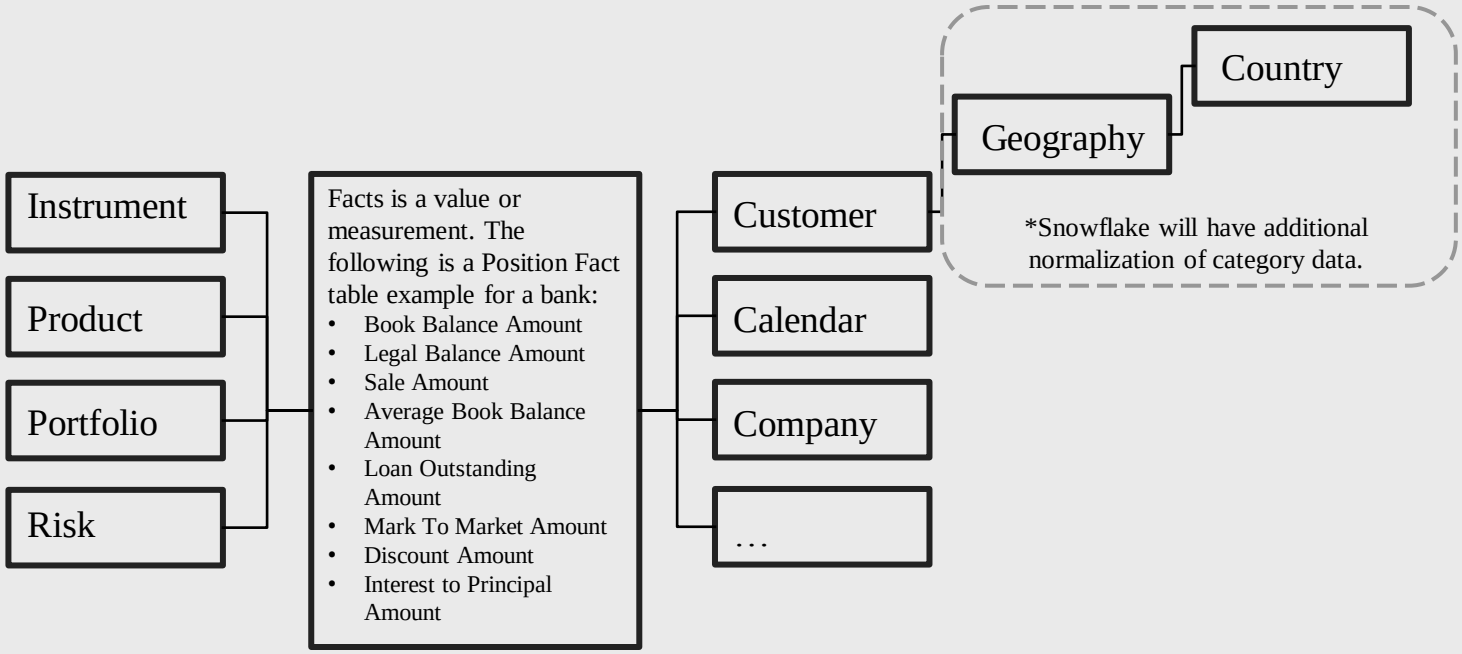
Following Inmon's approach was a bigger investment with many months spent on design which led to lower adoption by business as it was harder to see benefit in a timely manner. Although, theoretically, it made a lot of sense. He published *Building the Data Warehouse* in 1992. 20+ years after first coining the term.

In general, evolution of data warehouses, data analytics, and data management follow technology advancement. Punch cards (used up to about the 70s even though, I still used them in graduate school in the mid 90s) were replaced by magnetic tape (introduced in the 50s) for data storage. Magnetic tapes were replaced with hard disk drives etc. etc.

Today, you can load multiple Terabytes of data into memory for analysis.

Week 1 Topic:
Data Mart: Definition

A data mart is usually dimensional with facts (measures) and dimensions (categories) hence the term, dimensional modeling. We call this data model a star (and/or snowflake*):



A fact table is not normalized but rather designed to respond to queries fast. Dimensional data is normalized and represents a unique descriptive to measures.

Week 1 Topic: Data Solutions Comparisons

- Data Warehouse
  - Source of consistent, integrated, enterprise-wide data
  - Mechanism providing the analytical and decision support needs of the enterprise
  - Data is at multiple levels of granularity, including transaction-level and summarization
  - Data is typically retained for 3-7 years
  - Data may be sourced from the Operational environment and/or the ODS
- Data Mart
  - Mechanism providing the analytical and decision support needs of a business function
  - Data is highly summarized and is usually specific to the business function
  - Data is typically retained for 3-7 years
  - Data may be sourced from the ODS and/or the Data Warehouse
- Operational Data Store (ODS)
  - Mechanism enabling the collection, cleansing, and integration of operational data for population in either a Data Warehouse or a Data Mart
  - Data is typically at the transactional level of detail
  - Data may be retained for short time spans (90-120 days)

Week 1 Topic:
Data Warehouse: Now and Then

Data Warehouse – *the old days:*

- Multi-million $$, multi-year investment
- Batch-oriented
- Limited availability of data – high cost, limited processing power (software and hardware)

Data Warehouse – *new generation*:

- Speed of data has gotten faster e.g., streaming is not a wish, it's a must
- Amount of data being generated have increased e.g., in the petabyte range
- Type of data being processed is more diverse e.g., textural as well as tabular
- Reclaim of archive data ("dark data") – more availability
- More diverse delivery points e.g., handheld devices
- Cost of technology has plummeted

We will discuss influences of these trends on data management when we discuss Big Data towards the end of the semester.

Week 1 Topic: People Management

Qualitative items to consider when embarking on a new project is about People & Adoption in the data space. These measures will have an impact on what a developer does and chooses hence should not be ignored especially at initiation phase:

- Stakeholder Readiness - *understanding opinions*
  - Is the stakeholder sold on the goals, objectives, and value of the project?
  - Does the stakeholder understand or is willing to understand what's involved in implementing the analysis?
- Organizational Readiness – *understanding situational challenges*
  - Are the needed resources available in the organization? Is the organization receptive to external resources, if not?
  - How long does it take for the organization to adopt new technologies?
- Financial Readiness
  - How much and how long will the project cost?
  - What are the financial constraints for the project?
- Data & Technology Readiness
  - What technologies and methods does the organization use currently?
  - Is the data needed for the analysis available? How easily can it be obtained?
  - Is the hardware and software needed for the analysis available?

Week 1 Topic: Understanding data - retail

Data:

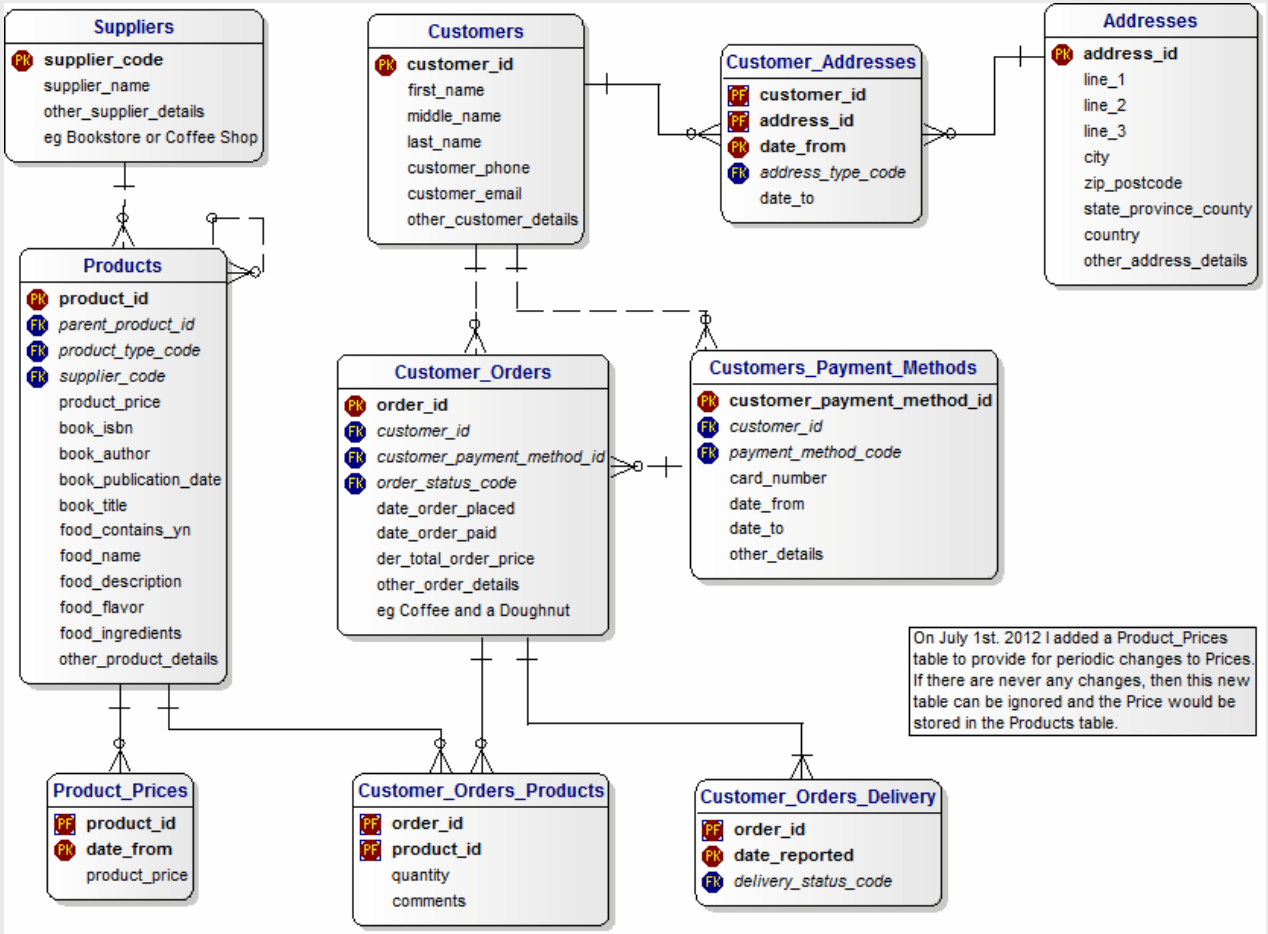
- (F) Transaction data: purchasing/orders, accounts payable, POS, sales projections, warehouse movements, employee shift records, returns
- (D) Product data: consumer merchandise, hardware, software, industrial raw materials, any tangible object or service that can be sold or bought along with SKU, EPC, etc.
- (D) Customer data: name, address, email, phone, demographic, behavioral, financial
- (D) Store/Branch data: location type, address, manager, size/resources
- (D) Others include sales reference data (e.g., account type, personnel) and supplier information

Example Analytics:

- Perform operational analytics to identify economies of scale, inventory management, cash flow analysis, optimal open hours
- Identify patterns, trends, and anomalies in transactions to mitigate risk and report fraudulent activities e.g., receipt fraud (falsified, stolen or reused receipts are used to return merchandise), price arbitrage (using higher priced product tags to return lower)
- Cross sell/Up sell using modeling techniques like market basket analysis or by simple product association e.g., diapers and diaper genie, movies with same actor, etc.
- Launch market campaigns by segmenting like customer groups together according to set criteria e.g., demographic, geographic, income, etc.

Week 1 Topic: Sample retail data model

http://www.databaseanswers.org/data_models/customers_and_orders/:



Week 1 Topic: Understanding data – banking/trading

Data:

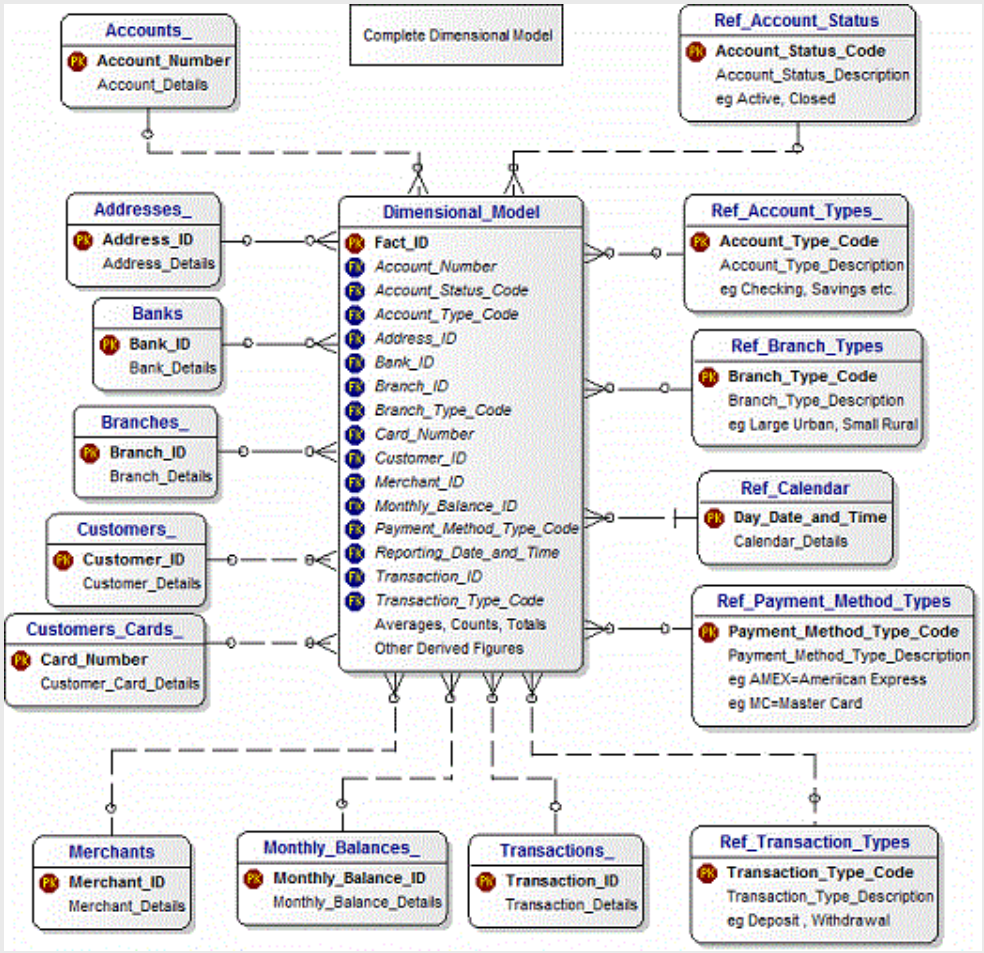
- (F) Transaction data: trades (price, size), quotes (bid price, ask price, bid size, ask size)
- (F) Position (financial) data: amount of securities or commodities held e.g., balances of accounts or portfolios
- (D) Product data (banking): savings, checking, mortgage, credit card, account
- (D) Instrument data (trading) : tradable assets of any kind e.g., securities, cash
- (D) Market data: curves, rates, prices, spreads
- (D) Customer/Obligor/Party data: in addition to the usual CUSIP/SIC/NAIC codes for businesses, SS#, Risk Rating, Obligor (bond issuer, borrower, debtor, contractually/legally obligated entity) Rating

Example Analytics:

- Banking: Compliance with regulatory measures e.g., AML, KYC, Volcker Rule, Basel, etc.
- Trading: Generate best execution outlier reports to identify trades that missed best price using transactions
- Trading: Calculate NBBO (National Best Bid and Offer - this is a regulation that requires brokers to execute customer trades at the best available ask price when buying securities, and the best available bid price when selling securities) matching trades to quotes for a given day

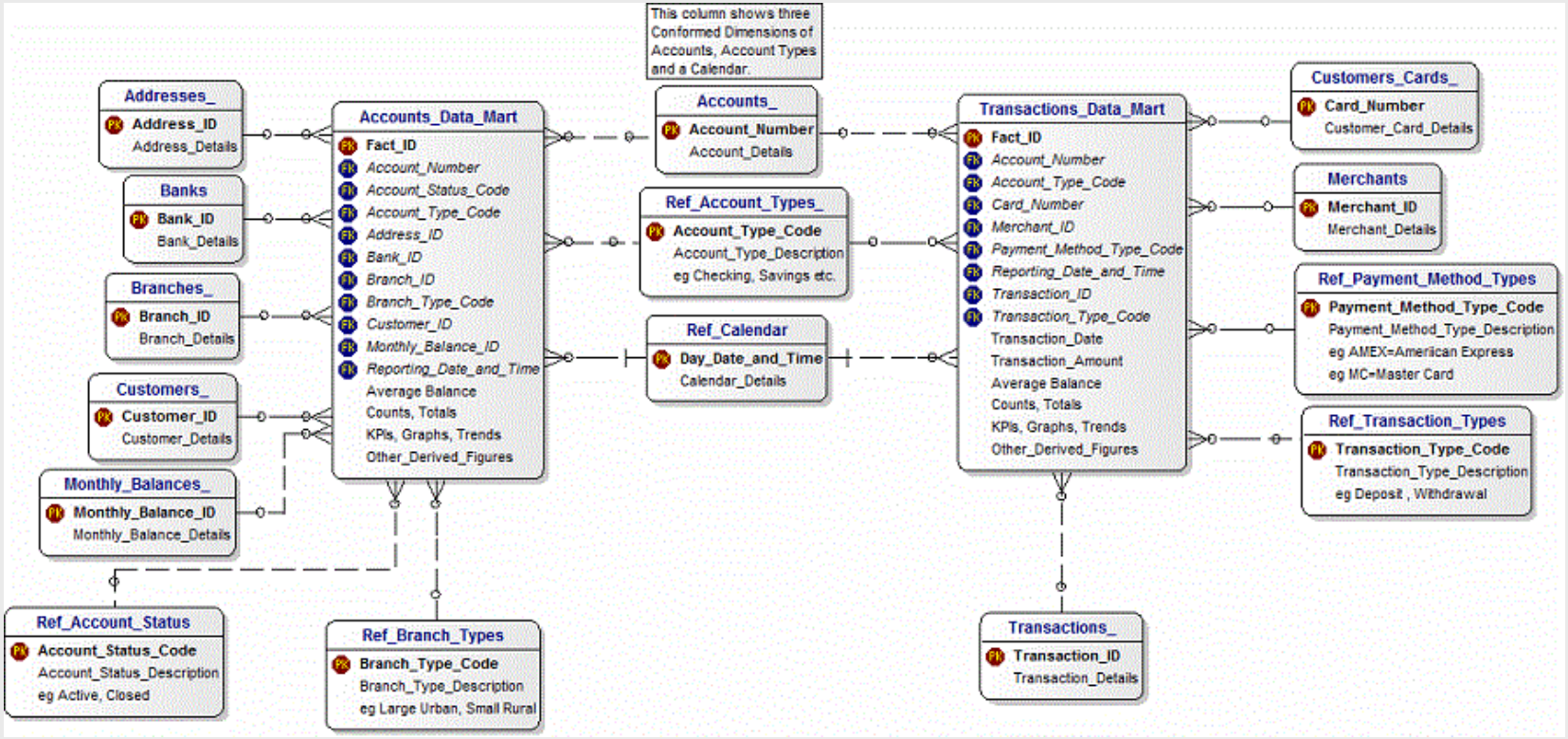
Week 1 Topic: Sample banking data model

http://www.databaseanswers.org/data_models/retail_banks/:



Week 1 Topic: Sample banking data model (cont.)

Another example with more facts and dimensions from http://www.databaseanswers.org/data_models/retail_banks/:



Week 1 Topic: Data Analysis Methodology

A. Inspect - All data is inspected and "cleansed"
   - Records are investigated and fixed where appropriate e.g., outliers, type check, range check
   - Consistent default values are assigned to "missing" data
   - Data validation codes are included in the data where appropriate (e.g. invalid zip code) to avoid/compensate/exclude during analysis

B. Transform - Data is standardized, improved or derived e.g., profitability, scores
   - Promotes consistent analysis using common business rules
   - Reduces analysis "programming" since necessary information is produced prior e.g. calculating months on books, banding score values, computing household product counts
   - Increases understanding of the data

C. Integrate - All of the required data is in one logical structure e.g., transaction, position, account, customer, household, product, instrument, branch
   - Simplifies data access because all data is located in one location
   - Reduces analysis time since all of the information can be retrieved from a single location through a single query

Week 1 Topic:
Data Analysis Methodology (cont.)

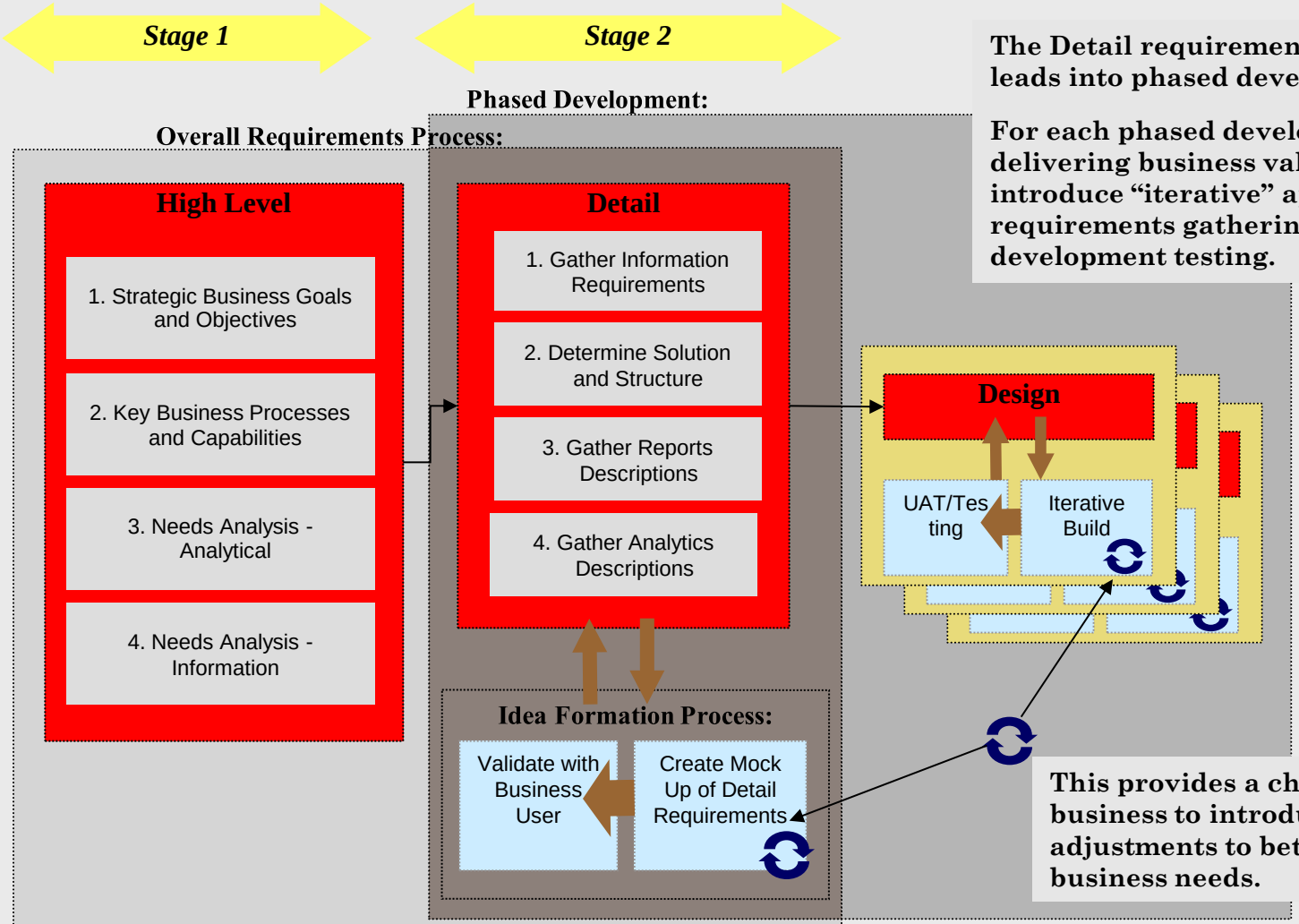
D. Organize BI - Data is stored dimensionally
   - Simplifies analysis by providing an intuitive, business oriented data design
   - Enables pre-stored aggregations (cubes w/drill through to detail) to be easily developed and managed

E. Organize A/DM – Data is stored in modeling specific data structure
   - This could be one de-normalized fact table with select dimensional information included in each row of data
   - Since data mining can only uncover patterns already present in the data, the sample should be large enough to contain significant information, yet small enough to process (dependent on resource capability)

F. Explore - Search for anticipated relationships, unanticipated trends and anomalies:
   - Clustering discovers groups or structures in the data that are similar, beyond the structures known in the data
   - Classification generalizes a known structure to apply to new data, such as classifying a customer as a good or poor credit risk

G. Document - Data definitions and transformation rules are documented and accessible
   - Able to understand the data and information gathered from the data

Week 1 Topic: Requirements Gathering Process



The Detail requirements gathering leads into phased development.

For each phased development delivering business value, we introduce "iterative" approach to requirements gathering and development testing.

This provides a chance for the business to introduce adjustments to better meet business needs.

Week 1 Topic: High Level Requirements Gathering

Strategic Business Goals & Objectives

- Obtain business goals and objectives for the reporting and analytic solutions
- Understand information priorities and challenges
- Identify must-haves from nice-to-haves
- Determine business value metrics to measure investment return

Key Business Processes & Capabilities

- Understand key capabilities and processes needed to satisfy the business goals and objectives e.g., functional and technology resource availability
- Understand key current issues and gaps e.g., timing of data, source system limitations, existing data quality issues
- Obtain current technical documentation e.g. data dictionary, business rules on data transformations

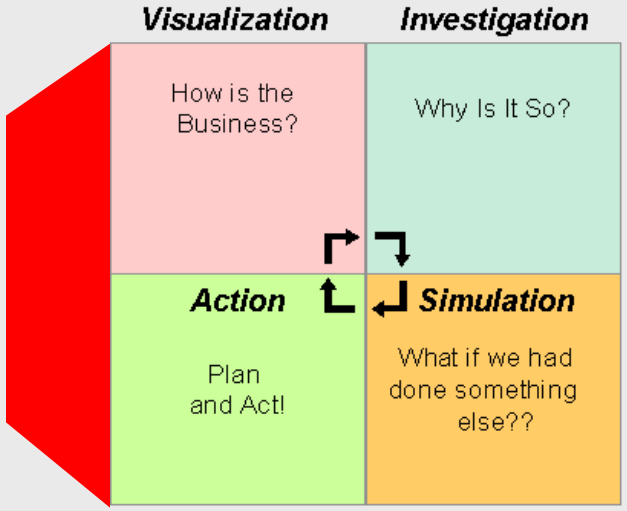
Needs Analysis - Analytical

- Determine analytic capabilities needed to enable processes and capabilities e.g. optimization modeling, segmentation analytics, statistics
- Identify tooling available to perform analytics
- Identify advanced modeling needs beyond tool functionality e.g., customization of modeling techniques in software

Needs Analysis - Information

- Determine information (key data elements & definitions, metrics, calculations, domain values, use cases, etc.) needed to enable processes and capabilities (conceptual and logical models)
- Identify high level information flows (from source to end users), interactions, and dependencies
- Identify tooling available to produce reports
- Identify ad-hoc reporting needs

Week 1 Topic: Understanding that it's iterative



An analytical environment supports not just reporting, but the full range of information usage listed below:

a) Basic Reporting

Periodic reports with the ability to change the report parameters by the users e.g., time period of reporting, metrics reported

b) Ad-hoc Analysis

Ability to access the data in free-form, create new aggregations and report definitions.

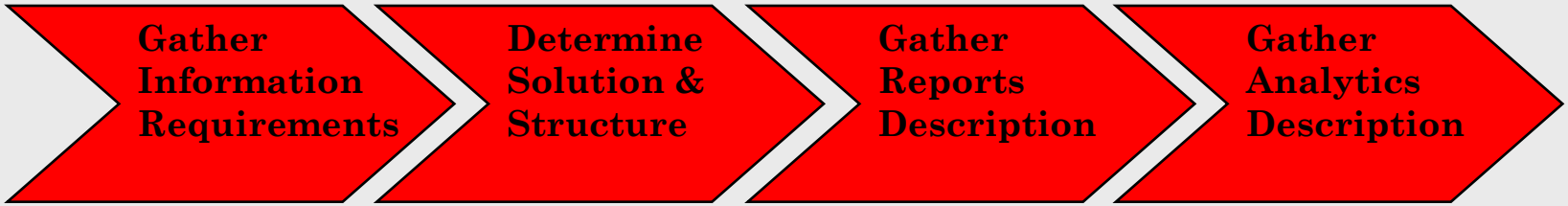
c) Custom Applications

Enterprise application specific functional use e.g., application packages that are used in individual business areas for forecasting, finance, campaign management

d) Intense Analytics

Statistical analysis, modeling and data mining done on an iterative basis to generate segment definitions, offer specifications, credit policies, etc. Requires robust sampling, modeling, scoring and testing capabilities.

Week 1 Topic: Detailed Requirements Gathering



- Identify reporting and analytic solutions and its data content currently in place, inventory, and leverage existing data sourcing solutions (physical models)
- Describe additional data sets. This may require requesting for additional source files, augmenting current source files, access to external data providers, or derivation of data from existing sources.

- Following the data analysis methodology, produce the standardized data model
- Map source data to target data structure and identify any transformations needed
- Perform functional analysis of toolsets
- Determine operational processes and user access needs

- Develop reports inventory including frequency of updates, data content, user access, query capability, owner, approval process, distribution scheme, etc.
- Prioritize reports for a phased rollout
- Identify whether a current report will satisfy requirements or a new report is needed

- Develop analytic solution inventory including analytic data needs, analytic capability (modeling, statistics, OLAP, etc.), parameters and conditions, etc.
- Understand and determine metrics to meet business goals
- Prioritize needs for a phased rollout

Week 1 Topic:
Excel Basics – fundamentals

Following lists some functions that is assumed:

1) Math Functions:
   a) basic math functions: *, /, +, - and use of = to define a formula for the cell.
   b) sum, min, max, count, counta, average, if, mod, trunc, standev
2) Data Functions:
   a) data formatting, conditional formats
   b) data importing, data connections
   c) concatenate, left, right
   d) cell references: use of $
   e) text to column, removing duplicates, data validation, showing all formulas
   f) sorting, filters, replace, find
3) Views & Printing:
   a) charts and graphs from table data
   b) freeze, split, hide, print areas, print formats
4) Navigation/Selection/Short Cuts
5) Vlookup and Hlookup

Week 1 Topic:
Excel Basics - cheat sheet

Navigation/Selection:

a) CTRL+Tab – rotate through open excel workbooks
(ALT+Tab flips through open application)

b) CTRL+PageUp/PageDown – move left/right inside workbook – between sheets

c) CTRL+Arrow – goes to end of cells of "continuous formatting"

d) CTRL+Home – top-left corner of worksheet (unless "freeze panes" is activated)

e) CTRL+End – bottom-right of worksheet

f) Holding shift key + arrow/mouse click (or CTRL+arrow) – selects continuous cells

g) Holding down CTRL key + mouse click – selects multiple cells (copying and pasting will eliminate any un-selected cells between the copied/cut selections)

h) CTRL+[(or CTRL+]) – navigate to cell linked to/from cell

Shortcuts:

a) CTRL+C = copy

b) CTRL+X = cut

c) CTRL+V = paste

d) CTRP+P = print

e) CTRL+S = save

f) CTRL+Z = undo

g) CTRL+Y = repeat (also F4 for multiple repeats)

h) CTRL+A = select all cells

i) F4 = repeat and quick absolute reference toggle (when in cell)

j) CTRL+1 = cell formatting menu

Week 1 Topic:
Excel Basics – further review items

We will review the following functions in more detail:

1) Pivot Tables using zip code and select use of Data Analysis (add-in)
2) Solver (add-in) using concession example from Data Smart

Good and Free references on Excel:

1) http://www.excel-easy.com/data-analysis.html
2) http://joelgrus.github.io/thinking-spreadsheet/
3) http://chandoo.org/wp/

Week 1 Topic:
Class Readings for this week

1) Read PDFs in ~/Readings/Week 1/ on data analytics and data management – *In Blackboard*
2) Read Data Smart Chapter 1. This covers Excel basics so any additional readings and reviews in Excel will be useful. Get familiar with Excel and its functions including Pivot Tables.

Week 1 Topic: Class Assignment

Worth 10 points:

1. Choose a data scientist.
2. Research their background and research area.
3. Determine applications of their works/research in current environments. Site any data analysis models, techniques, and/or theories used in their works.
4. Write a one-page (8-10 font, single spaced) report in word.
5. Site references.