Business Intelligence involves the acquisition, processing and transformation of raw data into meaningful information for business analysis purposes.  Business intelligence finds patterns in large volumes of data that helps make strategic decisions and optimize businesses.  Modern business intelligence software utilize a combination of database processing, analytics and statistics to aid in operations research.

Business intelligence systems are fed data from data warehouses. A data warehouse contains massive , time series data for decision support.Breslin notes that specialized software extracts data from operational databases, then summarizes, reconciles, and manipulates it. Then the data is ready to be stored in carefully designed relational database tables in the data warehouse. (p. 2)

I have used Oracle databases in the past at  work but these are mainly used for persisting user metadata. I have yet to use OLAP technologies and data warehouses.

References:

Breslin, M.(2004). Data Warehousing Battle of the Giants:Comparing the Basics of the Kimball and Inmon Models. Business Intelligence Journal. p 1-15.

subscription based services have come to dominate the market nowadays. I find that data mining is very interesting . Amazon for example maintains a user history of browsed items which help its machine learning programs make suggestions to shoppers, all this is possible after a diligent application of data processing and mining.  Analytics  is very interesting and has a lot of applications for data driven operations.

Kimball's introductory chapter  does a good job of outlining the data warehouse lifecycle. Kimball notes the project planning and business requirements definition should be conducted in tandem and these feed into the technical architecture design , dimensional modeling and BI app development stages in which products are selected, physical design and ETL is performed and BI apps developed. Once an iteration of these processes is complete the DW/BI system are deployed and monitored for growth and maintenance.

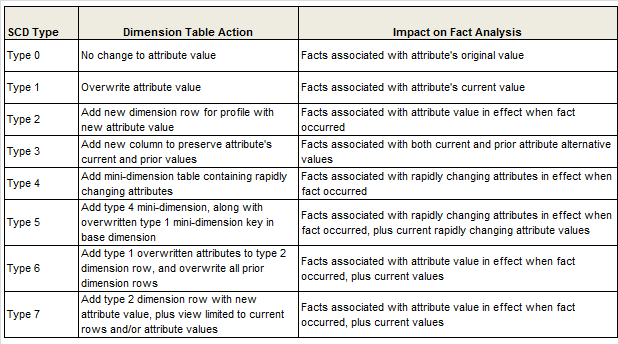
Kimball also gives an overview of how the data warehouse is the foundation for business intelligence whose primary role is query support.  Kimball stresses on the fact that we need to structure dimensional models around business processes and their data sources to promote consistent data views for all observers. Kimball emphasizes the need for identifying the enterprise data warehouse bus matrix which represents key business processes and dimensionality. Fact tables are highlighted as containing metrics from business processes or measurement events. Dimension tables are noted as having descriptive attributes used for constraining, grouping or labeling a query.  The star schema is introduced as instantiated dimensional models and OLAP cubes are discussed as stored dimensional models.

I found the restaurant analogy quite intuitive for understanding how raw data is prepared and presented to consumers. The ETL system is the back room kitchen where raw data is prepared the kitchen is designed for integrity and high quality and throughput is critical. Kimball introduces 34 subsystems as a checklist for anyone constructing the ETL kitchen.  Kimball also focusses on the restaurants front room where data is presented and notes that its based on preferences expressed by patrons.  Dimensional models are the front room deliverables in Kimballs model and facilitates ease of use and query performance.

References:

Kimball, R.(2008). The Data Warehouse Lifecycle Toolkit :Second edition. Wiley Publishing

Week 2



Week 3

The overall design of the dimensional model includes several important tasks such as :

1. Identifying the business process- we need to pick a process that embodies measurable physical actions in the real world.

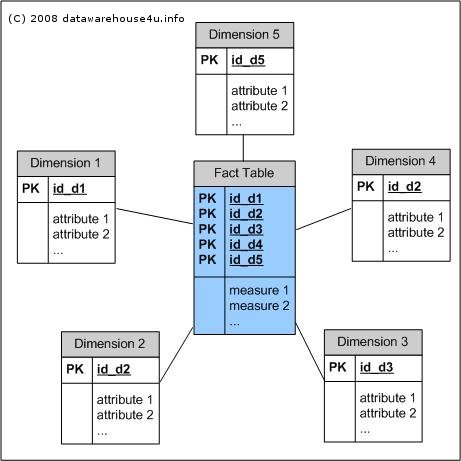
2. Establishing the grain of the business process.- we need to define the measure or grain of our recording process this can be periodic snapshot, transaction or accumalating snapshots

3. Identifying the dimensions- we need to identify the dimensions that describe the data qualitatively

4. Identifying the facts. - we need to identify the numeric data that identifies the facts for the model

Part of generating a enterprise data bus matrix we need to identify the business processes as rows and the dimensional attributes as columns.  It becomes important to keep in mind the simple star schema design of the fact table and dimension model integration.

I think that identifying the grain of the business process and selecting conformed dimensions are crucial tasks as they establish the consistency and granularity that makes historical summarization feasible.



References:

Kimball, R. and Ross, M. (2013). The data warehouse toolkit. Indianapolis, IN: Wiley.

Week 4:

The issue with ETL today is that it assumes that the data it receives is clean and consistent. ( Eckerson & White,  p.15)

In reality ETL  tools cannot assess the consistency or accuracy of source data and cannot handle specialized cleansing routines such as name and address scrubbing. To meet these challenges ETL tools partner with data quality vendors and integrate specialized data cleansing routines within ETL workflows.

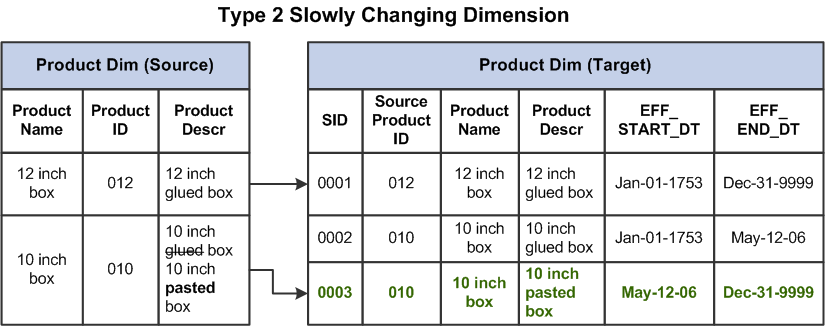
To avoid code, load and explode syndrome good data integration platforms provide support for data cleansing and profiling functionality. Data profiling tools allow us to generate a Rosetta Stone for our source data. We can identify the range and values of each field as well as column and row dependencies.

Data cleansing tools validate and correct business rules in source data. These tools provide specialized functions for scrubbing name and address data, parse and standardize format of names and addresses, verify the names and addresses against third party database and remove duplicates.

References:

Eckerson, W. and White,C. (2003). TDWI Report Series: Evaluating ETL and Data Integration Platforms. 101communications. Chatsworth, CA: TDWI reports

Week 5:



In dimension tables rows are uniquely identified by a sequential surrogate key. The use of surrogate keys promotes data integrity and performance. Since surrogate keys are system generated sequenctial numbers, these are non intelligent and non business related. With surrogate keys we can carry out more efficient joins, have smaller indexes. Keys in operational systems can be managed more effectively. Since we have multiple sources having a surrogate key ensures that there is no overlapping of source system primary keys.

Possible disadvantages would be having an additional load for ETL system as we choose a meticulous and resource intensive process to generate surrogate keys out of overlapping primary keys of source systems.

With slowly changing dimensions we have to consider if we want to

1. overwrite the dimension attribute(type1) - this might cause an issue with some aggregation reports and they probably need to be rebuilt.

2. add a new dimension row(type 2)- we have versioning of the change using version effective and end dates and a current indicator for the current record.

3. add a new dimension attribute(type 3)- we use a new attribute to record the changes.

Week 6:

Aggregations should be designed with performance criteria in mind. Using one base star schema many aggregates can be defined. Some dimensions could become lost, shrunken or collapsed due to this.  In general when defining aggregates

* Aggregates must be stored in their own fact tables; each distinct aggregation level must occupy its own unique fact table.
* The dimension tables attached to the aggregate fact tables must be shrunken versions of the dimension tables associated with the base fact table
* The base atomic fact table and all of its related aggregate fact tables must be associated together as a “family of schemas” so that the aggregate navigator knows which tables are related to one another
* We must force all SQL created by any end-user data access tool or application to refer exclusively to the base fact table and its associated full-size dimension tables

References:

Lawrence Corr

* *Aggregate improvement*
* *Lost, shrunken, and collapsed*

Ralph Kimball

* *Aggregate navigation with (almost) no metadata*

http://megaslides.com/doc/5206579/aggregation

Aggregation allows summarization in different levels for management to review and make decisions about a set of data. Most aggregations are pre calculated and loaded into the data warehouse during off hours. The roll up aggregation is a common form of summarizing data. Taking a SUM of daily sales and aggregating them to a monthly sales figures is one example of aggregation report.

Week 7:

Another database that is coming up is MemSQL which is an in memory database platform. It makes possible real time streaming and analytics on data. According to the founder, Eric Frenkiel ,"The Big Data market itself is rapidly shifting to real-time processing using main memory, and software like Apache Spark, Kafka, and MemSQL are enabling businesses to create and analyze real-time streams of data."

The companies clients include company’s Pinterest, Comcast, Akamai and Kellogg’s. MemSQL can ingest streams, process transactions, and perform analytics simultaneously in a single database, enabling businesses to analyze changing datasets. With immediate access to both real-time and historical data, MemSQL opens new opportunities for revenue and improved customer experiences.

References:

http://www.memsql.com/product/

Data mining is known as the process via which data analysts can search for patterns in a data set. Once a data set is cleaned, conformed and stacked it can be loaded to intelligent dashboards where specific computations are performed on this data. We can then obtain a timeline of the historical data and see patterns and trends within it that help us develop scenario based insights.

Today along with data processing and analytics software such as SAS, we have visualization tools such as Tableau. There is integration between the Tableau report server and the data sources which allows us to mark a  feed to the visualization tool. We can then get a rendered graph or chart of the data which makes business sense to the data analysts.

Data mining ties in with machine learning and statistical modeling in that it helps establish the direction that the data shapes the business towards. We can then create business processes in that direction and develop a infrastructure to cater new features by offering data more efficiently to clients in the road map. Amazon for example uses data mining to compute buying trends over a shopping sales event. They then use machine learning to develop a list of historical favorites that users purchased and recommendations for future purchases.

References:

http://www.theatlantic.com/technology/archive/2012/04/everything-you-wanted-to-know-about-data-mining-but-were-afraid-to-ask/255388/