We spent several weeks now talking about how to collect and organize data, and how it can be easily used and maintained.

How long do we keep data?

Only a few decades ago, we didn’t.
* Some companies shredded records as soon as they felt it was safe to do so.
* Other companies kept their written records indefinitely, but they were *really* hard to use and, essentially, unavailable.
* Private individuals only bothered to keep selected bills until the next one arrived, and retained their tax records for the legally required period.
* Banks and governments kept their data a longer, but it, too, was difficult to access.

Recall the raiders of the lost ark final scene

This began to change when businesses realized that their data have value beyond the few years after it was collected. One of the earliest examples of this was in grocery stores. When grocery stores learned that by offering simple perception discounts through the use of a loyalty card they could have the customer assist them in managing their inventory. Furthermore, they learned more about the buying habits of specific customers to the point that the graphics studies could be made by the "home office" to better manage the distribution of goods. For example:
* if marketing a Wheaties box with a swimmer on the front sold better in one part of the country then the same box with a football player on it, the company could ensure that the right product was in the right market.
* Similarly, a company might notice that sales of certain stimulant beverages increased at certain times of the year near colleges.

Lots of data

Data that doesn’t fit well into the rows and columns of a spreadsheet, prolly won’t fit on your computer anyway.
Doug Laney in 2001 defined these most common characteristics of Big Data: The 3 V's

* Volume
* Velocity
* Variety

This BIG data idea is revolutionizing the customer experience as well as providing greater insights into companies’ businesses.

- Customer advantages
- Business advantages
- R&D advantages

Big data is characterised by more than the 3 v’s

Jules Berman’s book on Big Data identified several others:

* Goals - small data gathered for a specific goal, Big data may have a goal in mind but other goals may emerge
* Location - SD is usually in one place, and one computer file - BD is in many files, computers, and locations
* Data Structure - SD is usually highly structured, while BD can have many formats, from a variety of sources and disciplines (even languages and formats), be linked to other resources, and be totally unstructured as with natural language (English, Chinese (with all the character issues that brings), ancient greek, etc.)
* Data prep - SD is usually prepared by the end user for their own purposes. With big data the data is usually prepared by one group of people, i analyzed by others, and used by a variety of other people with goals that might not have been in mind when the first two groups were working on the data.
* Longevity - small data is usually kept for a short period of time during the project and thereafter. In research the data may be retained for a bit longer, at least until the publication process completes or the grant ends. Research and Academic data is more often shared with others than business data. Thus, it will “live” longer. For BD, the data is kept for much longer, perhaps forever, in hope of learning more.
* measurements - small data is usually measured using one standard technique employing specific measurements normal for that technique. Since the sources of big data are by definition varied it is much more likely that the data will have different measurements, units or means of collection, and even cultural diversity issues. These differences have to be resolved during the ETL phase.
* Stakes - with SD, the loss or corruption of the data might be a setback, but can usually be recreated. With BD, the origins are so diverse, and often deleted once migrated to the data warehouse, any corruption of loss is likely unrecoverable and potentially catastrophic for the business.

So now everyone wants to keep their data indefinitely.

What problems does this raise?

- Storage space
- ease of access
- accuracy over time
- protection and privacy concerns

it’s no use keeping the data if you can’t use it, right?

Many companies just kept the data anyway, in hopes that it would prove valuable someday.

This has many implications, particularly when it comes to personal information.

So the new business of’s business intelligence, known as BI, was born.

To support this traditional databases needed many new features.

* The more data you put in the database the slower it gets. Thus, they needed the capability to export data from the main
A new process was introduced to migrate the operational data, the data used for day-to-day business, and the data warehouse, where data was stored.

* online vs. offline

This process is referred to as the ETL process: extraction, transformation, and loading.

**ETL**

**SLIDE 14–2**

1. Extraction - The selection of what data should be extracted (removed) from the operational database is not as easy as it sounds.
   1. you might not want all of the data since some of it would have no relevance in the future. For example, the time of day that particular sale was made may not be as relevant as the date that the sale was made.
   2. the data by becoming a variety of database, from potentially different vendors and with varying schemas
   2. Once the data is selected by the data management team, it often requires additional processing before it is loaded into the data warehouse.
      1. Filtering—where the interesting or relevant data is selected and the other data ignored or deleted.
      2. Transformation—where the data is transformed into a different format, unit (for example, converting English to metric measure), normalized, etc.
      3. Integrated—where data from a variety of databases is combined for use later in the data warehouse. This is somewhat like a JOIN where the common attributes of two tables hard disk carded keeping only one or maybe even none.
      4. classified—grouped and arranged subject to some common characteristics
      5. Aggregated—where data representing a specific attribute is combined, such as being summed or averaged, before it is placed in the data warehouse. This may also include external (noncompany) data, such as stock market information, industry-specific trends in information, and even news feeds. Whatever information that might shed light on the corporate data or make it more usable in the future.
      6. Summarized—where various techniques are used to reduce the data to a representative sample as opposed to the entire set.

   1. Finally the data is added to the data warehouse
      1. further integration is done between various data sources of necessary
      2. the data may be focused in particular subject areas. For example, individual transactions may be less important in the data warehouse then revenue summaries.
      3. The data warehouse may also make sure that the incoming data has a timestamp associated with and where it came from for potentially future use.
      4. The data warehouse purpose is to retain data. Thus, it needs to be well secured, nonvolatile, and in some cases backed up. The backup question is problematic, since the whole point here is to reduce the size of the operational database by removing inactive or less presently interesting data from that database and moving into the data warehouse. As a result the data warehouse becomes enormous over time. Backing up this resource, however valuable, becomes increasingly difficult and expensive. For these cases further reduction summarization is often employed.

**SLIDE 14–3**

Once the data is added to the data warehouse, it becomes essentially a read-only source. Periodic updates from the operational data are still allowed, but no updates to existing warehouse data is permitted.

The data flowing into the warehouse is often in different formats. Operational data tends to be highly structured, while data warehouse information may be structured partially structured or completely unstructured such as news feeds and English text.
The fact that the data warehouse is essentially read-only leads to many optimizations which are helpful given the size and cost of such installations.

In addition to the data from operations, the data warehouse maintains corresponding metadata which identify and define all of the data elements in the warehouse. Information like the source of the data what transformations were made to it with what other data was integrated or aggregated, the relationships between data groups, And of course the history of a particular data element.

Finally, since the usual purpose of the data warehouse is to guide business strategy, it is considered a business expense at the entire company must bear.

**Analysis**

All this analysis often leads to improved revenue, more efficient operations, and better customer service.

It can also lead to unexpected discoveries:

* A study of electrical power usage in hopes of optimizing delivery discovered unusual patterns of higher than normal use and one part of the city. Worried that this data indicated a problem, the company doing the big data study investigated. They were surprised to find not a failing transformer or faulty wiring, instead they found apartment buildings where a large number of residents had turned to growing their own crops in doors utilizing grow lamps 24 hours a day.
* A study of how employees responded to stock options led to the realization that if you were using them in such a way that it indicated a likely insider-trading fraud.

**BI**

The goal of all this activity is business intelligence.

The resulting data warehouse is used in the decision-support process.

By guiding the business to make better choices, more data may be requested and incorporated into the data warehouse.

The historical and nonvolatile nature of the data warehouse enables long-term evaluation of business decisions, further guiding future strategy.

The past behavior documented in the data warehouse is also used to predict future behavior and outcome with a high degree of accuracy.

- **Master data management (MDM)**
  - a collection of concepts, techniques, and processes for the proper identification, definition, and management of data elements within an organization.
  - The main goal is to provide a comprehensive and consistent definition of all data within an organization.
  - It ensures that all company resources (people, procedures, and IT systems) that work with data have uniform and consistent views of the company’s data.

- **Governance**
  - BI provides a method for controlling and monitoring business health and for consistent reporting (as to the government or other regulatory body) and for consistent decision-making.
  - Business intelligence often drives the key performance indicators (KPI) which are quantifiable numeric or scaled measurements that assess the company’s effectiveness or success in reaching its strategic and operational goals.
  - after a company’s main strategic and tactical and operational goals of them defined, these KPI’s are defined and monitored using the information flowing into the data warehouse.
  - Examples include:
    - year-to-year measurements of profit by line of business, store, product, sales due to promotions, employee performance
- earnings-per-share, profit margin, revenue per employee, expense per employee, etc.
- human resources - number of applicants to job openings, their skills, educational level, academic performance, and job experiences. Rates and trends of employee turnover and employee longevity.
- Education - test performance, graduation rates, population, publication rates, teacher evaluation scores, college acceptance rates, etc.

Data warehousing

The data warehouse is defined as “an integrated, subject-oriented, time-variant, nonvolatile collection of data supporting decision-making”.

* Integrated. The data warehouse was centralized, consolidated database that integrates data derived from the entire organization and from multiple sources within and external to the company potentially with diverse formats.
* Subject-oriented. Data warehouse data are arranged in optimized to provide answers to questions from diverse functional areas of the company. Data warehouse data are organized and summarized by topic, such as sales, marketing, finance, distribution, and transportation.
* Time-variant. In contrast to operational data, which focus on current transactions, warehouse data represents the flow of data through time. Because data in a data warehouse constitutes a snapshot of the company history is measured by its variables, a timestamp is crucial.
* Nonvolatile. Once data entered the data warehouse, they are never removed. Goes the data represents the company’s history, the operational data, which represent the near-term history, are always added to it. Thus, the data warehouse is always growing. Multiple terabytes and multiple system or more will be required.

One of the data modeling techniques used in data warehousing is called the star schema. The basic star schema has four components:

* facts.
* These are numeric measurements that represent a specific aspect of the business or inactivity.
* For example, sales figures, production efficiencies, revenue data.
* Facts are usually stored in a fact table that is in the center of the star schema.
* The fact table contains facts that are linked through their dimensions mature to find next.
* Dimensions.
* These are characteristics that provide additional perspectives to a given fact.
* Dimensions are important because decision-support data are almost always viewed in relation to other data.
* For instance, sales might be compared by product from region to region and from one time period to the next.
* In effect, dimensions are the magnifying glass to which you study the facts.
* Attributes.
* Each dimension table contains attributes.
* These are often used to search, filter, or classify facts.
* Dimensions provide descriptive characteristics about the facts through their attributes.

* SLIDE 14–4, 14–5, 14–6
* attribute hierarchies.
* Some attributes within dimensions will be ordered within a well-defined hierarchy. This provides a top-down data organization that can be used for two purposes: aggregation and drill down/rollup data analysis for example:

* SLIDE 14–7

Data analytics
this is a special area of business intelligence that uses a wide range of mathematical, statistical, and modeling techniques with the goal of extracting knowledge from data.

Analytics or all the rage nowadays. Data science and analytics for a great place to be.

Analytics discover characteristics, relationships, dependencies, or trends in the organization’s data, and then explains the discoveries and predicts future events based on the discoveries.

It is generally thought of as a spectrum: discovery to explanation to prediction.

There are two closely related areas of analytics:

* explanatory analytics focuses on discovering and explaining data characteristics of relationships.
* Predictive analytics focus on predicting future data outcomes with the highest degree of accuracy possible.

For example, what would next month’s sales be based on a given customer promotion?

The former explains the past and present, while the latter forecasts the future based on the explanatory analytics.

Analytics, or data mining, lead to a variety of discoveries and predictions:

* 65% of customers did not use a particular credit card in the last six months or 88% likely to cancel that account.
* If the applicant is under 30 and has an income less than 25,000 and a credit rating less than three and a credit amount greater than 25,000 then the minimum loan term is 10 years.

Social media has been a very rich source of big data.

* Mining Twitter feeds for product trends
* Spotify and its ability to understand what you want
* Yelp, Angie’s List, etc.
* and all the other reputation based data.

SLIDE 14–8

Tim Berners-Lee, the inventor of the Web, summarized it this way:

It's difficult to imagine the power that you're going to have when so many different sorts of data are available.

1. Lecture notes based on texts by Coronel, Widom, Ullman, Jukic, and Silberschatz.