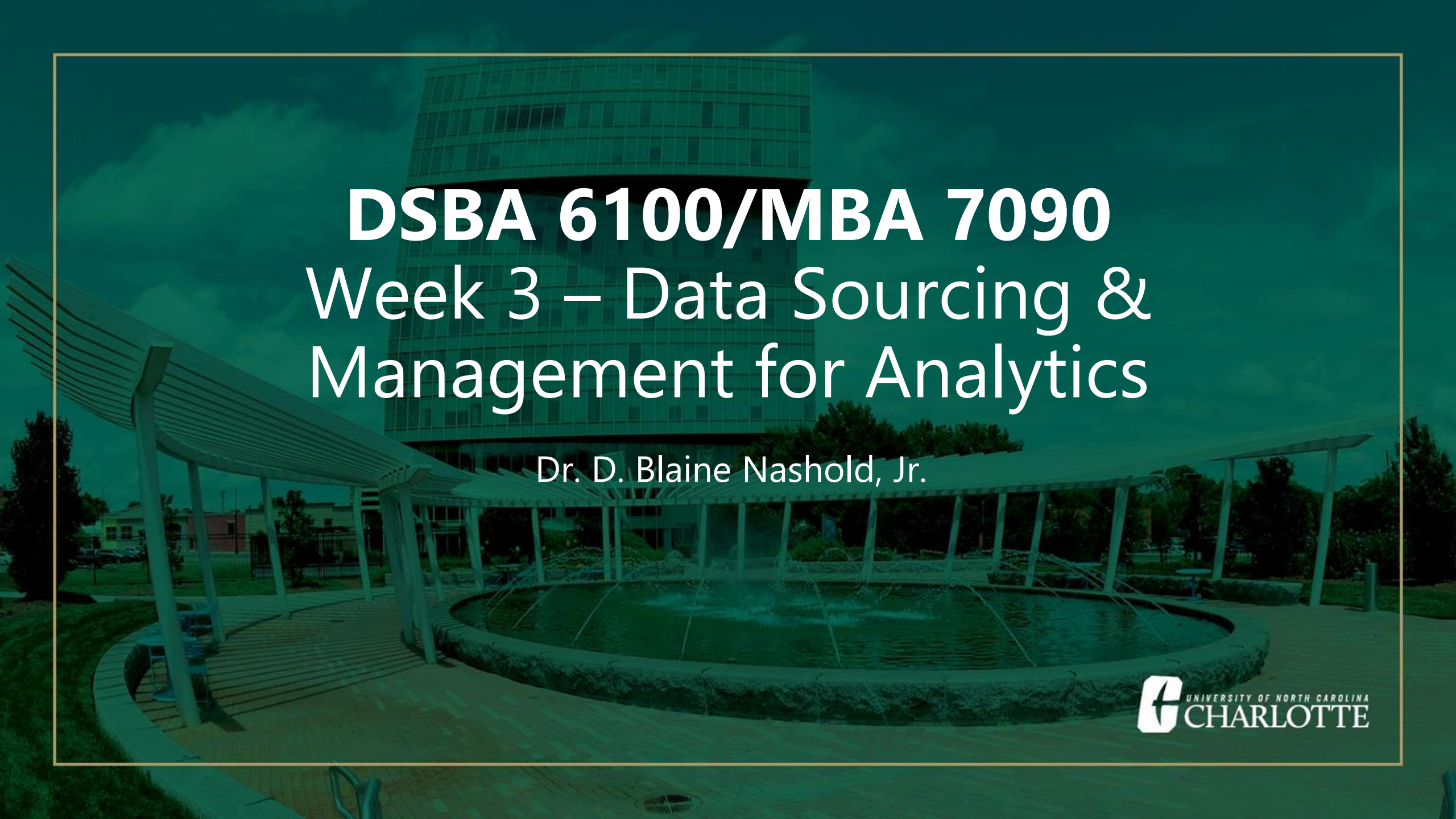


DSBA 6100/MBA 7090
Week 3 – Data Sourcing & Management for Analytics

Dr. D. Blaine Nashold, Jr.

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

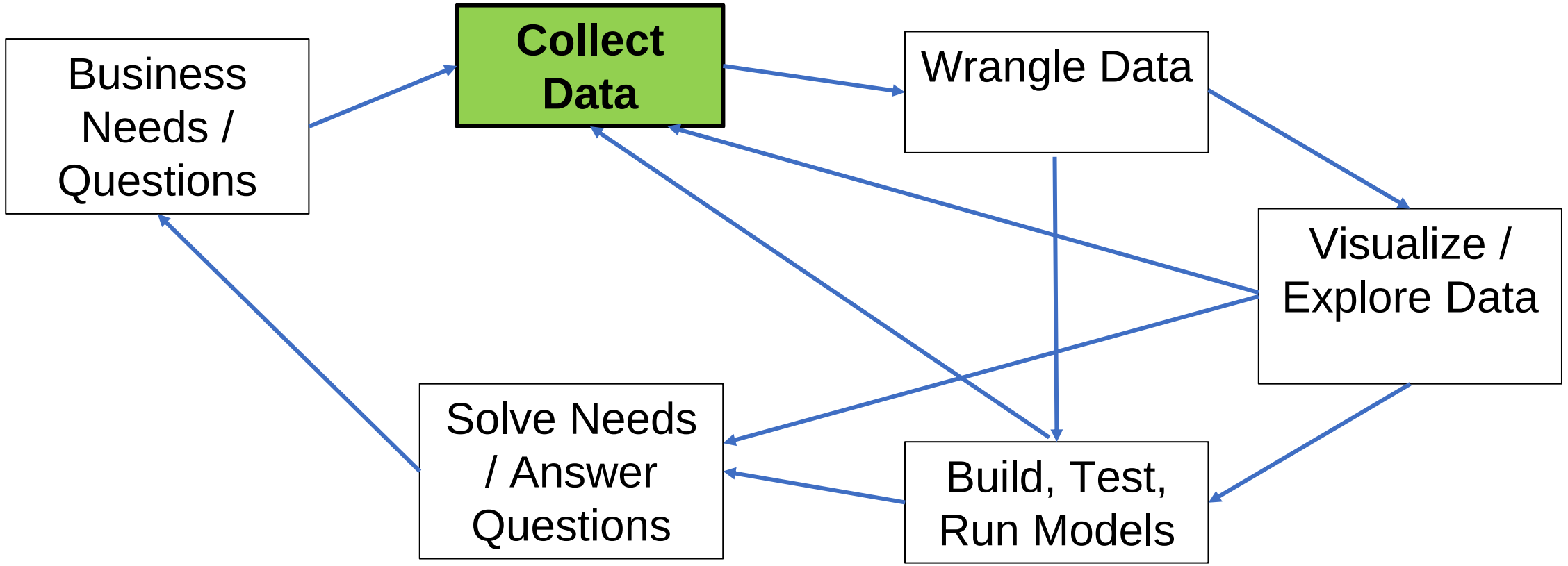
Discussion Outline

- Data Sources for Big Data Analytics
- Data Warehouse
- Hadoop and MapReduce
- Data Lake

Some questions to think about data!

- Why is data management important for analytics?
- What decisions are made (or supposed to be made) regarding data for analytics?
- Can’t the data needed for analytics simply be bought?
- Which areas of business should be involved in data management phase?

Big Data Analytics as a Process



Traditional Sources of Data

- Organizations have long relied on their internal systems to support decision-making
  - Point-of-sale systems
  - Enterprise systems (ERP, CRM)
  - Helpdesk tickets
  - Loyalty card data
- Much of this data is structured (or at least only the structured part was used for reporting and analysis)
- Much of it is delivered by data warehouses & data marts

New Data Sources – Internal Big Data

- Organizations can now use analytics on their internal semi-structured/unstructured data
- Internal semi-structured & unstructured data
  - Email
  - PDF documents
  - Web log
  - Internal blogs
  - Photos and videos
  - RFID data

New Data Sources – External Big Data

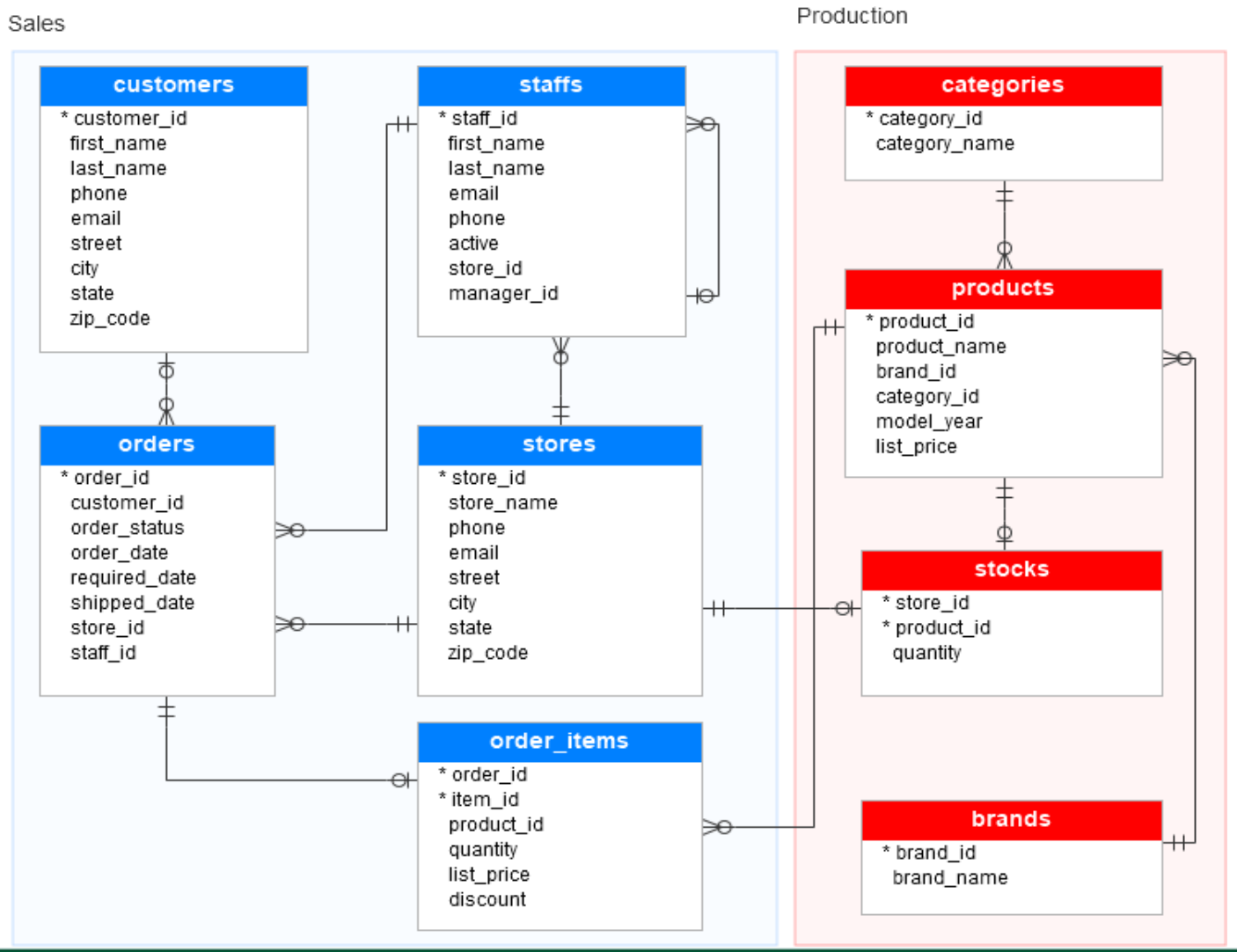
- Data from data aggregators
  - Acxiom
  - LexisNexis
  - MX
- Social media data – collected through APIs or third-party providers
  - GNIP (https://www.gnip.com/) – commercial twitter data provider
- Open data
  - Weather data
  - Census data
  - Federal/State/Local govt public data

Internet of Things & Big Data

- IoT sensors allow data to be gathered at unprecedented level and detail
- Emerging as an exciting new opportunity for many industries
- IoT data is being used to build intelligent response systems

Next → Storing & Accessing Structured Data

Databases to Store Structured Data



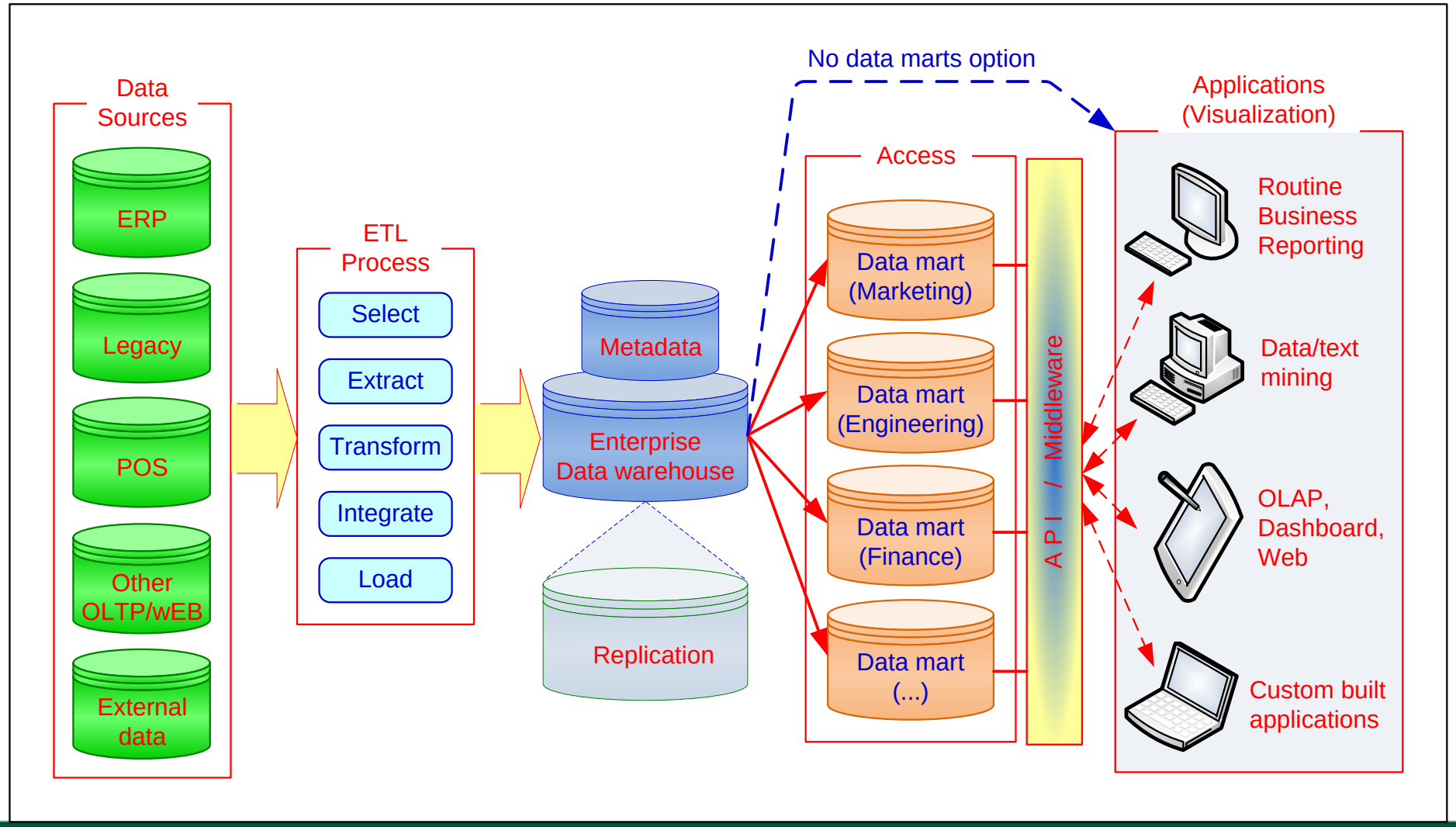
What is a Data Warehouse?

- A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format

- "The data warehouse is a collection of integrated, subject-oriented databases designed to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time"

Source: "Business Intelligence and Analytics" by Sharda et al., Pearson Education, 2015

A Generic DW Framework



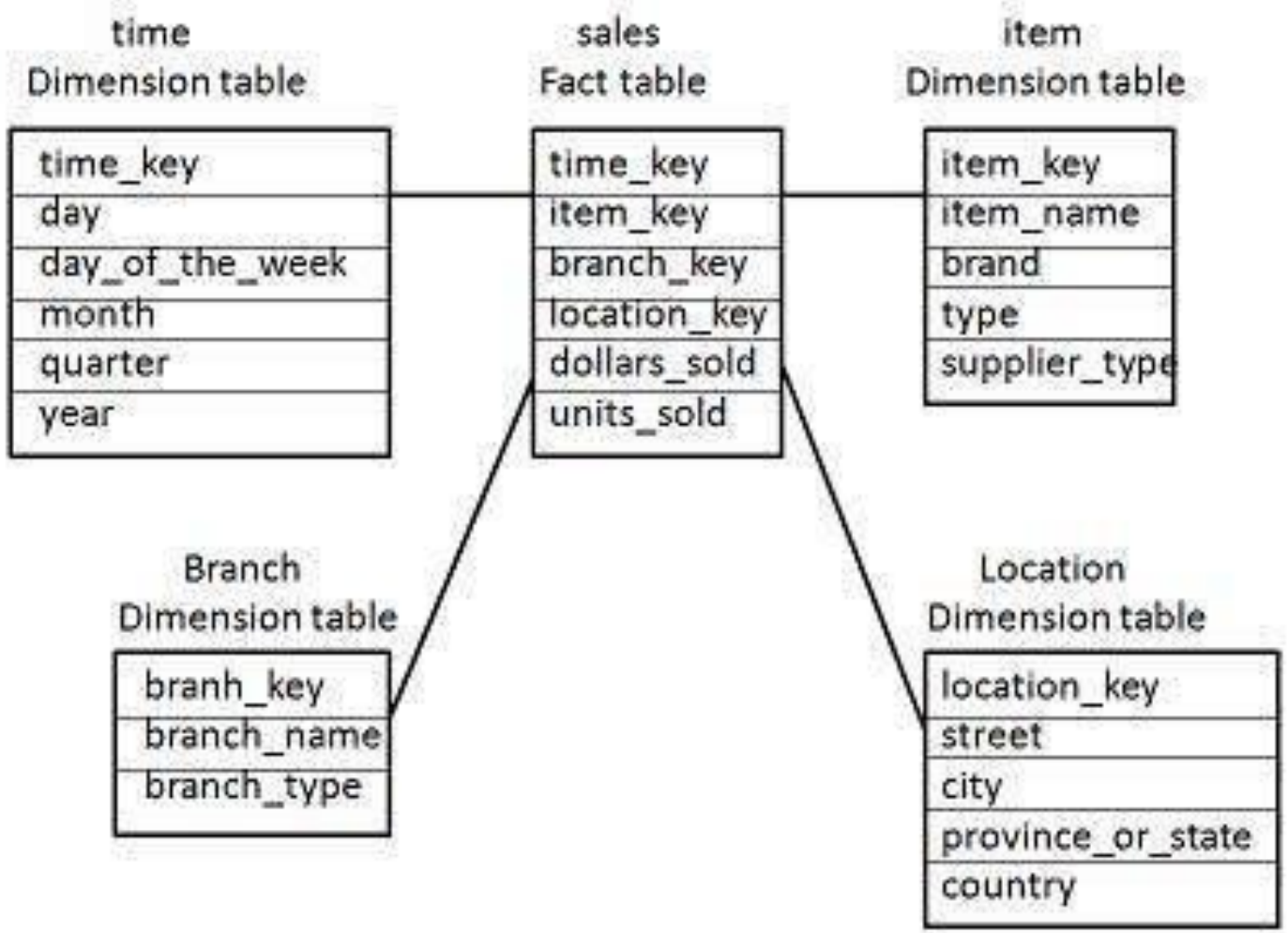
Source: “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015

Characteristics of DWs

Online Source: https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG-GUID-452FBA23-6976-4590-AA41-1369647AD14D

- Subject oriented
- Integrated
- Time-variant (time series)
- Nonvolatile
- Summarized
- Not normalized
- Makes use of Metadata

A Sample Data Warehouse Store



Source: https://www.tutorialspoint.com/dwh/dwh_schemas.htm

Multidimensionality

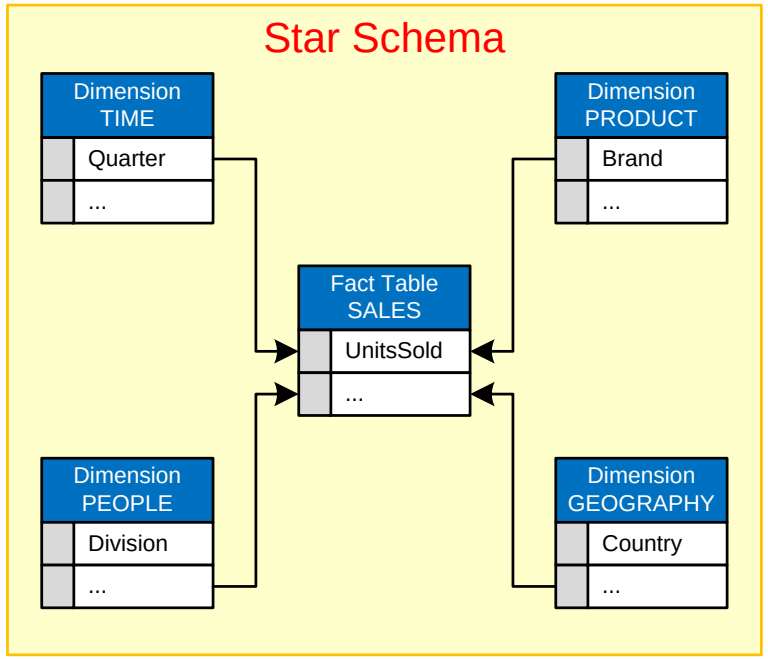
The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)

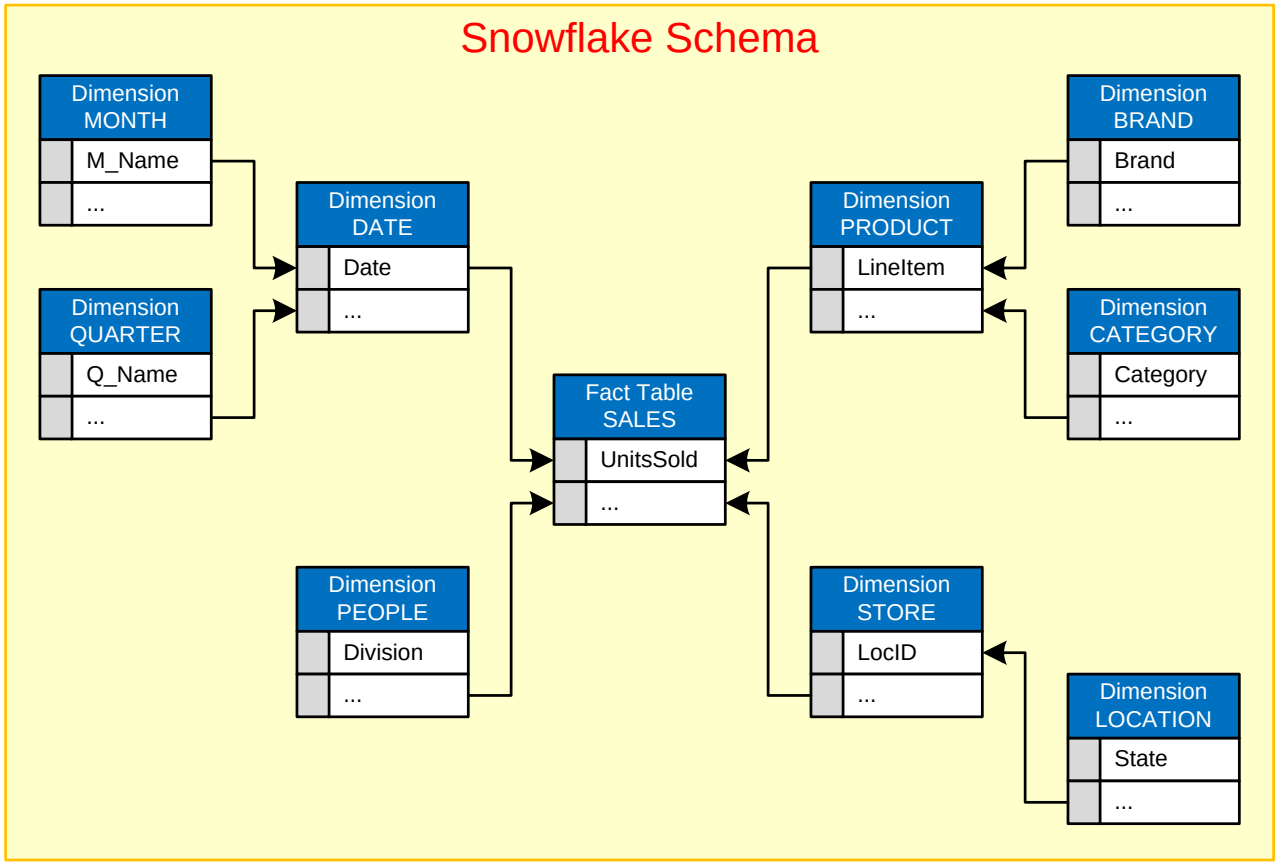
- Multidimensional presentation
  - Dimensions: products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry
  - Measures: money, sales volume, head count, inventory profit, actual versus forecast
  - Time: daily, weekly, monthly, quarterly, or yearly

Representation of Data in DW

- Dimensional Modeling
  - A retrieval-based system that supports high-volume query access
- Star schema
  - The most used and the simplest style of dimensional modeling
  - Contain a fact table surrounded by and connected to several dimension tables
- Snowflake's schema
  - An extension of star schema where the diagram resembles a snowflake in shape

Star versus Snowflake Schema in DW





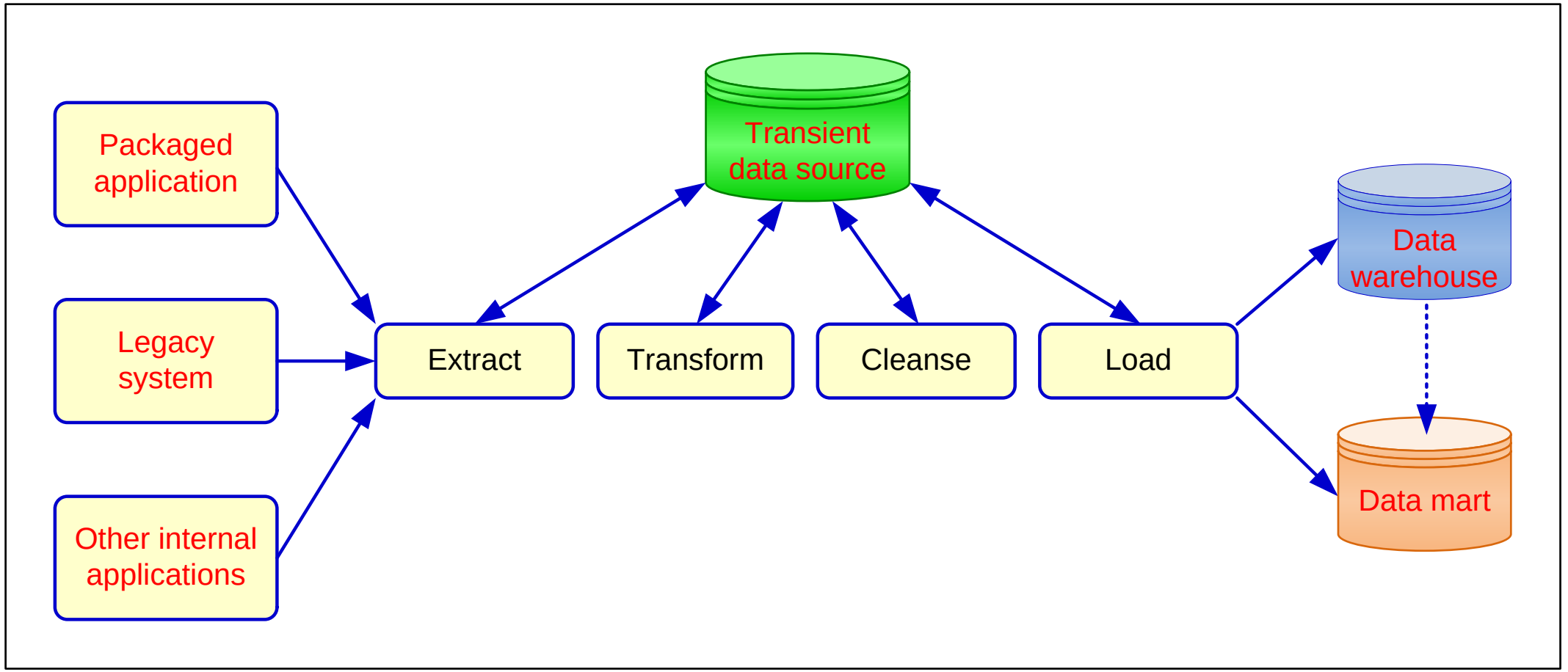
Source: “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015

The ETL Process

Online Source: https://docs.oracle.com/cd/B10500_01/server.920/a96520/ettoverv.htm

- **ETL = Extract Transform Load**
- ETL is a standard process to load data into a data warehouse according to the schema selected
- During extraction, desired data is identified and extracted from different sources
- More data than necessarily must be extracted since it is not possible to identify all questions of interest on the data in advance
- Some transformations take place during this process (e.g., transforming a unit of measurement to from English to metric)
- After extract & transform, the data is transported to a target system for storage

Data Integration and the ETL Process



Source: “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015

Next → Storing & processing unstructured data

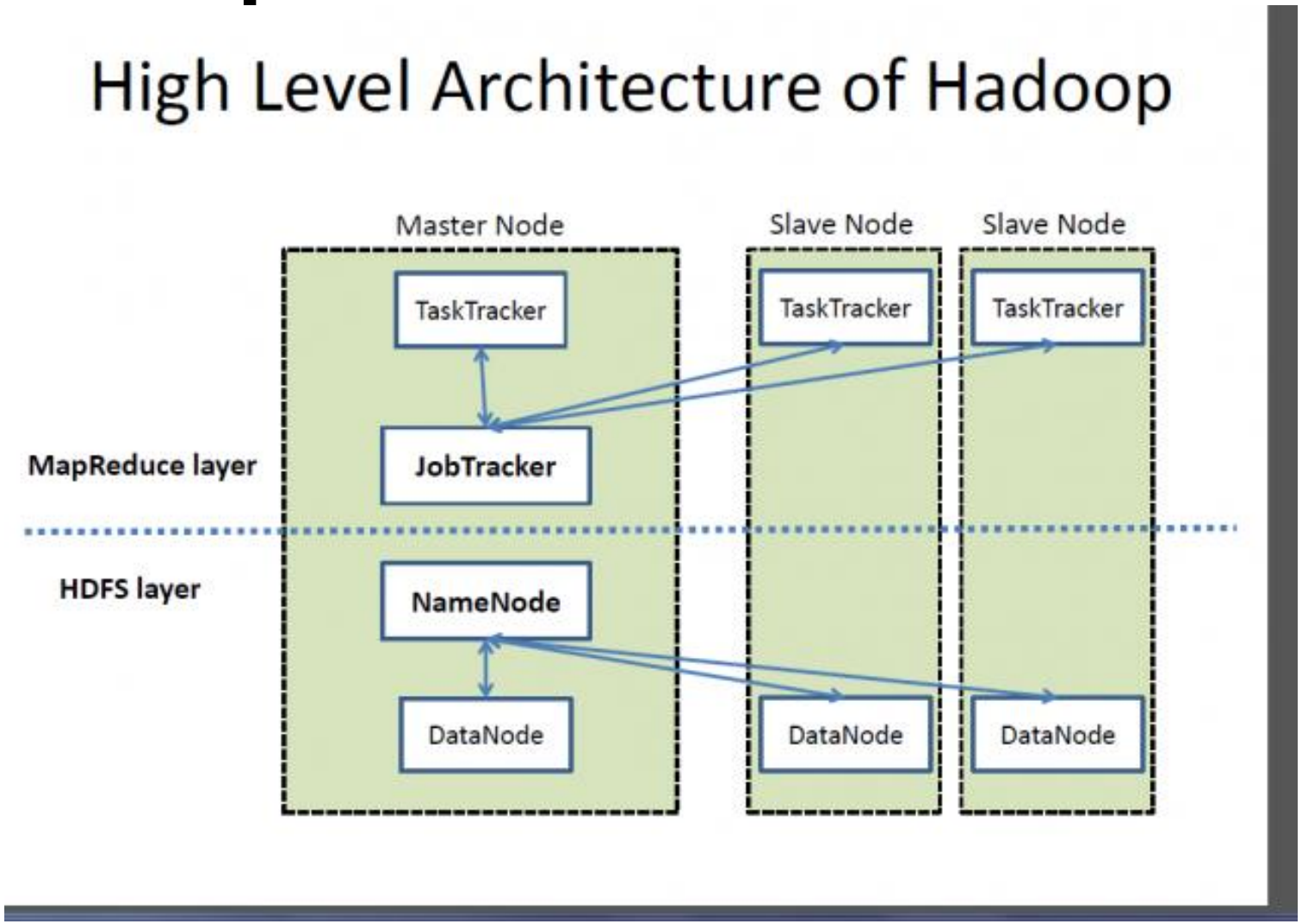
Big Data Technologies - Hadoop

- Hadoop is an open-source framework for storing and analyzing massive amounts of distributed, unstructured data
- Originally created by Doug Cutting at Yahoo!
- Hadoop clusters run on inexpensive commodity hardware so projects can scale-out inexpensively
- Hadoop is now part of Apache Software Foundation
- Open source - hundreds of contributors continuously improve the core technology
- MapReduce + Hadoop = Big Data core technology

Big Data Technologies - Hadoop

- How Does Hadoop Work?
  - Access unstructured and semi-structured data (e.g., log files, social media feeds, other data sources)
  - Break the data up into “parts,” which are then loaded into a file system made up of multiple nodes running on commodity hardware using HDFS
  - Each “part” is replicated multiple times and loaded into the file system for replication and failsafe processing
  - A node acts as the Facilitator and another as Job Tracker
  - Jobs are distributed to the clients, and once completed, the results are collected and aggregated using MapReduce

Hadoop & MapReduce



Source: https://opensource.com/life/14/8/intro-apache-hadoop-big-data

Big Data Technologies - Hadoop

- Hadoop Technical Components
  - Hadoop Distributed File System (HDFS)
  - Name Node (primary facilitator)
  - Secondary Node (backup to Name Node)
  - Job Tracker
  - Slave Nodes (the grunts of any Hadoop cluster)
  - Additionally, Hadoop ecosystem is made up of a number of complementary sub-projects: NoSQL (Cassandra, Hbase), DW (Hive), …
    - NoSQL = not only SQL

Big Data Technologies - MapReduce

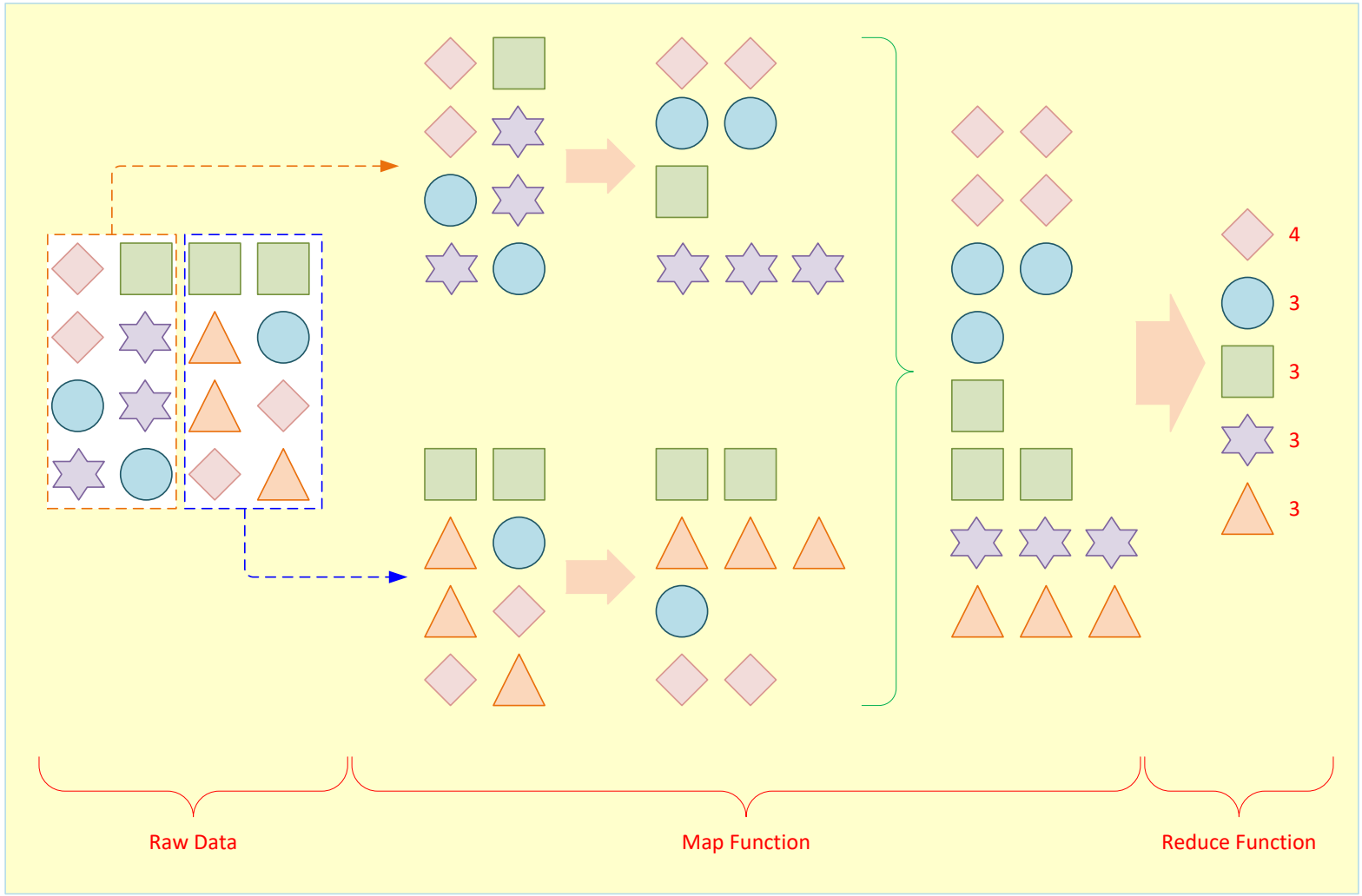
- MapReduce distributes the processing of very large multi-structured data files across a large cluster of ordinary machines/processors
- Goal - achieving high performance with “simple” computers
- Developed and popularized by Google
- Good at processing and analyzing large volumes of multi-structured data in a timely manner
- Example tasks: indexing the Web for search, graph analysis, text analysis, machine learning, …

How MapReduce Works – An example

- MapReduce first reads input file into HDFS which splits it into multiple pieces
- Each split data is processed by a map program (e.g., group by shape and color)
- Multiple map programs run in parallel on the nodes of the hadoop cluster
- Each map program returns its computation results on its part of the data (e.g., # of shapes and colors)
- The reduce program collects the outputs from the map programs and reduces to the final output (i.e., aggregates it as required)

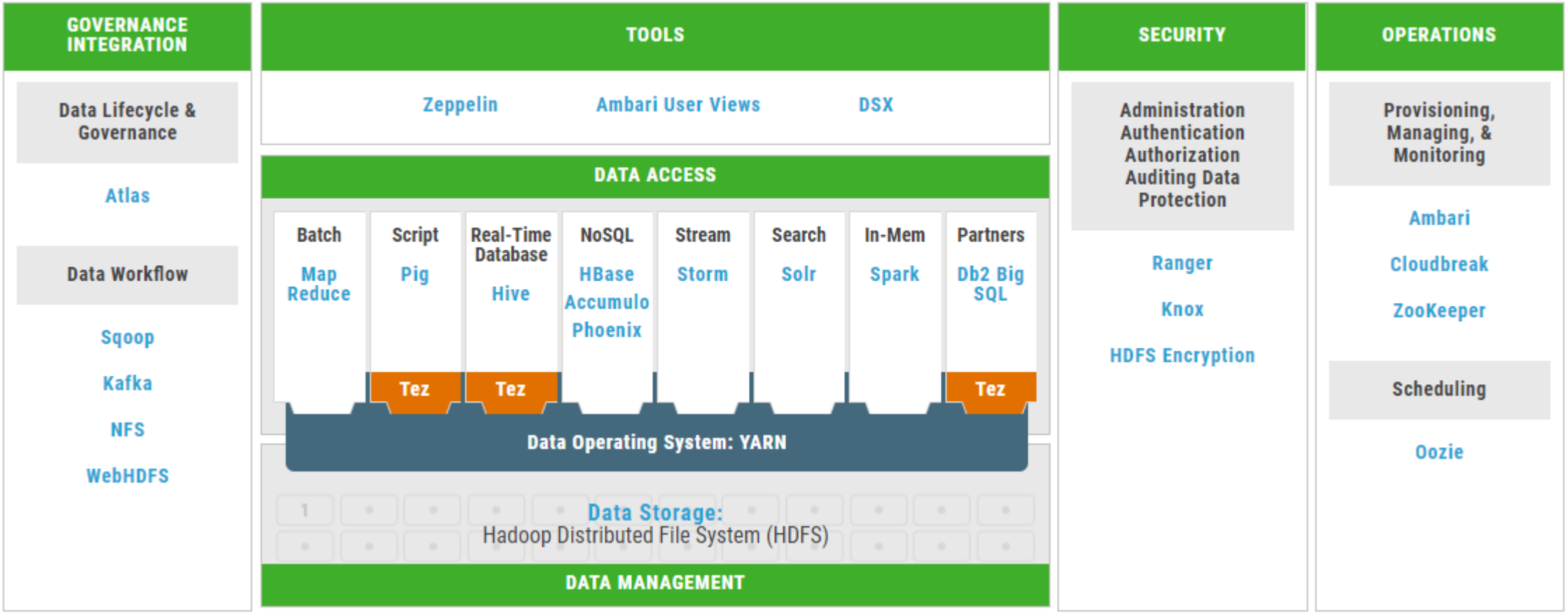
Big Data Technologies - MapReduce

How does MapReduce work?



Source: “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015

Apache Hadoop Ecosystem



Source: https://hortonworks.com/ecosystems/

Big Data Technologies - Facts about Hadoop

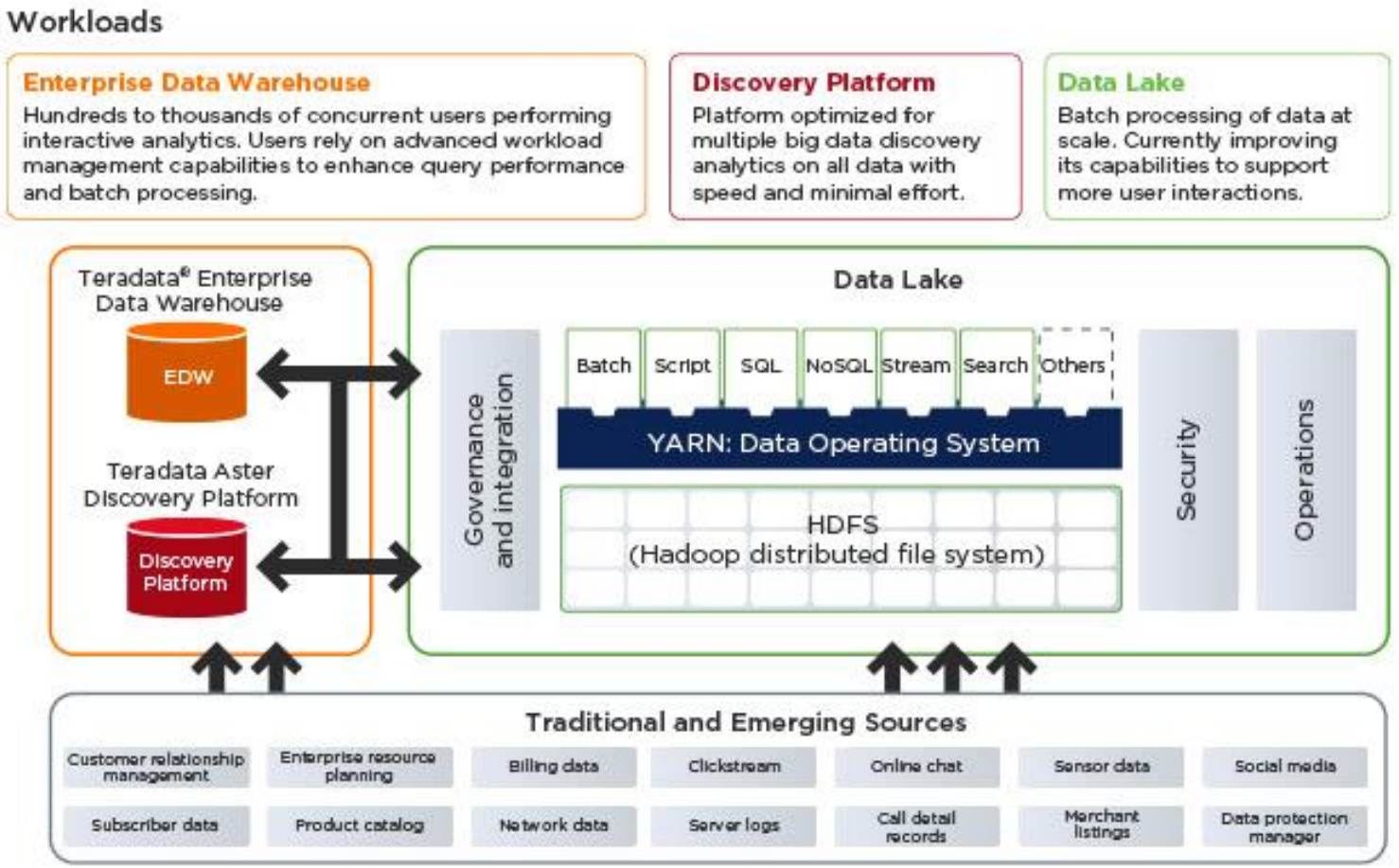
- Hadoop consists of multiple products
- Hadoop is open source but available from software providers too
- Hadoop is an ecosystem, not a single product
- HDFS is a file system, not a DBMS
- Hive resembles SQL but is not standard SQL
- Hadoop and MapReduce are related but not the same
- MapReduce provides control for analytics, not analytics
- Hadoop is about data diversity, not just data volume
- Hadoop complements a Data warehouse; it’s rarely a replacement

The Data Lake

- A storage repository to hold large amount of data in raw/native format
- Uses a flat architecture to store data
- Can store any type of data without enforcing restrictive schema
- Each element is assigned unique identifier and tagged with a set of metadata tags
- Data can be queried based on business need and the smaller dataset can be used for analytics
- Extract, Load & Transform data, instead of Extract, Transform, Load (ETL) of traditional data warehouses
- EMC, Microsoft Azure and Teradata are some providers

http://searchaws.techtarget.com/definition/data-lake

FIGURE **Data Lake Architecture**



An enterprise data warehouse, data lake and discovery platform facilitate analytics across the architecture to answer queries quickly.

Source: Teradata @ http://www.teradatamagazine.com/v14n04/Tech2Tech/Jump-in-a-Data-Lake/

Data Warehouse vs Data Lake

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Source: http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

Sources

- Chapter 2 (Data Analytics Life Cycle) from the book "Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing, and Presenting Data" by EMC Education Services, John Wiley & Sons, 2015.
- http://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#513ce6336796 (An article on free/open Big Data sources)
- https://docs.oracle.com/database/121/DWHSG/concept.htm#DWHSG-GUID-452FBA23-6976-4590-AA41-1369647AD14D
- https://docs.oracle.com/cd/B10500_01/server.920/a96520/ettoverv.htm (overview of ETL on Oracle website)
- “Business Intelligence and Analytics” by Sharda et al., Pearson Education, 2015