Distributed Big Data Management

Dr. Mohamed Elhoseny

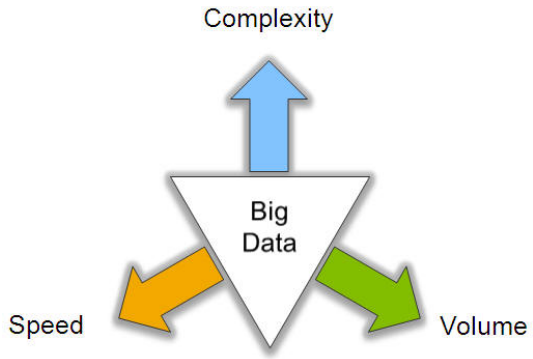
2019

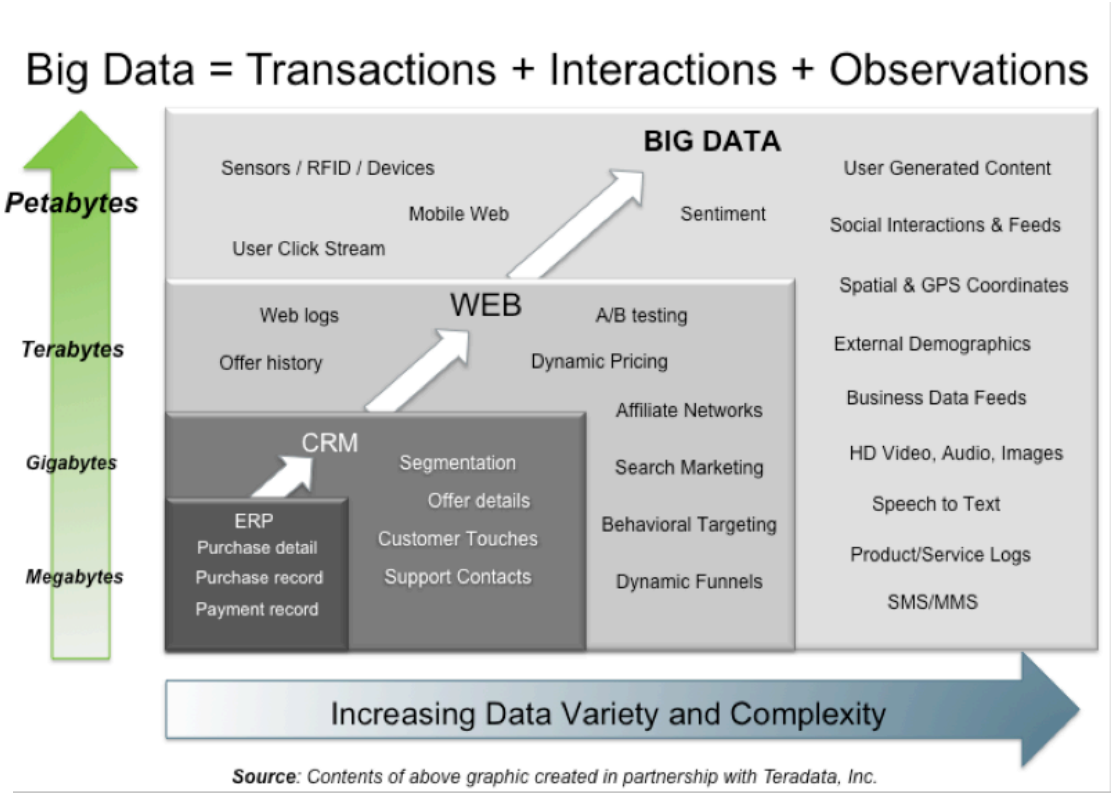
What's Big Data?

- **Big data** is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, curation, storage, search, sharing, transfer, analysis, and visualization.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

Big Data: 3V's


BIG
DATA?
VOLUME
Large amounts of data.
VELOCITY
Needs to be analyzed quickly.
VARIETY
Different types of structured and unstructured data.


Complexity
Big Data
Speed
Volume

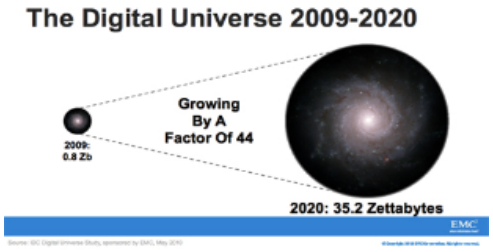

Big Data = Transactions + Interactions + Observations
Petabytes
Terabytes
Gigabytes
Megabytes
BIG DATA
Sensors / RFID / Devices
Mobile Web
User Click Stream
Sentiment
User Generated Content
Social Interactions & Feeds
Spatial & GPS Coordinates
External Demographics
Business Data Feeds
HD Video, Audio, Images
Speech to Text
Product/Service Logs
SMS/MMS
WEB
Web logs
Offer history
A/B testing
Dynamic Pricing
Affiliate Networks
Search Marketing
Behavioral Targeting
Dynamic Funnels
CRM
Segmentation
Offer details
Customer Touches
Support Contacts
ERP
Purchase detail
Purchase record
Payment record
Increasing Data Variety and Complexity
Source: Contents of above graphic created in partnership with Teradata, Inc.

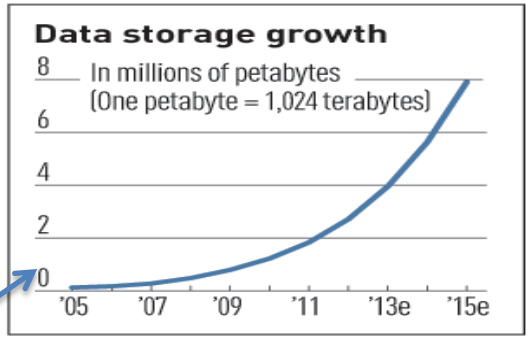
What is “big data”?

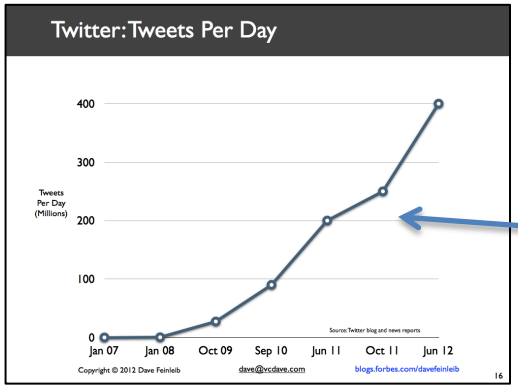
- "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”
- Complicated (intelligent) analysis of data may make a small data “appear” to be “big”
- Bottom line: Any data that exceeds our current capability of processing can be regarded as “big”

Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





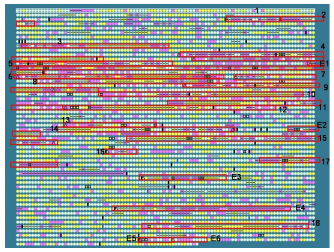


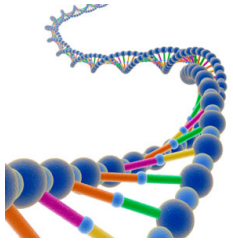
Exponential increase in collected/generated data

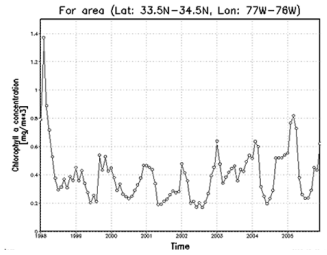

12+ TBs
of tweet data
every day
? TBs of
data every day
25+ TBs of
log data
every day
30 billion RFID
tags today
(1.3B in 2005)
4.6
billion
camera
phones
world wide
100s of
millions
of GPS
enabled
devices sold
annually
2+
billion
people on
the Web
by end
2011
76 million smart meters
in 2009…
200M by 2014

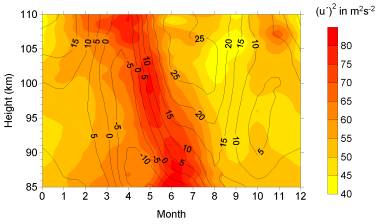
Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



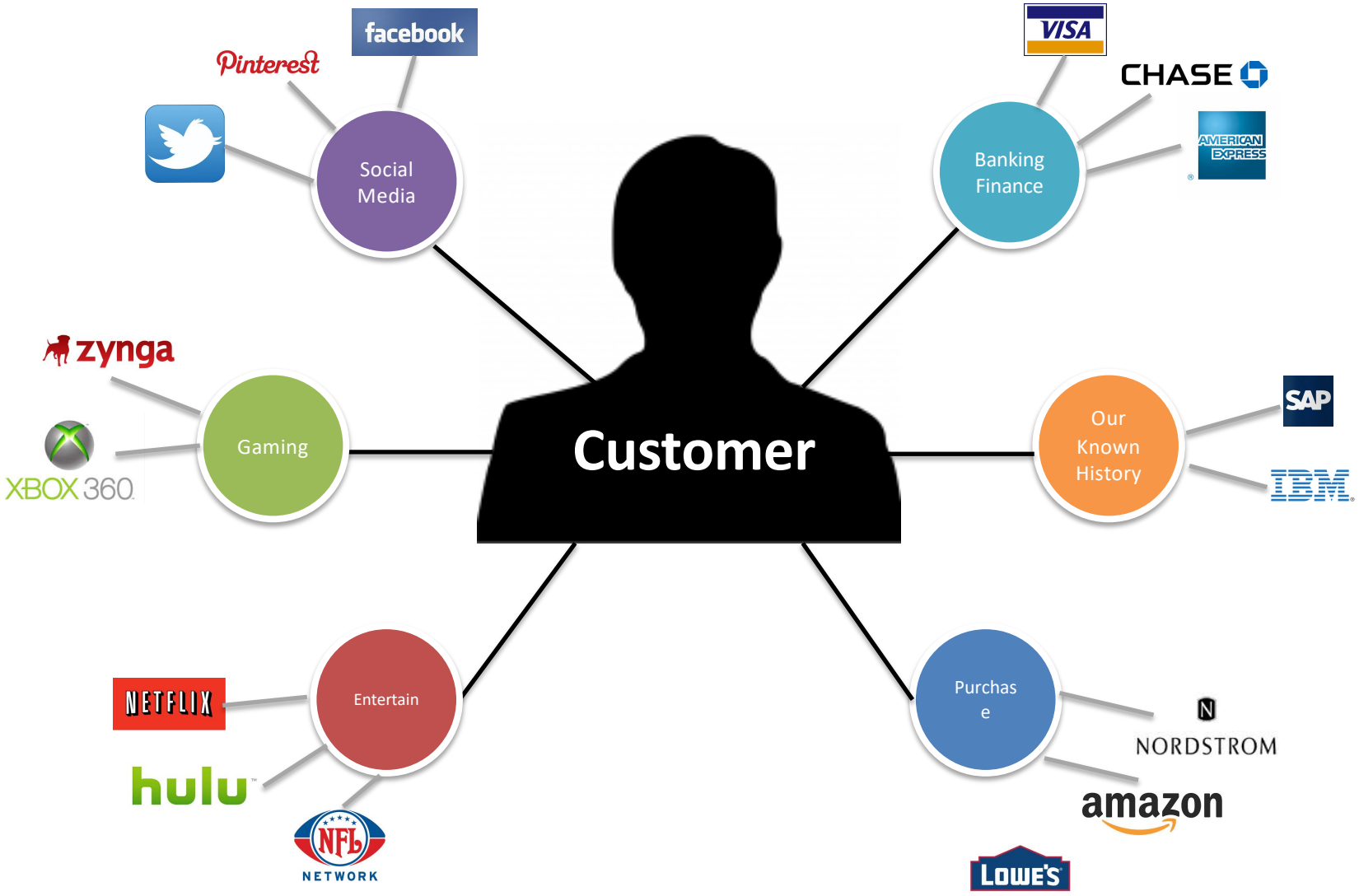






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

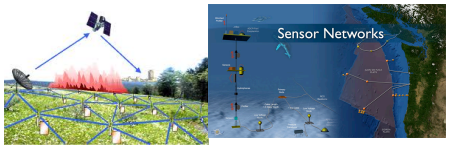
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

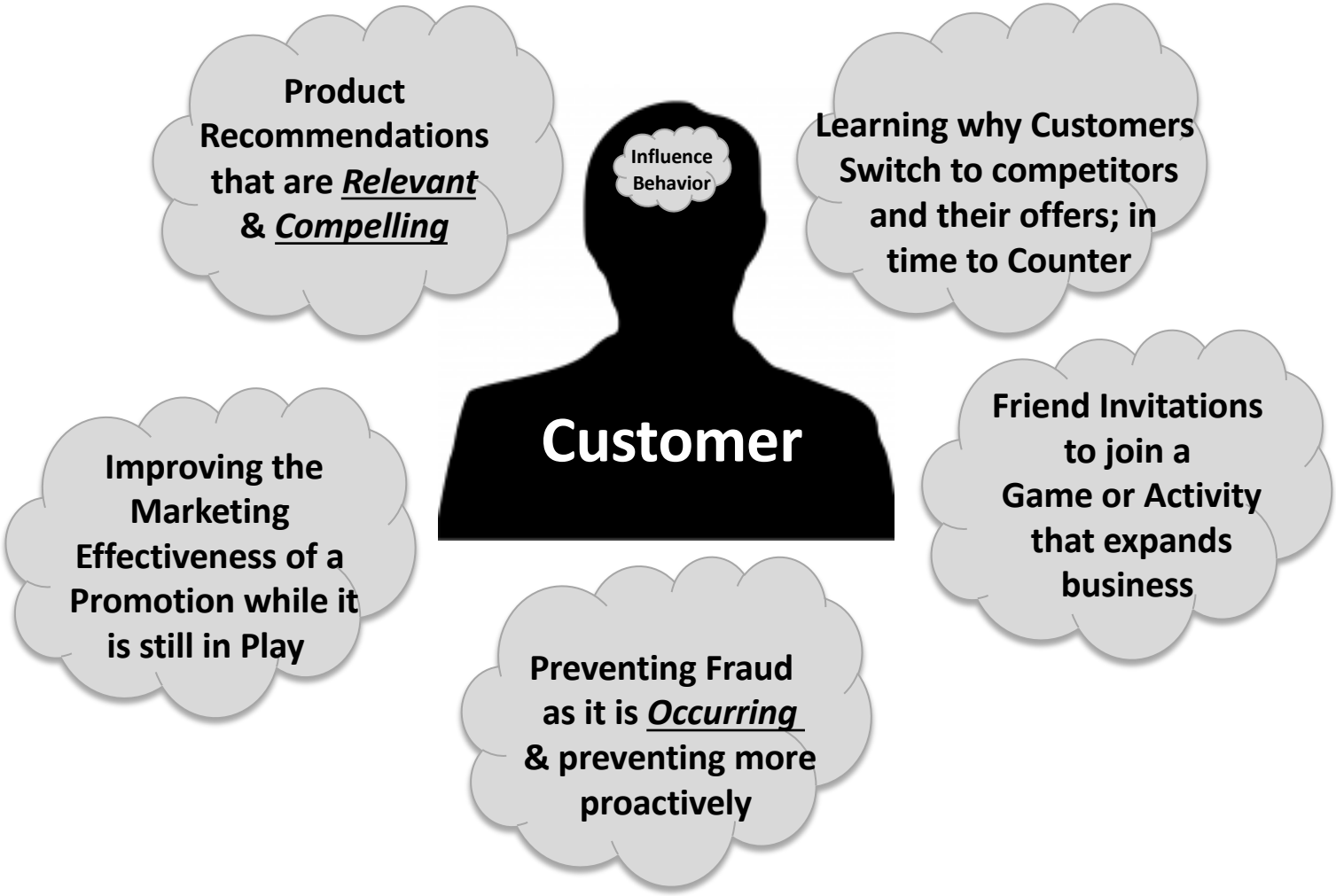
Mobile devices
(tracking all objects all the time)



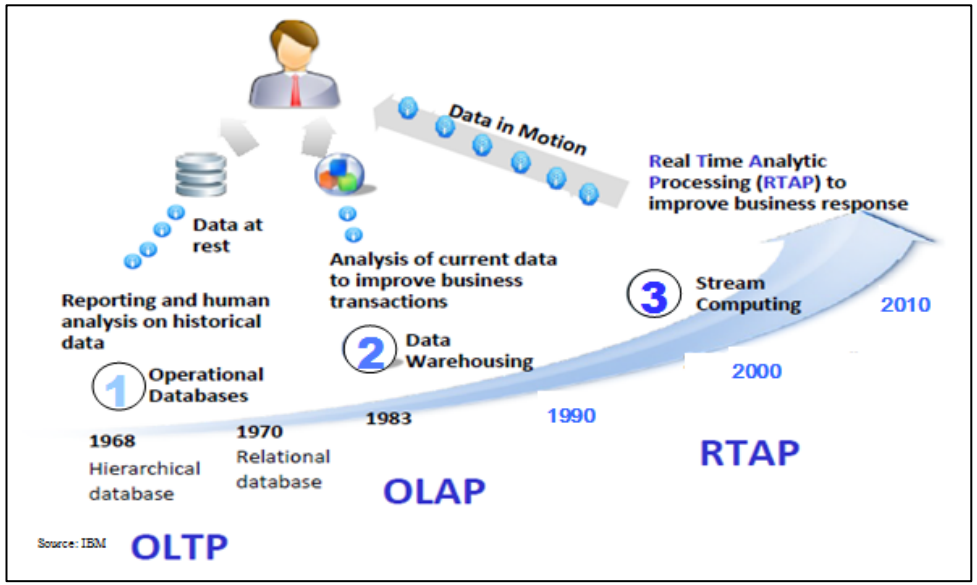
Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Real-Time Analytics/Decision Requirement



Harnessing Big Data



- **OLTP:** Online Transaction Processing (DBMSs)
- **OLAP:** Online Analytical Processing (Data Warehousing)
- **RTAP:** Real-Time Analytics Processing (Big Data Architecture & technology)

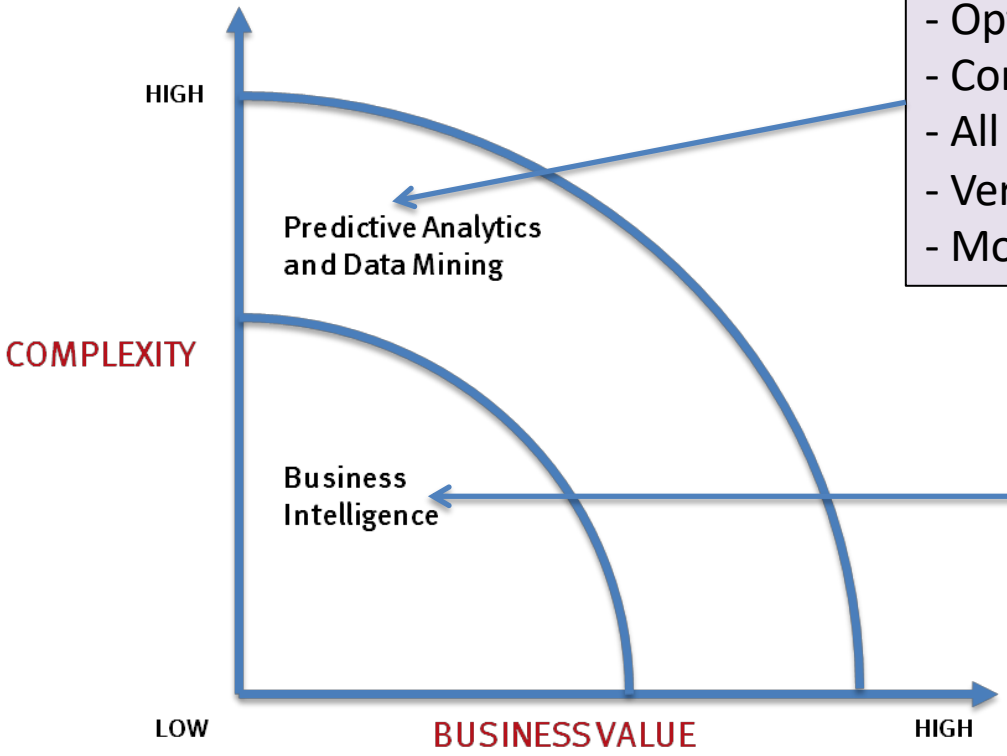
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

What's driving Big Data

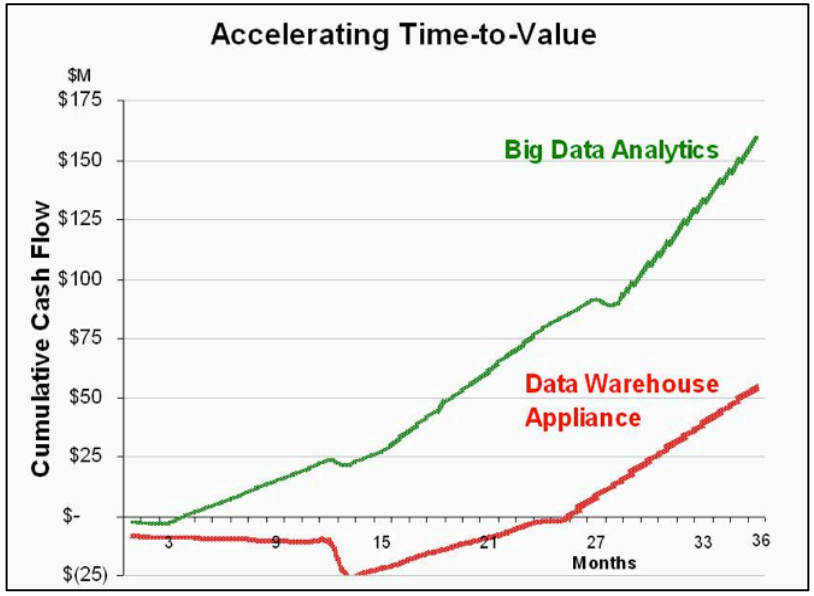


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Big Data Analytics

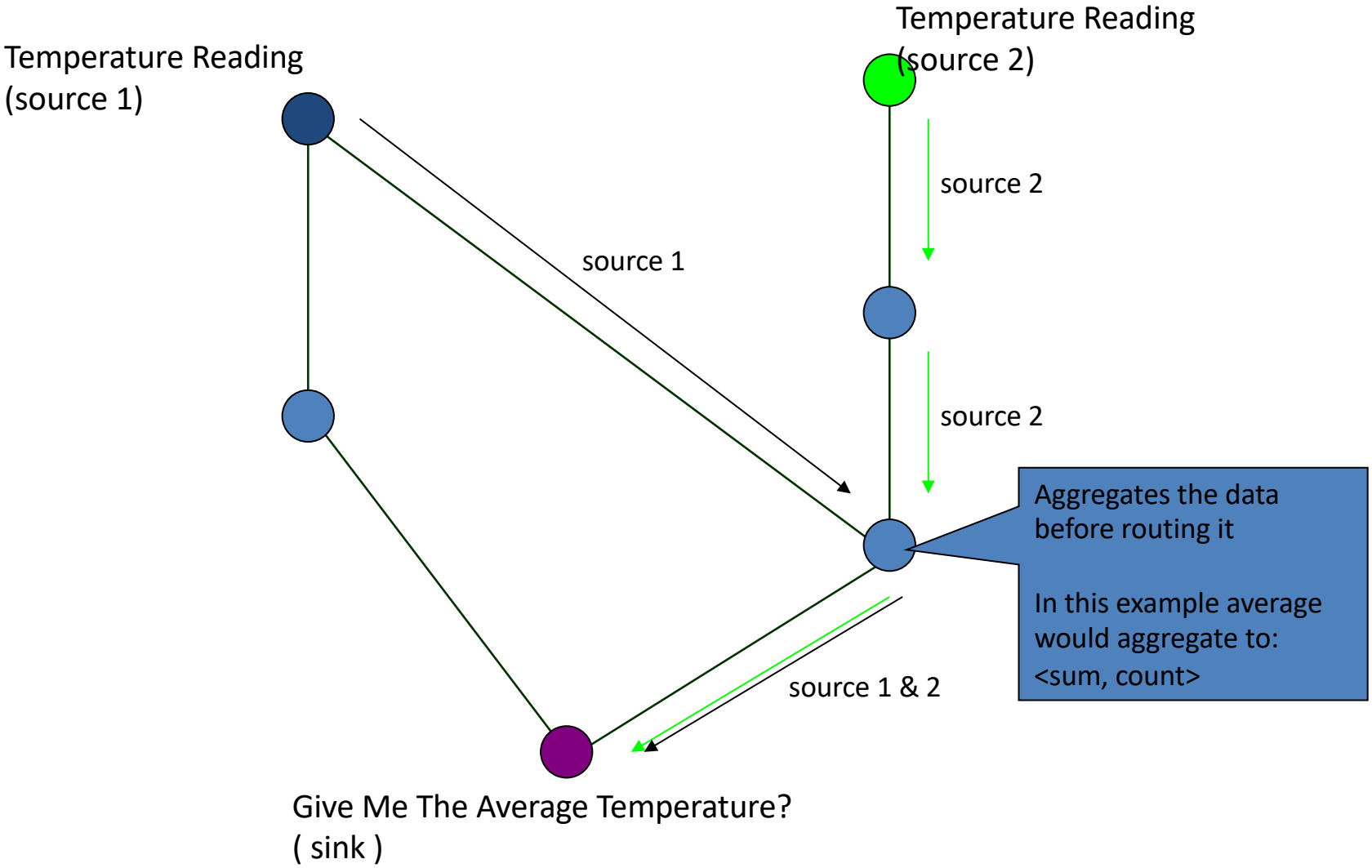
- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



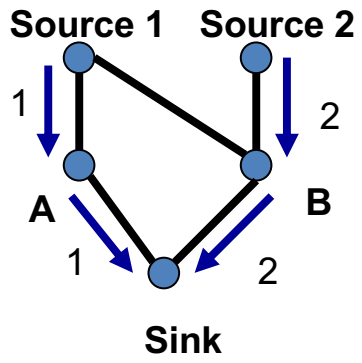
Data Aggregation in Sensor Networks

- Redundant Data/events
- Some services are amenable for in-network computations.
  - "The network is the sensor"
- Communication can be more expensive than computation.
- By performing "computation" on data in route to the sink, we can reduce the amount of data traffic in the network.
- Increases energy efficiency as well as scalability
  - The bigger the network, the more computational resources.

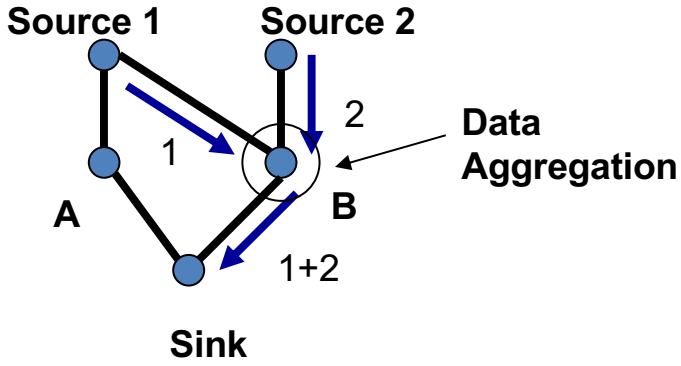
Data Aggregation



Transmission modes
AC vs DC



a) Address-Centric (AC) Routing (no aggregation)

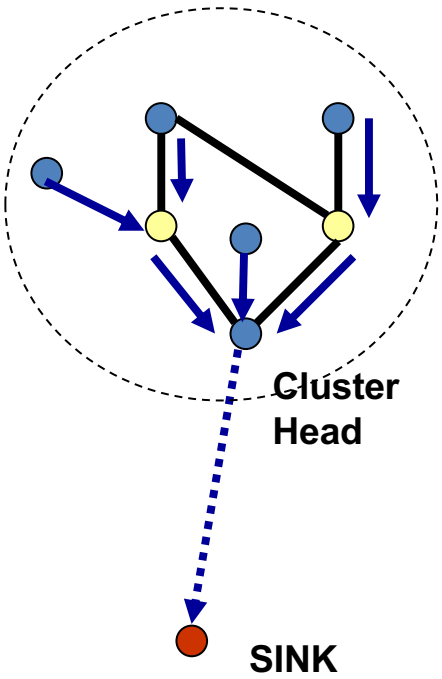


b) Data-Centric (DC) Routing (in-network aggregation)

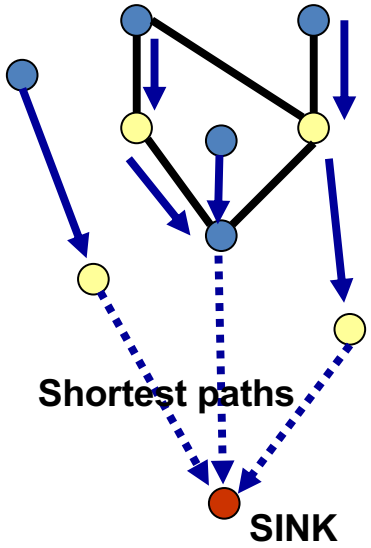
Aggregation Techniques

- Center at Nearest Source (CNSDC): All sources send the information first to the source nearest to the sink, which acts as the aggregator.
- Shortest Path Tree (SPTDC): Opportunistically merge the shortest paths from each source wherever they overlap.
- Greedy Incremental Tree (GITDC): Start with path from sink to nearest source. Successively add next nearest source to the existing tree.

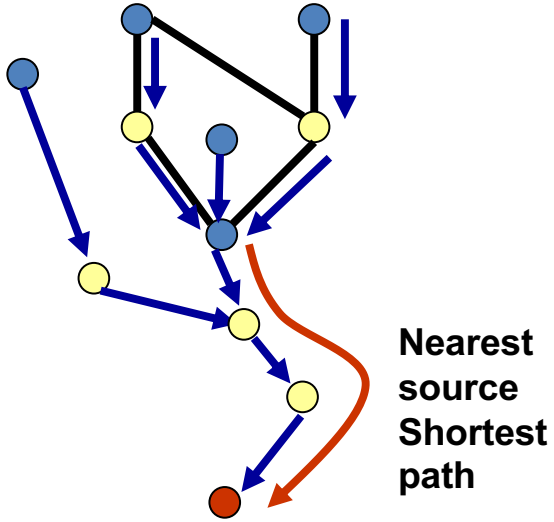
Aggregation Techniques



a) Clustering based CNS



b) Shortest Path Tree



c) Greedy Incremental

Data Storage in Sensors Model

- The data that is generated at one node is stored at another node determined by the name of the data.
  - Data must be named
- Data can be stored and retrieved by name. Generally speaking, a data-centric storage system provides primitives of the form:
  - *put (data)* and
  - *data = get (name)*.

- ***External Storage****: The cost of accessing the event is zero, while the cost of conveying the data to this external node is non-trivial, and significant energy is expended at nodes near the external node*
  - *Appropriate if the events are accessed far more frequently than generated.*
- ***Local Storage****: Incurs zero communication cost in storing the data, but incurs a large communication cost –a network flood– in accessing the data.*
  - *Feasible when events are accessed less frequently than they are generated.*
- ***Data-Centric Storage:*** *lies in between, incurs non-zero cost both in storing events and retrieving them.*