The Evolution of Analytic Scalability

Introduction- The Convergence of the Analytic and Data Environment

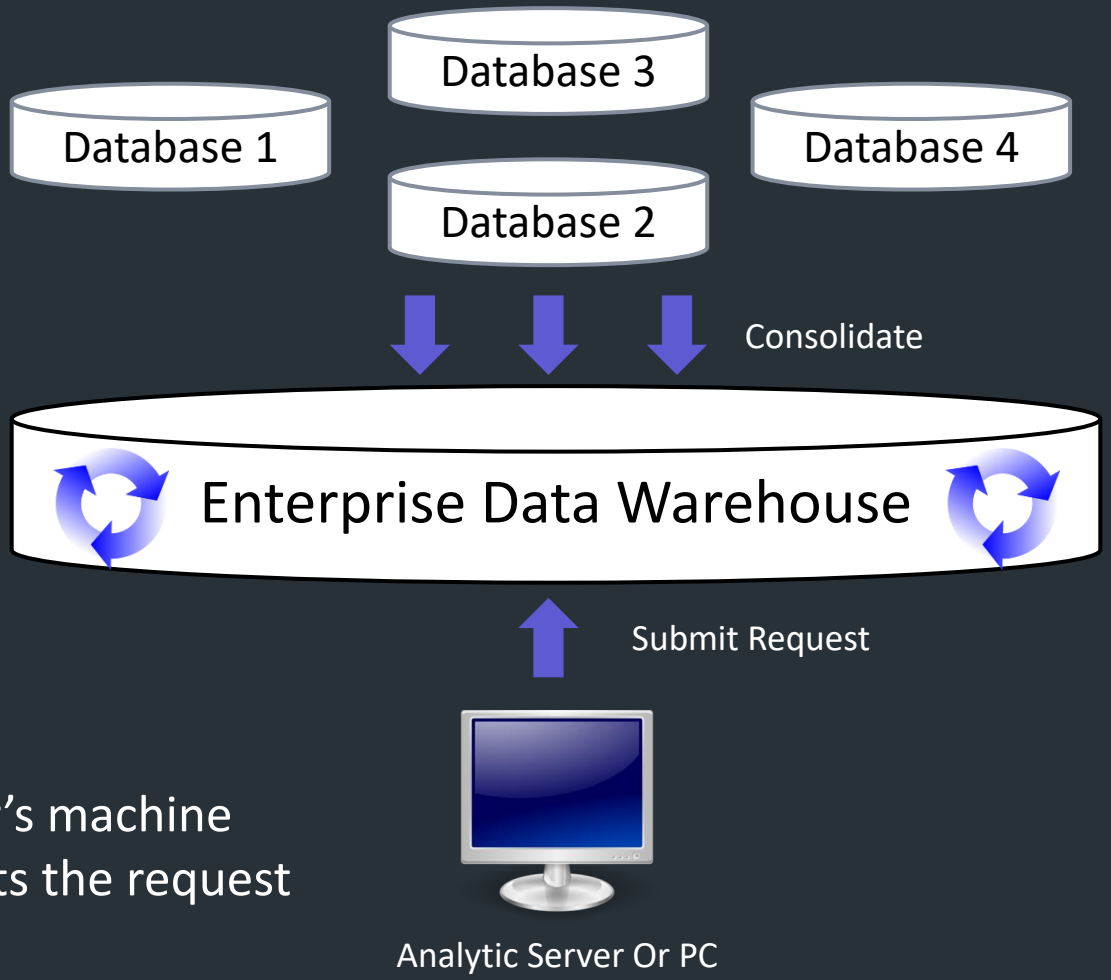
- The amount of data organizations process continues to increase

- Important technologies to tame the big data tidal wave possible

MPP | The cloud | Grid computing | MapReduce

Traditional Analytic Architecture

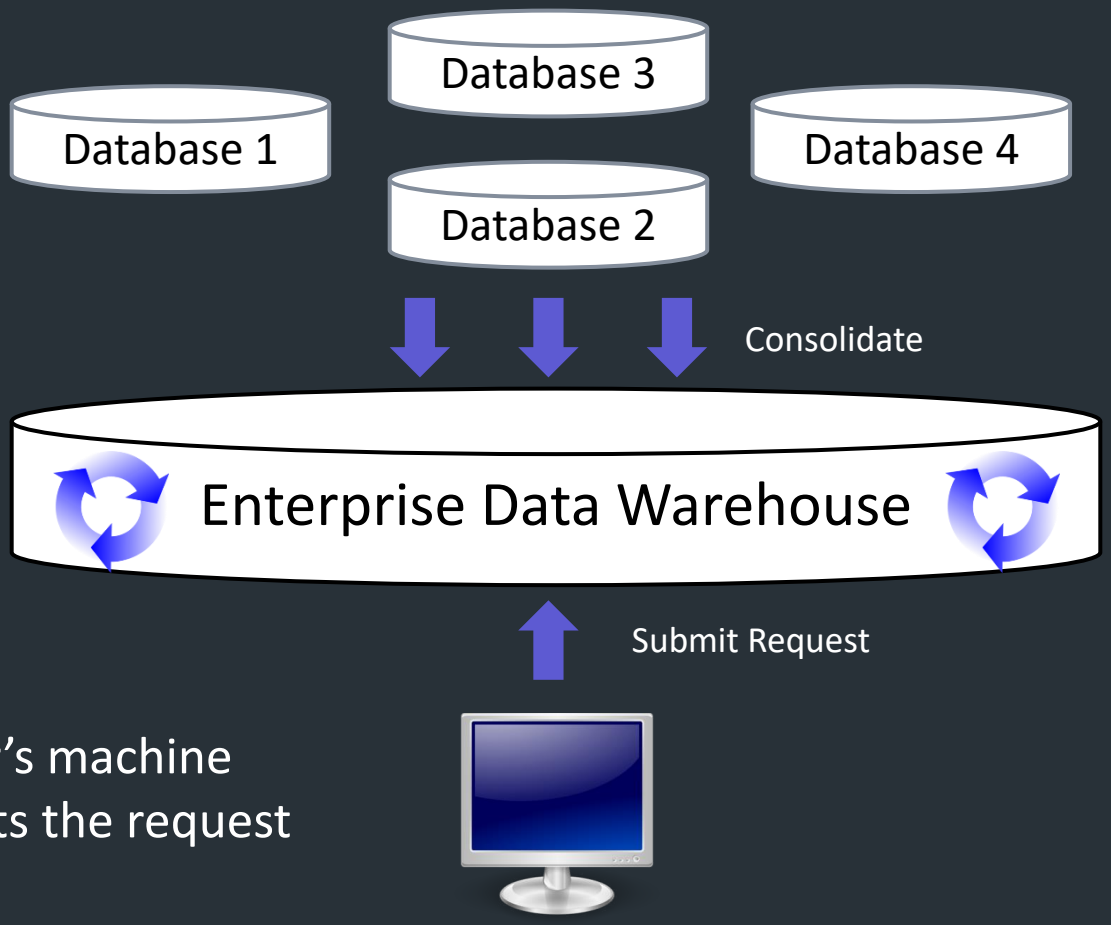
- We had to pull all data together into a separate analytics environment to do analysis



The user's machine just submits the request

Modern In-Database Architecture

- The processing stays in the database where the data has been consolidated



The user's machine
just submits the request

)

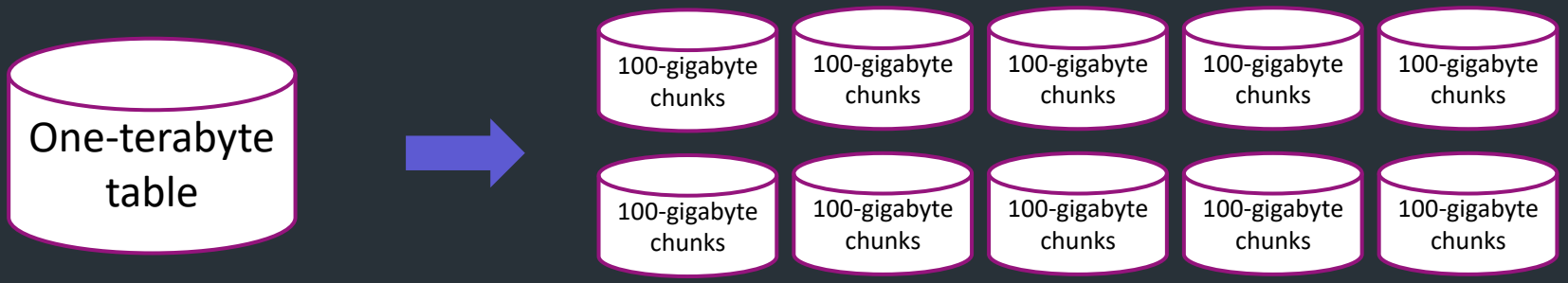
What is an MPP Database?

- An MPP database breaks the data into independent chunks with independent disk and CPU

Single overloaded server

Multiple lightly loaded servers

Shared Nothing!



A Traditional database will query a one-terabyte table one row at time

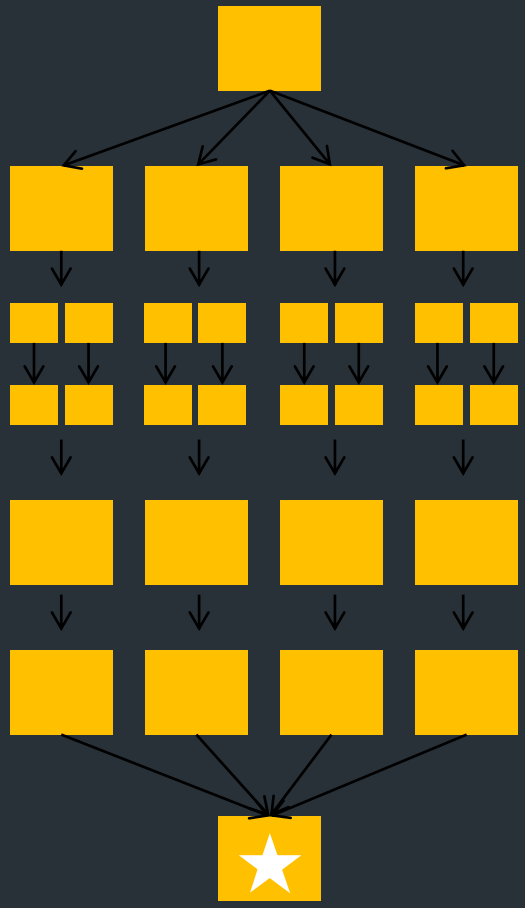
10 simultaneous 100-gigabyte queries

Concurrent Processing

- An MPP system allows the different sets of CPU and disk to run the process concurrently

Single Threaded Process

An MPP system breaks the job into pieces



Parallel Process

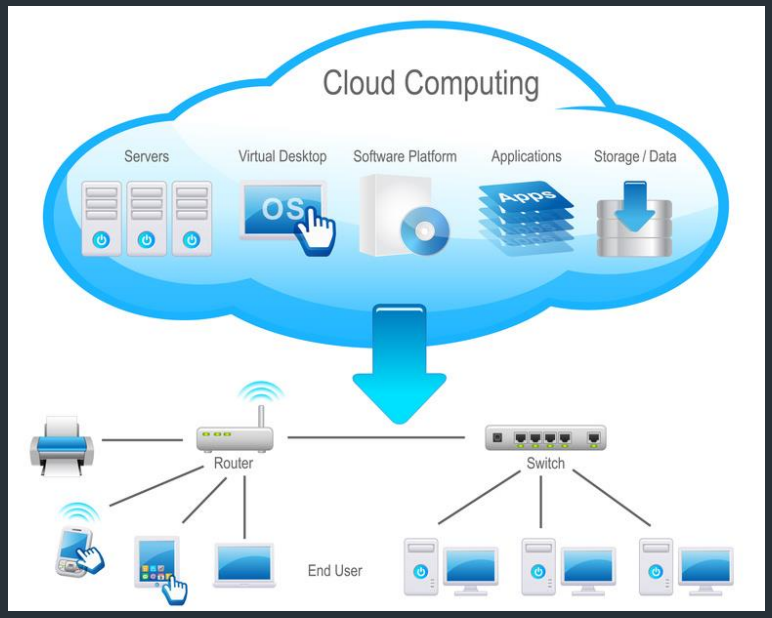
Others

- MPP systems build in redundancy to make recovery easy
- MPP systems have resource management tools
  - Manage the CPU and disk space
  - Query optimizer

What is Cloud Computing?

- McKinsey and Company paper from 2009[1]
  - Mask the underlying infrastructure from the user
  - Be elastic to scale on demand
  - On a pay-per-use basis

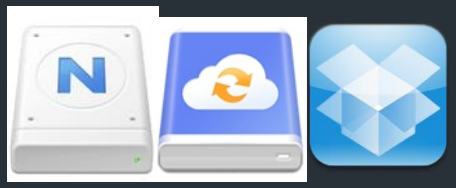
- National Institute of Standards and Technology (NIST)
  - On-demand self-service
  - Broad network access
  - Resource pooling
  - Rapid elasticity
  - Measured service



[1] McKinsey and Company, 'Clearing the Air on Cloud Computing," March 2009.

Two Types of Cloud Environment

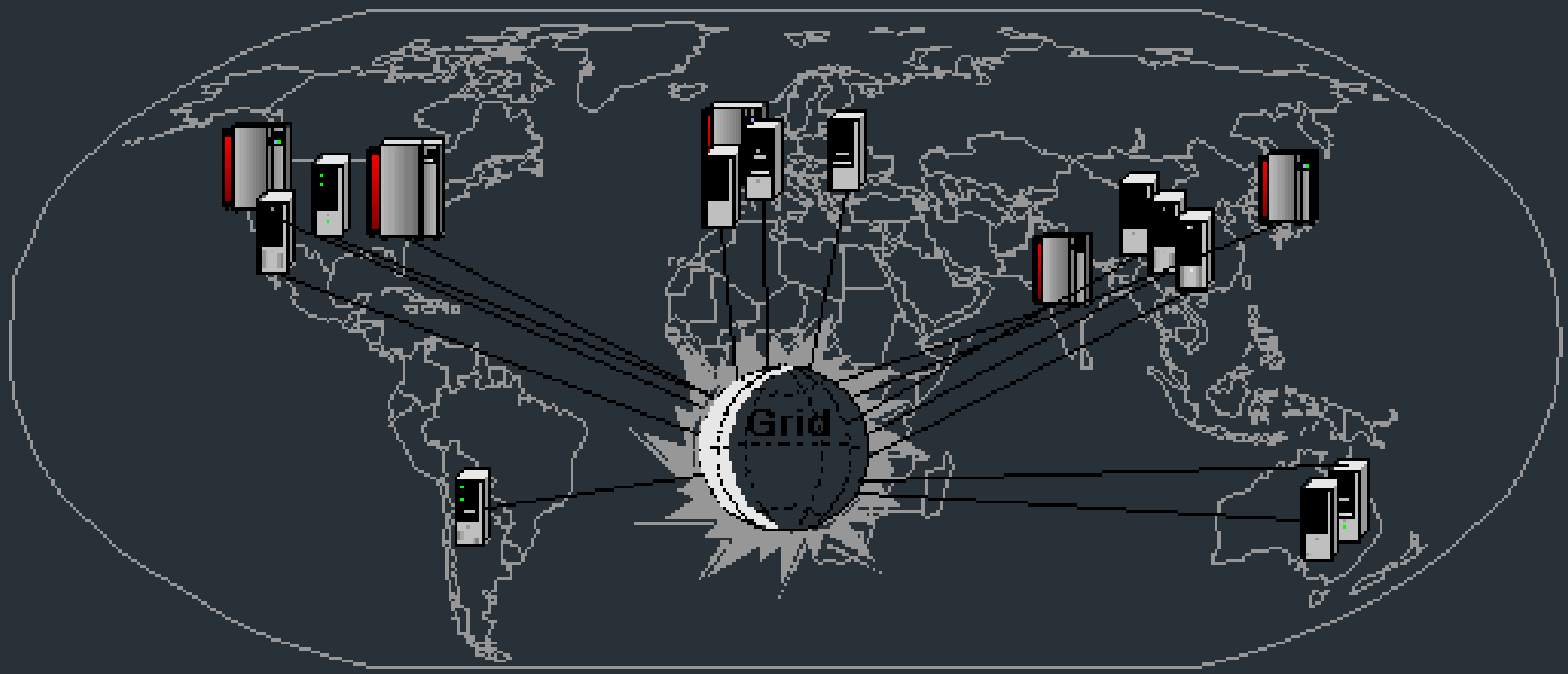
1. **Public Cloud**
   - The services and infrastructure are provided off-site over the internet
   - Greatest level of efficiency in shared resources
   - Less secured and more vulnerable than private clouds



2. **Private Cloud**
   - Infrastructure operated solely for a single organization
   - The same features of a public cloud
   - Offer the greatest level of security and control
   - Necessary to purchase and own the entire cloud infrastructure

Grid Computing

- The federation of computer resources to reach a common goal
  - E.g., SETI@Home (Search for Extraterrestrial Intelligence)
    - An Internet-based public volunteer computing project

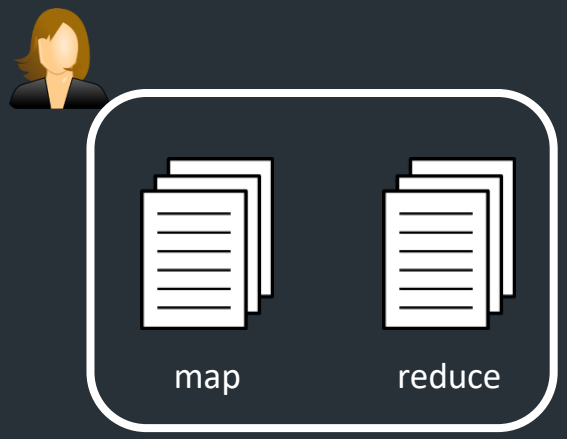


What is MapReduce?

- A Parallel programming framework[1]

Library

Parallelization
Fault-tolerance
Data distribution
Load balancing
……



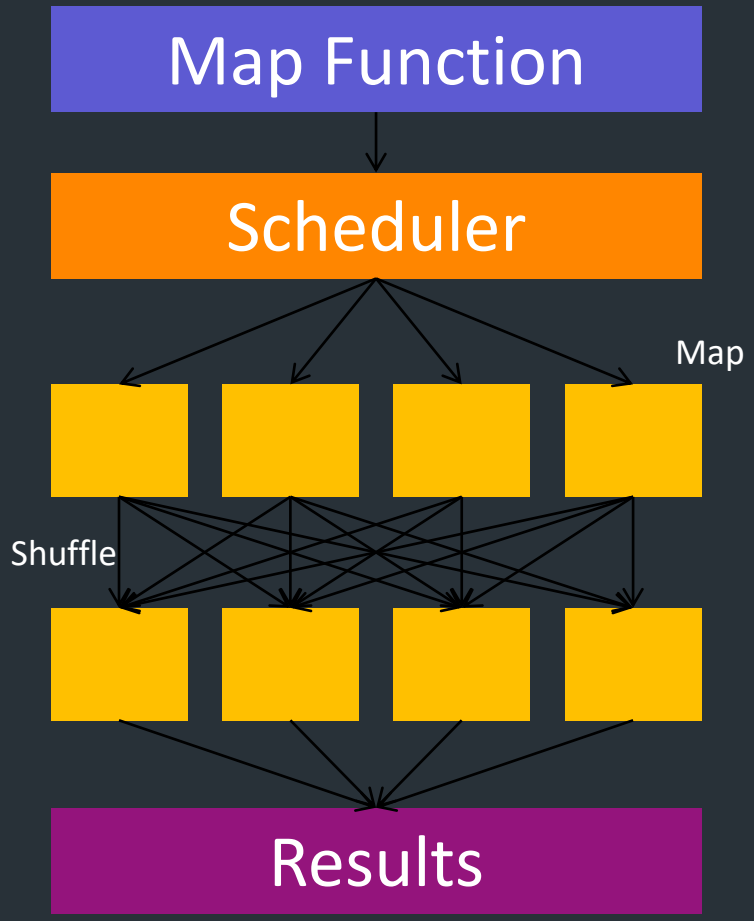
- *Map function*
  - Processing a key/value pairs to generate a set of intermediate key/value pairs
- *Reduce function*
  - Merging all intermediate values associated with the same intermediate key

[1] MapReduce: Simplified Data Processing on Large Clusters – OSDI 2004

How MapReduce Works

- Let's assume there are 20 terabytes of data and 20 MapReduce server nodes for a project

1. Distribute a terabyte to each of the 20 nodes using a simple file copy process
2. Submit two programs(Map, Reduce) to the scheduler
3. The map program *finds the data* on disk and *executes* the logic it contains
4. The results of the map step are then passed to the reduce process to *summarize* and *aggregate* the final answers



Strengths and Weaknesses

- **Good for**
  - Lots of input, intermediate, and output data
  - Batch oriented datasets (ETL: Extract, Load, Transform)
  - Cheap to get up and running because of running on commodity hardware

- **Bad for**
  - Fast response time
  - Large amounts of shared data
  - CPU intensive operations (as opposed to data intensive)
  - NOT a database!
    - No built-in security
    - No indexing, No query or process optimizer
    - No knowledge of other data that exists

The Evolution of Analytic Processes

Analytic Sandbox Benefits

View of Analytic Professionals

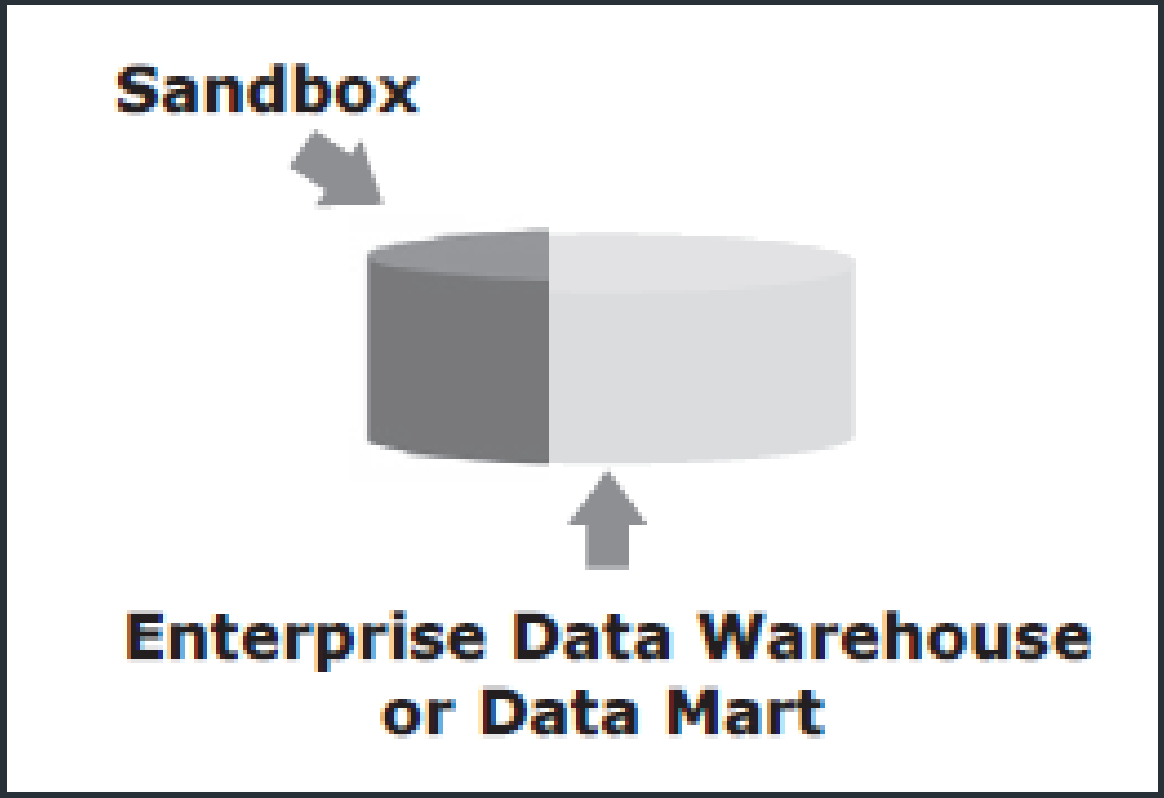
- Independence
- Flexibility
- Efficiency
- Freedom
- Speed

Analytic Sandbox Benefits

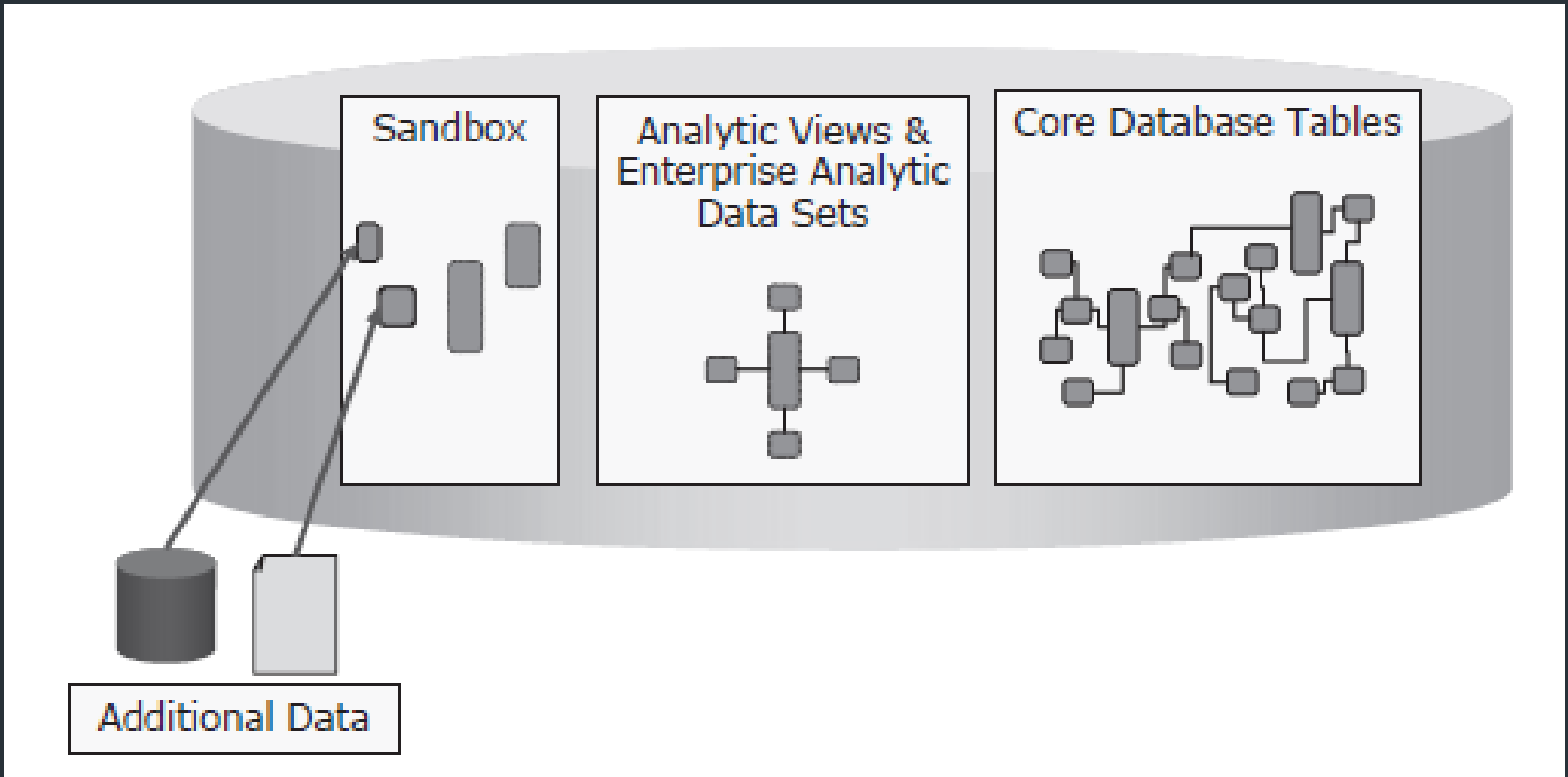
View of IT

Centralization.
Streamlining
Simplicity.
Control.
Costs:

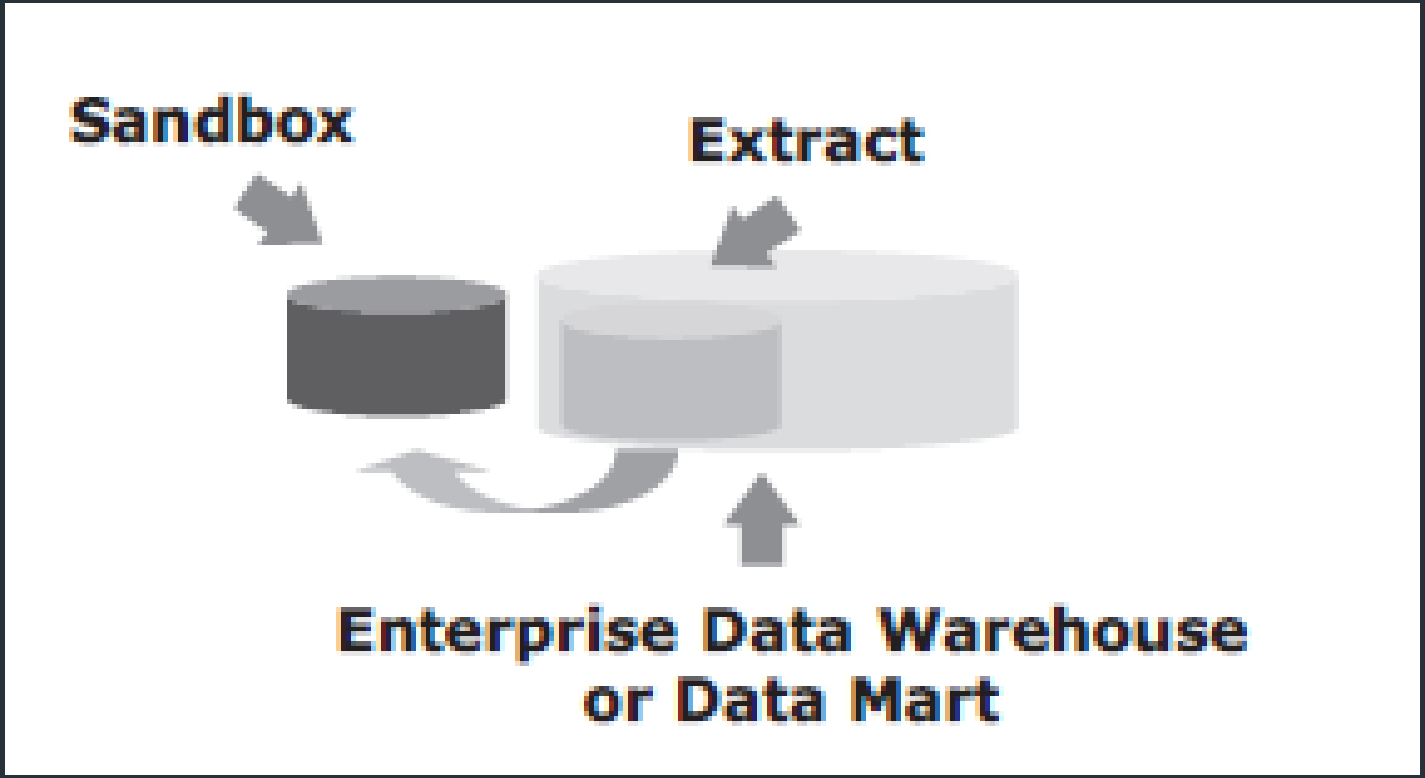
An Internal Sandbox



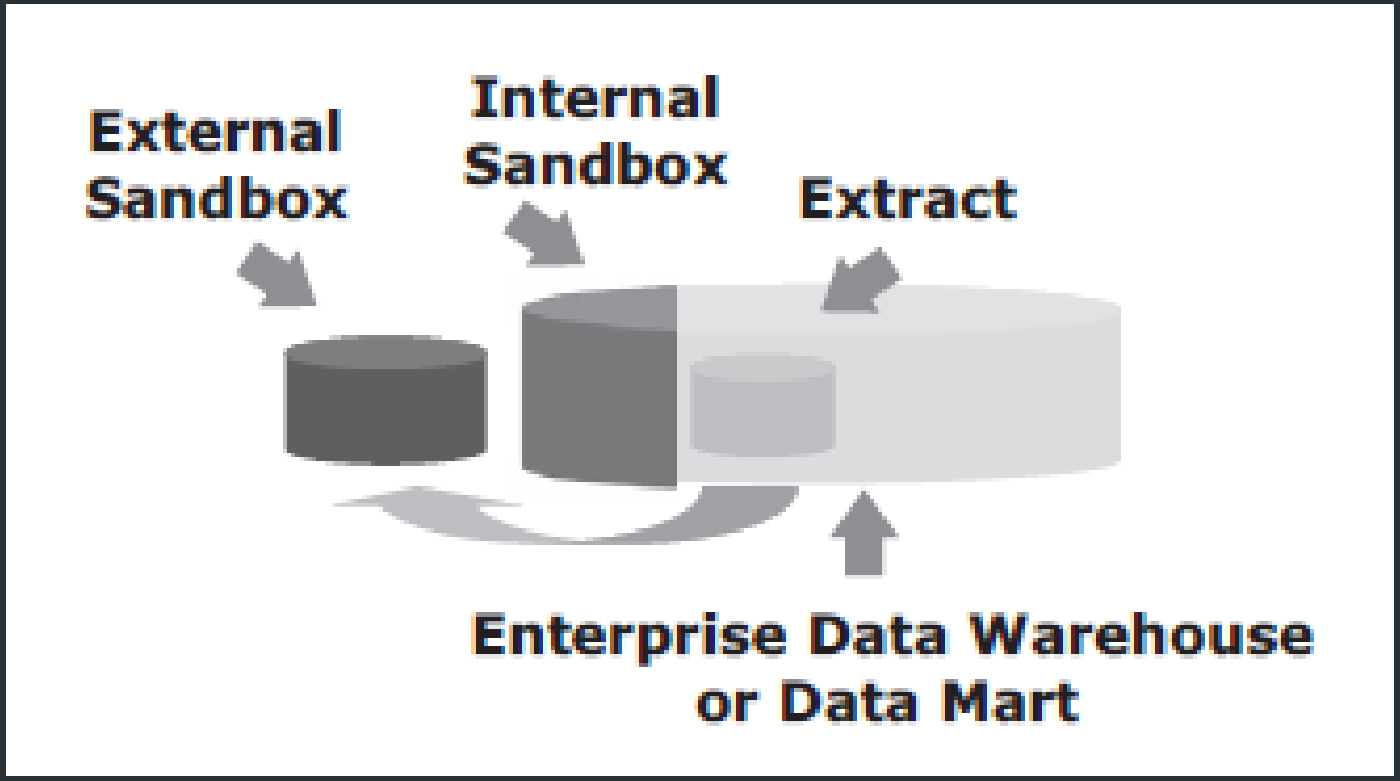
Internal Sandbox View



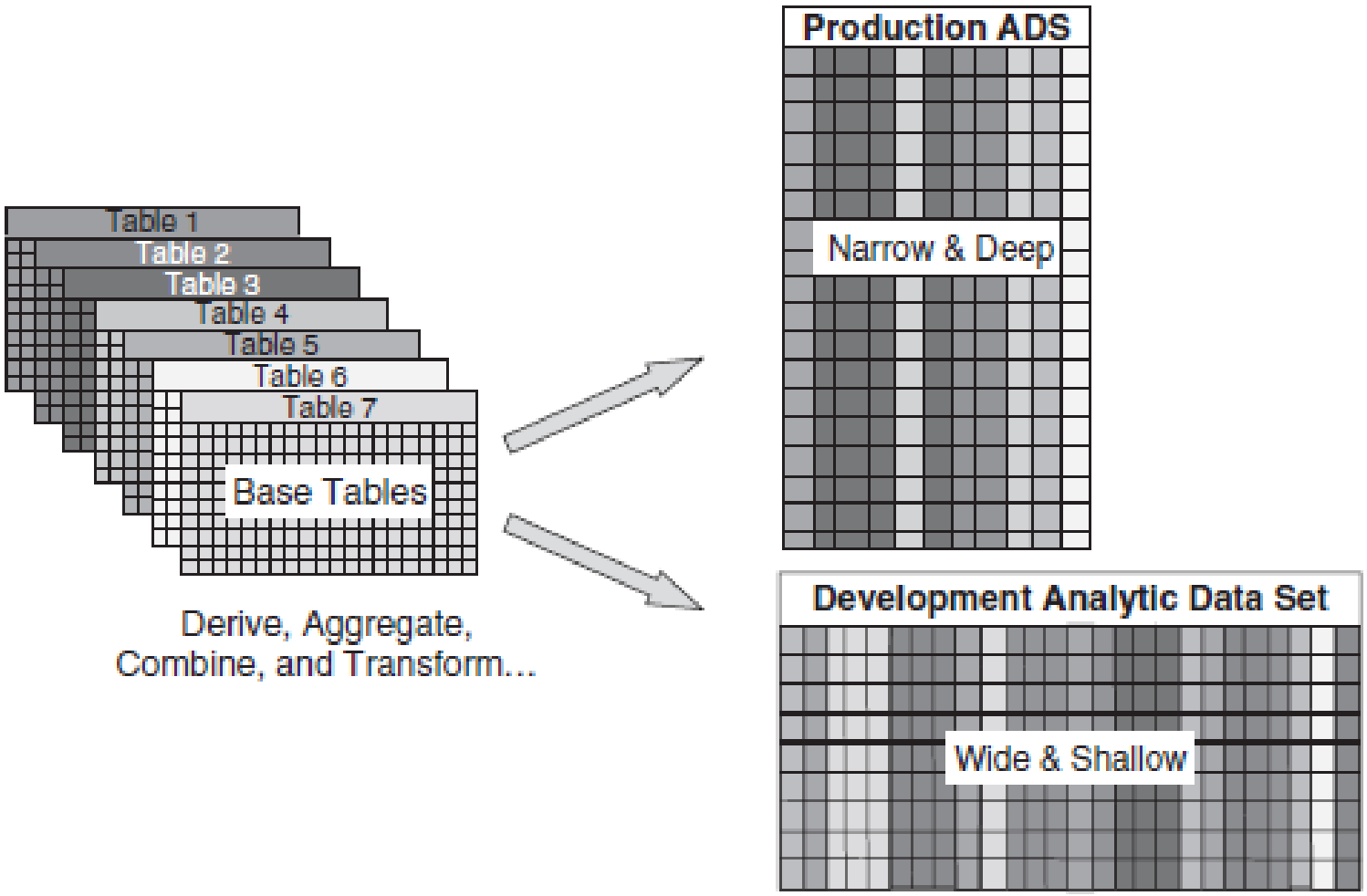
An External Sandbox



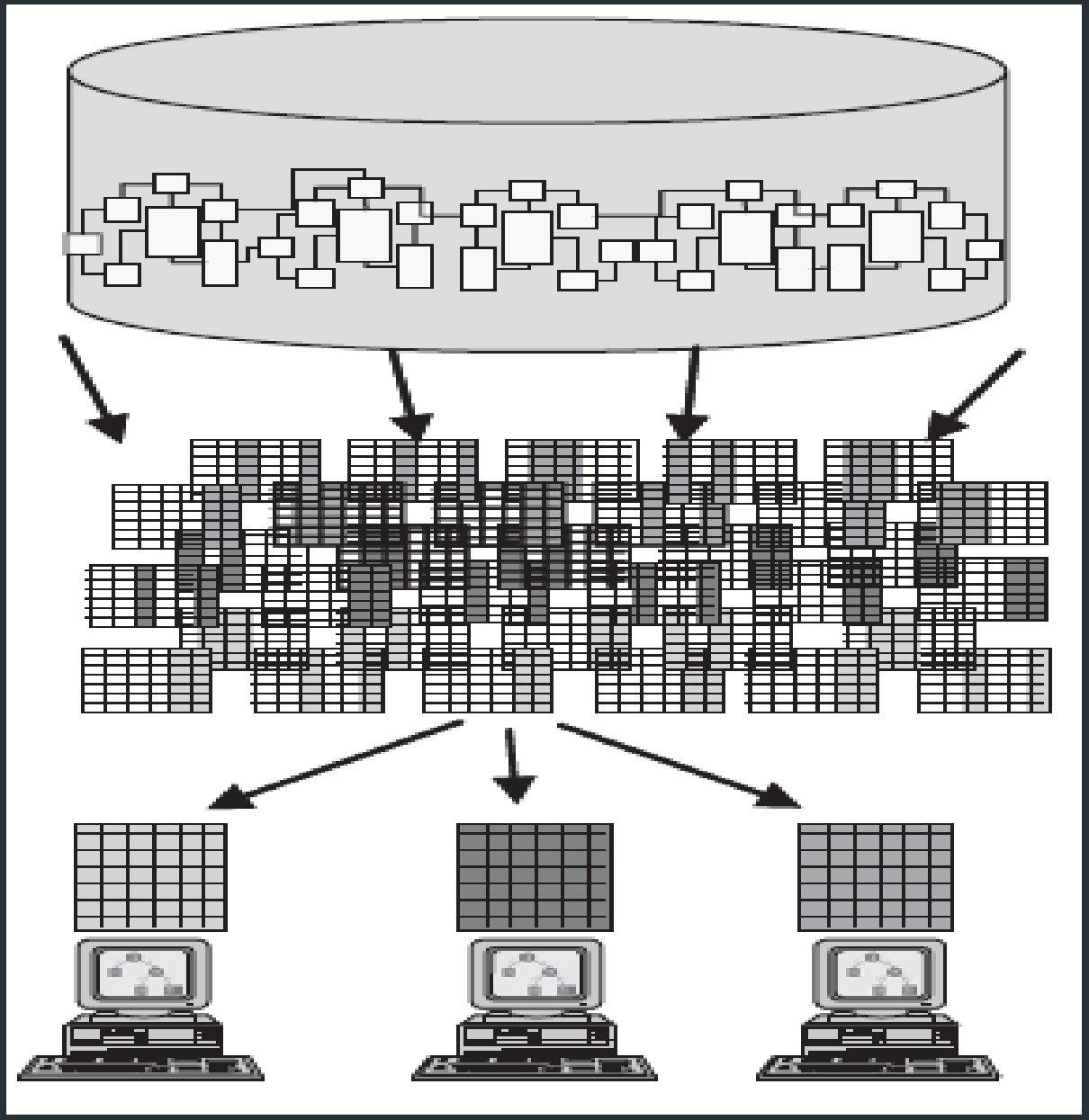
Hybrid Sandbox



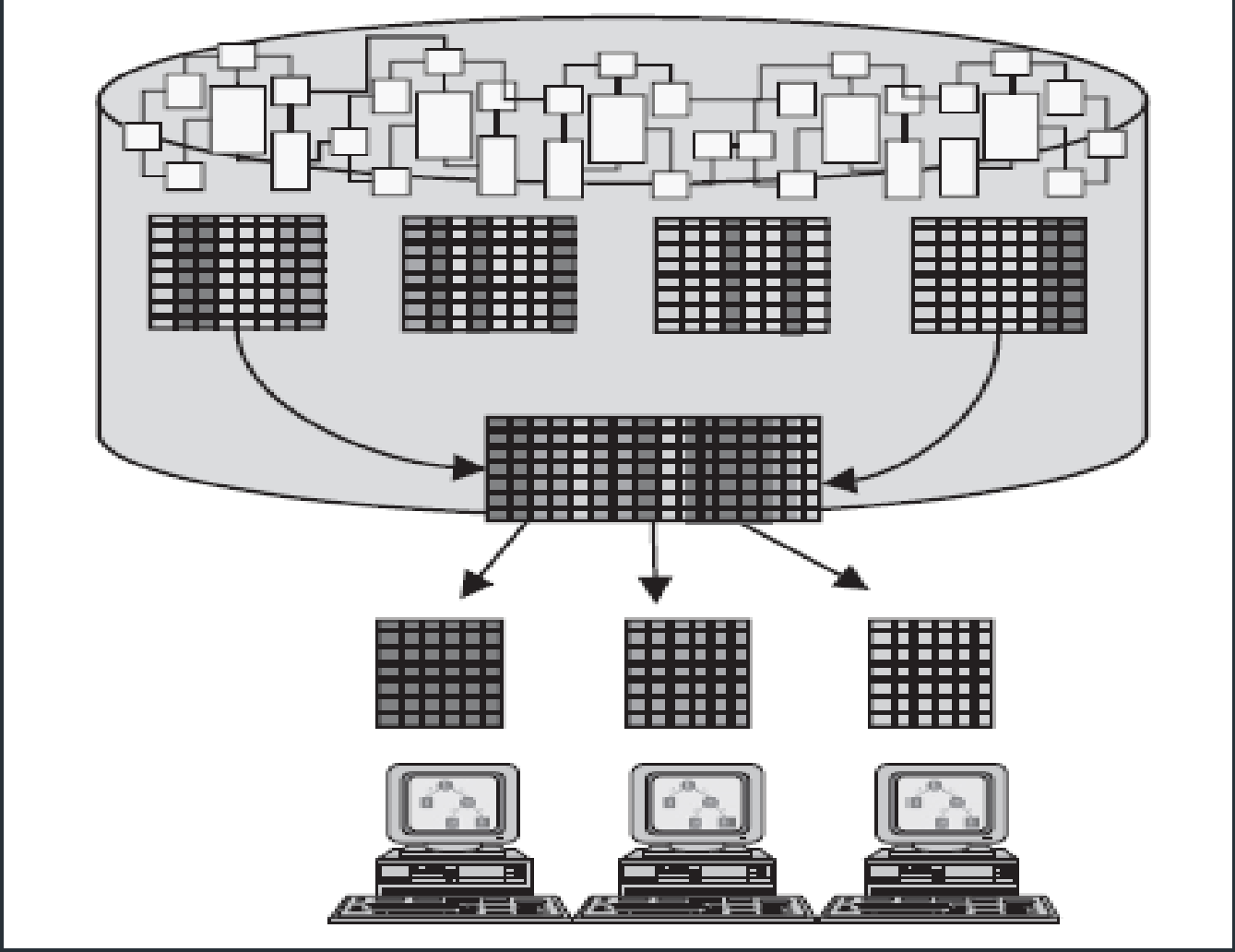
Development versus Production Analytic Data Sets



Traditional Analytic Data Set Process



Enterprise Analytic Data Set Process



Analytic Sandbox Benefits

View of IT

Centralization.
Streamlining
Simplicity.
Control.
Costs:

Analytic Sandbox Benefits

View of IT

Centralization.
Streamlining
Simplicity.
Control.
Costs:

Conclusion

- These technologies can integrate and work together
  - Databases running in the cloud
  - Databases including MapReduce functionality
  - MapReduce can be run against data sourced from a database
  - MapReduce can also run against data in the cloud

[Cloud Database]

[SQL-MapReduce]

[In-Database MapReduce][1]

[Running MapReduce in Database]

[Running MapReduce in Cloud][2]

[1] https://blogs.oracle.com/datawarehousing/entry/in-database_map-reduce
[2] http://code.google.com/p/cloudmapreduce/
Cloud mapreduce: a mapreduce implementation on top of a cloud operating system – CCGRID 2011, IEEE Computer Society