Introduction

School of Information Studies
Syracuse University

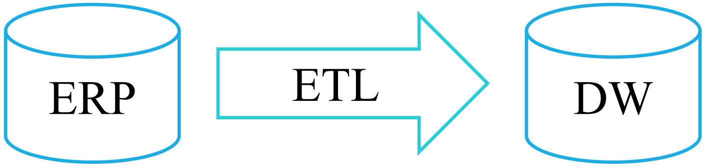
Agenda

- Understand the different types of data warehouse architectures.
- Understand the difference between technical architecture and systems architecture.
- Explain the components essential to all technical architectures.
- Learn how the technical architecture components integrate.
- Discuss systems architecture common to data warehousing.
- Understand key terminology related to both technical architecture and systems architecture.

Data Warehouse Architecture

Technical Architecture (Data Flow Architecture)

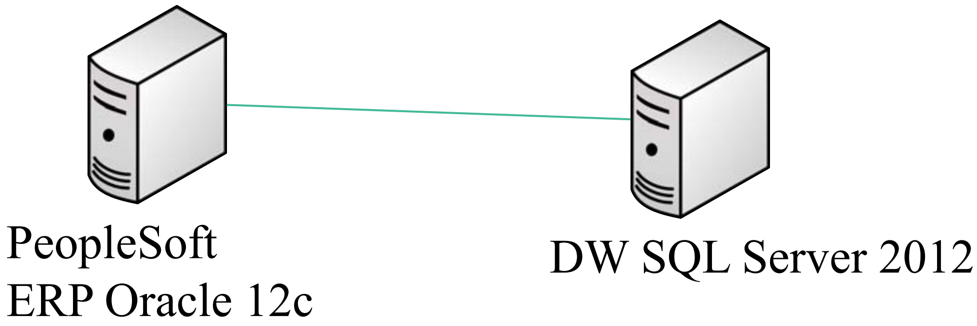
- How the data stores are arranged in the data warehouse and how data moves from data store to data store
- **Logical Architecture**



✚

System Architecture

- Physical configuration of systems, networks, and servers to support the technical architecture
- **Physical Infrastructure**

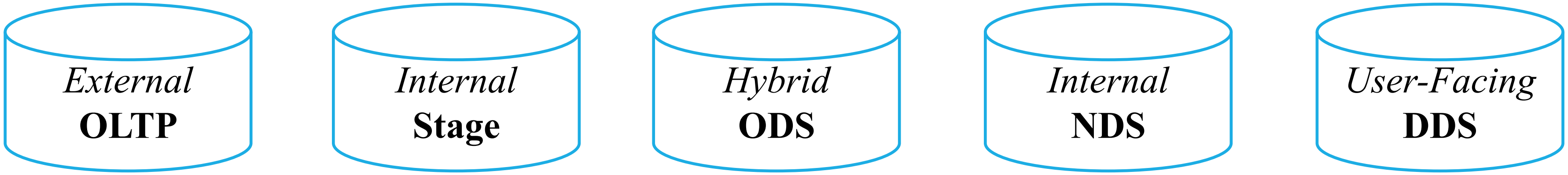


Components at a Glance

School of Information Studies
Syracuse University

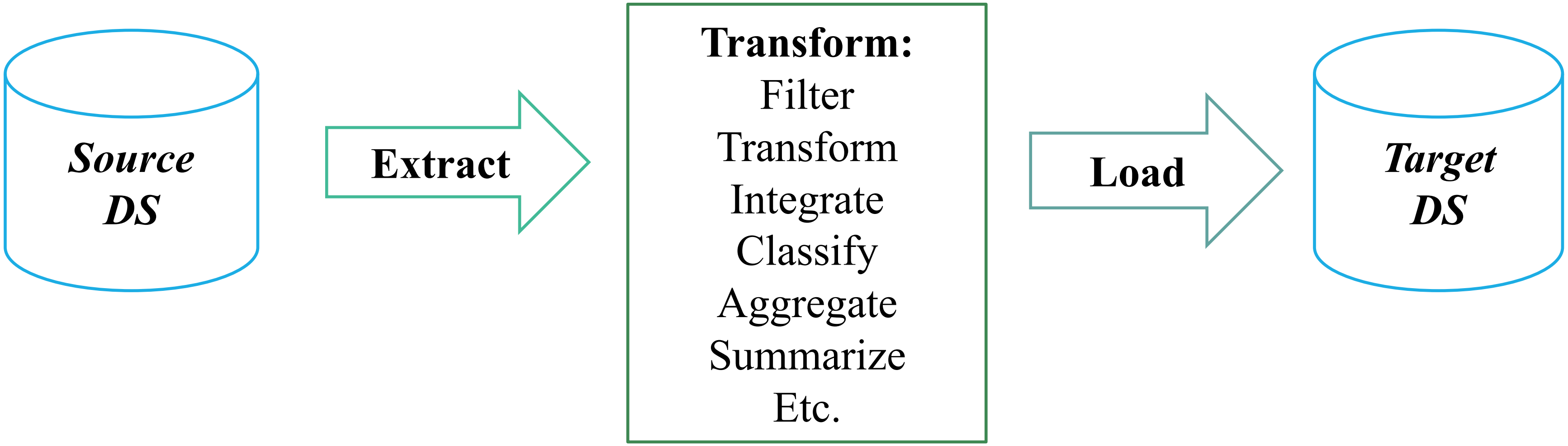
Data Stores: Data at Rest

- Typically stored in a DBMS but does not have to be:
  - Multidimensional database management systems
  - Hadoop HDFS
  - On the file system
  - Mainframe/legacy systems
  - Web services (Twitter, Weather, Etc…)
- Types:
  - **User Facing:** available to end users for query purposes via applications
  - **Internal:** used by the data warehouse only; not open to end users
  - **Hybrid:** combination of internal and user-facing
  - **External:** not part of the data warehouse

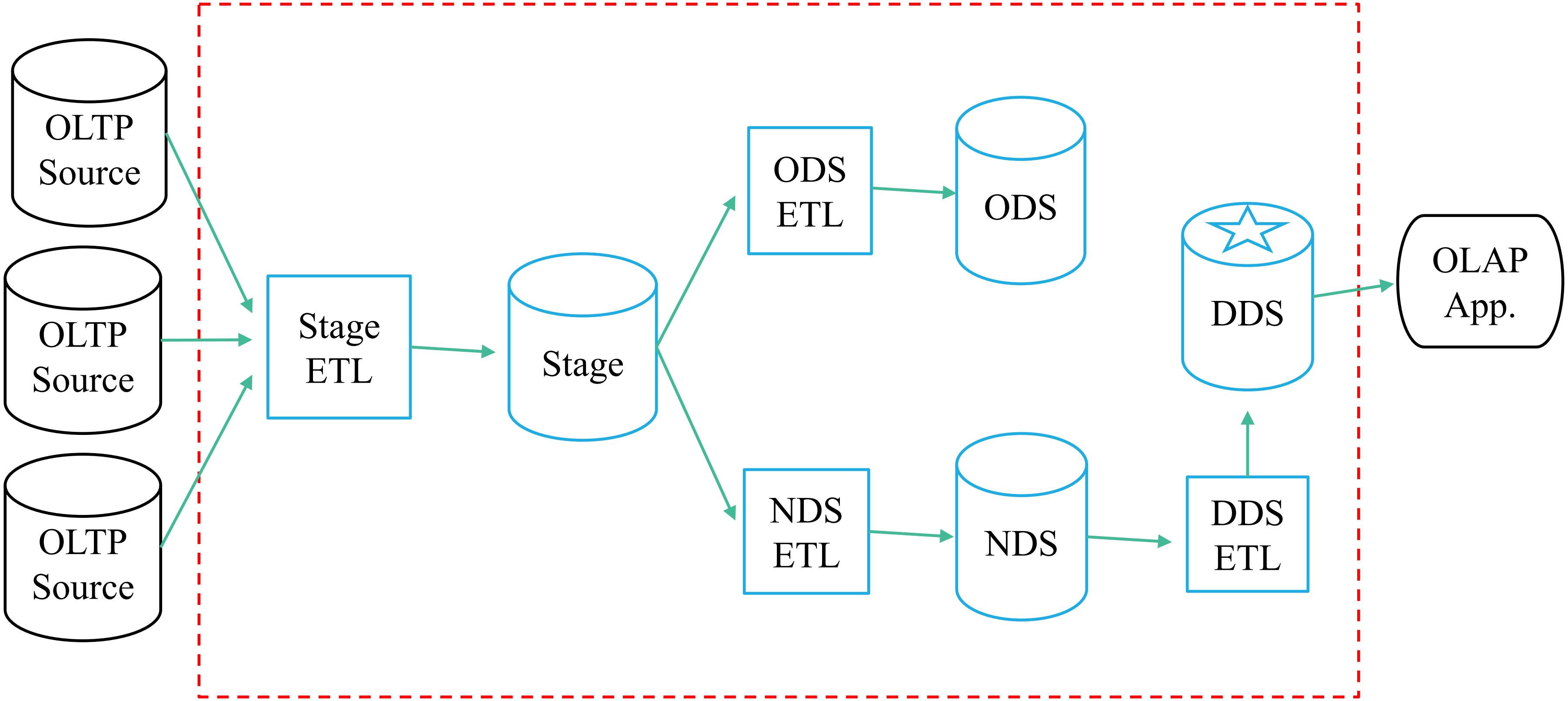


ETL: Data in Motion

- ETL = extract, transform, load.
- It is a process for moving data from one store (source) to another (target).
- It might be transformed along the way.



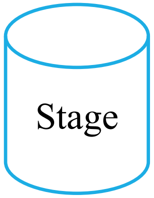
Data Architecture at a Glance



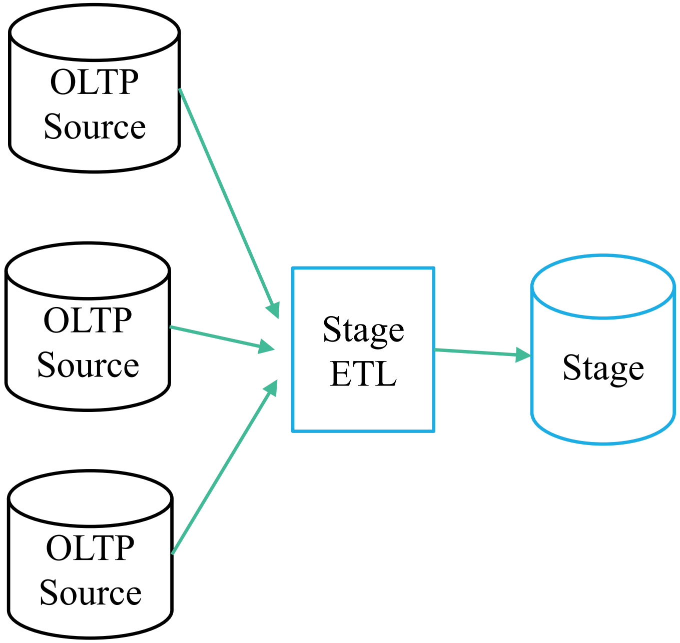
Stage

School of Information Studies
Syracuse University

Stage Data Store



- An ***internal*** data store. It is not user-facing.
- Stores extracts from source systems acting as a source for other systems in the data warehouse.
- Reduces contention with source systems.
- Consolidates data from multiple sources.
- Change detection.
- Snapshot data to a point in time.

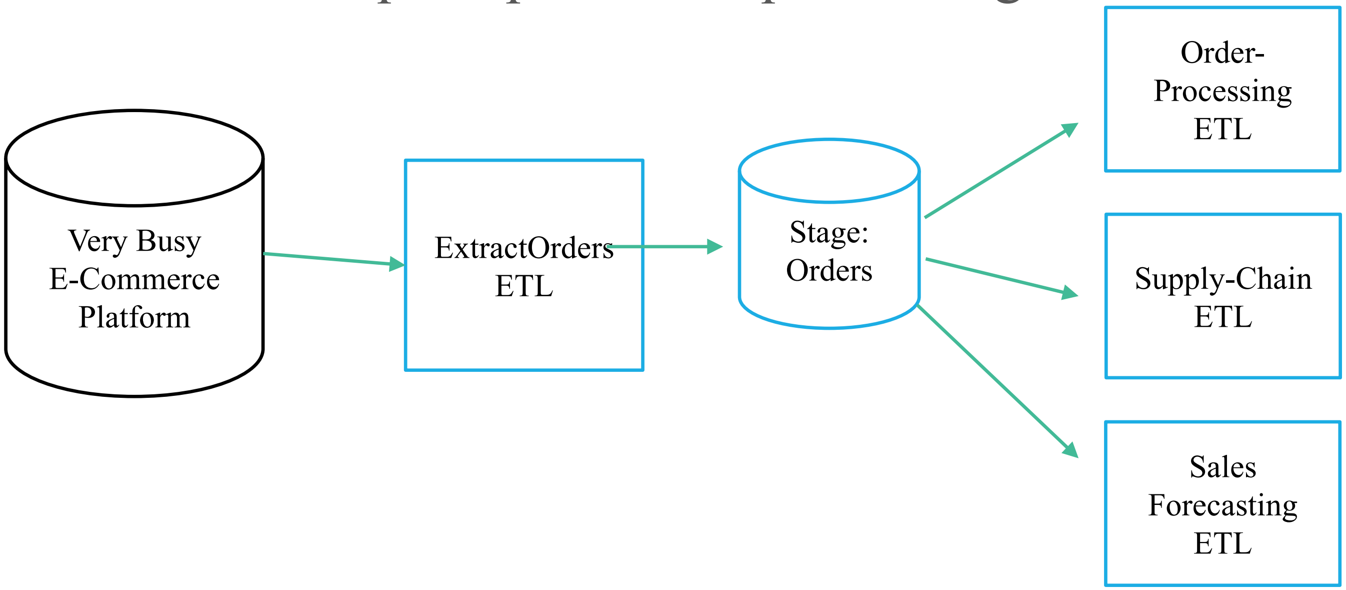


Four Reasons to Stage Data

1. Resource contention
2. Consolidation
3. Change detection
4. Snapshotting

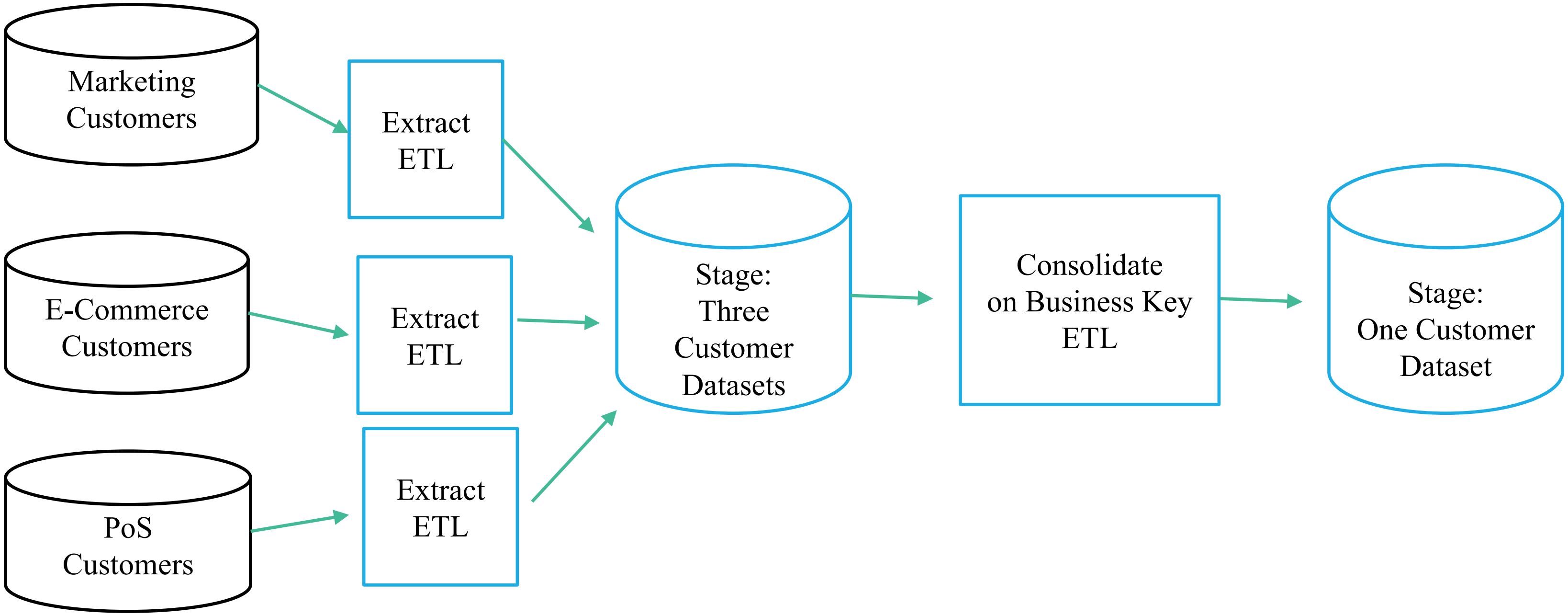
Resource Contention

- Staging data means we are not constantly querying the OLTP source for data.
- Stage queries the OLTP source.
- Each of the subsequent processes queries stage.



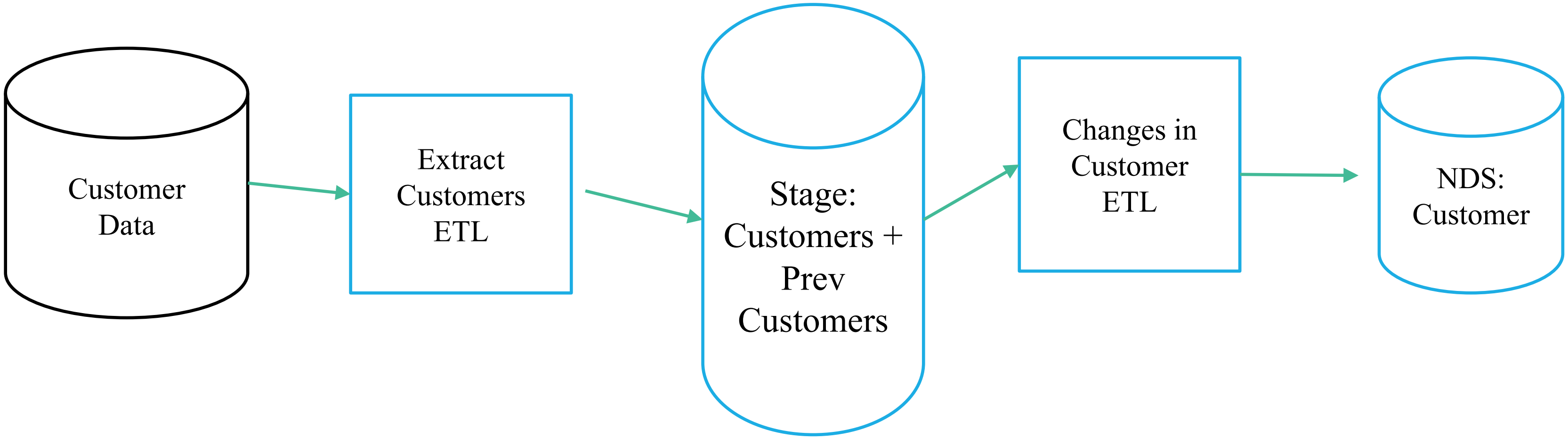
Consolidation

- Staging data provides us with a place where we can consolidate data from multiple sources.


Marketing Customers
E-Commerce Customers
PoS Customers
Extract ETL
Extract ETL
Extract ETL
Stage: Three Customer Datasets
Consolidate on Business Key ETL
Stage: One Customer Dataset

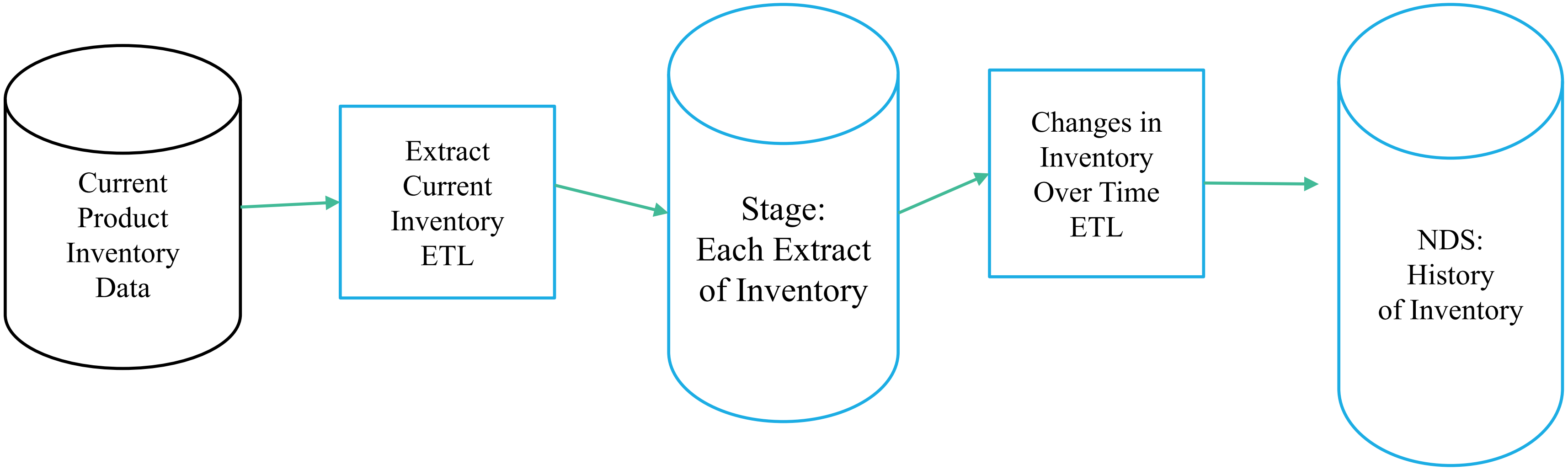
Change Detection

- Staging data gives us a reference point for detecting changes in new data.
- We can compare new data to data already have in stage to determine which data has changed.



Snapshotting

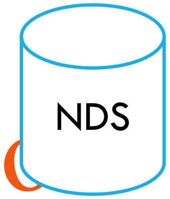
- We use snapshotting to build up time variance in data that is point in time.
- Allows us to build a history of data at source with no time variance.


Current Product Inventory Data
Extract Current Inventory ETL
Stage: Each Extract of Inventory
Changes in Inventory Over Time ETL
NDS: History of Inventory

Data Stores in the Data Warehouse

School of Information Studies
Syracuse University

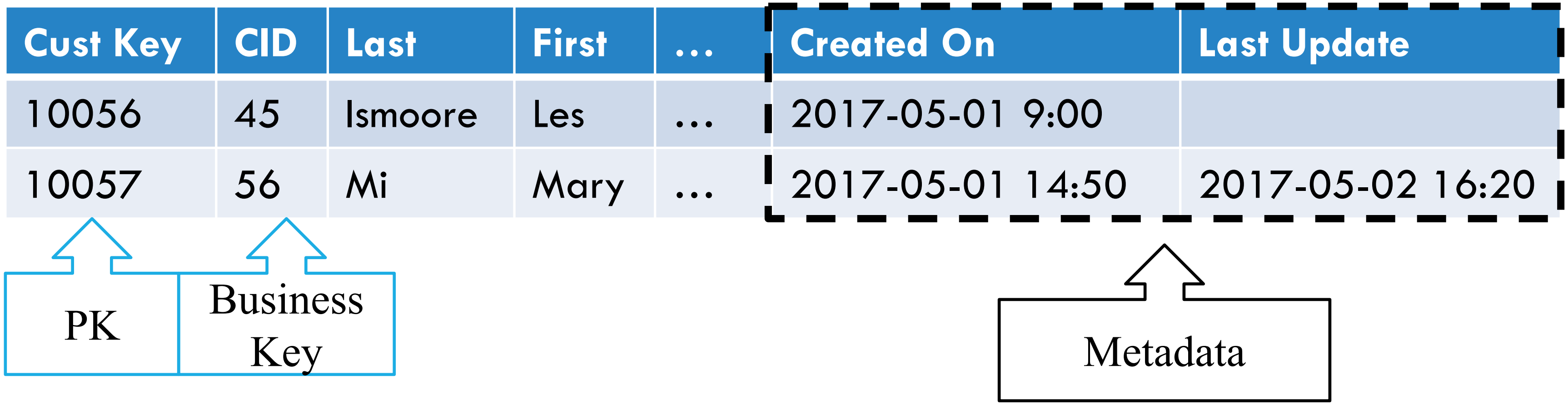
NDS: Normalized Data Sto



- An ***internal*** data store. Not user-facing.
- Used as the organization's source of a "***single version of the truth***" for other systems.
- **Subject-oriented, integrated, non-volatile, and time-variant** data from the OLTP sources.
- Stored in third normal form, to reduce redundancy.
- Use as a source for **data marts** and **decision support systems**, which use DDS.
- **Grows** in size over time due to **historical data**.

Example of NDS Data

- Normalized to 3NF
- PK different than source PK
- Metadata columns to track changes to data



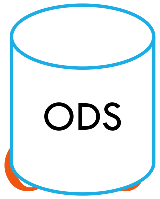
Cust Key	CID	Last	First	…	Created On	Last Update
10056	45	Ismoore	Les	…	2017-05-01 9:00	
10057	56	Mi	Mary	…	2017-05-01 14:50	2017-05-02 16:20

PK

Business Key

Metadata

ODS: Operational Data St



- A ***hybrid*** data store. Parts are internal, parts are user-facing.
- **Integrated, detailed, volatile, and current** data from source systems.
- Key differences:
  - **Volatile:** Data are updated and removed to reflect current.
  - **Consolidated** from disparate sources.
- Does not grow over time. References a point in time, which is typically "***now.***"
- Structured differently than NDS or DDS and, therefore, should be **stored as a separate DBMS**.

Example of ODS Data

- Consolidated from multiple sources
- Tells an important informational picture of "right now"
- Not time-variant but consolidated.

Order ID	Amount	Customer	Status	…	Last Update
10056	$1500.00	Les Ismoore	Packaging	…	2017-05-02 15:30
10057	$3500.00	Mary Mi	Shipping	…	2017-05-02 16:20

Metadata

DDS: Dimensional Data Sto

- A ***user-facing*** data store.
- **Subject-oriented, integrated, non-volatile, and time-variant** data from source systems.
- Stored in **dimensional format** to support ad hoc analytical query by end users and decision-support systems.
    - RDBMS → Star schema
    - MDBMS → Cube
- **Grows** in size over time due to **historical data**.
- Data are **consolidated** and denormalized. So no single version of the truth, but its easier for business users to query.

Example of DDS Data

- Same data in there more than once, for historical purposes.
- Only one row is current.

Product Key	Product Description	Product Code	Department	Effective Date	Expiration Date	Current Row
11981	Stapler, Red	ST901	Accessories	4/7/2010	9/1/2011	N
20342	Stapler, Red	ST901	Supplies	9/2/2011	3/31/2013	N
45393	Stapler, Red	ST901	Office Supplies	4/1/2013	12/31/9999	Y

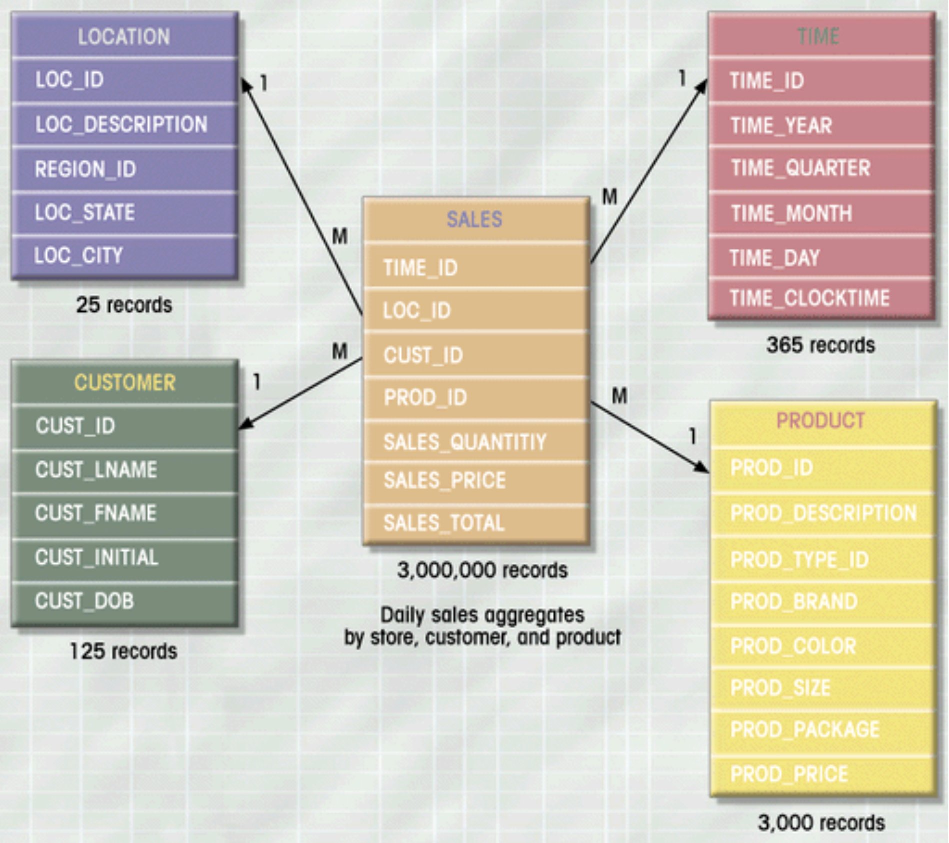
PK

Business Key

Metadata

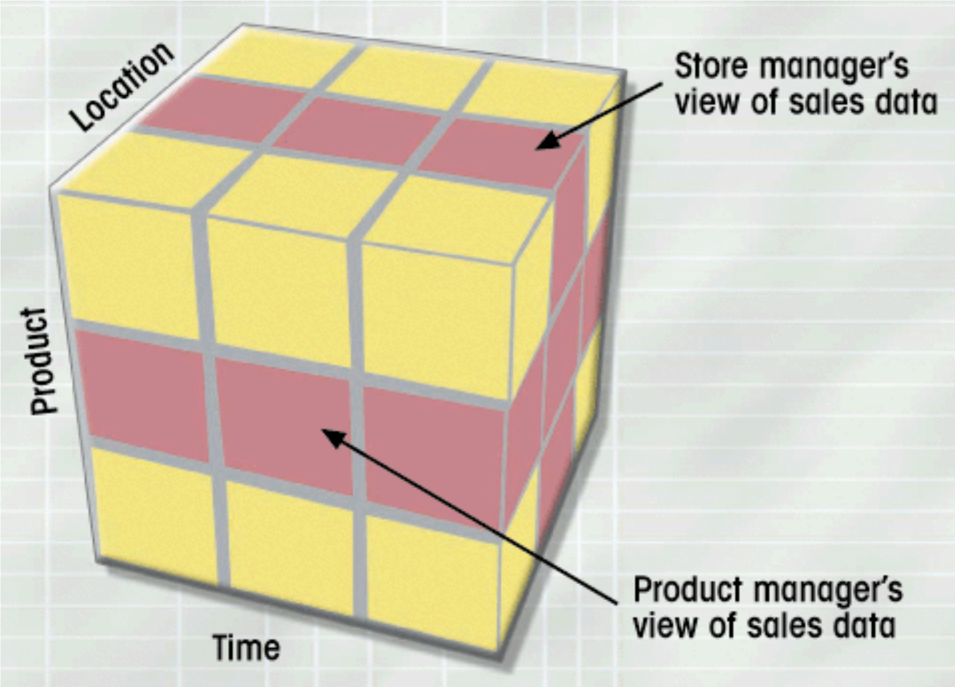
DDS: ROLAP/Star Schema/ Data Mart

- When the DDS is implemented in a **relational DBMS**, it is called ROLAP.
- Relational online analytical processing (ROLAP).
- The schema is a **star schema** because of the consistent M-1 structure between fact and dimension tables.
- A single-star schema is known as a **data mart**.



MDS: MOLAP/Cube

- When the DDS is implemented in a **Multi-Dimensional DBMS** it is called MOLAP.
- Multidimensional online analytical processing (MOLAP).
- Facts are pre-aggregated across all dimensions for improved performance.
- This is called a **cube**.


Location
Store manager's view of sales data
Product
Product manager's view of sales data
Time

- Faster query time and support for semantic metadata.

Example of Semantic Metadata

Relational (No Semantics)
Sorts Alphabetically ☹

Day of the Week ⬆
Friday
Monday
Saturday
Sunday
Tuesday
Thursday
Wednesday

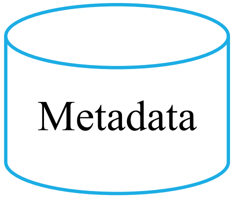
Multidimensional (Semantics)
Sorts by Day of the Week ☺

Day of the Week ⬆
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday

Metadata

School of Information Studies
Syracuse University

Metadata



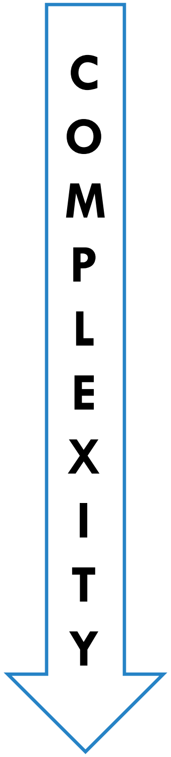
Metadata

- **Metadata** means “*data about the data*." It is an essential part of the data warehouse technical architecture.
- Metadata is **internal.**
- Three types:
  - **Technical metadata:** Infrastructure oriented. Indexes, table partitions, data types, data transformations. Security.
  - **Business metadata:** User oriented. Data structure definitions, data dictionaries, implicit data hierarchies, data quality screens.
  - **Process metadata:** System oriented. Performance metrics and measurements. Auditing the ETL processes.

Overview

School of Information Studies
Syracuse University

Common Technical Architectures



COMPLEXITY

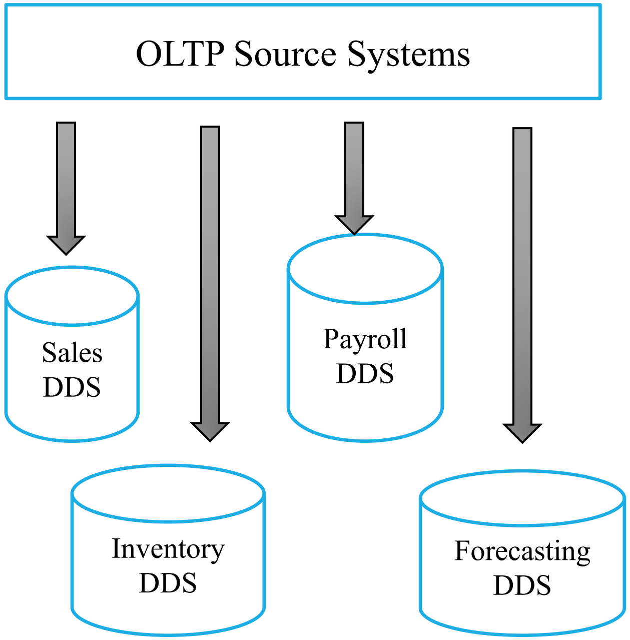
1. Independent data marts
2. Centralized
3. Enterprise bus architecture
    1. With ODS (ODS + DDS)
4. Hub and spoke
    1. With ODS
5. Federated With ETL
    1. Federate with EII

Independent Data Marts

School of Information Studies
Syracuse University

1. Independent Data Marts

- Ad hoc “grassroots” technical architecture.
- Easy to get started with, difficult to scale.
- Departmentalized, lacking enterprise focus.
- No data consistency or data integration between data marts.
- Data marts do not share dimensions
- Data are sourced independently for each data mart.

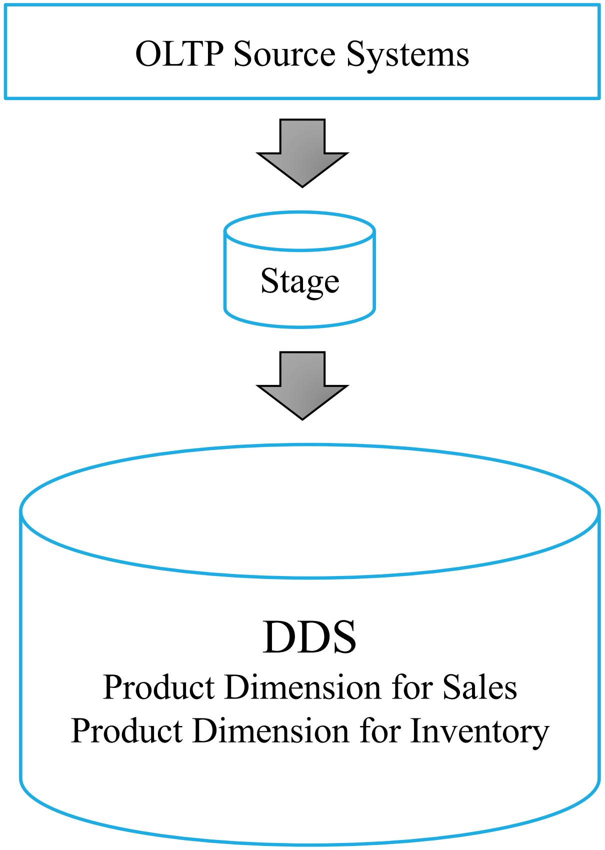


Centralized

School of Information Studies
Syracuse University

2. Centralized

- Next step up from independent data marts.
- Data marts are consolidated into a single DDS.
- There is still lack of integration among the dimensions, and there are copies of dimension for each data mart that requires them.
- More enterprise focus but still no data consistency among data marts.

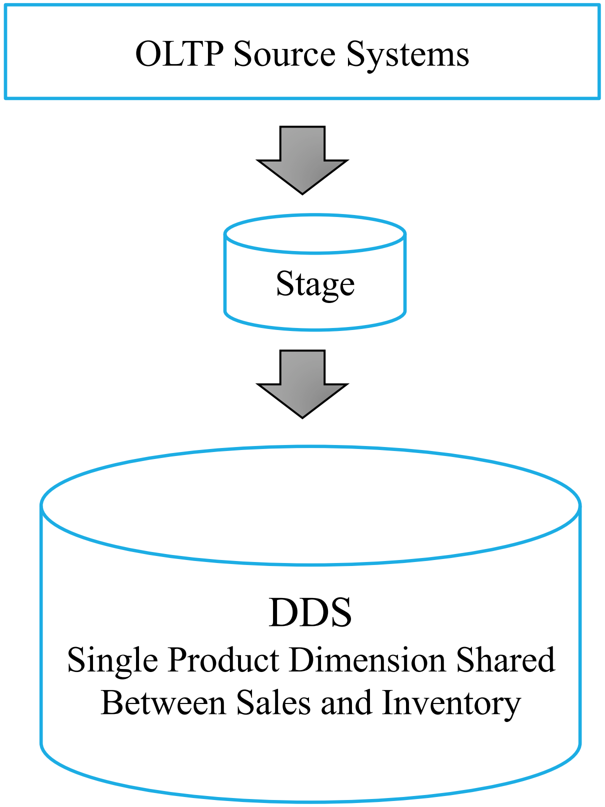


Enterprise Bus

School of Information Studies
Syracuse University

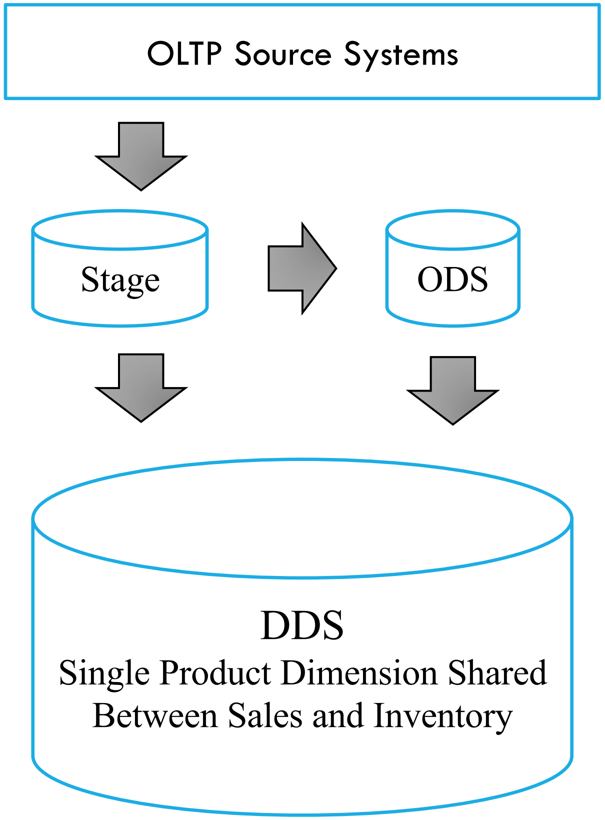
3. Enterprise Bus

- Next step up from centralized.
- Like centralized, all data marts in the DDS.
- Conformed dimensions, meaning they are reused across the data marts. Just a single dimension for master data.
- Difficult to achieve because enterprise focus is required when building data marts.
- This is the Kimball technical architecture



3.1 Enterprise Bus With ODS

- Variation on enterprise bus includes an ODS for reporting on current, consolidated data.
- ODS and stage if need be are the source of the DDS.
- Conformed dimensions like enterprise bus.

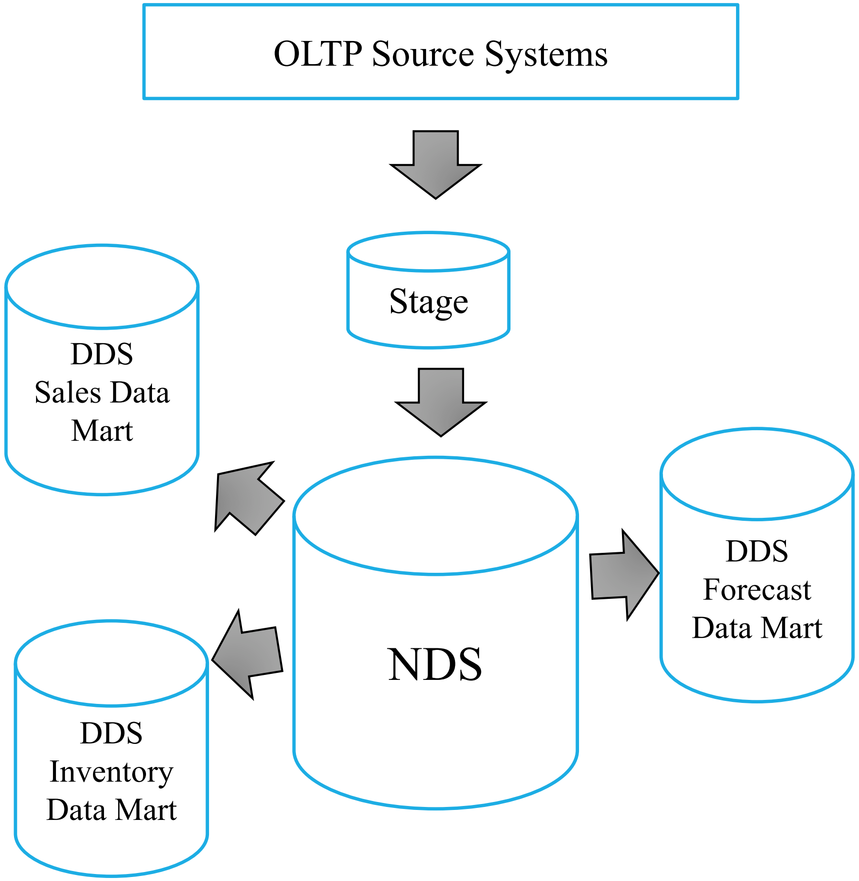


Hub and Spoke

School of Information Studies
Syracuse University

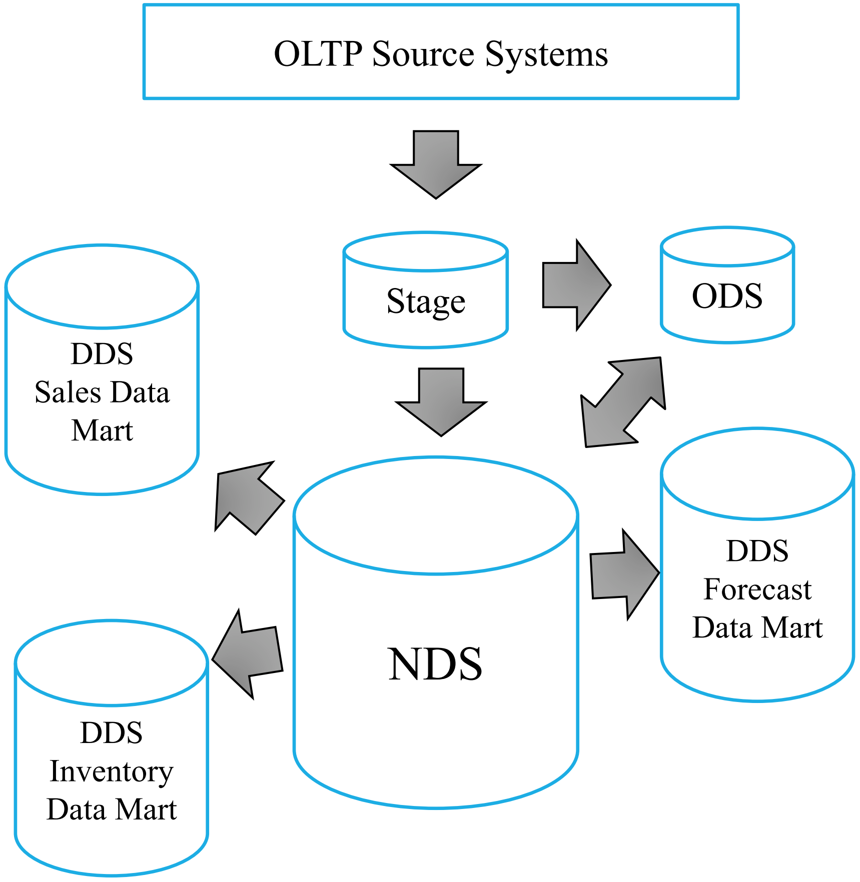
4. Hub and Spoke

- Next step up from enterprise bus.
- Data sourced systematically for “single version of the truth."
- Dimensional models in data marts are distributed and sourced from the NDS.
- Added complexity of 3NF data but reduced complexity of conformed dimensions.
- Inmon technical architecture.
- ODS can be added between stage and NDS just like with enterprise bus.



4.1 Hub and Spoke With ODS

- Full Inmon corporate information factory.
- ODS is consolidated and current version of data.
- ODS is sourced from stage or the NDS.
- ODS or stage can populate the NDS.

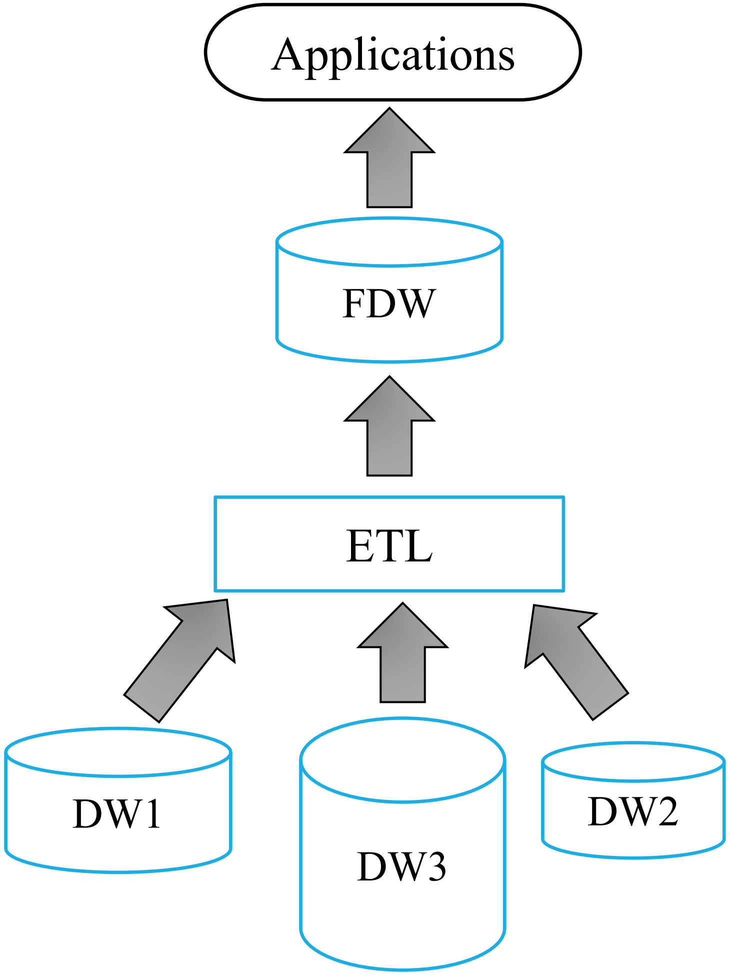


Federated Architectures

School of Information Studies
Syracuse University

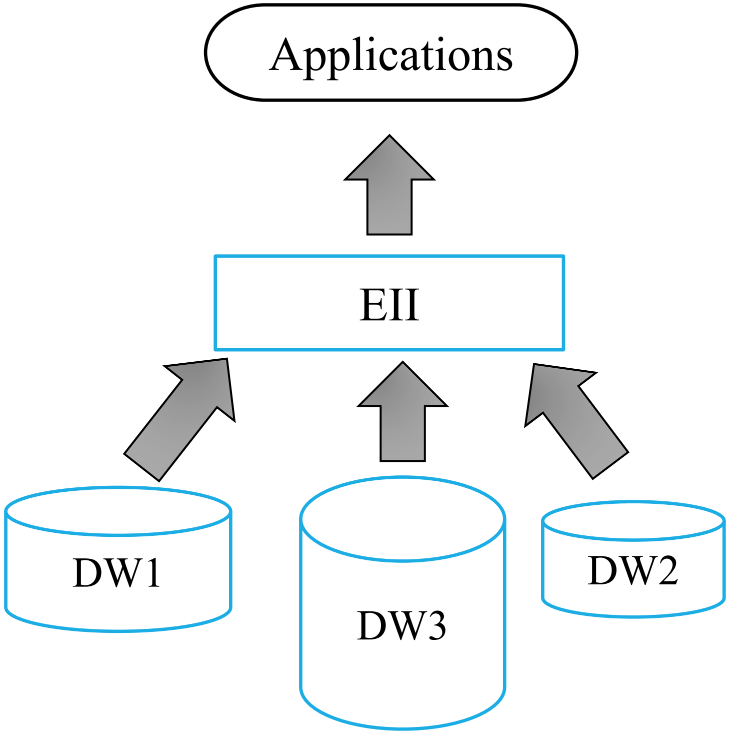
Federated With ETL

- Most complex technical architecture.
- Cases where you have several data warehouses, such as through mergers and acquisitions.
- ETL unifies disparate sources into a single federated data warehouse.
- Used to integrate existing data marts, warehouses, and legacy applications into a single logical data warehouse.



Federated With EII

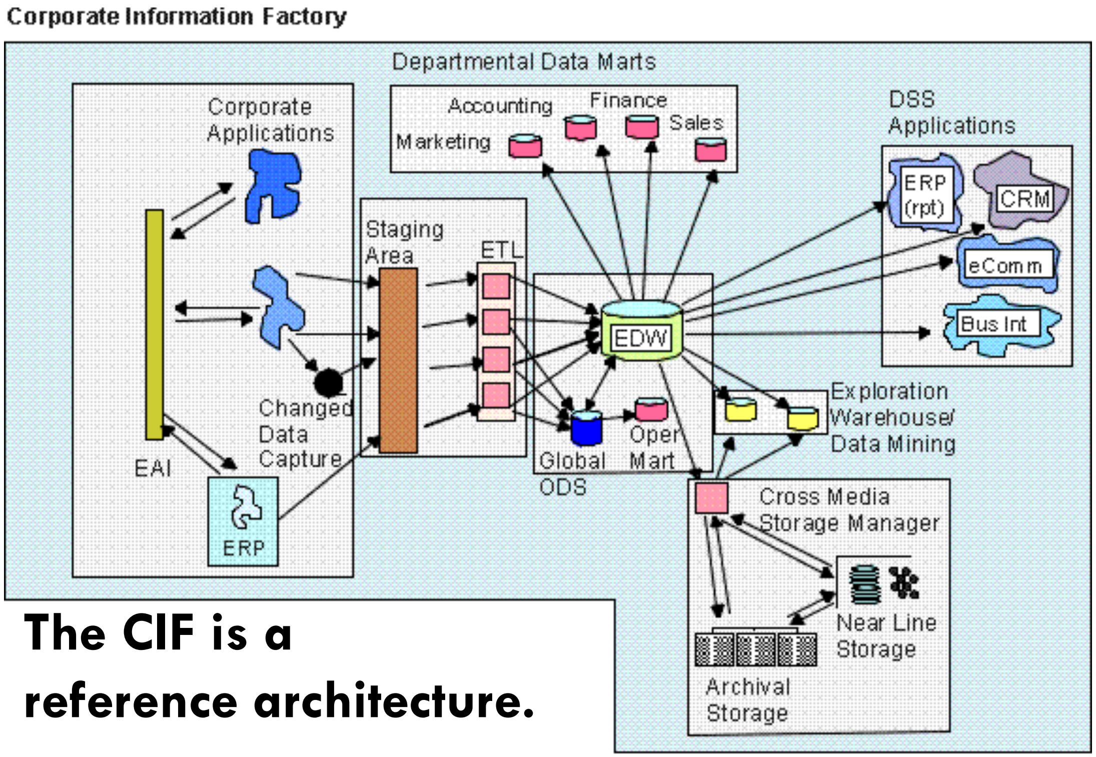
- EII = enterprise application integration.
- Federation is achieved through the EII application or by building your own services.
- Outputs are aggregated on the fly so there is no need to consolidate data into a single data warehouse.



What Is the CIF?

School of Information Studies
Syracuse University

Inmon's Corporate Information Factory

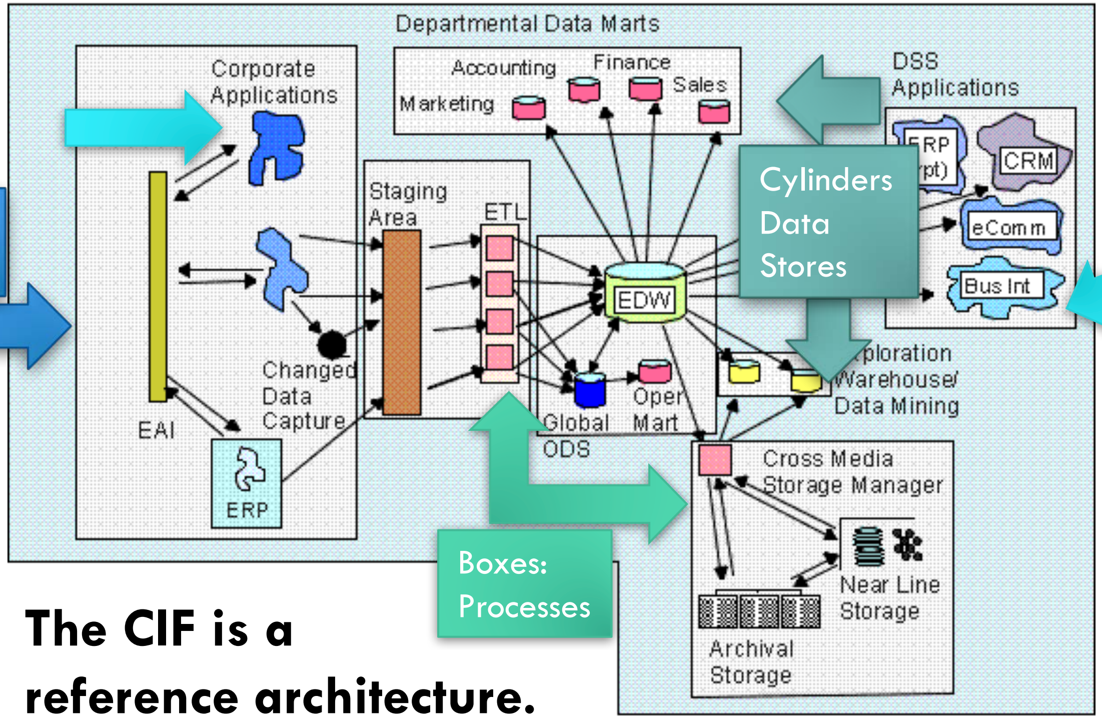


The CIF is a reference architecture.

by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

Understanding the Diagram

Corporate Information Factory



Rectangles: Components

Cylinders Data Stores

Blobs: Applications

Boxes: Processes

The CIF is a reference architecture.

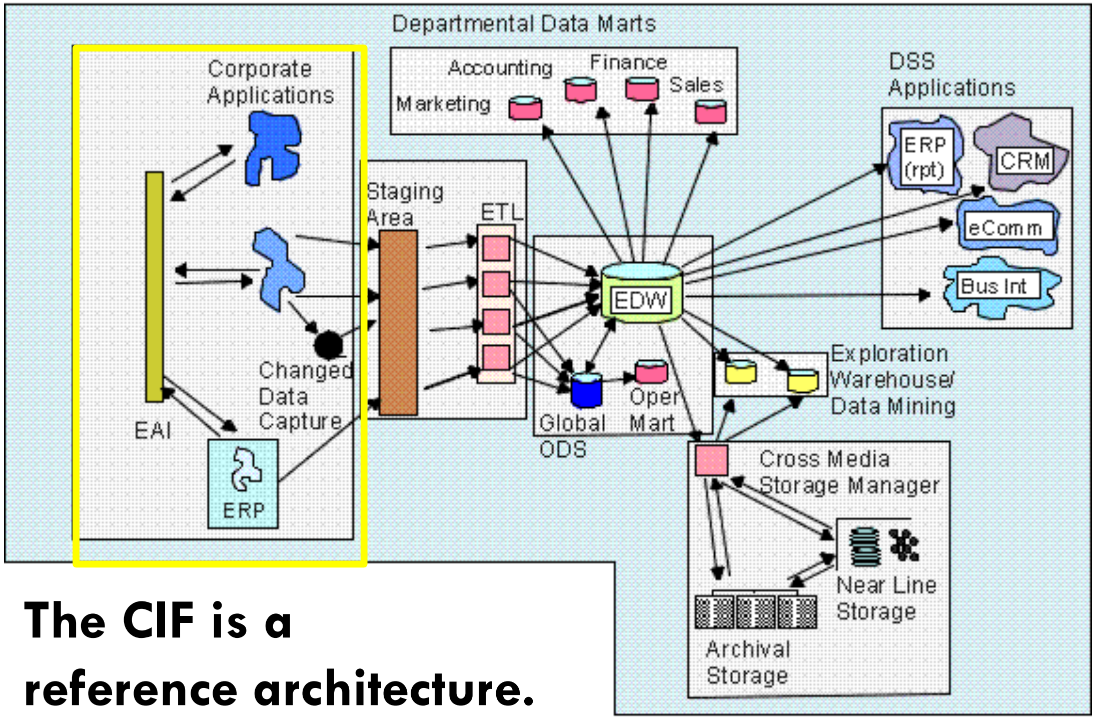
by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

External World

School of Information Studies
Syracuse University

External World and Applications

Corporate Information Factory



The CIF is a reference architecture.

by Bill Inmon and Claudia Imhoff

Copyright ©2001, all rights reserved.

External World and Applications

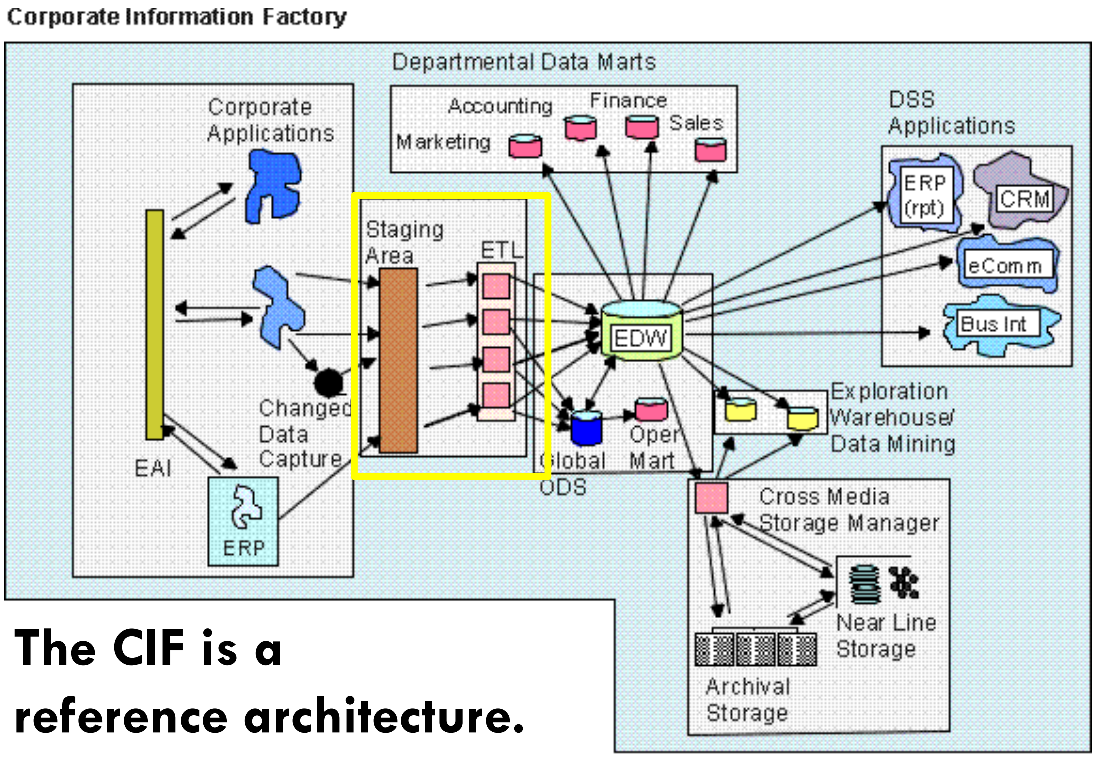
- **External world**: The people and systems that generate operational data. Transactional in nature. Called external world because they can come from *anywhere.*
- **Examples**: ERPs, business applications, Internet data, logs, external data streams (social media).
- The **data inputs/data sources** for the CIF.

❖ ***These are the OLTP source systems.***

IM&T Layer

School of Information Studies
Syracuse University

Integration and Transformation Layer



The CIF is a reference architecture.

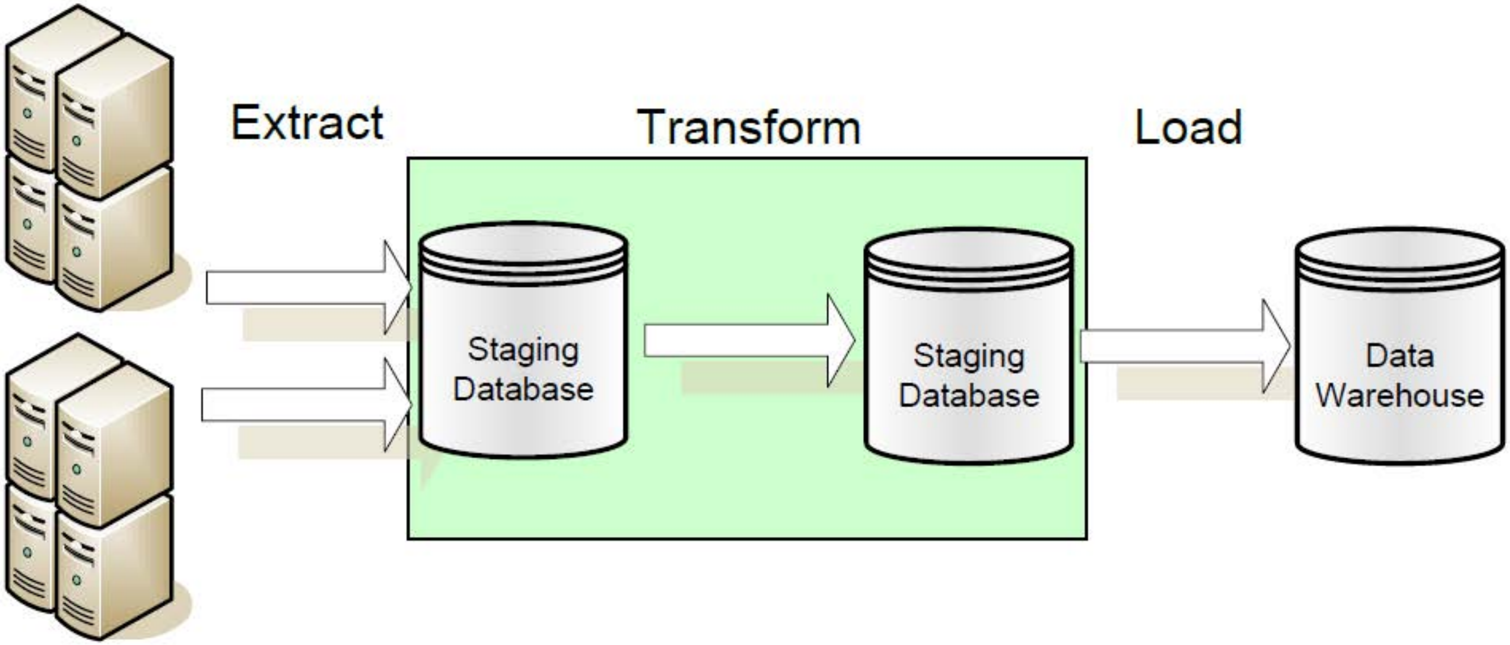
by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

Integration and Transformation Layer

- **I&T layer:** Takes unintegrated data from multiple sources and integrates and consolidates it.
- Computer programs are written to transform data from the **external world** into **corporate data**.
- The data come from a variety of sources and in both **structured** and **unstructured** formats.

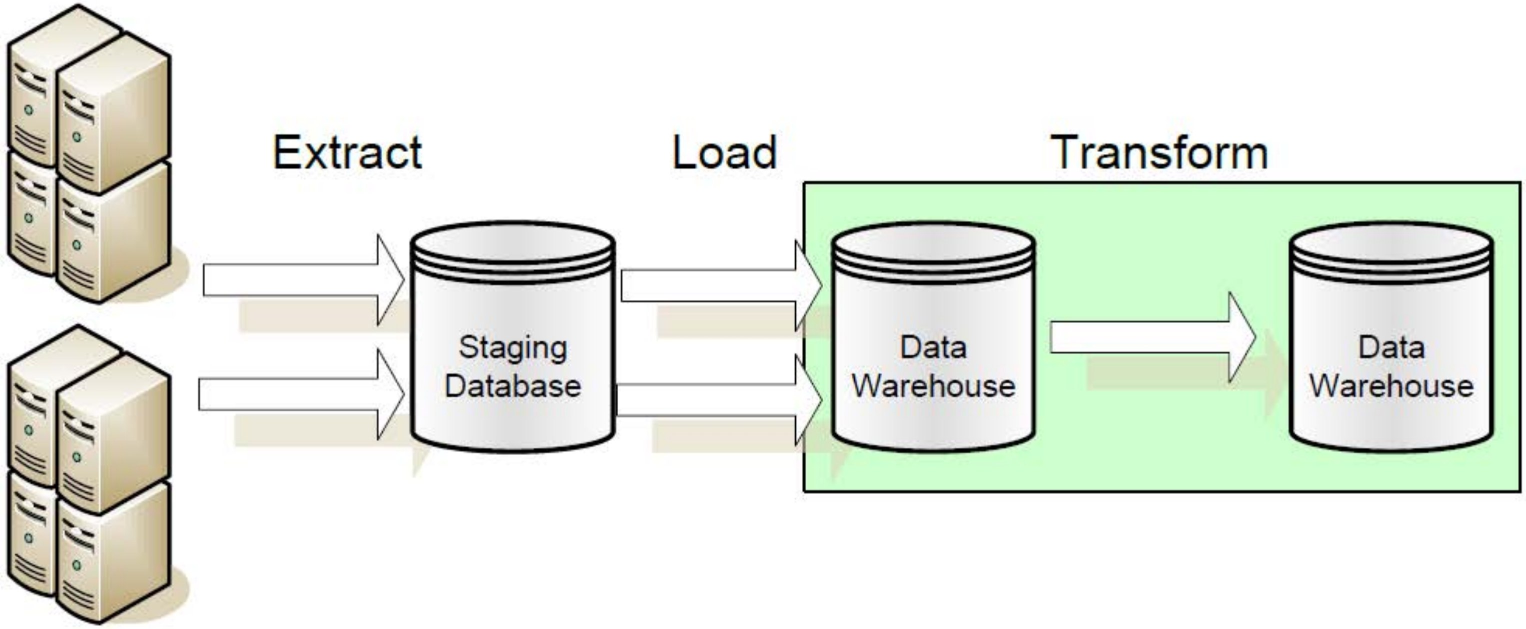
❖ ***This is staging data store and ETL.***

ETL: Extract, Transform, Load



- The data transformation occurs over **staged data**.
- The source data are **not stored in the warehouse**.
- Data transformation processing does not occur in the data warehouse.

ELT: Extract, Load, Transform



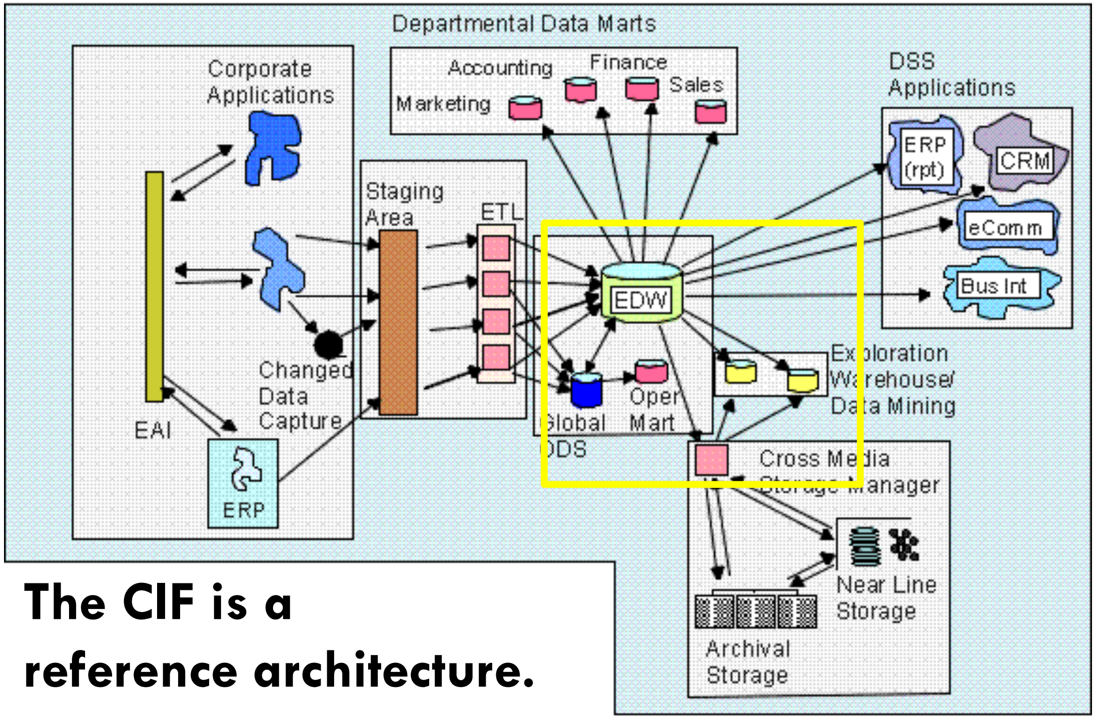
- The data transformation occurs over **warehoused** data.
- The staged data **are stored in the warehouse**.
- Data transformation processing occurs in the data warehouse.

EDW, ODS, and Data Marts

School of Information Studies
Syracuse University

ODS and EDW

Corporate Information Factory



The CIF is a reference architecture.

by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

ODS and EDW

- Same definition of ODS as before:
  - Current, consolidated data
  - Updated to be current; does not grow over time.
- Inmon's EDW is a DNS:
  - His notion of data warehouse is just NDS.
  - Most people think of the entire CIF as a data warehouse.

Enterprise Data Warehouse

- **Subject-oriented, integrated, summarized, and time-variant** data from the external world and applications.
- Stored in third normal form, to reduce redundancy, a NDS.
- Receives data from **I&T layer** and the **ODS.**
- Use as a source for **data marts** and **decision-support systems**, which are stored as dimensional models.
- **Grows** in size over time due to historical data.
- The heart of the CIF.

ODS vs. EDW (NDS)

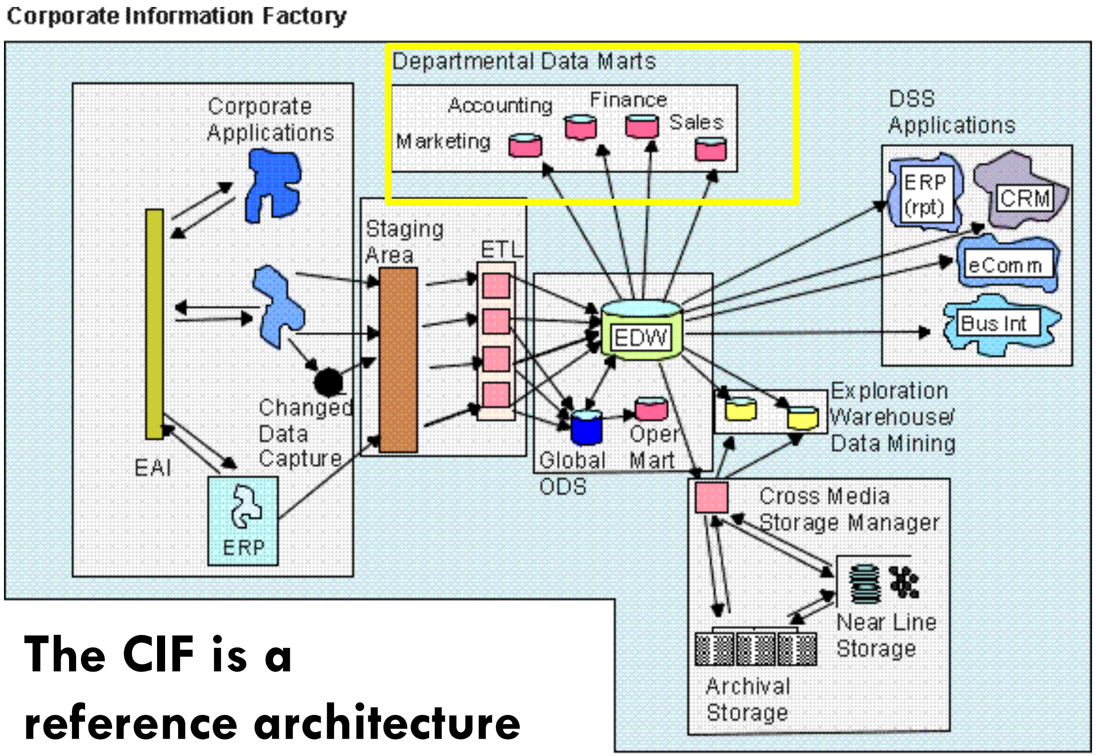
Characteristic	**ODS**	**EDW/NDS**
Primary Purpose	Run the business on a current basis	Support managerial decision-making
Design Goal	Performance throughput, availability of information	Single version of the truth
Subject-Oriented	Yes	Yes
Integrated	Yes	Yes
Detailed Data	Yes	Yes
Data Changes	**Yes**	**No**
Time of Data	**Current data**	**Historical snapshots**
Updates	**Frequent small updates**	**Periodic batch updates**
Queries	**Simple queries on a few rows**	**Complex queries on several rows**

ODS and EDW/NDS Cannot Share the Same System

I need fast updates!
I need query performance!

You can't have both in a relational DBMS! Why?

Data Marts



The CIF is a reference architecture

by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

Data Marts

- A collection of data tailored to the informational needs of a **department** or **business process**.
- Stored in **dimensional models**, with **fact** and **dimension** tables.
- Easy to control, low cost, and customizable due to their **limited scope**.
- Receive their data source from the "single version of the truth" **EDW.**
- Are source data for **online analytical processing** (ROLAP/MOLAP) engines.
- Like the DDS but does not necessarily have to be well integrated into a single data store.

OLAP: Online Analytical Processing

ROLAP

- Uses a **relational database management** system.
- Data design is the **star schema.**
- Built on well-known relational concepts.

MOLAP

- Uses a **multidimensional database management** system.
- Data design is the **cube.**
- Highly flexible, includes metadata.

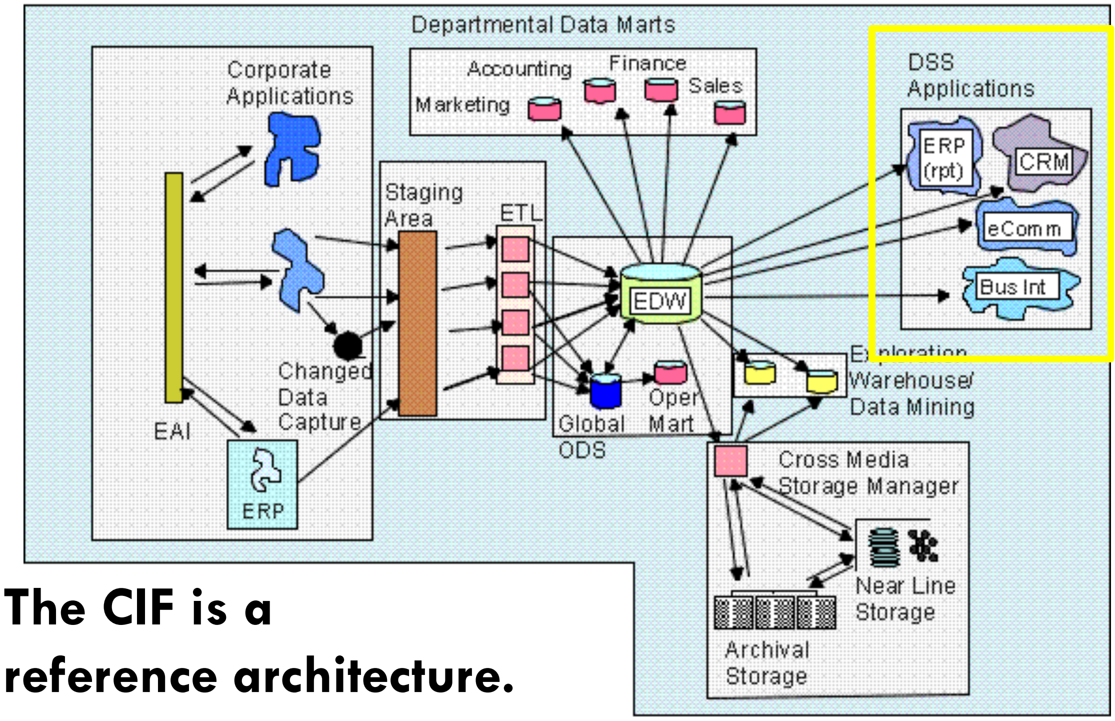
- Typical Kimball implementations have a ROLAP star schema feed the MOLAP cube.

Other Components

School of Information Studies
Syracuse University

DSS Applications

Corporate Information Factory



The CIF is a reference architecture.

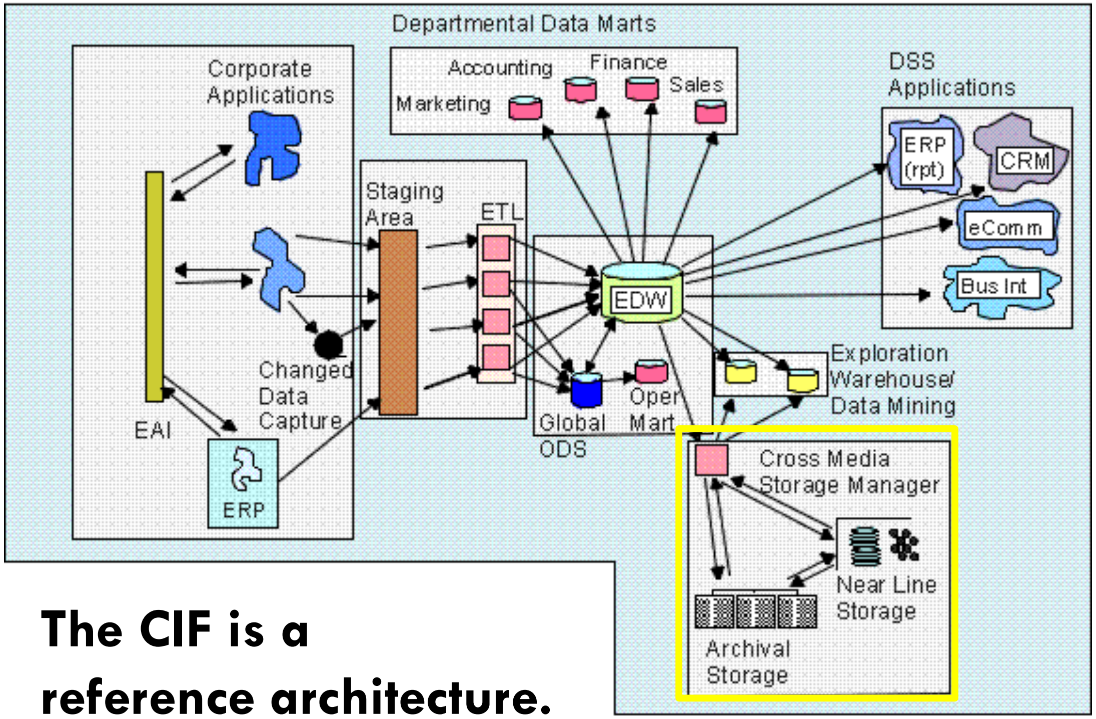
by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

Decision-Support Systems

- Assist with the decision-making process.
  - Do we extend credit to a customer?
  - Do we restock a product?
  - Which roads will likely require potholes to be filled this year?
- A form of business analytics
- Get their source data from the "single version of the truth" EDW.

Cross-Media Storage

Corporate Information Factory



The CIF is a reference architecture.

by Bill Inmon and Claudia Imhoff
Copyright ©2001, all rights reserved.

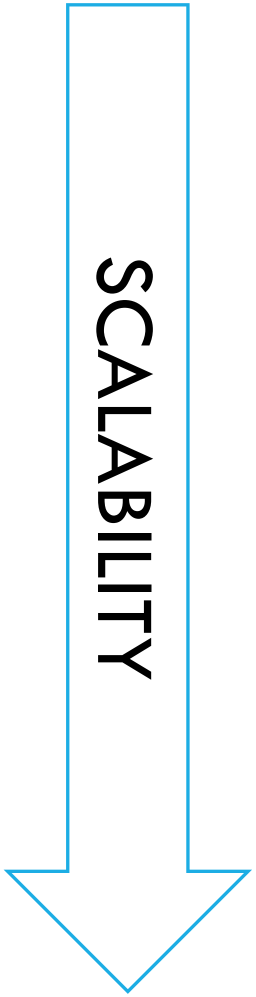
Cross-Media Storage Manager

- Stores historical data, which is infrequently accessed.
- Data are moved out of the EDW, which has high-end performant storage into more affordable storage with less performant access times.
- A process exists to enable some transparency in the retrieval process.
- Data movement can coincide with regulatory actions.

Overview

School of Information Studies
Syracuse University

System Architectures


SCALABILITY

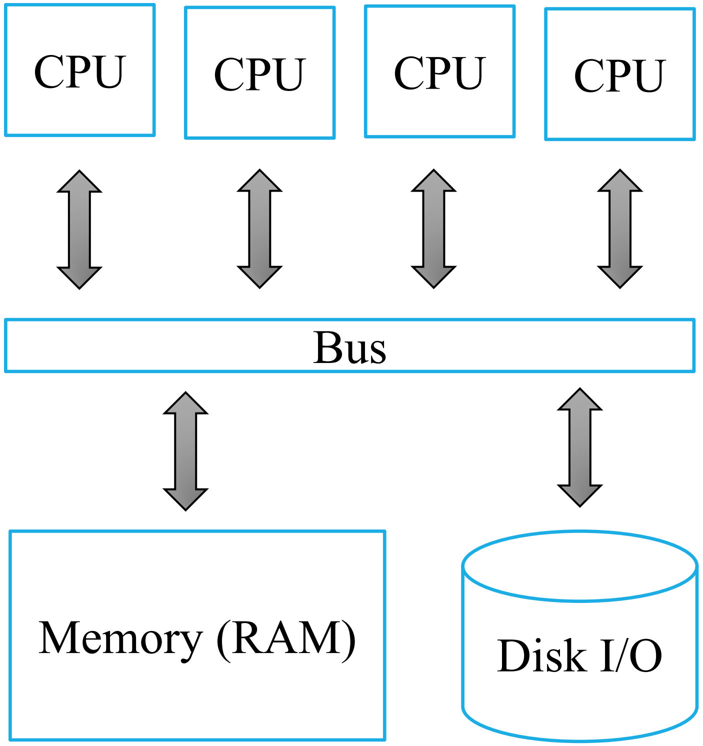
1. SMP
   symmetric multiprocessing
2. MPP
   massively parallel processing
3. Hadoop
   MapReduce/HDFS

SMP, MPP, and MapReduce

School of Information Studies
Syracuse University

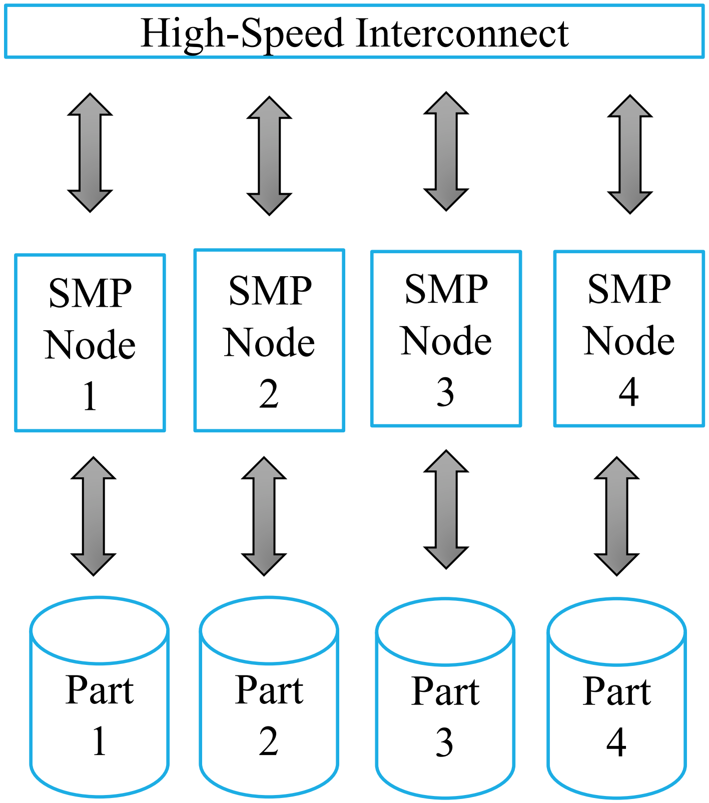
SMP

- Single system with multiple CPUs.
- Shared bus, memory, and I/O.
- CPUs share resources on a single system.
- Scales up but not out.
- Vendors: Microsoft, Oracle, IBM, Postgres



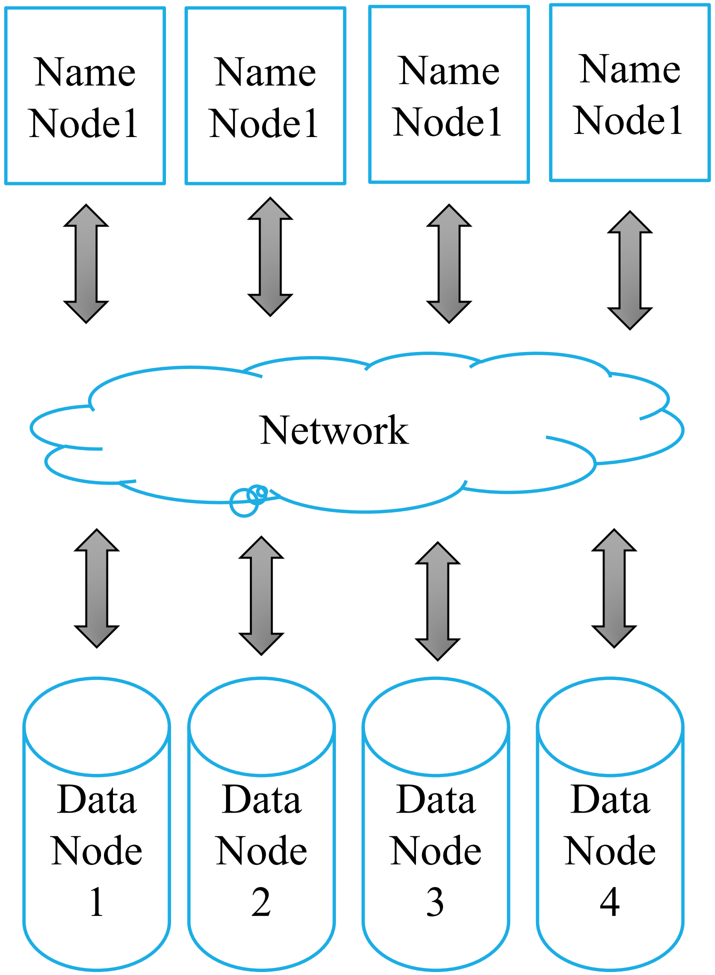
MPP

- Multiple SMP nodes interconnected to form a single cluster.
- Data are partitioned across nodes in the cluster.
- Single control node to orchestrate queries and data management.
- Processing and data partitions are tied together.
- Specialized hardware; difficult to scale out once configured.
- Vendors: Teradata, IBM Netezza, Vertica, Oracle, Microsoft



MapReduce

- General-purpose distributed batch processing framework.
- Fault tolerant.
- Runs on affordable commodity hardware.
- Processing and data are decoupled and distributed over the network.
- Slower query execution than MPP.
- Vendors: IBM, Cloudera, Hortonworks, MapR



Comparisions

Factor	SMP	MPP	MapReduce
Workloads	Small	Large	Very Large
Scale	Up	Up and Out	Up and Out
Technology Cost	Low	Very High	Low to High
Implementation Cost	Low	Moderate	High
Distributed	No	Processing and Data together	Processing and Data Independent
SQL Compliant	Yes	Yes	Somewhat
Fault-Tolerance	No	No	Yes
Nodes	1	100s	1000s
Hardware	Appliances and Commodity	Mostly Appliances	Commodity