Data Warehousing

UPDATED BY **ALON BRODY**
ORIGINAL BY **DAVID HAERTZEN**

CONTENTS

WHAT IS DATA WAREHOUSING?

Warehousing is a process for collecting, storing, and delivering decision-support data for some or all of an enterprise. Data warehousing is a broad subject that is described point by point in this Refcard. A data warehouse is one of the artifacts created in the data warehousing process.

William (Bill) H. Inmon has provided an alternate and useful definition of a data warehouse: "A subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

As a total architecture, data warehousing involves people, processes, and technologies to achieve the goal of providing decision-support data that is consistent, integrated, standardized, and easy to understand.

See the book *The Analytical Puzzle: Profitable Data Warehousing, Business Intelligence, and Analytics* (ISBN 978-1935504207) for details.

WHAT A DATA WAREHOUSE IS AND IS NOT

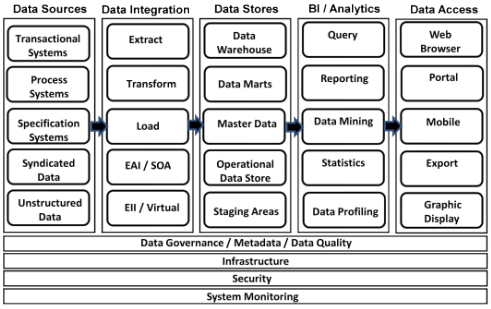
A data warehouse is a database whose data includes a copy of operational data. This data is often obtained from multiple data sources and is useful for strategic decision making. It does not, however, contain original data.

"Data warehouse," by the way, is not another name for "database." Some people incorrectly use the term "data warehouse" as if it's a generic name for a database. A data warehouse does not only consist of historic data, it can be made up of analytics and reporting data, too. Transactional data that is managed in application data stores will not reside in a data warehouse.

DATA WAREHOUSE ARCHITECTURE

DATA WAREHOUSE ARCHITECTURE COMPONENTS

The data warehouse's technical architecture includes: data sources, data integration, BI/Analytics data stores, and data access.



DATA WAREHOUSE TECH STACK

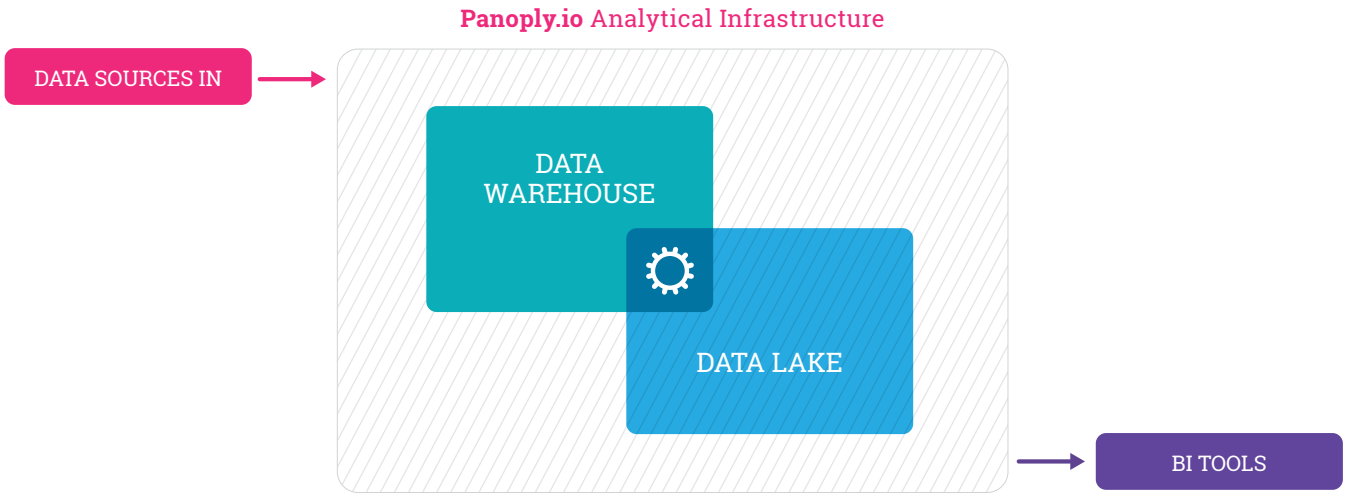
PLATFORM NAME	DESCRIPTION
Metadata	A software tool that contains data that describes other data. The two kinds of metadata are: business metadata and technical metadata.
Repository	A software tool that enables the design of data and databases through graphical means. This tool provides a detailed design capability that includes the design of tables, columns, relationships, rules, and business definitions.
Data Modeling Tool	A software tool that supports understanding data through exploration and comparison. This tool accesses the data and explores it, looking for patterns such as typical values, outlying values, ranges, and allowed values. It is meant to help you better understand the content and quality of the data.

Continued on next page

panoply.io

Self-Optimizing Data Warehouse/Data Lake

From ingestion to query, **Panoply** automates and simplifies data analytics, eliminating the overhead of preparing and modeling data, and managing cloud infrastructure.



Start Your Free Trial

Any Data. Any Scale

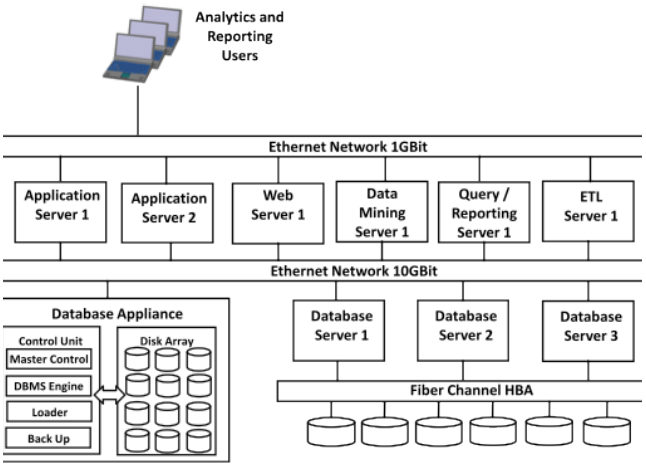
With its ETL-less integration pipeline, Panoply.io connects to all sorts of structured and semi-structured data sources – absorbing billions of writes daily without a line of code, and allowing you to capture and process your data at lightning speed.

Google Analytics · hadoop HDFS · Amazon S3 · MySQL · mongoDB. · Amazon Redshift · tableau · looker · Google AdWords · facebook · salesforce · MailChimp

PLATFORM NAME	DESCRIPTION
Data Profiling Tool	ETL (extract, transform, and load) tools, as well as realtime integration tools like the ESB (enterprise service bus) software tools. These tools copy data from place to place and also scrub and clean the data.
RDBMS (Relational Database Management System)	Software that stores data in a relational format using SQL (Structured Query Language). This is really the Database system that is going to maintain robust data and store it. It is also important to the expandability of the system.
MOLAP (Multidimensional OLAP)	Database software designed for data mart-type operations. This software organizes data into multiple dimensions, known as "cubes," to support analytics.
Big Data Store	Software that manages huge amounts of data (relational databases, for example) that other types of software cannot.
Reporting and Query Tools	Business-intelligence software tools that select data through query and present it as reports and/or graphical displays. The business or analyst will be able to explore the data-exploration sanction. These tools also help produce reports and outputs that are desired and needed to understand the data.
Data Mining Tools	Software tools that find patterns in stores of data or databases. These tools are useful for predictive analytics and optimization analytics.

INFRASTRUCTURE ARCHITECTURE

The data warehouse tech stack is built on a fundamental framework of hardware and software known as the infrastructure."

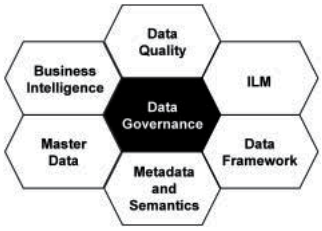


Using a data warehouse appliance or a dedicated database infrastructure helps support the data warehouse. This technique tends to yield the highest performance. The data warehouse appliance is optimized to provide database services using Massively Parallel Processing (MPP) architecture. It includes multiple, tightly coupled computers with specialized functions, plus at least one array of storage devices that are accessed in parallel. Specialized functions include: system controller, database access, data load, and data backup.

Data Warehouse Appliances provide high performance. They can be up to 100 times faster than the typical Database Server. Consider the Data Warehouse Appliance when more than 2TB of data must be stored.

DATA ARCHITECTURE

Data architecture is a blueprint for the management of data in an enterprise. The data architect builds a picture of how multiple sub-domains work. Some of these subdomains are data governance, data quality, ILM (Information Lifecycle Management), data framework, metadata and semantics, master data, and, finally, business intelligence.



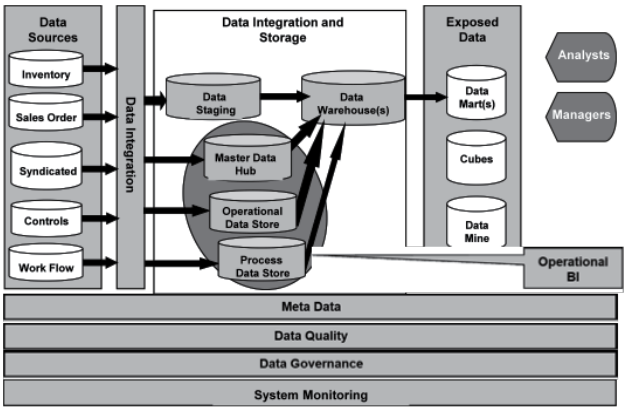
DATA ARCHITECTURE SUB-DOMAINS

SUB-DOMAIN	DESCRIPTION
Data Governance (DG)	The overall management of data and information includes people, processes, and technologies that improve the value obtained from data and information by treating data as an asset. It is the cornerstone of the data architecture.
Data Quality Management (DQM)	The discipline of ensuring that data is fit for use by the enterprise. It includes obtaining requirements and rules that specify the dimensions of quality required, such as accuracy, completeness, timeliness, and allowed values.
Information Lifecycle Management (ILM)	The discipline of specifying and managing information through its life from its conception to disposal. Information activities that make up ILM include classification, creation, distribution, use, maintenance, and disposal.
Data Framework	A description of data-related systems that is in terms of a set of fundamental parts and the recommended methods for assembling those parts using patterns. The data framework can include: database management, data storage, and data integration.

SUB-DOMAIN	DESCRIPTION
Metadata and Semantics	Information that describes and specifies data related objects. This description can include: structure and storage of data, business use of data, and processes that act on the data. “Semantics” refers to the meaning of the data.
Master Data Management (MDM)	An activity focused on producing and making available a “golden record” of master data and essential business entities, such as customers, products, and financial accounts. Master data is data describing major subjects of interest that is shared by multiple applications.
Business Intelligence	The people, tools, and processes that support planning and decision making, both strategic and operational, for an organization.

DATA FLOW

The diagram displays how data flows through the data warehouse system. Data first originates from the data sources, such as inventory systems (systems stored in data warehouses and operational data stores). The data stores are formatted to expose data in the data marts that are then accessed using BI and analytics tools.



DATA

Data is the raw material through which we can gain understanding. It is a critical element in data modeling, statistics, and data mining. It is the foundation of the pyramid that leads to wisdom and to informed action.

DATA ATTRIBUTE CHARACTERIST

CHARACTERISTIC	DESCRIPTION
Name	Each attribute has a name, such as “Account Balance Amount”. An attribute name is a string that identifies and describes an attribute. In the early stages of data design, you may just list names without adding clarifying information, called metadata.
Datatype	The datatype, also known as the "data format," could have a value such as a decimal (12.4). This is the format used to store the attribute. This specifies whether the information is a string, a number, or a date. In addition, it specifies the size of the attribute.
Domain	A domain, such as Currency Amounts, is a categorization of attributes by function.
Initial Value	An initial value such as 0.0000 is the default value that an attribute is assigned when it is first created.
Rules	Rules are constraints that limit the values that an attribute can contain. An example rule is, "the attribute must be greater than or equal to 0.0000." Use of rules helps to improve data quality.
Definition	A narrative that conveys or describes the meaning of an attribute. For example, Account Balance Amount is a measure of the monetary value of a financial account, such as a bank account or an investment account.

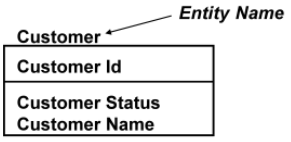
DATA MONITORING

Three levels of data modeling are developed in sequence:

1. **Conceptual Data Model** - A high level model that describes a problem using entities, attributes, and relationships.
2. **Logical Data Model** - A detailed data model that describes a solution in business terms, and that also uses entities, attributes, and relationships.
3. **Physical Data Model** - A detailed data model that defines database objects, such as tables and columns. This model is needed to implement the models in a database and produce a working solution.

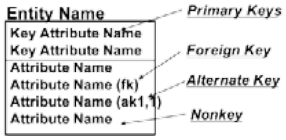
ENTITIES

An entity is a core part of any conceptual and logical data model. An entity is an object of interest to an enterprise—it can be a person, organization, place, thing, activity, event, abstraction, or idea. Entities are represented as rectangles in the data model. Think of entities as singular nouns.



ATTRIBUTES

An attribute is a characteristic of an entity. Attributes are categorized as primary keys, foreign keys, alternate keys, and non-keys, as depicted in the diagram to the right.

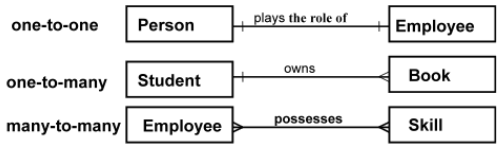


RELATIONSHIPS

A relationship is an association between entities. Such a relationship is diagrammed by drawing a line between the related entities. The following diagram depicts two entities — Customer and Order — that have a relationship specified by the verb phrase "places" in this way: Customer Places Order.

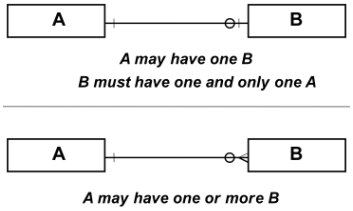
CARDINALITY

Cardinality specifies the number of entities that may participate in a given relationship, expressed as one-to-one, one-to-many, or many-to-many, as depicted in the following example:

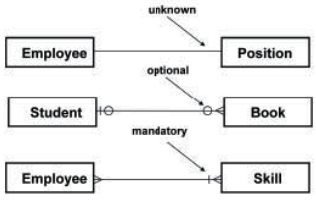


Cardinality is expressed as minimum and maximum numbers. In the first example below, an instance of entity A may have one instance of entity B, and entity B must have one and only one instance of entity A. Cardinality is specified by putting symbols on the relationship line near each of the two entities that are part of the relationship.

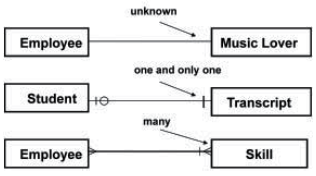
In the second case, entity A may have one or more instances of entity B, and entity B must have one, and only one, instance of entity A.



Minimum cardinality is expressed by the symbol farther away from the entity. A circle indicates that an entity is optional, while a bar indicates an entity is mandatory. At least one is required.



Maximum cardinality is expressed by the symbol closest to the entity. A bar means that a maximum of one entity can participate, while a crow's foot (a three-prong connector) means that many entities may participate, and thus denotes a large, unspecified number.



NORMALIZED DATA

Normalization is a data modeling technique that organizes data by breaking it down to its lowest level, i.e., its "atomic" components, to avoid duplication. This method is used to design the Atomic Data Warehouse part of the data warehousing system. Here are the first three normalized levels by Edgar F. Codd. There are other normalized levels that you can read more about here. A relational database is considered normalized when it reaches the third normal form.

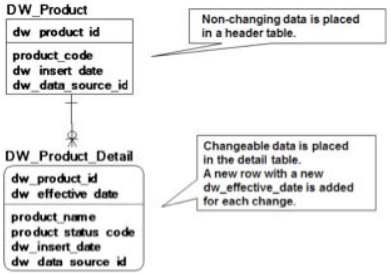
SUB-DOMAIN	DESCRIPTION
First Normal Form (1NF)	Entities contain no repeating groups of attributes.
Second Normal Form (2NF)	Entity is in the first normal form and attributes that depend on only part of a composite key are separated into new entities.
Third Normal Form (3NF)	The entity is in the second normal form and non-key attributes representative of an entity's facts are separated to more independent, multi-valued entities.

ATOMIC DATA WAREHOUSE

The atomic data warehouse (ADW) is an area where data is broken down into low-level components in preparation for export to data marts. The ADW is designed using normalization and methods that make for speedy history loading and recording.

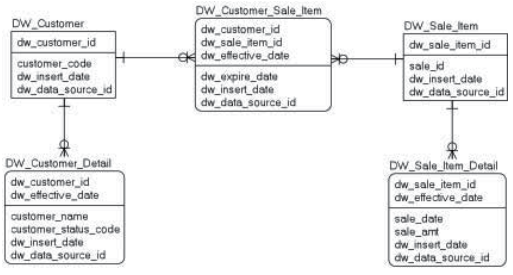
HEADER AND DETAIL ENTITIES

The ADW is organized into non-changing data with logical keys and changeable data that supports tracking of changes and rapid load/insert. Use an integer as the primary surrogate key. Then add the effective date to track changes.



ASSOCIATIVE ENTITIES

Track the history of relationships between entities using an associative entity with effective dates and expiration dates.



ATOMIC DW SPECIALIZED ATTRIBUTES

Use specialized attributes to improve ADW efficiency and effectiveness. Identify these attributes using a prefix of ADW_.

ATTRIBUTE NAME	DESCRIPTION
dw_xxx_id	Data Warehouse assigned surrogate key. Replace 'xxx' with a reference to the table name, such as 'dw_ customer_ dim_id'.
dw_insert_date	The date and time when a row was inserted into the data warehouse.
dw_effective_date	The date and time when a row in the data warehouse began to be active.
dw_expire_date	The date and time when a row in the data warehouse stopped being active.
dw_data_process_ log_id	A reference to the data process log. The log is a record of the process of how data was loaded or modified in the data warehouse.

SUPPORTING TABLES

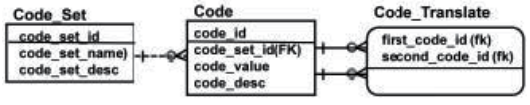
Supporting data is required to enable the data warehouse to operate smoothly. Some examples of supporting data are:

- Code Management and Translation
- Data Source Tracking and Logging
- Message Logging

CODE MANAGEMENT AND TRANSLATION

Data warehousing requires that codes, such as gender code and unit of measure, be translated to standard values aided by code-translation tables, like these:

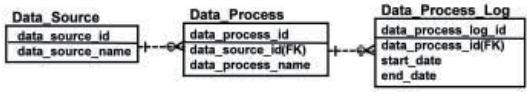
- **Code Set** – A group of codes, such as "Gender Code."
- **Code** – An individual code value.
- **Code Translation** – Mapping between code values.



DATA-SOURCE TRACKING AND LOGGING

Data-source tracking provides a means of tracing where data originated within a data warehouse:

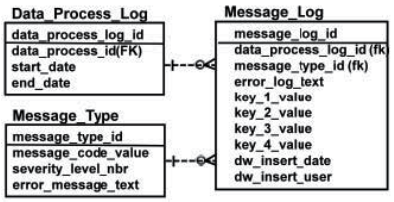
- **Data Source** – Identifies the system or database.
- **Data Process** –Traces the data-integration procedure.
- **Data Process Log** – Traces each data warehouse load.



MESSAGE LOGGING

Message logging provides a record of events that occur while loading the data warehouse:

- **Data Process Log** – Traces each data warehouse load.
- **Message Type** – Specifies the kind of message.
- **Message Log** – Contains an individual message.

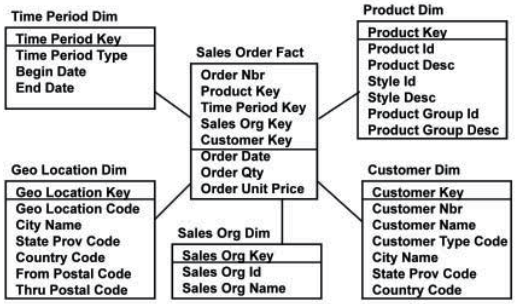


DIMENSIONAL DATABASE

A Dimensional Database is a database that is optimized for query and analysis and is not normalized like the Atomic Data Warehouse. It consists of fact and dimension tables, where each fact is connected to one or more dimensions.

SALES ORDER FACT

The Sales Order Fact includes the measurer's order quantity and currency amount. Dimensions of Calendar Date, Product, Customer, Geo Location, and Sales Organization put the Sales Order Fact into context. This star schema supports looking at orders in a cubical way, enabling slicing and dicing by customer, time, and product.



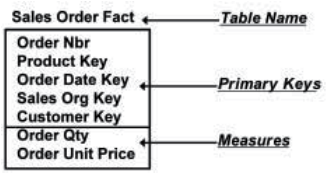
FACTS

A fact is a set of measurements. It tends to contain quantitative data that gets presented to users. It often contains amounts of money and quantities of things. Facts are surrounded by dimensions that categorize the fact.

ANATOMY OF A FACT

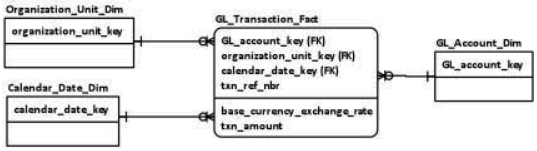
Facts are SQL tables that include:

- **Table Name** – A descriptive name usually containing the word 'Fact.'
- **Primary Keys** – Attributes that uniquely identify each fact occurrence and relate it to dimensions.
- **Measures** – Quantitative metrics.



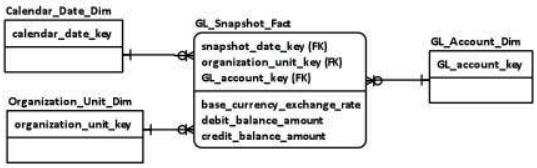
EVENT FACT EXAMPLE

Event facts record single occurrences, such as financial transactions, sales, complaints, or shipments.



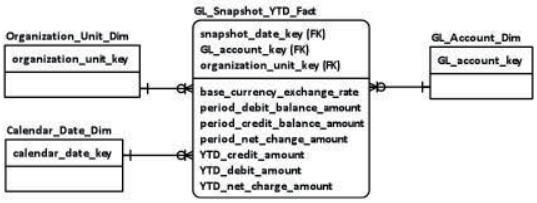
SNAPSHOT FACT

The snapshot fact captures the status of an item at a point in time, such as a general ledger balance or inventory level.



CUMULATIVE SNAPSHOT FACT

The cumulative snapshot fact adds accumulated data, such as year-to-date amounts, to the snapshot fact.



AGGREGATED FACT

Aggregated facts provide summary information, such as general ledger totals during a period of time, or complaints per product per store per month.

FACT-LESS FACT

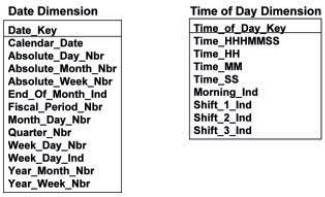
The fact-less fact tracks an association between dimensions rather than quantitative metrics. Examples include miles, event attendance, and sales promotions.

DIMENSIONS

A dimension is a database table that contains properties that identify and categorize. The attributes serve as labels for reports and as data points for summarization. In the dimensional model, dimensions surround and qualify facts.

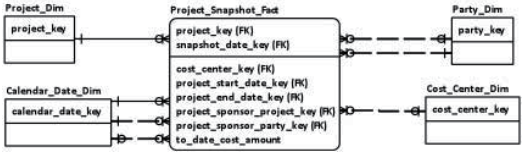
DATE AND TIME DIMENSIONS

Date dimensions support trend analysis, and include the date and its associated week, month, quarter, and year. Time-of-day dimensions are used to analyze daily business volume.



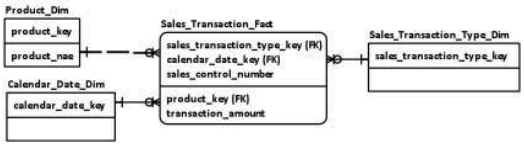
MULTIPLE-DIMENSION ROLES

One dimension can play multiple roles. The date dimension could play the roles of a snapshot date, a project start date, and a project end date.



DEGENERATE DIMENSION

A degenerate dimension has a dimension key without a dimension table. Examples include transaction numbers, shipment numbers, and order numbers.



SLOWLY-CHANGING DIMENSIONS

Changes to dimensional data can be categorized into levels:

SCD TYPE	DESCRIPTION
SCD Type 0	Data is non-changing. It is inserted once and never changed.
SCD Type 1	Data is overwritten. New data will overwrite and update existing values.
SCD Type 2	Adds a new row. Each change will add a new row where all the values will be the same except for the changed fields. This will mean that a new field(s) will be added to mark the rows and state which one is effective.
SCD Type 3	Adds a new attribute. Changes will be tracked in another field(s). In addition to the effective field, we will have the last X values stored in other fields (each one in a separate field) where usually only one history record will be saved for each field.

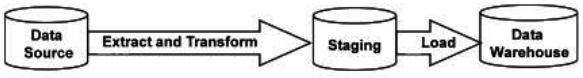
SCD TYPE	DESCRIPTION
SCD Type 4	History table. In addition to the effective table we keep a history table that will hold all the changes that occurred in the table. It will create a snapshot of every row that was changed and save it with the relevant timestamp.
SCD Type 6	Hybrid method of types 1, 2, and 3.

DATA INTEGRATION

Data integration is a technique for moving data or otherwise making data available across data stores. The data integration process can include extraction, movement, validation, cleansing, transformation, standardization, and loading.

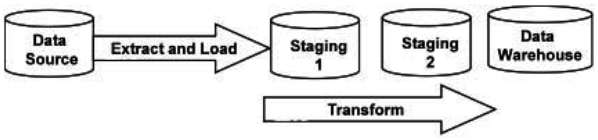
EXTRACT TRANSFORM LOAD (ETL)

In the ETL pattern of data integration, data is extracted from the data source and then transformed while in flight to a staging database. Data is then loaded into the data warehouse. ETL is strong for batch processing of bulk data.



EXTRACT LOAD TRANSFORM (ELT)

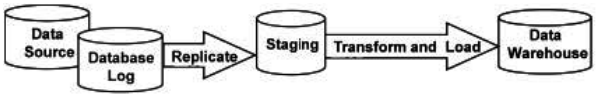
In the ELT pattern of data integration, data is extracted from the data source and loaded to staging without transformation. After that, data is transformed within staging and then loaded to the data warehouse.



This type of integration together with the use of views/queries as the transform part will help the data to be available faster for use by the end user.

CHANGE DATA CAPTURE (CDC)

The CDC pattern of data integration is strong in event processing. Database logs that contain a record of database changes are replicated near real time at staging. This information is then transformed and loaded to the data warehouse.



CDC is a great technique for supporting real-time data warehouses.

ABOUT THE AUTHOR

As a kid, **ALON BRODY** thought he'd be a police detective when he grew up. But Panoply.io's Tel Aviv-born and bred Lead Data Architect was always crazy about computers, and after serving as a combat platoon leader in the Israel Defense Forces and then completing a degree in industrial engineering and management, he moved into a career in tech, starting out as a BI analyst – first at Optimove and then at Win.com. A fun-loving coder, Alon spends his free time hanging out with his fabulous wife Adi – and indulging his addiction to PlayStation. He's a big joker who, by his own admission, is only really serious about three things: animal care, excellent food, and Manchester United.

DZONE, INC.
150 PRESTON EXECUTIVE DR.
CARY, NC 27513

888.678.0399
919.678.0300

REFCARDZ FEEDBACK WELCOME
refcardz@dzone.com

SPONSORSHIP OPPORTUNITIES
sales@dzone.com