Managing and Working with Data

Joe Cline

DATA MODELER & ENGINEER

@mrjoedata www.linkedin.com/in/josephcline www.datanomicon.blog

In This Module

What database administration is

The DBA and what they do

Data quality through data governance

Compliance auditing

Database development and the DAO

Data integration developers and ETL

Database Administration and the DBA

Create Table Statement

FQDN (Fully qualified domain name)

```
CREATE TABLE IF NOT EXISTS `mydb`.`dbo`.`Hotel` (  `Hotel_ID` INT NOT
NULL,  `Hotel_Name` VARCHAR(45) NULL,  `Hotel_Phone` VARCHAR(45) NULL,
`Hotel_Address` VARCHAR(45) NULL,  `Postal_Code` VARCHAR(45) NULL,
PRIMARY KEY (`Hotel_ID`),  UNIQUE INDEX `Hotel_ID_UNIQUE` (`Hotel_ID`
ASC),  INDEX `FK_Postal_Code_idx` (`Postal_Code` ASC),  CONSTRAINT
`FK_Postal_Code`  FOREIGN KEY (`Postal_Code`)  REFERENCES
`mydb`.`Postal_Code` (`Postal_Code`)  ON DELETE NO ACTION  ON
UPDATE NO ACTION)ENGINE = InnoDB;
```

Example DDL SQL Generated by a Modeling Tool

Responsibilities of a DBA

- Selecting database server(s)
- Connection to the network, storage and application
- Run and maintain backups of the database
- Run occasional database restore tests
- Automate maintenance tasks and other scheduled jobs
- Manage database user and group accounts
- Secures the data

Responsibilities of a DBA

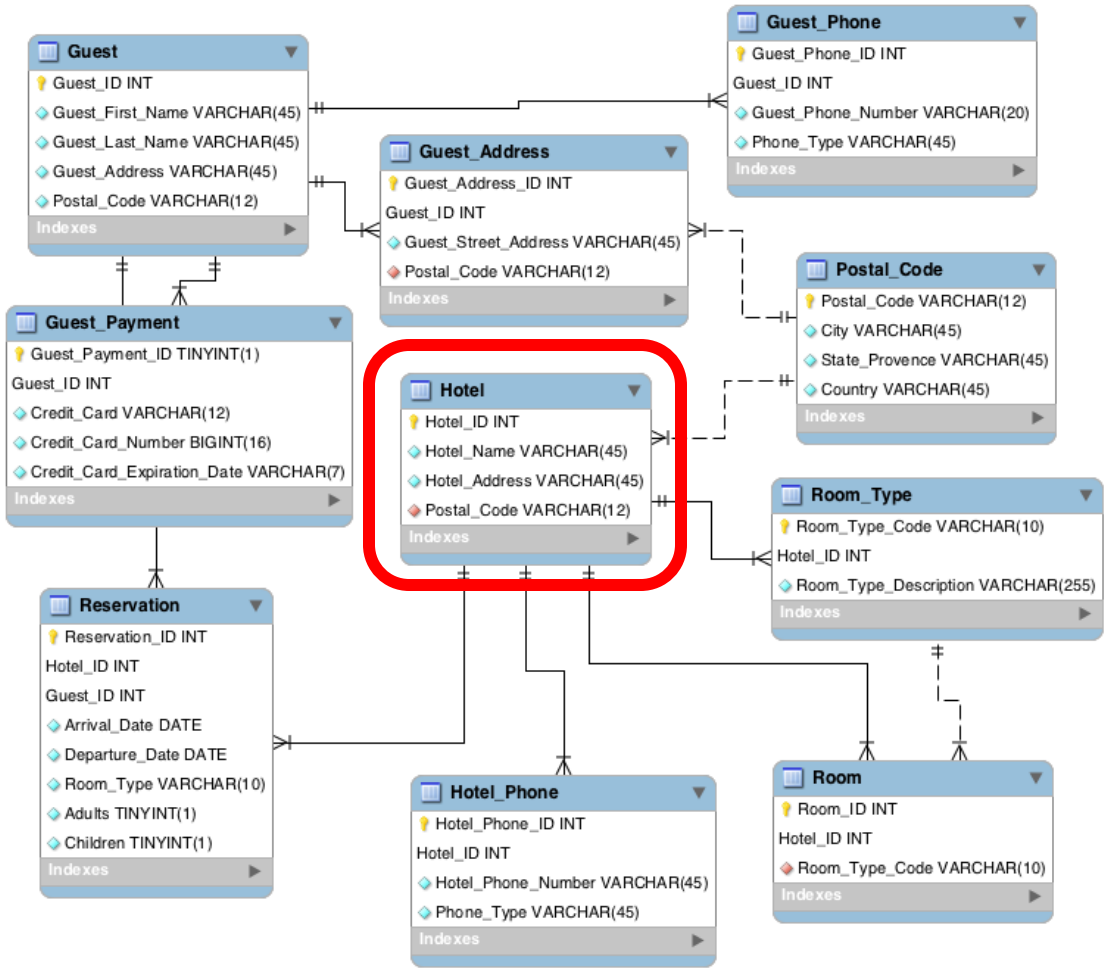
Monitors for performance and troubleshoots issues

Maintaining table indexing

Suggesting changes to the database schema

And more...

Entity Relationship (ER) Diagram Example



Database Development and the DAO

Is One DAO Better Than the Other?

Embedded SQL	Stored Procedures
Dedicated class as DAO	Dedicated by function or business logic
A function of an application	Executable object in the RDBMS
Modular	Modular
Encapsulated	Encapsulated
Compile and deploy app = downtime	No need to compile app = no downtime
SQL injection attacks	No SQL Injection
Dynamic SQL	Not dynamic – defeats purpose
App processed = network load	Processed on DB server cluster

Governance – Data Quality and Compliance

Data Quality

Governance standards and best practices

- Naming conventions
- Data types
- Lineage is documented

Data Quality

Data profiling and scoring

- More than just NULL values
- Statistics
- Know your data

Data Type Profiling

Date-time datatype

1-1-1907 ~~00:00:00.000~~

Date datatype

1-1-1907

Data Type Profiling

Date-time as a string datatype

“1-1-1907 00:00:00.000”

CAST string to date

Date datatype

1-1-1907

Data Quality

Data profiling and scoring

- More than just NULL values
- Statistics
- Know your data

Compliance Audit

HIPPA

Sarbanes-Oxley

PII - Personally identifiable information

PCI-DSS

- Payment Card Industry Data Security Standard

Jane's Story

This is Jane

This is Jane's store

Jane's store accepts credit cards

Jane's customers are happy

Jane's Story

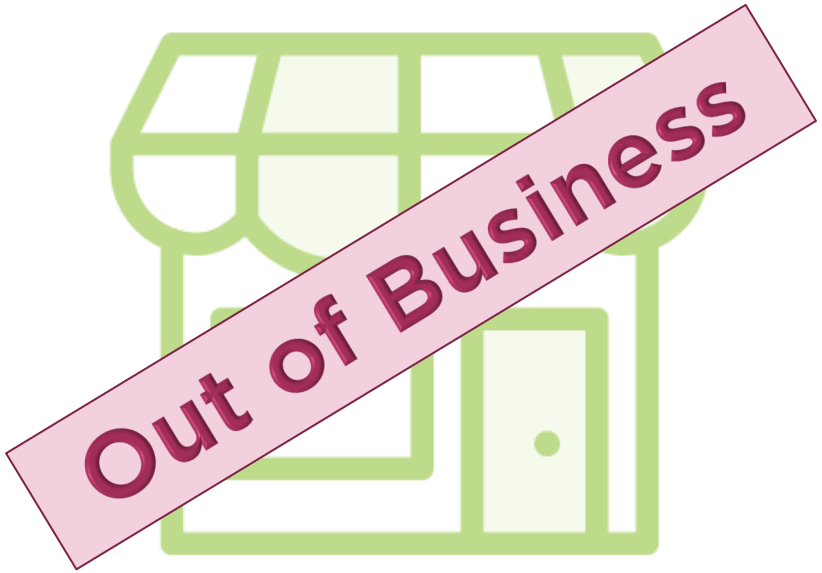
Jane is happy because her customers are happy

Jane doesn't have a policy for her customers' PII and credit card data

Jane gets hacked and now her customer's data on the dark web

Jane's customers are not happy and sue for damages

Jane's Story



Jane goes out of business and files for bankruptcy

Jane is not happy

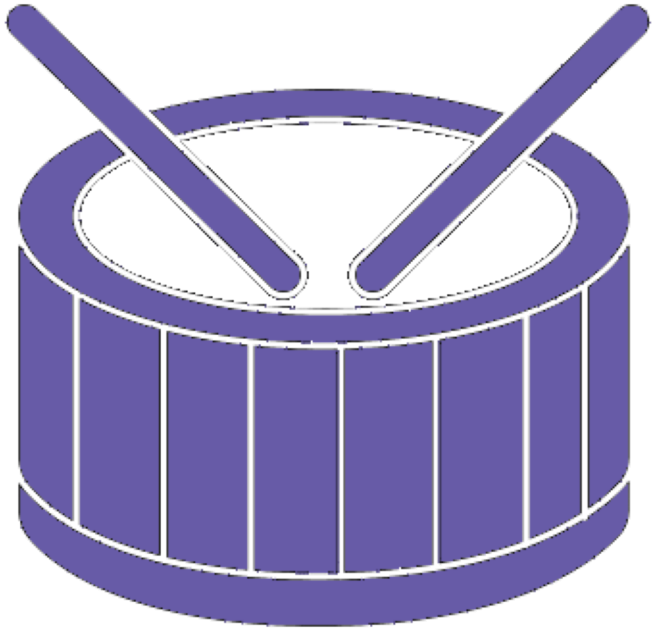
Don’t be Jane

Compliance Audit

Database access

- Failed login attempts
- Principle of least privilege

(Drum Roll Sound)



Data Governance!

Compliance Audit

Database access

- Failed login attempts
- Principle of least privilege
- Change management records

Subjected data

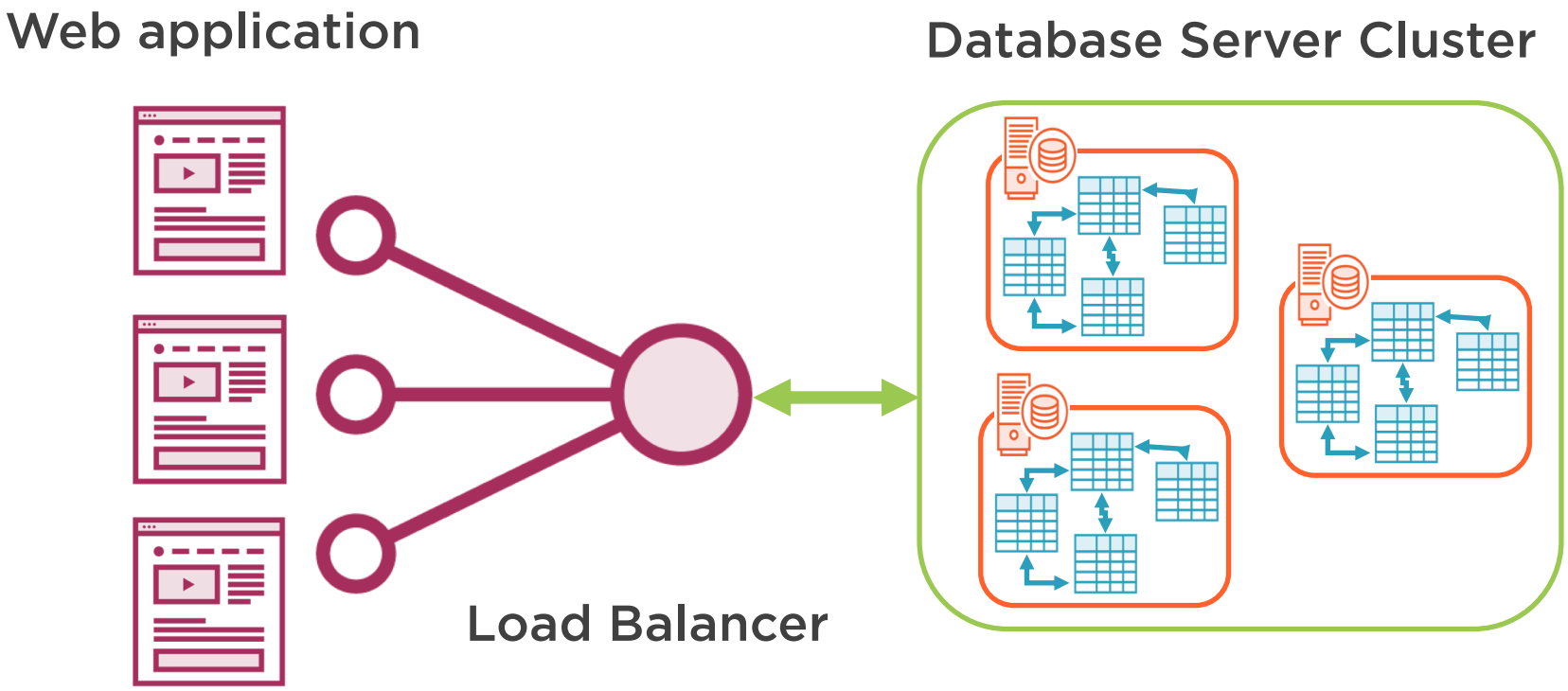
- Who has access
- How it's secured
- Where it's stored
- Retention

Where to DQ?

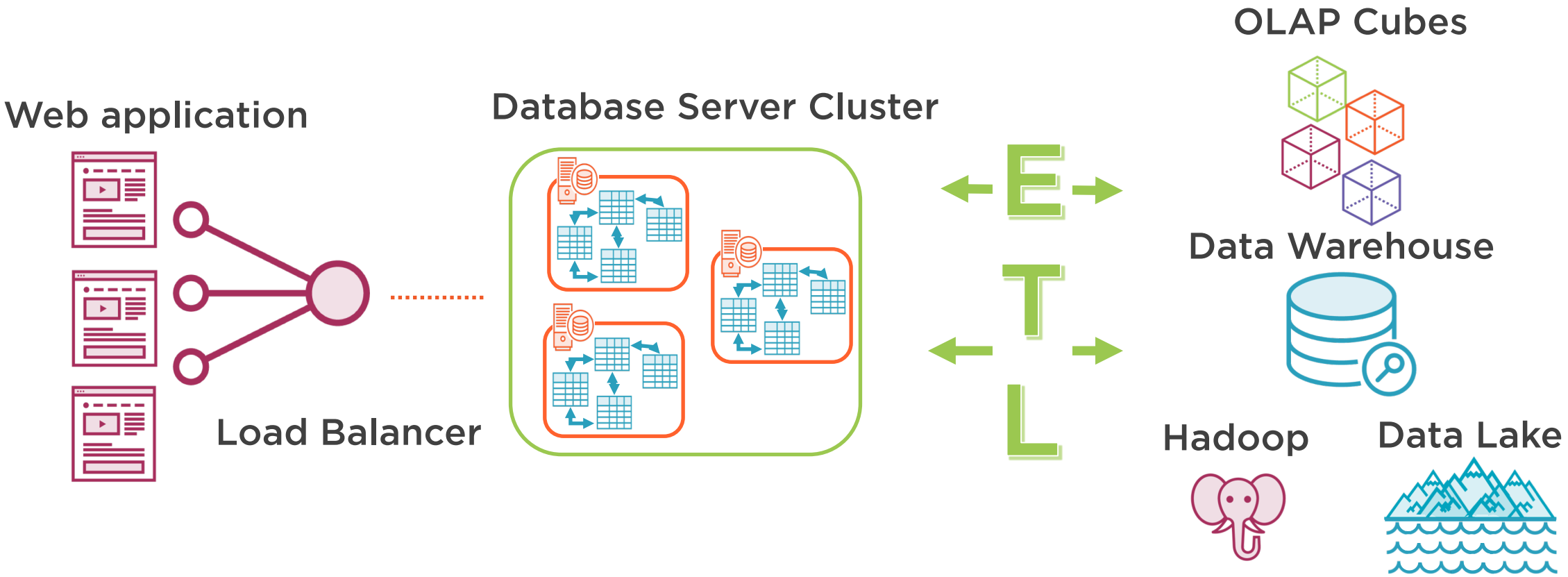
At points in the flow where data is transformed or used for reports

Data Integration Development and ETL

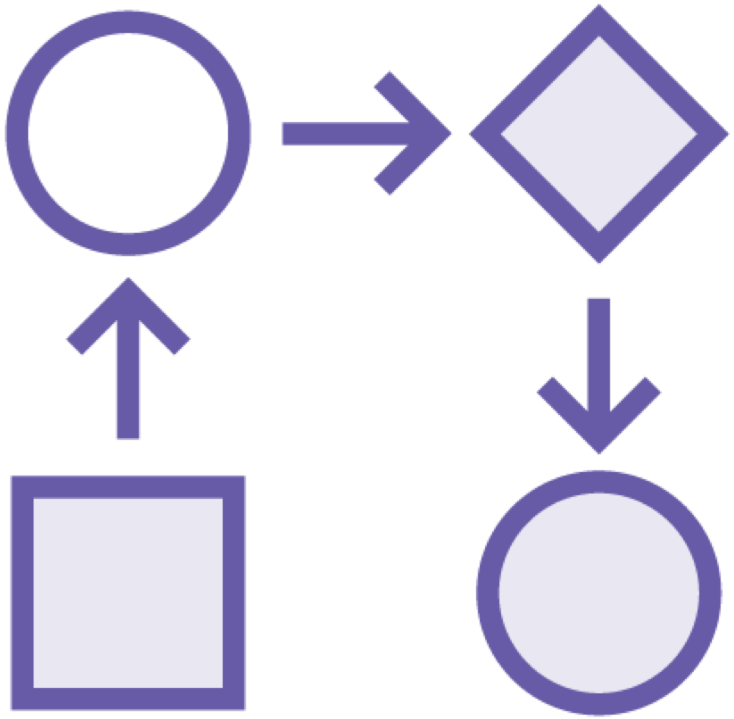
Where We Are, So Far



Generic Architecture



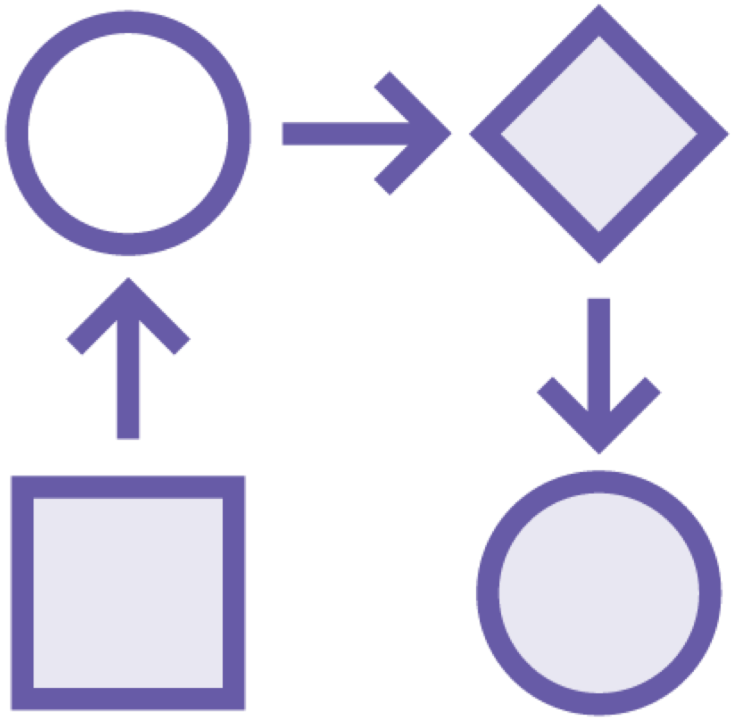
Data Integration



ETL - Extract, Transform, Load

ELT - Extract, Load, Transform

Popular Off-the-shelf ETL Tools



Informatica

SyncSort DMX and DMX-h

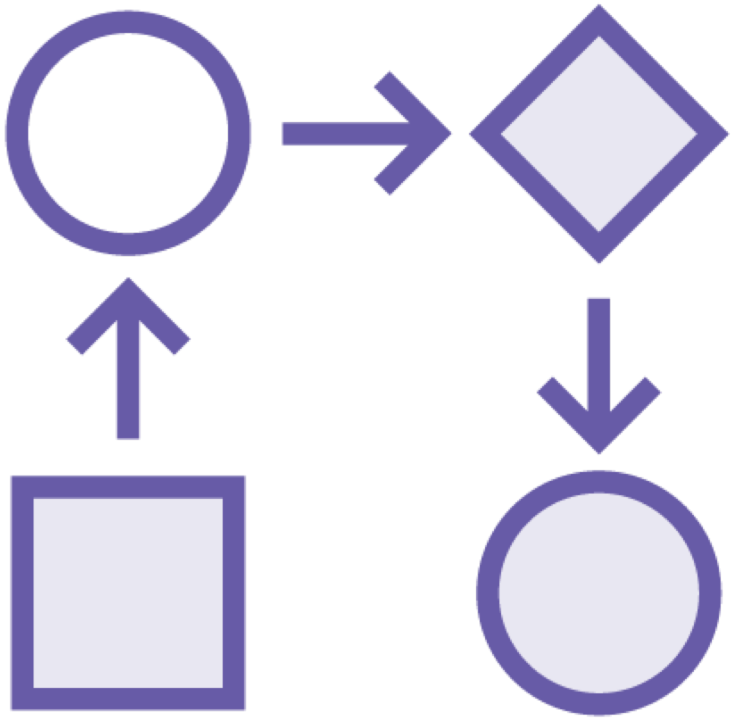
- The “h” = Hadoop

Microsoft SSIS

- SQL Server Integration Service

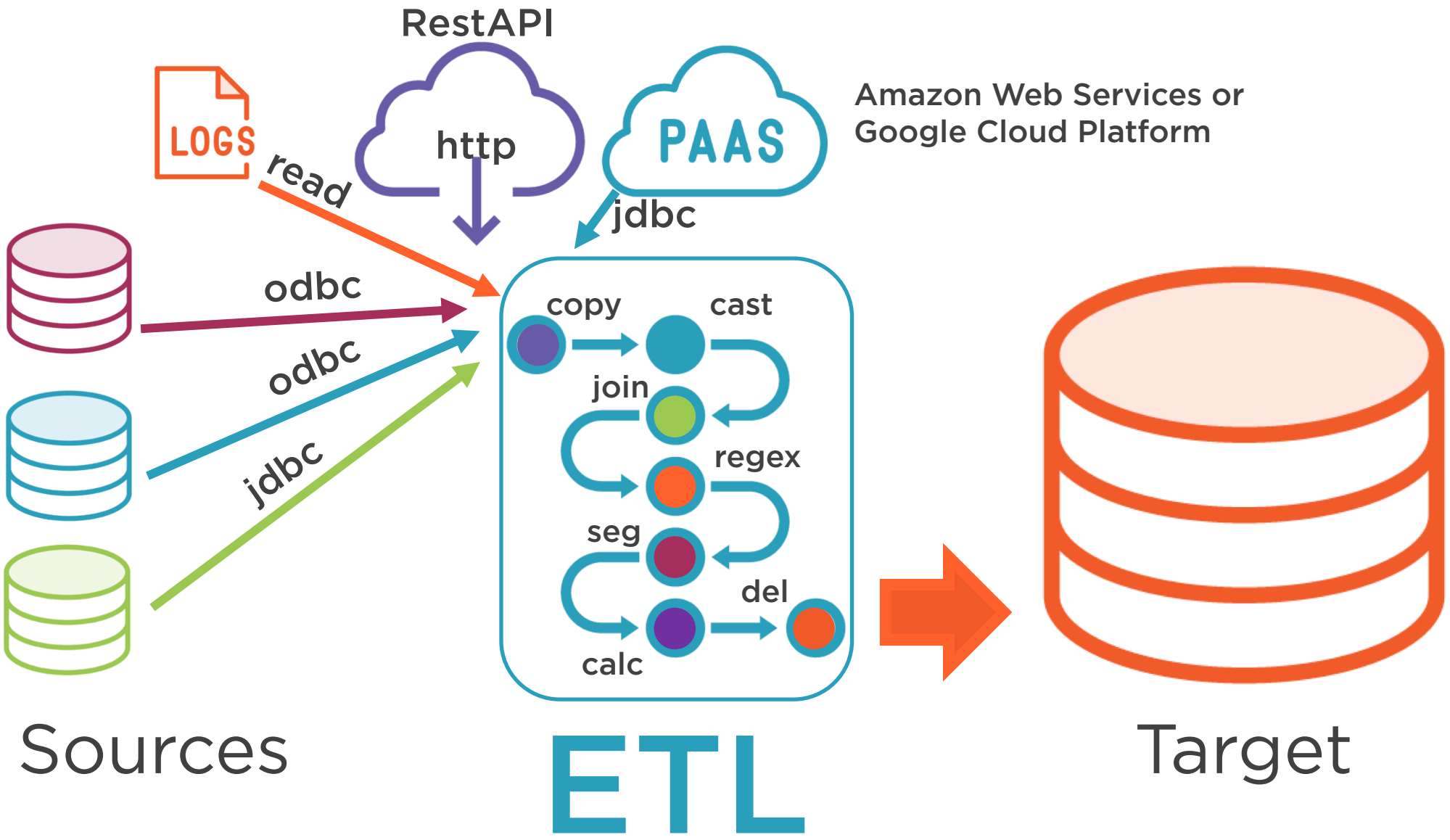
IBM Infosphere Datastage

Open Source ETL Tools



- Talend
- Pentaho Data Integration
- Jasper
- Apache Camel

An ETL Example



There's Gold in Them Thar Hills ...Of Data

Summary

What database administration is

Responsibilities of a DBA

How governance policies are enforced

- Data quality
- Government and industry compliance

Database app development (DAO)

Responsibilities of a data integration developer

Extract, Transform and Load (ETL)

Next up: Getting Value from Data