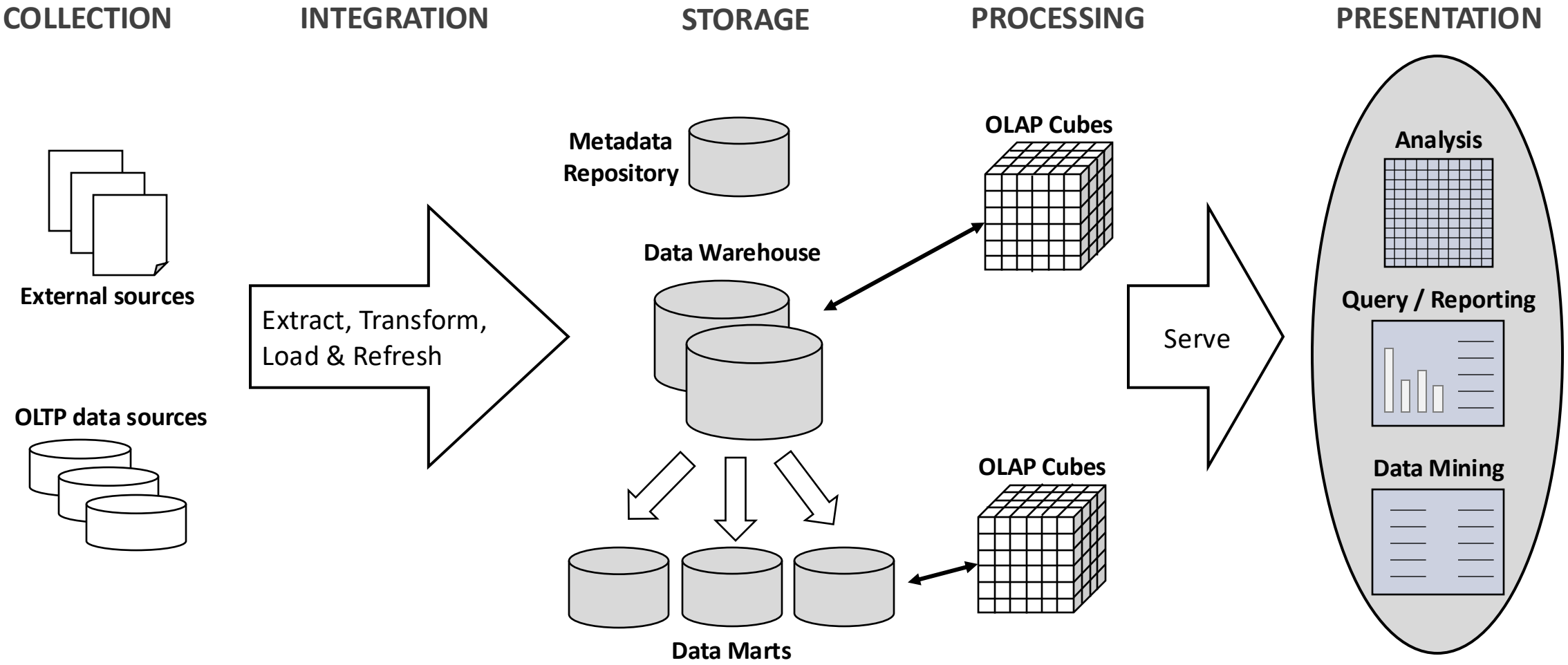
# The concepts behind “Data Integration”

## Data Ingestion, ETL, ELT, Zero-ETL & Pipelines

Abdel Dadouche
DJZ Consulting

adadouche@hotmail.com
@adadouche

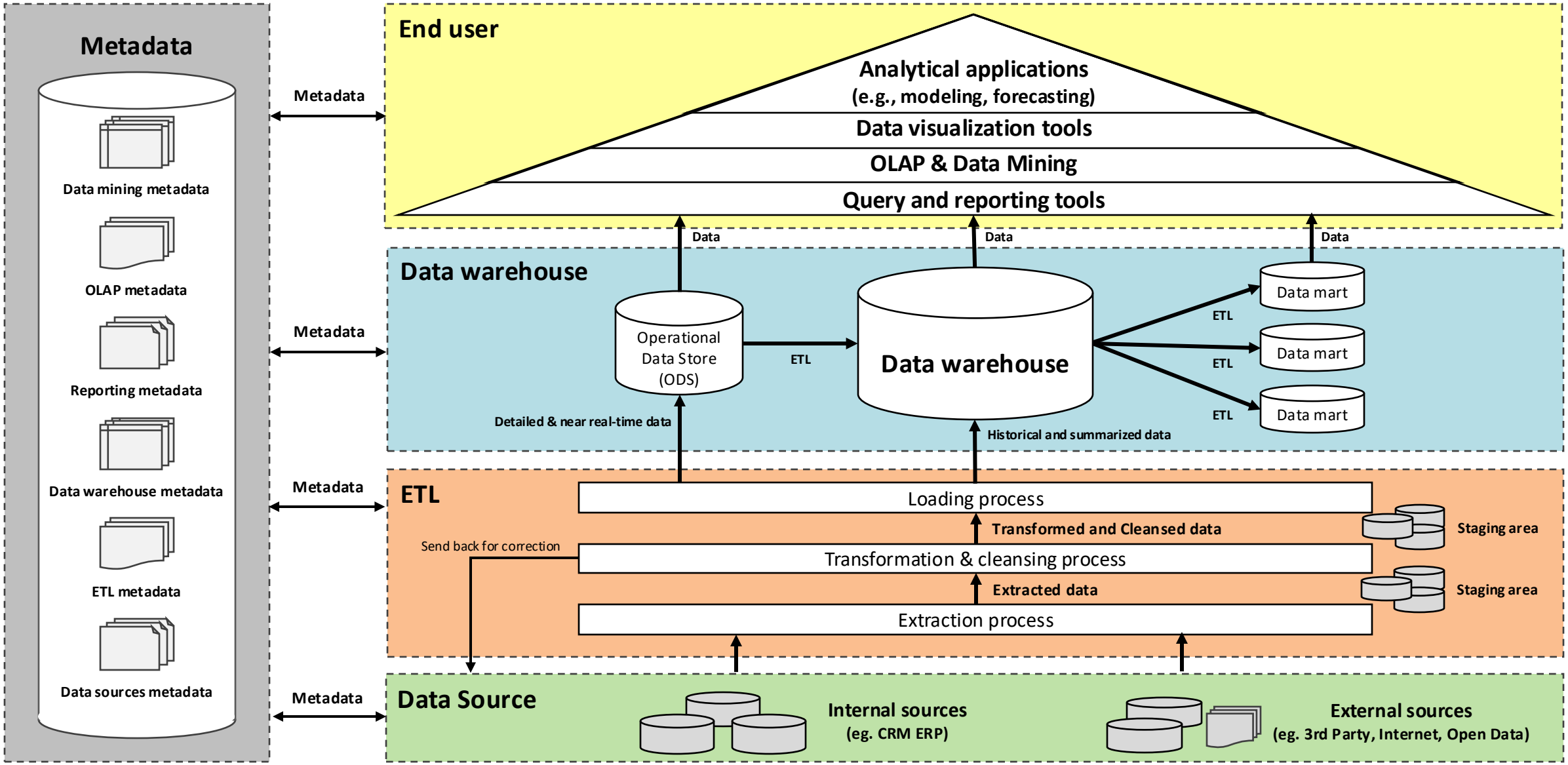
# A « Modern » Analytics architecture




Business Intelligence Process (Wan, J., Chen, T. and Chiu, S. (2005)

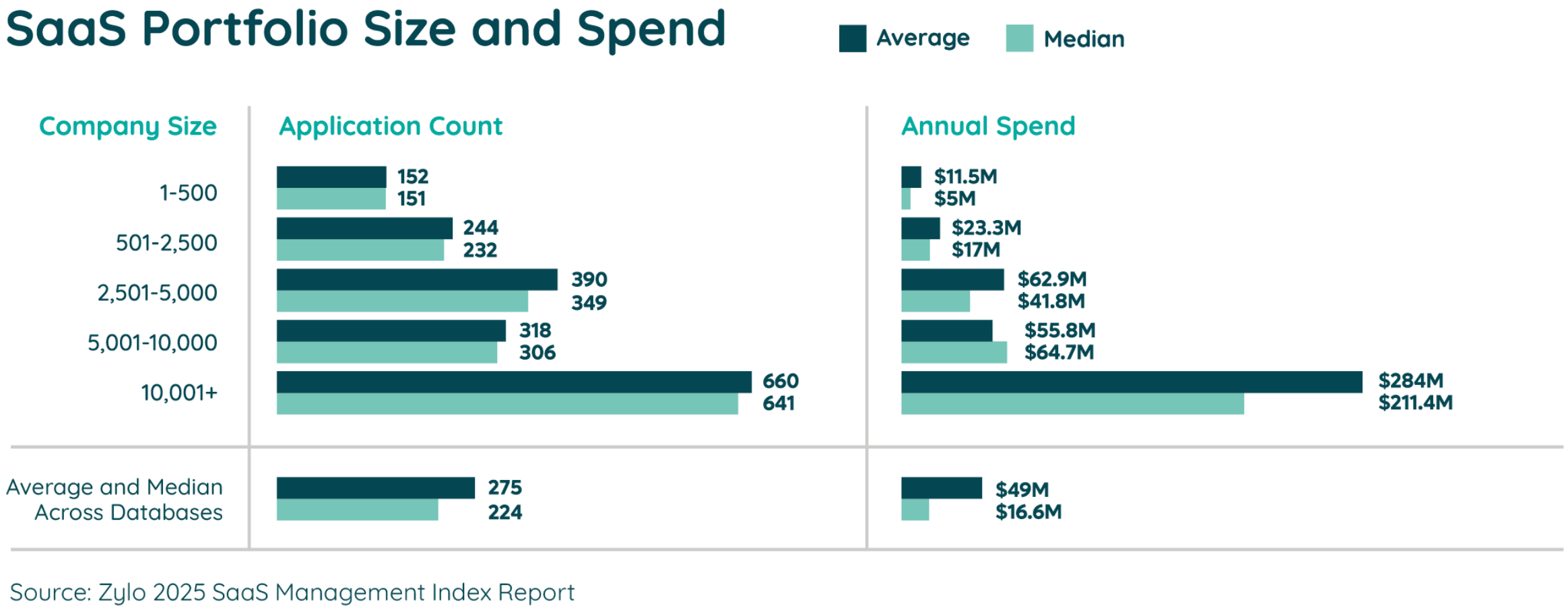
Source: A Five-Layered Business Intelligence Architecture (2011)

# The « modern » Analytics architecture by layers



# A few facts

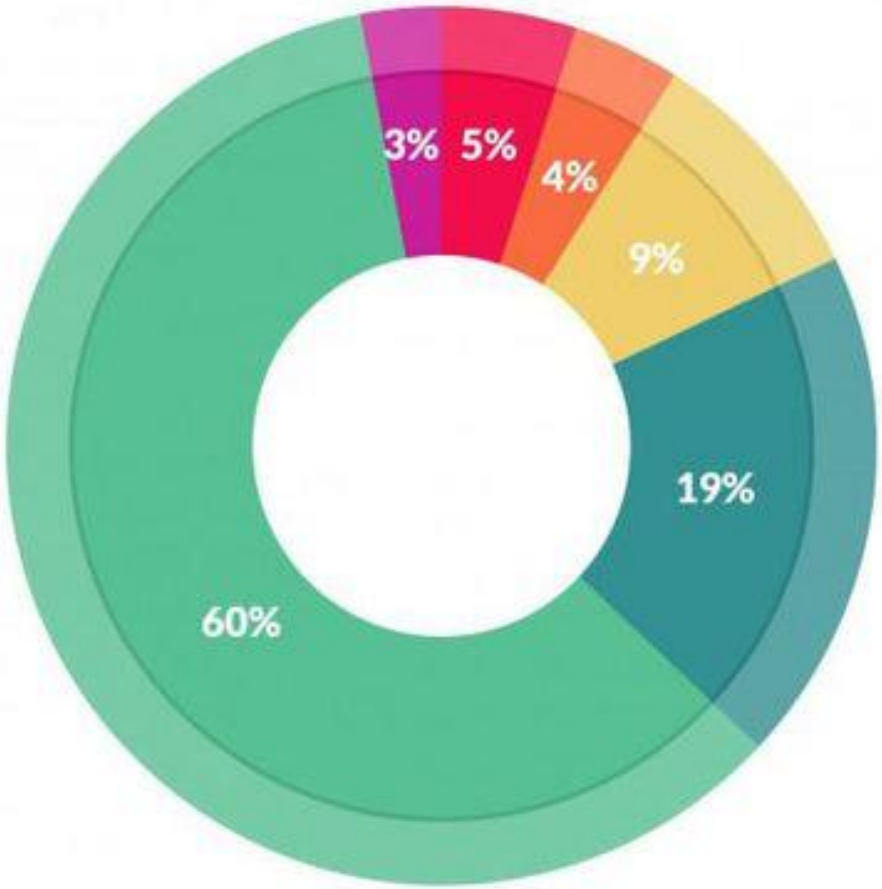
# Bigger companies means more application sources




Source: Zylo 2025 SaaS Management Index Report

Source: https://zylo.com/blog/saas-statistics/

# The reality of data scientist life

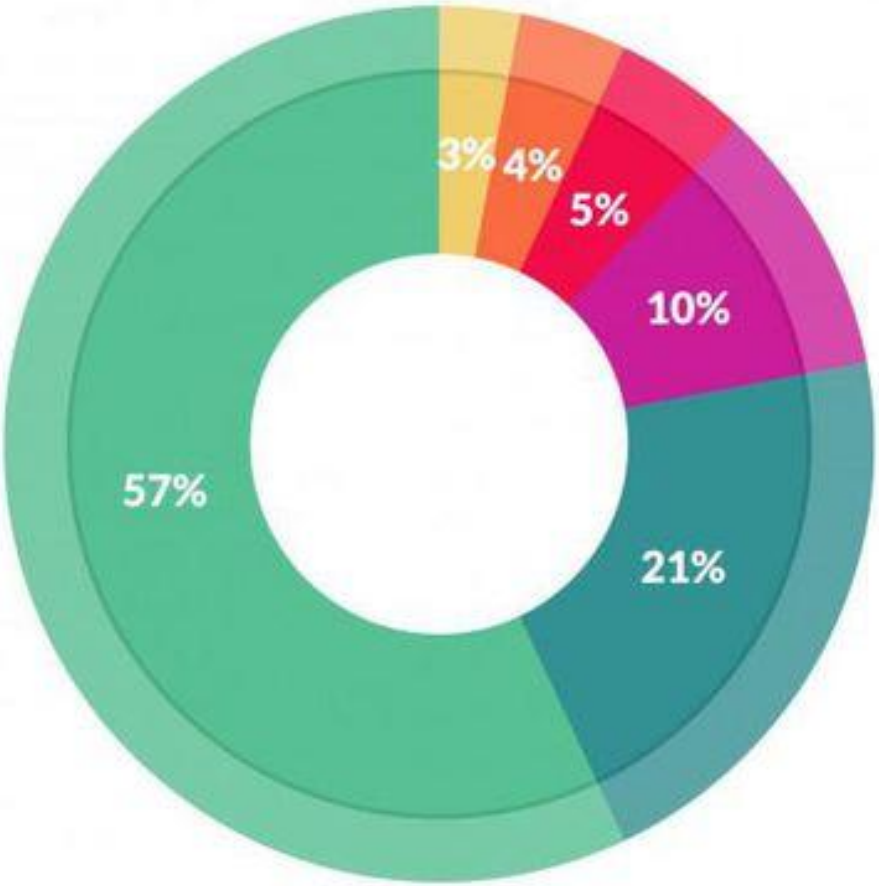



**What data scientists spend the most time doing**

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

# The reality of data scientist life




**What's the least enjoyable part of data science?**

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

# Where Does Data Come From?

- Sensor inputs (ex: scans at a checkout / cashier desk / phones)
- Manual data entry (ex: Census, forms)
- Digital content (ex: social media)
- Digital activity record (ex: clicks)

# How Does Data Are Exposed?

- **API**: Enable data ingestion and interoperability between software applications by exchanging data in formats like JSON or XML
- **Data files**: From manual data collection to ad hoc calculations or extraction in formats such as CSV, XLSX and TSV
- **Database logs and query results:** Generated by operational databases
- **Event tracking**: User-triggered code snippets embedded in web pages and applications that generates a granular record of how users interact with a website or application

# Data Integration Approaches

# Data integration

- Processes used to manage and centralize flows of data from various sources, in order to use that data in your decision-making process
- Allows your organization to maintain a view all of its data in a single environment, so you can create a comprehensive view of your business operations and customer interactions.
- Remember, that the quality of your insights are highly bound to the quality of your data and therefore the efficiency of your data integration

# The “Ad hoc” way

- Many organizations still rely on a manual & ad hoc approach to data integration
- A recent study of 1.048 executives by Deloitte finds that 62% still rely on spreadsheets for their insights
- This involves downloading files, manually altering or cleaning values, producing intermediate files, and similar actions.
- Ad hoc data integration implies:
  - Suitable only for very small volumes of data
  - Slow
  - Prone to human error
  - Insufficiently secure for sensitive information
  - Often unreproducible

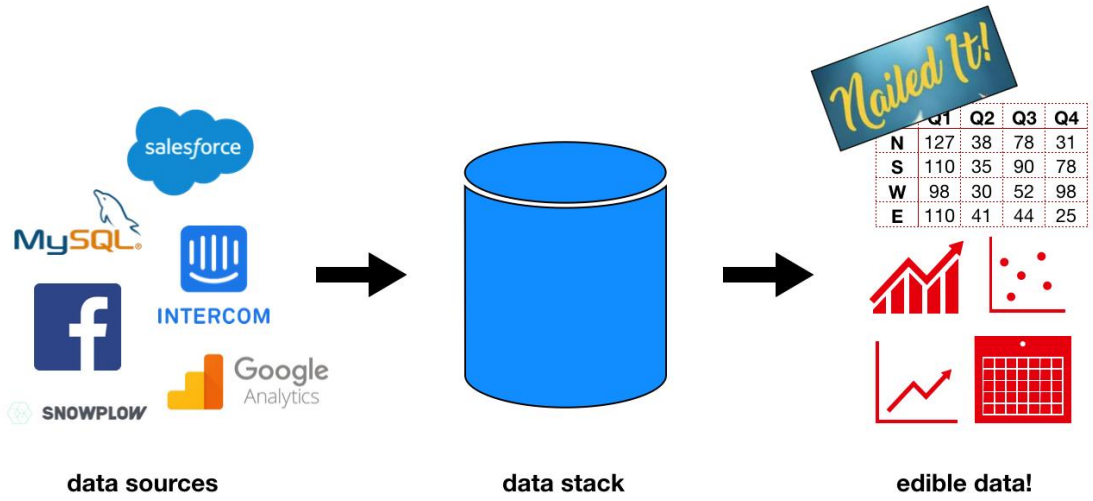
Source: https://www.zdnet.com/article/spreadsheets-still-dominate-business-analytics/

# The “federated” way

- Maintain the silos between separate data sources
- Bridge silos with “federated” queries, by querying multiple source systems directly and then merge the result data on the fly
- No need to replicate data anymore, consume them where they are
- Requires you to “trust” the data where they are
- Poor performance at large scale as it requires to move large chucks of data across location (security!) and may impact the source system performance

# The “data stack” way

- A data stack makes data edible and is like a “kitchen for data” but using a systematic & replicable approach
- Consists of tools and technologies that collectively integrate and analyze data from a variety of sources




Source: https://medium.com/@foundinblank/whats-a-data-stack-7c96f7a15fe8

# CASE STUDY: Zoopla Uses a Data Pipeline to Unify Data Integration Efforts

- Zoopla is an online property market enabling users to buy, sell or rent residential or commercial property in the UK.
- Analysts and engineers used to build a variety of undocumented custom scripts written in different languages for ad hoc needs  to extract and analyze its data
- This quickly became unsustainable as the company continued to grow and add data sources, and as it strove to quantify its progress.
- After adopting a modern data stack, Zoopla was able to combine data from its ERP and CRM software to produce a dashboard that automatically refreshes weekly and features more than 40 separate KPIs across the entire company. These KPIs are continuously displayed around the office, and used by senior leadership and rank-and-file employees alike to guide decisions.

# CASE STUDY: DiscoverOrg Stopped Using OLTP for Analytics

- *DiscoverOrg is a B2B lead-generation platform that profiles individuals and companies in order to enable smarter sales and marketing efforts.*
- *DiscoverOrg use to drew its analytics data from a copy of its OLTP production database, and excluded data from any third-party applications. Queries could take up to 36 hours, or crash the system.*
- *With data integration tool, DiscoverOrg was able to combine its production data with third-party sources into a data warehouse, spare the work of two or three data engineers, generate reports in a matter of minutes instead of days, and develop a lead-routing algorithm to acquire contracts with an 80-90% higher average value.*
- *More recently, DiscoverOrg has begun embedding analytics dashboards within its platform for the benefit of its customers*

# Data Integration Challenges

# Data Fragmentation

- Two types of fragmentation:
  - Sources are not designed for analytics queries and therefore generate data that lacks important context requiring extensive data modeling to make sense of the data
  - Sources are not designed for interoperability with other source systems requiring huge effort for establishing necessary context across sources
- Can cause the creation of “dark data” (when a large volume of collected data is unused)

# Accuracy

- Data can be inaccurate in two ways (and maybe more):
  - Use of faulty measurement or recording, especially if the data was entered by hand or transcribed from non-digital media
  - From calculations or transformations performed on raw data using different calculations
- Data can be “massaged” in many ways but every calculation / transformation takes you one step further from the original values which can take you to radically different versions of the truth

# Stale Data

- As external conditions changes faster, you cannot spare spending weeks or months assembling a report
- Working with staled data will impact your ability to make better decisions as you will be taking your decision on data that can be considered as already obsolete
- Access to data in “near real time” has become the new normal

# Effort Costs

- Historically, analysts and engineers spent the vast majority of their time building and maintaining sophisticated tools & software to wrangle data

- Data science is commonly associated with cutting-edge predictive modeling and machine learning

- But about 80% of an average data scientist's time is spent finding and integrating data rather than analyzing it

Source: https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html

# Data Integration with ETL

# ETL - Extract-Transform-Load

- Predominant approach to data integration since the 1970s
- Acronym used describe data integration activities in general.
- It is the industry “standard” among established organizations
- ETL was born when computing power, storage and bandwidth were scarce and expensive (really expensive!!)

# ETL Limitations

- Orchestration and transformation before loading causes a critical process vulnerability
- Transformations is usually specifically tailored to the source and destination. Any downstream or upstream data schema changes can break the process
- Data will become unusable by analyst unless repaired

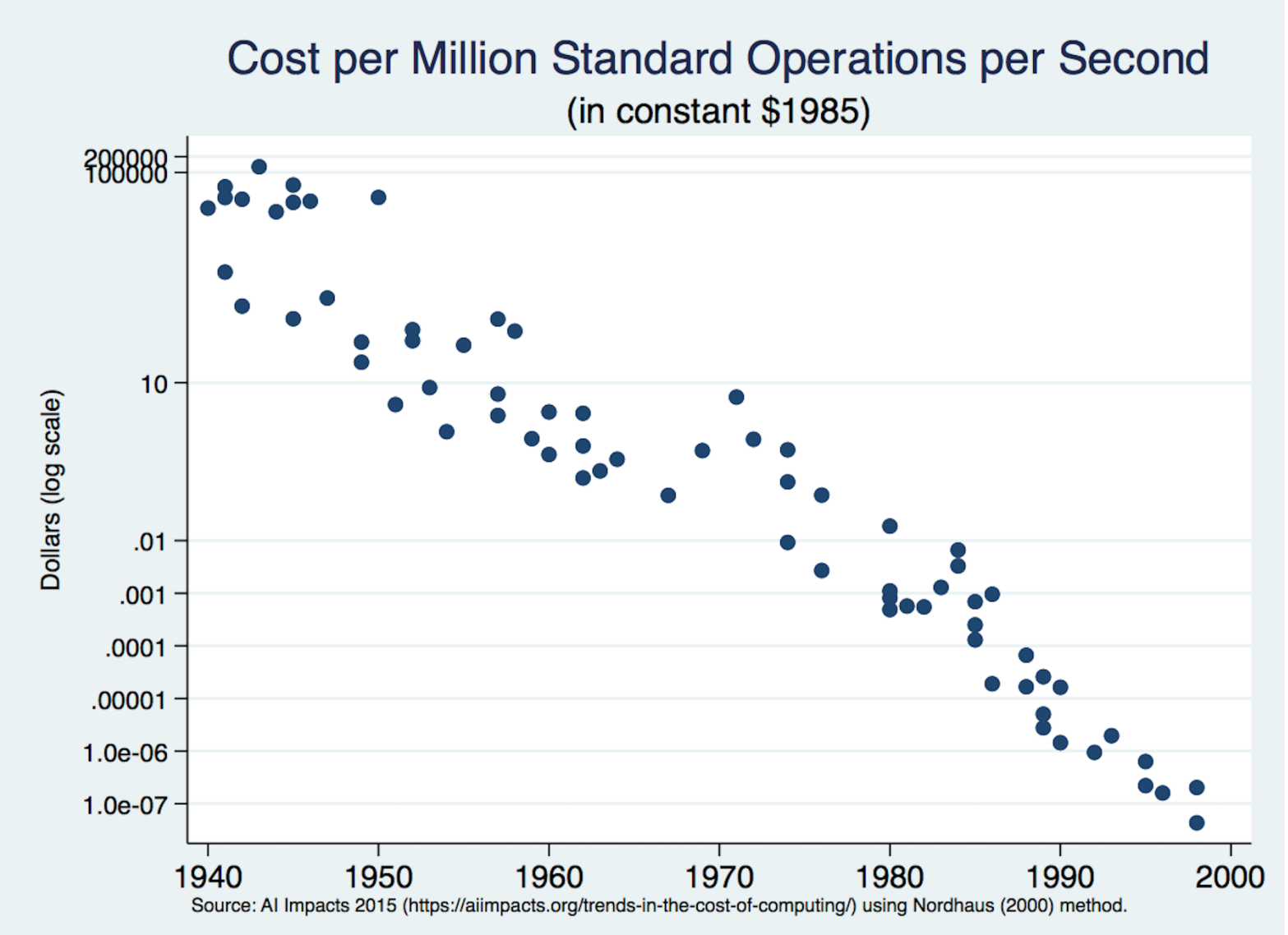
# ETL Downsides

- Complexity
  - Specific transformation needs will require custom code to be developed by highly specialized data engineers with hardly transferable skills and code base
- Easily break
  - Quick adjustments are costly and risky because of the complexity or simply impossible as extensive revisions of the code is required.
- Inaccessibility
  - Hard and expensive to implement for smaller organizations without dedicated data engineers, forcing them into using sampled data extracted manually for ad hoc reporting

# The impact of the Web and the Cloud on ETL

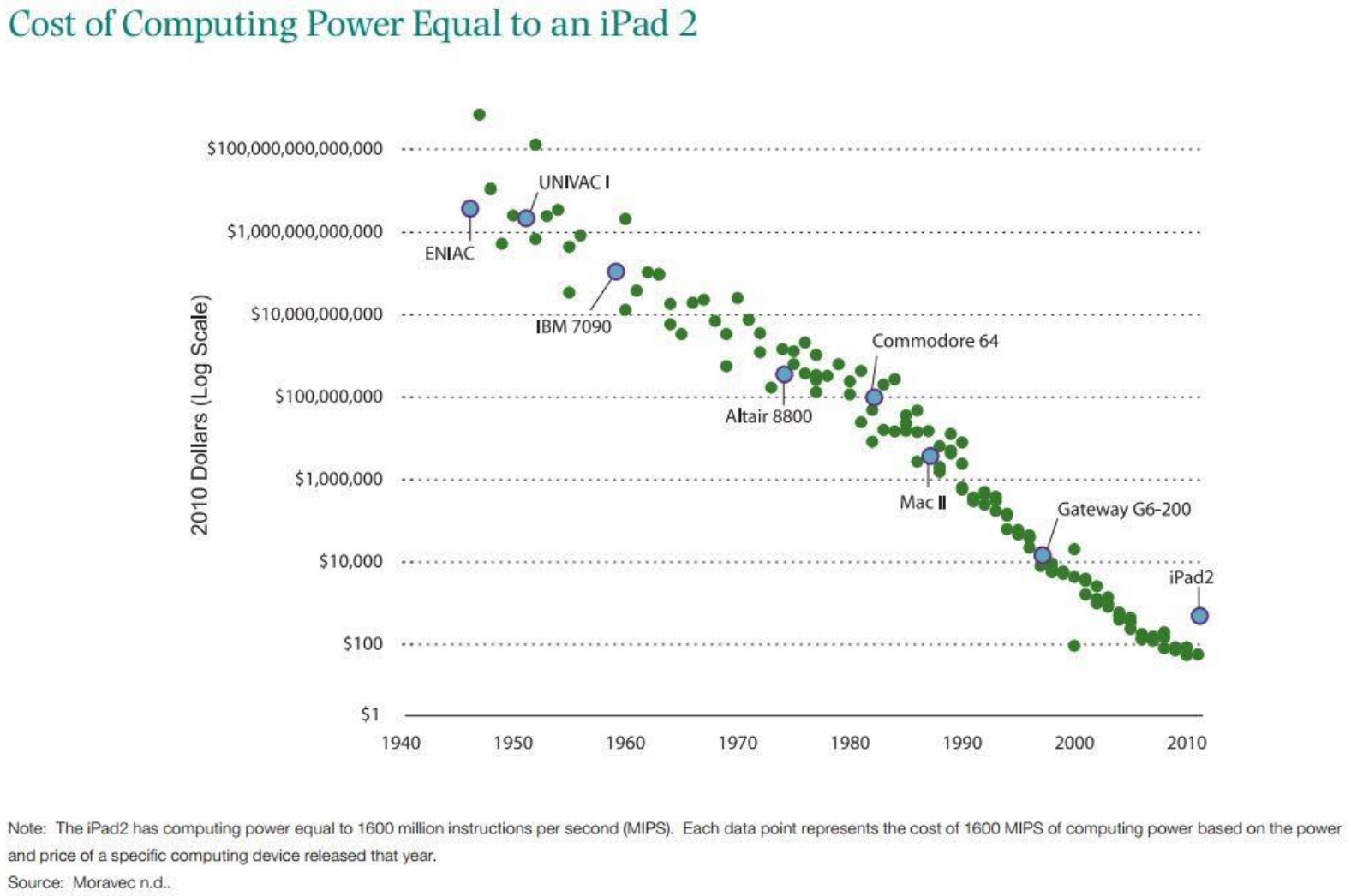
Source: https://techpolicyinstitute.org/2018/07/19/cloud-computing-co-invention-for-the-masses/

# The impact of the Web and the Cloud on ETL



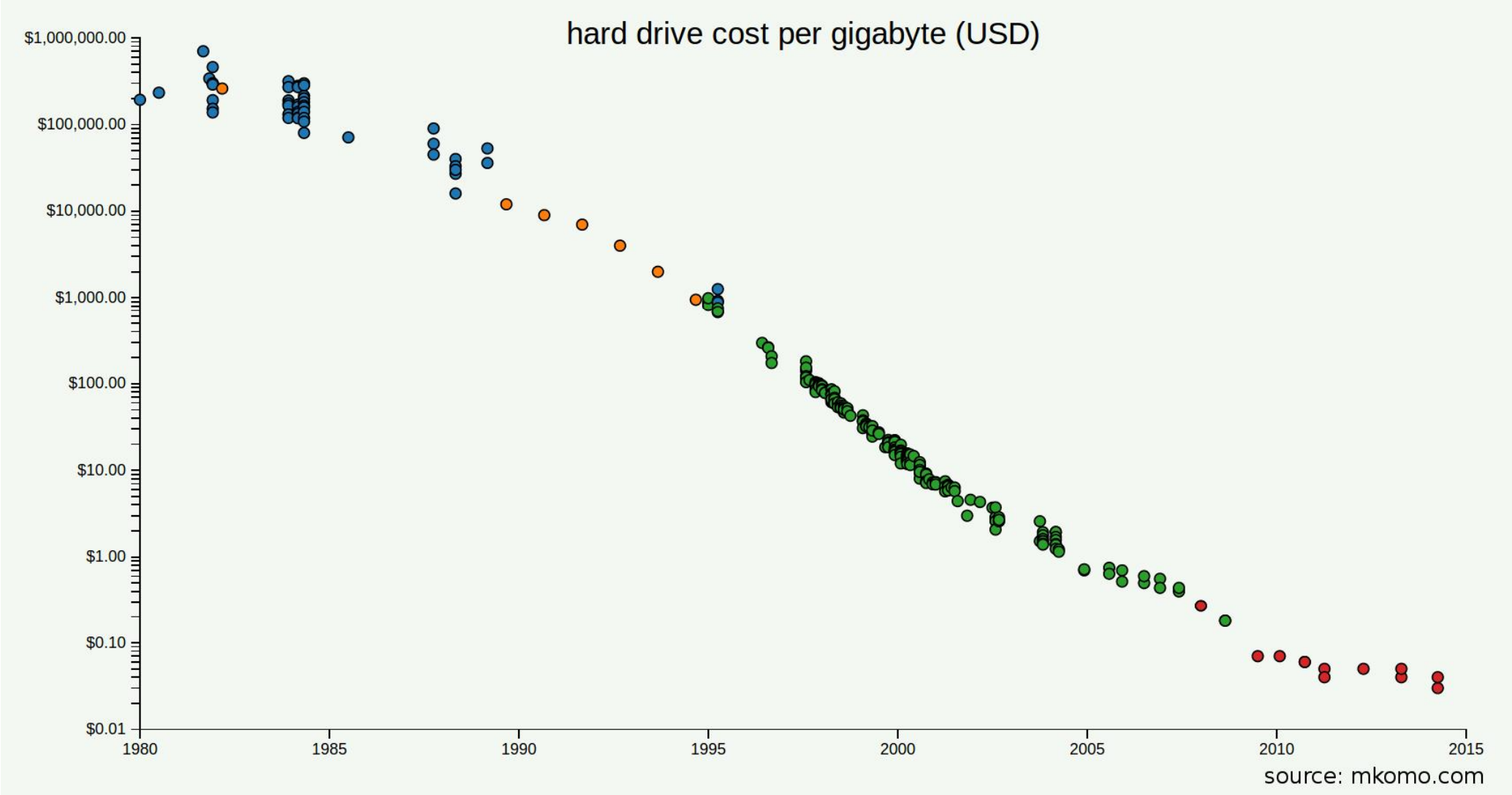
Source: https://www.hamiltonproject.org/charts/cost_of_computing_power_equal_to_an_ipad2

# The impact of the Web and the Cloud on ETL



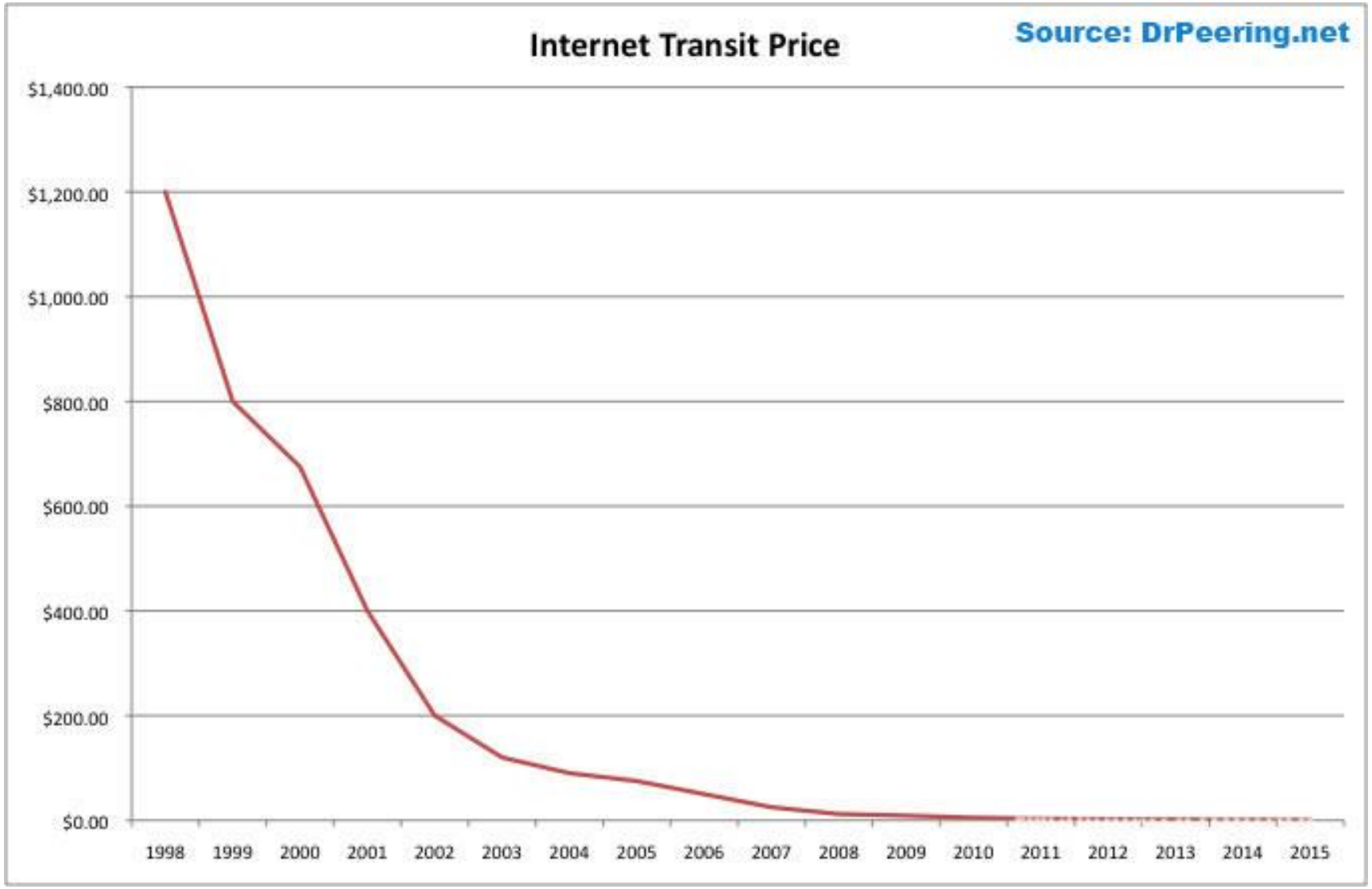
Source: https://mkomo.com/cost-per-gigabyte-update

# The impact of the Web and the Cloud on ETL



Source: https://drpeering.net/white-papers/Internet-Transit-Pricing-Historical-And-Projected.php

# The impact of the Web and the Cloud on ETL



# The impact of the Web and the Cloud on ETL

- Computation, storage and bandwidth have become cheap and ubiquitous
  - Prior 1985, compute cost halved every 17.1 months and every 10.3 months after 1985
  - The cost of a gigabyte of storage has dropped from nearly $1M to a few cents in the last 35 years
  - Over the last 20 years, bandwidth cost dropped from about $1,200/Mpbs to a few of cents
- Thanks to these cost reduction, much larger volumes of data including new data sources can be handled and pre-aggregation is no longer required
- Deeper and more comprehensive analysis can now be performed on raw data

# ELT, The Modern Approach to Data Integration

# ELT - Extract-Load-Transform

- Data can be streamed and loaded before it is transformed using data connectors
- Transformation only happens in the DWH when needed by the analysts using derivative tables, called “views,” without altering the source data
- Allows the creation of a repository of record not affected by upstream schema changes or downstream business needs
- Data can be applied to multiple use cases
- Reduces the data engineers workload. Once warehoused, analysts can use SQL to perform transformations at will.
- Complex transformations might still be required but can be orchestrated and planned so that failures no longer impact the system

# A new way forward with Automated ELT?

# Automated ELT

- Require a deeper knowledge of the data sources, extensive data modeling and analytics expertise
- Automation includes detecting and replicating data changes, light cleaning and normalization process of the data, update or creation of new tables
- Ability to leverage the expertise of outside parties who understand every aspect of the underlying data sources (ex. SalesForce, Workday, SAP) — and have stress-tested their connectors against a much wider range of corner cases than you likely ever will

# CASE STUDY: DocuSign Uses Automated Data Integration to Triple Number of Data Sources

- DocuSign is the world leader in e-signature technology, helping individuals and organizations automatically prepare, sign, act on and manage agreements.
- Formerly, DocuSign used SQL Server as a data warehouse, with a set of 6 data sources managed by 1 engineer. Each connectors took 3 to 6 months to build and up to 20 hours a week to maintain.
- This workload became untenable as the company continued to grow, especially as engineers were needed for core projects, and business teams needed to model and catalogue data from applications.
- With an automated data integration solution and a more elastic cloud data warehouse, DocuSign was able to save all 20 hours and triple the number of its data sources from 6 to 18.
- The sudden increase in scale and savings of time and labor have accompanied another highly positive development — people from all teams across the company now use over 100 active dashboards in their BI tool.

# What about SaaS Data?

# What about SaaS Data?

- One of the growing predominant source of business data
- Spans a huge range of operations and industries: marketing, ERP, CRM, eCommerce…
- SaaS applications generate a continuous, high-volume data throughput recording every user actions offers more capabilities to identify patterns and causal relationships
- At that scale, manual data integration is virtually impossible

# Choosing the right tool, How?

# Data Connector

- Basic component of every ELT data pipeline
- Ingests data from a predefined source, clean, normalize then loads the data into the DW
- Criteria to evaluate before choosing:
  - Open-source vs. proprietary
  - Standardized and normalized schemas
  - Incremental vs. full updates
  - Sources and Destinations
  - Custom data integration

# Configuration vs. Zero-Touch (or Zero-ETL)

- Configurable & customizable tools is best suited for organizations with trained and skilled engineers with deep understanding of the data source and destination

- The fully managed tools, zero-touch, are more accessible, standardized, stress-tested and maintenance-free where analysts can use SQL to perform additional transformations

# Automation and Pipelines

- Remove all manual/human intervention from the process by using **API** to programmatically control and administer your platform:
  - Handling data type changes
  - Schedule continuous synchronization using streams at short, regular intervals
  - Migrate schemas automatically
  - Manage general performance

# Recovery From Failure

**“Everything fails all the time”, Werner Vogels, Amazon CTO**

- Build with idempotence — the ability to repeatedly attempt the same process and produce the same result each time
- Build with “additive” integration – when a value is deleted in the source data, it’s retained (but flagged) in the DW to preserves historical records
- With fully managed tools, what is the service-level agreements?

# Security and Regulatory Compliance

- Regulatory compliance & standards like GDPR, SOC 2, HIPPA…
- Encryption or redaction of personally identifiable information (PII)
- Data sovereignty & retention
- Manage users authorization and access (GRC)

# Summary

- ETL or ELT will set the foundation to your Analytics landscape
- Use to be very constrained but is now free from resource limitation
- From highly configurable to fully managed / pre-packaged
- Must be build for failures and recovery

Source for this module: https://get.fivetran.com/rs/353-UTB-444/images/fivetran-2020-essential-guide-digital.pdf