Week 1
Where Does Data Come From & Tools for Data Engineering

ISTA 322 - Data Engineering

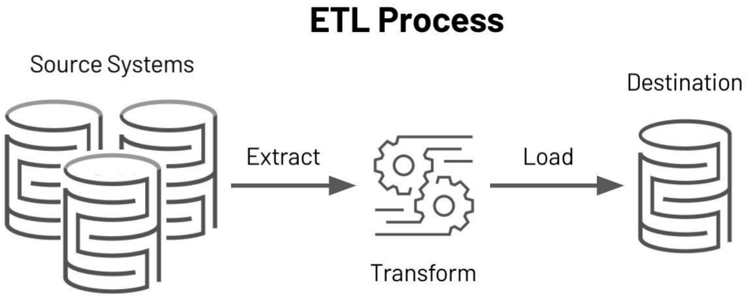
Recap of last lecture

- Data volumes have boomed in the last 20 years
- Some early companies were effective in using this
- This and other things (media, research) drove the potential for data scientists to use big data
- Early efforts did not go well as data is often in messy, not in immediately useful formats
- *And* there was lots of it which limited ability to process
- Early DS roles involved a lot of data engineering
- Now, there are explicit DE roles
- **DE is all about making data useful for analysis**

Where are we going today?

- Talk about where all these data are coming from
- The (generally) main job of a DE - making *ETLs
- Technologies using in DE and what subset we'll use

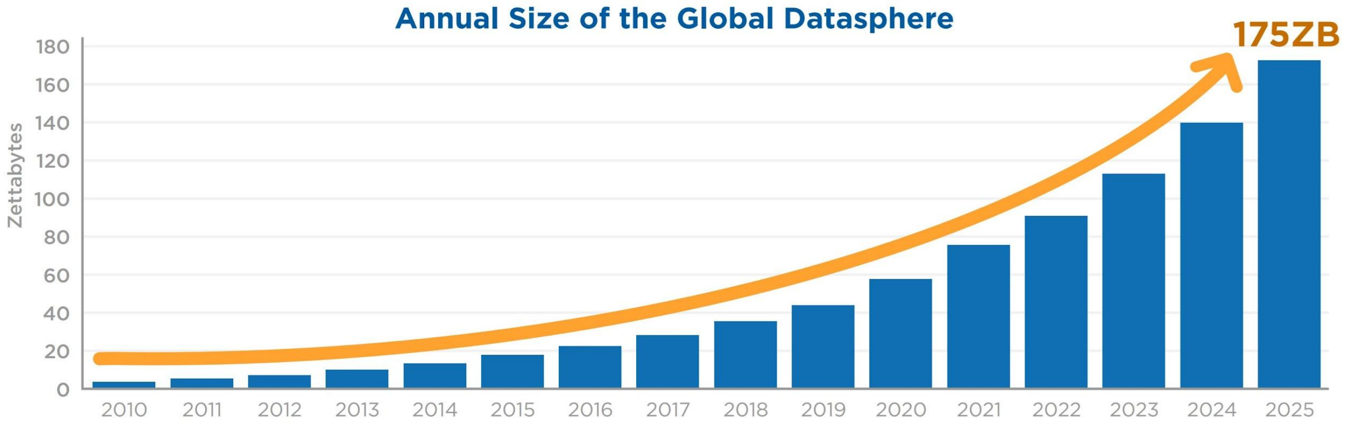
*ETL: Extract, transform, load



https://databricks.com/glossary/extract-transform-load

Where do data come from?

- There are tons of data and the amount being collected is exploding.
- What generates these data?
- Events!



1 Zetabyte is
- 350 trillion songs
- 100k copies of wikipedia

How events create data

- Events - data of actions performed by entities
  - Clicked on an ad
  - Left the page
  - Searched for something
  - Tried to log in
  - Made a transaction
  - Scrolled up or down
  - Liked, reacted, retweeted, hearted, shared
  - Uploaded a photo or video
  - Doesn’t have to be human… temperature probe recording, machine finishing a job, airplane sensors measuring tons of stuff, etc.
- Also will record time and who did it

How events create data

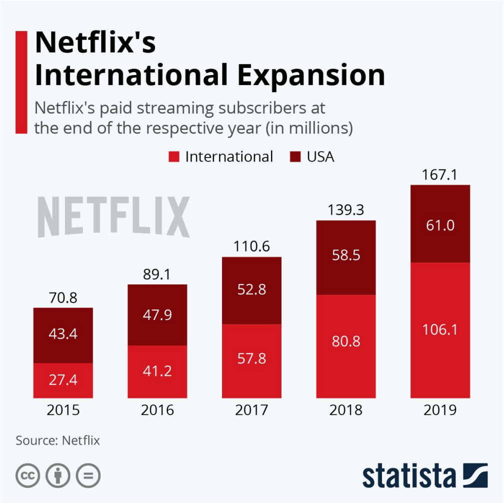
- Events will then be linked to other data that’s collected
- e.g. you click on Tiger King
  - {time : 07:26, event : click_watch, show_id : tk_S1E1, user_id : x88}
- These events are linked to other data that’s known about you or the show
- There will be a table that contains show info
  - {show_id : tk_S1E1, tags : [‘drama’, ‘reality’], runtime : 55min }
- And user info
  - {user_id : x88, age : 35, gender : ‘M’, OS : [‘wind’, ‘andro’]}

Event data at Netflix

- With this data arriving at over **2 million events per second**, getting it into a database that can be queried quickly is formidable. We need sufficient dimensionality for the data to be useful in isolating issues and as such **we generate over 115 billion rows per day**.
  - This was in 2020 - Ref @ Netflix Tech Blog
- They ‘*only*’ generated 10 billion rows a day in 2015
  - Ref @ Netflix Tech Blog

How events create data

- In just 5 years Netflix increased the amount of data collected by 10x (115 billion vs 10 billion)
  - Number of subscribers only increased by 2.5 - ref
  - Increased granularity of data collected
- This allows for more complex models & better analytics
  - “We need sufficient dimensionality for the data to be useful in isolating issues”
  - Remember, models need n x m matrix
  - More dimensions = more features in the matrix
  - More features = more models & better predictions
  - ∴ more money



Not all events

- Of course, not all data collected is structured like this
- Some is just stored in a database across multiple tables
  - Each transaction in a convenience store

TABLE ID: STORE		
store_id	**store_state**	**country**
az_23	AZ	USA
az_45	AZ	USA
ca_12	CA	USA
to_39	Ontario	Canada

TABLE ID: TRANSACTIONS			
transact_id	**store_id**	**UPC**	**price**
x88943	az_23	49914	2.57
x88943	az_23	99371	1.99
a85921	to_39	95831	8.99
a85921	to_39	99492	5.49
a85921	to_39	27482	4.49
z88930	az_45	33491	0.99

Not all events

- Of course, not all data collected is structured like this
- Some is just stored in a database across multiple tables
  - Each transaction in a convenience store
  - And data collected might not be optimized

A	B	C	D	E	F	G	H	I	J
id	name	host_id	host_name	neighbourhood_g	neighbourhood	latitude	longitude	room_type	price
2539	Clean & quiet apt	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149
2595	Skylit Midtown Ca	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225
3647	THE VILLAGE O	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150
3831	Cozy Entire Floor	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89
5022	Entire Apt: Spaci	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80
7322	Chelsea Perfect	18946	Doti	Manhattan	Chelsea	40.74192	-73.99501	Private room	140
7726	Hip Historic Brow	20950	Adam And Charit	Brooklyn	Crown Heights	40.67592	-73.94694	Entire home/apt	99
7750	Huge 2 BR Uppe	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190
7801	Sweet and Spaci	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.95718	Entire home/apt	299
8024	CBG CtyBGd He	22486	Lisel	Brooklyn	Park Slope	40.68069	-73.97706	Private room	130
8025	CBG Helps Haiti	22486	Lisel	Brooklyn	Park Slope	40.67989	-73.97798	Private room	80
8110	CBG Helps Haiti	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Private room	110

So what does a DE do again?

- Takes data from these various databases that are recording events/transactions/information
- Reorganizes it in some way or another into a format that lets people do analytics or data science
- Puts it in a database for them to use.
- This process has a general name - **ETL**
  - **Extract - Transform - Load**

ETL

- **ETLs are essentially the core of DE**
- That raw data in structured, semi-structured, or unstructured format is all stored in a **data lake**
- The transform step is going to remove errors, create features, scale values, aggregate data for metrics and whatever else is needed to support analytics and DS
- The transformed data is stored in a **data warehouse**

ETL

- From reading - Ch1 Data Mining Concepts and Techniques

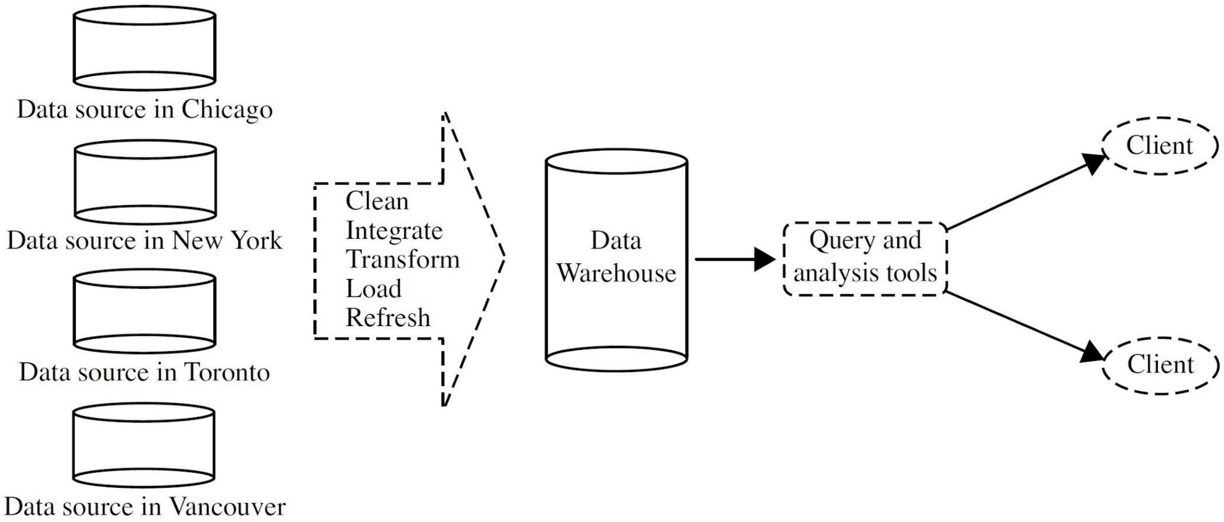


Figure 1.6 Typical framework of a data warehouse for *AllElectronics*.

But how to deal with so many events?

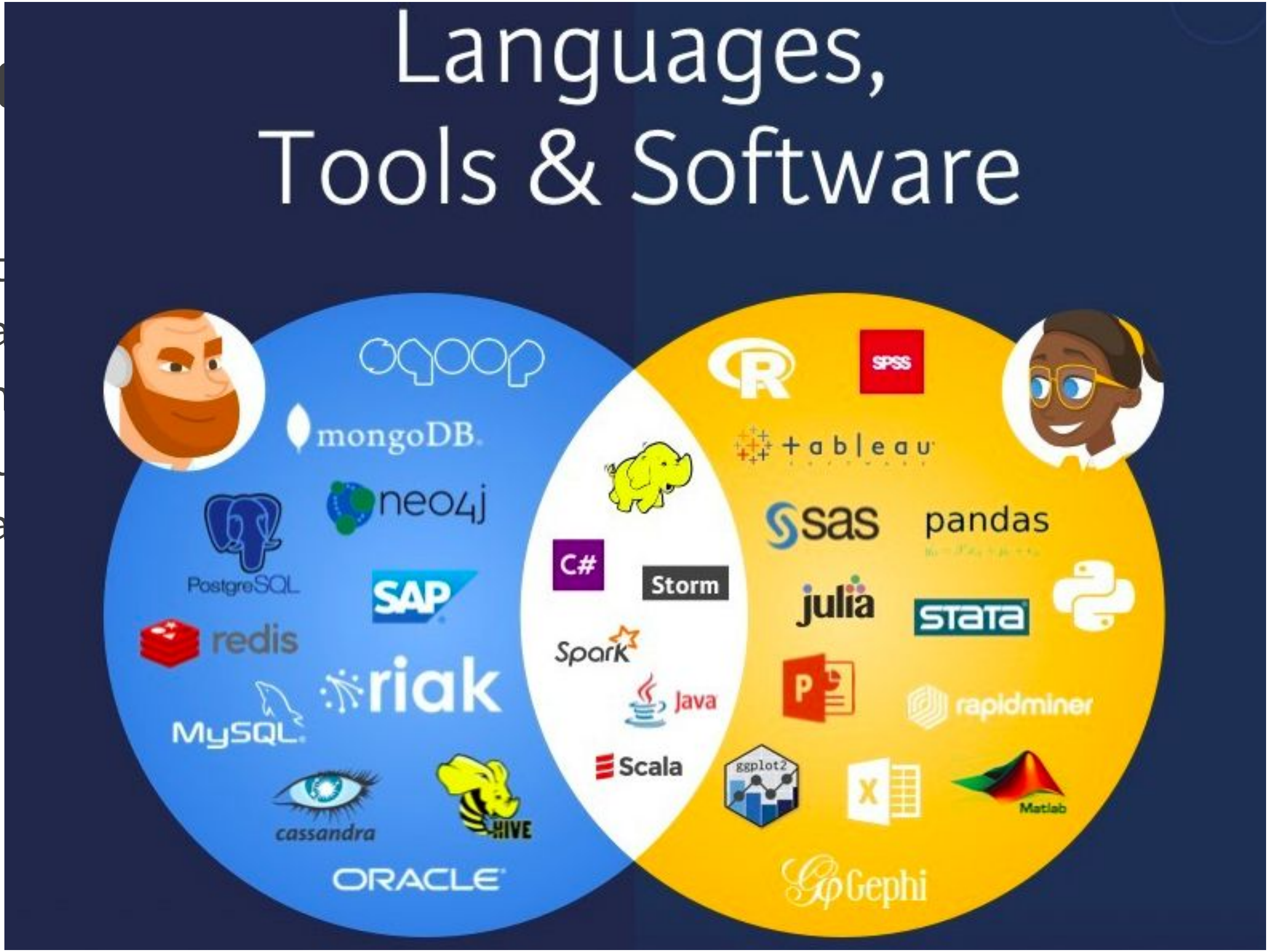
- OK, our goal is to get the data into a useful format
- But we're dealing ***lots*** of data
- Average computer has say 16gb of memory
  - A decade ago Facebook was dealing with 10+ gb of processed data a day
  - Amazon's daily login datafile alone is 1tb
- Obviously this is the other challenge of DE
  - How to deal with massive volumes of data fast enough to be useful
  - Can't let it take hours/days/weeks to process on one machine

Enter big data technologies

- The other part of of being a DE is using big data processing frameworks that allow for much, much faster data processing
- Technologies like hadoop/mapreduce and Spark utilized clusters of machines to distribute framework and optimize speed

Enter big data tec

- The other part
  processing fra
  data processin
- Technologies l
  clusters of ma
  speed



Enter big data technologies

- The other part of of being a DE is using big data processing frameworks that allow for much, much faster data processing
- Technologies like hadoop/mapreduce and Spark utilized clusters of machines to distribute framework and optimize speed
- It’s a massive ecosystem of tools - We’re only going to learn some of the essential tools

A bit more about the technologies we're going to use

- Languages / technologies
  - Python and pandas
  - SQL - Likely PostgreSQL
  - Pyspark locally
  - Pyspark via Databricks
- Environments
  - We'll be working in Jupyter Notebooks
  - Use Google Colaboratory - Google cloud based Jupyter Notebook
    - You'll download a notebook, upload and open there
  - You're welcome to use a local install, but I won't be providing tutorials (I can't troubleshoot 40 installs of all the libraries)
  - AWS - Pull from and set up database on AWS
  - Databricks - Cloud notebook based analytics/DS platform