
DATA & DATA INTENSIVE SYSTEMS

INTRO

- Computer Engineering at Georgia Tech
- Product management at Intel (datacenter, mobile)
- Kellogg MBA
- Built a social insights and analytics company now a part of Material
- Led many client engagements
  - 1st party & social data developing customer / market insights

brook@4brook.com

TA: AZIZA

- Bachelors in Finance and Data Science at MTSU
- Currently pursuing Masters in Artificial Intelligence at Northwestern
- Keenly involved in conducting research where she was part of:
    - Business and Economic Research Center at MTSU as a Research Assistant
    - Recipient of 2021 Station1 Frontiers Fellowship to participate in socially directed science and technology research.
    - Currently contributing to research at Tiilt Lab (Technological Innovations for Inclusive Learning & Teaching) from NLP point of view
- Developed EduAnalytics start up project aimed to automate the data analysis process for edu centers in Uzbekistan.

ROADMAP FOR OUR FIRST CLASS

- Course overview
- Twilio breakout
- Data models and Databases
- Database breakout
- 15 min break
- Lab intro
- Lab

OUR GOALS

Building skills for manipulating and interpreting data

Understanding data intensive systems (tech, process, people)

Delivering business outcomes

Where does the data come from?

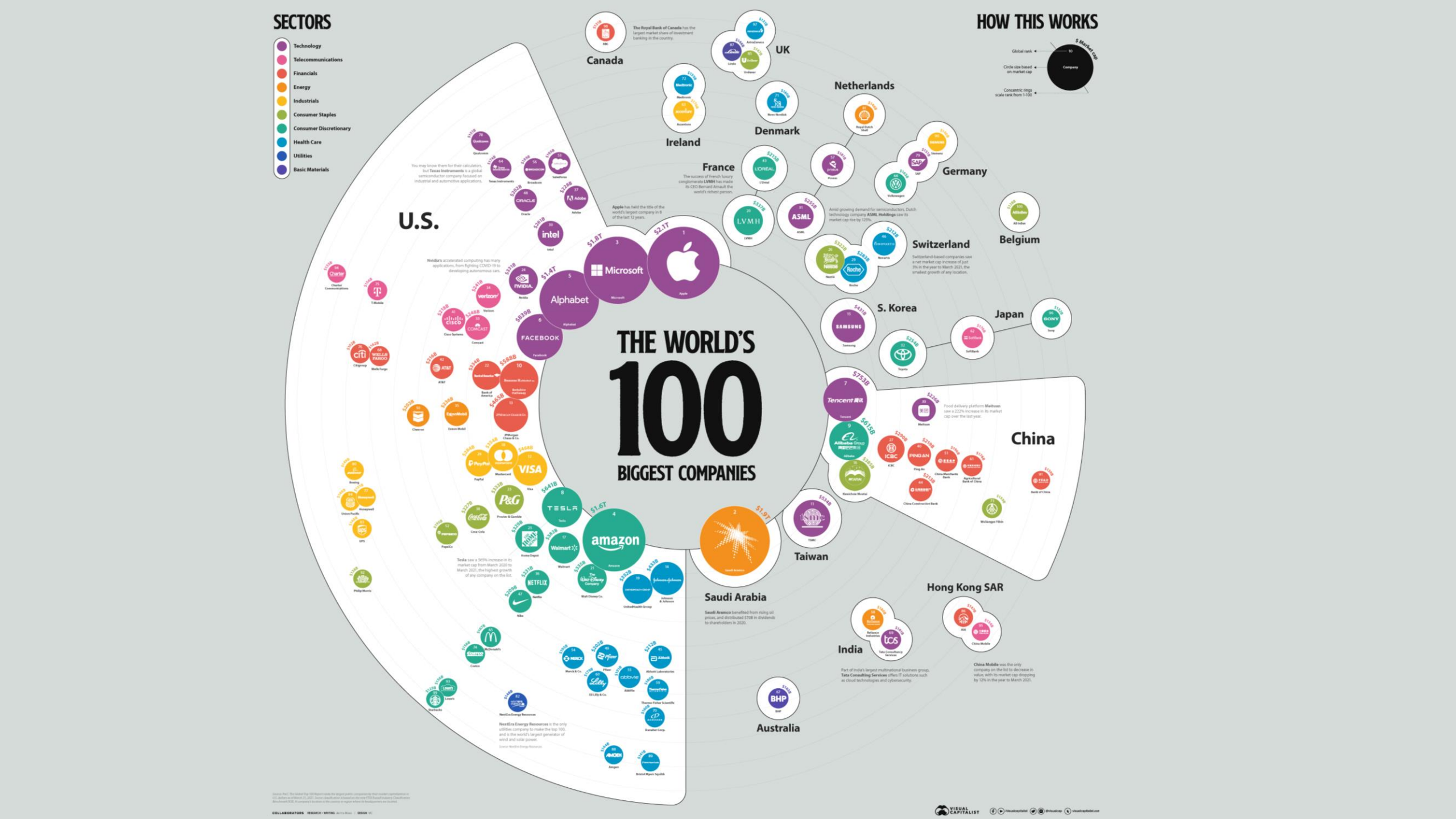
What happened to it along the way?

Does this create decision / operational / legal risks?

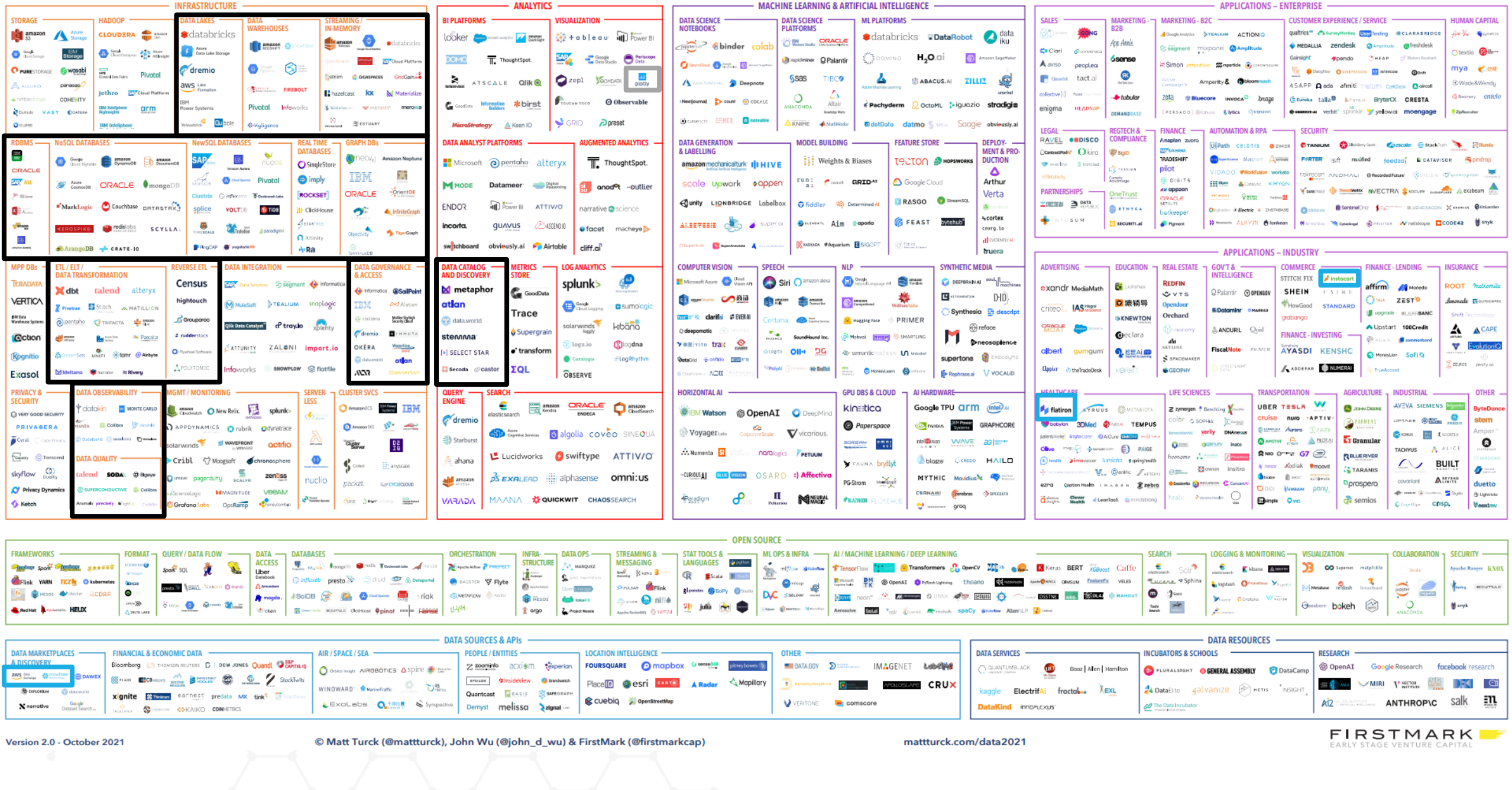

BUILDING SKILLS FOR MANIPULATING DATA - INEVITABLE
Intuition > Python
Why > What > How

DATA INTENSIVE SYSTEMS

- Product manager / working with / leading teams building (building on top of) data intensive systems
- Working at a data intensive company
- Analyst / consumer of analysis of data


SECTORS
Technology
Telecommunications
Financials
Energy
Industrials
Consumer Staples
Consumer Discretionary
Health Care
Utilities
Basic Materials
HOW THIS WORKS
$ Market cap
Global rank
Circle size based on market cap
Concentric rings scale rank from 1-100
Company
THE WORLD'S 100 BIGGEST COMPANIES
U.S.
Canada
The Royal Bank of Canada has the largest market share of investment banking in the country.
UK
AstraZeneca
Linde
Unilever
Ireland
Medtronic
Accenture
Denmark
Novo Nordisk
Netherlands
Royal Dutch Shell
Prosus
France
The success of French luxury conglomerate LVMH has made its CEO Bernard Arnault the world's richest person.
L'Oréal
LVMH
ASML
Amid growing demand for semiconductors, Dutch technology company ASML Holdings saw its market cap rise by 125%.
Germany
Siemens
SAP
Volkswagen
Switzerland
Switzerland-based companies saw a net market cap increase of just 3% in the year to March 2021, the smallest growth of any location.
Nestlé
Roche
Novartis
Belgium
AB InBev
S. Korea
Samsung
Japan
Toyota
SoftBank
Sony
China
Food delivery platform Meituan saw a 222% increase in its market cap over the last year.
Tencent
Alibaba Group
Kweichow Moutai
ICBC
Ping An
China Merchants Bank
Agricultural Bank of China
China Construction Bank
Bank of China
Wuliangye Yibin
Taiwan
TSMC
Saudi Arabia
Saudi Aramco benefited from rising oil prices, and distributed $70B in dividends to shareholders in 2020.
Hong Kong SAR
AIA
China Mobile
China Mobile was the only company on the list to decrease in value, with its market cap dropping by 12% in the year to March 2021.
India
Reliance Industries
Tata Consultancy Services
Part of India's largest multinational business group, Tata Consulting Services offers IT solutions such as cloud technologies and cybersecurity.
Australia
BHP
Apple has held the title of the world's largest company in 8 of the last 12 years.
Nvidia's accelerated computing has many applications, from fighting COVID-19 to developing autonomous cars.
You may know them for their calculators, but Texas Instruments is a global semiconductor company focused on industrial and automotive applications.
Tesla saw a 365% increase in its market cap from March 2020 to March 2021, the highest growth of any company on the list.
NextEra Energy Resources is the only utilities company to make the top 100, and is the world's largest generator of wind and solar power.
Source: NextEra Energy Resources
Microsoft
Apple
Alphabet
Facebook
Tesla
Amazon
Visa
Mastercard
PayPal
Procter & Gamble
Coca-Cola
Home Depot
Walmart
Walt Disney Co.
Netflix
Nike
UnitedHealth Group
Johnson & Johnson
Pfizer
Merck & Co.
Eli Lilly & Co.
AbbVie
Abbott Laboratories
Thermo Fisher Scientific
Danaher Corp.
Amgen
Bristol Myers Squibb
Philip Morris
McDonald's
Costco
Starbucks
Lowe's
Union Pacific
Honeywell
UPS
PepsiCo
Exxon Mobil
Chevron
Bank of America
Berkshire Hathaway
JPMorgan Chase & Co.
AT&T
Wells Fargo
Citigroup
Comcast
Cisco Systems
Verizon
Nvidia
T-Mobile
Charter Communications
Qualcomm
Texas Instruments
Broadcom
Salesforce
Oracle
Adobe
Intel
Boeing
Source: PwC, The Global Top 100 Report ranks the largest public companies by their market capitalization in U.S. dollars as of March 31, 2021. Sector classification is based on the new FTSE Russell Industry Classification Benchmark (ICB). A company's location is the country or region where its headquarters are located.
COLLABORATORS RESEARCH + WRITING Jenny Ross | DESIGN VC
VISUAL CAPITALIST
@visualcapitalist
@thevisualcap
visualcapitalist.com

MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021



Version 2.0 - October 2021

mattturck.com/data2021

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

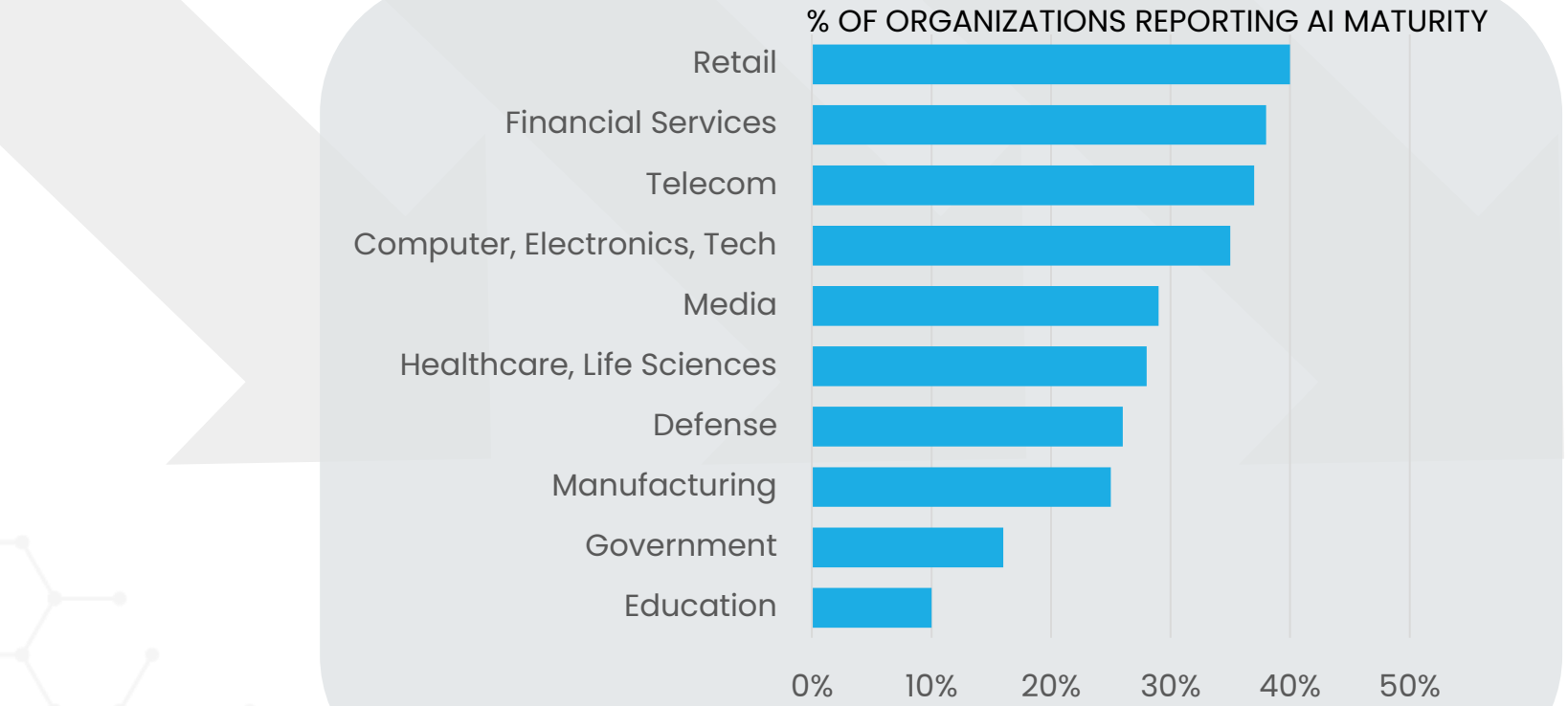
MEGA-TRENDS

2005
internet native
oss

2010
big data
cloud

2015
ML/AI

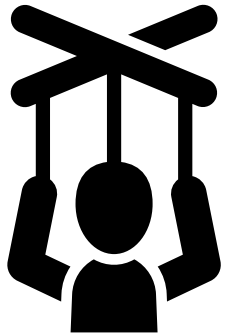
2020+
cloud ML/AI
+ data types

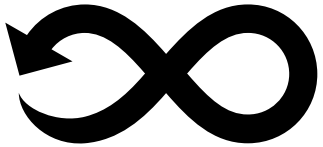


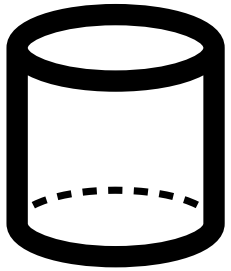
https://www.oreilly.com/radar/ai-adoption-in-the-enterprise-2021/


BUSINESS OUTCOMES

INCREASED SOPHISTICATION IN HOW WE THINK ABOUT DATA







Usage	Lifecycle	Modality	Challenges
Reporting Analytics Decision Support Data Driven Products	Acquire Analyze Activate Archive	Data Model Structured Unstructured Files APIs	Access Scale (Latency,Throughput,Volume) Quality Governance Discoverability Bias


HOUSEKEEPING

CLASS PARTICIPATION

- Grade: 10% participation, 10% attendance
  - Quality >> quantity
- Use office hours if you would like to discuss topics that seem too narrow or in-depth for class Q&A, discussion
  - We are covering a lot of ground in this class
- Avoid acronyms and jargon, like Thanos it is a bit inevitable
- Vendor agnostic, but we'll use some examples from key vendors that highlight important details
- Learning to drive != Building a car engine
  - Eventually relevant - an F1 engineer will have a different perspective to the driver

REFLECTIONS

- 1 per class, 20% of grade. Groups of ~4. ~750 words +/-250, prefer link to google colab markdown doc.
  - Template
- Due by Tuesday at 10PM Central, same as homework, submit one copy via Canvas. (same as homework assignment)
- Twilio - Alfredo Sone, Taylor Baker, Cameryn Steiger, Tushita Singhel
- Databases - Jack Van Vleck, Hannie Gulcin, Benjamin Housley, Tiffany Dsouza
- Colab & SQL- Meet Patel, Katie Williamson, Tianyi He, Jason Feng
- Reflection will be shared to all students so you can keep copies

HOMEWORK

- 4 Paired assignments
- Available Thursday 8AM, Due Tuesday by 10pm
- Explore the data in a notebook based environment, accomplishing the key elements / techniques demonstrated in class
- Add a final markdown cell of < 250 words with your team's recommendation

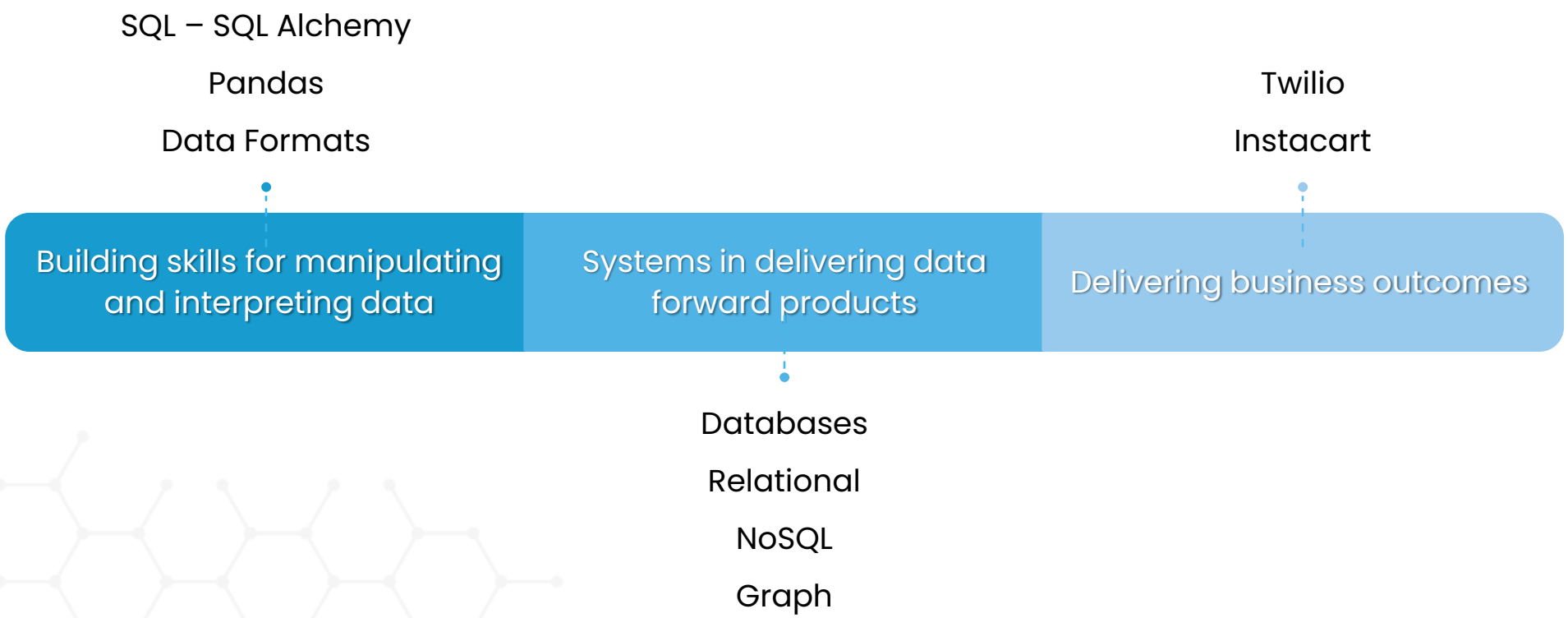
MORE INFO

- I'll post the class slides and additional readings after class

- Class liaison?
  - Please let me know the most useful time for office hours (tentatively Thursday 2PM)
  - Thursday 10:30AM central to 5:30PM central.

DATA MODELS & DATABASES

OUR GOALS – THIS CLASS

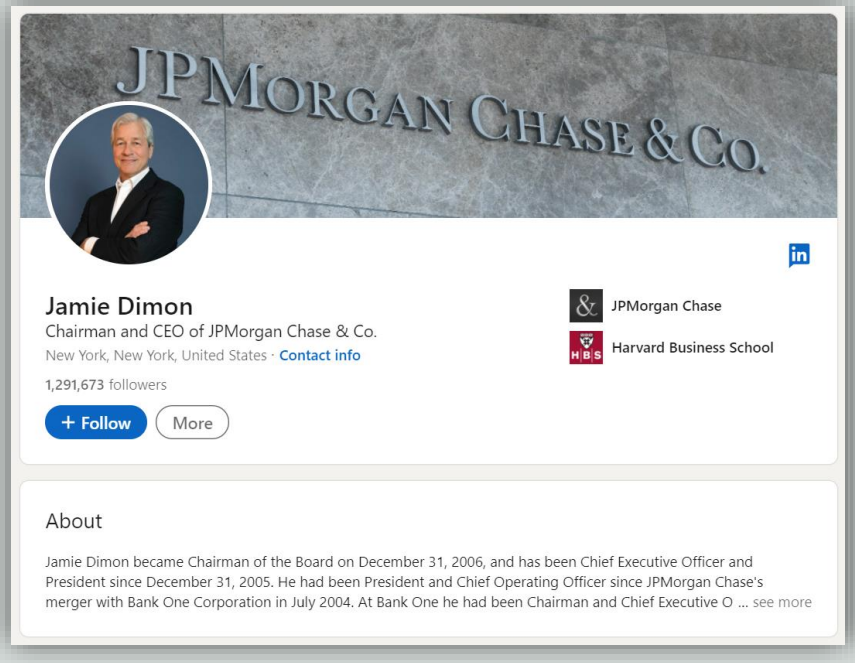


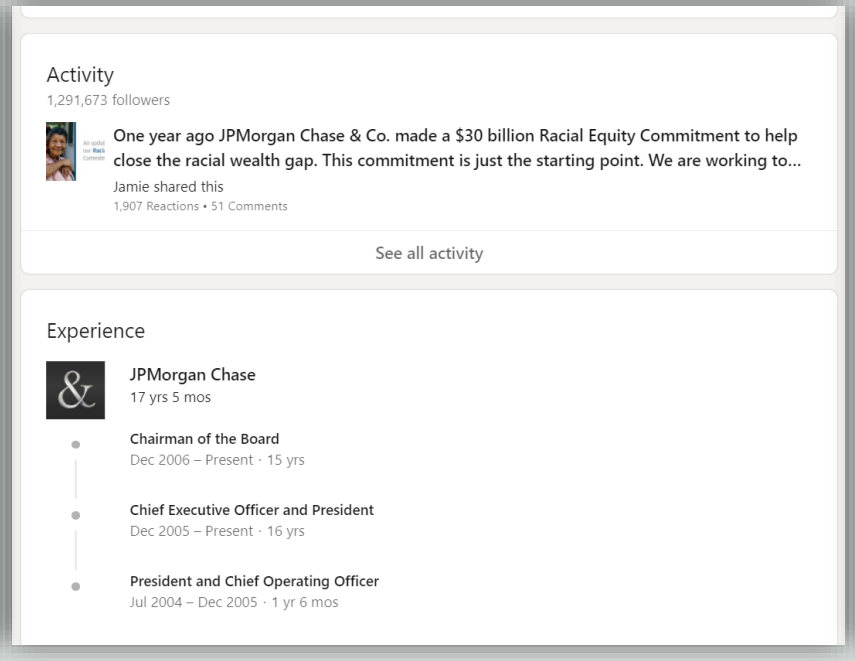
LINK

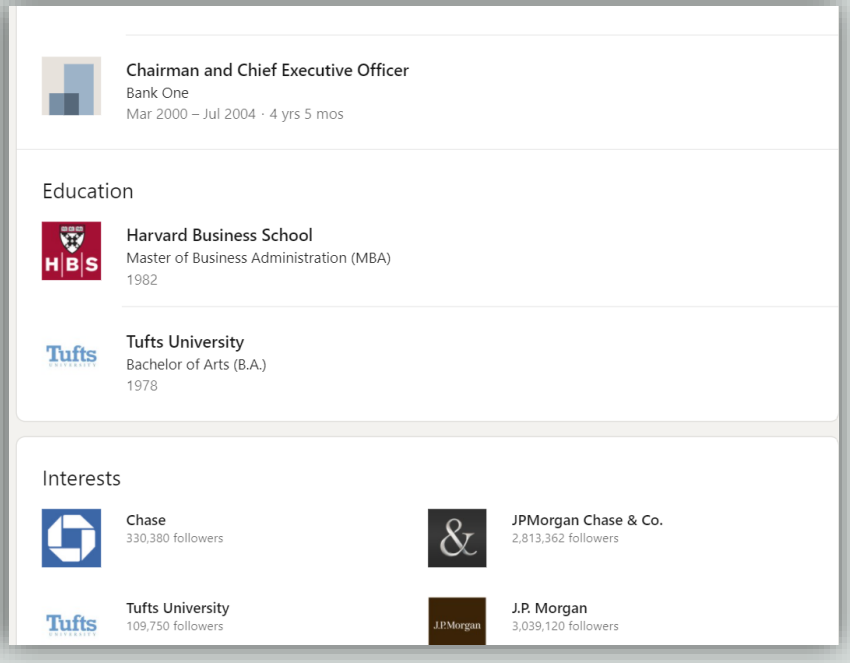
https://s21.q4cdn.com/963721274/files/doc_presentations/2020/10/2020-Twilio-Investor-Day-FINAL-WEBSITE.pdf

DATA MODELS

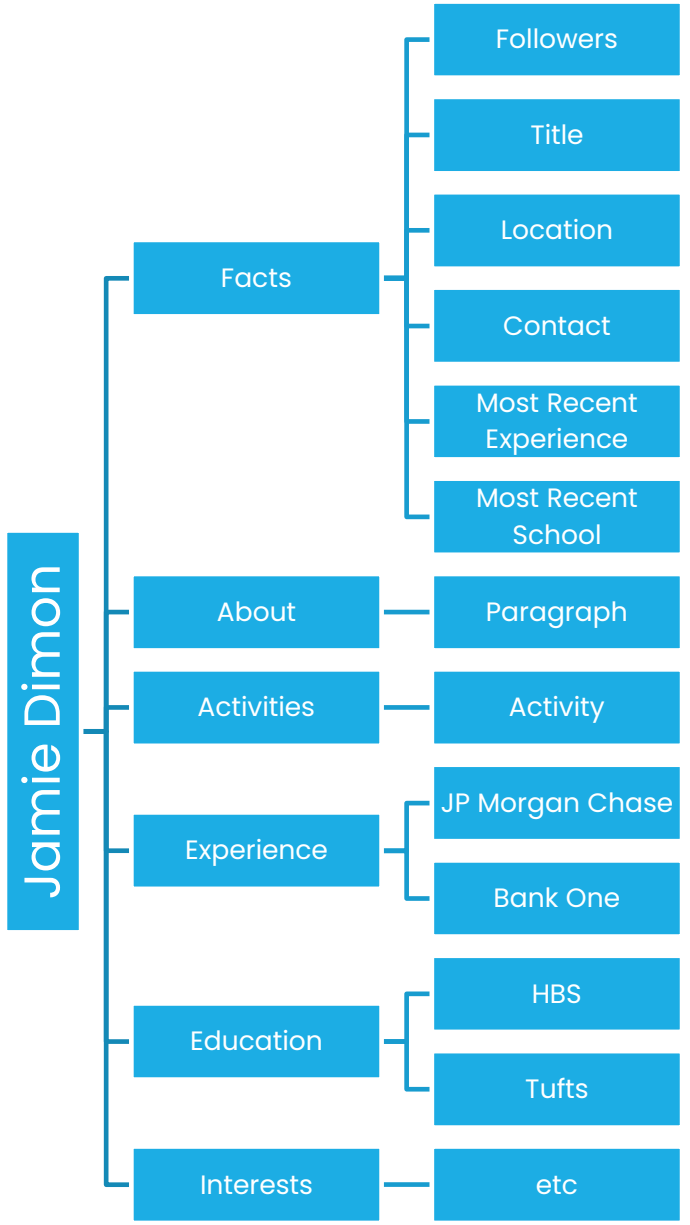
LINKEDIN PROFILE







THE DOCUMENT MODEL



THE SPREADSHEET

id	first_name	last_name	email	skills
1	Rusty	Sheldrick	rsheldrick0@barnesandnoble.com	Phased systemic capability
2	Gabi	Grenshields	ggrenshields1@51.la	Enterprise-wide context-sensitive initiative
3	Godwin	Brandli	gbrandli2@infoseek.co.jp	Proactive upward-trending initiative
4	Linette	Tease	ltease3@devhub.com	Vision-oriented reciprocal parallelism
5	Yetta	Plastow	yplastow4@wired.com	Devolved zero administration intranet
6	Vick	Folshom	vfolshom5@youtube.com	Upgradable coherent framework
7	Morlee	Earlam	mearlam6@ning.com	Innovative stable website
8	Wilmer	Brandolini	wbrandolini7@washingtonpost.com	Secured analyzing workforce
9	Chaim	Van der Baaren	cvanderbaaren8@webnode.com	Realigned incremental access
10	Hedy	Cuttle	hcuttle9@liveinternet.ru	Managed coherent core
11	Alikee	O'Halloran	aohallorana@sun.com	Multi-lateral tertiary knowledge base
12	Melloney	Danbury	mdanburyb@people.com.cn	Extended content-based time-frame
13	Mallory	Hargey	mhargeyc@nyu.edu	Integrated empowering data-warehouse
14	Charyl	Ralling	crallingd@angelfire.com	Business-focused 5th generation middleware
15	Arther	Kidston	akidstone@printfriendly.com	De-engineered mission-critical implementation

Data from: Mockaroo

EXTENDING THE SPREADSHEET FOR EDUCATION

Query

Update

id	first_name	last_name	email	education1	education2	education3
331	Scottie	Mellmer	smellmer96@icio.us	Stephens College		Gujarat Ayurved University
332	Frankie	Cramp	fcramp97@state.tx.us	Universidad Nacional de General Sarmiento	Chulalongkorn University	Rajasthan Technical University
333	Nigel	Templeman	ntempleman98@theglobeandmail.c	Ulsan University		
334	Bastien	Cersey	bcersey99@latimes.com	Ilam University	Tianjin University of Commerce	
335	Auberon	Crosham	acrosham9a@opera.com	Nkumba University		
336	Elyn	Lightfoot	elightfoot9b@cdc.gov	Technological University (Taungoo)		
337	Wilona	Borrott	wborrott9c@hp.com	Polonia University in Częstochowa		
338	Dasie	Acum	dacum9d@soup.io	Liaoning Technical University	Medaille College	
339	Lesli	Robathon	lrobathon9e@youtube.com	National Law School of India University		
340	Frank	Chaldecott	fchaldecott9f@home.pl	Southern California Institute of Architecture		
341	Iain	Roylance	iroylance9g@foxnews.com	Hanoi University of Mining and Geology	Universidad Industrial de Santander	
342	Yolane	Dorrity	ydorrity9h@sbwire.com	Perm State University		
343	Tammy	de Mendoza	tdemendoza9i@bloomberg.com	Fachhochschule Biberach, Hochschule für Bauwesen (	Chunchon National University of Education	
344	Kerry	Snoday	ksnoday9j@sphinn.com	Kirikkale University	Nizhny Novgorod State Architectural - Building	
345	Robinett	Stegel	rstegel9k@baidu.com	Lagos State Polytechnic	Surugadai University	
346	Lorri	Grinham	lgrinham9l@pbs.org	Jamia Hamdard University	Life University	Universidad Peruana de Ciencias Aplicadas
347	Clem	Bartunek	cbartunek9m@tiny.cc	Florida Community College at Jacksonville	University of Minnesota - Crookston	
348	Humfried	Sinfield	hsinfield9n@cbsnews.com	Istanbul Ticaret University		
349	Jud	Posvner	jposvner9o@pen.io	Gorno-Altaisk State University	Kodolanyi Janos University College	
350	Kalina	Attock	kattock9p@npr.org	University College Dublin		
351	Burlie	Moseley	bmoseley9q@rambler.ru	Universiteit Doesburg (UNDO)	Ecole Nationale Supérieure du Pétrole et des Mc	
352	Jeffy	Balding	jbalding9r@meetup.com	Wah Medical College	Al-Buraimi University College	Hardin-Simmons University

if (student.education1 == target & student.education2 == ...

THE RELATIONAL MODEL

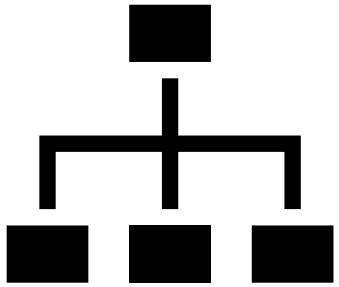
id	first_name	last_name	email
331	Scottie	Mellmer	smellmer96@icio.us
332	Frankie	Cramp	fcramp97@state.tx.us
333	Nigel	Templeman	ntempleman98@theglobeandmail.c
334	Bastien	Cersey	bcersey99@latimes.com
335	Auberon	Crosham	acrosham9a@opera.com
336	Elyn	Lightfoot	elightfoot9b@cdc.gov
337	Wilona	Borrott	wborrott9c@hp.com
338	Dasie	Acum	dacum9d@soup.io
339	Lesli	Robathon	lrobathon9e@youtube.com
340	Frank	Chaldecott	fchaldecott9f@home.pl
341	Iain	Roylance	iroylance9g@foxnews.com
342	Yolane	Dorrity	ydorrity9h@sbwire.com
343	Tammy	de Mendoza	tdemendoza9i@bloomberg.com
344	Kerry	Snoday	ksnoday9j@sphinn.com
345	Robinett	Stegel	rstegel9k@baidu.com
346	Lorri	Grinham	lgrinham9l@pbs.org
347	Clem	Bartunek	cbartunek9m@tiny.cc
348	Humfried	Sinfield	hsinfield9n@cbsnews.com
349	Jud	Posvner	jposvner9o@pen.io
350	Kalina	Attock	kattock9p@npr.org
351	Burlie	Moseley	bmoseley9q@rambler.ru
352	Jeffy	Balding	jbalding9r@meetup.com

student_id	institution_id
331	45
331	12

Id	Institution
1	3rd Military Medical University
2	École Nationale Supérieure d'Ingénieurs de Constructions Aéronautiques
3	Abasyn University Peshawar
4	Abubakar Tafawa Balewa University
5	Academy of Economics in Wroclaw
6	Academy of Public Administration
7	Adamson University
8	Adeyemi College of Education
9	Afghan University
10	Air University
11	Akhbar El Yom Academy
12	Al Nahrain University
13	Alabama Agricultural and Mechanical University
14	Alain University of Science and Technology
15	Al-Birony University
16	Alfaisal University
17	Alhamd Islamic University
18	All India Institute of Medical Sciences
19	American College of Thessaloniki
20	American University Extension, Okinawa
21	American University of the Caribbean

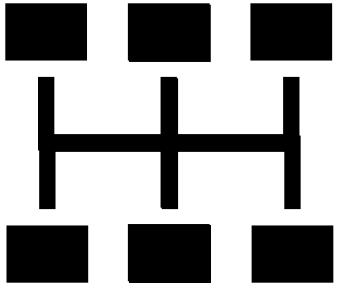
SPARESENESS -> REDUCED RESOURCES
DRY – NORMALIZED
CONSISTENT UPDATES

RELATIONSHIPS



One to many

- Customer -> Orders



Many to many

- People -> Institutions

What constraints need to be enforced to ensure our relationships are valid?

ACQUIRING DATA

STORING

FILES

CONSUMING

DATABASES

MOVING

APIS

STORING DATA: FILES

- Common desktop: Xlsx, google sheets, csv
- Meta data: Exif for images, parquet for data
- Compression, Row vs. column locality
- Many proprietary and performance optimized formats
  - Commonly used and supported by pandas: csv, json arrays, parquet

JSON

(TWITTER TIMELINE)


```
}, {
    "entryId": "tweet-1478814463629271040",
    "sortIndex": "1478814463629271040",
    "content": {
        "entryType": "TimelineTimelineItem",
        "itemContent": {
            "itemType": "TimelineTweet",
            "tweet_results": {
                "result": {
                    "__typename": "Tweet",
                    "rest_id": "1478814463629271040",
                    "core": {
                        "user_results": {
                            "result": {
                                "__typename": "User",
                                "id": "VXNlcjo4NjYyNjg0NQ==",
                                "rest_id": "86626845",
                                "affiliates_highlighted_label": {},
                                "has_nft_avatar": false,
                                "legacy": {
                                    "blocked_by": false,
                                    "blocking": false,
                                    "can_dm": false,
                                    "can_media_tag": false,
                                    "created_at": "Sun Nov 01 00:02:55 +0000 2009",
                                    "default_profile": false,
                                    "default_profile_image": false,
                                    "description": "physician-scientist, author, editor  \nhttps://t.co/vwstu2BYrC",
                                    "entities": {
                                        "description": {
                                            "urls": [{
                                                "display_url": "drerictopol.com",
                                                "expanded_url": "http://drerictopol.com",
                                                "url": "https://t.co/vwstu2BYrC",
                                                "indices": [38, 61]
                                            }]
                                        },
                                        "url": {
                                            "urls": [{
                                                "display_url": "scripps.edu/translational",
                                                "expanded_url": "http://www.scripps.edu/translational",
                                                "url": "https://t.co/Z53npawhDU",
                                                "indices": [0, 23]
                                            }]
                                        }
                                    },
                                    "fast_followers_count": 0,
                                    "favourites_count": 40634,
                                    "follow_request_sent": false,
                                    "followed_by": false,
                                    "followers_count": 593000,
                                    "following": true,
                                    "friends_count": 534,
                                    "has_custom_timelines": true,
                                    "is_translator": false,
                                    "listed_count": 11192,
                                    "location": "La Jolla, CA",
                                    "media_count": 24527,
                                    "muting": false,
                                    "name": "Eric Topol",
                                    "normal_followers_count": 593000,
```

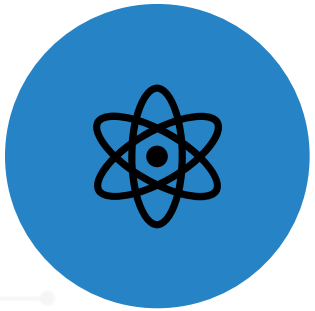
RELATIONAL DATABASES

- Schema of tables and relations
- SCHEMA: This describes how real-world entities are modeled in the database. Tables, relations, and constraints (what data will we not allow in)
- Query with SQL
- Most support: text, JSON, binary data types / columns
- Optimized for storage, compute & query

Source: https://mattturck.com/data2021/

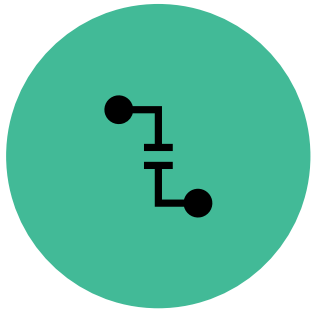
PostgreSQL

ACID TRANSACTIONS



ATOMICITY

CONSISTENCY



ISOLATION

DURABILITY

DOCUMENT ORIENTED DATABASES (TYPICALLY NoSQL)

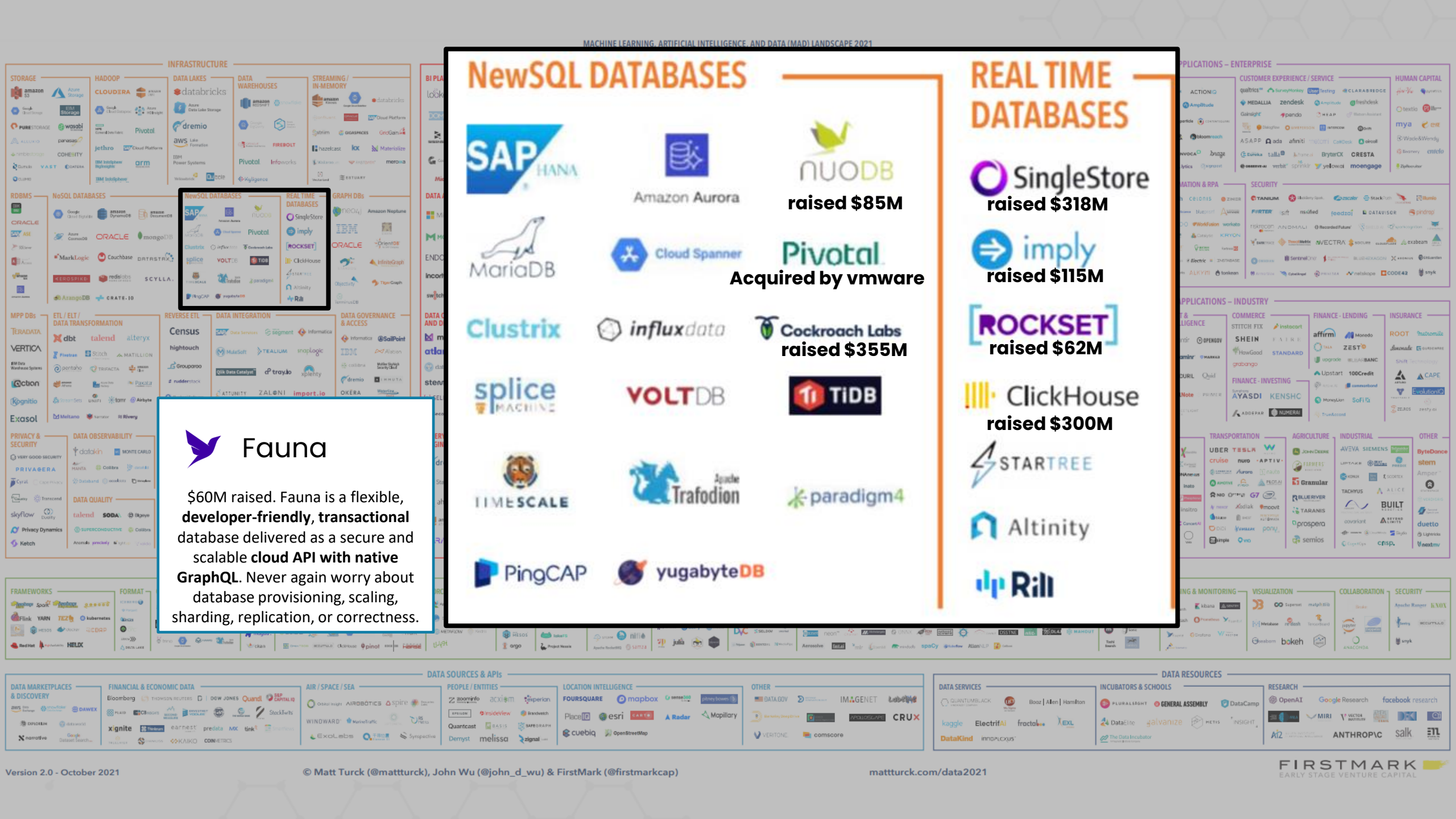
- Data typically stored in json / document model
- API: simplest query using key / id fetch typically support indexing and partitioning across pre-defined keys
- Many now support sql

Source: https://mattturck.com/data2021/

cassandra

GRAPH DATABASES

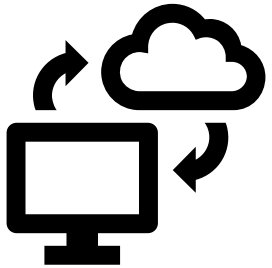
- A graph database stores nodes and relationships instead of tables, or documents.
- Nodes (data) and edges (relationships)
- Data is stored without restricting it to a pre-defined model, perhaps multiple data models
- Facebook Tao


MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, AND DATA (MAD) LANDSCAPE 2021
NewSQL DATABASES
SAP HANA
Amazon Aurora
NUODB
raised $85M
MariaDB
Cloud Spanner
Pivotal
Acquired by vmware
Clustrix
influxdata
Cockroach Labs
raised $355M
splice MACHINE
VOLTDB
TiDB
TIMESCALE
Apache Trafodion
paradigm4
PingCAP
yugabyteDB
REAL TIME DATABASES
SingleStore
raised $318M
imply
raised $115M
ROCKSET
raised $62M
ClickHouse
raised $300M
STARTREE
Altinity
Rill
Fauna
$60M raised. Fauna is a flexible, developer-friendly, transactional database delivered as a secure and scalable cloud API with native GraphQL. Never again worry about database provisioning, scaling, sharding, replication, or correctness.
INFRASTRUCTURE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
DATA SOURCES & APIs
DATA RESOURCES
Version 2.0 - October 2021
© Matt Turck (@mattturck), John Wu (@john_d_wu) & FirstMark (@firstmarkcap)
mattturck.com/data2021
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

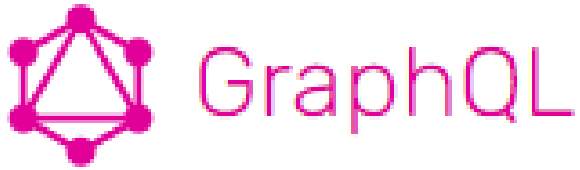

DATABASE BREAKOUT

APIS

- Databases typically support proprietary data connectors that are language / library dependent
  - ODBC/JDBC
- REST - JSON using http verbs
  - CRUD
- Protobuf
- GraphQL



REST



BREAK


USING SQL TO ANALYZE GROCERY SALES

Brandon Rohrer
@_brohrer_

ML strategy tip

When you have a problem, build two solutions - a deep Bayesian transformer running on multicloud Kubernetes and a SQL query built on a stack of egregiously oversimplifying assumptions. Put one on your resume, the other in production. Everyone goes home happy.

3:45 AM · Aug 12, 2021 · Twitter for iPhone

696 Retweets **114** Quote Tweets **4,081** Likes

orders

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2398795	1	prior	2	3	7	15
2254736	1	prior	4	4	7	29
3367565	1	prior	6	2	7	19
3108588	1	prior	8	1	14	14
2550362	1	prior	10	4	8	30
2168274	2	prior	1	2	11	nan
1901567	2	prior	3	1	10	3
1673511	2	prior	5	3	11	8
3194192	2	prior	7	2	12	14
1718559	2	prior	9	2	9	8
1402090	2	prior	11	1	10	30

primary key

aisles

aisle_id	aisle
1	prepared soups salads
2	specialty cheeses
3	energy granola bars
4	instant foods
5	marinades meat prepar
6	other
7	packaged meat
8	bakery desserts
9	pasta sauce
10	kitchen supplies

foreign key

order_id	product_id	add_to_cart_order	reordered
2	33120	1	1
2	9327	3	0
2	30035	5	0
2	40141	7	1
2	43668	9	0
3	24838	2	1
3	21903	4	1
3	46667	6	1
3	32665	8	1
4	26434	2	1
4	27761	4	1

order_products

product_id	product_name	aisle_id	department_id
2	All-Seasons Salt	104	13
4	Smart Ones Classic Favor…	38	1
6	Dry Nose Oil	11	11
8	Cut Russet Potatoes Stea…	116	1
10	Sparkling Orange Juice &…	115	7
12	Chocolate Fudge Layer Ca…	119	1
14	Fresh Scent Dishwasher C…	74	17
16	Mint Chocolate Flavored …	103	19
18	Pizza for One Suprema Fr…	79	1
20	Pomegranate Cranberry & …	98	7
22	Fresh Breath Oral Rinse	20	11

products

department_id	department
1	frozen
2	other
3	bakery
4	produce
5	alcohol
6	international
7	beverages
8	pets
9	dry goods pas…
10	bulk

departments