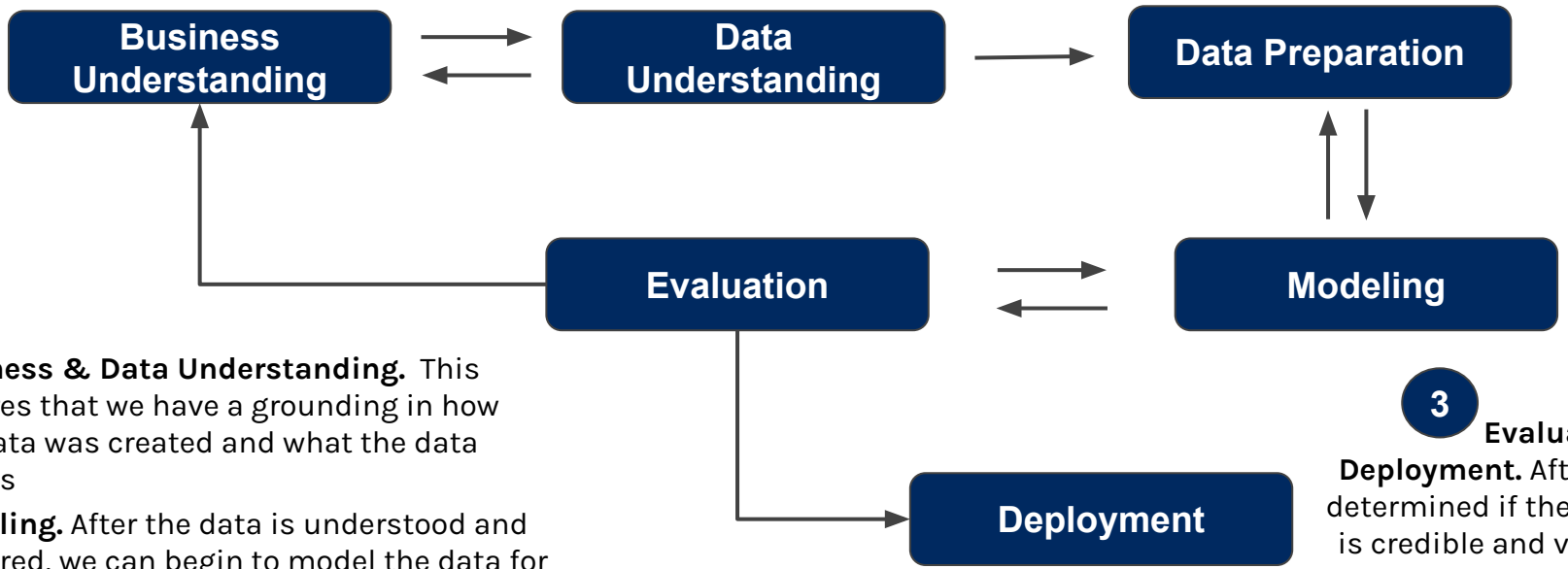
HST 953/6.8850: Clinical Data Learning

Ned McCague

Cross-industry standard process for data mining (CRISP-DM) with new data



1. **Business & Data Understanding.** This ensures that we have a grounding in how the data was created and what the data means
2. **Modeling.** After the data is understood and prepared, we can begin to model the data for insights
3. **Evaluation & Deployment.** After we've determined if the insight is credible and valuable, we can enhance our own business understanding or deploy the finding in some fashion

Reminder

- PSET Three is due next week!
- Reflections are still due
- Final presentations AND the final report are fast approaching

Data and Healthcare in the news

Speculation about the impact of election

Humana Inc
NYSE: HUM

Overview | Compare | Financials

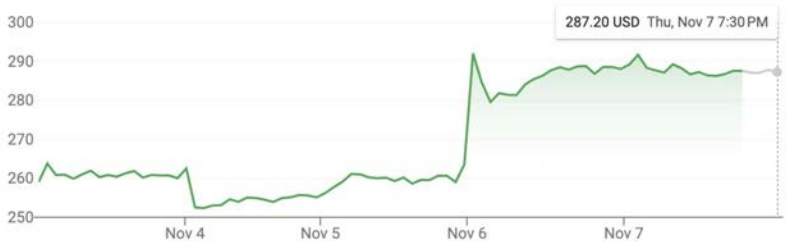
Market Summary > Humana Inc

287.76 USD

+28.76 (11.10%) ↑ past 5 days

Closed: Nov 8, 8:43 AM EST • Disclaimer
Pre-market 288.10 +0.34 (0.12%)

1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max



Open	-	Mkt cap	34.65B	CDP score	B-
High	-	P/E ratio	25.48	52-wk high	527.18
Low	-	Div yield	1.23%	52-wk low	213.31

Trump Will Create New Winners and Losers in Healthcare

Market reactions sharply diverge among healthcare companies as investors expect changes to Obamacare, Medicare

By David Wainer Follow
Nov. 6, 2024 11:17 am ET

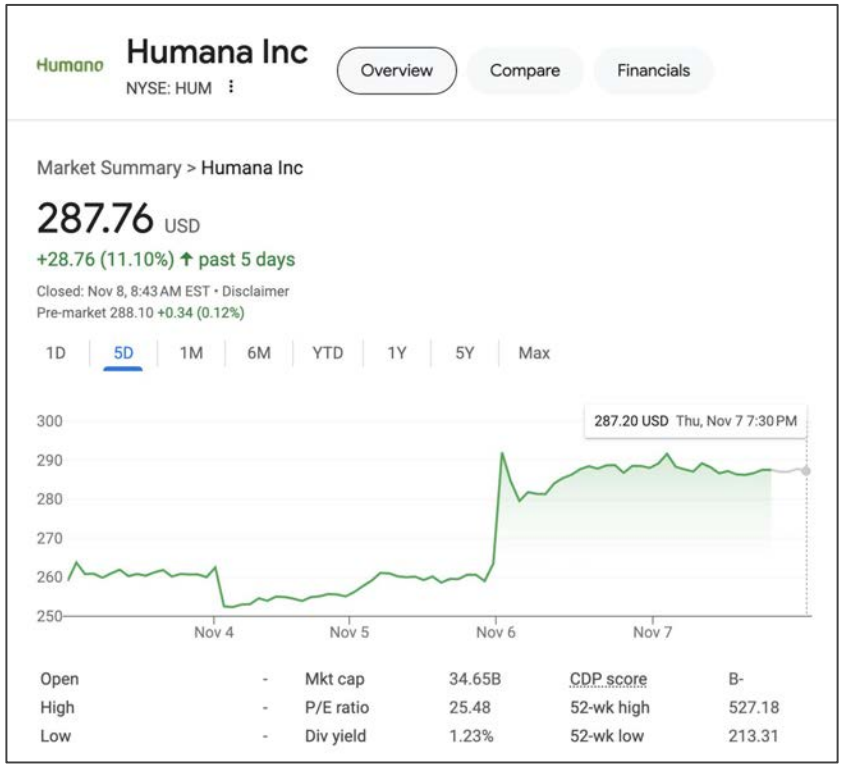
Share | Resize | Listen (3 min)

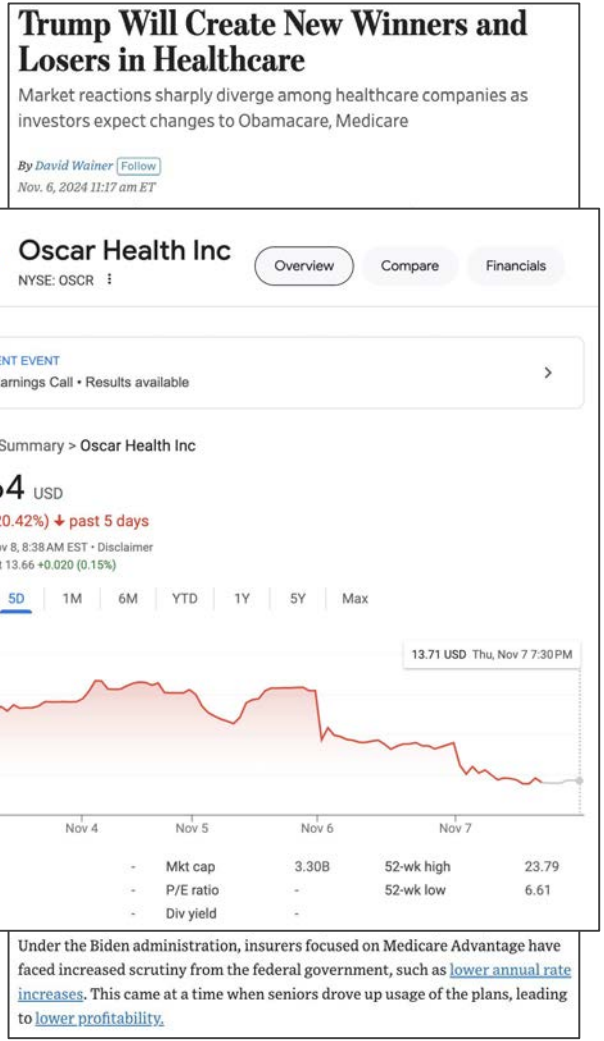
HCA Healthcare is a hospital chain that benefits from more government money pouring into Medicaid and the exchanges. PHOTO: DOUGLAS R. CLIFFORD/ZUMA PRESS

For healthcare companies, Donald Trump's victory means very different things depending on which part of the sector they operate in.

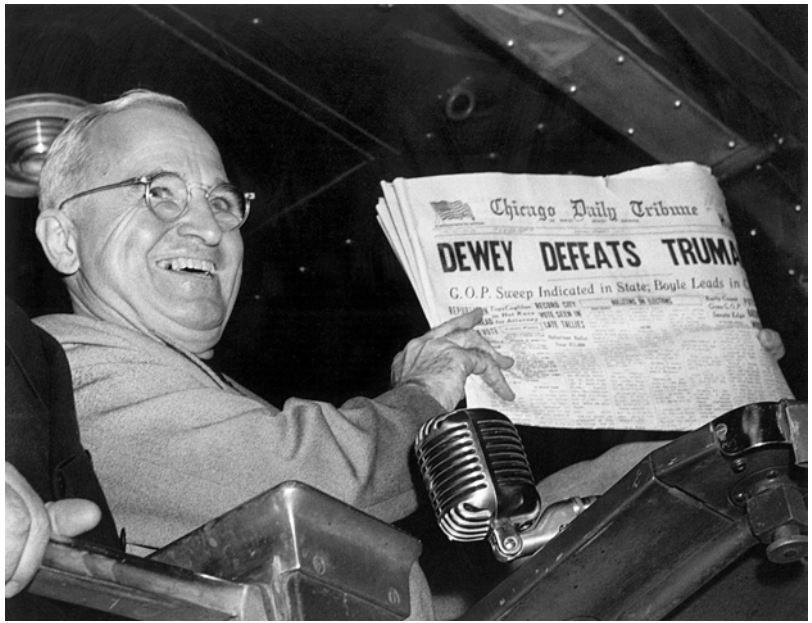
For firms offering plans in the exchanges created by the Affordable Care Act (aka Obamacare), as well as Medicaid plans, it could be bad news. That explains why Oscar Health OSCR -12.28% ▼, which derives most of its business from Obamacare marketplaces, was down 8% Wednesday morning while Centene CNC -0.06% ▼, a big Medicaid operator, was down 5%. But for businesses operating in Medicare Advantage, the privately run system that mainly serves seniors, a Republican victory is expected to provide major regulatory benefits. Under the Biden administration, insurers focused on Medicare Advantage have faced increased scrutiny from the federal government, such as lower annual rate increases. This came at a time when seniors drove up usage of the plans, leading to lower profitability.

Speculation about the impact of election

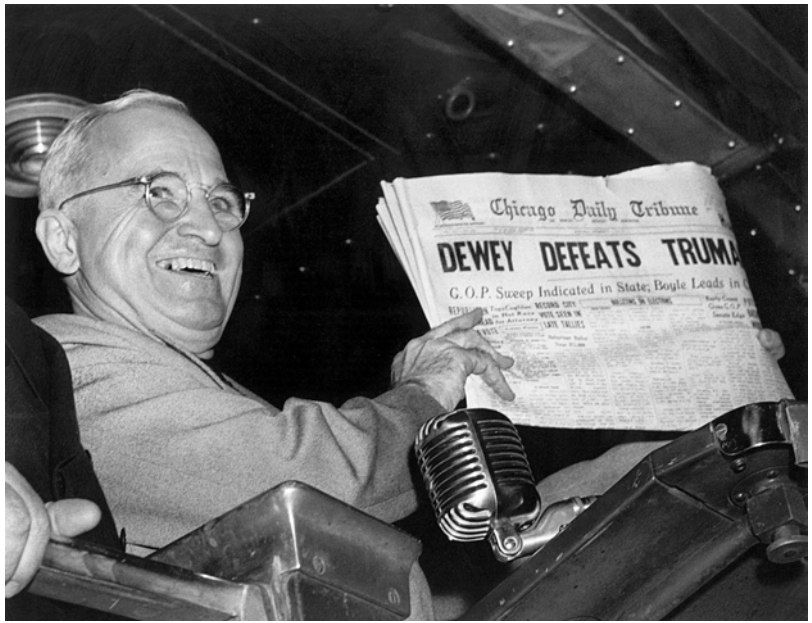




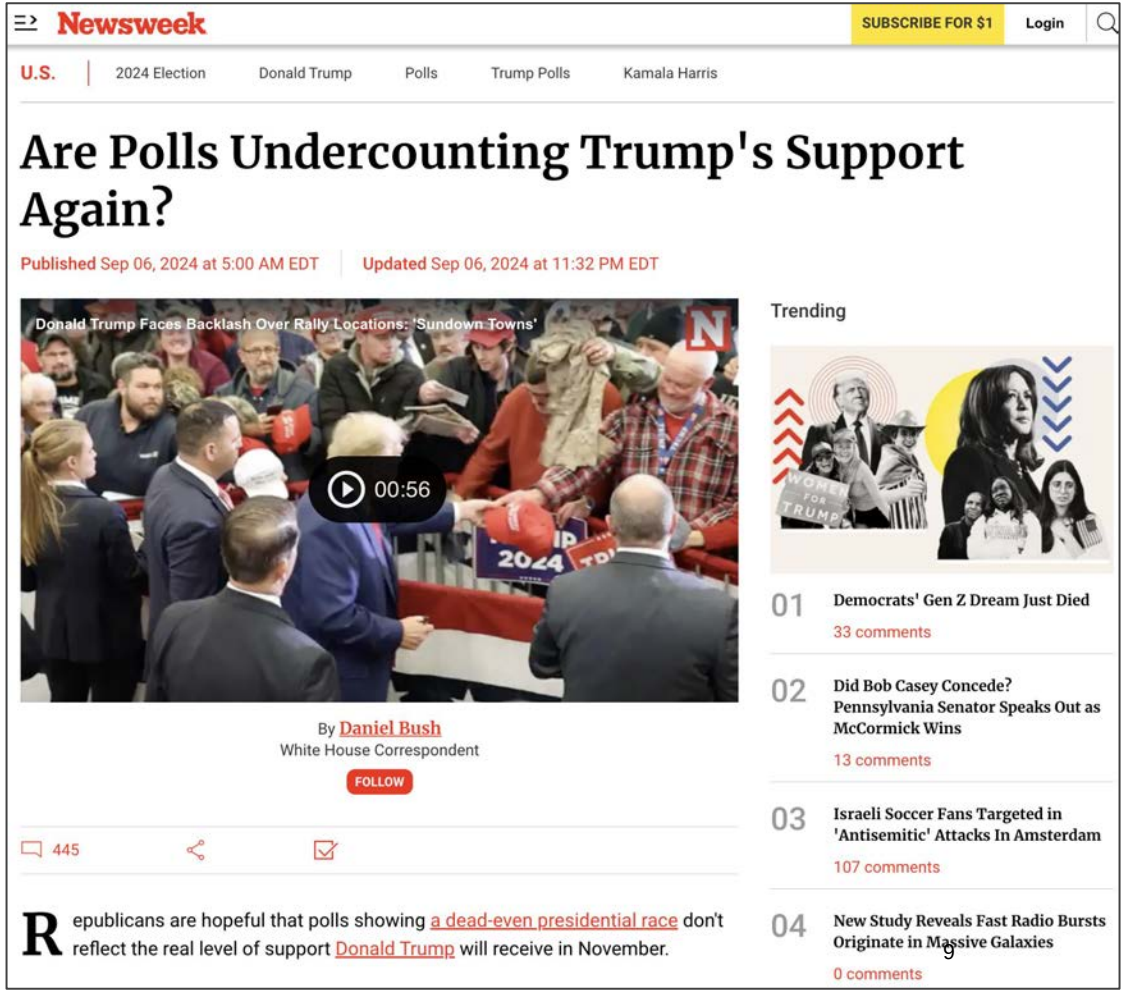
Does anyone know this photo?



Does anyone know this photo?



Systematic undercounting?



Newsweek

SUBSCRIBE FOR $1 Login

U.S. | 2024 Election Donald Trump Polls Trump Polls Kamala Harris

Are Polls Undercounting Trump's Support Again?

Published Sep 06, 2024 at 5:00 AM EDT | Updated Sep 06, 2024 at 11:32 PM EDT

Donald Trump Faces Backlash Over Rally Locations: 'Sundown Towns'

00:56

By Daniel Bush
White House Correspondent

FOLLOW

445

Republicans are hopeful that polls showing a dead-even presidential race don't reflect the real level of support Donald Trump will receive in November.

Trending

01 Democrats' Gen Z Dream Just Died
33 comments

02 Did Bob Casey Concede? Pennsylvania Senator Speaks Out as McCormick Wins
13 comments

03 Israeli Soccer Fans Targeted in 'Antisemitic' Attacks In Amsterdam
107 comments

04 New Study Reveals Fast Radio Bursts Originate in Massive Galaxies
0 comments

Today’s agenda

- Heather Mattie
- Ned McCague

An KDNuggets article on infrastructure from 2023

"To make the most out of data, organizations need efficient and scalable solutions that can store, process, and analyze data effectively. From ingesting data from multiple sources through transformation and serving, data storage underpins the data architecture.

So choosing the right data storage solution while factoring in how you'll access the data and the specific use case is important. In this article, we'll explore three popular data storage abstractions: data warehouses, data lakes, and data marts.

We'll go over the basics and compare these data storage abstractions across features like access patterns, schema, data governance, use cases, and more."

Source: https://www.kdnuggets.com/data-warehouses-vs-data-lakes-vs-data-marts-need-help-deciding

Data Infrastructure!

Why does infrastructure matter so much?

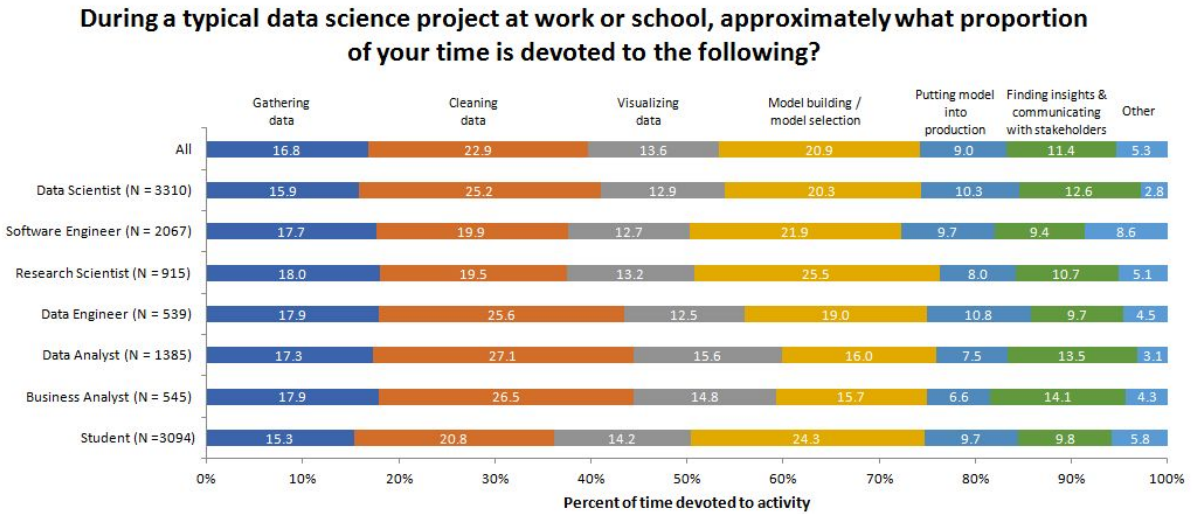
There are a few reasons:

- Most of “the job” of data professionals is **not** in the modeling
- Most of the work data professionals perform is actually working with the infrastructure or working with people
  - Data gathering
  - Data cleaning
  - Visualizing data
  - Problem definition
  - Inter-team coordination
  - Communicating insights

Simply put, you need to understand the system that generates, processes, stores,[13] and visualizes your data

This is not simply my opinion

This narrative is backed up by data

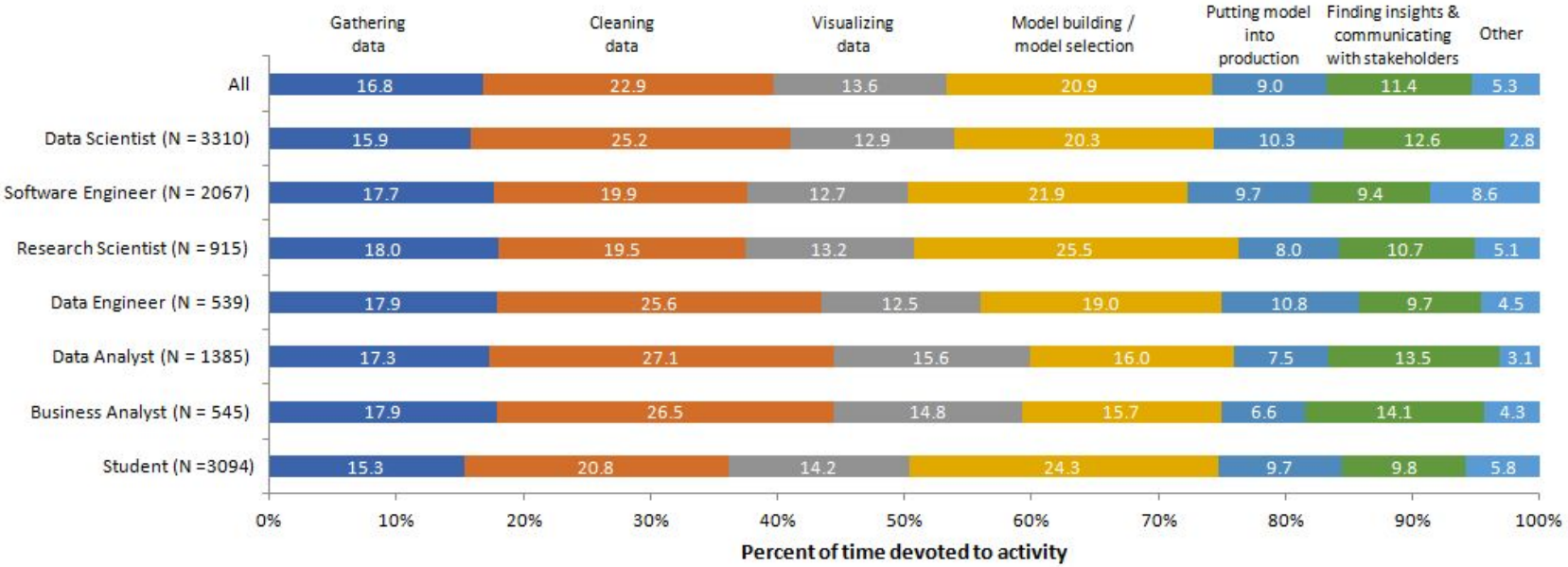


Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018.
A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 15937 respondents who provided an answer to this question. Only selected job titles are presented.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

Source: https://businessoverbroadway.com/2019/02/19/how-do-data-professionals-spend-their-time-on-data-science-projects/

During a typical data science project at work or school, approximately what proportion of your time is devoted to the following?



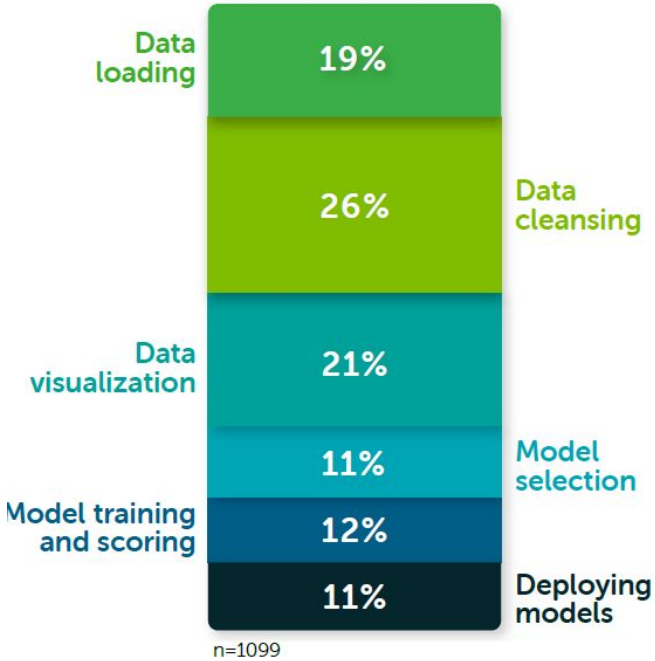
Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: http://www.kaggle.com/kaggle/kaggle-survey-2018.
A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 15937 respondents who provided an answer to this question. Only selected job titles are presented.

BUSINESS BROADWAY
DATA SCIENCE | CUSTOMER ANALYTICS | MACHINE LEARNING

Source: https://businessoverbroadway.com/2019/02/19/how-do-data-professionals-spend-their-time-on-data-science-projects/

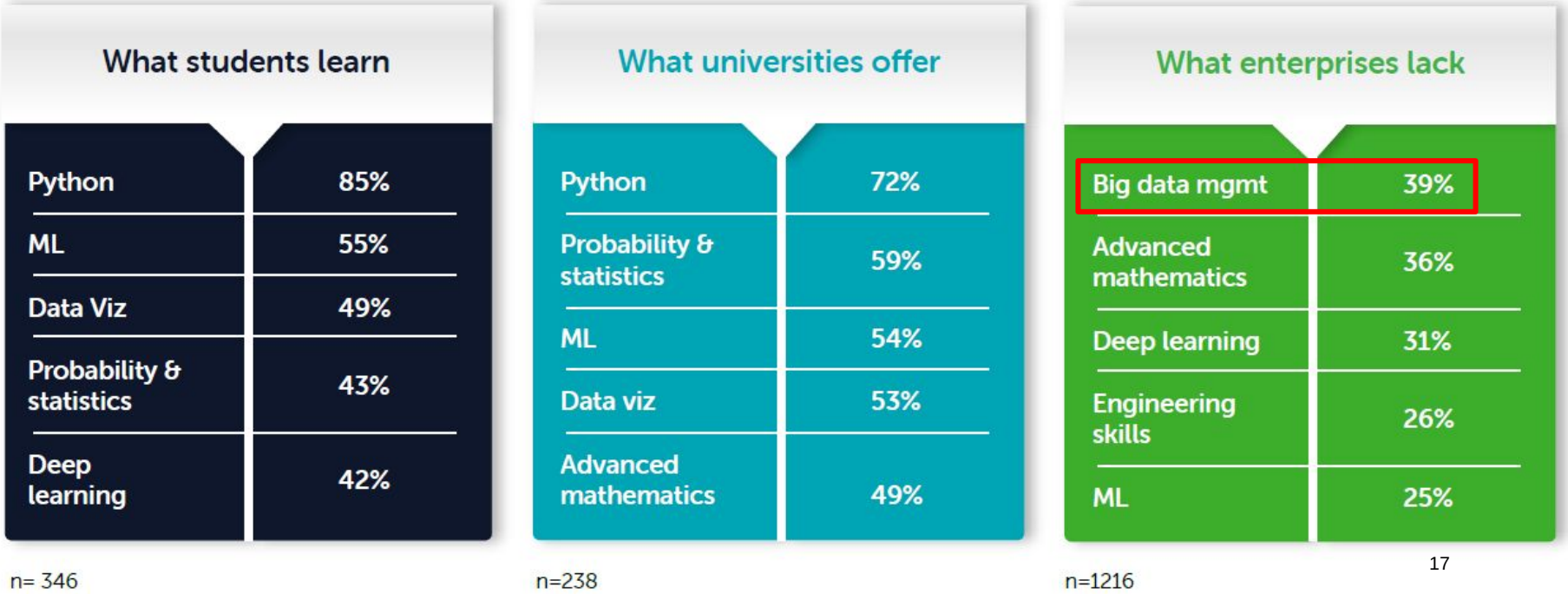
A more recent survey showed similar results

"THINKING ABOUT YOUR CURRENT JOB, HOW MUCH OF YOUR TIME IS SPENT IN EACH OF THE FOLLOWING TASKS?"



Source: https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/

This same report shows a problem with how we train data professionals



Source: https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/

What do I do when I join a new company?

What do we think?

What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs

What is data governance?

Description of Data Governance:

- Data Governance is a catch-all term that describes the process of managing, improving, monitoring, maintaining, and protecting data at an enterprise level. It includes procedures, a plan to execute those procedures, and a governance framework.

  Less formally, we can think of Data Governance as the proactive management of data to support the business achieve its strategy and vision.

What is data governance?

Considerations:

- **Data Quality**. Are there tools for data cataloging? Does that company have a data governor, a data management office, a data council, or any data stewards? Are there valid values for every field? Are there formal quality checks on the data? Do those quality checks have contain compliance and outlier checks? Is there automated alerting at various stages of the data ecosystem? Are there formal or informal relationships within the company to ensure that data creation conforms to appropriate norms and expectations? Does the company even have any norms or expectations when it comes to data creation?

What is data governance?

Description of Data Governance:

- Data Governance is a catch-all term that describes the process of managing, improving, monitoring, maintaining, and protecting data at an enterprise level. It includes procedures, a plan to execute those procedures, and a governance framework.

  Less formally, we can think of Data Governance as the proactive management of data to support the business achieve its strategy and vision.

What is data governance?

Considerations:

- **Data Access**. Who has access to what data? Is this access role-based? Is there formal training to gain access to data? Is that training overseen by a centralized team with sign offs from legal, security, and IT? What are the procedures to gain access to data? How is data access revoked? Do we give external stakeholders access to data?

What is data governance?

Considerations:

- **Regulatory Compliance**. Is the company compliant with any and all relevant regulations, including
  - Sarbanes-Oxley Act,
  - General Data Protection Regulation (GDPR),
  - California Consumer Privacy Act (CCPA),
  - The Payment Card Industry Data Security Standard (PCI-DSS), and
  - Canada's Personal Information Protection and Electronic Documents Act (PIPEDA)?
- Is there formal training on these regulations? Are there documented policies supporting and enforcing them?

What is data governance?

Considerations:

- **Data Literacy**. Data literacy ensures that the organization is educated to a sufficient level to consume data with confidence. Are there formal trainings on the data in the company? Are there major KPIs that are owned, managed, and reported through a centralized team to key stakeholders? Is there alignment within the organization around which metrics matter and how accurate the numbers are? Do the key leaders in the organization agree on the major KPIs of the company?

What is data governance?

Considerations:

- **Policies, Procedures, Programs.** The 3Ps describe the formal rules by which requests are made, access is granted, and use cases are defined. They provide the organizational red tape that prevents mishandling of data, governs access to data, and ensures that there is visibility into data lineage and quality.
- If Data Governance is strategic in scope, then the Policies, Procedures, and Programs are operational and tactical in depth. The two go hand-in-glove.

What is data governance?

Considerations:

- **Extra.** There are other topics, both formal and informal, that should be considered, including Data Stewardship and Ownership, Data Policies and Guidelines, Data Standards for Presentation of Data, Metadata Management, Data Lineage, Data Cataloging, Data Quality, and Data Security.

What I do when I join a new company

A few things:

- Conduct an audit of the internal **Data Governance**, including Policies, Procedures, and Programs

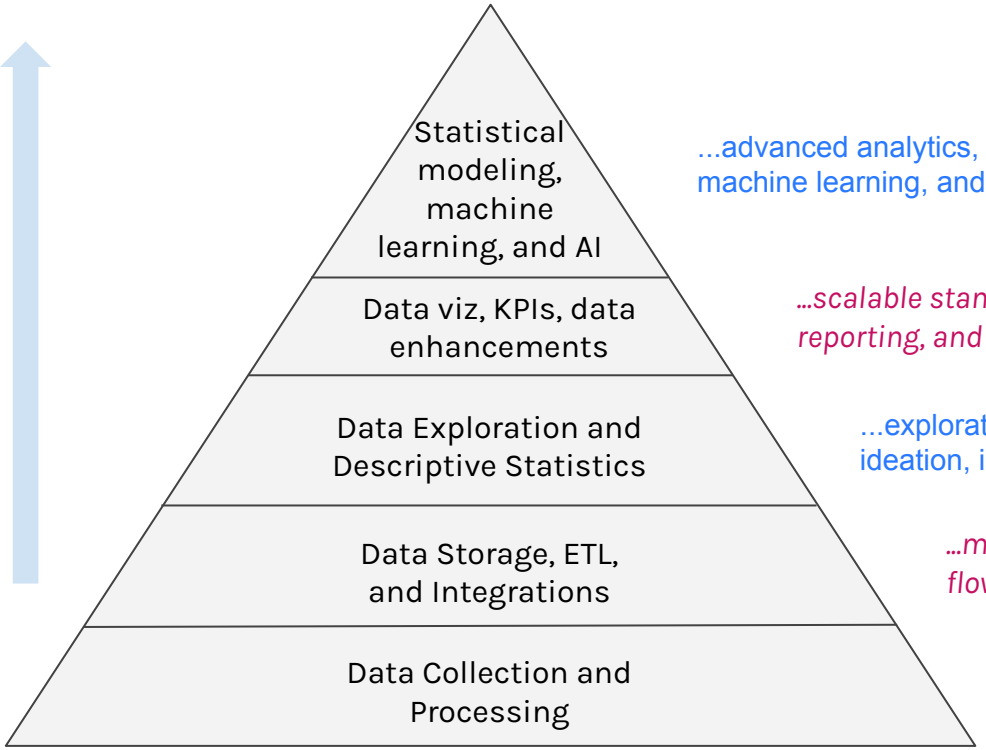
What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs
- Review the **data sufficiency pyramid** with key leaders within the company

Data Sufficiency Pyramid

Scalable, robust reporting depends on a strong foundation of data collection, processing, storing, and management



...advanced analytics, regression analysis, machine learning, and **strategic analyses**

...scalable standardized in-app reporting , longitudinal reporting, and data enhancements (i.e. back into the DW)

...exploratory data analysis, metric exploration and ideation, internal vetting, and internal alignment

...multisource data connections, reliable data flows, relational data modeling, and MDM

...logging, instrumentation, data acquisition, and simple storage

What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs
- Review the data sufficiency pyramid with key leaders within the company
- Conduct an **Analytics Maturity Assessment** on the company

Analytics Maturity Assessment

Level 0

–Struggling to get basic information
–Chaotic/Inconsistent
–No analytic structure
–Severely limited analytic capability
–Reporting is difficult and dubious

Level 1

–Concerned with current fires and issues
–Descriptive in nature
–Historically focused
–No integration
–Print outs
–Reactive reporting

Level 2

–Operational reports
–Descriptive reporting in a timely manner
–Manual integration
–Manual processes
–Focus on performance evaluation

Level 3

–Concerned with current plans
–Diagnostic reporting
–Concern about what is happening and why
–Mature data warehousing exists
–Uneven competency across the organization

Level 4

–Insight into what is likely to happen
–Organizational dashboards/scorecards
–Managed and measured
–Widespread analytic capabilities
–Consistent production

Level 5

–Integrated into business planning and decision making
–Shapes actions and perceptions
–Analytics is a differentiator
–Real-time analysis

What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs
- Review the data sufficiency pyramid with key leaders within the company
- Conduct an Analytics Maturity Assessment on the company
- Create a diagram of the **data sources** in the company (i.e. a data ecosystem map or data architecture diagram)

Understanding the data sources at a company

Types of data you'll see:

- 3rd party vendors
- Product usage
- Customer files
- Manual data uploads

Understanding the data sources at a company

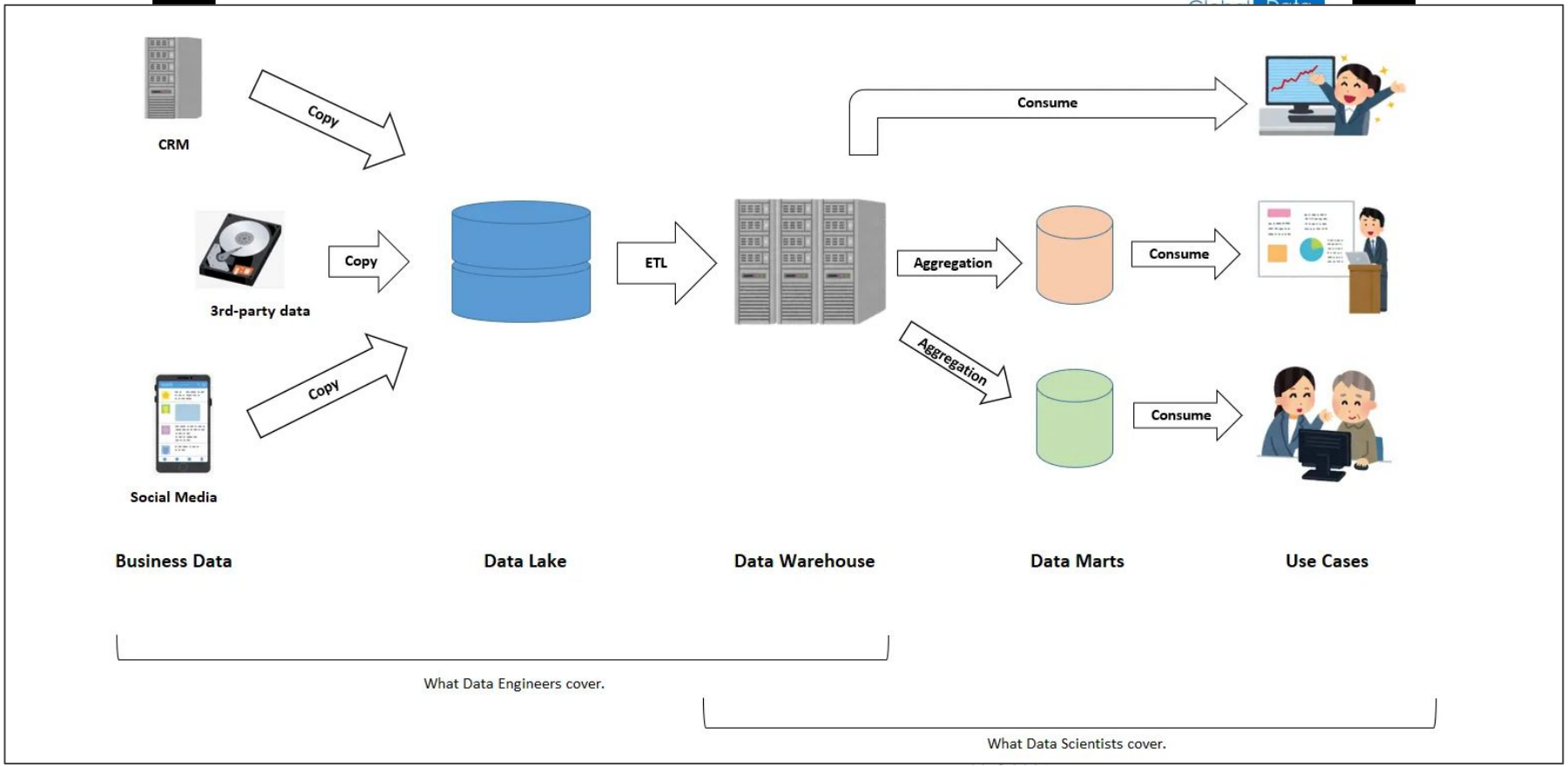
Types of data you'll see:

- 3rd party vendors
- Product usage
- Customer files
- Manual data uploads

Questions to consider:

- What do those data sources feed into? And then what happens?
- Where does the data go? How does it get there?
- What is the full data ecosystem that you're walking with?
  - What tools do you have?
  - What teams do you have?
  - Who leads those teams?
  - Who is a peer-leader you can learn from?

The point is to enable you to have deeper level conversations with key stakeholders



The point is to enable you to have deeper level conversations with key stakeholders

Source: (link)

What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs
- Review the data sufficiency pyramid with key leaders within the company
- Conduct an Analytics Maturity Assessment on the company
- Create a diagram of the **data sources** in the company (i.e. a data ecosystem map or data architecture diagram)

What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs
- Review the data sufficiency pyramid with key leaders within the company
- Conduct an Analytics Maturity Assessment on the company
- Create a diagram of the data sources in the company (i.e. a data ecosystem map or data architecture diagram)

Broad-Stroke Goals:

- Get a lay of the land
- Formalize things that are often informal
- Communicate expertise and competence
- Identify gaps
- Build relationships

What I do when I join a new company

A few things:

- Conduct an audit of the internal Data Governance, including Policies, Procedures, and Programs
- Review the data sufficiency pyramid with key leaders within the company
- Conduct an Analytics Maturity Assessment on the company
- Create a diagram of the data sources in the company (i.e. a data ecosystem map or data architecture diagram)

Broad-Stroke Goals:

- Get a lay of the land
- Formalize things that are often informal
- Communicate expertise and competence
- Identify gaps
- Build relationships
- Create inputs for roadmap
- Identify resourcing needs
- Clarify organizational structure
- Identify adversaries in the org
- Create a strategic framework

Case Example

McKinsey & Company

Operations Practice

Toward smart production: Machine intelligence in business operations

A detailed study of machine intelligence in industrial and manufacturing operations reveals the surprisingly different paths companies can take. But a group of leaders shares similar characteristics.

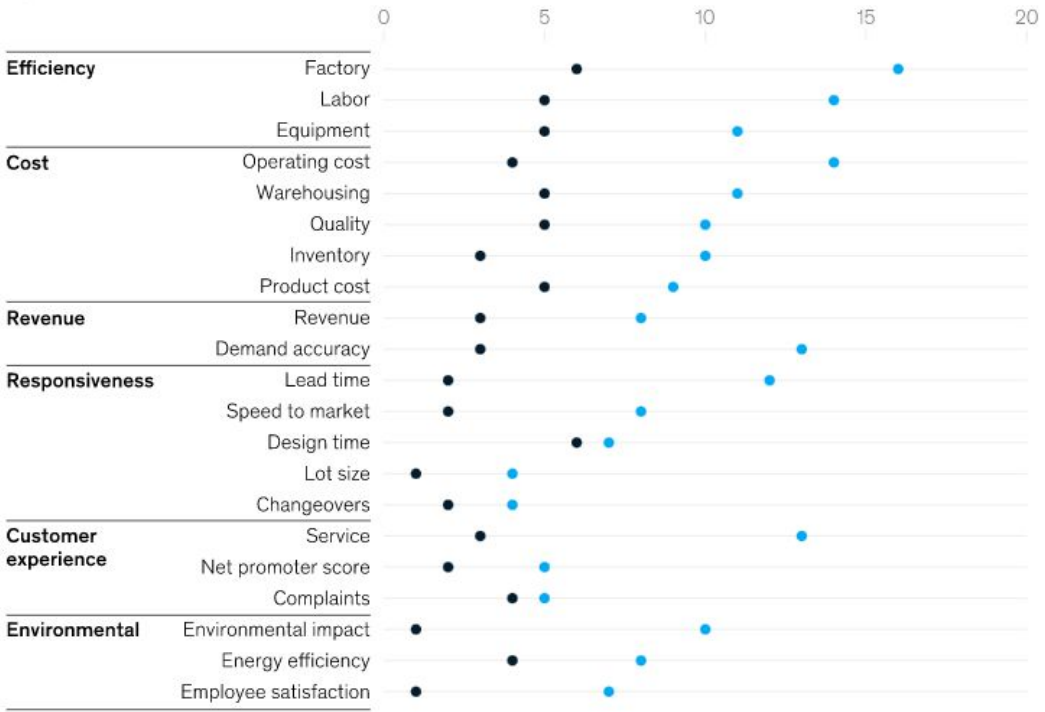
This article is a collaborative effort by Duane S. Boning, Vijay D'Silva, Pete Kimball, Bruce Lawler, Retsef Levi, and Ingrid Millan, representing views from McKinsey's Operations Practice and the Massachusetts Institute of Technology's Machine Intelligence for Manufacturing and Operations program.

Source: https://www.mckinsey.com/capabilities/operations/our-insights/toward-smart-production-machine-intelligence-in-business-operations

Case Ex

Across a broad range of metrics, machine-intelligence leaders achieve triple the improvement of other companies.

Average improvement through machine intelligence, by KPI, %

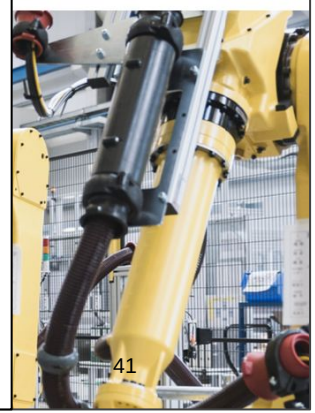


Note: All variables are normalized to a 0 to 1 scale.
Source: MIT Machine Intelligence for Manufacturing and Operations (MIMO) program; McKinsey Machine Intelligence Survey

duction:
ce in
s

and manufacturing
mpanies can take.

mball, Bruce Lawler, Retsef Levi,
nd the Massachusetts Institute of
am.

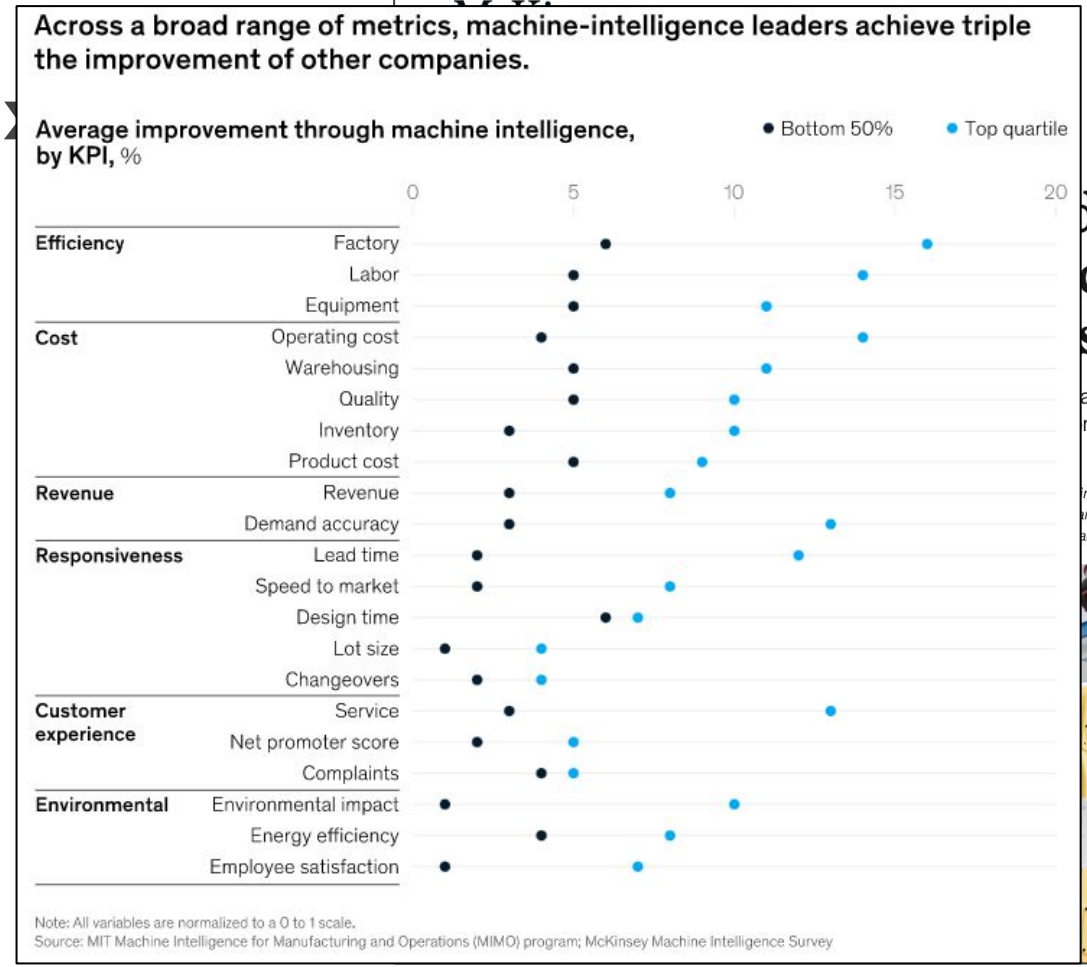


Source: https://www.mckinsey.com/capabilities/operations/our-insights/toward-smart-production-machine-intelligence-in-business-operations

Case Ex

An interesting quote:

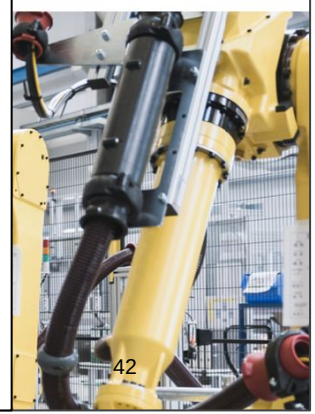
- The key differentiator between top performers and mediocre performers was the difference between their “[l]evel of investment in infrastructure to support MI implementation”



duction:
ce in
s

and manufacturing
mpanies can take.

mball, Bruce Lawler, Retsef Levi,
nd the Massachusetts Institute of
am.



Source: https://www.mckinsey.com/capabilities/operations/our-insights/toward-smart-production-machine-intelligence-in-business-operations

So let’s talk about infrastructure!

Activating data in an organization

Mature organizations need to be able to...

- Create data
- Collect data
- Process data
- Activate data

Activating data in an organization

Mature organizations need to be able to...

- Create data
- Collect data
- Process data
- Activate data

You need data tools that allow you do this at scale

Activating data in an organization

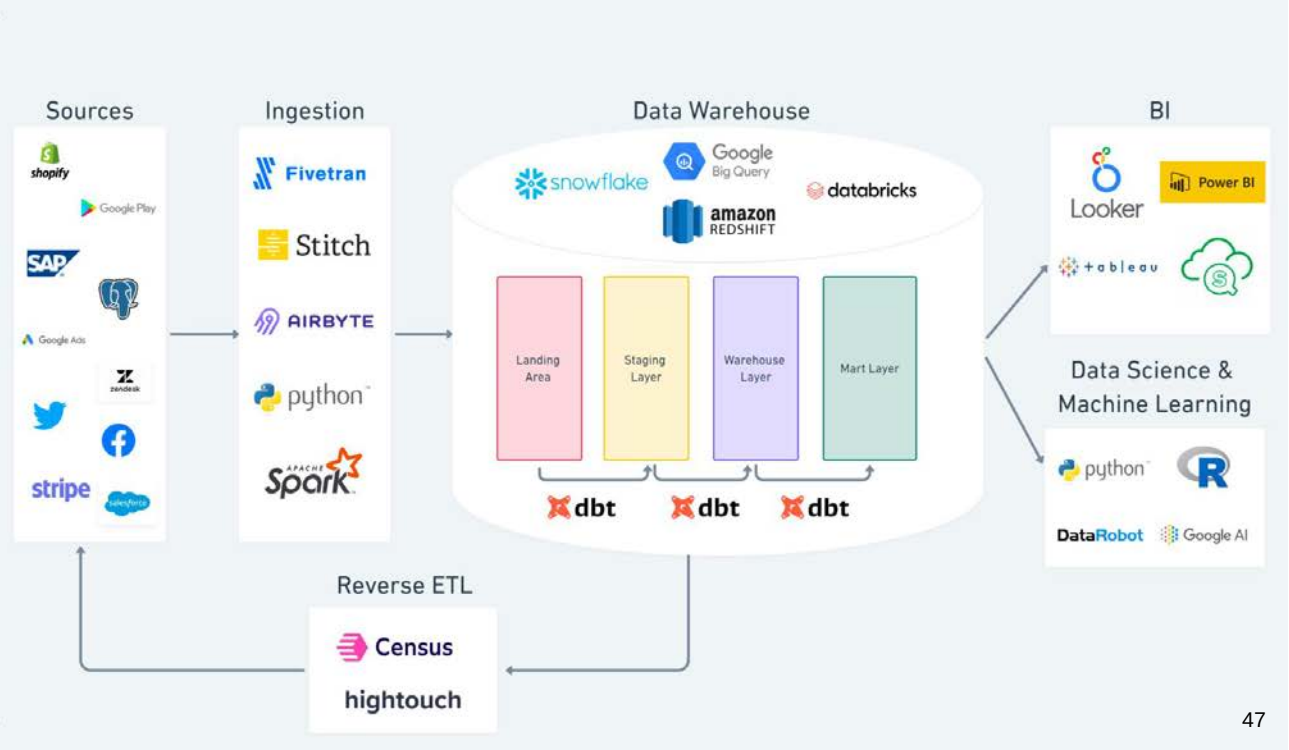
Mature organizations need to be able to...

- Create data
- Collect data
- Process data
- Activate data

This means you need tools to help you extract, transform, load, store, analyze, visualize, and monitor your data

You need data tools that allow you do this at scale

Modern Data Stack

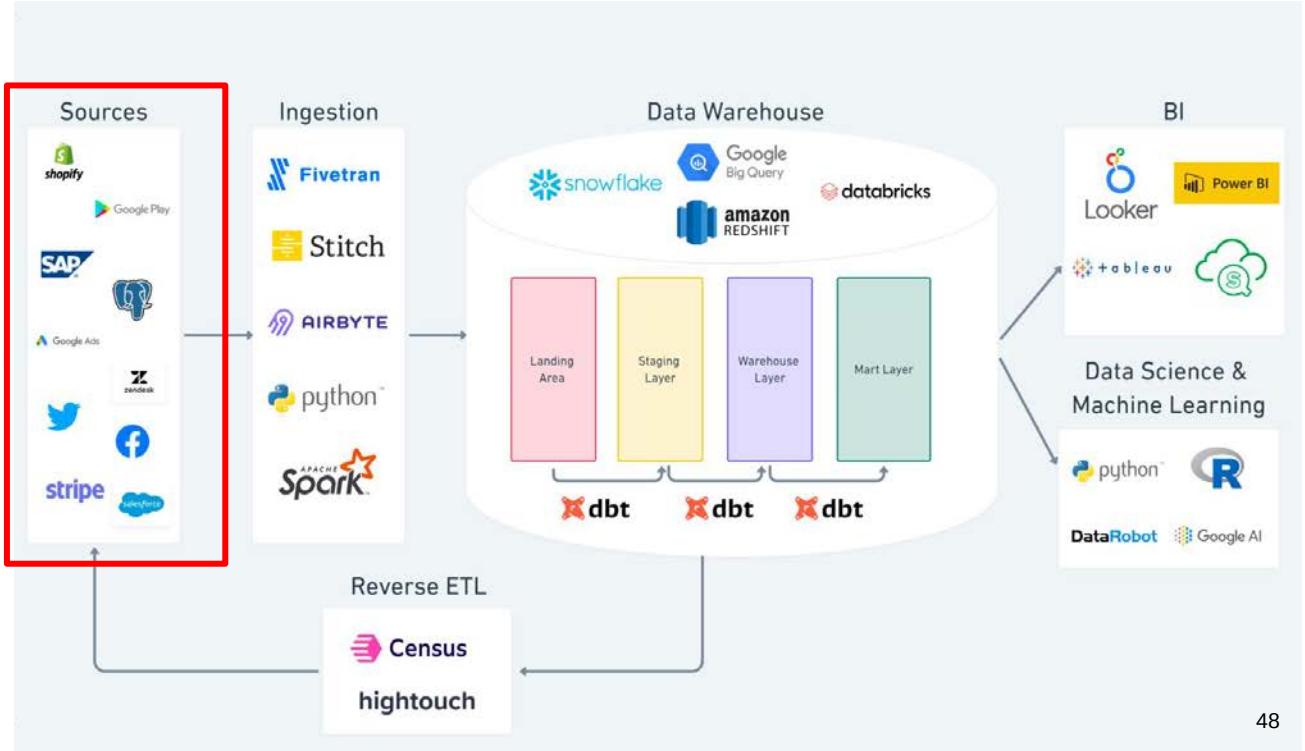


Source: (licnk)

Modern Data Stack

Data Sources

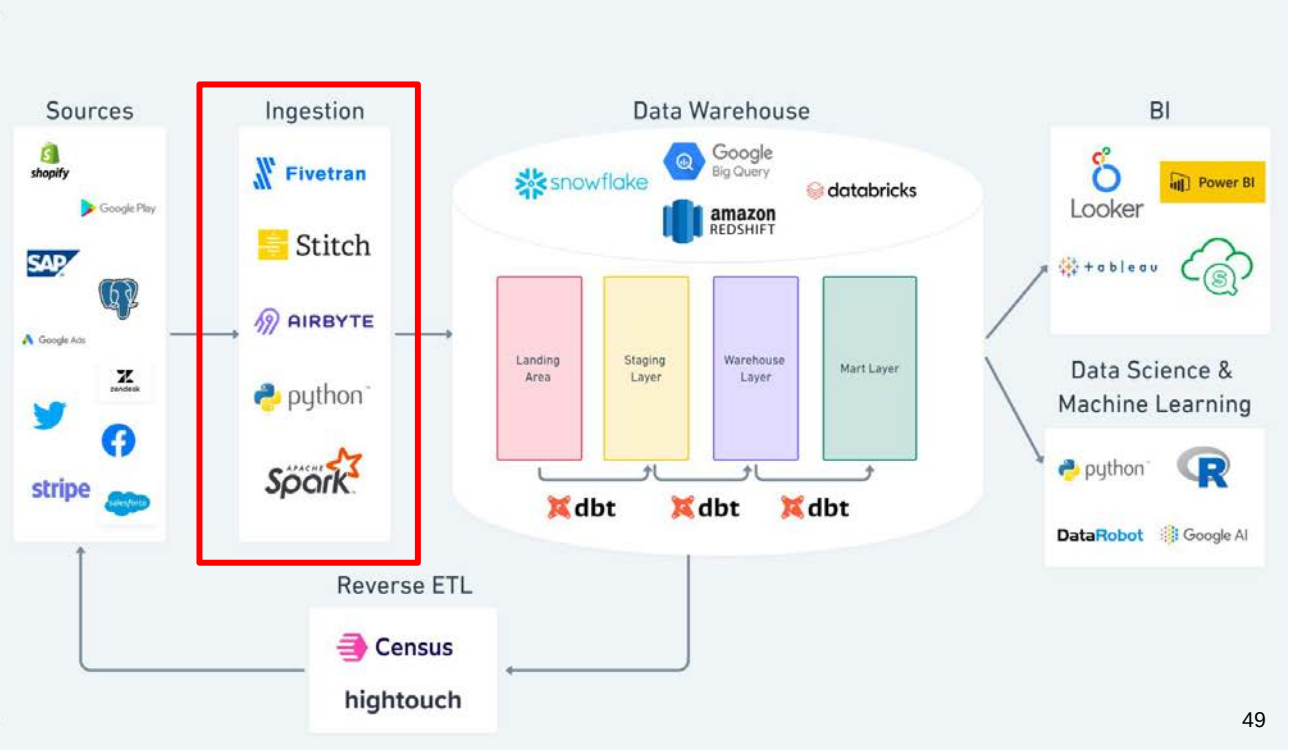
These represent the different data stores that your business relies on. These can include:
3rd party vendors,
Product usage,
Customer files,
Manual data uploads



Modern Data Stack

Ingestion

You need some type of tool to get the data out of the source systems. These tools allow you to extract the data you need.

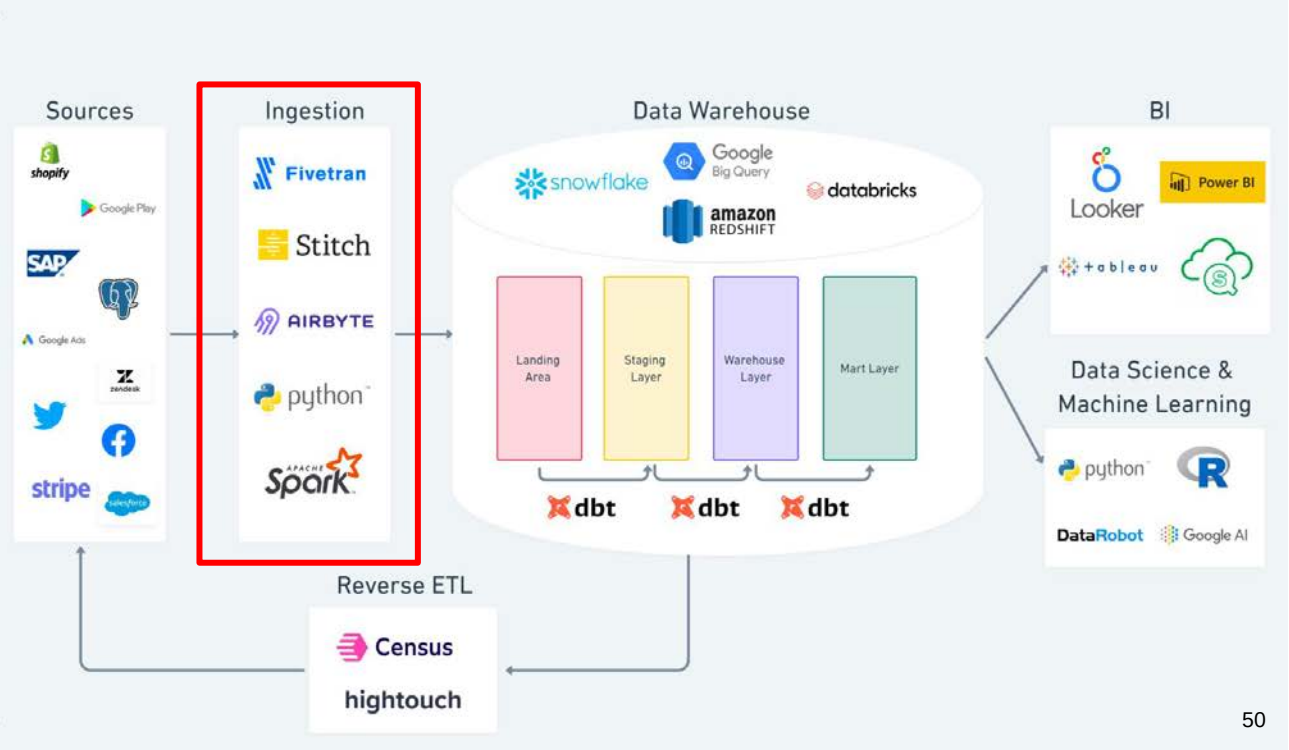


Modern Data Stack

Ingestion

You need some type of tool to get the data out of the source systems. These tools allow you to extract the data you need.

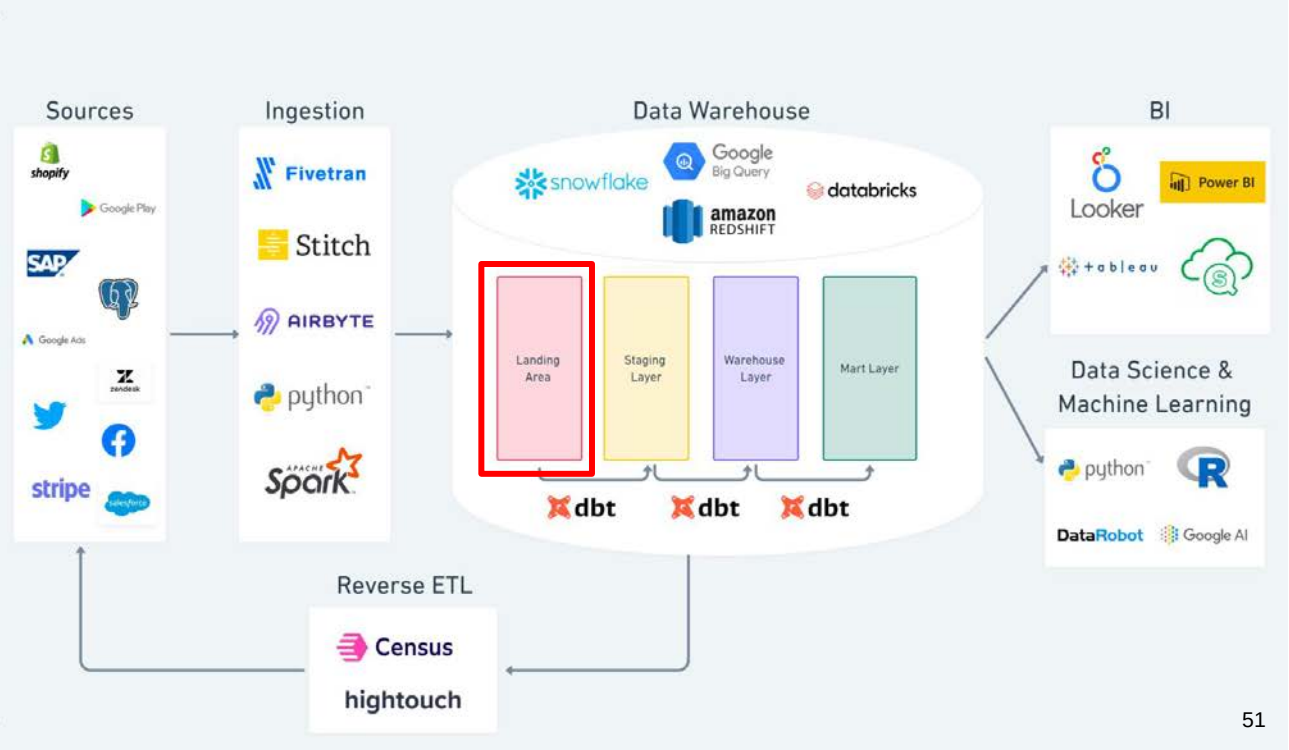
Example with Fivetran® and Google Ads™ (link; link2)



Modern Data Stack

Data Lake

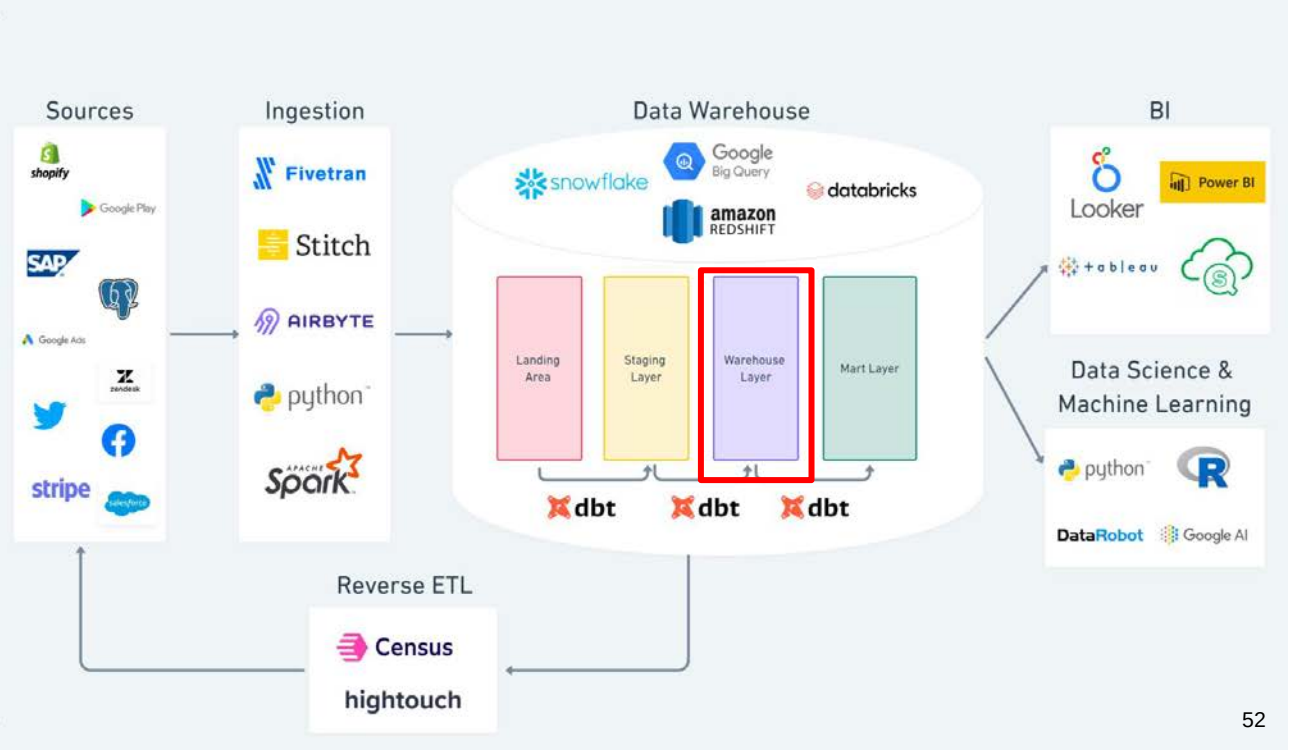
Conceptually, the data lake is a central repository that accepts data of all types (i.e. relational, structured, semi-structured, and non-structured) in a low-to-no modeling framework



Modern Data Stack

Data Warehouse

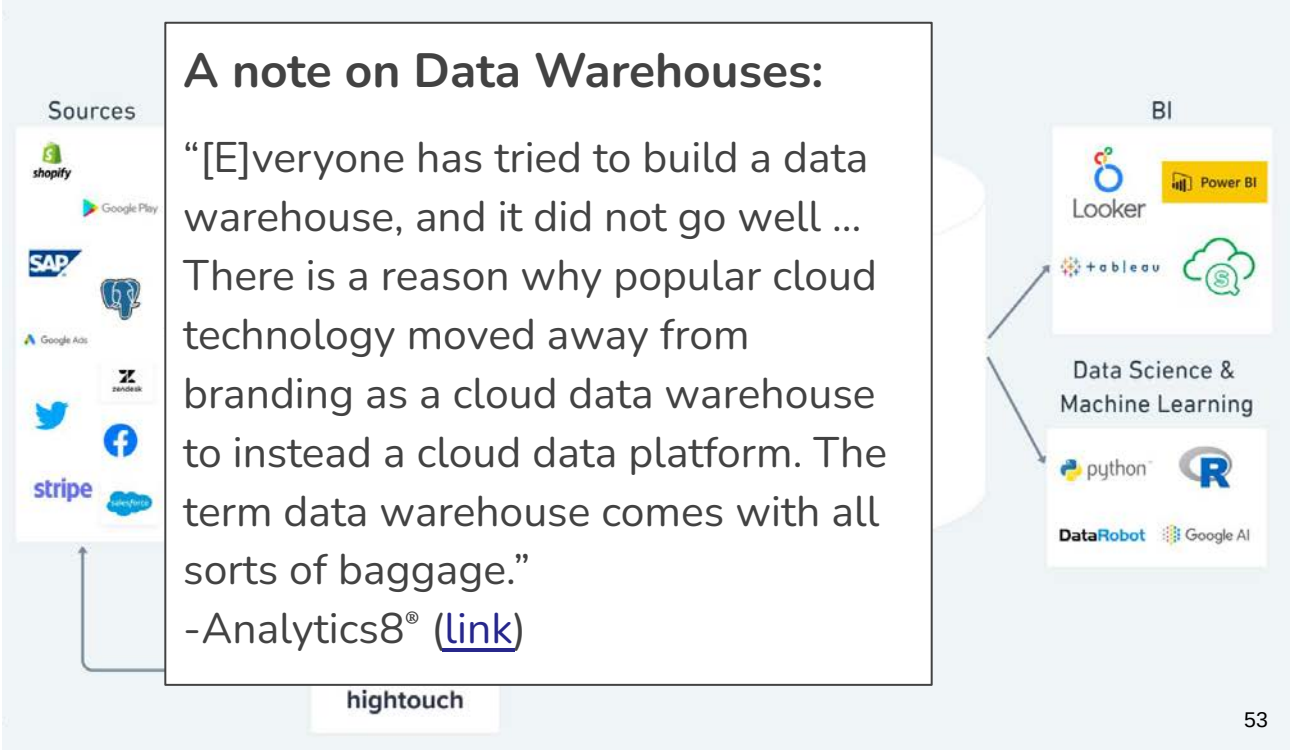
This is a centralized repository that stores clean, integrated, and structured data from one or multiple sources – it's designed for querying, analysis, and reporting



Modern Data Stack

Data Warehouse

This is a centralized repository that stores clean, integrated, and structured data from one or multiple sources – it's designed for querying, analysis, and reporting



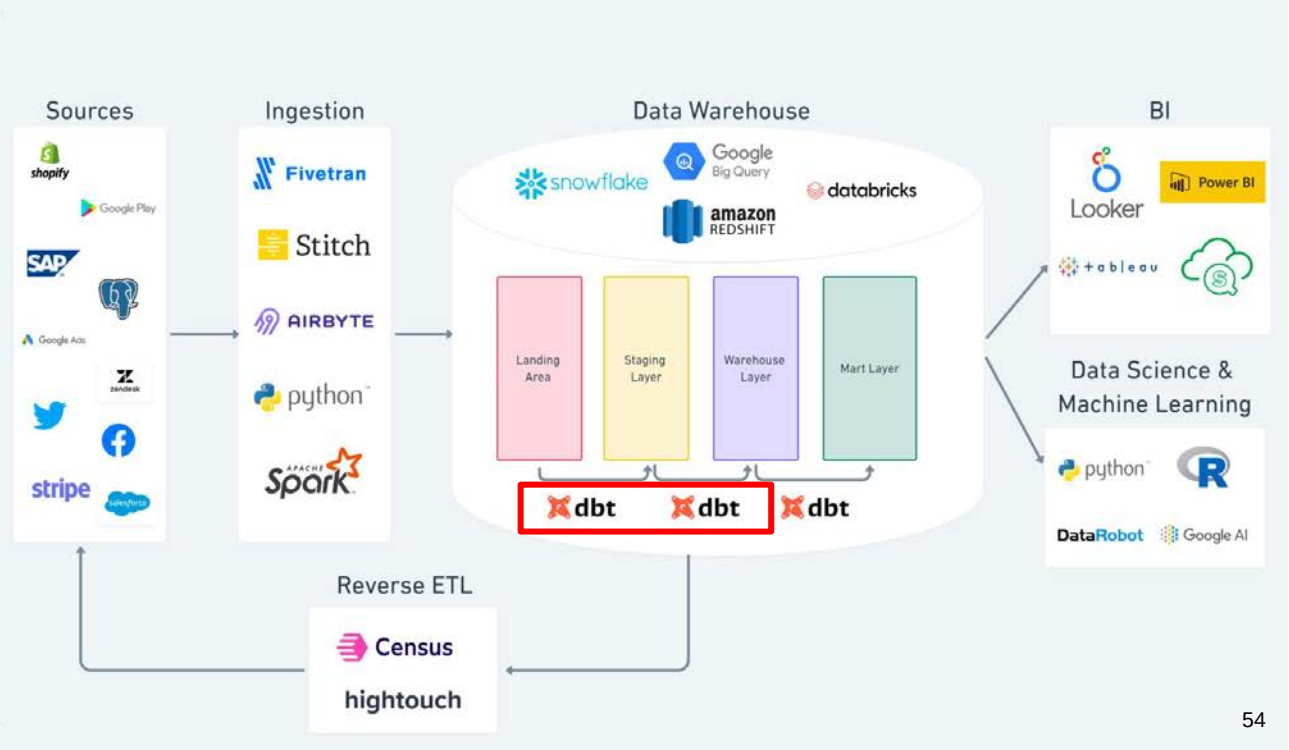
A note on Data Warehouses:

"[E]veryone has tried to build a data warehouse, and it did not go well … There is a reason why popular cloud technology moved away from branding as a cloud data warehouse to instead a cloud data platform. The term data warehouse comes with all sorts of baggage."
-Analytics8® (link)

Modern Data Stack

Transformation Tool

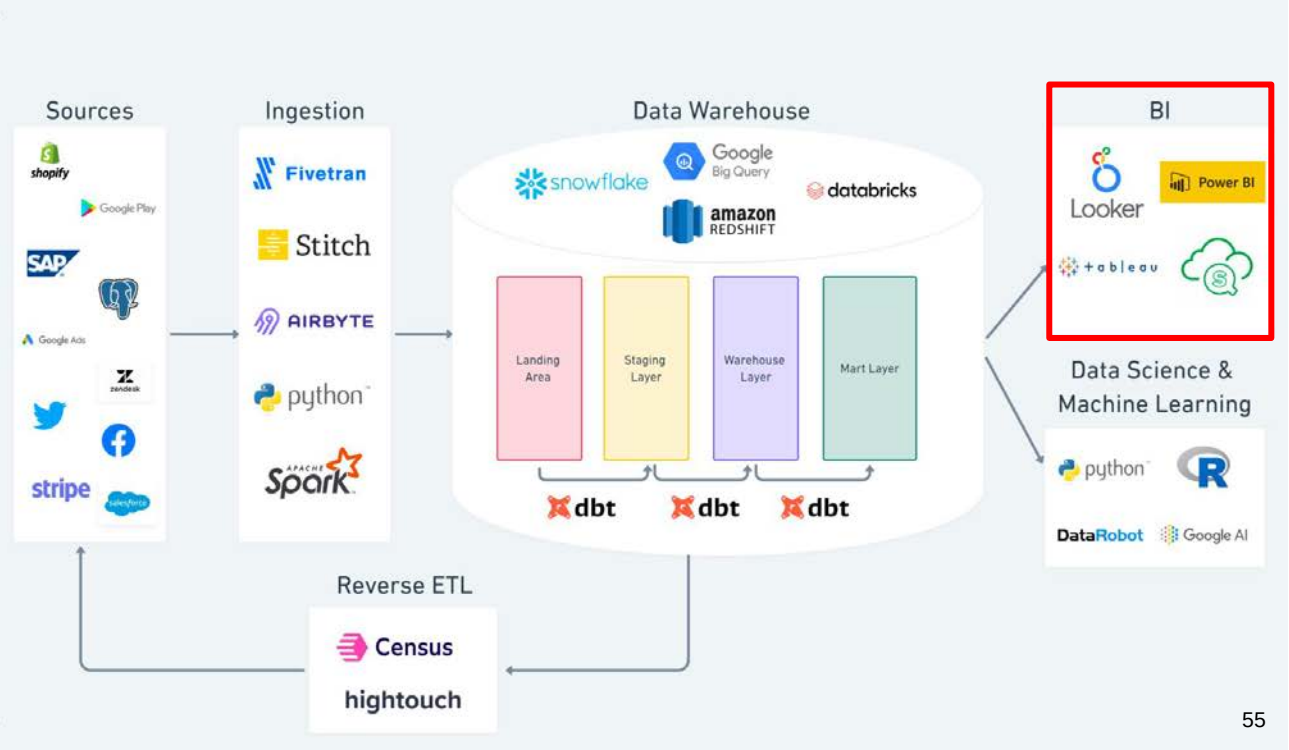
You need something that can turn your raw data into structured data. For many, dbt is the best tool on the market to manage these transformations.



Modern Data Stack

Business Intelligence Tool

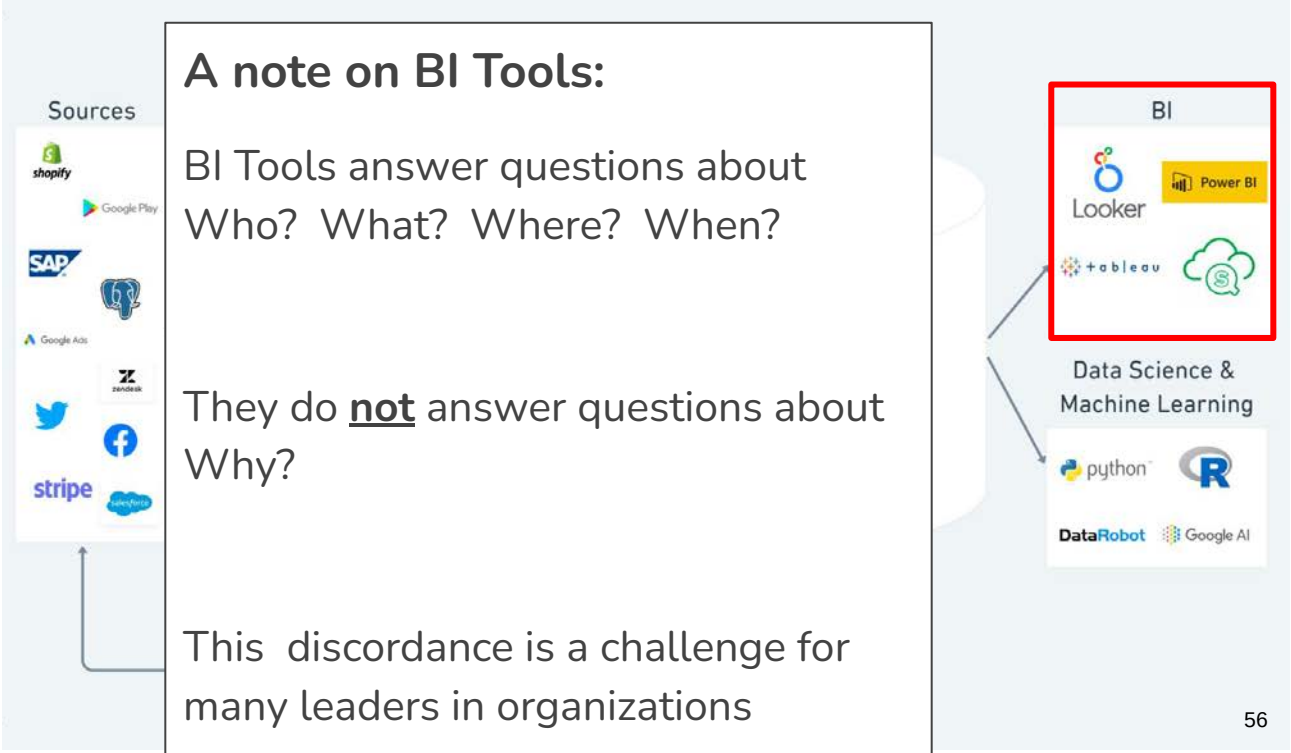
You likely need a tool to create scalable analytics (read: dashboards) for the organization. Do you need to give customers access to the data as well?



Modern Data Stack

Business Intelligence Tool

You likely need a tool to create scalable analytics (read: dashboards) for the organization. Do you need to give customers access to the data as well?



A note on BI Tools:

BI Tools answer questions about Who? What? Where? When?

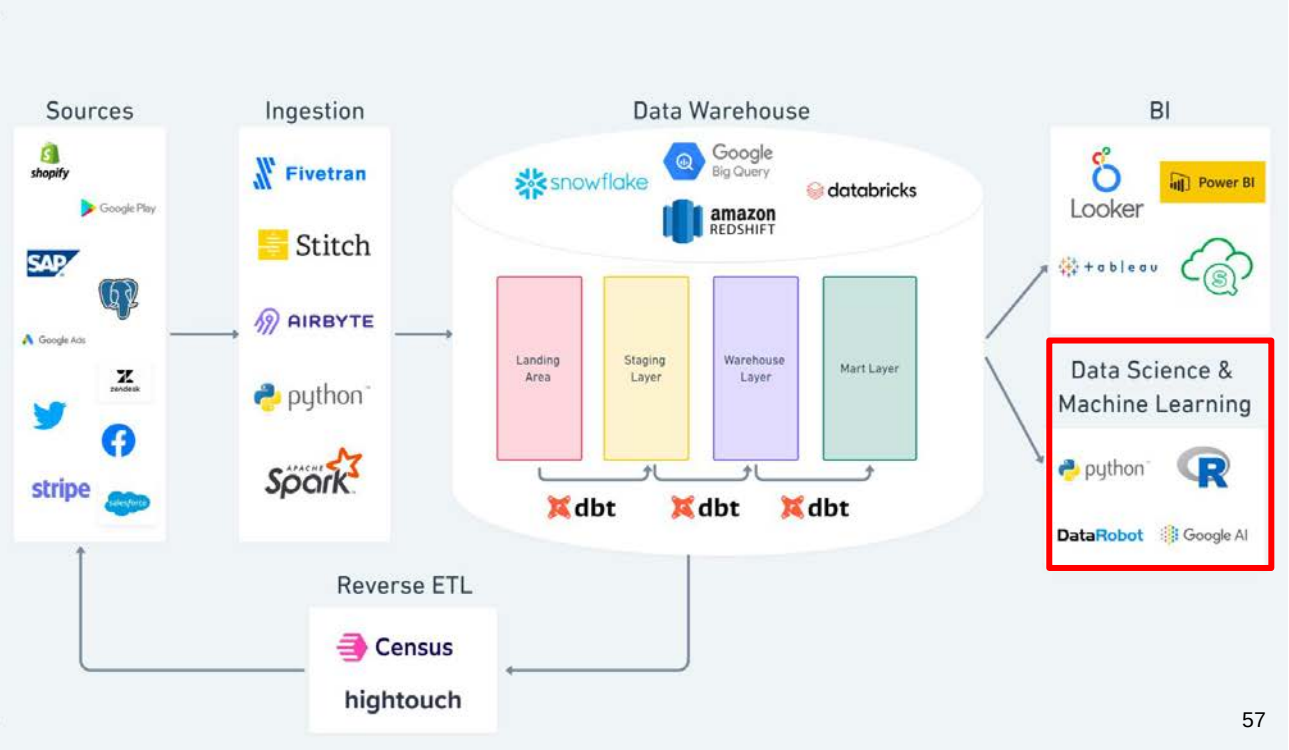
They do **not** answer questions about Why?

This discordance is a challenge for many leaders in organizations

Modern Data Stack

DS & ML Tools

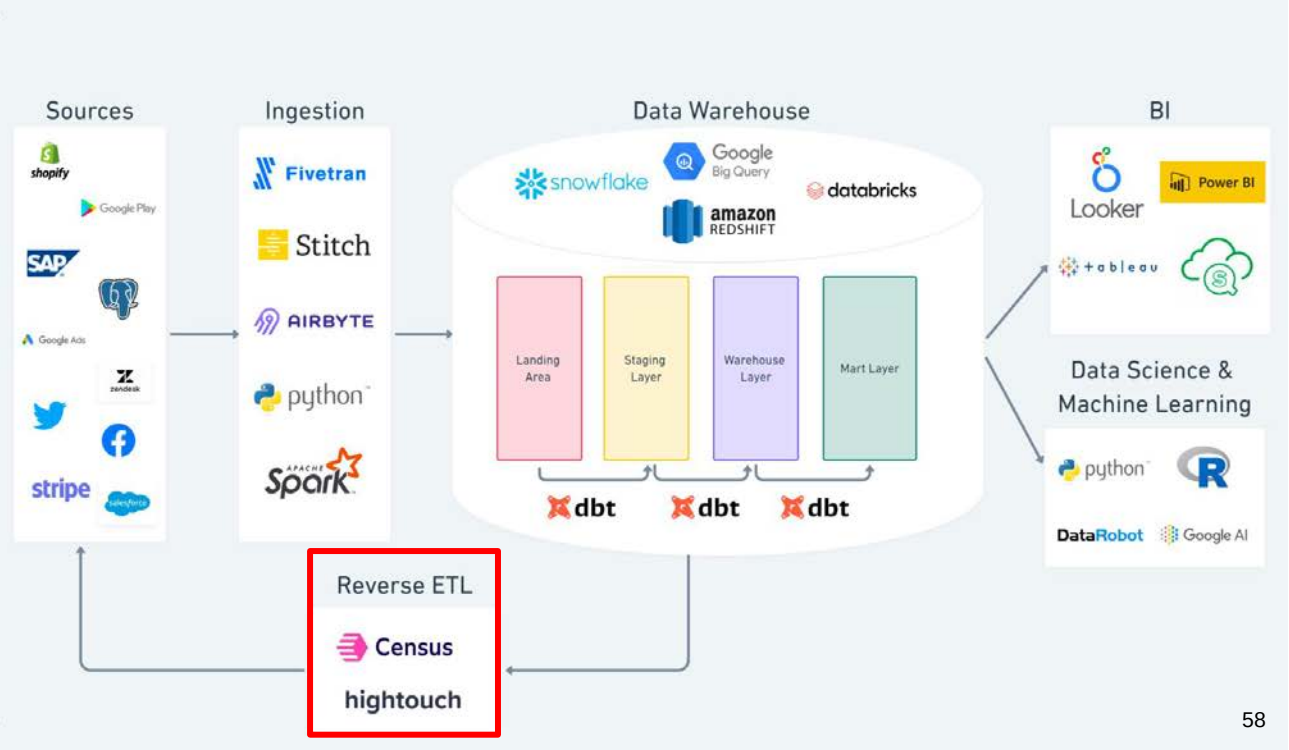
You need a tool to perform exploratory data analysis (EDA). Python®, R™, and SAS® are key players in the space, but there are others.



Modern Data Stack

Reverse ETL

This occurs when you feed clean, processed data back into your source applications – for example, Customer LTV into Salesforce® or Lead Scoring into Marketo®

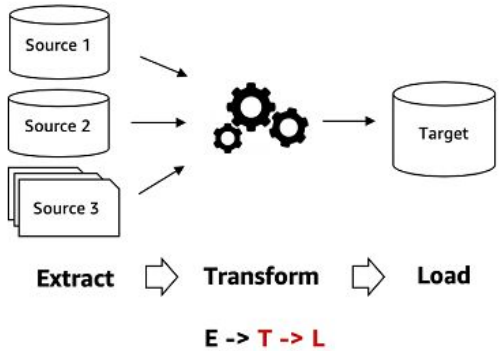


Practical Details

ETL vs ELT

ETL

- Extract, Transform, and Load
- Emerged in the 1970s
- Transforms data on a separate processing server
- Slower, less accommodating with real-time data



ELT

- Extract, Load, and Transform
- Emerged in the 2000s
- Sends raw data into the RDMS and transforms data **within** the data warehouse itself
- Faster and easier for GDPR compliance

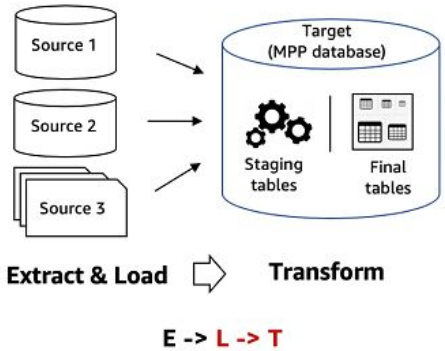


Image source: (link)

Types of databases

OLTP (Online Transaction Processing)

- OLTP databases are designed to perform very well on row-based queries with single-row operations
- For example, many online applications (banking, e-commerce, booking, etc.) operate with OLTP database (e.g. Postgres, MySQL, etc.)
- These databases are very fast in query retrieval, especially if combined with indexes to prevent full-table scans
- These databases struggle with complex queries and large data volumes (i.e. rows of data)
- Often have normalized data schemas

Types of databases

OLAP (Online Analytical Processing)

- OLAP databases are designed to handle complex data analysis queries with many rows of data
- For example, many analytic database (Snowflake, BigQuery, etc.) are used by data professionals in many organizations
- These databases are columnar, meaning that they are designed to store, manage, and retrieve data by columns, rather than by rows
- Often have denormalized data schemas

Types of databases

Time Series Database

- Time Series Databases are a very niche type of database that are used almost exclusively with time-series data
- They become valuable in situations where you have huge volumes of data, skinny tables (i.e. many rows, but few columns), and time-dependent needs
- Reasonable use-cases include application server monitoring, fintech, sensor data, some IoT situations, longitudinal outlier detection, and self-driving cars

Organizing Data:
Star Schema vs Snowflake Schema

Star Schema Design

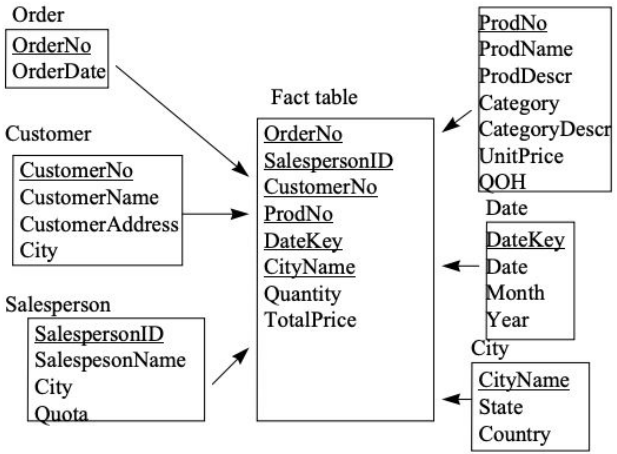


Figure 3. A Star Schema.

Snowflake Schema Design

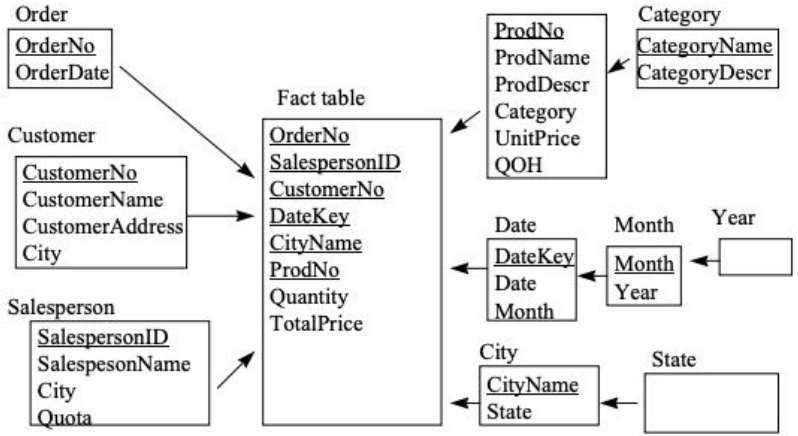


Figure 4. A Snowflake Schema.

Source: https://dl.acm.org/doi/pdf/10.1145/248603.248616

Team Organization

How do you structure a data team?

Obviously, the answer is “it depends”. But what are the variables to consider?

How do you structure a data team?

Obviously, the answer is “it depends”. But what are the variables to consider?

- Stage of the organization
- Size of the organization
- International presence
- Industry
- Risk tolerance of leadership
- Competing analytics teams

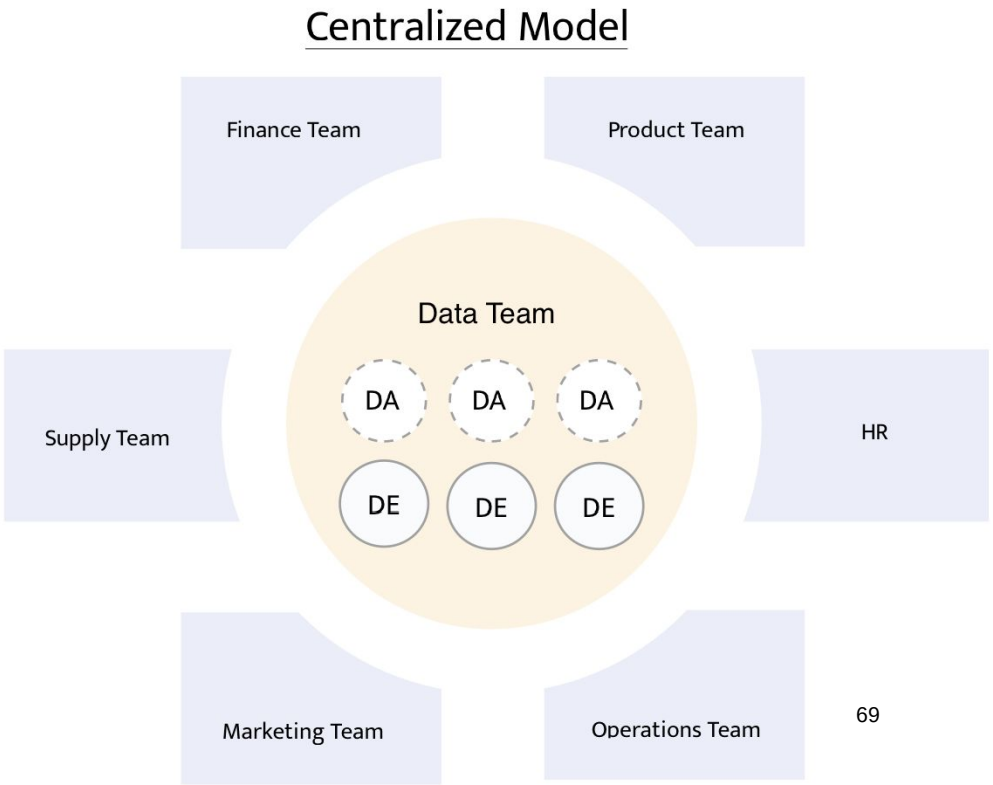
How do you structure a data team?

Obviously, the answer is “it depends”. But what are the variables to consider?

- Stage of the organization
- Size of the organization
- International presence
- Industry
- Risk tolerance of leadership
- Competing analytics teams
- Maturity of the data tech stack
- Budgetary restrictions
- Existing skills and expertise
- Data complexity
- Organizational structure
- Previous data failures

Centralized Team Structure

Centralized teams report to a single leader and coordinate with other teams

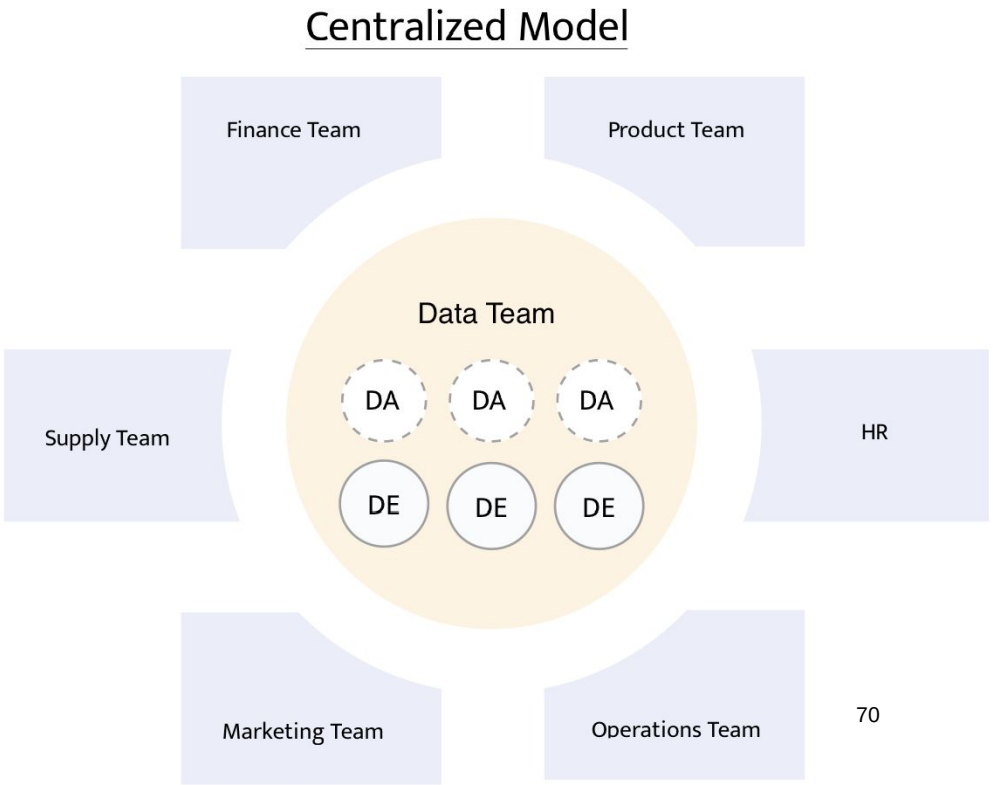


Source: https://www.getdbt.com/data-teams/centralized-vs-decentralized

Centralized Team Structure

Centralized teams report to a single leader and coordinate with other teams

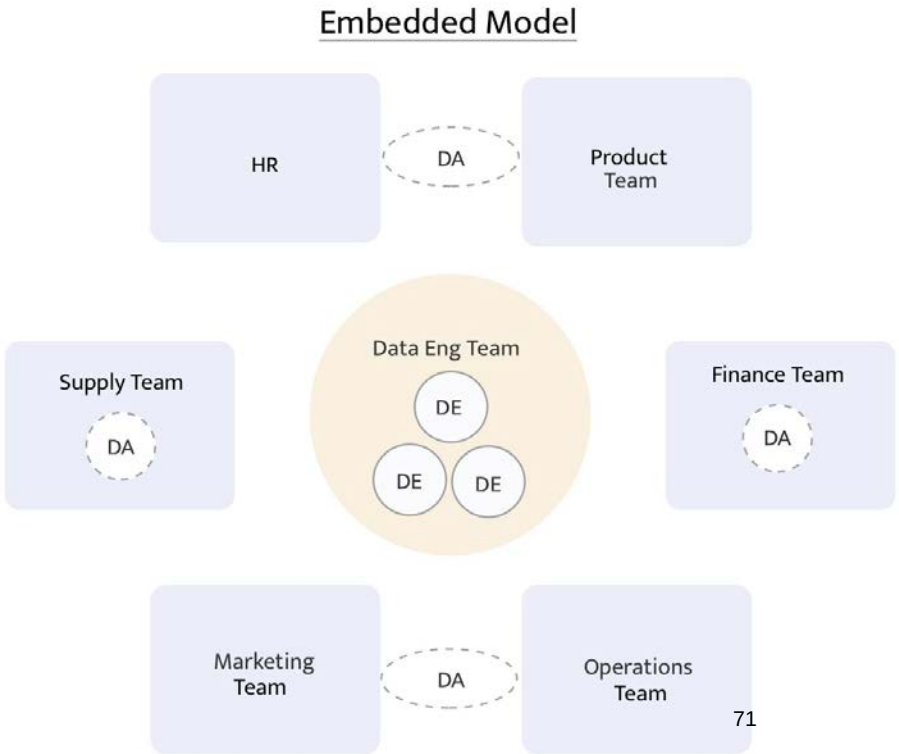
- Eliminates conflict of interest when teams do their own analyses
- Improves knowledge sharing
- Easier team alignment



Source: https://www.getdbt.com/data-teams/centralized-vs-decentralized

Embedded Team Structure

Embedded team structure creates more direct accountability and lines of support

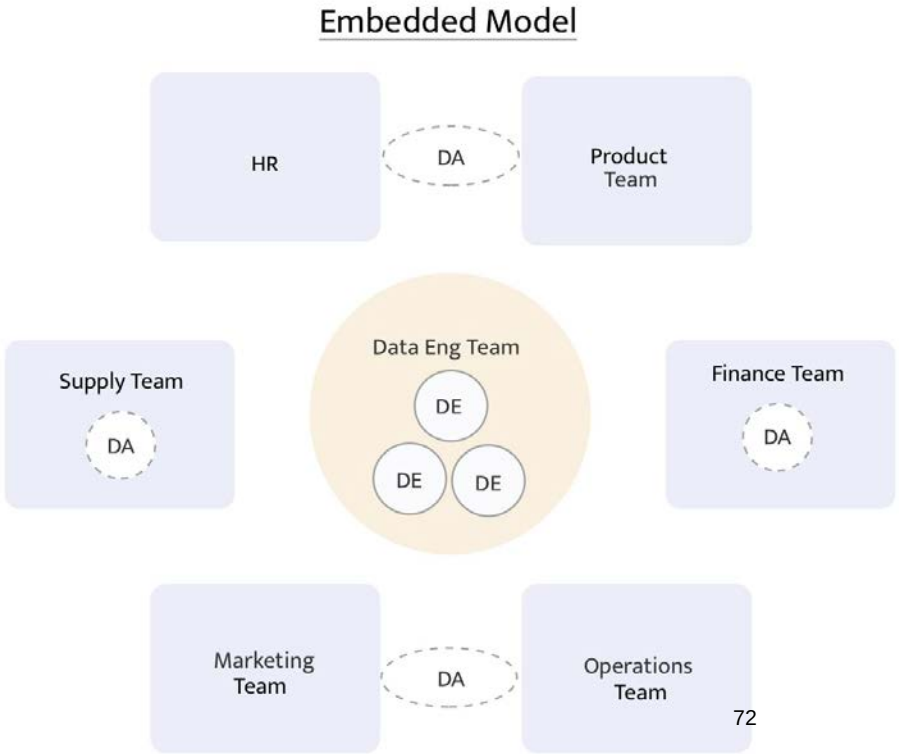


Source: https://www.getdbt.com/data-teams/centralized-vs-decentralized

Embedded Team Structure

Embedded team structure creates more direct accountability and lines of support

- Gives other teams a point person they can work with
- Creates direct alignment with a vertical team
- Can create single points of failure and limit collaboration

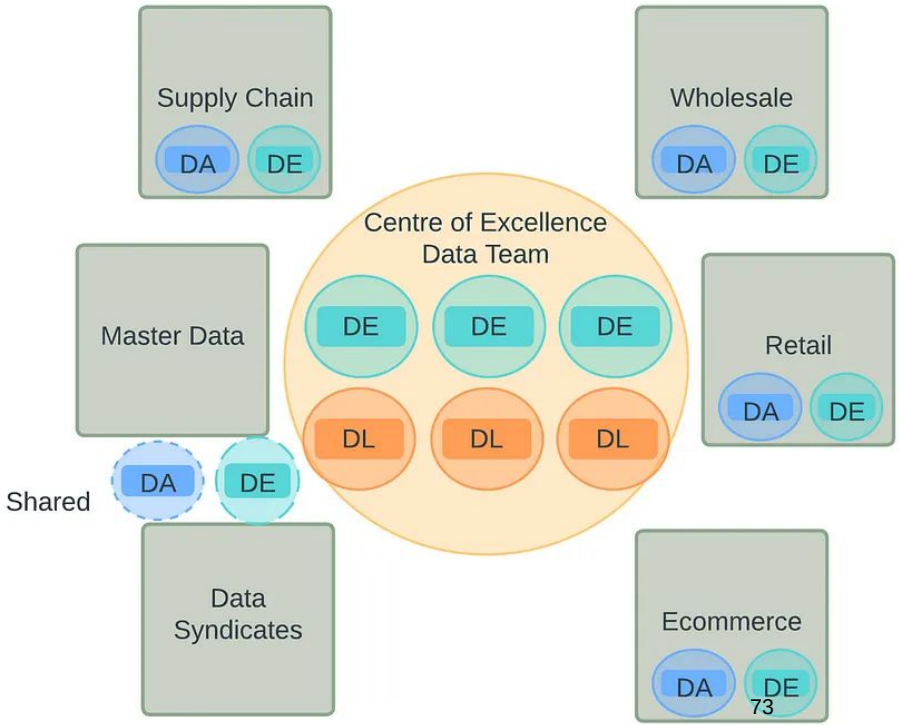


Source: https://www.getdbt.com/data-teams/centralized-vs-decentralized

Hybrid (Hub and Spokes) Team Structure

A hybrid team structure relies on a COE and a hub and spokes framework

- Creates an embedded framework without losing cross collaboration
- Centralizes a team that is focused on coordination and standardization

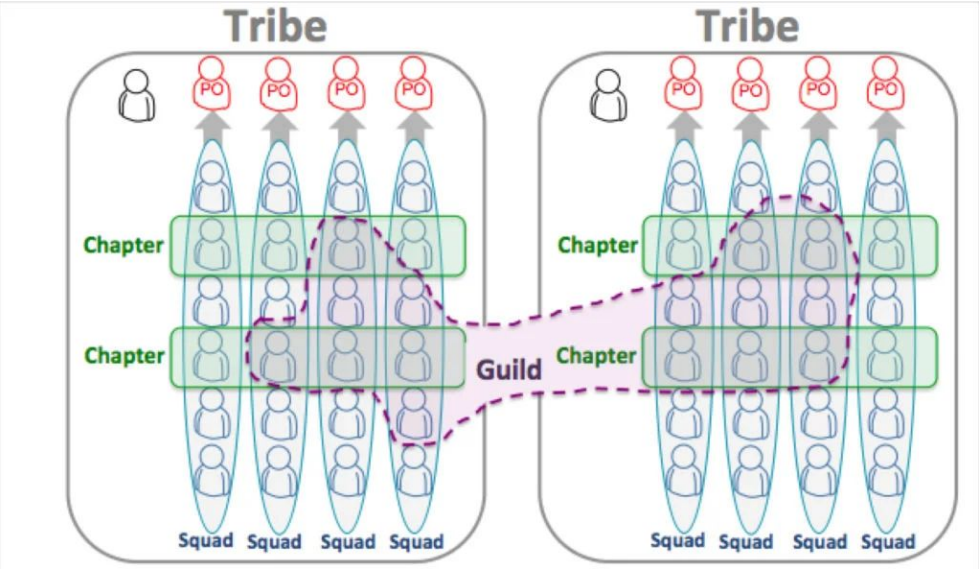


Source: https://medium.com/@DataEnthusiast/building-a-data-informed-organization-inside-look-at-the-data-teams-structure-and-roles-84d4df4c2160

The Guild Model was popularized at Spotify

This model doesn't work.

- I promise you, it doesn't work



Source: https://www.montecarlodata.com/blog-data-center-of-excellence/

Recommended reading

- Unraveling the Concepts of a Data Warehouse and a Data Lake (link)
- Quickstart for dbt Cloud and BigQuery (link)
- Data Team Handbook at GitLab (link)
- What's the Difference Between ETL and ELT? (link)
- The Analytics Engineering Podcast – Ep38: A romp through database history (w/ Postgres co-creator Mike Stonebraker + Andy Palmer) (link)
- The Data Engineering Podcast – Reflecting on The Past 6 Years of Data Engineering (link)
- Conceptual vs logical vs physical data models (link)
- The Analytics Setup Guidebook (link)
- Is the Modern Data Stack Still a Useful Idea? (link)
- Data or Pokemon (link)

Image Credits

Slide 5:

HCA Florida Citrus Hospital photo by Douglas R. Clifford, Zuma Press. In: Wainer, David, Nov. 6, 2024. "Trump Will Create New Winners and Losers in Healthcare." *The Wall Street Journal.* © Dow Jones & Company, Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see: https://ocw.mit.edu/help/faq-fair-use/

Slide 7:

Photo of Harry S. Truman, by Byron Rollins, "A presidential election history lesson: Americans often waited days or weeks for the outcome", The Washington Post, November 4, 2020. Public domain.

Slide 8:

President Harry S. Truman at Union Station, St. Louis, Missouri. Truman Library, Accession Number 58-777 © source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Image Credits

Slide 9:
Screenshot of Trump campaign rally footage © source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/
Headline from Newsweek article, Sept. 06, 2024 © Newsweek Digital LLC. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 11:
Image & text excerpt from KDnuggets.com © Guiding Tech Media. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slides 14 & 15:
Figure 1 in: "How do Data Professionals Spend their Time on Data Science Projects?" Bob Hayes, February 19, 2019 © Business Over Broadway. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 16:
How data scientists spend their time, In: Anaconda “2020 State of Data Science: Moving From Hype Toward Maturity.” © Anaconda Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Image Credits

Slide 17:

What Students learn, What Universities Offer, What Enterprises Lack, In: Anaconda "2020 State of Data Science: Moving From Hype Toward Maturity." © Anaconda Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 30:

Data Sufficiency Pyramid © source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 36:

Fundamentals of Data Architecture diagram, by MOTO DEI on Medium.com, Sept. 11, 2020, using materials from Irasuto-ya © Irasutoya. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 40:

Article excerpt: Toward smart production: Machine intelligence in business operations, by Duane S. Boning, Vijay D'Silva, et al. © McKinsey & Company. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Photo of machinery © Export Development Canada (EDC). All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Image Credits

Slides 41 & 42:
Improvement metrics chart © McKinsey & Company. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Quote from article © McKinsey & Company. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slides 47-58:
Modern Data Stack diagram, in: Densmore, J. *Data Pipelines Pocket Reference*. O'Reilly Media, ISBN: 978-1-492-08783-0. © James Densmore. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Trademarks in slides 47-58:
Logos are trademarked/registered to their respective owners.

Image Credits

Slide 60:

Diagram of ETL & ELT processes © Amazon Web Services, Inc. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 64:

Fig. 3 Star Schema and Fig. 4 Snowflake Schema © ACM. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slides 69 & 70:

Centralized Team Model © David Murray, via Medium. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slides 71 & 72:

Embedded Team Model © David Murray, via Medium. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 73:

Hybrid Team Structure © Himanshu Gaurav, via Medium. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

Slide 74:

The Guild Model © Functionly. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

HST.953 Clinical Data Learning, Visualization, and Deployments
Fall 2024