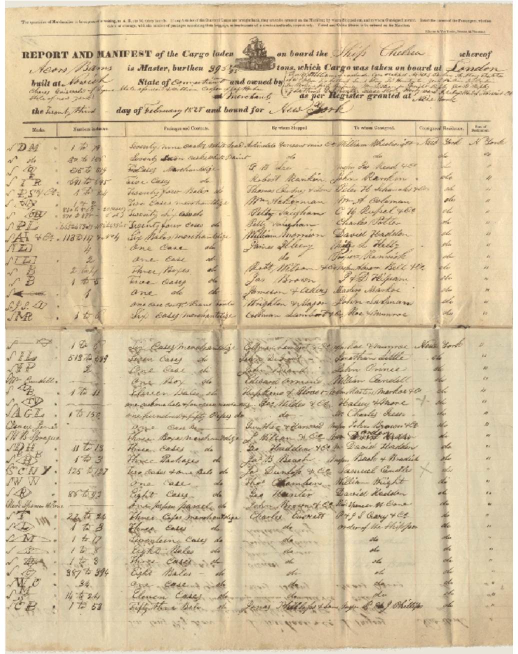
Intro to BI:
History and Evolution

CS 459 Business Intelligence

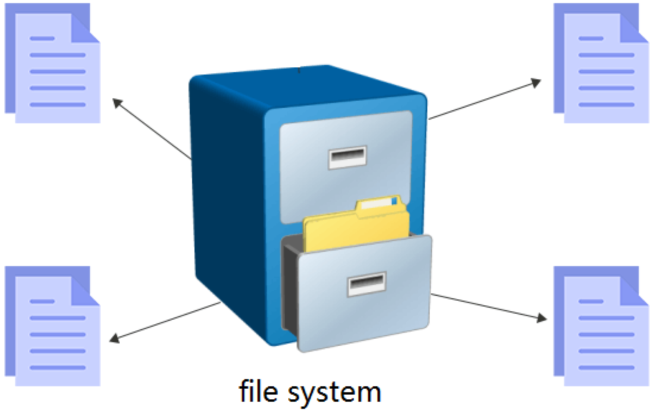


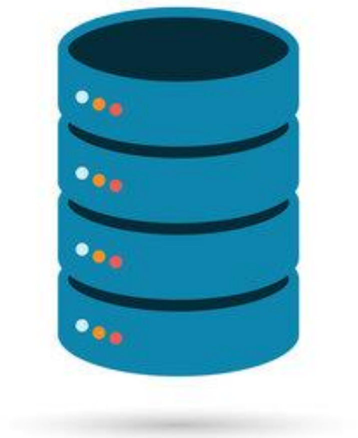
REPORT AND MANIFEST of the Cargo laden on board the Ship Chelsea whereof Acors Barns is Master, burthen tons, which Cargo was taken on board at London

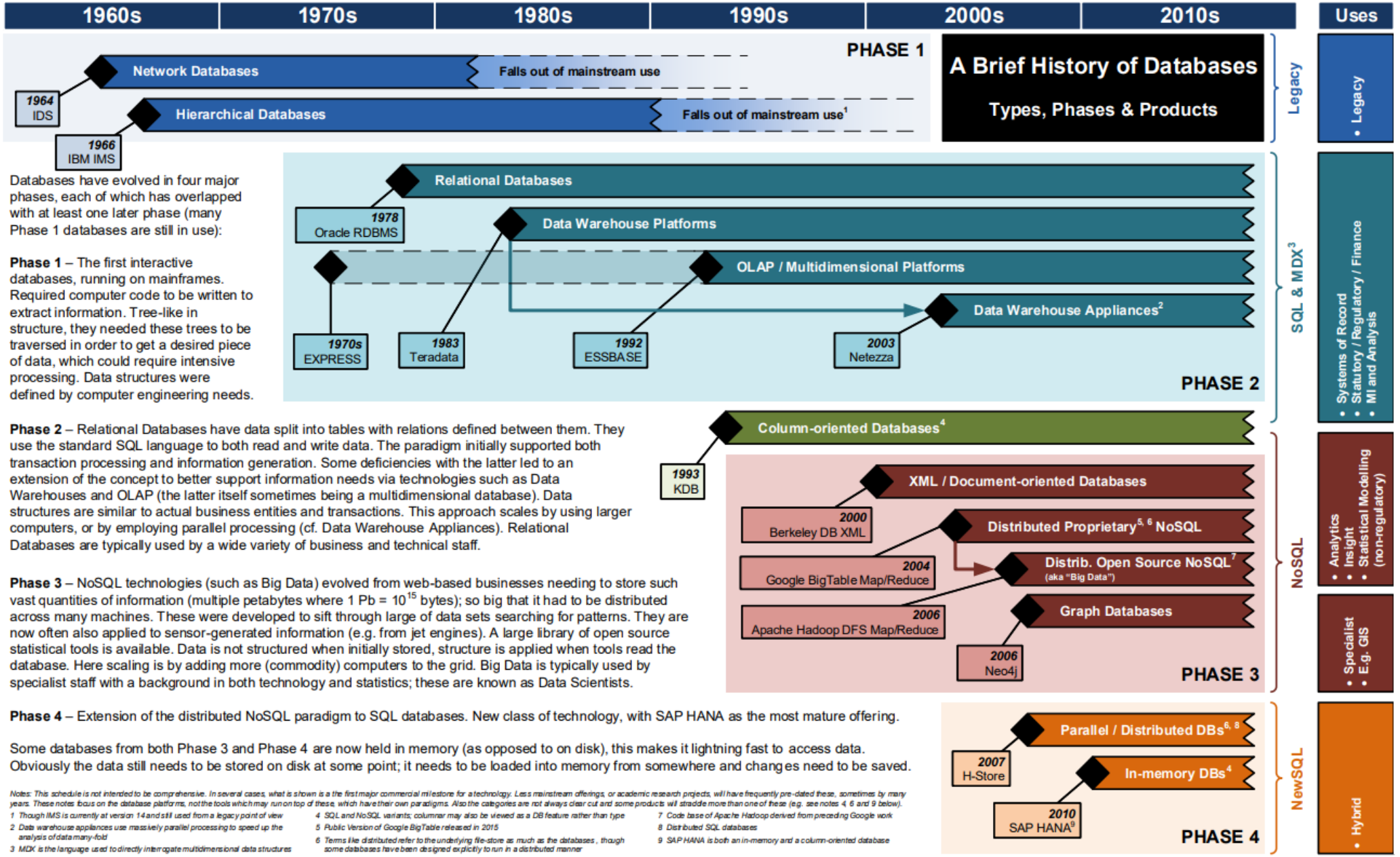
built at Norwich State of Connecticut and owned by Merchants as per Register granted at New York

the Twenty third day of February 1828 and bound for New York

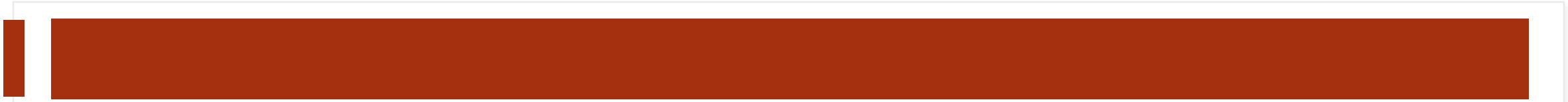
Marks.	Numbers inclusive.	Packages and Contents.	By whom Shipped.	To whom Consigned.	Consignees' Residence.	Place of Destination.
DM	1 to 79			William Wheelwright	New York	N York
	691 to 695	Five Cases	Robert Rankin	John Rankin		






1960s
1970s
1980s
1990s
2000s
2010s
Uses
PHASE 1
Network Databases
Falls out of mainstream use
1964 IDS
Hierarchical Databases
Falls out of mainstream use[1]
1966 IBM IMS
A Brief History of Databases
Types, Phases & Products
Legacy
• Legacy
Databases have evolved in four major phases, each of which has overlapped with at least one later phase (many Phase 1 databases are still in use):
Phase 1 – The first interactive databases, running on mainframes. Required computer code to be written to extract information. Tree-like in structure, they needed these trees to be traversed in order to get a desired piece of data, which could require intensive processing. Data structures were defined by computer engineering needs.
Relational Databases
1978 Oracle RDBMS
Data Warehouse Platforms
OLAP / Multidimensional Platforms
Data Warehouse Appliances[2]
1970s EXPRESS
1983 Teradata
1992 ESSBASE
2003 Netezza
PHASE 2
SQL & MDX[3]
• Systems of Record
• Statutory / Regulatory / Finance
• MI and Analysis
Phase 2 – Relational Databases have data split into tables with relations defined between them. They use the standard SQL language to both read and write data. The paradigm initially supported both transaction processing and information generation. Some deficiencies with the latter led to an extension of the concept to better support information needs via technologies such as Data Warehouses and OLAP (the latter itself sometimes being a multidimensional database). Data structures are similar to actual business entities and transactions. This approach scales by using larger computers, or by employing parallel processing (cf. Data Warehouse Appliances). Relational Databases are typically used by a wide variety of business and technical staff.
Column-oriented Databases[4]
1993 KDB
XML / Document-oriented Databases
2000 Berkeley DB XML
Distributed Proprietary[5, 6] NoSQL
2004 Google BigTable Map/Reduce
Distrib. Open Source NoSQL[7] (aka "Big Data")
2006 Apache Hadoop DFS Map/Reduce
Graph Databases
2006 Neo4j
PHASE 3
NoSQL
• Analytics
• Insight
• Statistical Modelling (non-regulatory)
• Specialist
• E.g. GIS
Phase 3 – NoSQL technologies (such as Big Data) evolved from web-based businesses needing to store such vast quantities of information (multiple petabytes where 1 Pb = 10^{15} bytes); so big that it had to be distributed across many machines. These were developed to sift through large of data sets searching for patterns. They are now often also applied to sensor-generated information (e.g. from jet engines). A large library of open source statistical tools is available. Data is not structured when initially stored, structure is applied when tools read the database. Here scaling is by adding more (commodity) computers to the grid. Big Data is typically used by specialist staff with a background in both technology and statistics; these are known as Data Scientists.
Phase 4 – Extension of the distributed NoSQL paradigm to SQL databases. New class of technology, with SAP HANA as the most mature offering.
Some databases from both Phase 3 and Phase 4 are now held in memory (as opposed to on disk), this makes it lightning fast to access data. Obviously the data still needs to be stored on disk at some point; it needs to be loaded into memory from somewhere and changes need to be saved.
Parallel / Distributed DBs[6, 8]
2007 H-Store
In-memory DBs[4]
2010 SAP HANA[9]
PHASE 4
NewSQL
• Hybrid
Notes: This schedule is not intended to be comprehensive. In several cases, what is shown is a the first major commercial milestone for a technology. Less mainstream offerings, or academic research projects, will have frequently pre-dated these, sometimes by many years. These notes focus on the database platforms, not the tools which may run on top of these, which have their own paradigms. Also the categories are not always clear cut and some products will straddle more than one of these (e.g. see notes 4, 6 and 9 below).
1 Though IMS is currently at version 14 and still used from a legacy point of view
2 Data warehouse appliances use massively parallel processing to speed up the analysis of data many-fold
3 MDX is the language used to directly interrogate multidimensional data structures
4 SQL and NoSQL variants; columnar may also be viewed as a DB feature rather than type
5 Public Version of Google BigTable released in 2015
6 Terms like distributed refer to the underlying file-store as much as the databases, though some databases have been designed explicitly to run in a distributed manner
7 Code base of Apache Hadoop derived from preceding Google work
8 Distributed SQL databases
9 SAP HANA is both an in-memory and a column-oriented database
Acknowledgements:
a. Neil Raden (@neilraden) from Hired Brains (hiredbrains.com) – For both reviewing earlier drafts and providing significant input
© peterjamesthomas.com 2017 - 2018
Creative Commons Attribution 4.0 International License

Recall:
OLTP vs OLAP



Operational/Transactional Database

- Stored in a Relational Database or files.
- Highly Normalized (Data stored as efficiently as possible, lots of tables.)
- Optimized for processing speed and handling the "now".
- Designed for capturing data, not for reporting on it.
- Designed to support the operational needs of the organization.

Walmart Story

Walmart Story

- 1962 Founded by Sam Walton
- 1967 Expanded to 24 stores
- 1970 Company goes nation-wide

Going Big!

- Breaking all records
- First company to grow that fast $1billion in annual sales by 1980
- America's top retailer by 1990

- *Approximate statistics of today:* 2.2 million employees, 200 million customers a week, 10,000 stores in 27 countries more than $1billion per day sales

Enters Teradata

- 1976-1979: In the parallel world, enters **Teradata -** Startup from Caltech (California Institute of Technology)
- 1984: Released first parallel processing architecture for DWH

Walmart Meets Teradata (1990)

- Walmart CIO: *I want to keep track of sales in all my 1900 stores simultaneously*
- Teradata consultant: *You need our RDBMS software. You can stuff data in as sales are rung up at cash registers and simultaneously query data right in your office*
- So, Walmart buys a $1 million Sun E10000 multi-CPU server, a $500000 Teradata license, expensive technical consultants, and builds a **normalized SQL data model**

DWH Story

- After a few months of collecting large amounts of data a Walmart executive asks…
  *I have noticed that there was a **Colgate promotion** recently, directed to people who live in small towns. How much toothpaste did we sell in those towns yesterday?*

```
select sum(sales.quantity_sold) from sales, products, product_categories, manufacturers,
stores, cities where manufacturer_name = 'Colgate'
and product_category_name = 'toothpaste'
and cities.population < 40 000
and trunc(sales.date_time_of_sale) = trunc(sysdate-1)
and sales.product_id = products.product_id
and sales.store_id = stores.store_id
and products.product_category_id = product_categories.product_category
and products.manufacturer_id = manufacturers.manufacturer_id
and stores.city_id = cities.city_id
```

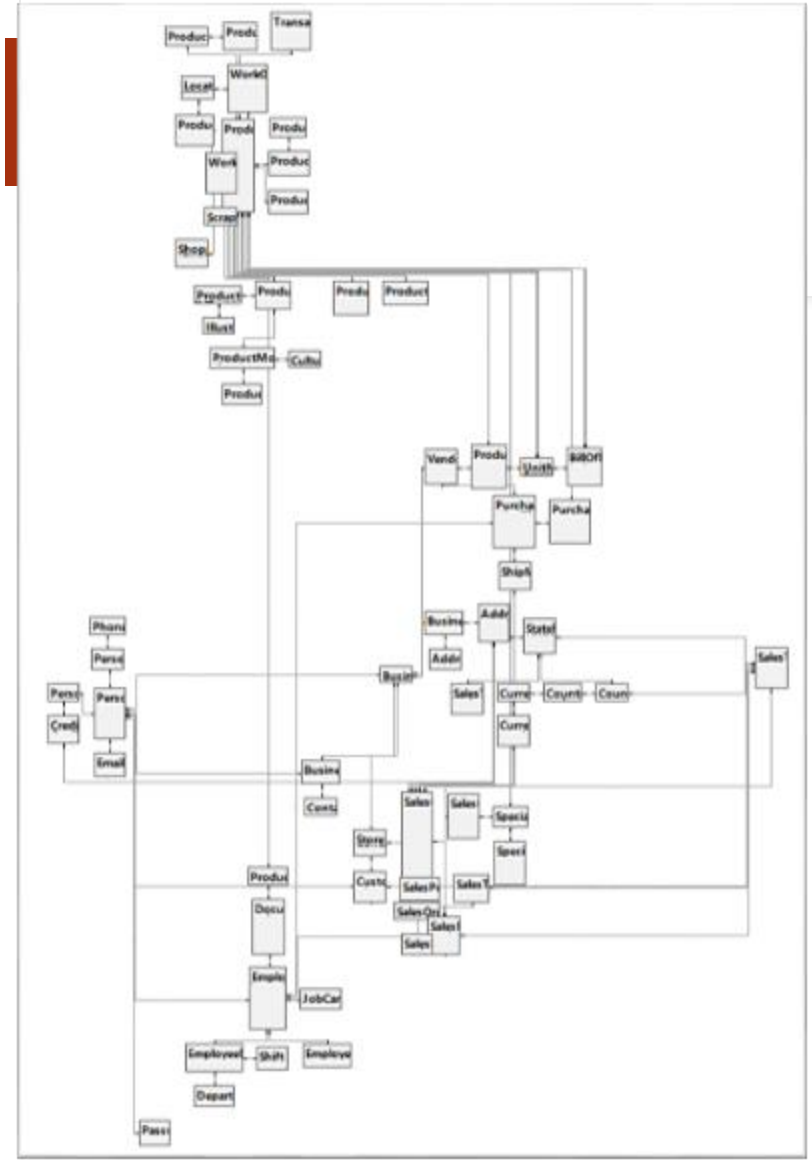
6 tables join
*Computer freezes for **20 minutes** and processing a sale becomes impossible*

DWH Story

- Walmart CIO: *WE TYPE IN THE TOOTHPASTE QUERY AND OUR SYSTEM HANGS!!!*
- Teradata support: *Of course it does! You built an **on-line transaction processing (OLTP)** system. You can't feed it a **decision support system (DSS)** query and expect things to work!*
- Walmart CIO: *<furious> I thought this was the whole point of SQL and your RDBMS…to query and insert simultaneously!!*
- Teradata support: *Uh, not exactly. If you're **reading** from the database, nobody can **write** to the database. If you're **writing** to the database, nobody can **read** from the database. So if you've got a query that takes 20 minutes to run and don't specify **special locking instructions**, nobody can update those tables for 20 minutes.*

DWH Story

- Walmart CIO: *It sounds like a bug*.
- Teradata support: *Actually, it is a feature. We call it* ***pessimistic locking***.
- Walmart CIO: *Can you fix your system so that it doesn't lock up???*
- Teradata support: *No. But we made this great loader tool so that you can copy everything from your OLTP system into a separate Data Warehouse system*



Transactional Databases Are *Complex*

- ←Adventure works fictitious bicycle manufacturer.
  72 tables.
- Blackboard Learning Management System.
  592 tables.
- Typical Oracle PeopleSoft ERP Implementation
  40,000+ tables.

Example: A Query of "iSchool Students"

- Students in the current term with gpa, demographics, major, minor, program of study, etc… Either enrolled in one of our programs or taking one of our courses.

```
select distinct s.term,
    e.emplid,  e.netid, e.email_published_addr, e.name_last_first_mid,
    case when (s.acad_prog_primary in (select distinct d.acad_prog from DBUSER.v_sis_stdnt_full_acad_prog_deg d  where (1=1)
                                       and ((d.acad_prog_org = 'IST') or (d.acad_prog like '%IS%' and d.acad_prog <> 'CIS' and d.acad_career='UGRD'))
                                      ) ) then 'iSchool Student' else 'Non-iSchool Student' end as IN_IST_PROG,
    s.total_cumulative, s.total_inprog_gpa, s.total_transfer, s.curr_gpa, s.cum_gpa, s.acad_career, s.acad_career_desc,
    s.acad_prog_primary, s.acad_prog_primary_desc, b.last_acad_term,  s.acad_level_begin_term, s.acad_level_begin_term_desc, s.acad_load, s.acad_load_sh_desc,
    p.acad_plans,
    (select max(d.matric_term)
        from dbuser.v_sis_stdnt_max_acad_prog_deg d
        where  d.acad_prog_status = 'AC' and d.emplid = s.emplid and d.acad_prog = s.acad_prog_primary) as matric_term_primary,
    (select max(d.admit_term)
        from dbuser.v_sis_stdnt_max_acad_prog_deg d
        where  d.acad_prog_status = 'AC' and d.emplid = s.emplid and d.acad_prog = s.acad_prog_primary) as admit_term_primary,
    (select max(d.expected_grad_term)
        from dbuser.v_sis_stdnt_max_acad_prog_deg d
        where  d.acad_prog_status = 'AC' and d.emplid = s.emplid and d.acad_prog = s.acad_prog_primary) as expected_grad_term_primary,
  b.citizenship_code, b.citizenship_desc,
  x.ECS_UGRD_EC_IS, x.IST_GRAD_CU07C, x.IST_GRAD_DA50C, x.IST_GRAD_DI10C, x.IST_GRAD_ES30C, x.IST_GRAD_GL60C, x.IST_GRAD_IN26C, x.IST_GRAD_IN31D,
  x.IST_GRAD_IN31M, x.IST_GRAD_IN32D, x.IST_GRAD_IN32M, x.IST_GRAD_IN34C, x.IST_GRAD_IN37C, x.IST_GRAD_IN40M, x.IST_GRAD_LI25M, x.IST_GRAD_LI27M,
  x.IST_GRAD_SC35C, x.IST_GRAD_TE10M, x.IST_UGRD_IS, x.IST_UGRD_IS_MG, x.PC_UGRD_PC_IS

  from DBUSER.v_sis_stdnt_term_summary_22 s
    join DBUSER.v_sis_stdnt_bio_data_2 b on b.emplid = s.emplid
    join DBUSER.v_sis_stdnt_max_acad_prog_deg d on d.emplid = s.emplid  --and s.acad_prog_primary = d.acad_prog
    join DBUSER.v_sec_student_email e on e.emplid =s.emplid
    join DBUSER.v_sis_term t on t.term = s.term
    join ( select d.emplid, d.acad_career,
            listagg(d.acad_plan_type_sh_desc || ': ' || d.acad_plan_desc, ' / ') within group (order by d.student_career_nbr) as acad_plans
            from dbuser.v_sis_stdnt_max_acad_prog_deg d where  d.acad_prog_status = 'AC'
            group by d.emplid, d.acad_career
          ) p on s.emplid = p.emplid and s.acad_career = p.acad_career
    left join (
      with pivot_data as (
          select distinct s.acad_career, s.emplid, s.acad_prog_org || '_' || s.acad_career || '_' || s.acad_prog as acad_org_career_prog, 1 as prog_count
            from DBUSER.v_sis_stdnt_max_acad_prog_deg s
            where (s.acad_prog_status ='AC')
            and (    (s.acad_prog_org = 'IST')
                  or (s.acad_prog like '%IS%' and s.acad_prog <> 'CIS' and s.acad_career='UGRD')
                )
            order by acad_org_career_prog
```

Issues Reporting with Transactional Databases

- **Difficult, Time-consuming & Error prone.**
  - Many joins, sub-selects, due to vast number of tables.
  - How do you know your query is correct?
- **Resource-intensive**
  - The database is not optimized for this purpose.
  - Multi table joins take over the RAM and CPU processing
- **Impossible**
  - Transactional systems are flushed or archived frequently to maintain performance.
  - You can't query data you no longer have

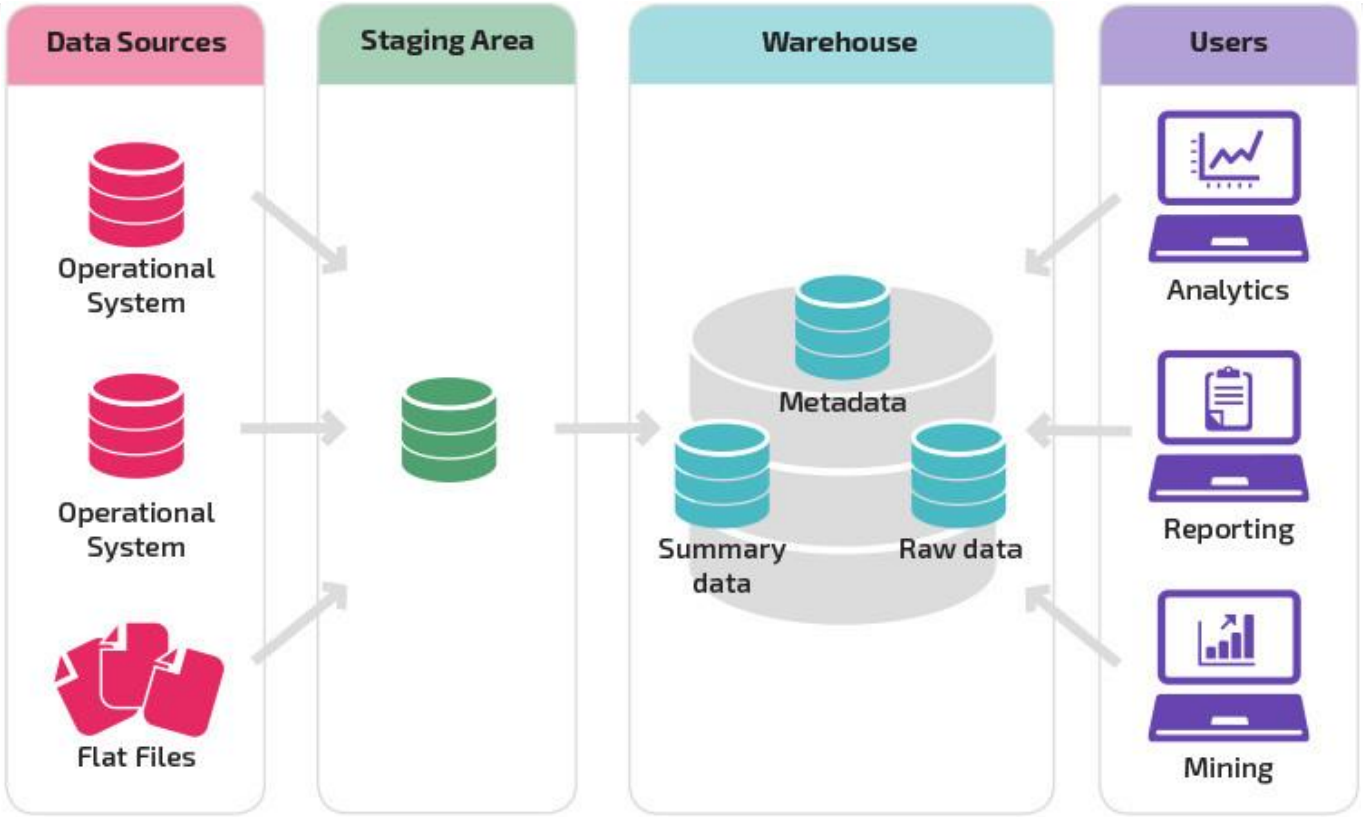
Solution? The Data Warehouse

- Designed to support an organization's **informational needs**.
- Data is re-structured for reporting and analytic applications.
- Transactional databases are data sources for the Data Warehouse.
- Data grows over time; existing data in the warehouse very seldom changes.

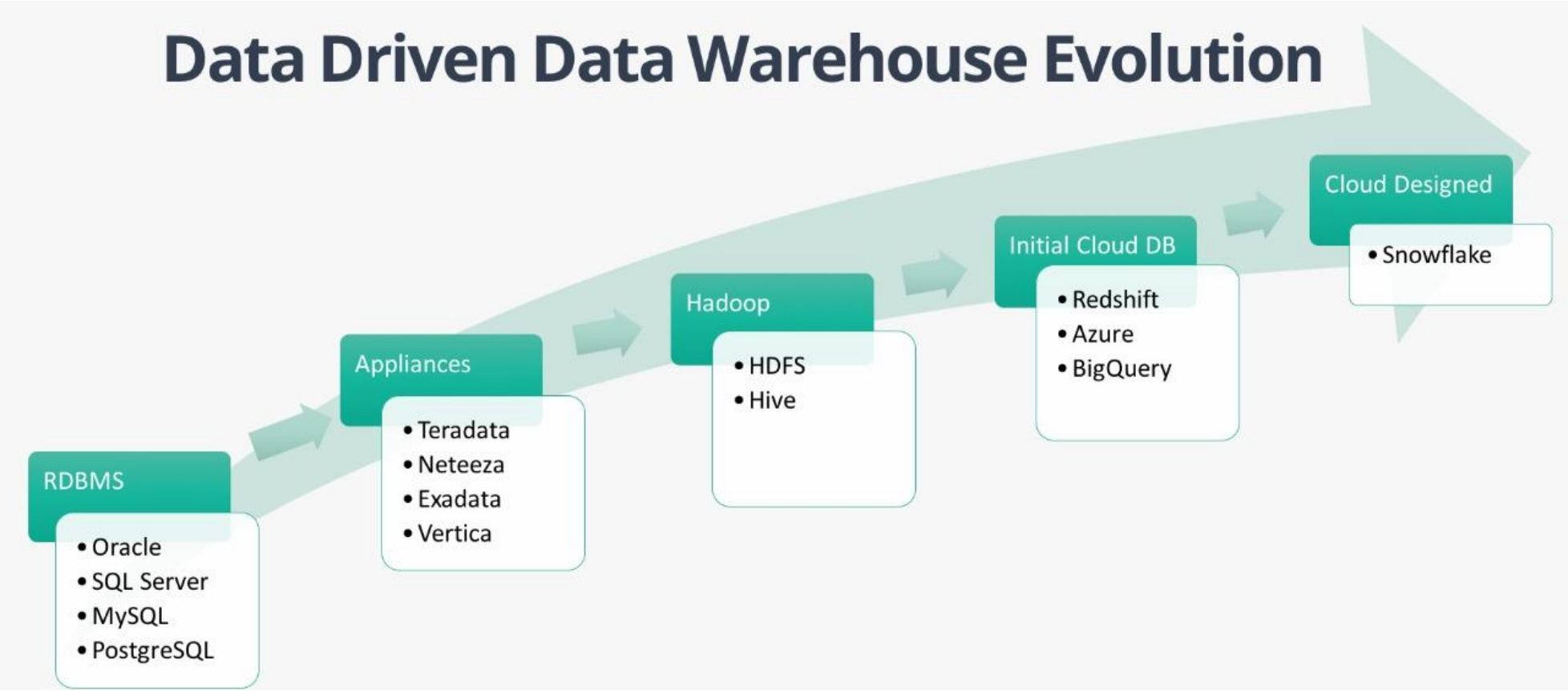
Session 02

2000s - Data Warehouses

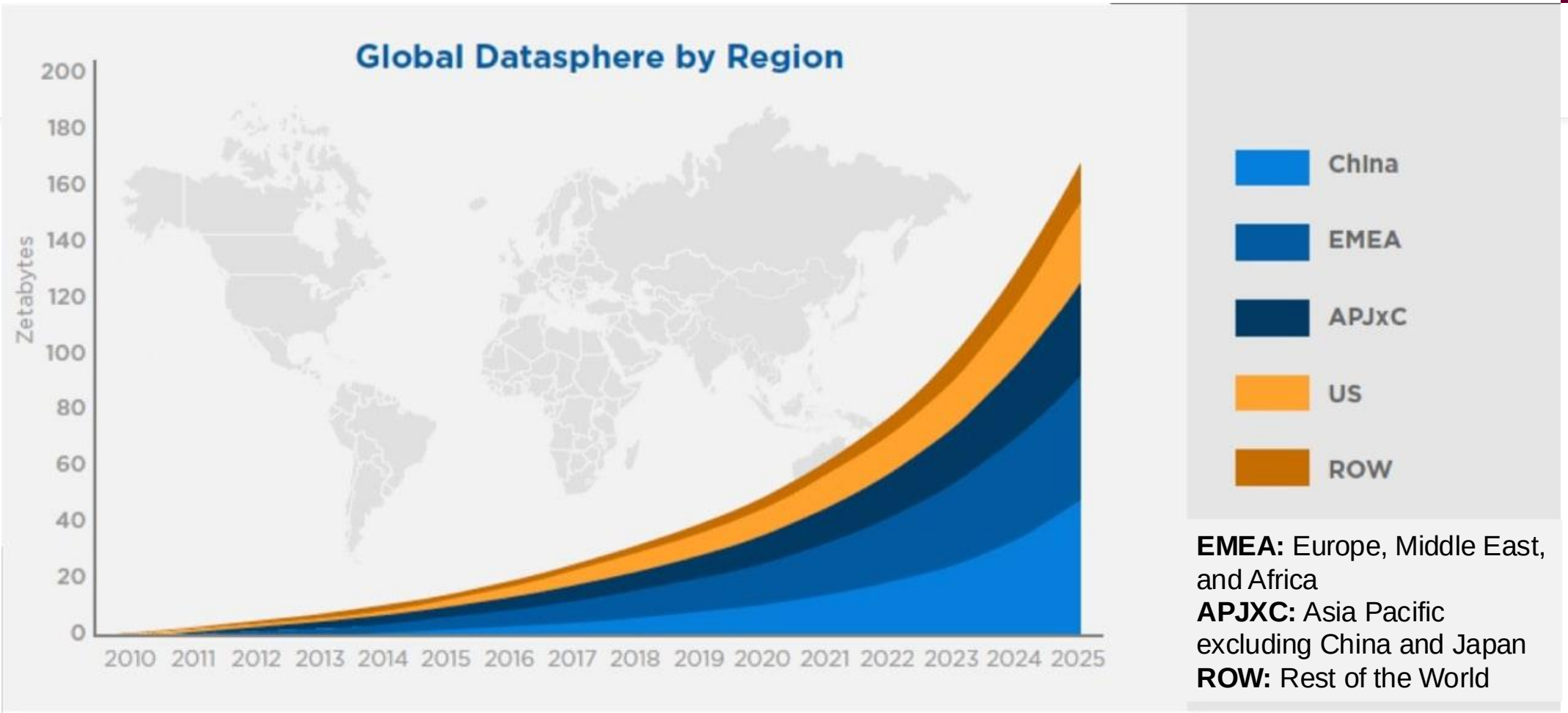
Developed by businesses to consolidate the data from a variety of databases to help support strategic decision-making.



Evolution of Data Warehousing



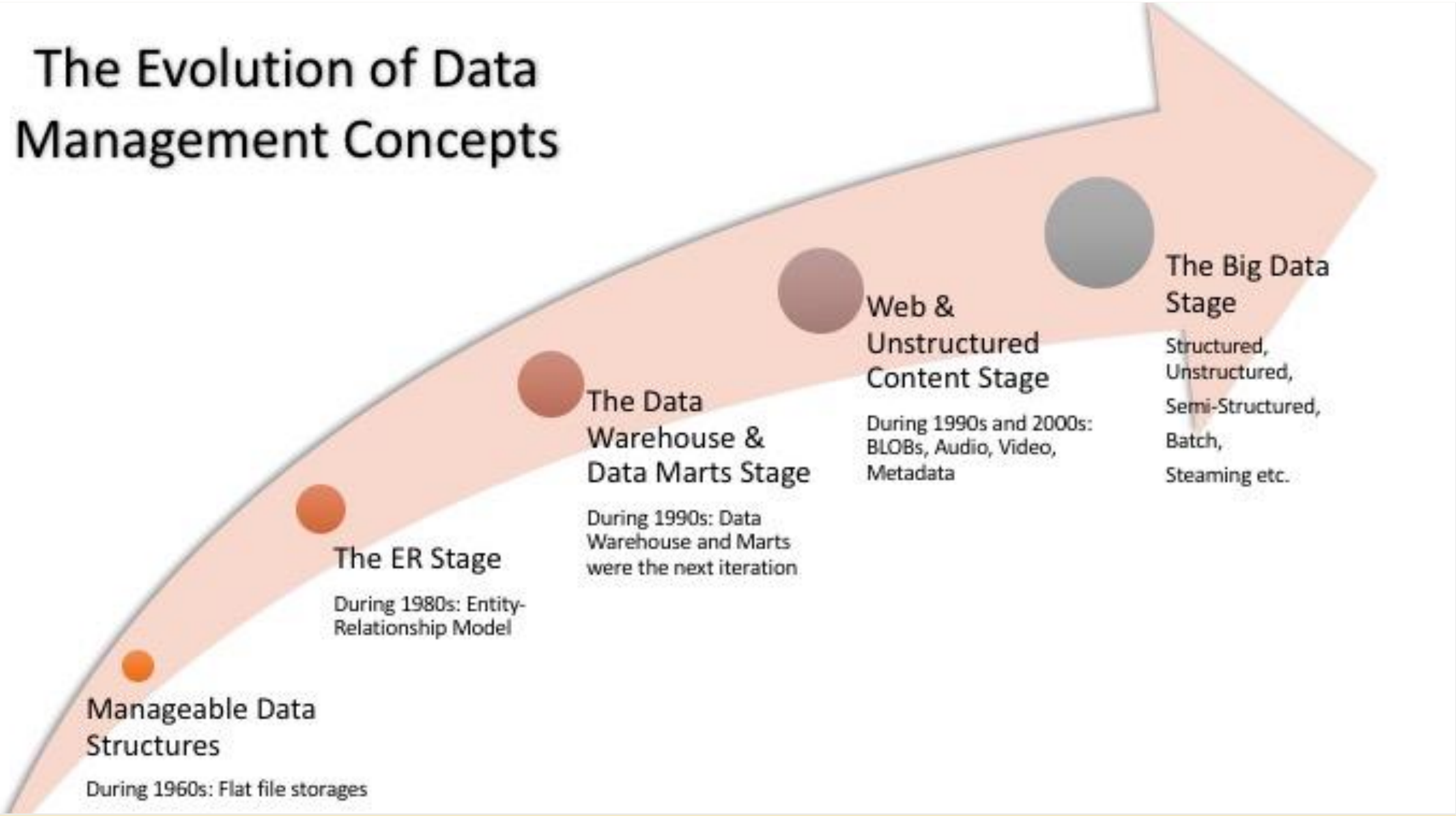
Datasphere Today



EMEA: Europe, Middle East, and Africa
APJXC: Asia Pacific excluding China and Japan
ROW: Rest of the World

Source: IDC's Data Age 2025 study, sponsored by Seagate

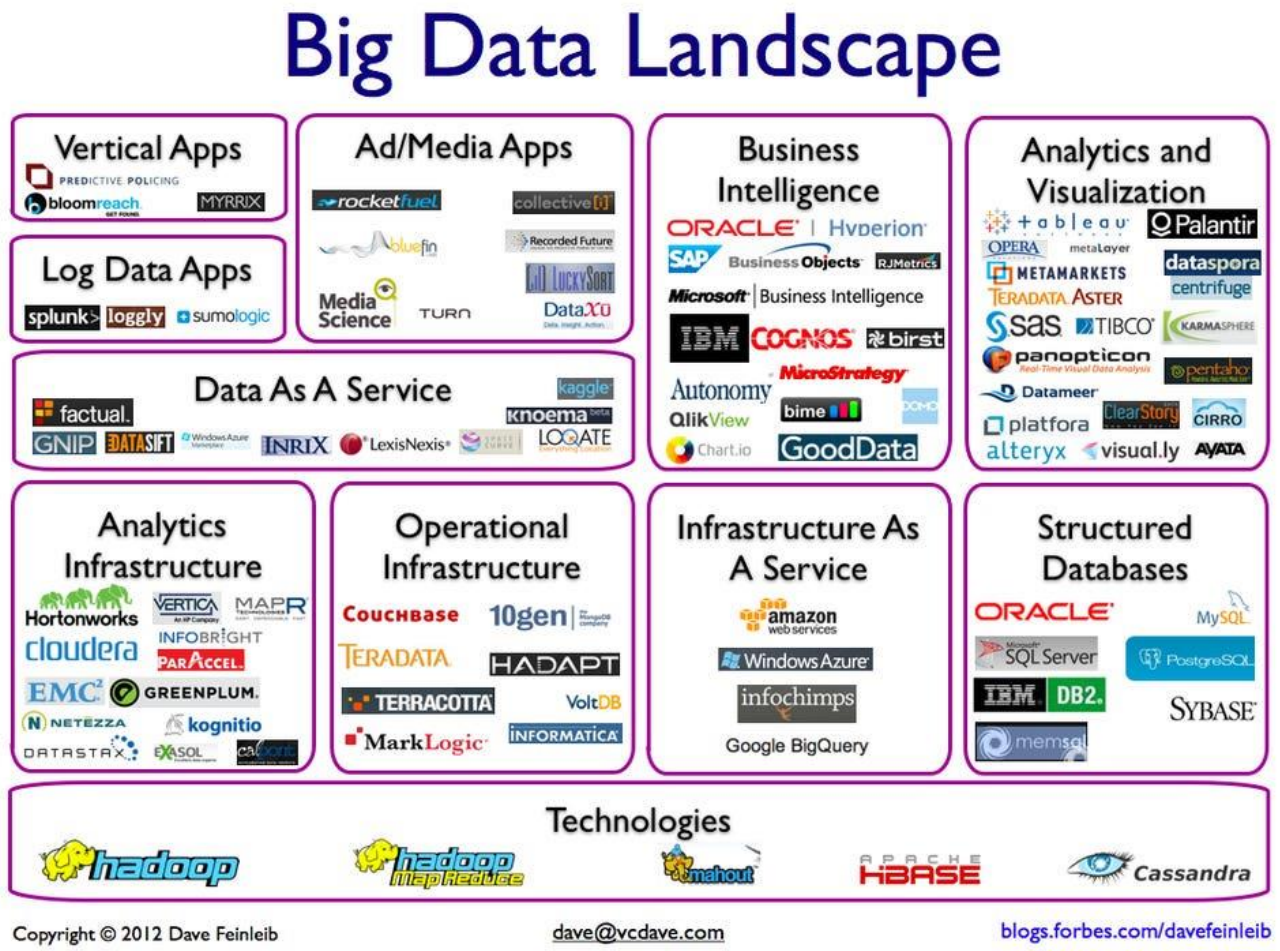
The Evolution of Data Management Concepts



Era of Big Data

3 Vs – Volume, Velocity, Variety

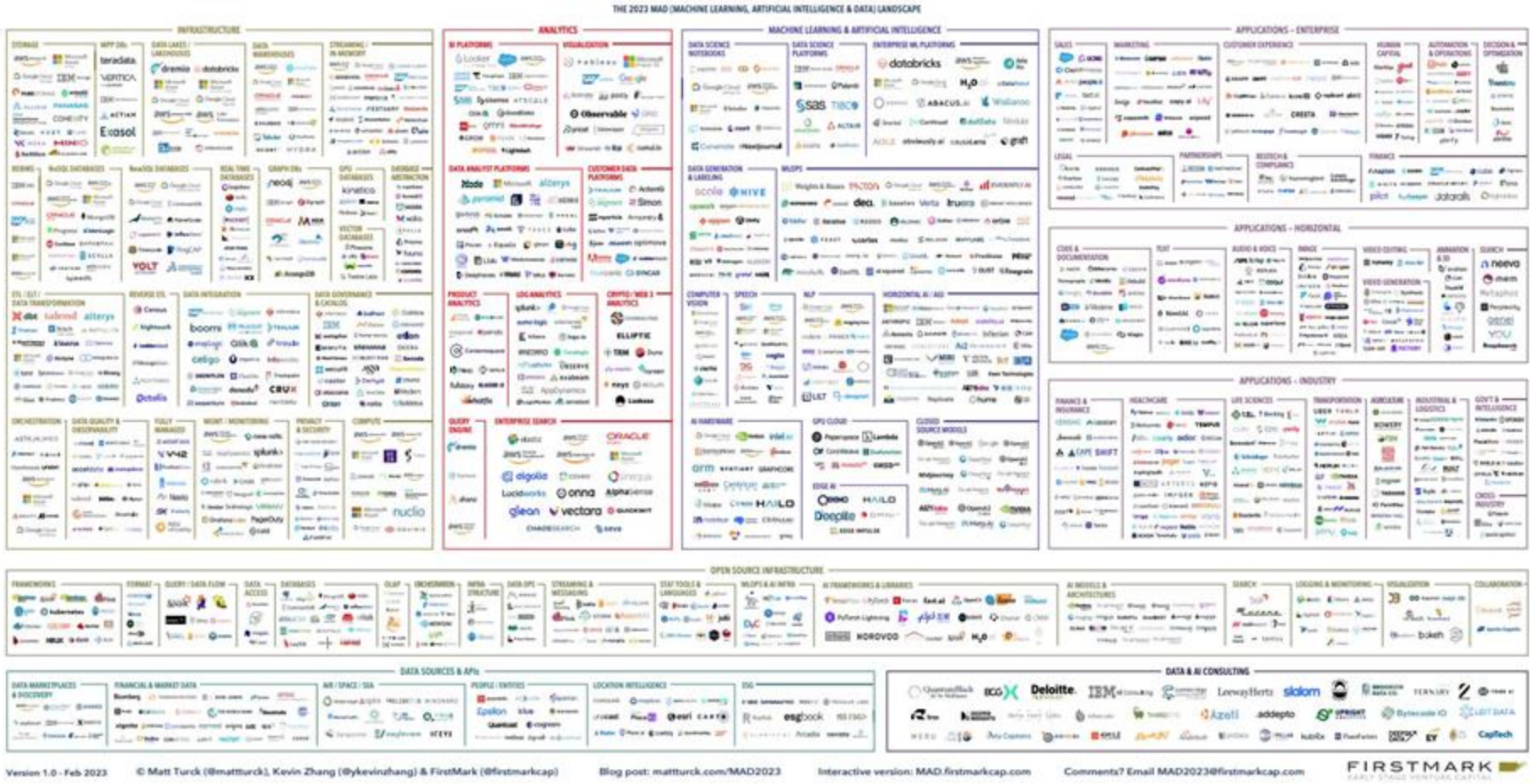
Big Data Landscape in 2012



MAD Landscape 2023

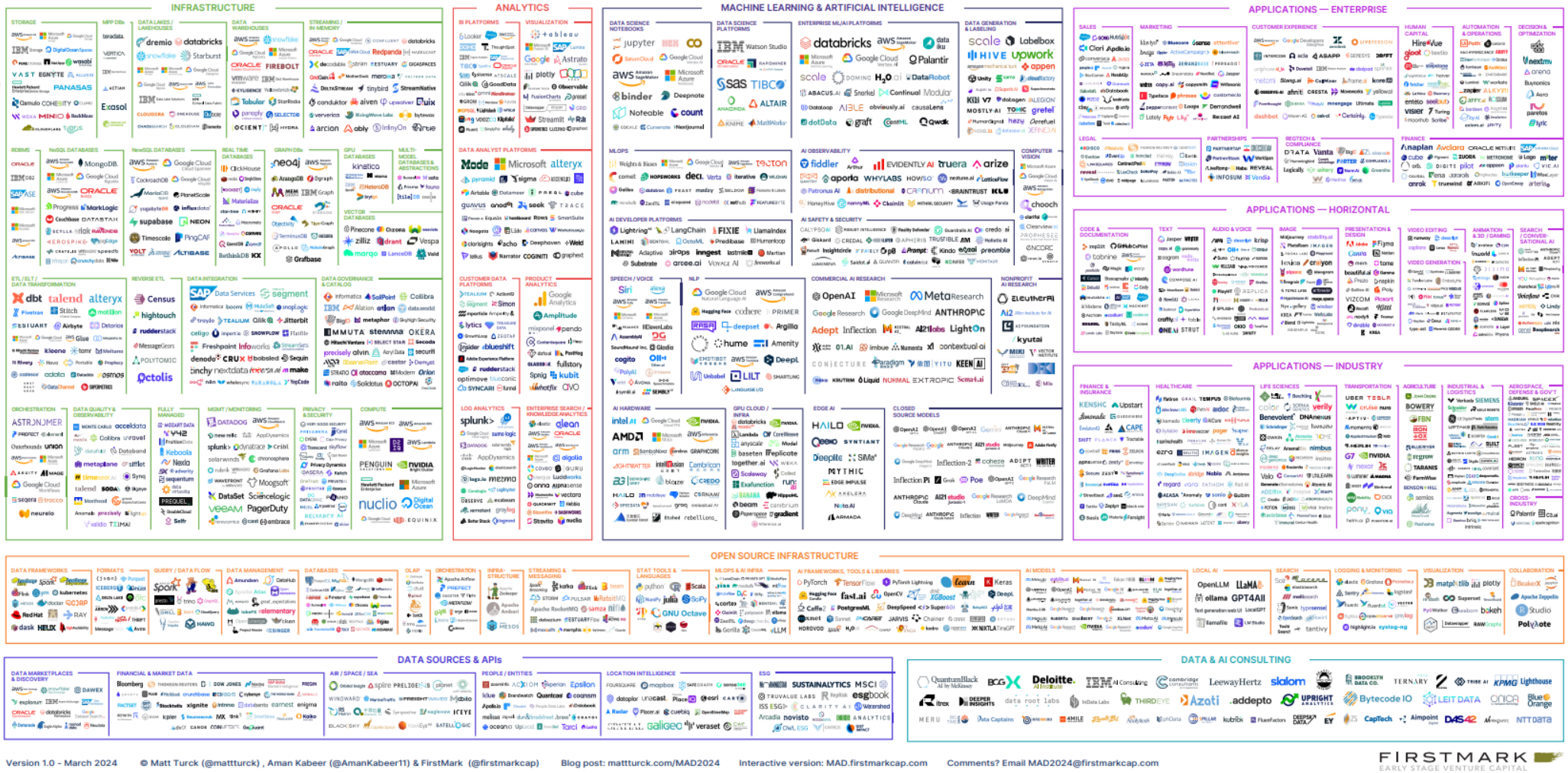
MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA

https://mattturck.com/landscape/mad2023.pdf



MAD Landscape 2024

FirstMark | 2024 MAD (ML/AI/Data) Landscape



BI Evolution

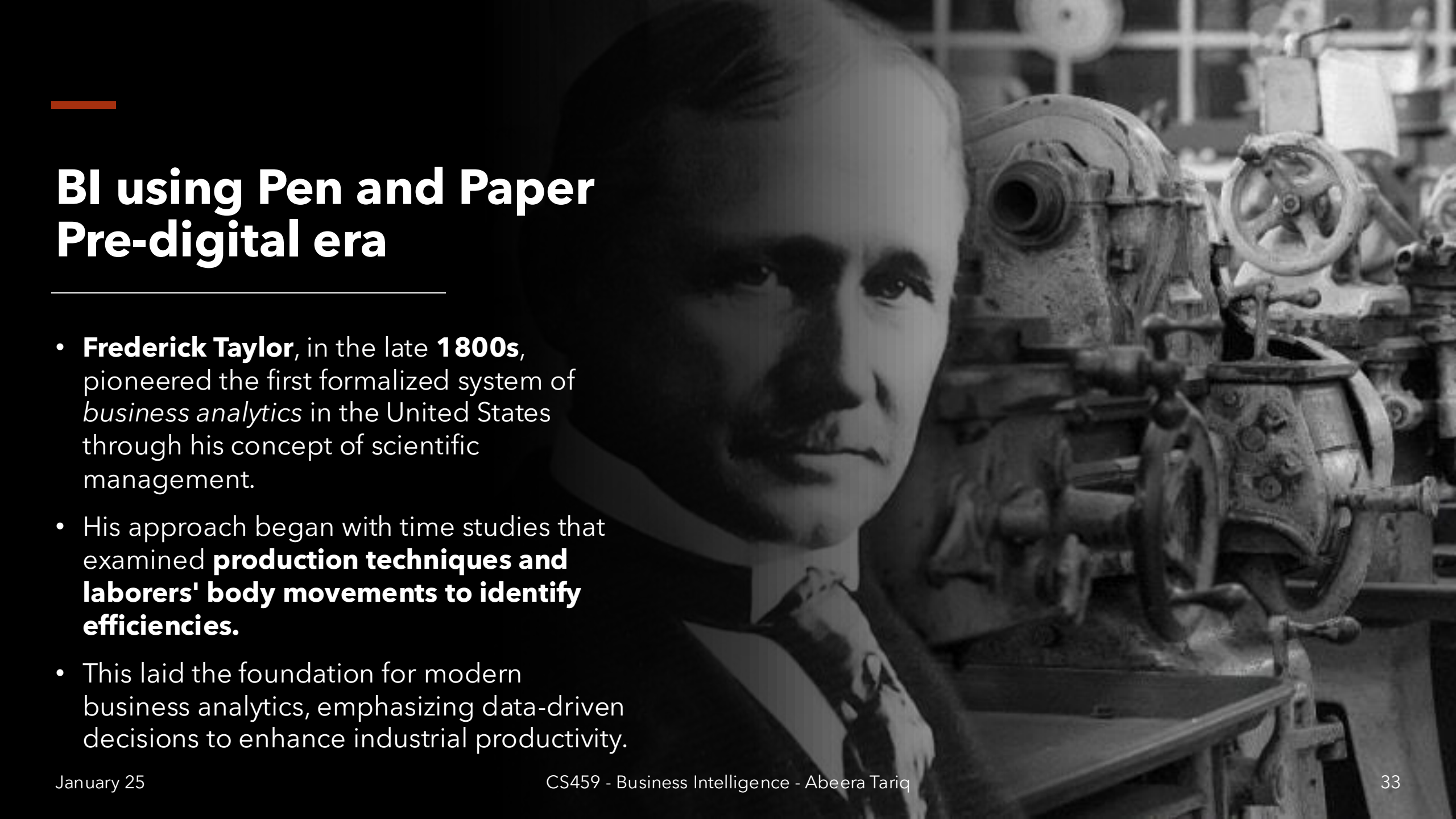
BI using Pen and Paper
Pre-digital era

- The term "business intelligence" was first introduced by **Richard Miller Devens** in *Cyclopedia of Commercial and Business Anecdote* (1865).
- Sir Henry Furnese profited by using timely information from a network of informants across Europe.
- Acting quickly on insights about political and market changes gave him a competitive edge.

Generated by DallE

BI using Pen and Paper Pre-digital era

- **Frederick Taylor**, in the late **1800s**, pioneered the first formalized system of *business analytics* in the United States through his concept of scientific management.
- His approach began with time studies that examined **production techniques and laborers' body movements to identify efficiencies.**
- This laid the foundation for modern business analytics, emphasizing data-driven decisions to enhance industrial productivity.



THE BEGINNING

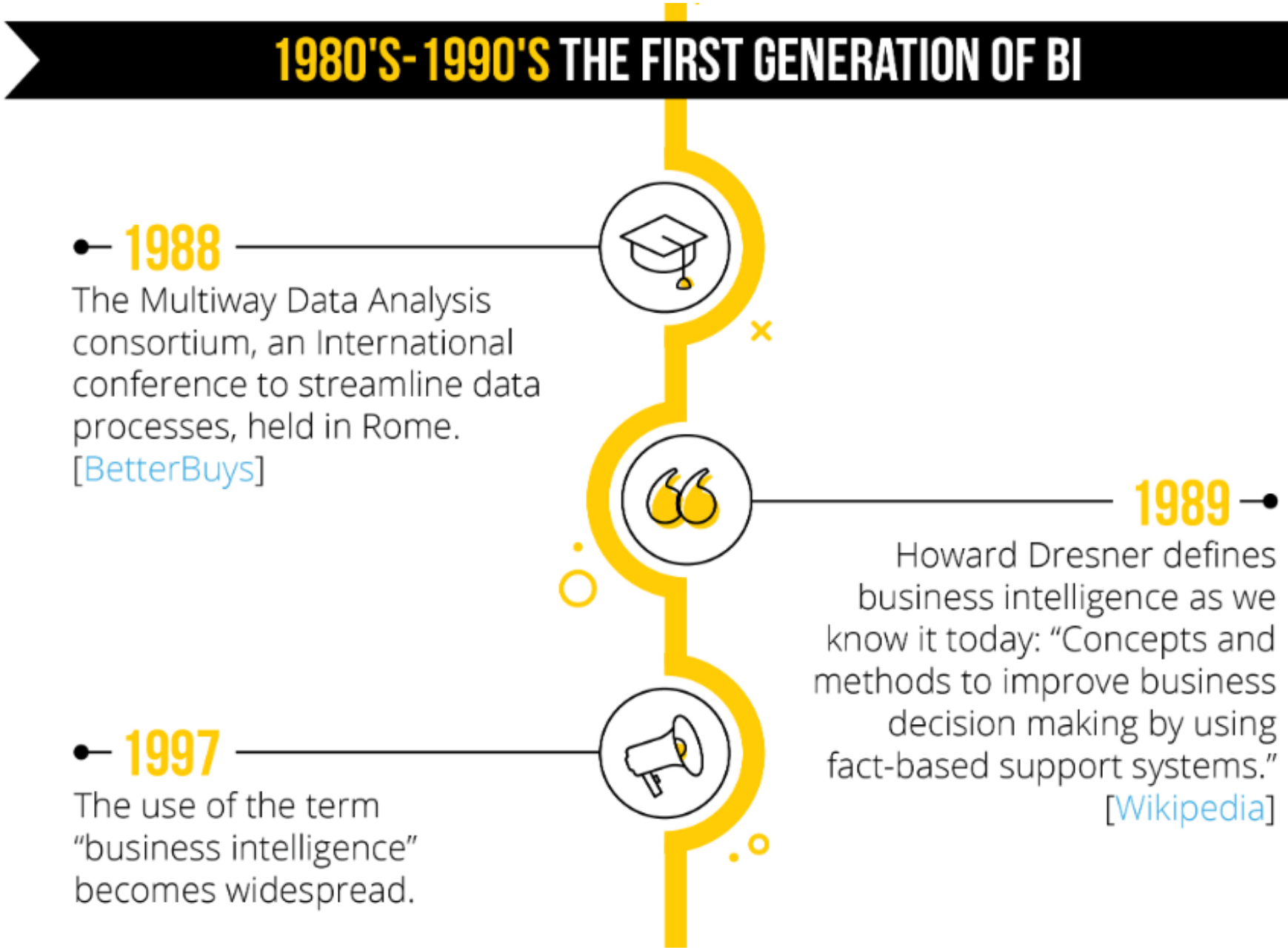
1958

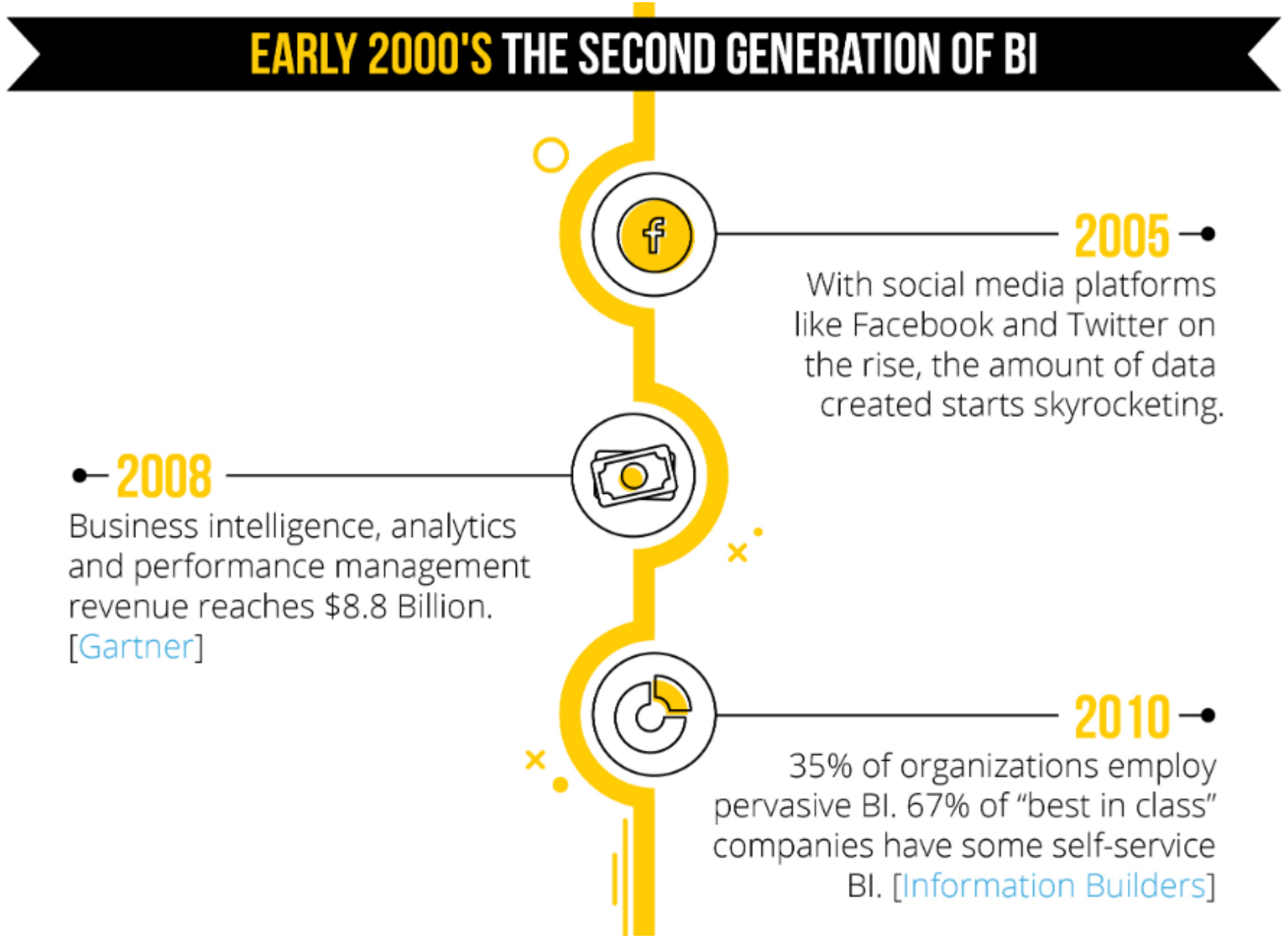
IBM Researcher Hans Peter Luhn publishes "A Business Intelligence System." Hans is later named the Father of Business Intelligence.

1970'S

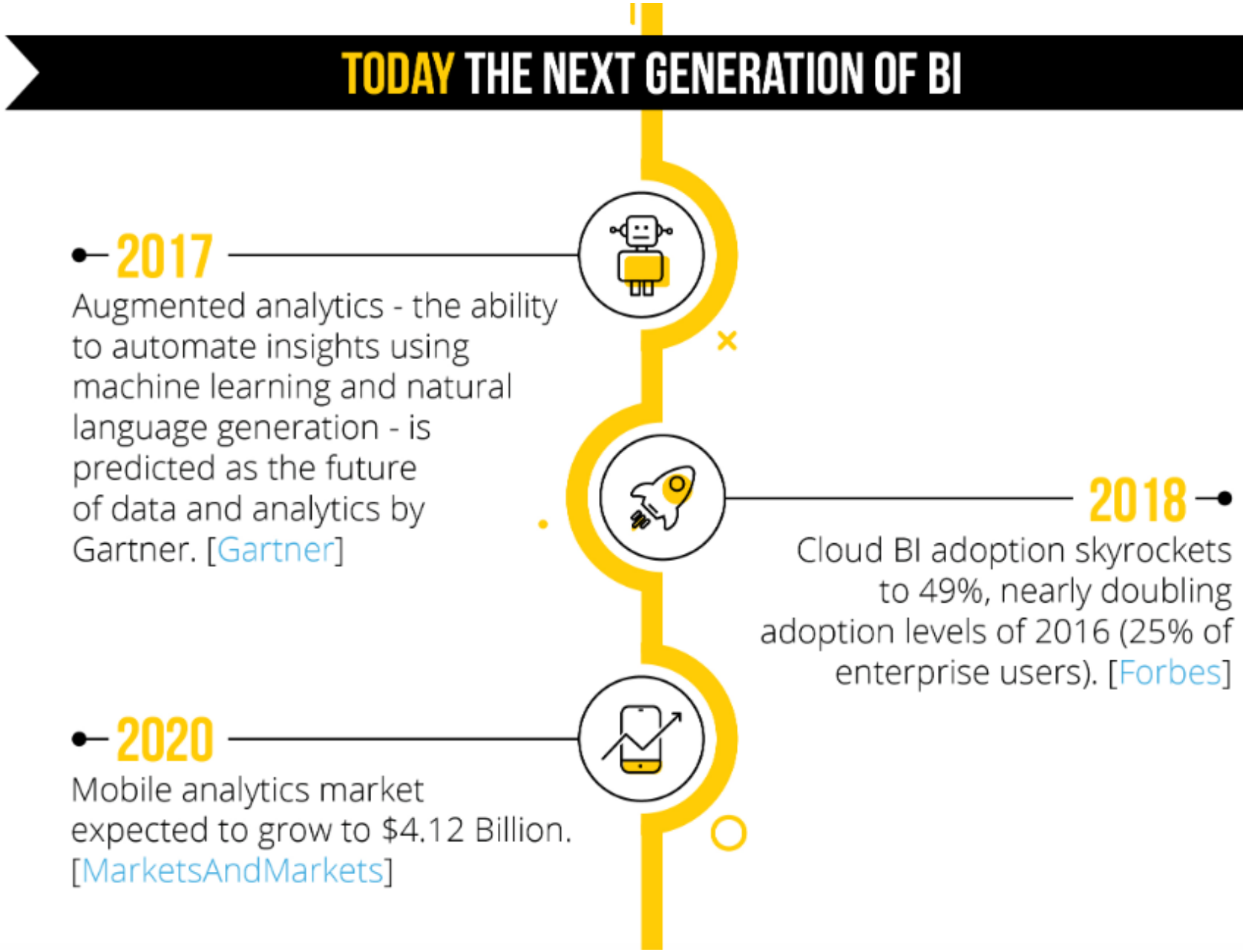
The first few BI vendors pop up with tools that are meant to help in accessing and organizing data. [BetterBuys]

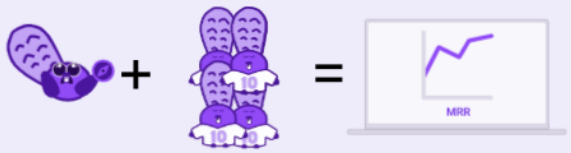
1980'S-1990'S THE FIRST GENERATION OF BI




EARLY 2000'S THE SECOND GENERATION OF BI
2005
With social media platforms like Facebook and Twitter on the rise, the amount of data created starts skyrocketing.
2008
Business intelligence, analytics and performance management revenue reaches $8.8 Billion. [Gartner]
2010
35% of organizations employ pervasive BI. 67% of "best in class" companies have some self-service BI. [Information Builders]

TODAY THE NEXT GENERATION OF BI



Traditional BI	Self-service BI	Augmented Analytics
Business user gathers requirements for a report or a dashboard.	Data team gathers user requests for self-service platform	Business user defines his data needs for a given report
User submits request to data team	Self-service tool implemented, giving business users access to data.	User creates a self service data pipeline
Data team extracts the data & loads it into a data warehouse	Business user directly accesses data	The automated pipeline extracts data from various sources and transforms it into the right format
Data team creates a data model	Business user selects which data to include	The automated pipeline loads data into the data warehouse
User approves report or dashboard, or requests changes. 	User builds his own report/ dashboard 	User builds his own report/ dashboard 

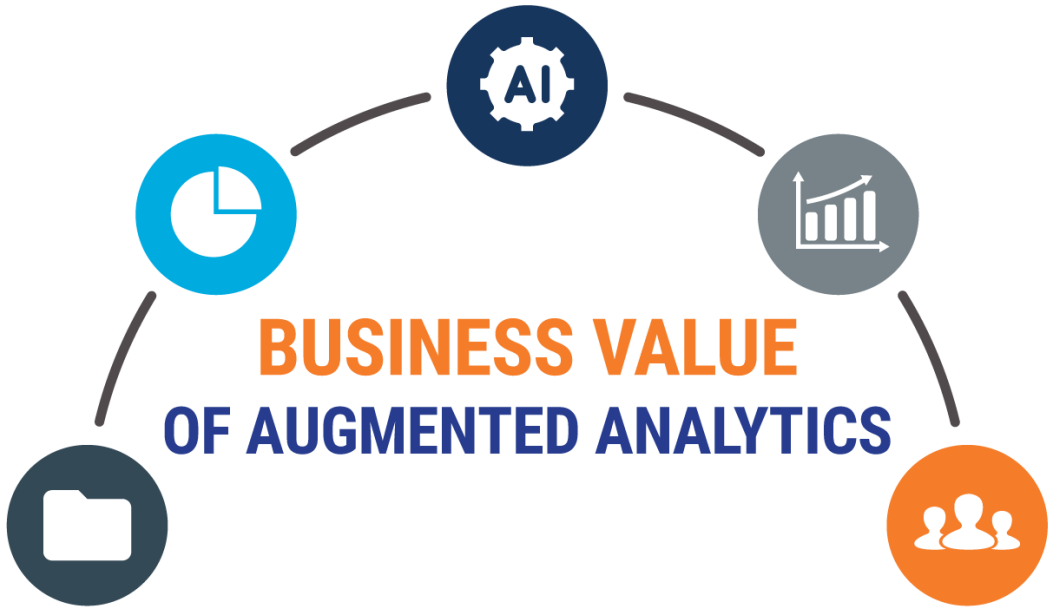
Augmented Analytics

REDUCED BIAS

AI-based platforms provide a more thorough data-capture process, ensuring that insights take all possibilities into account.

AUTOMATED DATA VISUALIZATION

Data visualization solutions offered by augmented analytics allows users to view and analyze information, making it easy to identify a problem and its root cause and start moving toward a solution.



DEMOCRATIZED BI

Augmented analytics democratize business intelligence by allowing non-technical users to access capabilities that previously required sophisticated data science skills.

ENSURES DATA ACCURACY–AT SCALE

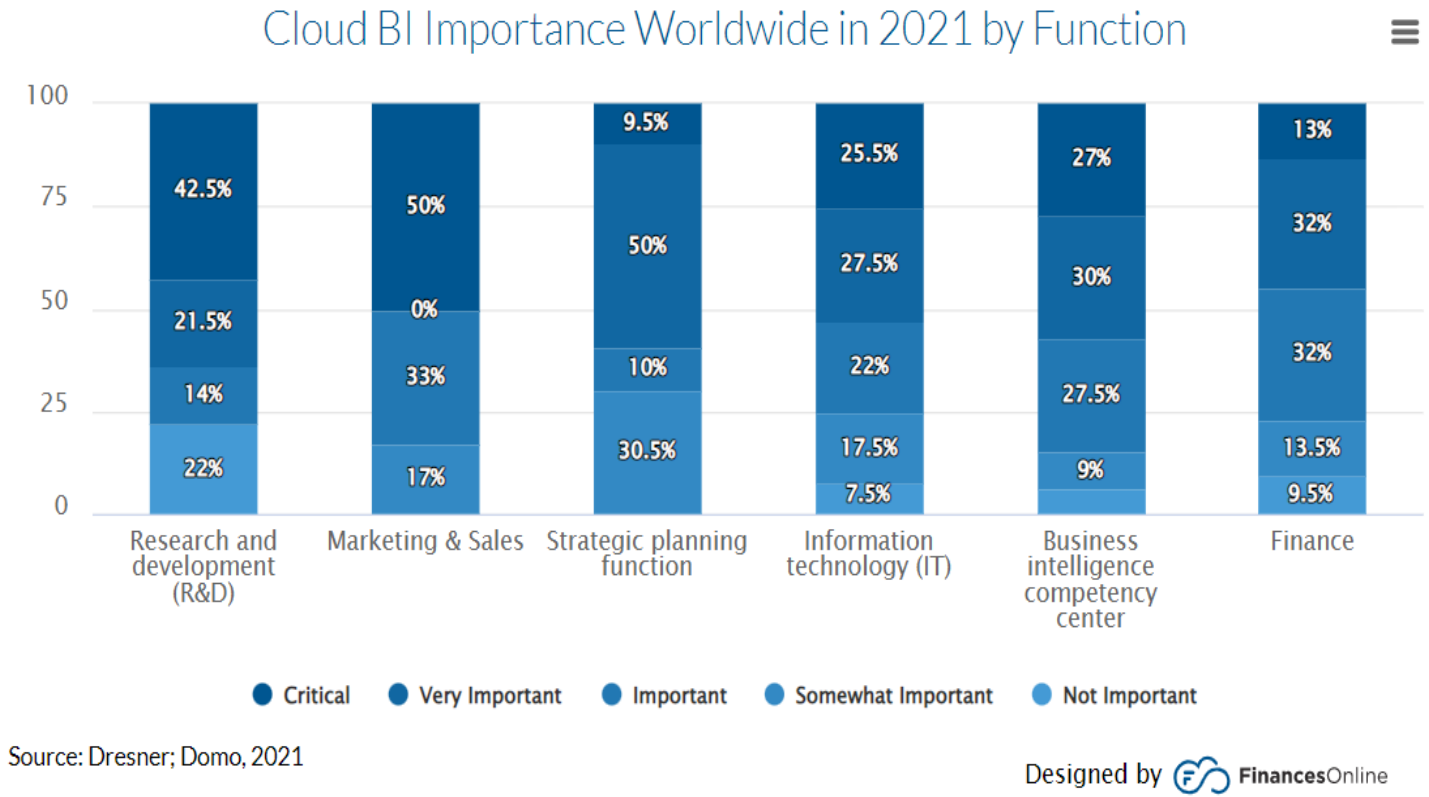
By applying machine learning to data management processes, organizations can reduce instances of error–thus ensuring that all decisions are made based on accurate information.

SUPPORT FOR DATA SCIENTISTS & IT

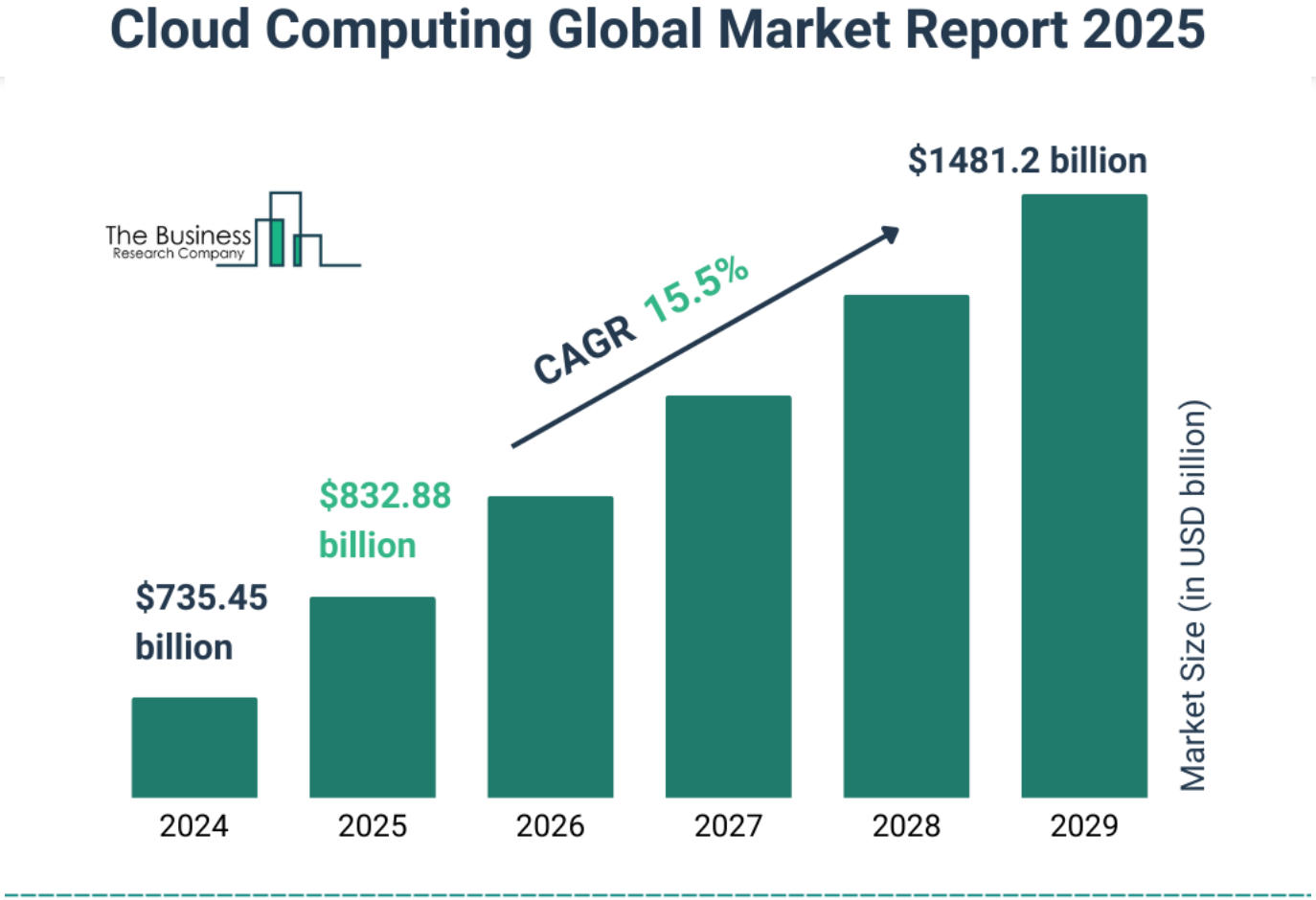
Augmented analytics enables both teams to focus on higher-value tasks. IT teams will have more time to focus on hardware and connectivity support, while data scientists can look for ways to capture even deeper insights.

Rise of Cloud based BI platforms

- Cloud-based business intelligence, or cloud BI, describes the process of transforming data into actionable insights either *partially or fully* within a **cloud environment**.
- Get the information they need to make data-driven decisions without the cost or hassle of physical hardware.



Cloud Computing Global Market



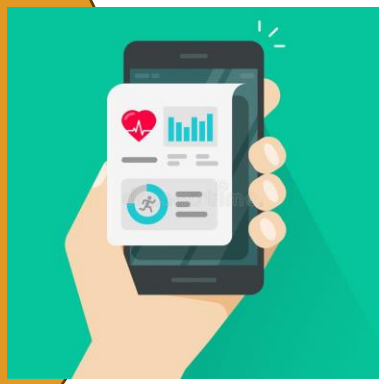
*CAGR: Compound annual growth rate

Mobile BI

- *Mobile Business Intelligence (BI) refers to the ability to access and perform BI-related data analysis on mobile devices and tablets.*
- Easier to display KPIs, business metrics, and dashboards.
- Portable screen - same features and capabilities as desktop/app-based BI software.

Where do you think this will be most useful?

Healthcare: A doctor using a mobile app to access patient records, monitor vital signs in real-time, and make informed treatment decisions.

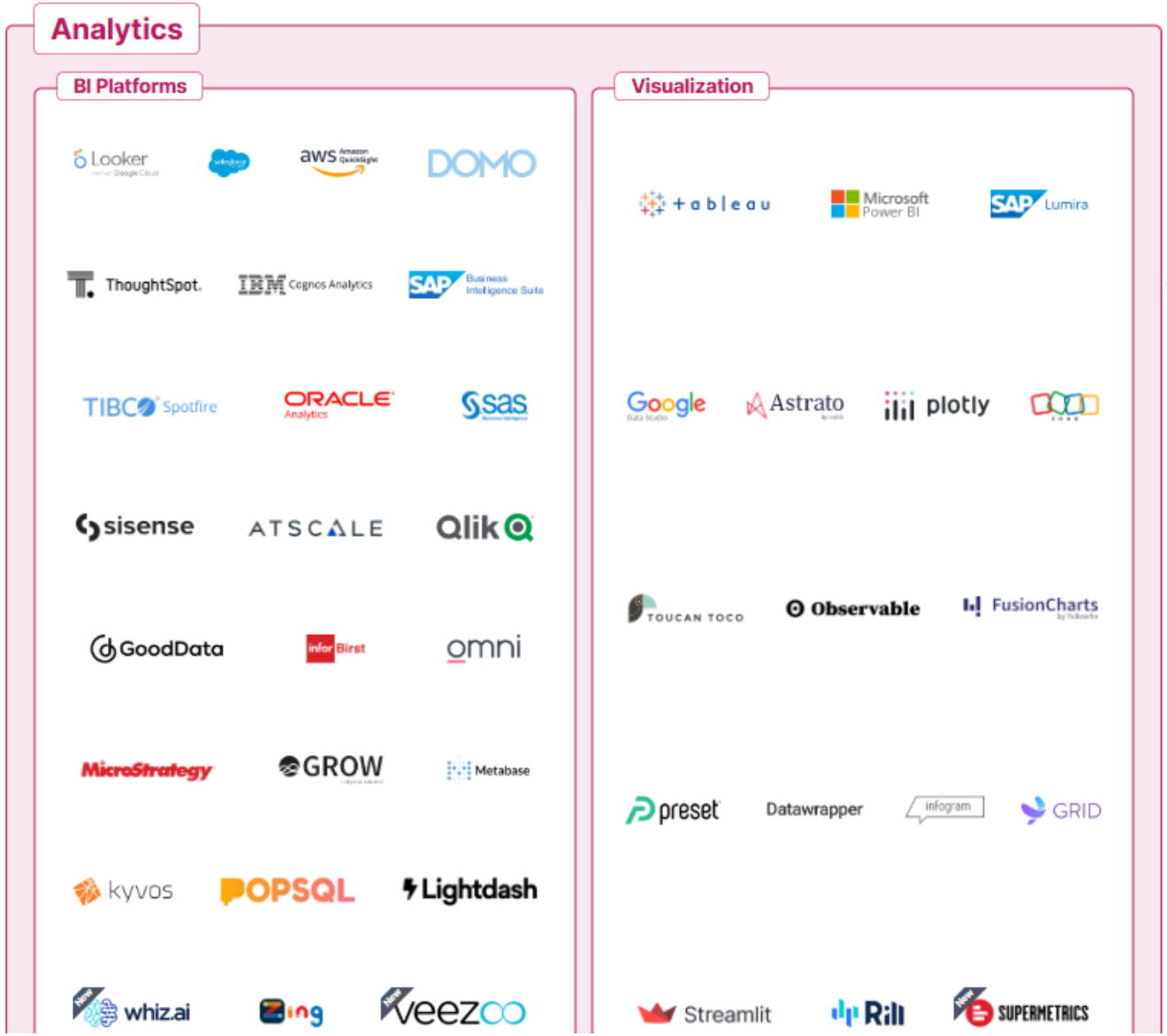


****And more…*

BI Tools


Microsoft
Power BI


+ a b | e a u


Analytics
BI Platforms
Looker
aws
DOMO
ThoughtSpot
IBM Cognos Analytics
SAP
TIBCO Spotfire
ORACLE Analytics
sas
sisense
ATSCALE
Qlik
GoodData
Birst
omni
MicroStrategy
GROW
Metabase
kyvos
POPSQL
Lightdash
whiz.ai
Bing
Veezoo
Visualization
+ a b l e a u
Microsoft Power BI
SAP Lumira
Google
Astrato
plotly
TOUCAN TOCO
Observable
FusionCharts
preset
Datawrapper
infogram
GRID
Streamlit
Rill
SUPERMETRICS

Why use BI tools?

- **Understand** the business
- **Improve** the business.
- *Informed data-driven decision making*

Advantages of using BI Tools

Activity Next class: Pre-work

- Select 2 online stores
  - (of different types with at least one store also having a physical presence)
    (e.g., Amazon, Daraz, Foodpanda, Imtiaz, Naheed, or any other ecommerce site etc.)
- Form groups of 3 or 4
- Investigate the site – take screenshots and put down your observations