DBW624 – Lecture 1
Introduction to Data Warehousing

Basic Concepts

- Companies run their businesses with OLTP (On-line Transaction Processing) systems – also known as "Systems of Record"
    - Sales support, order tracking, banking actions, customer support, etc

- Companies want to know things about their business (perform analytics)
    - Report on results
    - Historical analysis
    - Identify trends
    - Predict future results

- This can be accomplished in a couple of ways:
    - 1. Directly against the OLTP system
    - 2. Building a dedicated (Data Warehouse) system

- This course is primarily focused on #2.

- What is a Data Warehousing ?
    - Data Warehousing is a system used for reporting and data analysis
    - Data Warehouses are a central repository for data from one or more sources
    - Data Warehouses store current and historical data

Basic Concepts

- Terms used with Data Warehousing ?
  - Warehousing
  - Analytics
  - Operational Analytics
  - Mining
  - Enterprise Data Warehouses (EDW)
  - Operational Data Store (ODS)
  - Physical Data Marts
  - Logical Data Marts
  - On-line Analytic Processing (OLAP)

- Raw Data vs Business Information

- ETL and ELT (Extract Transform Load and Extract Load Transform)
  - How data moves in the infrastructure which includes a Data Warehouse

- DSS (Decision Support System)
  - Turning data into information

Basic Concepts

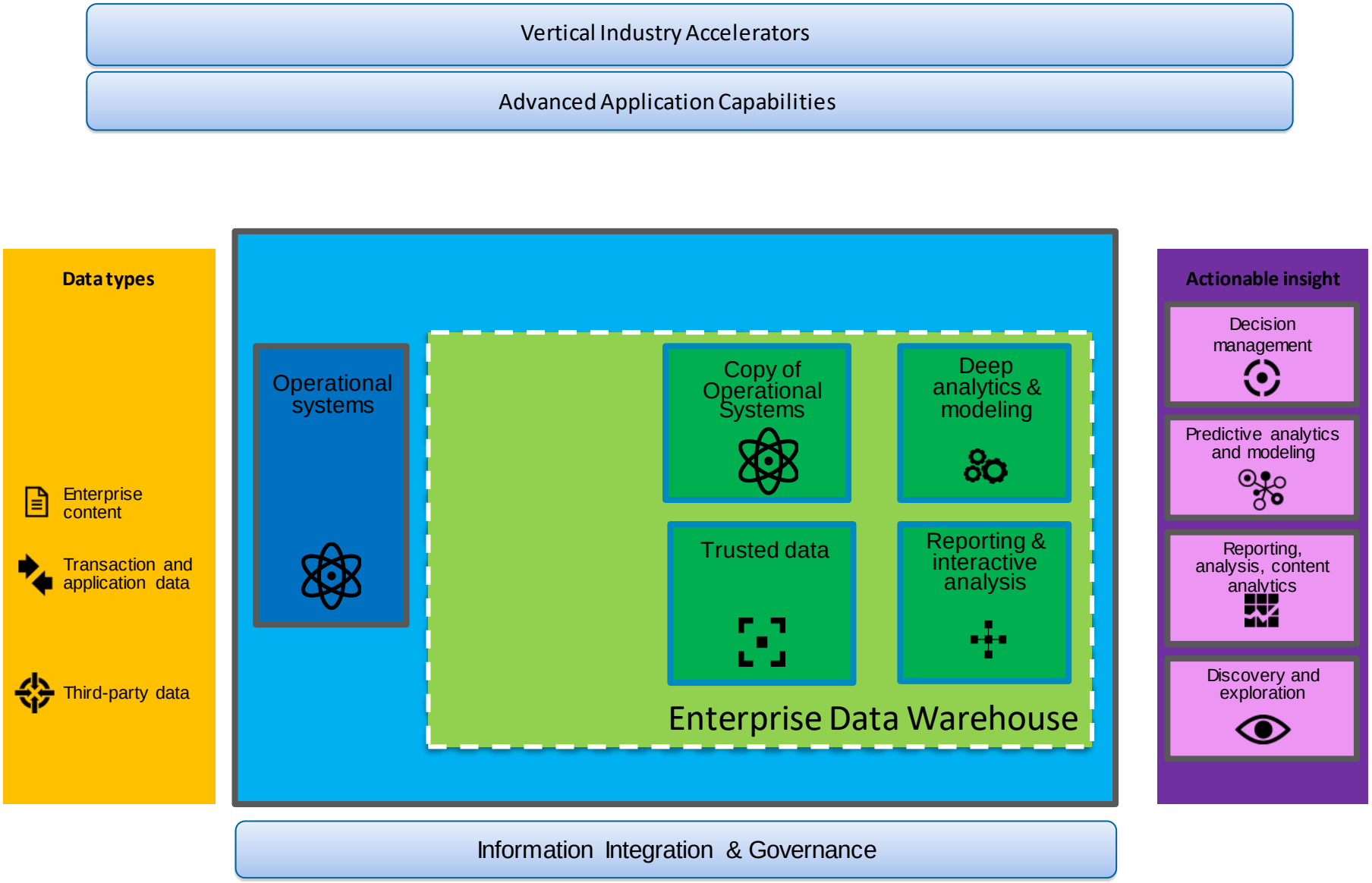
- Mixed workloads vs dedicated analytic platforms
  - Benefits of optimizing systems for a particular workload

- Data Warehouse characteristics
  - Batch based vs continual data ingest
  - Cleansed
  - Re-structured
  - Optimized for reporting, querying, analytics
  - Organizes data into non-volatile, subject-specific groupings
  - Multiple data sources

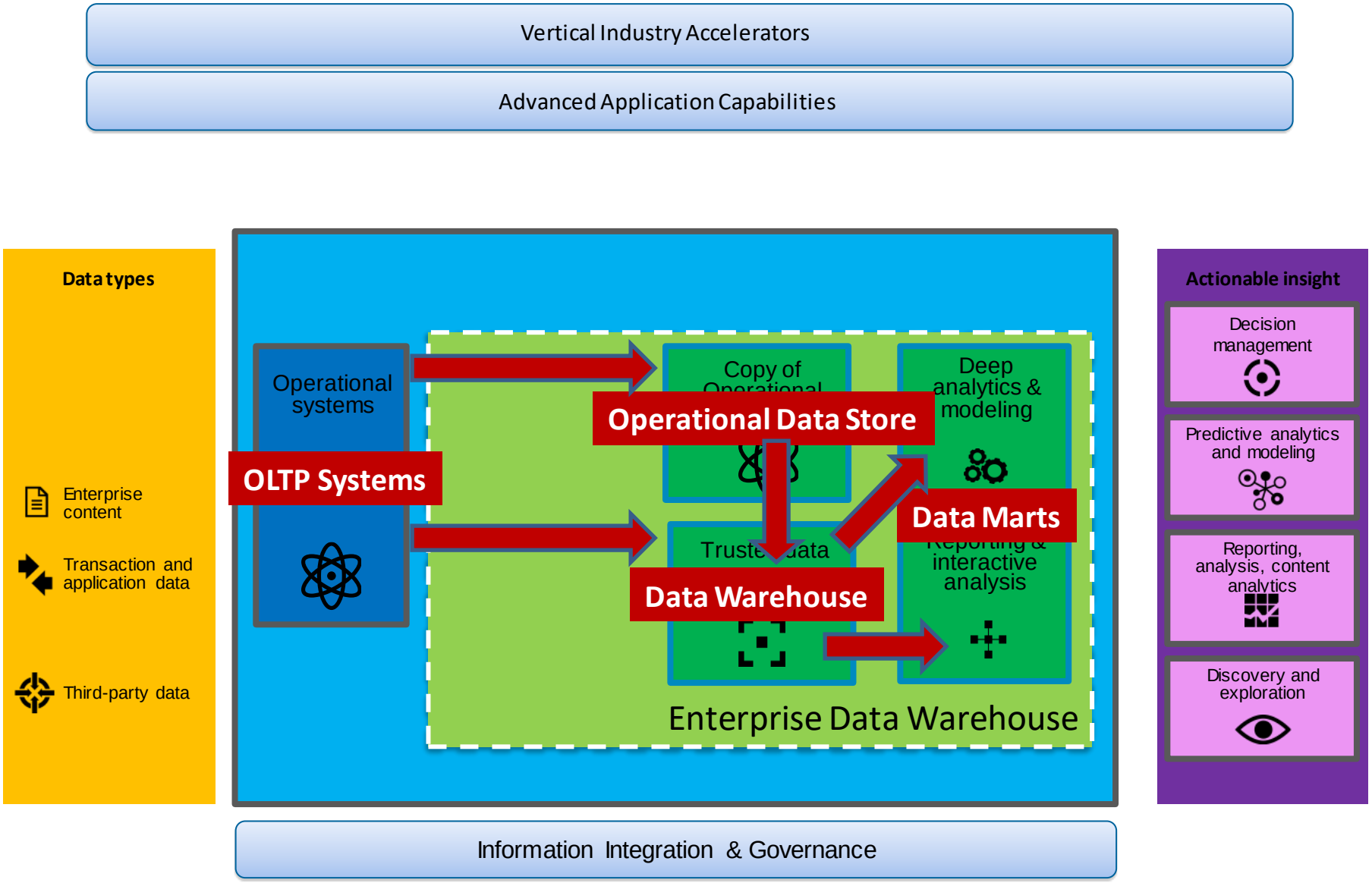
- On-line Analytic Processing (OLAP)
  - Relational OLAP (ROLAP)
  - Multi-dimensional OLAP (MOLAP)

- Facts, Dimensions, Star-schema, Snowflake, Hierarchies, Cubes

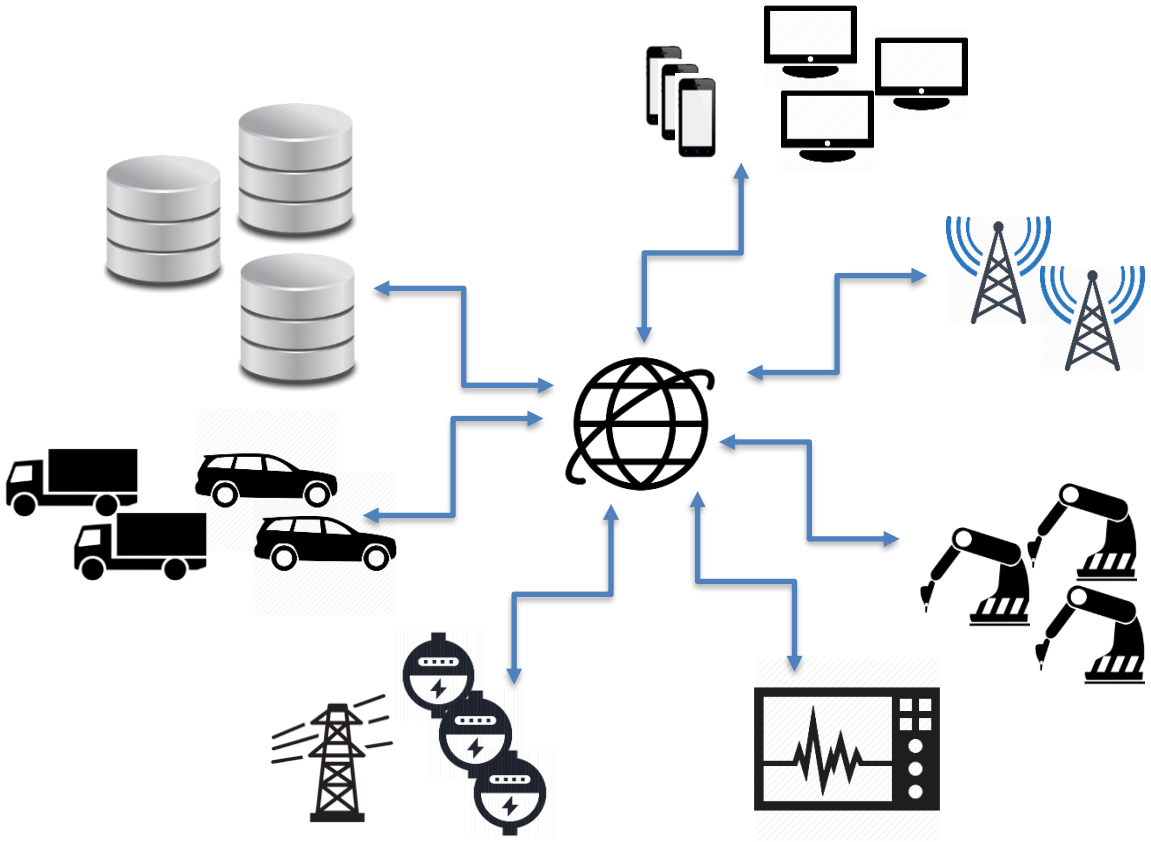
Analytics Platform: Traditional Components



Analytics Platform: Traditional Components



Data is Everywhere



Businesses are Becoming Data Driven

Warehousing Landscape Being Re-Shaped

Big Data – Impact Across The Market Place

The shift of power to the consumer

Creating the need for organizations to understand and anticipate customer behavior and needs based on customer insights across all channels

Disruptive Force in some industries

UBER

NETFLIX

Evolving every industry

Banking
- Optimizing Offers and Cross-sell
- Customer Service and Call Center Efficiency

Insurance
- 360˚ View of Domain or Subject
- Catastrophe Modeling
- Fraud & Abuse

Telco
- Pro-active Call Center
- Network Analytics
- Location Based Services

Energy & Utilities
- Smart Meter Analytics
- Distribution Load Forecasting/Scheduling
- Condition Based Maintenance

Media & Entertainment
- Business process transformation
- Audience & Marketing Optimization

Retail
- Actionable Customer Insight
- Merchandise Optimization
- Dynamic Pricing

Travel & Transport
- Customer Analytics & Loyalty Marketing
- Predictive Maintenance Analytics

Consumer Products
- Shelf Availability
- Promotional Spend Optimization
- Merchandising Compliance

Government
- Civilian Services
- Defense & Intelligence
- Tax & Treasury Services

Healthcare
- Measure & Act on Population Health Outcomes
- Engage Consumers in their Healthcare

Automotive
- Advanced Condition Monitoring
- Data Warehouse Optimization

Chemical & Petroleum
- Operational Surveillance, Analysis & Optimization
- Data Warehouse Consolidation, Integration & Augmentation

Aerospace & Defense
- Uniform Information Access Platform
- Data Warehouse Optimization

Electronics
- Customer/ Channel Analytics
- Advanced Condition Monitoring

Life Sciences
- Increase visibility into drug safety and effectiveness

Digital Disruption Is Upon All of Us…

World's Largest Accommodations Company
Owns No Real Estate

World's Largest Taxi Company
Owns No Vehicles

World's Largest Retailer
Carries No Inventory

World's Largest Media Company
Creates No Content

Market Observations

1. There is increasing pressure to perform analytics where data gets created
   "Point-of-decision HTAP promises to simplify the information infrastructure by removing unnecessary data marts and, potentially, data warehouses." – Gartner

2. Event-driven applications will enable new analytic use cases
   "Event-driven real-time digital business is poised to become a priority for mainstream business " Gartner
   "In-process HTAP could potentially redefine the way some business processes are executed" Gartner

3. Business applications are leveraging both *SQL and NoSQL* data in structured repositories for analytics
   "Top relational database solutions are now offering a wide range of new features to combine structured and unstructured data types " …. Database decision-makers need to look at investing in these database technologies. – Forrester

4. *Hybrid cloud* capabilities of software support economies of scope
   Public cloud adoption has stalled for the time being, signaling enterprises are moving to the hybridization phase of their IT transformations. TBRI 2H 2016

5. Private cloud needs cloud-scale convenience
   IDC "by end of 2016 38% of the IT Market spend will be private hosted or private on Prem Cloud with On-Demand Convenience future growth point within private cloud. Skills, timing or cost to effectively procure, assemble, run, manage disperse infrastructure resource require integrated versatile platform offerings with appliance-like simplicity" – A client

6. *Diverse data sources* support an ecosystem of innovation
   Established vendors…, have continued their cloud-focused innovation around hybrid cloud for both cost and workload optimization. Many have added open-source products to their portfolio — usually by acquisition — in an attempt to capture a new generation of buyers .

Some Interesting Data Points

Data is proliferating, often stored in different locations and formats.
It's getting more difficult to provide data access and analytics to the business.

15% of organizations fully leverage data and analytics — Forbes

0.5% of all data is actually analyzed — MIT Technology review

80% of all data is stored by corporations — Baseline Magazine

50% of large enterprises will had hybrid cloud deployments by the end of 2018 — IBM Institute for Business Value

10% increase in data accessibility will result in more than $65 M additional net income — Baseline Magazine

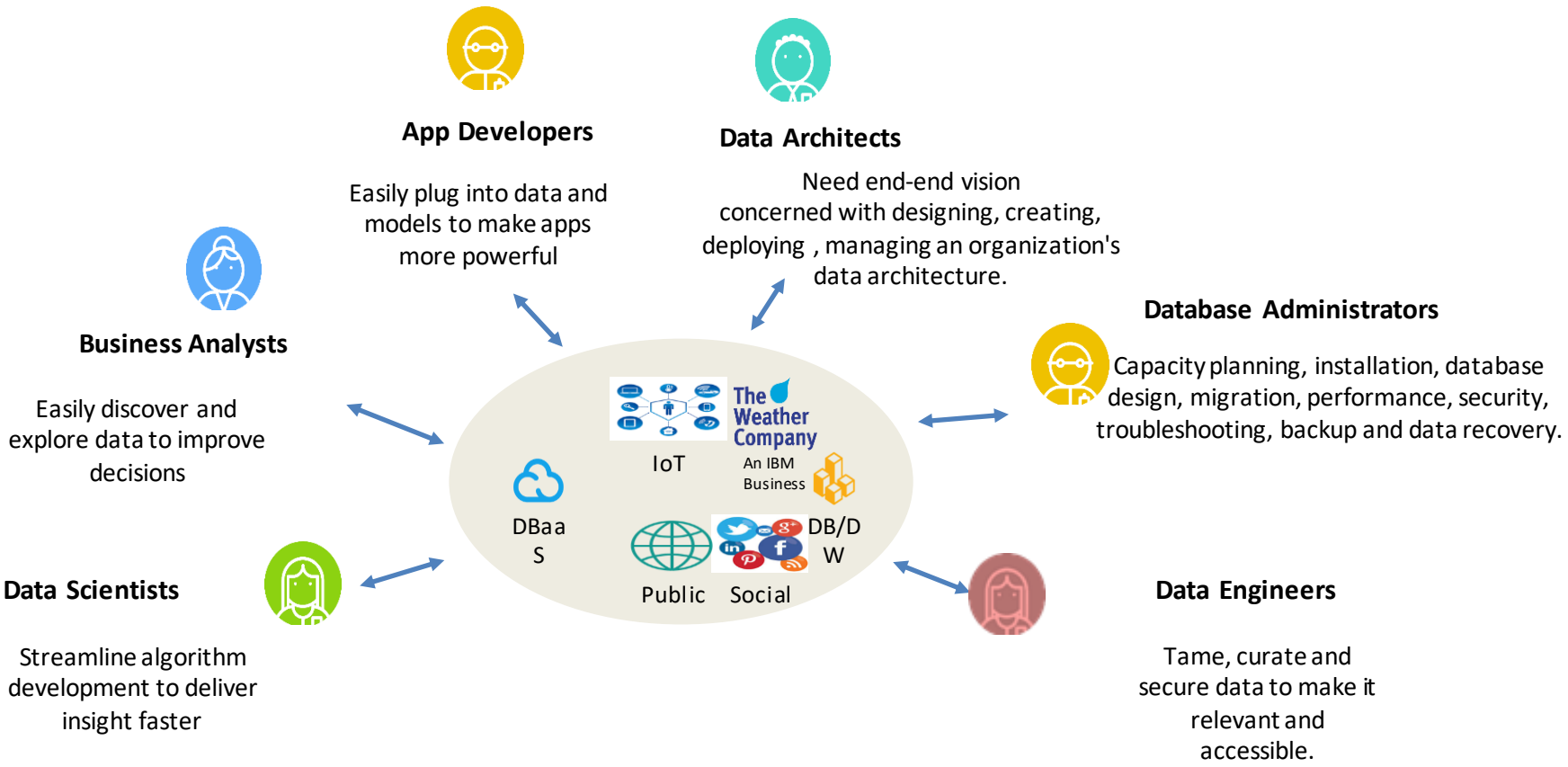
76% of the FTSE 100 companies have been replaced since 1984

81% believe AI to be very important or extremely important to the future of their organizations

80% of enterprise data sits behind the firewall

Data Professionals – Needs are Evolving

As data maturity increases, so does the number of data professionals who are hungry to put data to work



Hybrid – NOT “One Size Fits All”

NOT about Cloud **OR** On-premises	It’s about Cloud **AND** On-premises
NOT about traditional relational **OR** open source	It’s about traditional relational **AND** open source
NOT about SQL **OR** NoSQL	It’s about SQL **AND** NoSQL
NOT about structured **OR** unstructured data	It’s about structured **AND** unstructured data
NOT about data at rest **OR** data in motion	It’s about data at rest **AND** data in motion

Cloud Computing – The Value

30% to 50% of all servers within a typical IT environment are dedicated to test

Most test servers run at less than 10% utilization, if they are running at all!

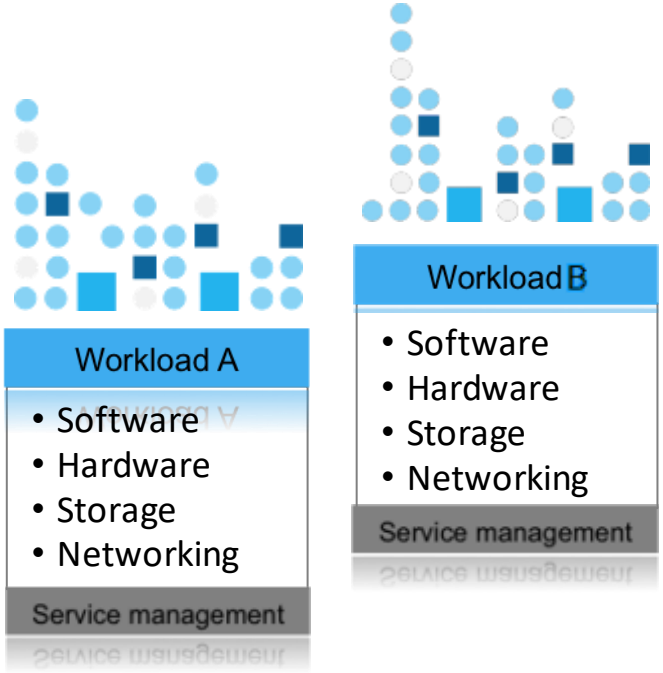
In distributed computing environments, up to 85% of computing capacity sits idle.

Silos of people, process, and projects

Complex Infrastructure	High Costs	Chaos
▪ Lengthy on-boarding	▪ Low utilization rates	▪ Weak project governance
▪ Acquiring, installing, configuring and managing environments	▪ Cost inefficiencies	▪ Lack of domain expertise
	▪ Poor LOB oversight	▪ Inflexible tooling integration
		▪ Incompatible tools / repositories

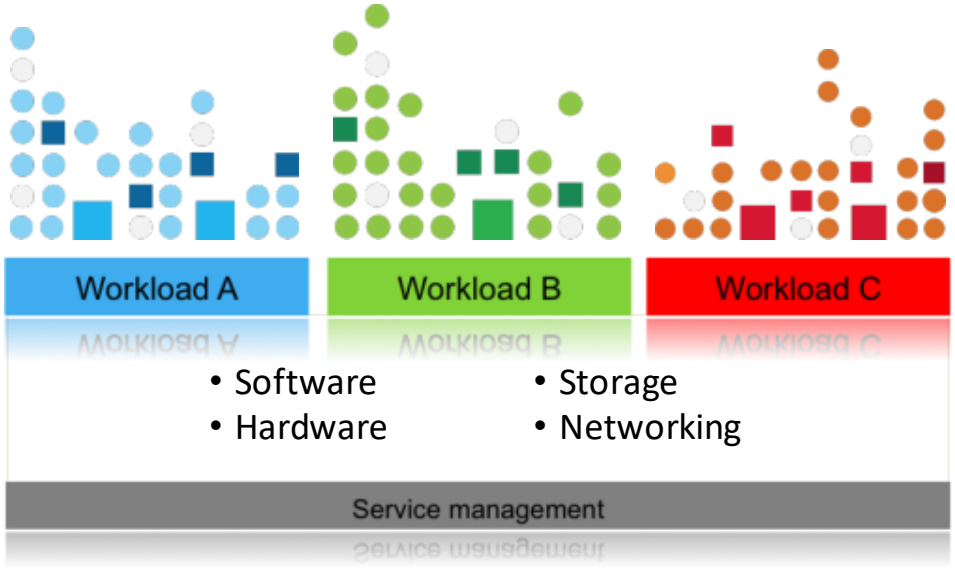
Cloud Computing - Outsourcing

Without cloud computing



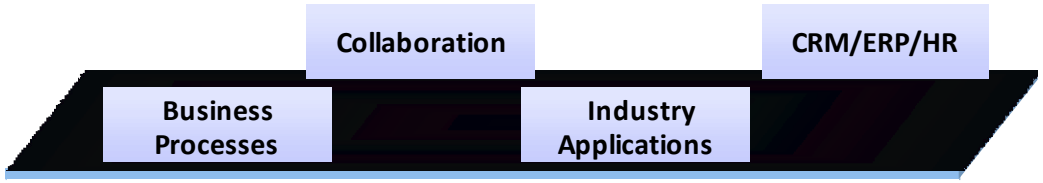
With cloud computing

- Virtualized resources
- Automated service management
- Standardized services
- Location independent
- Rapid scalability
- Self-service



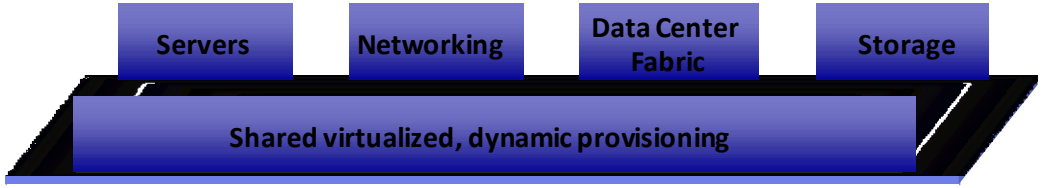
Note: Elements of cloud computing taken from NIST, Gartner, Forrester and IDC cloud computing definitions

Cloud Computing - Different Types of Services



Software as a Service

Platform as a Service



Infrastructure as a Service

Cloud Computing – Moving to Multi-Cloud Strategy

Artificial Intelligence - Massive Driving Force

A prescriptive approach to accelerating your journey to AI

INFUSE - Automate and scale across your processes

TRUST - Achieve trust and transparency in outcomes

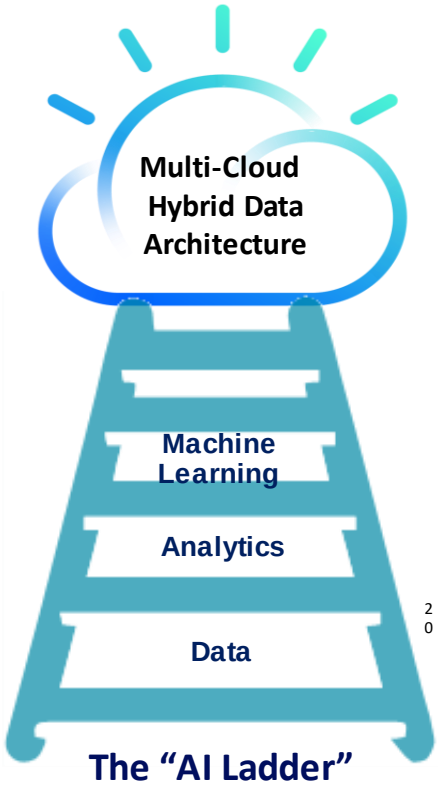
ANALYZE - Scale insights with ML everywhere

ORGANIZE - Create a trusted analytics foundation

COLLECT - Make data simple and accessible

Data of every type, regardless of where it lives

MODERNIZE your data estate for an AI and multi-cloud world



The "AI Ladder"

Artificial Intelligence - Inhibitors

AI adoption is Accelerating …. but there is a set of unique challenges

94%

of companies believe that AI is key to competitive advantage

AI associated with CRM activities will boost global business revenue by **$1.1T** from 2017 to 2021

Only **1 in 20** companies have extensively incorporated AI in offerings or processes

Top reasons for lack of AI Adoption

- **Skills**

Lack of requisite talent to drive AI adoption

- **Data**

Only 19% respondents strongly agreed that their organizations understand the data required to train AI algorithms. Data used is not of high quality or trusted.

- **Trust**

Only 35% of IT and Business decision makers had a high level of trust in their own organization's analytics. AI insights not well integrated into current processes

Artificial Intelligence - Inhibitors

49%

of C-level execs and IT Decision makers reported that their organization is unable to deploy the AI technologies they want because their data is not ready to support them.

Age of AI, Infosys survey is of 1,000+ c-level execs and IT decision makers from 500 to 5,000 employees

Fast Data – Internet of Things

Customers situation:

High Speed IoT Data

Data arriving faster than ever before
Value is in deriving fast and deep insights
High data volume requires efficient storage

Problems:

Can’t land data fast enough

Unable to derive insights fast enough
Existing architectures are too complex and expensive

Fast Data – Internet of Things

Digital business is event-driven, so organizations need to invest in event-centric design practices and technologies to exploit digital business moments. Enterprise architecture and technology innovation leaders must champion event thinking across business and IT.

Key Findings

- Legacy application architectures lack support for continuous innovation and global scale, both of which are essential in digital business.
- A well-designed event-driven model supports scalability, resilience and operational efficiency.
- Event-driven architecture is a natural fit for several digital business use cases, including real-time decision making, Internet of Things (IoT) initiatives and agile microservices design.
- Most organizations already use event-processing technology, but few take an event-driven architecture approach to application design.

Strategic Planning Assumptions

By 2022, event notifications will form part of over 60% of new digital business solutions.

By 2022, over 50% of business organizations will participate in event-driven digital business ecosystems.

By 2022, 50% of organizations managing APIs will incorporate mediation of event notifications into their operations.

By 2022, most leading providers of application platforms will include high-productivity tools for event-driven design.

Self Service

Self-service

= No dependency on IT

= Empowered end-user

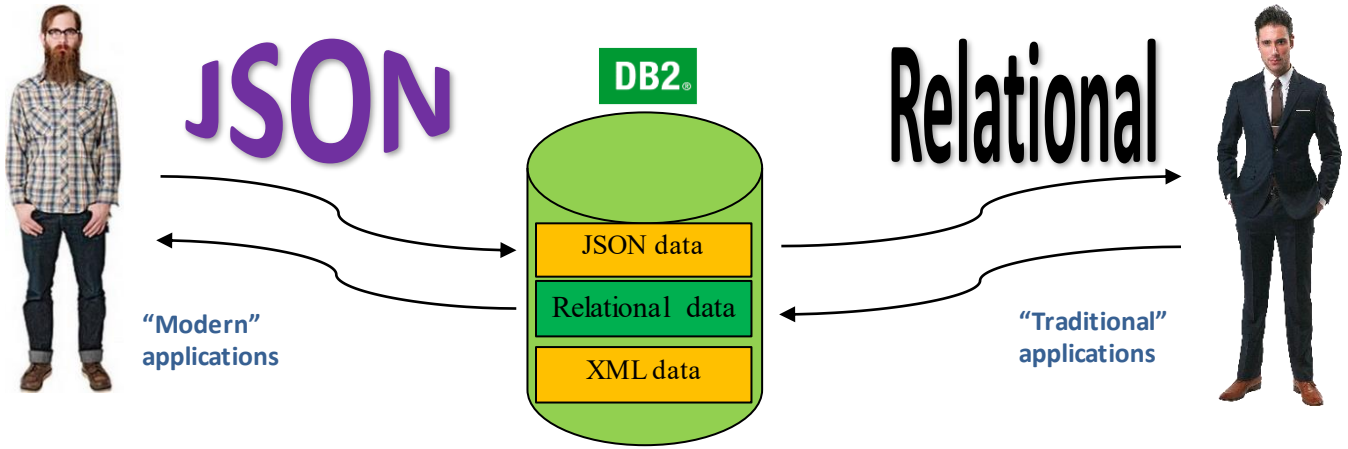
= Visibility into all data

= Access to the right data

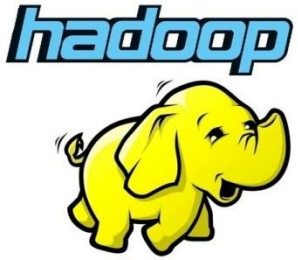
= Trust your data

= Availability of data

SQL and NOSQL (Not Only SQL)

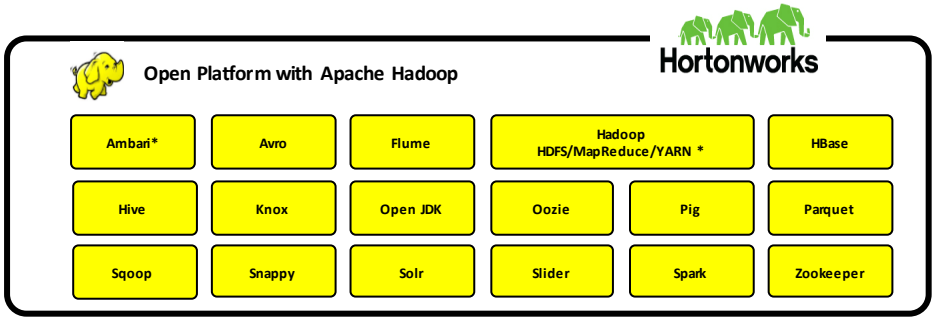


Key Vendors – Not Exhaustive



Key Vendors - Open Source

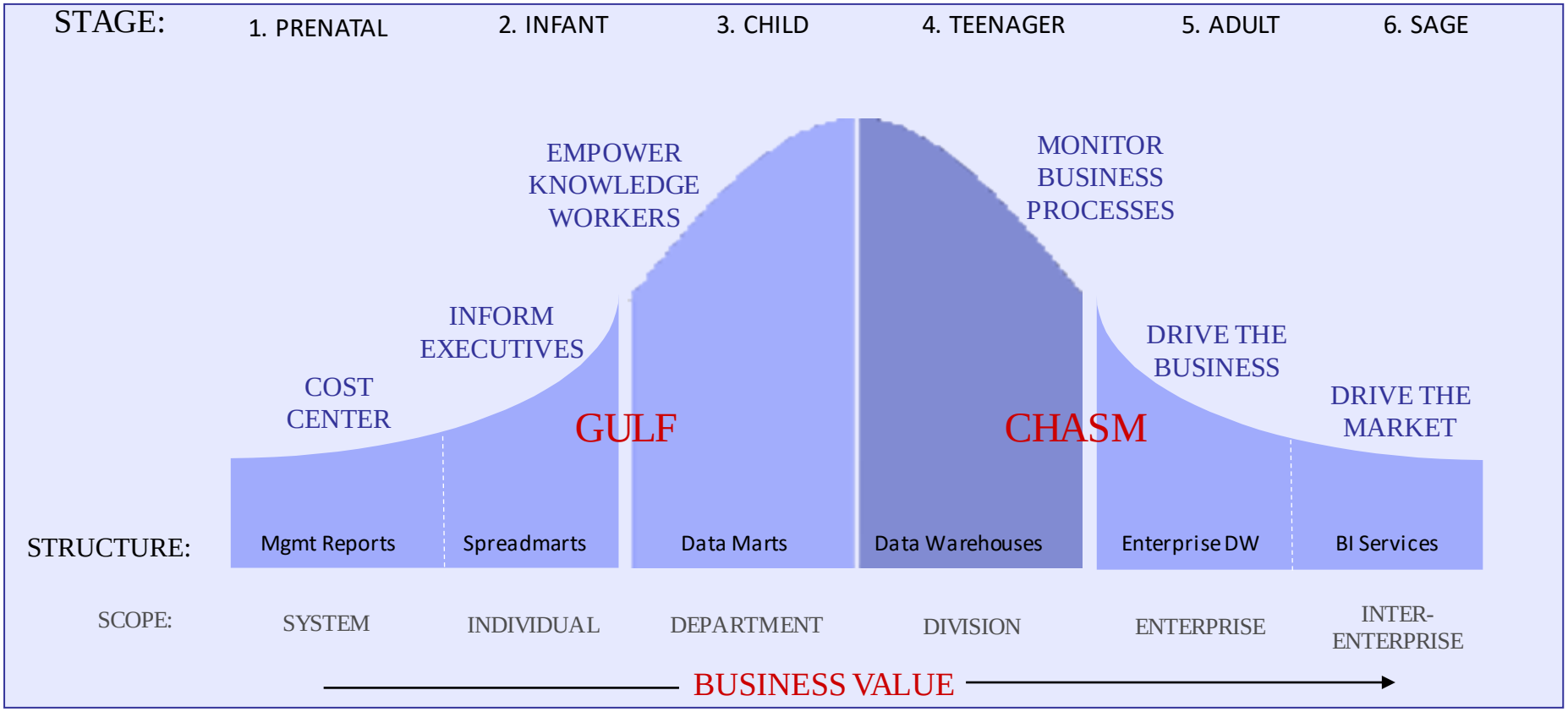
MongoDB	Redis	Elasticsearch	PostgreSQL	RethinkDB	RabbitMQ	etcd	MySQL	ScyllaDB
NoSQL (BSON) Document DB	Data Caching Key-Value DB	Full-Text Search Indexing Engine	Extensible and Secure Object Relational DB	Scalable JSON Database for Real-Time Apps	Asynchronous Messaging Layer	Key-Value Database for Distributed DBs	Extensible and Secure Relational Data Store	Distributed NoSQL data store



Spark

Hortonworks + cloudera

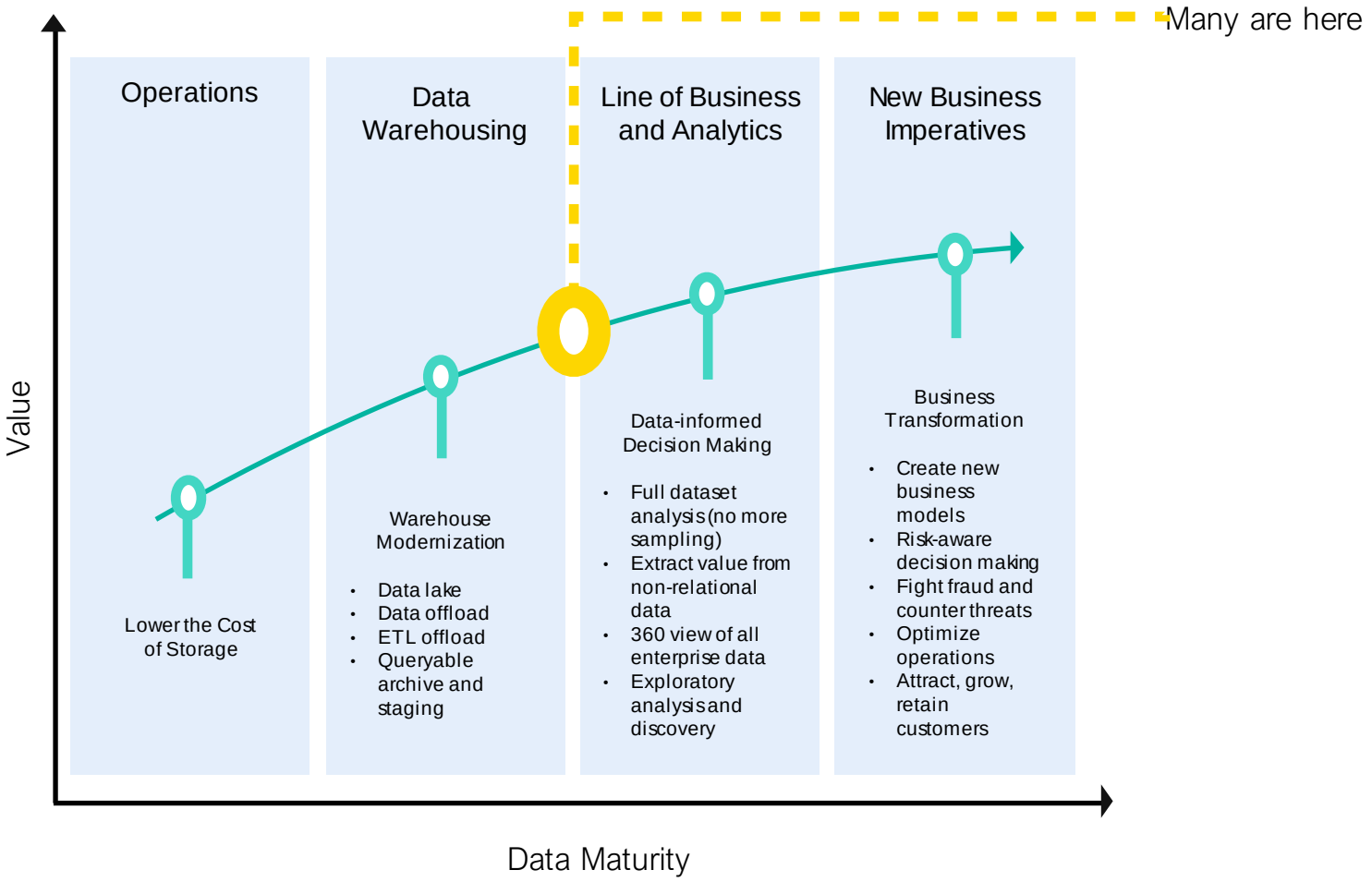
Data Warehousing Maturity Model



An ever increasing demand for enterprise information requires a decision to proceed further along the maturity model to meet business need.

Source: "Gauge Your Data Warehouse Maturity, Wayne Eckerson, DMReview, Nov 2004

Approach to Data is Evolving

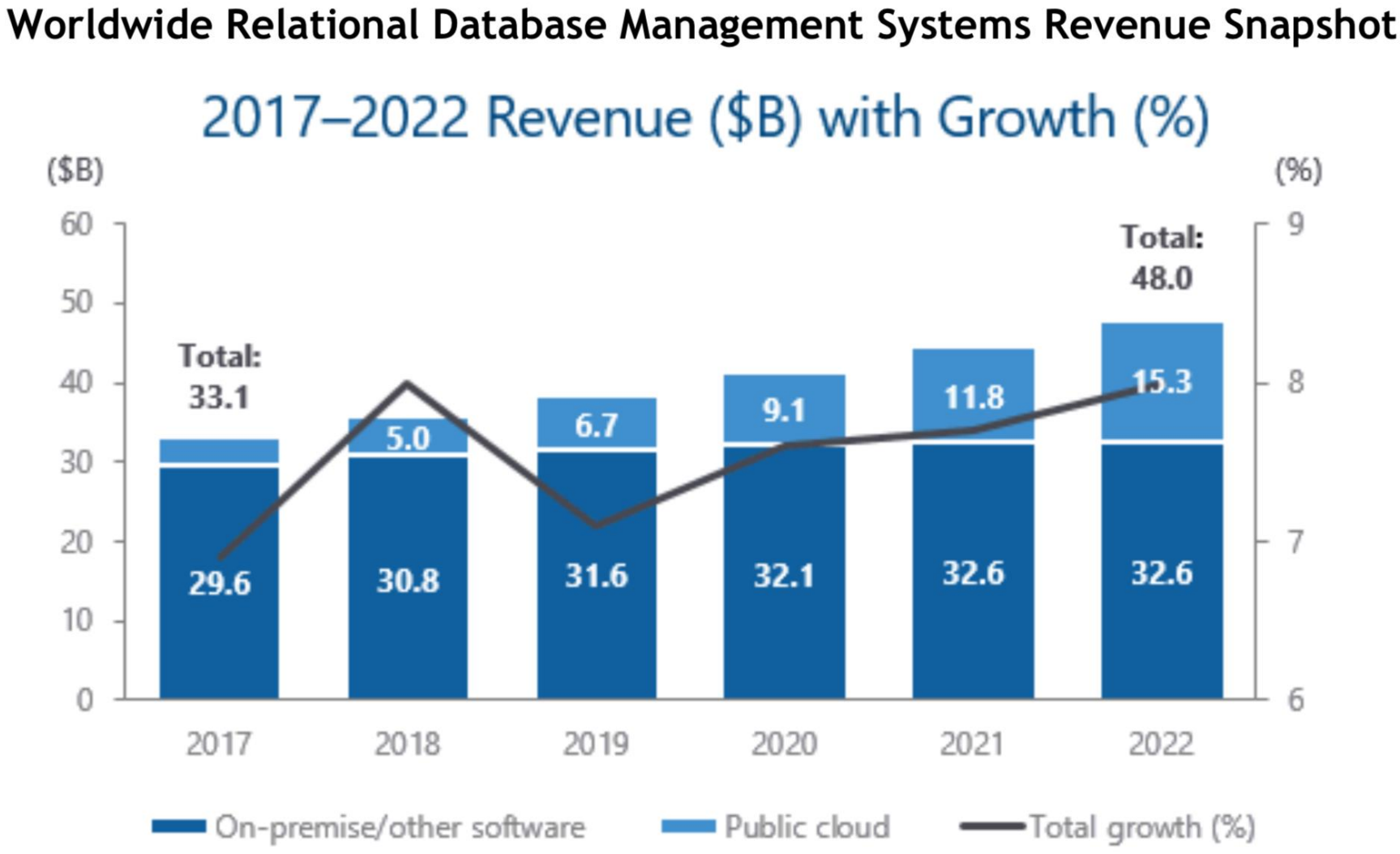


Analytics Revenue Spending

IDC 2018 forecasts attribution

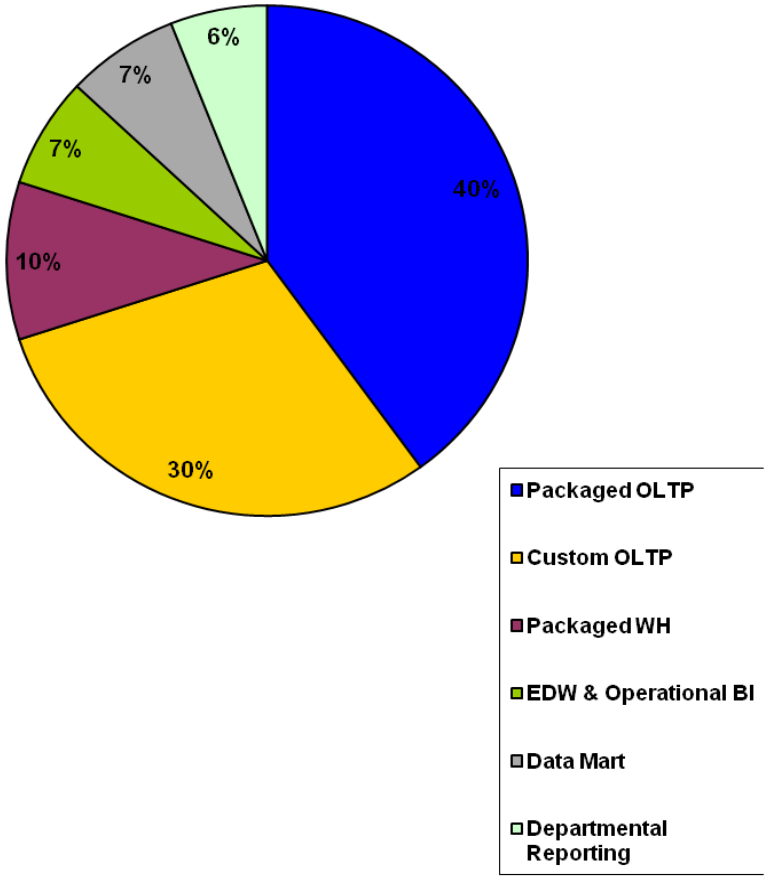
Column1	Column2							Column9	Column10
RDBMS		2018	2019	2020	2021	2022	2023	Growth 2019-2023	Cagr
Operational/Mixed	On Premise	$ 22.54	$ 22.53	$ 22.81	$ 23.16	$ 23.28	$ 23.31	$ 0.78	$ 0.15
	Cloud	$ 2.25	$ 3.31	$ 4.98	$ 7.03	$ 9.57	$ 12.54	$ 9.23	$ 205.46
	Total	$ 24.79	$ 25.84	$ 27.79	$ 30.19	$ 32.85	$ 35.84	$ 10.01	$ 2.70
RDBMS									
Pure Analytics (Deep Analytics/Data marts/ Operation	On Premise	$ 8.83	$ 9.02	$ 9.27	$ 9.32	$ 9.29	$ 9.28	$ 0.25	$ 0.12
	Cloud	$ 2.87	$ 3.78	$ 4.93	$ 6.18	$ 7.61	$ 9.22	$ 5.45	$ 34.50
	Total	$ 11.70	$ 12.80	$ 14.20	$ 15.50	$ 16.90	$ 18.50	$ 5.70	$ 3.36
Total RDBMS	On Premise	$ 31.37	$ 31.55	$ 32.07	$ 32.48	$ 32.57	$ 32.58	$ 1.03	$ 0.14
	Cloud	$ 5.12	$ 7.09	$ 9.92	$ 13.21	$ 17.18	$ 21.76	$ 14.68	$ 87.93
	Total	$ 36.49	$ 38.64	$ 41.99	$ 45.69	$ 49.75	$ 54.34	$ 15.71	$ 2.91
Nonrelational/Dynamic Analytics	On Premise	$ 0.9	$ 1.1	$ 1.1	$ 1.2	$ 1.3	$ 1.3	$ 0.27	$ 1.46
	Cloud	$ 2.5	$ 3.8	$ 5.3	$ 6.9	$ 8.6	$ 10.7	$ 6.83	$ 58.81
	Total	$ 3.4	$ 4.9	$ 6.4	$ 8.1	$ 9.9	$ 12.0	$ 7.10	$ 34.97
Non-Relational/Dynamic Other(dynamic DS)	On Premise	$ 0.49	$ 0.47	$ 0.57	$ 0.65	$ 0.74	$ 0.82	$ 0.34	$ 7.71
(includes document DBs bulk of revenue MS/Amazon)	Cloud	$ 1.36	$ 1.71	$ 2.62	$ 3.73	$ 5.08	$ 6.52	$ 4.81	$ 210.52
	Total	$ 1.85	$ 2.19	$ 3.19	$ 4.37	$ 5.82	$ 7.34	$ 5.15	$ 126.21
Total Dynamic Systems	On Premise	1.38	1.54	1.70	1.85	1.99	2.15	$ 0.61	$ 2.80
	Cloud	3.87	5.55	7.89	10.63	13.73	17.19	$ 11.64	$ 91.29
	Total	5.25	7.09	9.59	12.47	15.72	19.34	$ 12.25	$ 54.50

Relational Database Revenue Spending

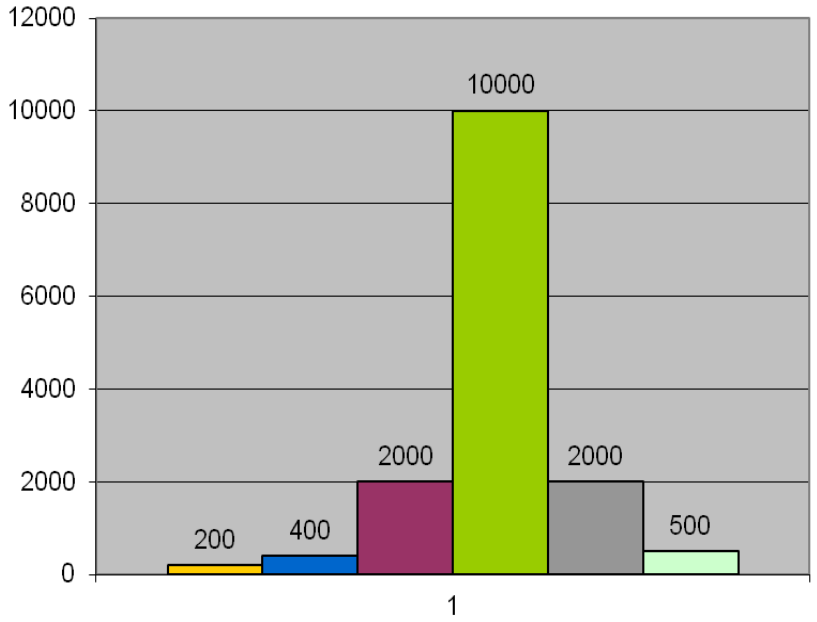


Traditional Database Market Overview

Database Market Breakdown



Average Database Size (GB)



Analytics Platform: New Components