Daimler TSS

A company of Daimler AG

LECTURE @DHBW: DATA WAREHOUSE

PART I: INTRODUCTION TO DWH AND DWH ARCHITECTURE

ANDREAS BUCKENHOFER, DAIMLER TSS

DHBW Duale Hochschule Baden-Württemberg

ABOUT ME

Andreas Buckenhofer
Senior DB Professional
andreas.buckenhofer@daimler.com

Since 2009 at Daimler TSS
Department: Big Data
Business Unit: Analytics

in https://de.linkedin.com/in/buckenhofer

https://twitter.com/ABuckenhofer

DOAG https://www.doag.org/de/themen/datenbank/in-memory/

DHBW Duale Hochschule Baden-Württemberg http://wwwlehre.dhbw-stuttgart.de/~buckenhofer/

XING https://www.xing.com/profile/Andreas_Buckenhofer2

- 1995 **Datenbanken**
  - Admin
  - Tuning
  - Entwicklung
  - Modellierung
- 1998 **Oracle DB**
  - ab V7.3.4
- 1998 **DWH / CRM**
  - Architektur
  - Integration
  - Reporting
- 2004 **RZ Migration**
  - 2 Jahre UK
- 2006 **Siebel CRM**
  - ab V7.8
- 2006 **Replication / CDC**
  - Oracle CDC
- 2007 **SQL Server**
  - ab V2005
  - SSIS
- 2009 **Data Vault**
- 2010 **Cognos**
  - ab V8.4
- 2012 **DB2**
  - ab V9.7
- 2013 **Hadoop**
  - Architektur
  - Datenintegration
  - SQL-on-Hadoop
- 2015 **IIDR**
  - 10.2.1

NOT JUST AVERAGE: OUTSTANDING.

As a 100% Daimler subsidiary, we give 100 percent, always and never less. We love IT and pull out all the stops to aid Daimler's development with our expertise on its journey into the future.

Our objective: We make Daimler the most innovative and digital mobility company.

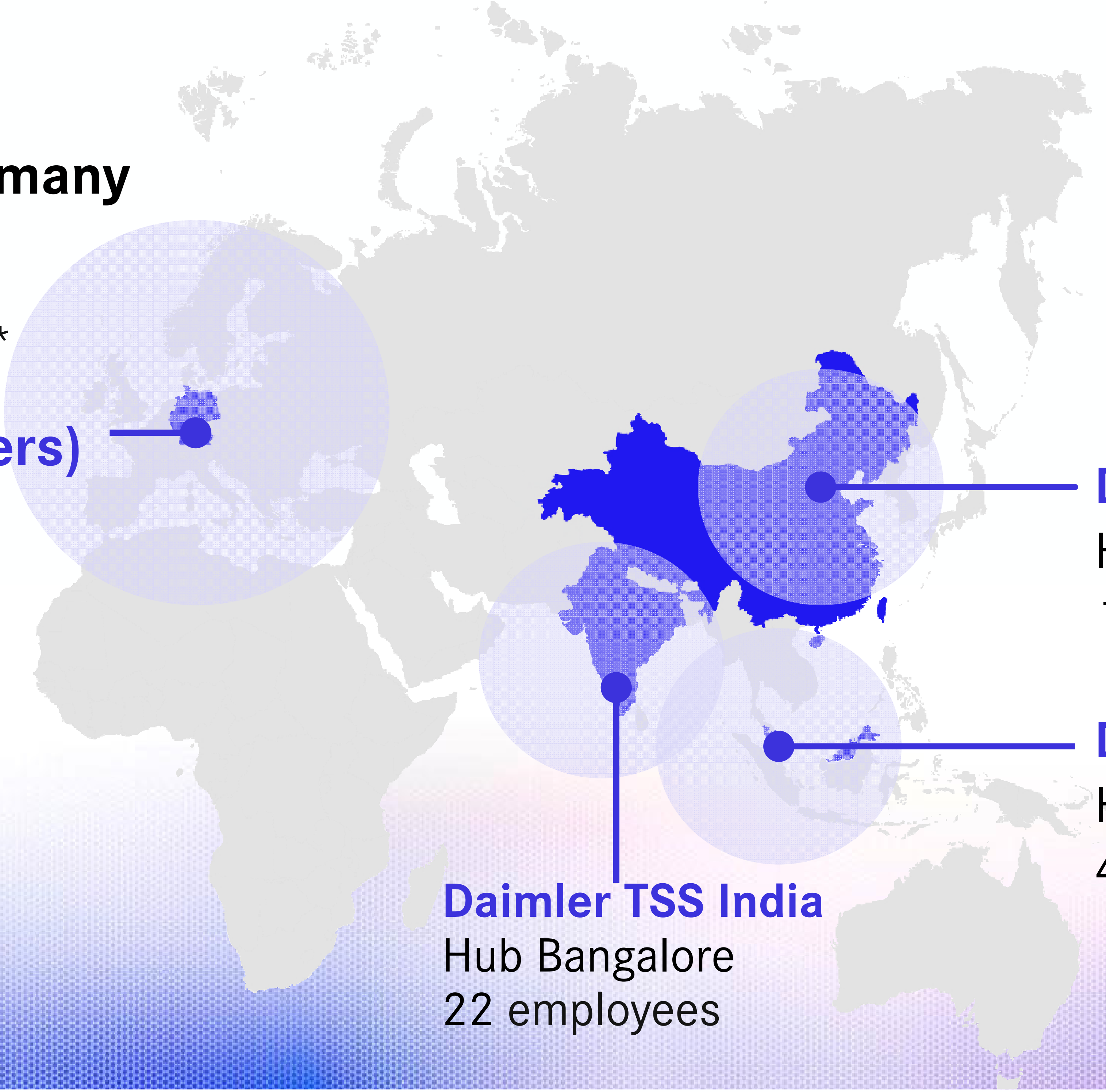
INTERNAL IT PARTNER FOR DAIMLER

- Holistic solutions according to the Daimler guidelines
  - IT strategy
  - Security
  - Architecture
- Developing and securing know-how
- TSS is a partner who can be trusted with sensitive data

As subsidiary: **maximum added value** for Daimler

- Market closeness
- Independence
- Flexibility (short decision making process, ability to react quickly)

LOCATIONS



Daimler TSS Germany

7 locations

1000 employees*

Ulm (Headquarters)

Stuttgart

Berlin

Karlsruhe

Daimler TSS China
Hub Beijing
10 employees

Daimler TSS Malaysia
Hub Kuala Lumpur
42 employees

Daimler TSS India
Hub Bangalore
22 employees

* as of August 2017

DWH, BIG DATA, DATA MINING

This lecture is about the **classical DWH**

- 6 sessions

Mr. Bollinger’s/Roth's lecture is about **Data Mining**

DWH LECTURE - LEARNING TARGETS

- Describe different DWH architectures
- Explain DWH data modeling methods and design logical models
- Name DB techniques that are well-suited for DWHs
- Explain ETL processes
- Specify reporting & project management & meta data requirements
- Name current DWH trends

OVERVIEW OF THE LECTURE

Introduction to Data Warehouse	• 11.10.2017
DWH Architectures, Data Modeling	• 19.10.2017
Data Modeling, OLAP	• 26.10.2017
OLAP, ETL	• 02.11.2017
ETL, Metadata, DWH Projects	• 09.11.2017
DWH Projects, Advanced Topics	• 30.11.2017

ABOUT THE LECTURE

Structure of the lecture

- Review of the preceding lecture
- Presentation of content
- Group tasks, exercises
- Comprehensive case study

13:15 – 15:45 or 16:00 – 18:30

- 1x15min break

COURSE MATERIAL

- Download slides from (theory)
  http://wwwlehre.dhbw-stuttgart.de/~buckenhofer/
- Additional material for case study (practice)
  https://github.com/abuckenhofer/dwh_course

HOW TO CONTACT THE COURSE?

Who is the class representative?

- Please send me an email so that I have your contact data

Do you have a class email address?

WHAT YOU WILL LEARN TODAY

Data Warehousing is a major topic of computer science

After the end of this lecture you will be able to

- Understand the basic business and technology drivers for data warehousing
- Describe the characteristics of a data warehouse
- Describe the differences between production and data warehouse systems
- Understand logical standard DWH architecture
- Describe different layers and their meaning
- Describe advantages and disadvantages of further DWH architectures

MANY EMPLOYMENT OPPORTUNITIES

DWH department in every (bigger) **end user company**, also in many medium-sized or small-sized companies

DWH department in every (bigger) **consulting company**

DWH-only specialized consulting companies

DWH **tool vendors**

MANY EMPLOYMENT OPPORTUNITIES – CHALLENGING JOB REQUIREMENTS

DWHs are complex, much more complex compared to most OLTP systems

Challenging job profiles with comprehensive requirements

- Data Architecture
- Data Integration / ETL
- Data Modeling (not only 3NF)
- Data Visualization
- Data Quality
- Data Security
- Requirements Engineering
- Project Management

JOB DESCRIPTION EXAMPLES

Ihre Aufgaben:

- Sie übernehmen die Implementierung von Projekten und Teilprojekten zur Anpassung und Weiterentwicklung unserer Data Warehouse Systeme im internationalen Umfeld
- Sie entwickeln Metamodelle und Lademechanismen und programmieren interne und externe Kundenservices
- Sie erstellen komplexe Analysen und Reports und das Design für eine skalierbare, analytische Data Warehouse Umgebung
- Sie erarbeiten Lösungen in den Bereichen Performance Tuning, Datenmodellierung und Datenmodelloptimierung
- Sie unterstützen bei der korrektiven Wartung der Data Warehouse Systeme

Ihr Profil:

- Erfolgreich abgeschlossenes Studium der Informatik/ Wirtschaftsinformatik oder vergleichbare Aus- und Weiterbildung mit entsprechender Berufserfahrung
- Erste Berufserfahrung als Data Warehouse Entwickler
- Kenntnisse von OLAP-Tools (z.B. Microstrategy) und analytischen Systemarchitekturen
- Erfahrung im Aufbau von Datenbanken, idealerweise Oracle, SQL und PL/SQL
- Fundierte Projektmanagementkenntnisse
- Eigenverantwortliche, strukturierte und zielorientierte Arbeitsweise
- Teamfähigkeit und Flexibilität

Ihre Aufgaben – damit zaubern Sie Kunden ein Lächeln aufs Gesicht

- Datenherkunftsanalyse und Entwicklung des logischen Datenmodells sowie Umsetzung in physisches Design
- Konzeption, Design und Implementierung von ETL Prozessen zur Datenintegration in Datawarehouses
- Optimierung und Weiterentwicklung des Data Warehouse hinsichtlich Architektur, Datenorganisation und Performance
- Auswahl von relevanten Daten gemeinsam mit Business Analysten, Business Partnern und IT-Einheiten
- Extraktion, Validierung, Aufbereitung und Analyse großer, (un)strukturierter, interner und externer Datenbestände mit geeigneten Verfahren (z.B. ETL, EAI, Hadoop)
- Programmierung unter Verwendung der eingesetzten Data Warehouse und Big Data Werkzeuge wie Informatica, TIBCO, Hadoop, o.ä.
- Durchführung und Leitung von IT-Projekten
- Dokumentation, Wissenstransfer und Monitoring für Systemabläufe, Schnittstellen und Prozesse

Ihre Qualifikation – damit verzaubern Sie uns

- Abgeschlossenes Hochschulstudium (Mathematik, Informatik) oder vergleichbare in der Praxis erworbene IT-Kenntnisse
- Nachweisbare Praxiserfahrungen im Betrieb von Data Warehouses sowie mindestens einem der folgenden BI-Tools: Informatica, Tibco
- Breite Kenntnisse von klassischen DWH-/BI Konzepten
- Ausgeprägte Datenbank- und Programmierkenntnisse, inklusive Hadoop MapReduce, Java, SQL
- Erfahrungen in der agilen Softwareentwicklung (z.B. Scrum) und/oder dem agilen Projektmanagement
- Ausgeprägtes Qualitätsbewusstsein sowie Teamfähigkeit
- Sehr gutes Organisations- und Problemlösungsvermögen
- Analytisches Denkvermögen und die Fähigkeit zum konzeptionellen Arbeiten

JOB DESCRIPTION EXAMPLES

AUFGABEN

- Entwicklung von ETL Software mit PL/SQL, SQL und SAS für DW, sowie anderen Lösungen, die die Bereitstellung von Daten erfordern
- Zusammenarbeit mit dem Produkt Eigentümer und den Business Analysten, damit die Business Anforderungen vollständig verstanden werden und sichergestellt werden kann, dass die Projektleistung diese erfüllt
- Teil eines sich selbst organisierenden SCRUM Teams werden
- Vorbereiten von Test Szenarien sowie Strategien für Modul-, System- und Integrationstests
- Überprüfung von Code um grundlegende technische und logische Fehler zu identifizieren
- Identifizierung von Schwachstellen, Abweichungen des Systems nach der Veröffentlichung, zusätzlich zu der normalen Produktentwicklung

Weitere Aufgaben

- Teilnahme an Implementierungsprojekten für Kunden als Product Consultant
- Einhalten von bestehenden Software Development Prozessen wie Versions-, Build- und Deployment Management
- Aktives Vorantreiben von Standardisierung und Optimierung von Software Entwicklung
- Zusammenarbeiten und Unterstützung von neuen Kollegen und dem Near-Offshore Entwickler Teams
- Entwickeln von best practice Ansätzen, die die Produktivität und den Nutzen von Programmierungsgrundsätzen, Tools und Techniken für die Entwicklung von Solution Code verbessern

Ihr Aufgabengebiet umfasst im Wesentlichen:

- Erstellung und Weiterentwicklung von Lösungen im bestehenden BI-Umfeld (Oracle PL/SQL)
- Teilprojektleitung bei Konzeption und Umsetzung von BI-Projekten in Abstimmung mit Fachabteilungen und der Bankorganisation
- Anbindung operativer Datenquellen sowie Aufbereitung und strukturierte Bereitstellung der Daten in Oracle
- Eigenverantwortliche Weiterentwicklung der datentechnischen Architektur
- Enge Abstimmung mit unseren Fachabteilungen hinsichtlich Anforderung, Spezifikation und der Umsetzung
- Daten- und Prozessmodellierung bestehender wie neuer Systeme und Beladungsprozesse
- Zusammenarbeit mit dem Anwendungsbetrieb bei der Administration der Datenbanken und dem Releasemanagement

Unser System-Umfeld (BI):
Datenbanken: Oracle 12c
BI-Tools: SAS-BI-Server 9.4
Betriebssysteme: SUSE Linux ES 10 und 11
Entwicklungssprachen: SQL, PL/SQL, SAS, Shell

Ihr Eignungsprofil:

- Abgeschlossenes Studium der (Wirtschafts-) Informatik, Mathematik oder vergleichbare Qualifikation
- Ausgeprägte Kenntnisse im Umgang mit relationalen Datenbanken, Datenmodellierung, Data Warehouse-Konzepten und -Architekturen, die Sie schon erfolgreich in Projekten unter Beweis stellen konnten
- Mehrjährige Erfahrung in der Datenbankentwicklung unter Oracle mit PL/SQL
- Kenntnisse Oracle spezifischer Features und Konzepte: Z.B. Performance Tuning, Indexing, Partitioning, Materialized Views, Role Manager
- LINUX-Betriebssystem-Kenntnisse (inkl. Scripting)
- Idealerweise Branchenwissen im Bereich Banken und Finanzdienstleistungen
- Analytisches Denkvermögen, eigenständige Arbeitsweise und schnelles Auffassungsvermögen

JOB DESCRIPTION EXAMPLES

Aufgaben

Als Teil des [redacted] Business Intelligence und Analytics Team, stehen Sie im regelmäßigen Wissensaustausch mit allen Teammitgliedern und stellen sich ständig neuen Herausforderungen. Im Rahmen Ihrer Tätigkeit haben Sie die Möglichkeit sowohl Ihre technische als auch methodische Expertise kontinuierlich auszubauen sowie Ihre eigenen Kenntnisse weiterzugeben.

- Sie unterstützen die technische Betreuung, Entwicklung und Integration von großen und komplexen Data-Warehousing und Business Intelligence-Projekten
- Sie entwickeln neue Konzepte für unser Data-Warehouse mithilfe von Microsoft SQL Server, sowie anderen verwandten Microsoft-Technologien
- Sie identifizieren Potenziale durch neue Ansätze, sowohl hinsichtlich der Architektur als auch der verwendeten Technologien
- Sie konzipieren und entwickeln funktionale und nicht-funktionalen Anforderungen an neue MS BI- Lösungen
- Sie sind verantwortlich für die Implementierung und Institutionalisierung von Standards und Methoden für eine umfassende und effektive Qualitätssicherung in der Software-Entwicklung

Qualifikationen

- Sie haben Ihr Masterstudium im Bereich Informatik erfolgreich abgeschlossen oder haben eine vergleichbare Ausbildung absolviert
- Sie haben bereits mehrjährige Berufserfahrung als Microsoft Business Intelligence Entwickler gesammelt
- Sie haben Erfahrung im Bereich Data Architectur and Modelling (DataVault und Dimensional Data Warehouse nach Kimball)
- Sie haben sehr gute Kenntnisse im Umgang mit Microsoft SQL Server 2014 und dem Microsoft BI Stack (SSIS, SSDB, SSRS, SSAS, Tabular Model, PowerBI,)
- Sie bringen idealerweise Kenntnisse im Bereich der Data Warehouse Automatization mit Tools wie Talend mit und kennen sich mit .NET Programmierung aus
- Sie haben bereits in dynamischen und agilen Teams (SCRUM, KANBAN) gearbeitet
- Sie arbeiten gerne in einem dynamischen Umfeld und zeichnen sich durch effektives, effizientes und selbstmotiviertes Arbeiten aus
- Sie weisen fließende Deutsch- und Englischkenntnisse auf (in Wort und Schrift)
- Sie besitzen ein analytisches Denkvermögen und haben eine Affinität zu Daten und Fakten

DATA WAREHOUSE (DWH) OR BUSINESS INTELLIGENCE (BI)?

Often used as synonym

DWH more **technical focus**

BI more **business / process focus**

- “Business intelligence is a set of **methodologies**, **processes**, **architectures**, and **technologies** that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision making.” (Boris Evelson, Forrester Research, 2008)

INFORMATION TECHNOLOGY (1960'IES – 80'IES)

Many systems throughout the enterprises for dedicated purposes

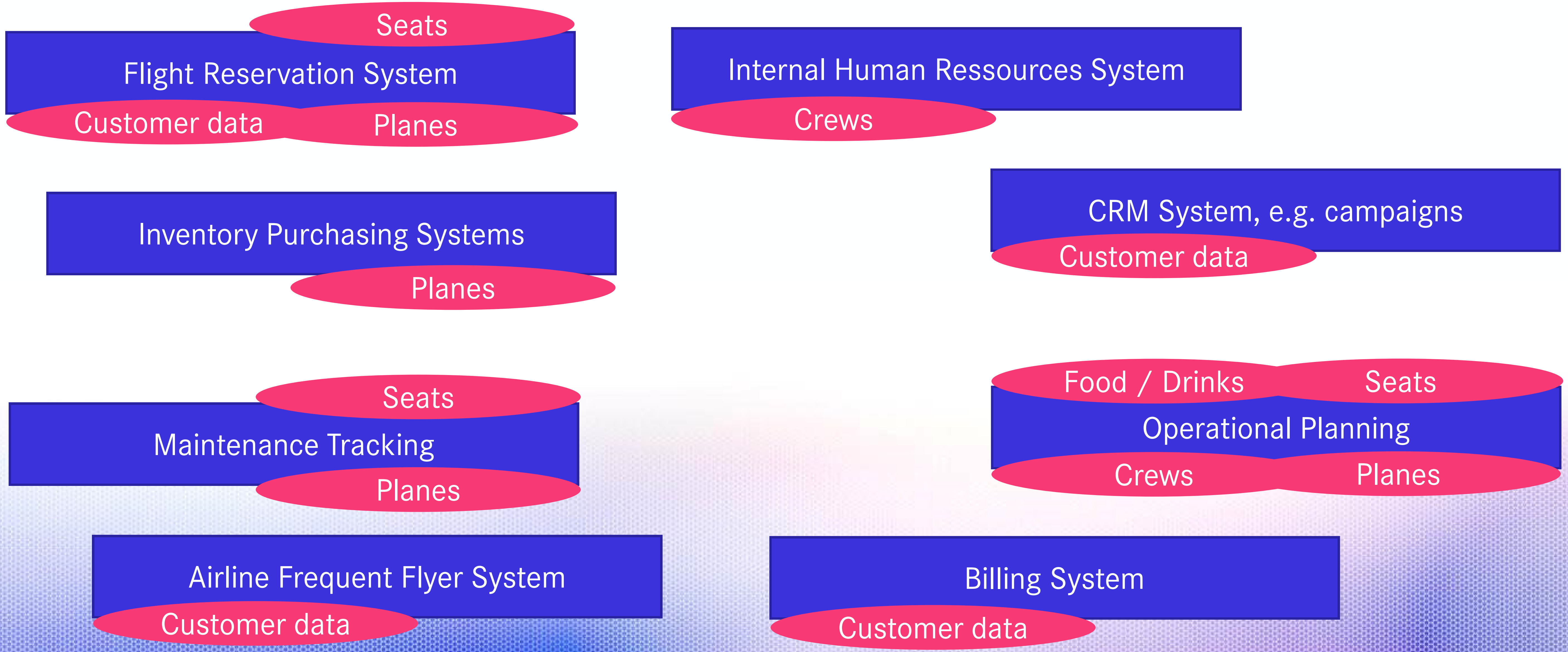
- Support daily transactions / day-to-day business
- Target: replace manual and time consuming activities

Data embedded in process-specific application

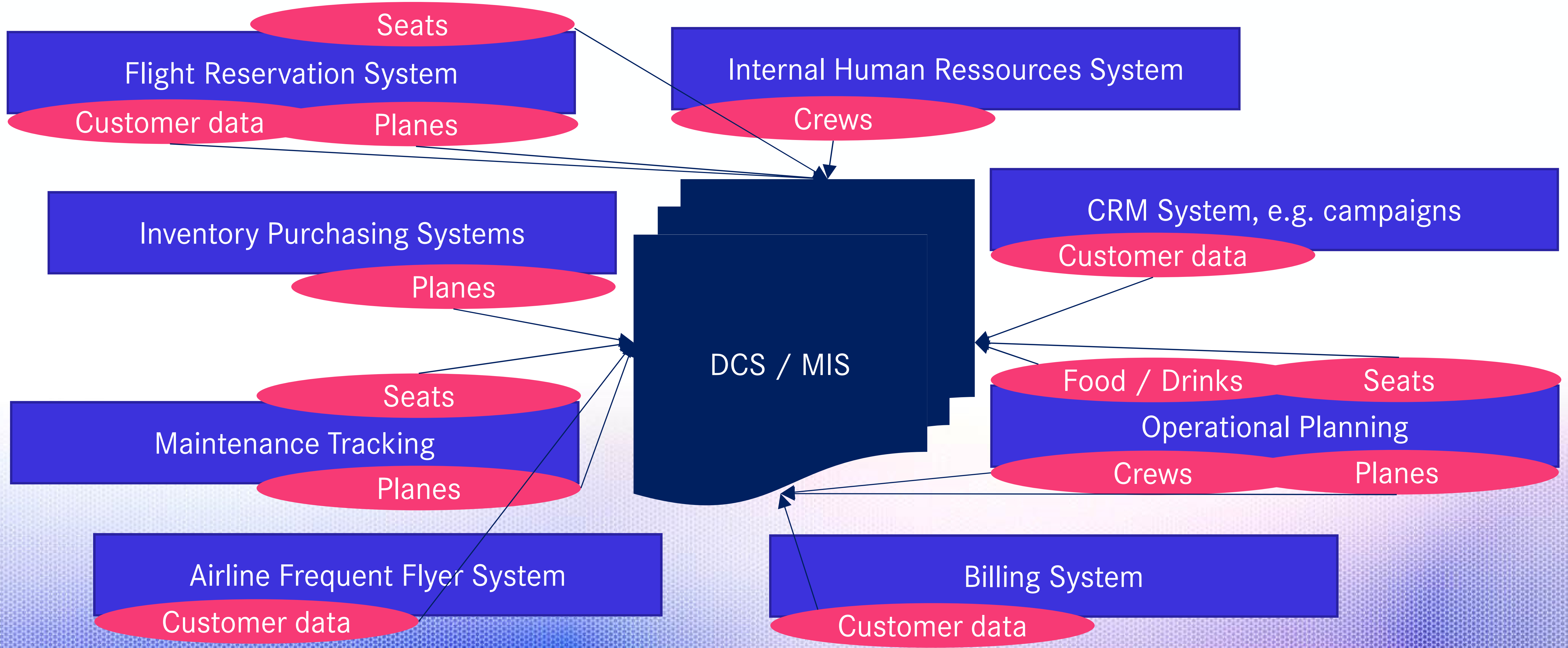
- **Process-orientation + dedicated purpose**

Customer data, order data, etc. spread over many systems in many variations and with contradictions

SAMPLE APPLICATIONS FOR AN AIRLINE



NEED FOR DECISION SUPPORT SYSTEM / MANAGEMENT INFORMATION SYSTEM



EARLY DECISION SUPPORT SYSTEMS (1960’IES – 80‘IES)

Can be characterized as “Unplanned decision support” or “Unplanned Management Information Systems (MIS)”

- Management needs reports / combined data from different systems to make decisions for company
- Reports are manually written by IT people
  - Extract, combine, accumulate data
  - Can take several days to write report and to get the data
- Error prone and labour-intensive
  - Relevant information may be forgotten or combined in a wrong way

Did not really work

INFORMATION TECHNOLOGY TODAY
FURTHER REQUIREMENTS

Data still spread across many applications, but additional requirements

Data as Asset, getting more and more important also in production industries

- Not only classical data-intensive companies like Google or Facebook
- Increasing interest e.g. in insurance, health care, automotive, …
  - Connected cars, Smart Home, Tailor-made insurances, etc.

Hype technologies

- New databases technologies like NoSQL and Big Data

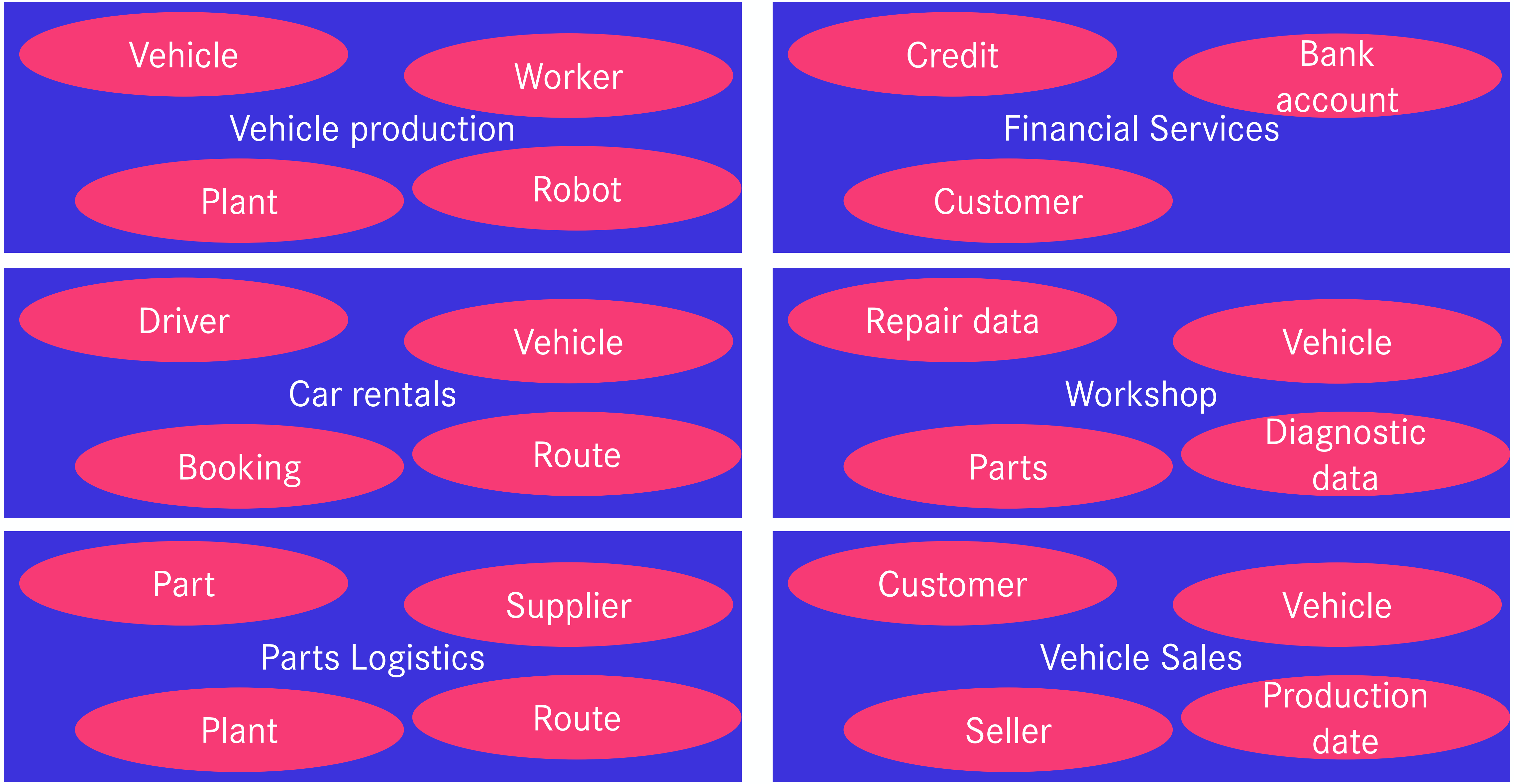
DWH still booming with additional stimuli coming from Big Data, Digitization, Internet Of Things IOT, Industry 4.0, Real Time, Time To Market, etc.

EXERCISE – OLTP SYSTEMS

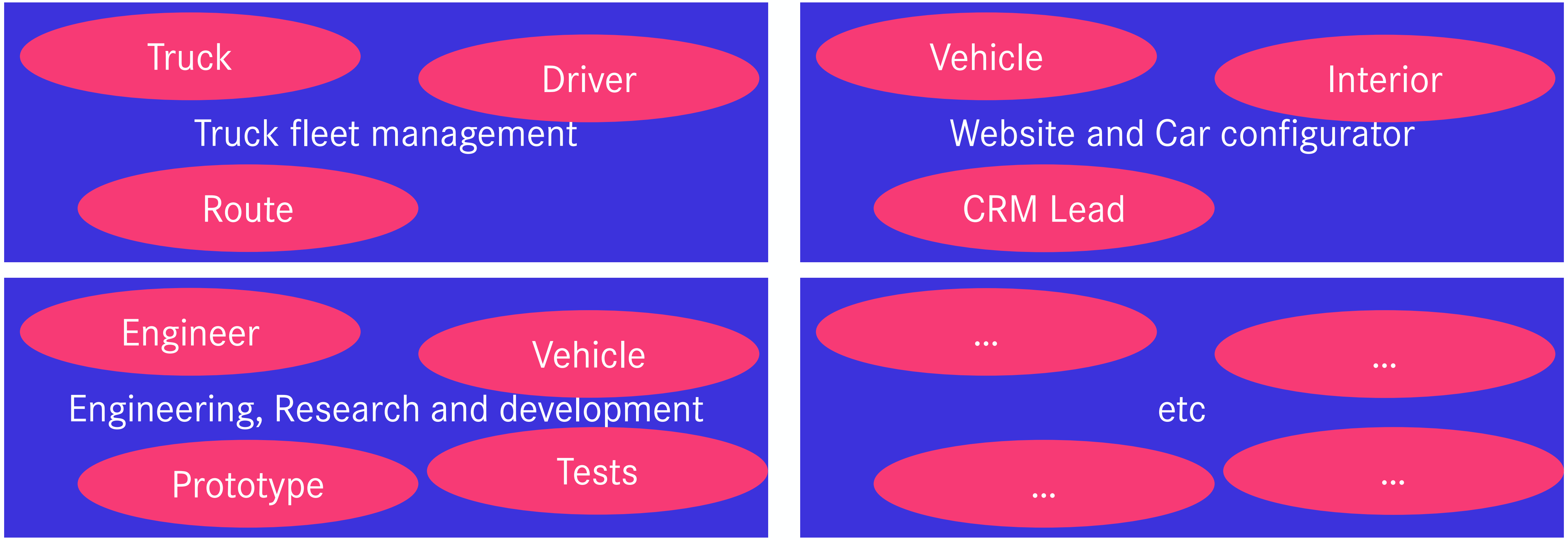
Outline the at least 5 operational systems for a vehicle manufacturer

- which data is stored by these systems
- characterize which operations are performed by them
- which questions can be answered by these systems (and which questions can not be answered = major problems for decision support)

SAMPLE OLTP SYSTEMS



SAMPLE OLTP SYSTEMS



CHALLENGE

How to get an overall view

across OLTP applications / functions that works?

MAJOR PROBLEMS FOR EFFECTIVE DECISION SUPPORT

- Distributed data
- Different data structures
- Historic data
- System workload
- Inadequate technology

DISTRIBUTED DATA

Problem: Data resides on

- different systems / storages
- different applications
- different technologies

Solution: Data has to be accumulated on one system for further analysis

- Data is inhomogeneous, e.g. each system has their own customer number or order number, etc.
- How to combine the data?
- Data must be ingested regularly, e.g. daily and not ad-hoc

DIFFERENT DATA STRUCTURES

Problem: Systems developed independently from each other

- Different data types
  - E.g.: zip-code as integer or character string
- Different encodings
  - E.g.: kilometer or miles
- Different data modeling
  - E.g.: last name / first name in different fields vs last name / first name (badly modelled) in one single field

Solution: Dedicated system required that harmonizes / standardizes the data

ISSUES WITH HISTORIC DATA

Problem: Data is updated and deleted or archived after max. 3 months

- daily transactions produce lots of data
- limited size of storage → high amounts of data fill up systems

Historic data is required for decision support

- e.g. how did sales figures develop compared to last month / last year / etc.

Solution: All data (changes) have to be stored in a system capable of dealing with huge amounts of data

ISSUES WITH SYSTEM WORKLOAD

Problem: Performance not optimized for new workloads

- Systems stressed by additional load (due to reports)
- Not optimized for this kind of workload
- Performance of daily transaction business jeopardized
- May possibly lead to system failure!
- Imagine what happens if a system like Amazon gets slow

Solution: Dedicated system that handles complex (arithmetic) queries on huge amounts of data. A system that is optimized for that kind of workloads.

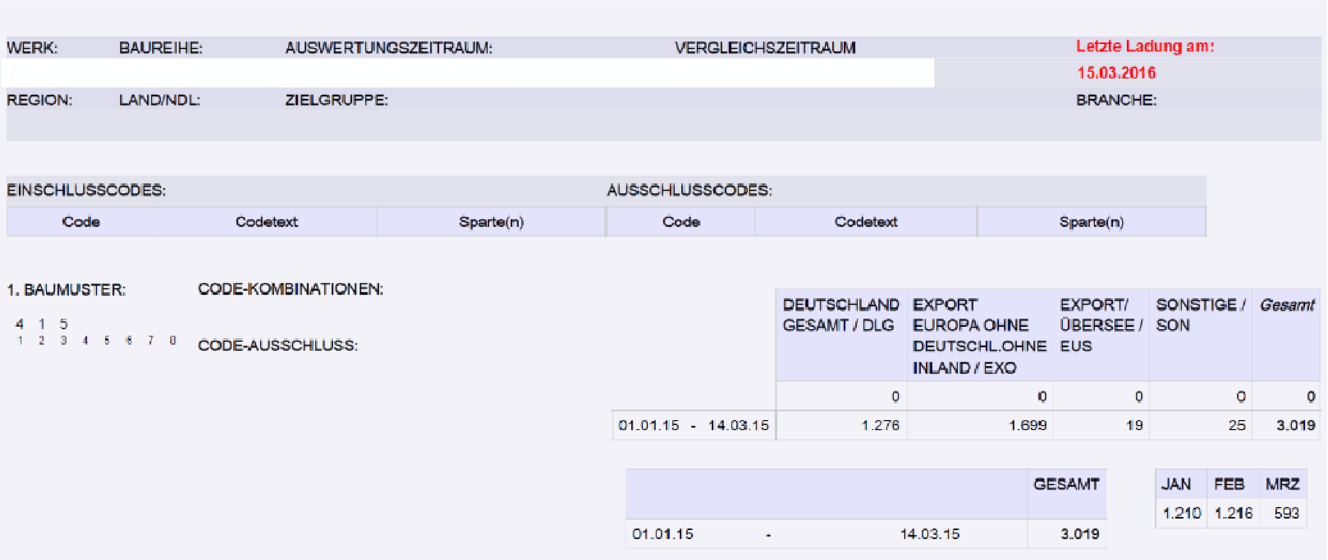
INADEQUATE TECHNOLOGY

Problem: Tooling and technology different from OLTP

- Inadequate tools for data integration and analysis
- Infrastructure configured for OLTP transactions and not for DWH load
- Storage systems and processors to weak to fulfill the requirements

Solution: Standard Tools and technology that help to increase productivity and solve such problems, e.g. Reporting Tools for Data Analysis or ETL tools for Data ingestion/load

TEXTUAL AND VISUAL REPORTS



WERK:	BAUREIHE:	AUSWERTUNGSZEITRAUM:	VERGLEICHSZEITRAUM	Letzte Ladung am: 15.03.2016
REGION:	LAND/NDL:	ZIELGRUPPE:		BRANCHE:

EINSCHLUSSCODES:			AUSSCHLUSSCODES:		
Code	Codetext	Sparte(n)	Code	Codetext	Sparte(n)

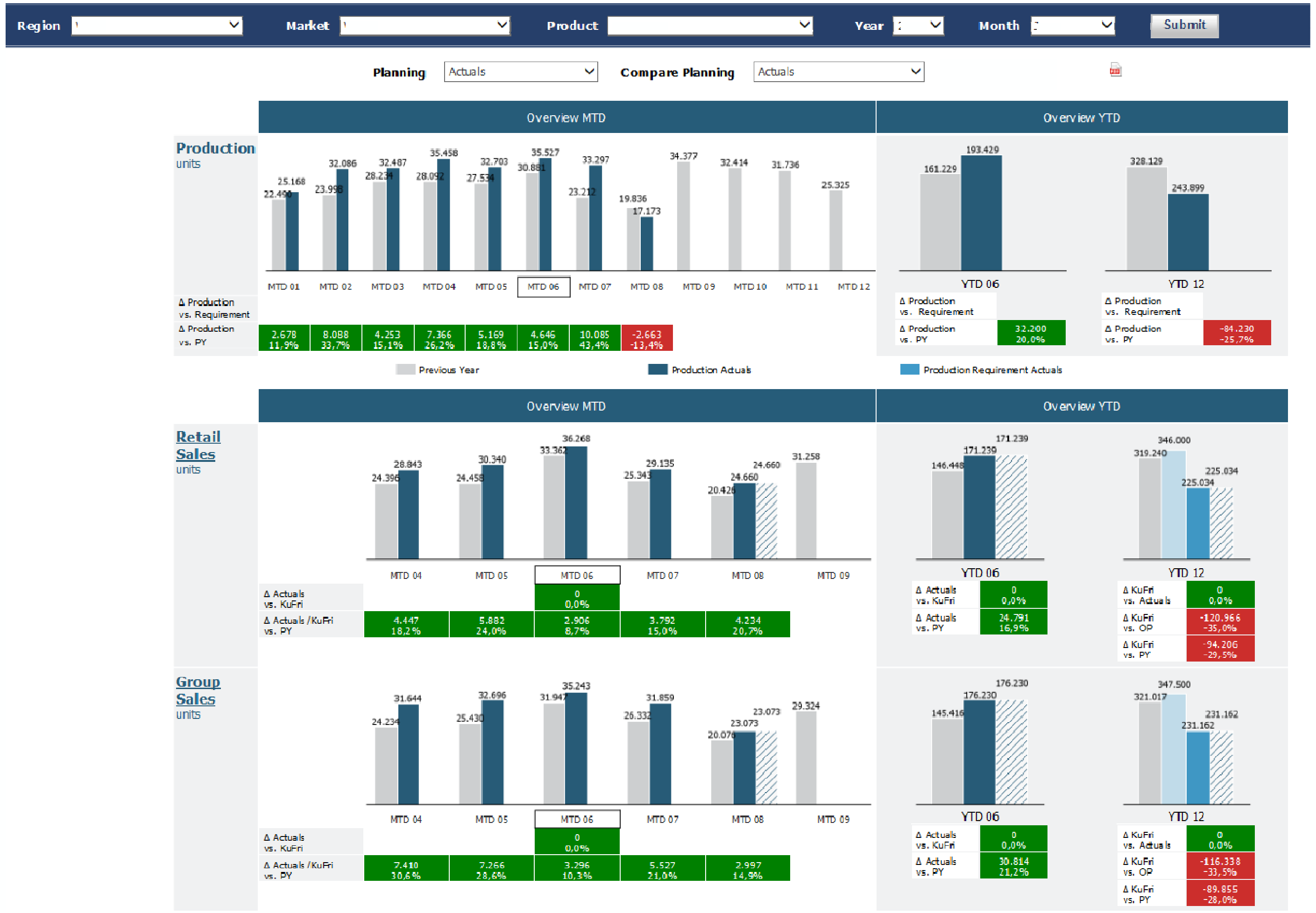
1. BAUMUSTER: CODE-KOMBINATIONEN:

4 1 5
1 2 3 4 5 6 7 8

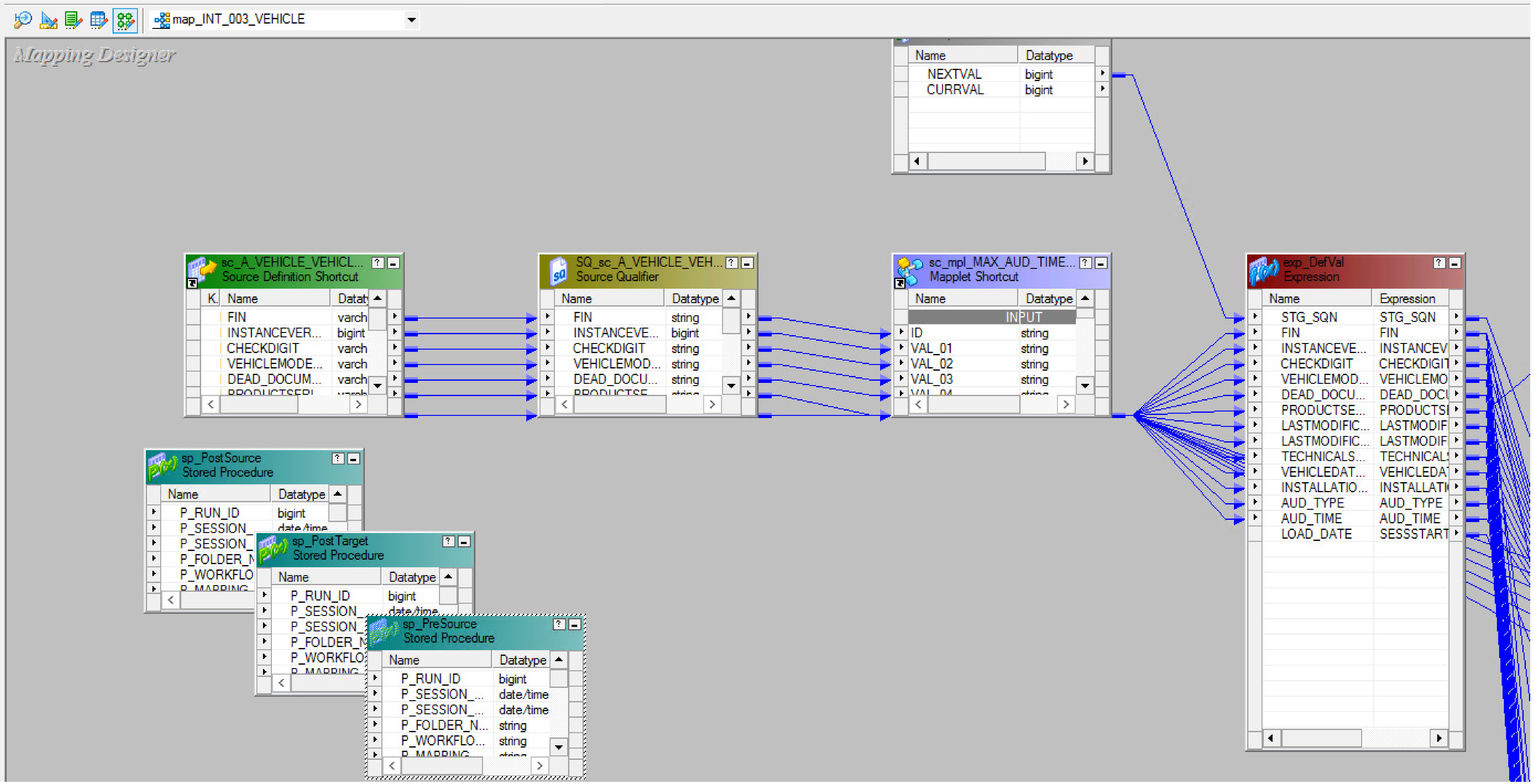
CODE-AUSSCHLUSS:

	DEUTSCHLAND GESAMT / DLG	EXPORT EUROPA OHNE DEUTSCHL.OHNE INLAND / EXO	EXPORT/ ÜBERSEE / EUS	SONSTIGE / SON	*Gesamt*
	0	0	0	0	0
01.01.15 - 14.03.15	1.276	1.699	19	25	3.019

	GESAMT	JAN	FEB	MRZ
01.01.15 - 14.03.15	3.019	1.210	1.216	593



VISUAL DATA INTEGRATION TOOL



CHALLENGE

How to get an overall view
across OLTP applications / functions that works?

CONCLUSION

Operative systems not suitable for analytical evaluations

Need for a new, separated system

- fast answers, ad-hoc questions possible
- no interference with daily transaction business

→ Data Warehouse

EXERCISE

List possible (functional and non-functional) requirements for a data warehouse end-user. Think of deficiencies of transactional systems like

- Distributed data
- Different data structures
- Problem with historic data
- Problem with system workload
- Inadequate technology

What are requirements from a Data Warehouse user perspective? (List at least 5 requirements)

DATA WAREHOUSE USER

- Wants to trust the data: **quality assured data**
- Wants to access and analyze all data in a **single database**
- Wants to get a complete analysis including **history**, e.g. where did the customer live 5 years ago or how did bookings develop the last 10 days?
- Wants **fast** data access for his queries
- Wants to **understand the data model** = one single and easy data model and not many different applications
- Wants to browse through combined data sets to identify correlations or **new insights**

DATA WAREHOUSE

Contains data from different systems

Imports data from different systems on a regular basis

- detailed data and summarized data
- provide historic data
- generate metadata

OLTP applications remain, DWH is a completely new system

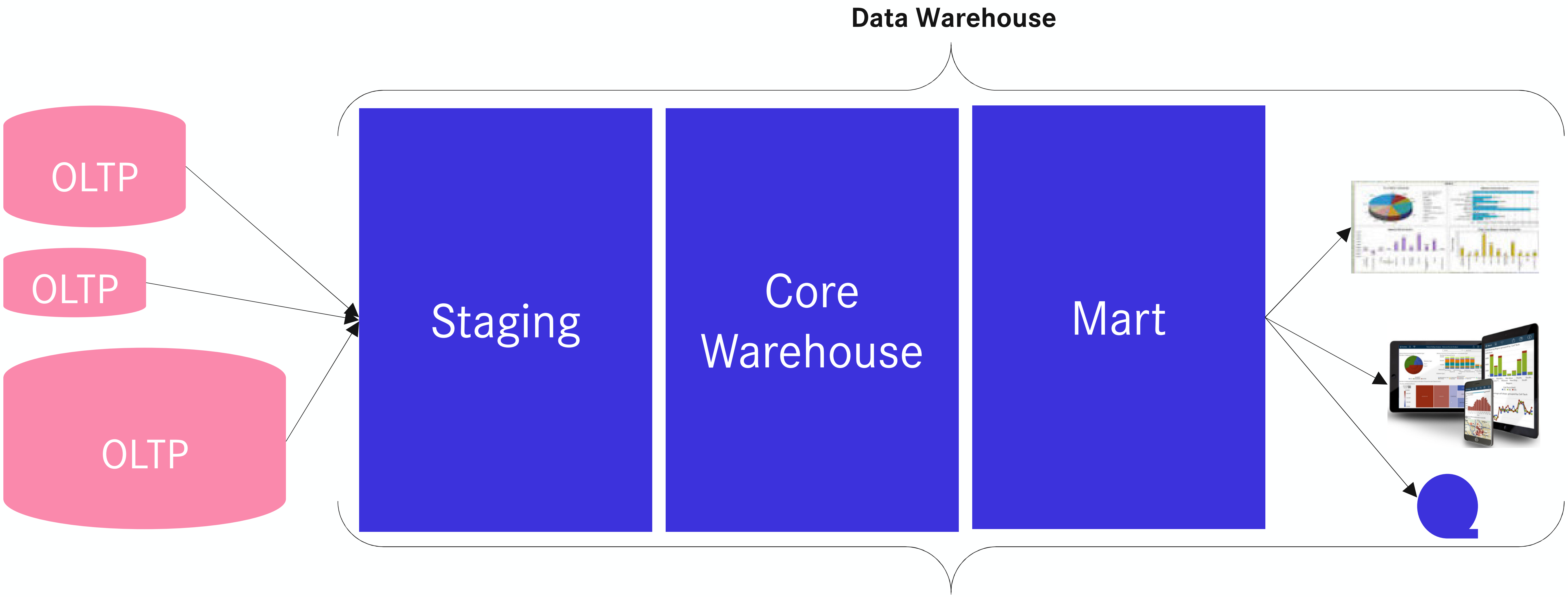
Overcomes difficulties when using existing transaction systems for those tasks

DATA WAREHOUSE

Not a product, but a overall concept

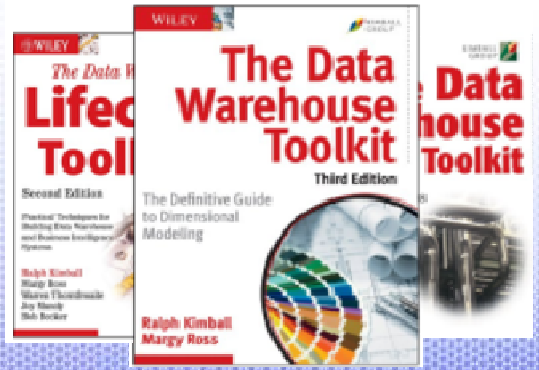
Applications come, applications go. The data, however, lives forever. It is not about building applications; **it really is about the data** underneath these application (Tom Kyte)

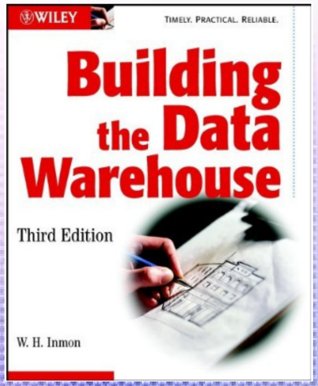
HIGH-LEVEL DATA WAREHOUSE ARCHITECTURE



DATA WAREHOUSE DEFINITIONS BY TWO "FATHERS" OF THE DWH

Ralph Kimball	William Harvey „Bill" Inmon
„A data warehouse is a copy of transaction data specifically structured for querying and reporting"	"A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process"





SUBJECT-ORIENTED

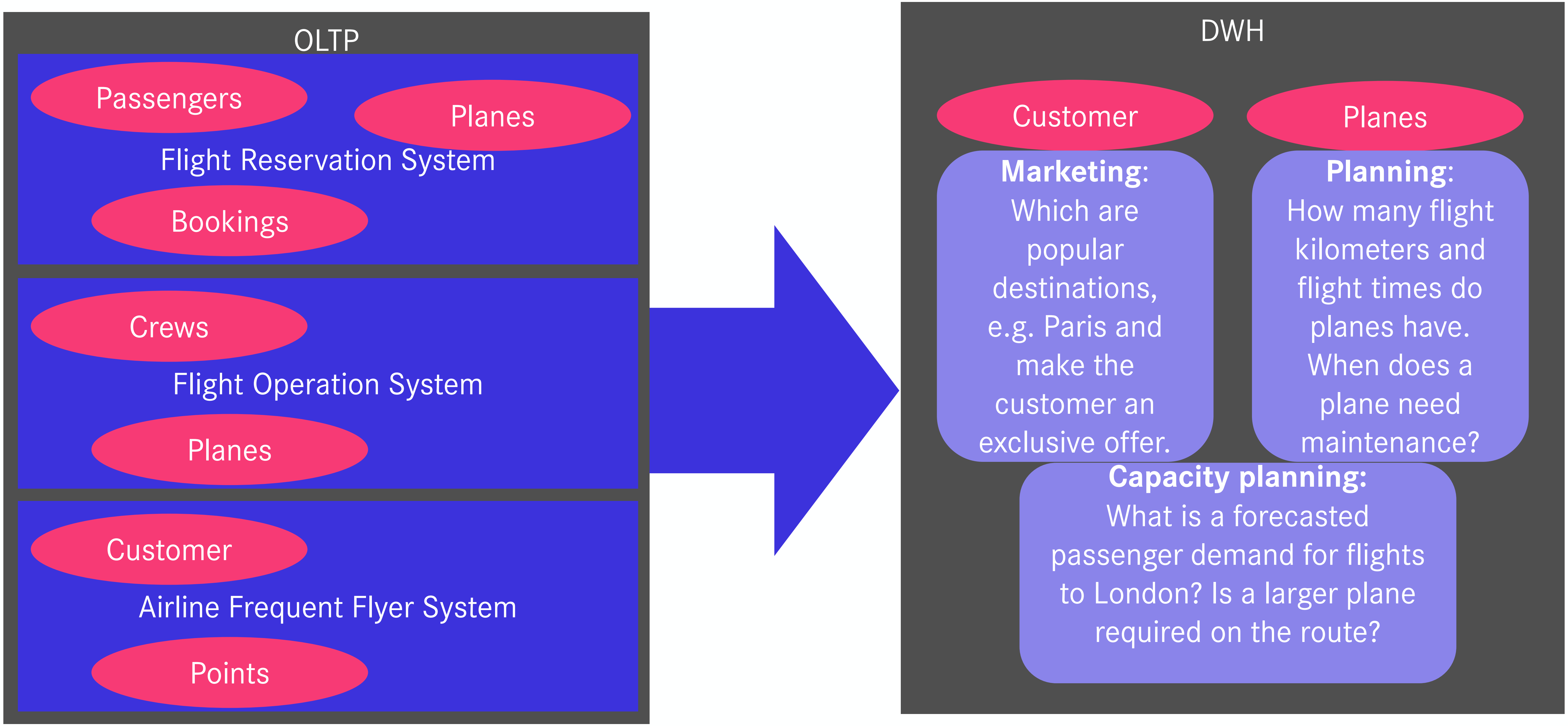
A data warehouse is organized around the major subjects (business entities) of the enterprise like

- Customer
- Vendor
- Car
- Transaction or activity

In contrast to the application/process/functional orientation such as

- Booking application
- Delivery handling

SUBJECT-ORIENTED - EXAMPLE



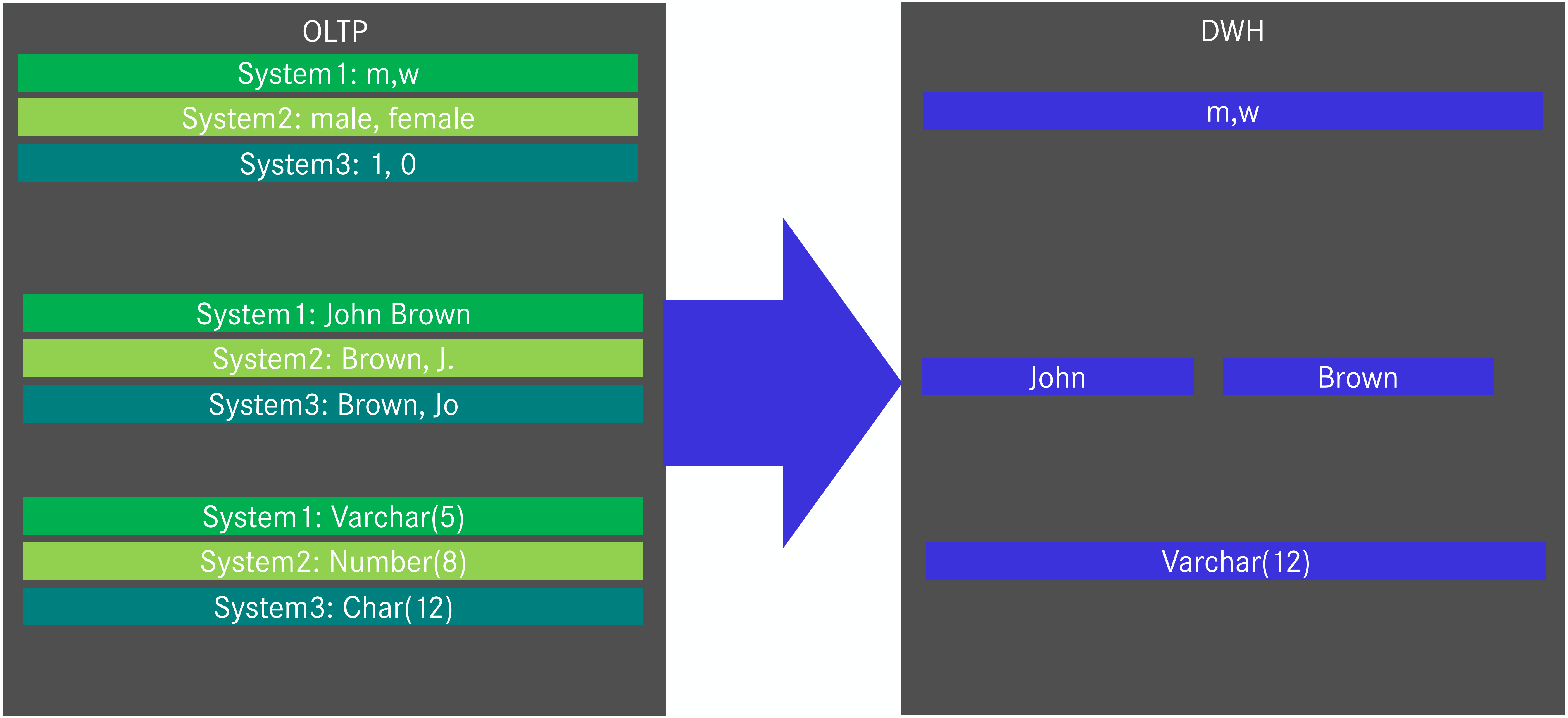
INTEGRATED

Data contained in the warehouse are integrated.

Aspects of integration

- consistent naming conventions
- consistent measurement of variables
- consistent encoding structures
- consistent physical attributes of data

INTEGRATED - EXAMPLE



NONVOLATILE

Operations in operational environment

- Insert
- Delete
- Update
- Select

Operations in a data warehouse

- Insert: the initial and additional loading of data by (batch) processes
- Select: the access of data
- (almost) **no updates and deletes** (technical updates / deletes only)

NONVOLATILE - EXAMPLE

OLTP

Flight Reservation System

Passenger John flies from Stuttgart to London on 15.02 at 06:00

Insert into DB:
Passenger John, From Stuttgart to London, 15.02. 06:00

Passenger John changes his mind and flies at 10:00

Update in DB:
Passenger John, 15.02. 10:00

DWH

Insert into DB:
Passenger John, From Stuttgart to London, 15.02. 06:00

Insert into DB:
Passenger John, From Stuttgart to London, 15.02. 10:00

NONVOLATILE - EXAMPLE

What happens in the OLTP system if the customer cancels his booking?

- Delete operation in OLTP
- Seat gets available again and can be sold to another passenger

What happens in the DWH?

- Insert operation in DWH with e.g. a flag indicating that the customer cancelled/deleted his booking
- Business can make analysis about cancelled booking: why might the customer have cancelled? How to prevent the customer or other customers to cancel next time?

TIME-VARIANT

All data in the data warehouse is accurate as of some moment in time

- Has to be associated with a **time stamp**

Once data is correctly recorded in the data warehouse, it cannot be updated or deleted

- Data warehouse data is, for all practical purposes, a long **series of snapshots**

In the operational environment data is accurate as of the moment of access

Operational data, being accurate as of the moment of access, can be updated as the need arises

TIME-VARIANT - EXAMPLE

DWH

Insert into DB	DB insert timestamp
Insert into DB: Passenger John, From Stuttgart to London, 15.02. 06:00	DB insert timestamp: 02.02. 15:03:21
Insert into DB: Passenger Jim, From Hamburg to Munich, 18.02. 15:00	DB insert timestamp: 02.02. 15:04:29
Insert into DB: Passenger John, From Stuttgart to London, 15.02. 10:00	DB insert timestamp: 05.02. 12:15:03
Insert into DB: Passenger Mike, From Hamburg to Munich, 15.02. 10:00	DB insert timestamp: 05.02. 12:15:11
Insert into DB: Passenger John, From Stuttgart to London, 15.02. 10:00, Cancel Flag	DB insert timestamp: 08.02. 09:52:33

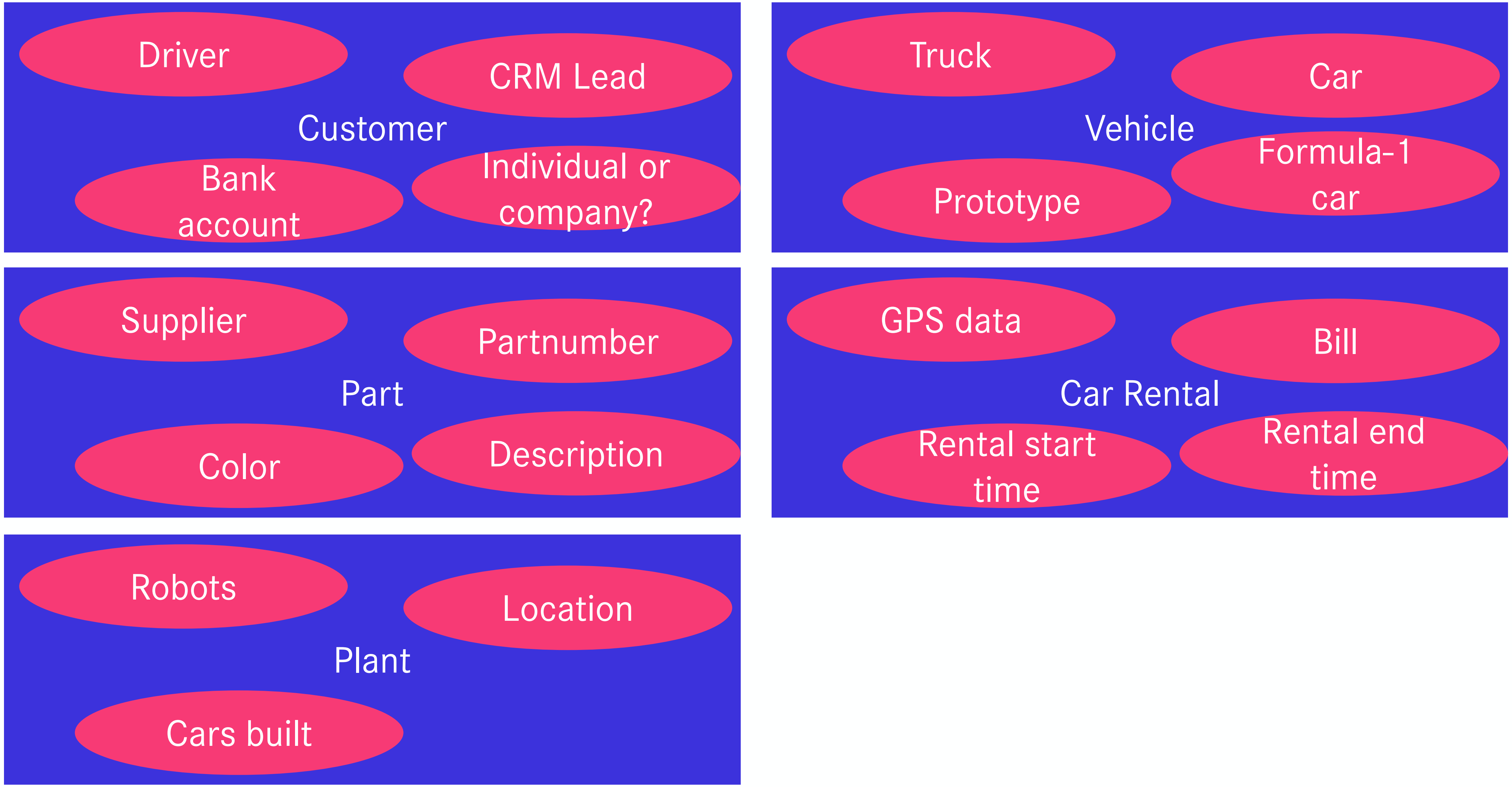
EXERCISE - DWH

You outlined OLTP systems for a vehicle manufacturer in an earlier exercise.

Now start designing a Data Warehouse:

- Describe what data can be stored in it. Define at least 5 subject-areas!
- Which questions can/should be answered with this information

DWH – SUBJECT AREAS



EXERCISE – SAMPLE QUESTIONS

Which customers own a car and use car rental regularly?

Which parts have the most defects? Can diagnostic data be used to predict potential defects and warn customers?

Which areas and times are popular for car rentals? Does it make sense to relocate cars to these areas? (e.g. cinema in the evening/night)

OLTP VS OLAP

Online Transaction Processing	Online Analytical Processing
Transaction-oriented system	Query-oriented system
Optimized for insert and update consistency	Optimized for complex queries with short response times; ad-hoc queries
Many users change data	Only ETL process writes data
Selective queries on the data	Evaluations of all data including history (complex queries)
Avoid redundancy	Redundant data storage
Normalized data management 3NF	De-normalized data management
Relational Data Modeling	Several layers with different data models, one model usually Dimensional Data Modeling

OPERATIVE VS INTEGRATED DATA

	Operative data	Integrated data
Handling	Structured, parallel processes with short and isolated ("atomic") transactions	Information for management (decision support)
Modeling	Process- and function oriented, individual for each application	Different data models in one DWH; historic, stable and summarized, data
# users	Many	Few(er) but increasing user base
System return time	Milliseconds	Seconds to minutes (even hours)

OPERATIVE VS ANALYTICAL DATABASES

	Operative DBs	Analytical DBs
Purpose	Processing of daily business transactions	Information for management (decision support)
Content	Detailed, complete, most recent data	Historic, stable and summarized data
Data amount	Small amount of data per transaction. Nested Loop Joins	Large amount of data for load, and often per query. Hash Joins common
Data structure	Suitable for operational transactions	Several models; suitable for long term storage and business analyses
Transactions	ACID; very short read/write transactions	Long load operations, longer read transactions

SUMMARY

Which challenges could not be solved by OLTP? Why is a DWH necessary?

- Integrated view, distributed data, historic data, technological challenges, system workload, different data structures

Name two “fathers” of the DWH

- Bill Inmon and Ralph Kimball

Which characteristics does a DWH have according to Bill Inmon?

- Subject-oriented, integrated, non-volatile, time-variant

DWH ARCHITECTURE

PURPOSE: WHY ARE DWH ARCHITECTURES USEFUL?

- Specific implementation can follow an architecture
  - Architecture describes an ideal type. Therefore an implementation may not use all components or can combine components
- Better understanding, overview and complexity reduction by decomposing a DWH into its components
  - Can be used in many projects: repeatable, standardizable
- Map DWH tools into the different components and compare functionality
- Functional oriented as it describes data and control flow

EXAMPLES OF DATA WAREHOUSES IN THE INDUSTRY

Apple: multiple Petabytes

- Customer insights: who's who and what are the customers up to

Walmart: 300TB (2003), several PB today

- It tells suppliers, "You have three feet of shelf space. Optimize it."

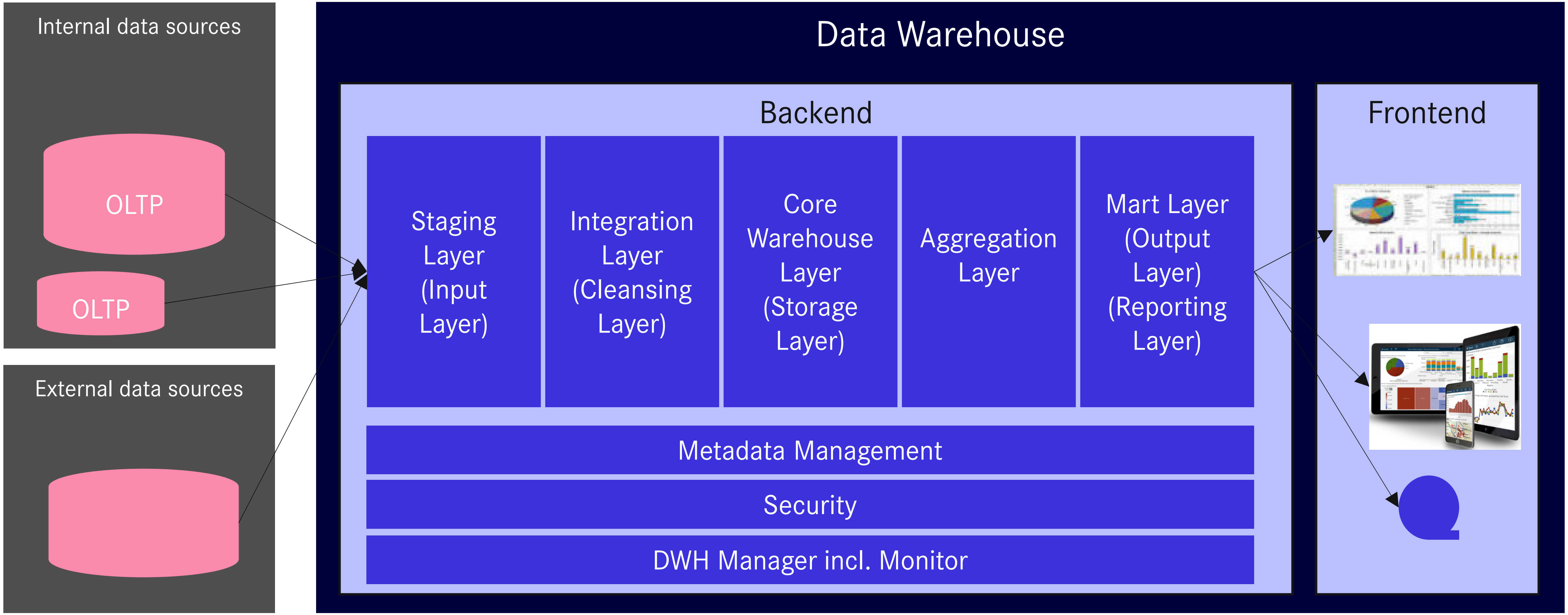
eBay: >10PB, 100s of production DBs fed in

- Get better understanding of customers

Most DWHs are much smaller though. For huge and small DWHs: High challenges to **architect** + develop + maintain + run such complex systems

https://gigaom.com/2013/03/27/why-apple-ebay-and-walmart-have-some-of-the-biggest-data-warehouses-youve-ever-seen/ and http://www.dbms2.com/2009/04/30/ebays-two-enormous-data-warehouses/

LOGICAL STANDARD DATA WAREHOUSE ARCHITECTURE



DATA SOURCES

- Providing internal and external data out of the source systems
- Enabling data through Push (source is generating extracts) or Pull (BI Data Backend is requesting or directly accessing data)
  - Example for Push practice (deliver csv or text data through file interface; Change Data Capture (CDC))
  - Example for Pull practice (direct access to the source system via ODBC, JDBC, API and so on)

STAGING LAYER

- “Landing Zone” for data coming into a DWH
- Purpose is to increase speed into DWH and decouple source and target system (repeating extraction run, additional delivery)
- Granular data (no pre-aggregation or filtering in the Data Source Layer, i.e. the source system)
- Usually not persistent, therefore regular housekeeping is necessary (for instance delete data in this layer that is few days/weeks old or – more common - if a correct upload to Core Warehouse Layer is ensured)
- Tables have no referential integrity constraints, columns often varchar

INTEGRATION LAYER

- Business Rules, harmonization and standardization of data
- Classical Layer for transformations: ETL = Extract – TRANSFORM – Load
- Fixing data quality issues
- Usually not persistent, therefore regular housekeeping is necessary (for instance after a few days or weeks or at the latest once a correct upload to Core Warehouse Layer is ensured)
- The component is often not required or often not a physical part of a DB

CORE WAREHOUSE LAYER

- Data storage in an integrated, consolidated, consistent and non-redundant (normalized) data model
- Contains enterprise-wide data organized around multiple subject-areas
- Application / Reporting neutral data storage on the most detailed level of granularity (incl. historic data)
- Size of database can be several TB and can grow rapidly due to data historization
- Write-optimised layer

AGGREGATION LAYER

- Preparing data for the Data Mart Layer to the required granularity
  - E.g. Aggregating daily data to monthly summaries
  - E.g. Filtering data (just last 2 years or just data for a specific region)
- Harmonize computation of key performance indicators (measures) and additional Business Rules
- The component is often not required or often not a physical part of a DB

DATA MART LAYER

- Read-optimised layer: Data is stored in a denormalized data model for performance reasons and better end user usability/understanding
- The Data Mart Layer is providing typically aggregated data or data with less history (e.g. latest years only) in a denormalized data model
  - Created through filtering or aggregating the Core Warehouse Layer
- One Mart ideally represents one subject area
- Technically the Data Mart Layer can also be a part of an Analytical Frontend product (such as Qlik, Tableau, or IBM Cognos TM1) and need not to be stored in a relational database

METADATA MANAGEMENT, SECURITY, MONITOR

- Metadata Management
  - “Data about Data”, separate lecture
- Security
  - Not all users are allowed to see all data
  - Data security classification (e.g. restricted, confidential, secret)
- DWH Manager incl. Monitor
  - DWH Manager initiates, controls, and checks job execution
  - Monitor identifies changes/new data from source systems, separate lecture

EXERCISE: CLASSICAL DWH ARCHITECTURES

The article

http://www.kimballgroup.com/2004/03/differences-of-opinion/

compares THE two classic DWH architectures.

Read the paper and complete the table / questions on the next slide.

(Caution: The paper is biased / favors one approach; you may want to read other/more papers for a neutral view.)

EXERCISE: CLASSICAL DWH ARCHITECTURES

How are the approaches called?		
Who “invented” the approach?		
How many layers are used and how are the layers called?		
Which data modeling approaches are used in which layer?		
In which layer are atomic detail data stored?		
In which layer are aggregated / summary data stored?		
List at least 2 advantages		
List at least 2 disadvantages		

EXERCISE: CLASSICAL DWH ARCHITECTURES

How are the approaches called?	**Kimball Bus Architecture**	**Corporate Information Factory**
Who “invented” the approach?	• Ralph Kimball	• Bill Inmon
How many layers are used and how are the layers called?	• Data Staging • Dimensional Data Warehouse	• Data Acquisition • Normalized Data Warehouse • Data Delivery / Dimensional Mart
Which data modeling approaches are used in which layer?	• Data Staging: variable, corresponds to source system • Dimensional Data Warehouse: Dimensional Model	• Data Acquisition: variable, corresponds to source system • Normalized Data Warehouse: 3NF • Data Delivery: Dimensional Model
In which layer are atomic detail data stored?	• Dimensional Data Warehouse	• Normalized Data Warehouse
In which layer are aggregated / summary data stored?	• Dimensional Data Warehouse	• Data Delivery / Dimensional Mart

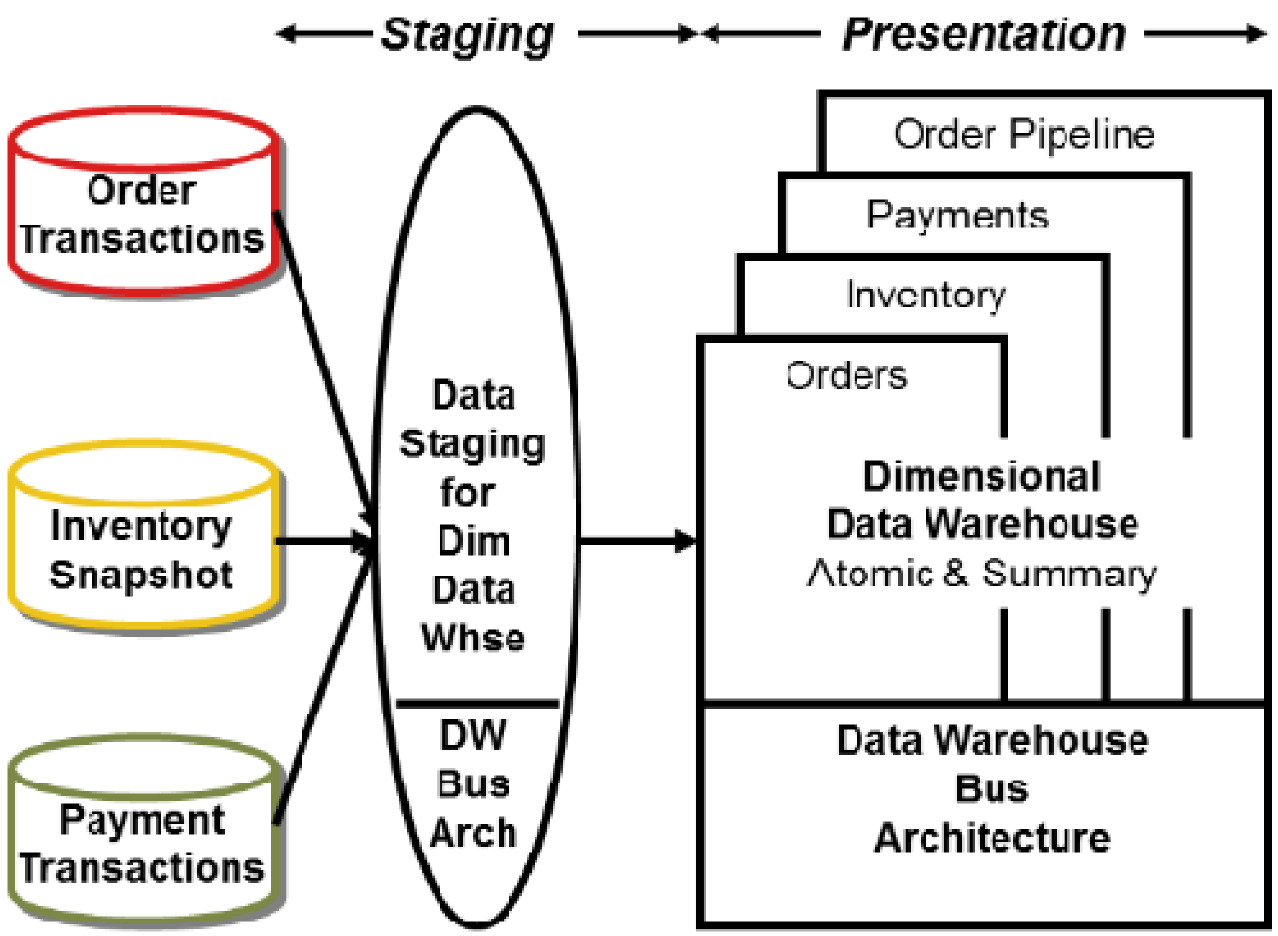
EXERCISE: CLASSICAL DWH ARCHITECTURES

	Kimball Bus Architecture	**Corporate Information Factory**
Advantages	• Two layers only mean faster development and less work • Rather simple approach to make data fast and easily accessible • Lower startup costs (but higher subsequent development costs)	• Separation of concerns: long-term enterprise data storage separated from data presentation • Changes in requirements and scope are easier to manage • Lower subsequent development costs (but higher startup costs)
Disadvantages	• If table structures change (instable source systems), high effort to implement the changes and reload data, especially conformed dimensions ("Dimensionitis" desease) • Non-metric data not optimal for dimensional model • Dimensional model (esp. Star Schema) contains data redundancy	• Data model transformations from 3NF to Dimensional model required • More complex as two different data models are required • Larger team(s) of specialists required

OTHER ARCHITECTURES

- Kimball Bus Architecture (Central data warehouse based on data marts)
- Inmon Corporate Information Factory
- Data Vault 2.0 Architecture (Dan Linstedt)
- DW 2.0: The Architecture for the Next Generation of Data Warehousing
- Virtual Data Warehouse
- Operational Data Store (ODS)

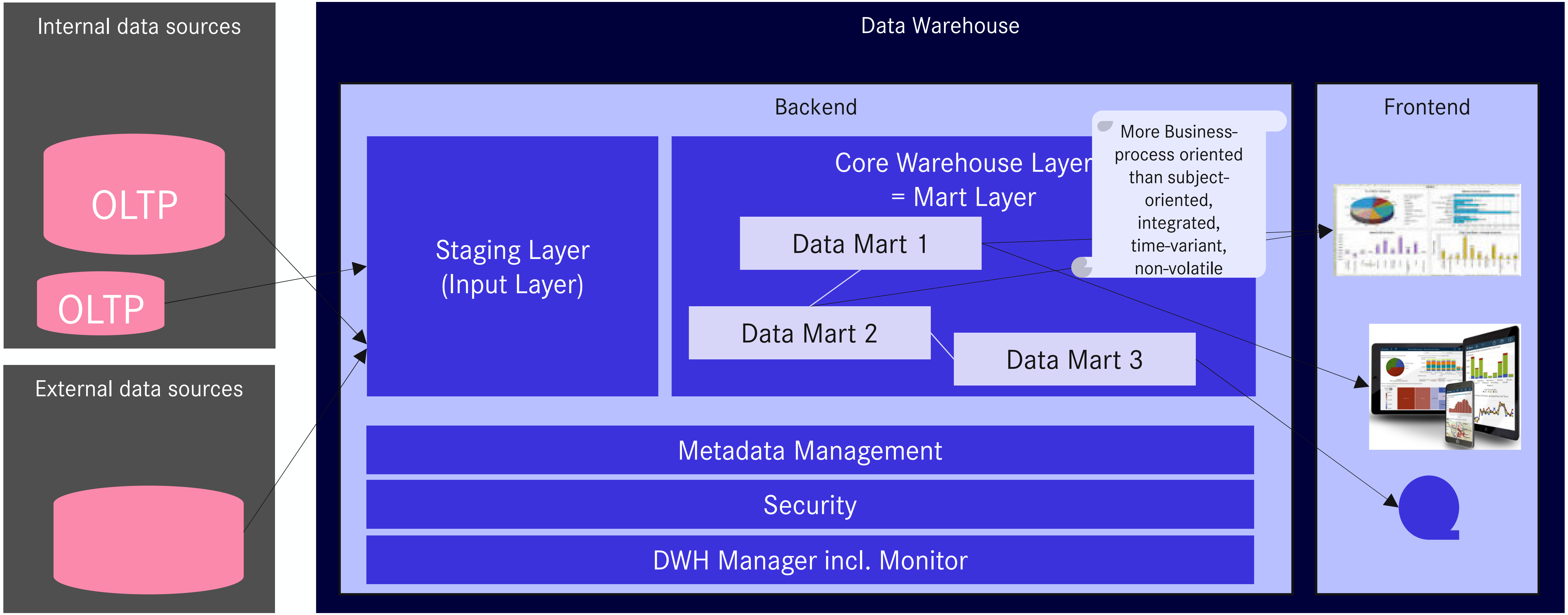
KIMBALL BUS ARCHITECTURE (CENTRAL DATA WAREHOUSE BASED ON DATA MARTS)



Dimensional data warehouse.

Source: http://www.kimballgroup.com/2004/03/differences-of-opinion/

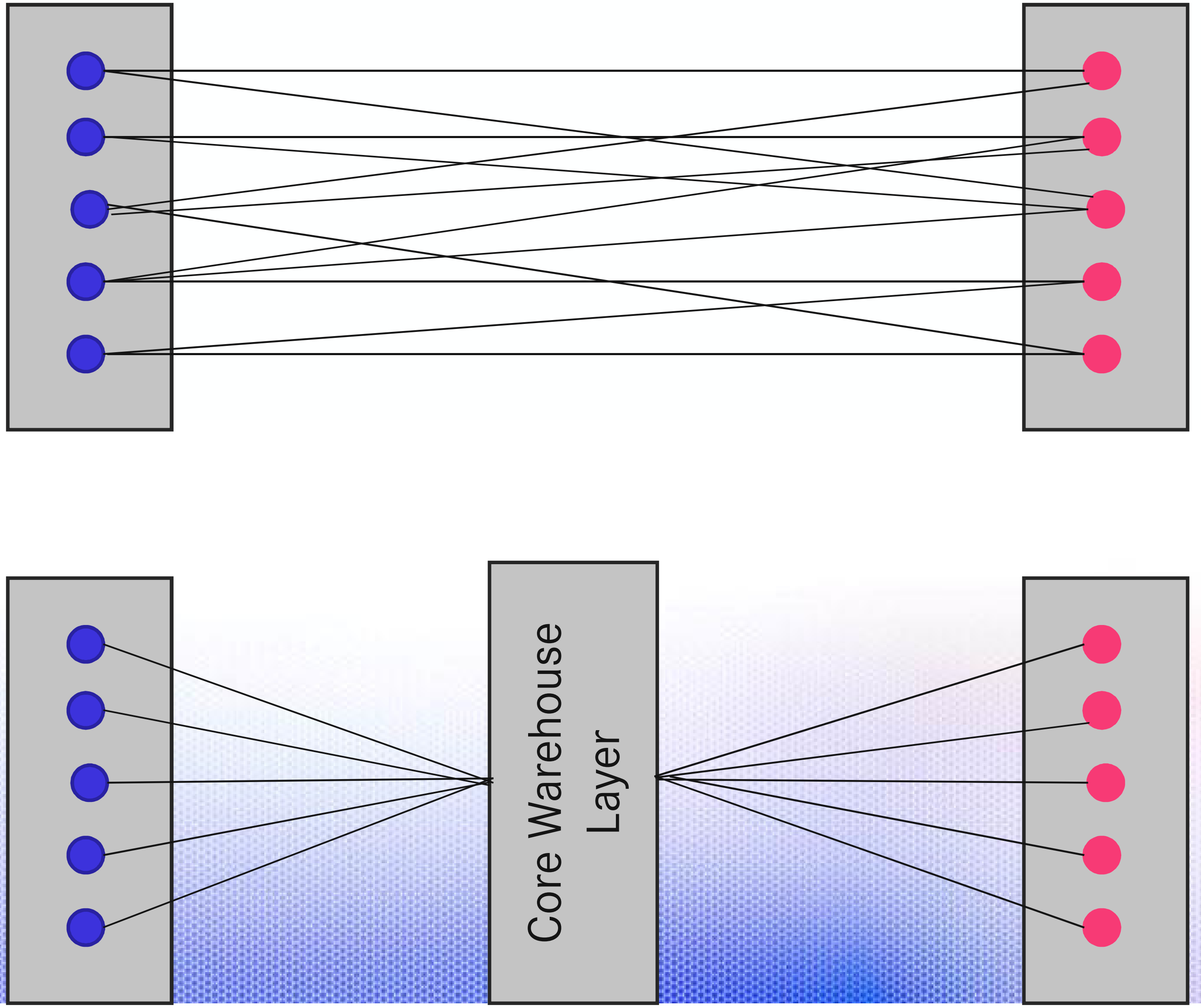
KIMBALL BUS ARCHITECTURE (CENTRAL DATA WAREHOUSE BASED ON DATA MARTS)



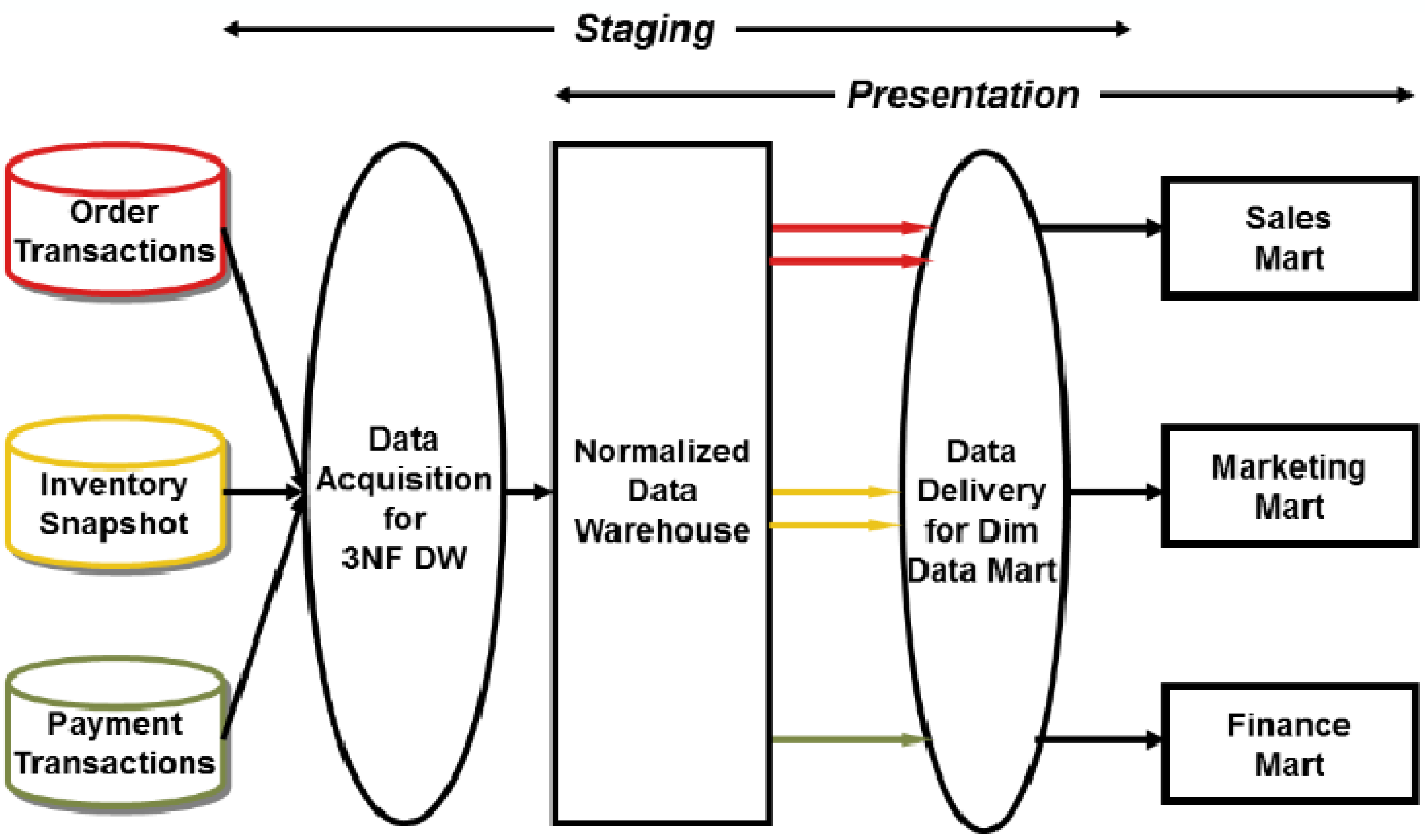
KIMBALL BUS ARCHITECTURE (CENTRAL DATA WAREHOUSE BASED ON DATA MARTS)

- Bottom-up approach
- Dimensional model with denormalized data
- Sum of the data marts constitute the Enterprise DWH
- Enterprise Service Bus / conformed dimensions for integration purposes
  - (don’t confuse with ESB as middleware/communication system between applications)
- Kimball describes that agreeing on conformed dimensions is a hard job and it’s expected that the team will get stuck from time to time trying to align the incompatible original vocabularies of different groups
  - Data marts need to be redesigned if incompatibilities exist

DATA INTEGRATION WITH AND WITHOUT CORE WAREHOUSE LAYER



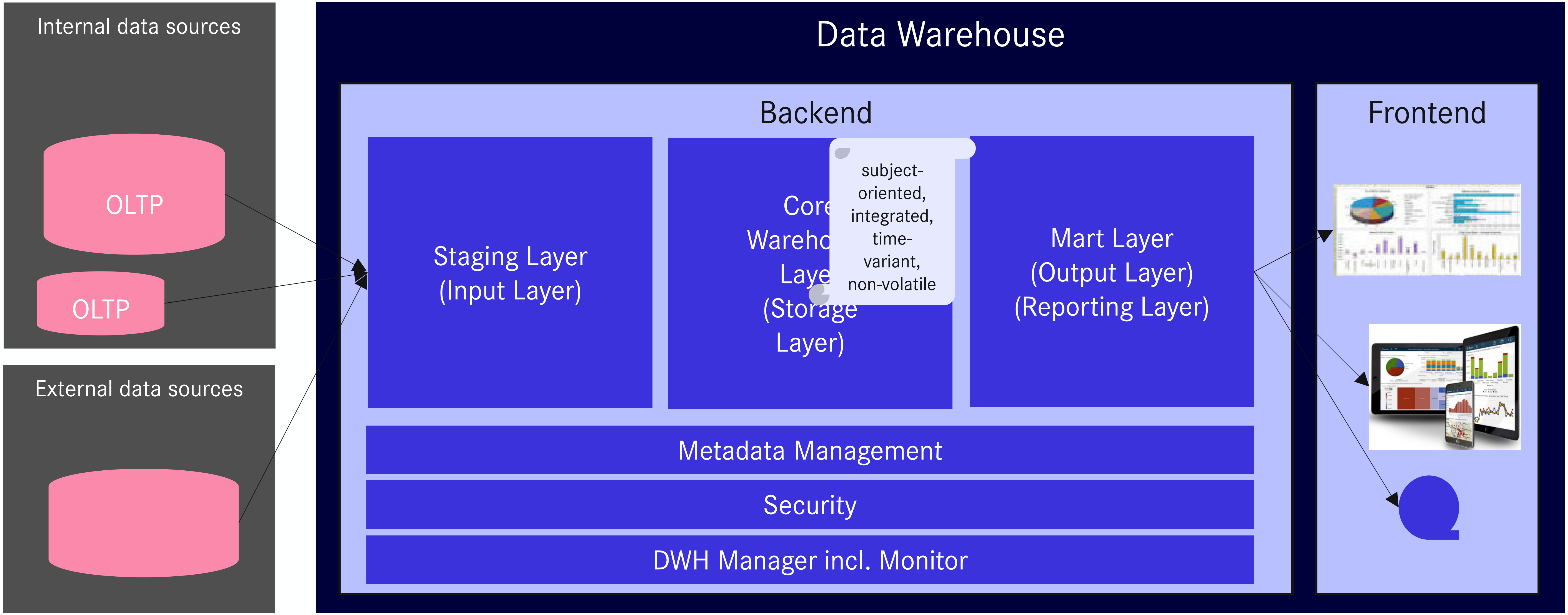
INMON CORPORATE INFORMATION FACTORY



Normalized data warehouse with summary dimensional marts (CIF).

Source: http://www.kimballgroup.com/2004/03/differences-of-opinion/

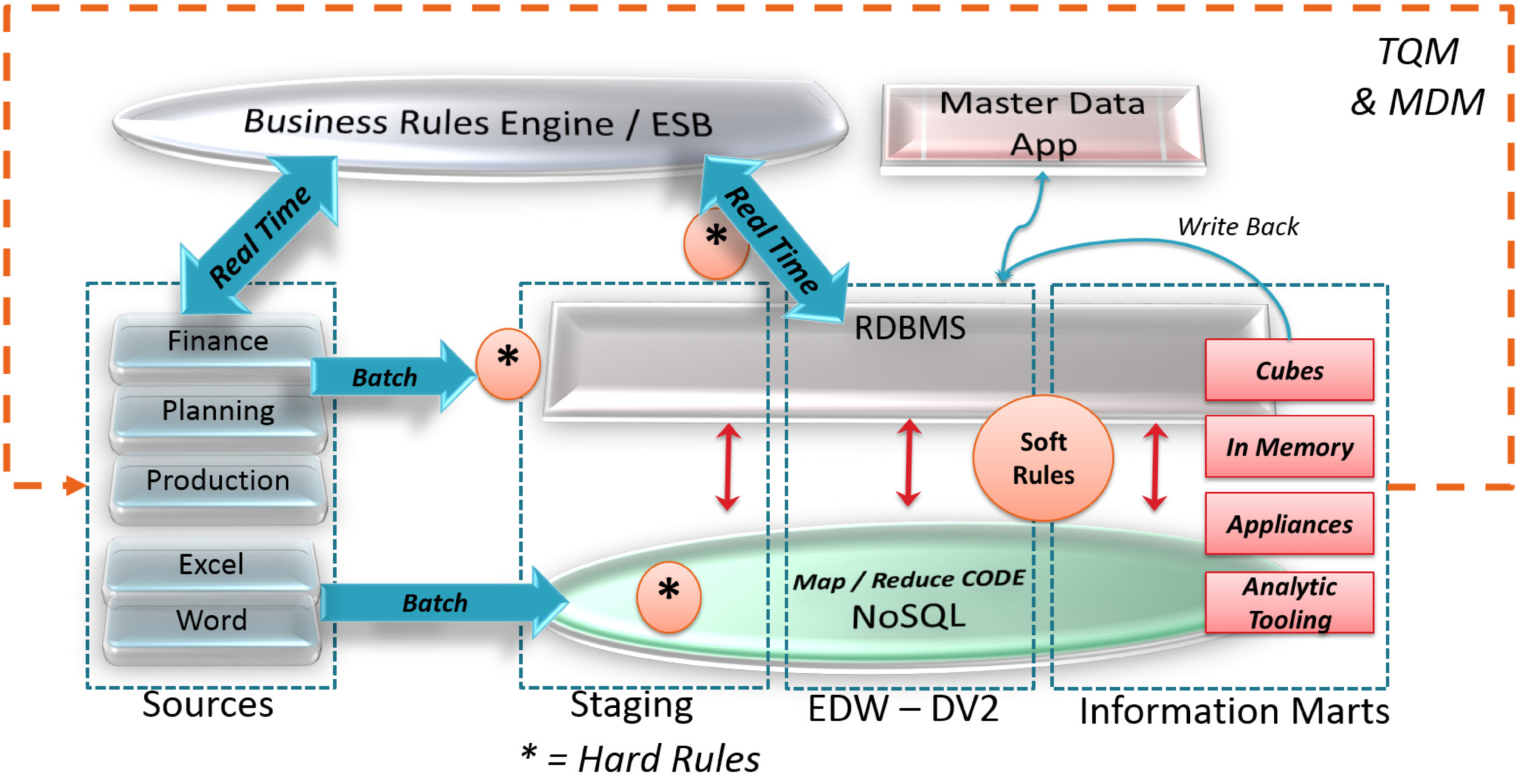
INMON CORPORATE INFORMATION FACTORY



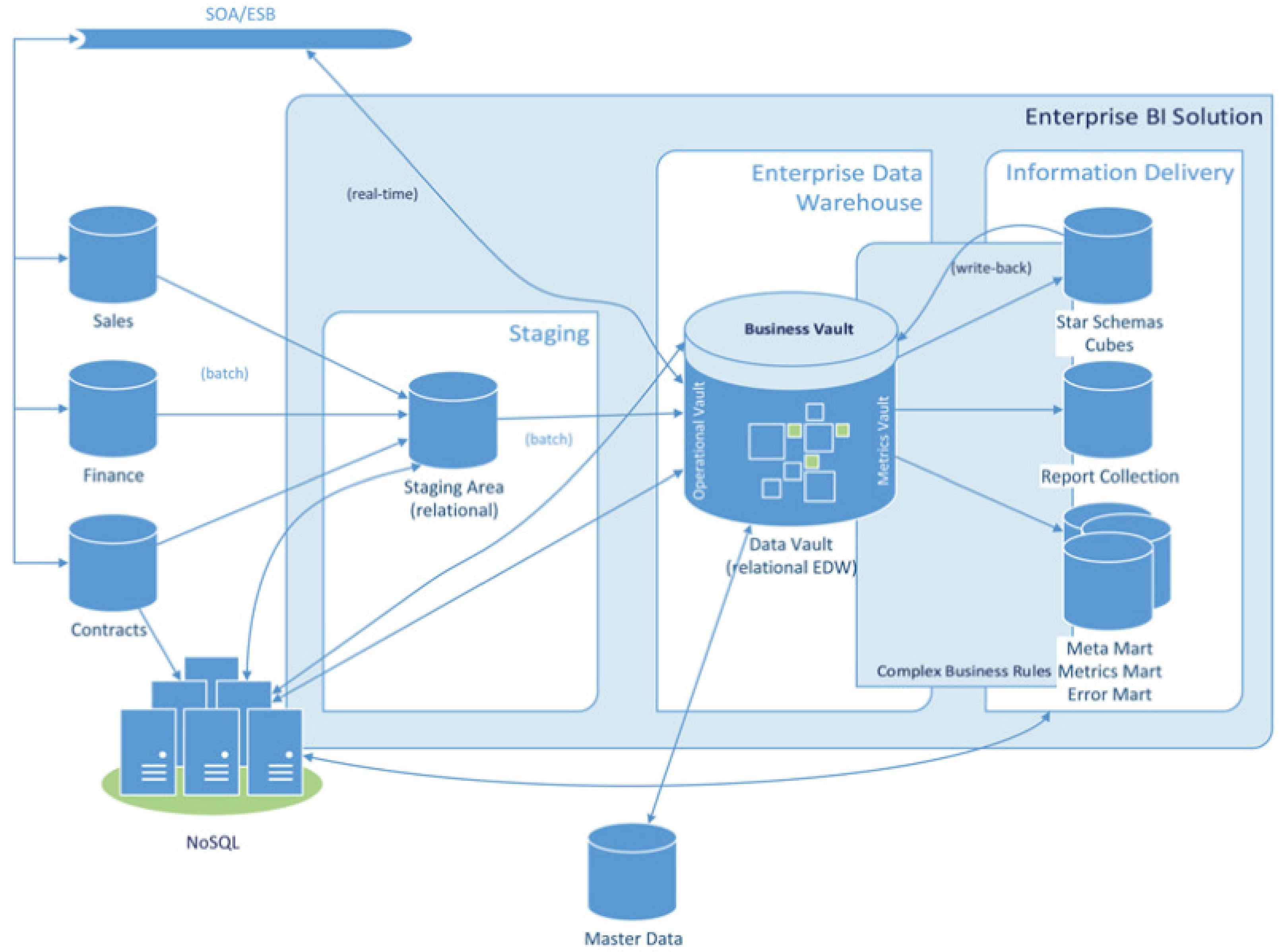
INMON CORPORATE INFORMATION FACTORY

- Top-down approach
- (Normalized) Core Warehouse is essential for subject-oriented, integrated, time-variant and nonvolatile data storage
- Create (departmental) Data Marts as subsets of Core Enterprise DWH as needed
- Data Marts can be designed with Dimensional model
- The logical standard architecture is more general compared to CIF, but was mainly influenced by CIF

DATA VAULT 2.0 ARCHITECTURE – TODAY'S WORLD (DAN LINSTEDT)

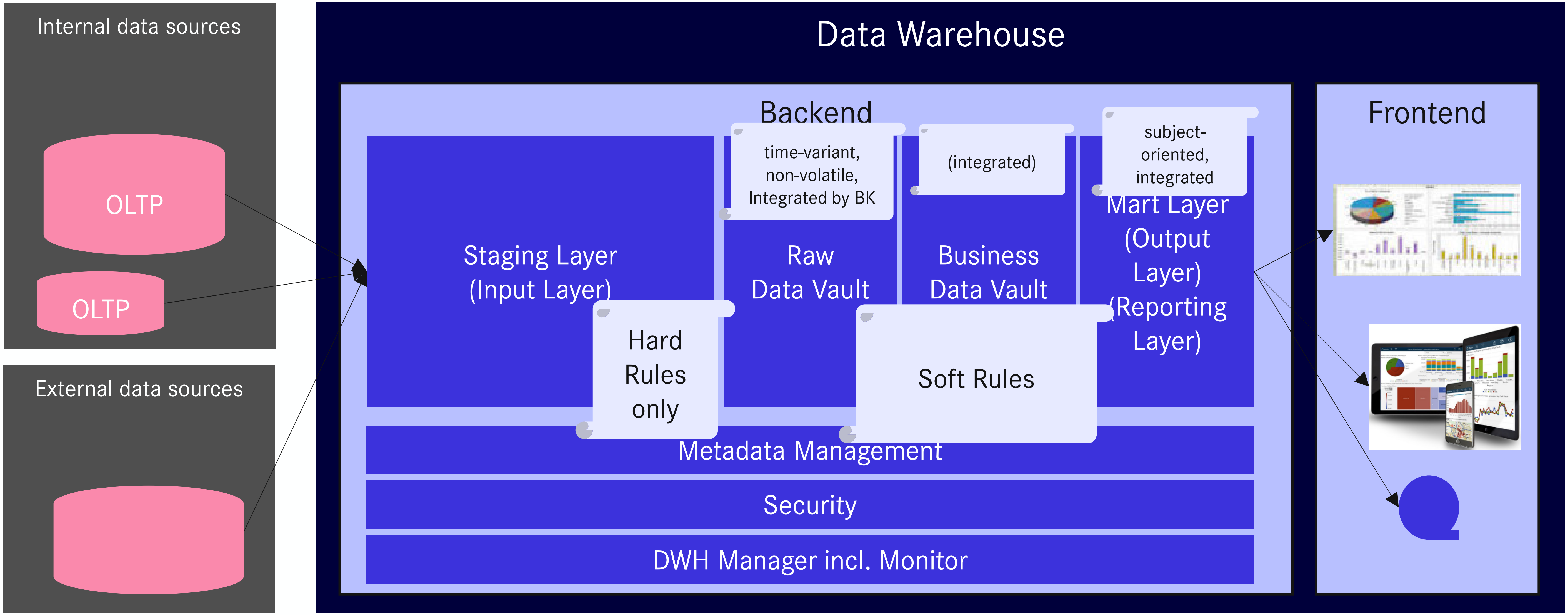


DATA VAULT 2.0 ARCHITECTURE (DAN LINSTEDT)



Michael Olschimke, Dan Linstedt: Building a Scalable Data Warehouse with Data Vault 2.0, Morgan Kaufmann, 2015, Chapter 2.2

DATA VAULT 2.0 ARCHITECTURE (DAN LINSTEDT)



DATA VAULT 2.0 ARCHITECTURE (DAN LINSTEDT)

- Core Warehouse Layer is modeled with Data Vault and integrates data by BK (business key) “only” (Data Vault modeling is a separate lecture)
- Business rules (Soft Rules) are applied from Raw Data Vault Layer to Mart Layer and not earlier
  - Alternatively from Raw Data Vault to additional layer called Business Data Vault
- Hard Rules don’t change data
  - Data is fully auditable
- Real-time capable architecture
- Architecture got very popular recently; also applicable to BigData, NoSQL

DATA VAULT 2.0 ARCHITECTURE (DAN LINSTEDT)

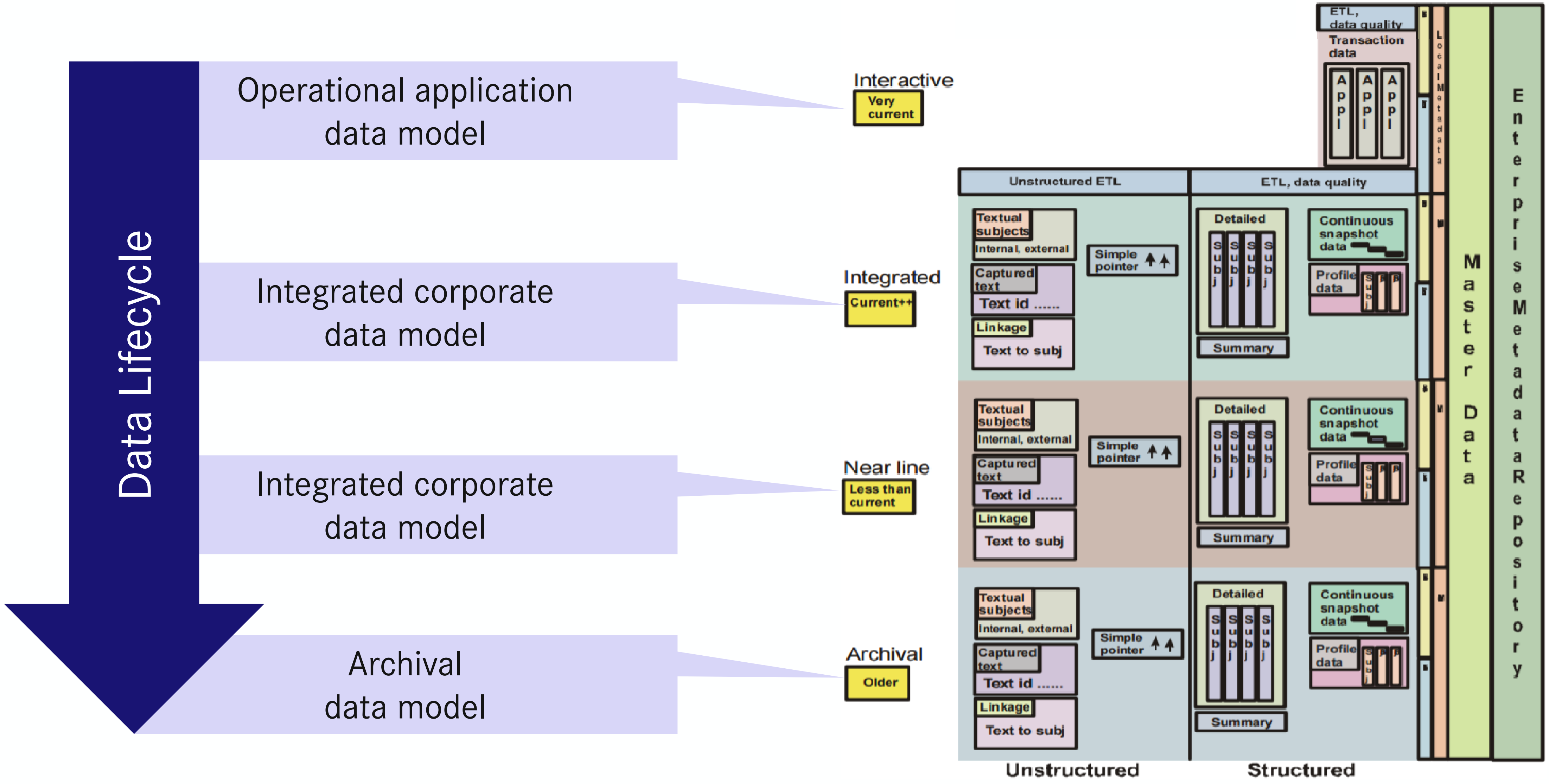
- In the classical DWHs, the Core Warehouse Layer is regarded as "**single version of the truth**"
  - Integrates + cleanses data from different sources and eliminates contradiction
  - Produces consistent results/reports across Data Marts
  - But: cleansing is (still) objective, Enterprises change regularly, paradigm does not scale as more and more systems exist
- Data in Raw Data Vault Layer is regarded as "**Single version of the facts**"
  - 100% of data is loaded 100% of time
  - Data is not cleansed and bad data is not removed in the Core Layer (Raw Vault)

DATA VAULT 2.0 ARCHITECTURE (DAN LINSTEDT)

- Data Vault is optimized for the following requirements:
  - Flexibility
  - Agility
  - Data historization
  - Data integration
  - Auditability

- Bill Inmon wrote in 2008: “Data Vault is the optimal approach for modeling the EDW in the DW2.0 framework.” (DW2.0)

DW 2.0: THE ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING

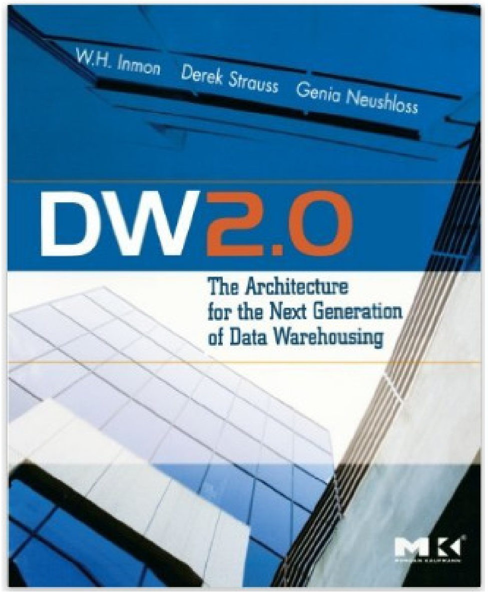


Source: W.H. Inmon, Dan Linstedt: Data Architecture: A Primer for the Data Scientist, Morgan Kaufmann, 2014, chapter 3.1

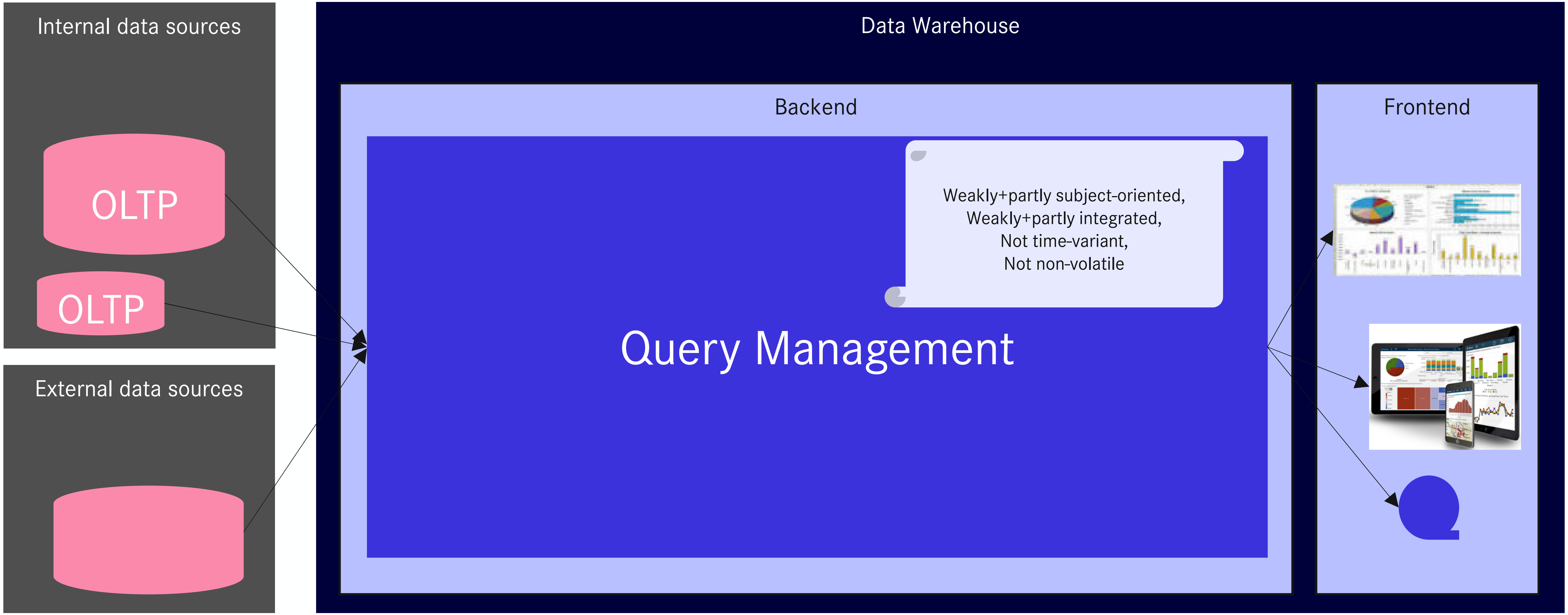
DW 2.0: THE ARCHITECTURE FOR THE NEXT GENERATION OF DATA WAREHOUSING

Main characteristics:

- Structured and “unstructured” data, not just metrics
- Life Cycle of data with different storage areas
    - Hot data: High speed, expensive storage (RAM, SSDs) for most recent data
    - …
    - Cold data: Low speed, inexpensive storage (e.g. hard disks) for old data; archival data model with high compression
- Metadata is an integral part of the DWH and not an afterthought



VIRTUAL DATA WAREHOUSE



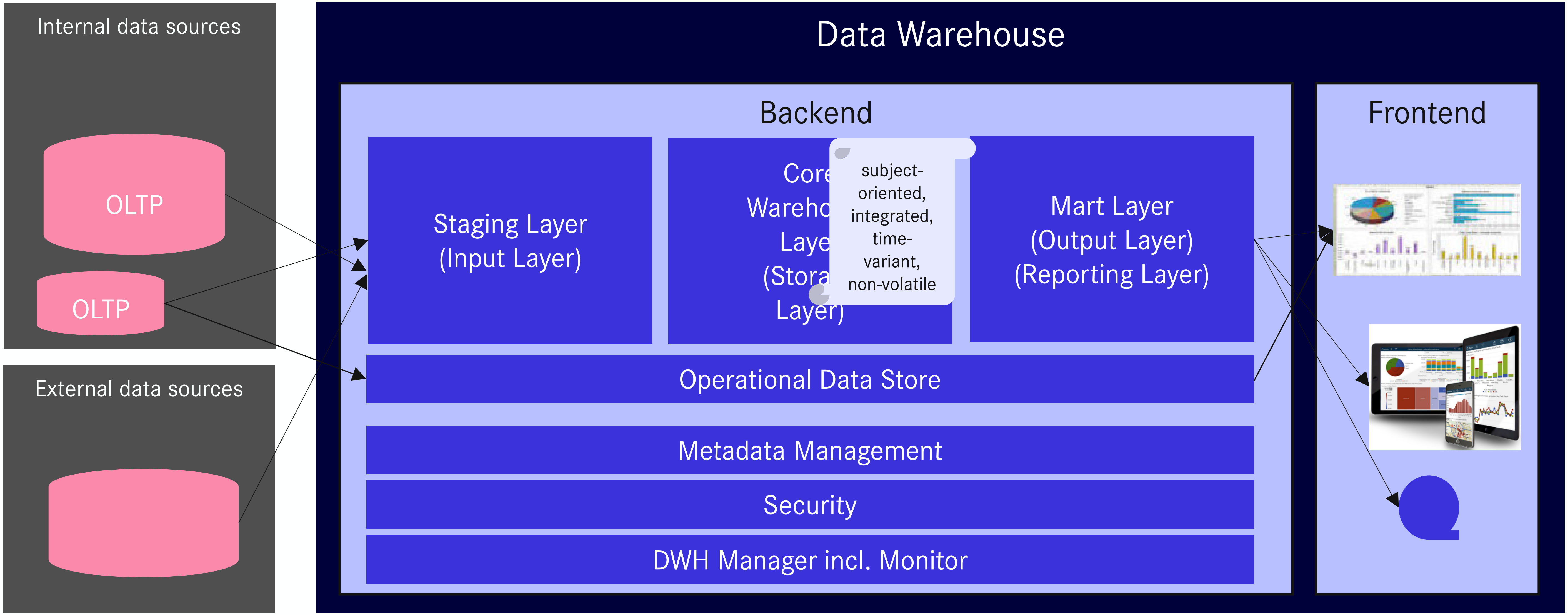
VIRTUAL DATA WAREHOUSE

- Data not extracted from operational systems and stored separately
- Standardized interface for all operational data sources
  - One "GUI" for all existing data
  - Generates combined queries
  - Query Processor joins query result data from different sources
  - Can also access data in Hadoop (Polybase, Big SQL, BigData SQL, etc)

VIRTUAL DATA WAREHOUSE

- Query Management manages metadata about all operational systems
  - (physical) location of data and algorithms for extracting data from OLTP system
  - Implementation easier
  - Low cost: can use existing hardware infrastructure
- Queries cause significant performance problems in operational systems
- Known problems when analyzing operational data directly
- Same query is processed multiple times (if queried multiple times)
- Same query delivers different results when processed at different times

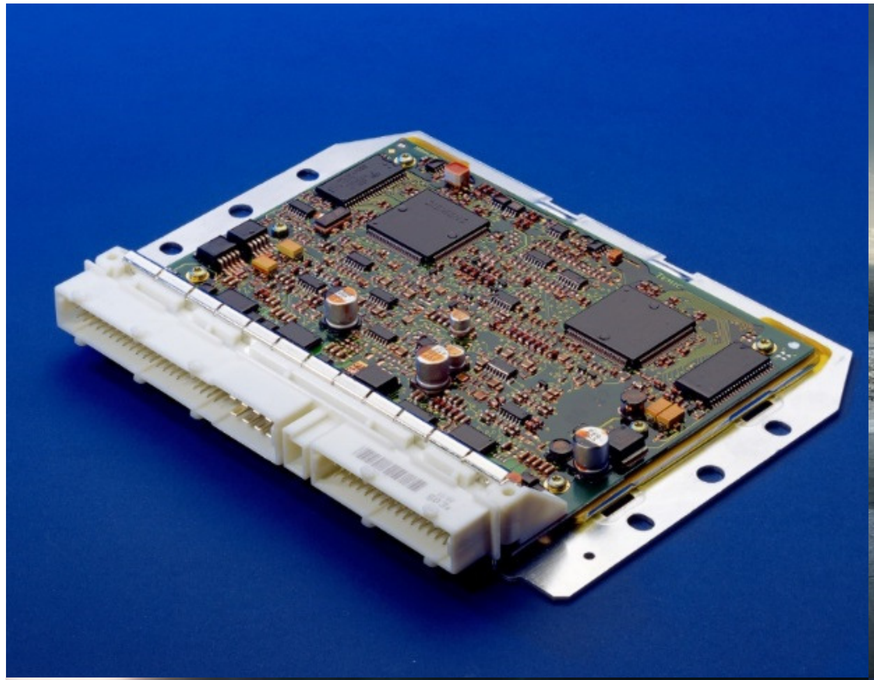
OPERATIONAL DATA STORE (ODS)

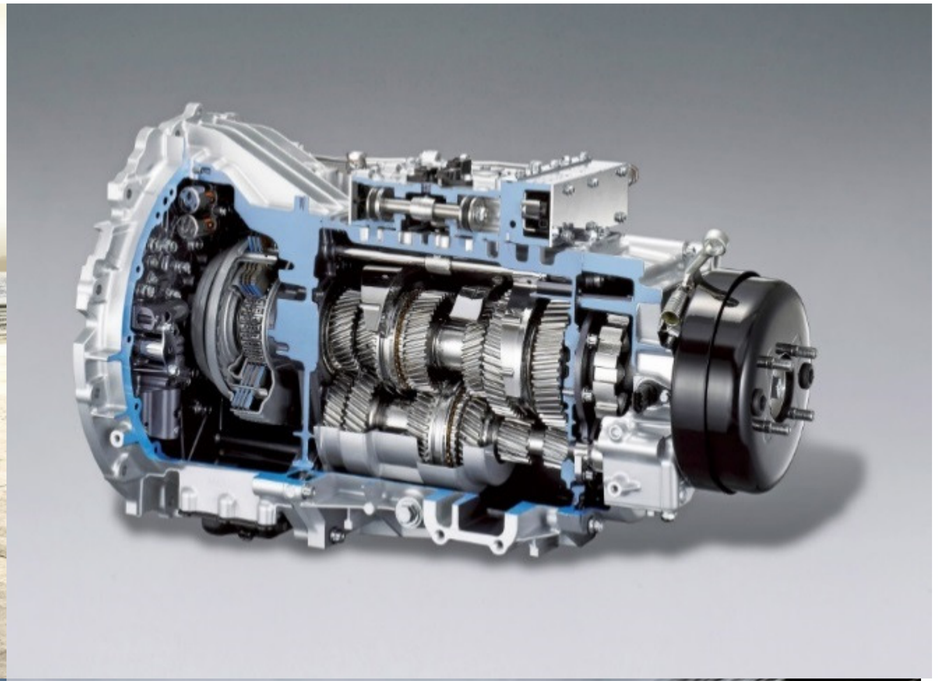


OPERATIONAL DATA STORE (ODS)

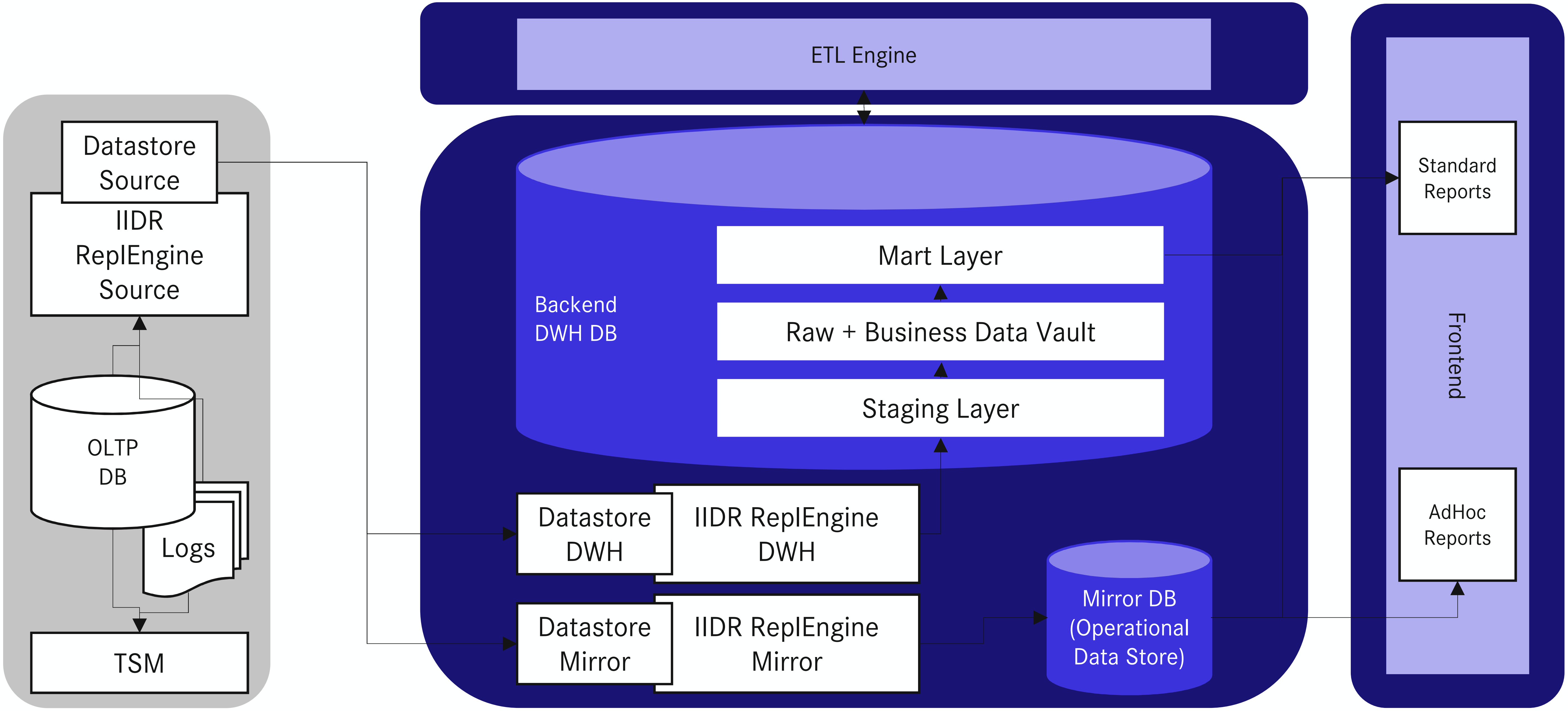
- ODS: Real-time/Right-time layer
- Replication techniques used to transport data from source database to ODS layer with minimal impact on source system
- Data in the ODS has no history and is stored without any cleansing and without any integration (1:1 copy from single source)
- DWH performance not optimal as data model is suited for OLTP and not for reporting requirements
- ODS normally additionally to Staging / Core DWH / Mart Layer but can exist alone without other layers

EXAMPLE DWH FOR STATE OF CONSTRUCTION DOCU






ARCHITECTURE FROM AN ACTUAL PROJECT IN THE AUTOMOTIVE INDUSTRY



END USER SAMPLE QUESTIONS

Which vehicles or aggregates are documented incompletely? (Data quality)

Which vehicles / which control units require SW updates?

Which interiors are most common by region?

Supply data for external simulations, customs clearance, spare part planning, etc.

EXERCISE: RECOMMEND AN ARCHITECTURE

Review the presented data warehouse architectures.

Which architecture would you recommend for

- A holding of 3 telecommunication companies
- An online store with real/right-time data integration needs
- Marketing department of a bank

List advantages and drawbacks of your proposal.

EXERCISE: RECOMMEND AN ARCHITECTURE

A holding of 3 telecommunication companies

- Architecture: Virtual Data Warehouse
- + Companies may not want to provide their data to a new storage
- + Can easily be extended if new companies join the holding or reduced if a company leaves the holding
- - Bad performance
- - Not really data integration achieved, low Data Quality
- - Firewalls have to be opened

EXERCISE: RECOMMEND AN ARCHITECTURE

An online store with real-time/right-time data integration needs

- Architecture: Data Vault 2.0
- + Integration of many internal and external source systems (e.g. integrate social media data about the online store)
- + Fast data delivery in Raw Vault Layer (Real-time/Right-time data integration). Complex data cleansing / transformation / soft rules are delayed downstream towards Mart Layer
- - Transformation overhead (Source system data model to Data Vault data model to Dimensional data model)

EXERCISE: RECOMMEND AN ARCHITECTURE

Marketing department of a bank

- Architecture: Kimball Bus architecture
- \+ Start small for a department. If other departments are interested, new data and new Marts can be added on demand
- \- High risk to loose the Enterprise view and several DWHs are built

That's still quite a common scenario nowadays. A single Enterprise DWH is often not achieved (e.g. Mergers & Acquisitions, inflexibility due to a single centralized DWH, rapidly changing conditions, etc.) and therefore very often several DWHs with different architectures exist in parallel within a company.

EXERCISE - INTRODUCTION AND DWH ARCHITECTURE GROUP TASK

- Now imagine that you prepare an exam.
- Identify 1-3 questions about DWH architecture (and/or DWH introduction) that you would ask in an exam.
- Write down the questions on stick-it cards.

SUMMARY

Which layers does the logical standard architecture have?

- Staging (Input), Integration (Cleansing), Core Warehouse (Storage), Aggregation, Mart (Reporting, Output) and additionally Metadata, Security, DWH Manager, Monitor

Which other architectures exist?

- Kimball Bus Architecture (Central data warehouse based on data marts)
- Inmon Corporate Information Factory
- Data Vault 2.0 Architecture (Dan Linstedt)
- DW 2.0: The Architecture for the Next Generation of Data Warehousing
- Virtual Data Warehouse
- Operational Data Store (ODS)

SUPPLEMENT: POLYGLOT DATA ARCHITECTURE

ARCHITECTURES AROUND BUZZWORDS LIKE BIG DATA, STREAMING & DATA LAKES

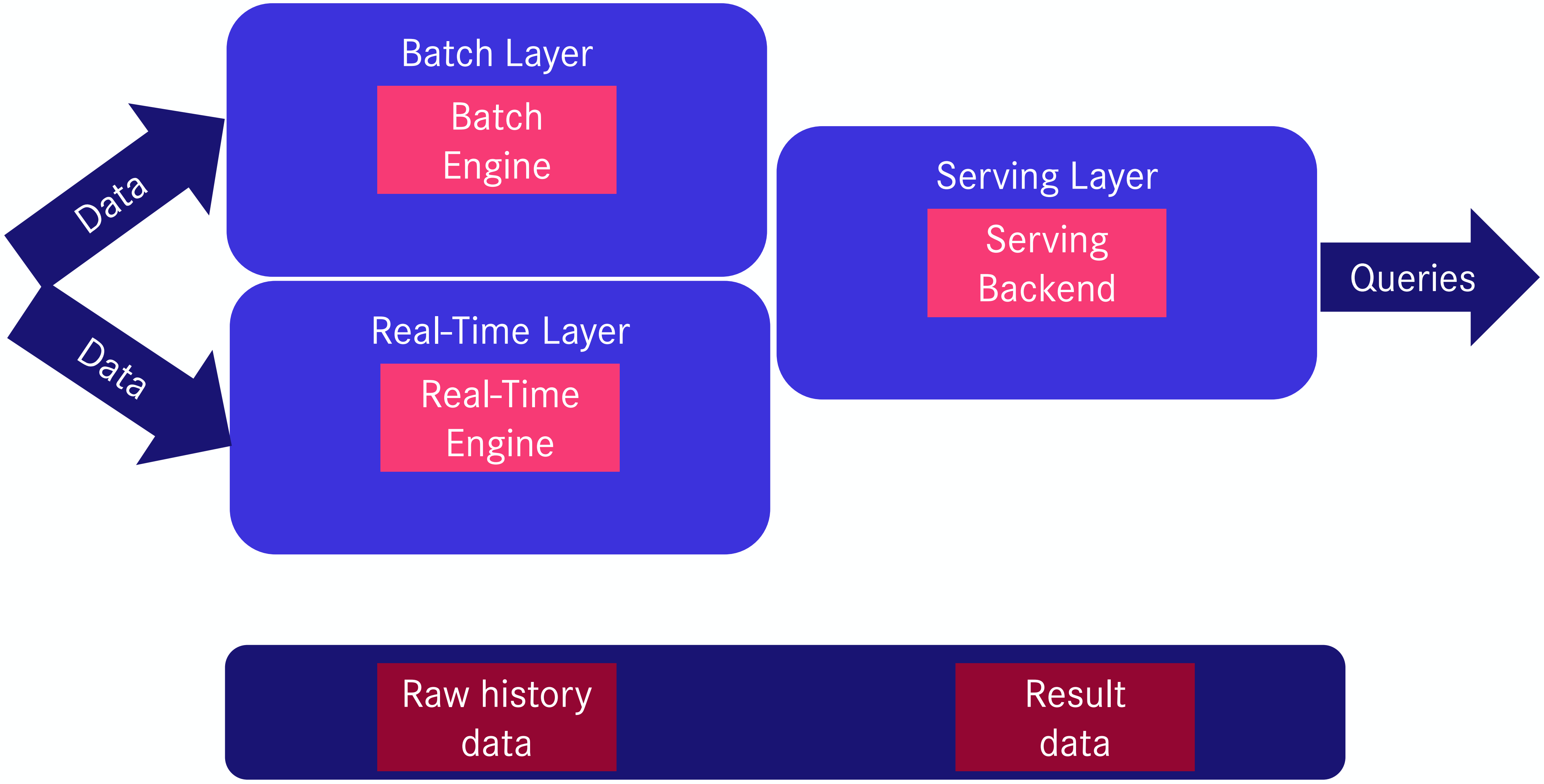
BIG DATA / DATA LAKE ARCHITECTURES
INTRODUCTION

- There exist well-known reference architectures for Data Warehouses
- Many tools and schema-on-read came with the Hadoop ecosystem
  - Was a “black box” at the beginning
  - Gets more and more structure with different layers instead of a “black box”
  - Structure, modeling, organization, governance instead of tool-only focus
- The slides provide some architectures with links to more information

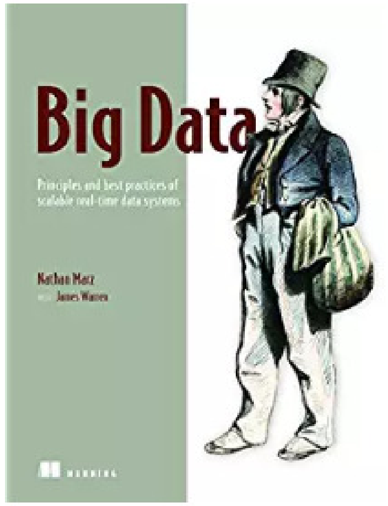
LAMBDA ARCHITECTURE

- Architecture by Nathan Marz
- Realtime and batch processing
- Batch layer stores and historizes raw data
- Serving layer has to union batch and realtime layer
- Rather complex
- Author recommends graph data model and advises against schema-on-read

LAMBDA ARCHITECTURE



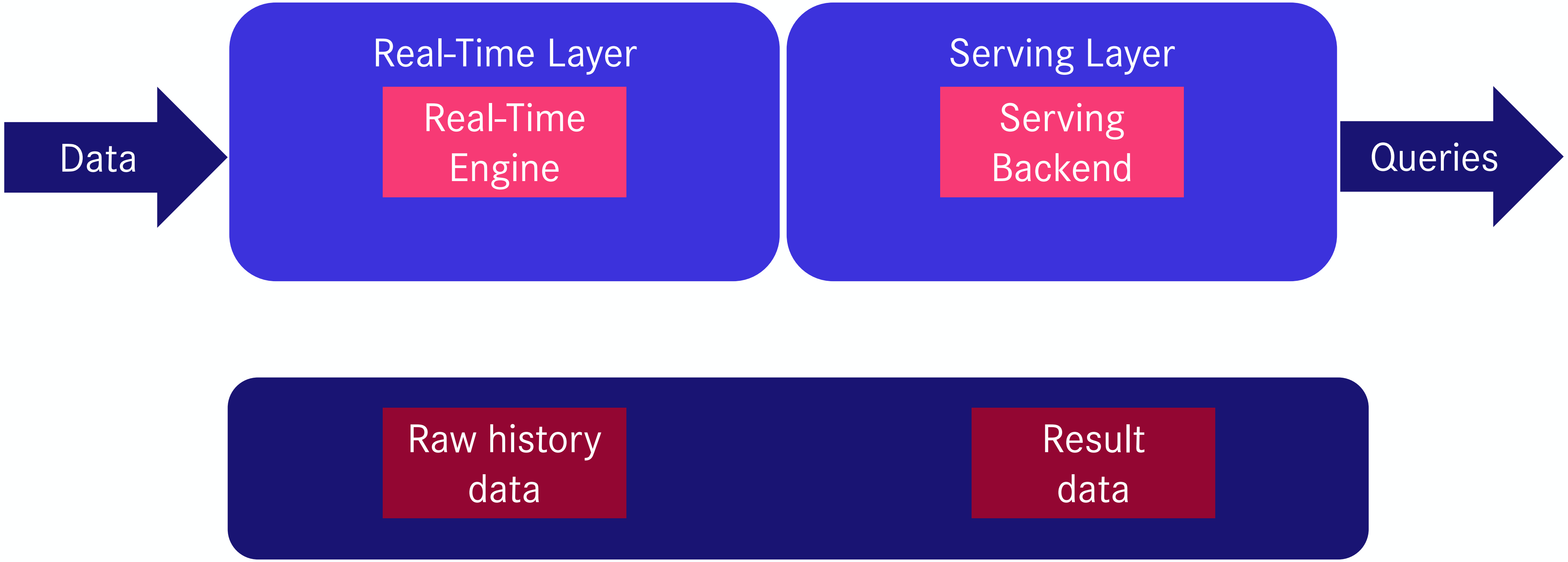
Source:



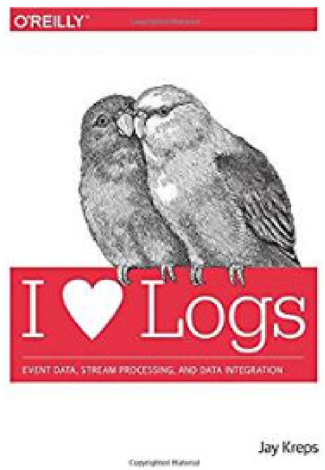
KAPPA ARCHITECTURE

- Architecture by Jay Kreps
- Logcentric, write-ahead logging
- Each event is an immutable log entry and is added to the end of the log
- Read and write operations are separated
- Materialized views can be recomputed consistently from data in the log

KAPPA ARCHITECTURE



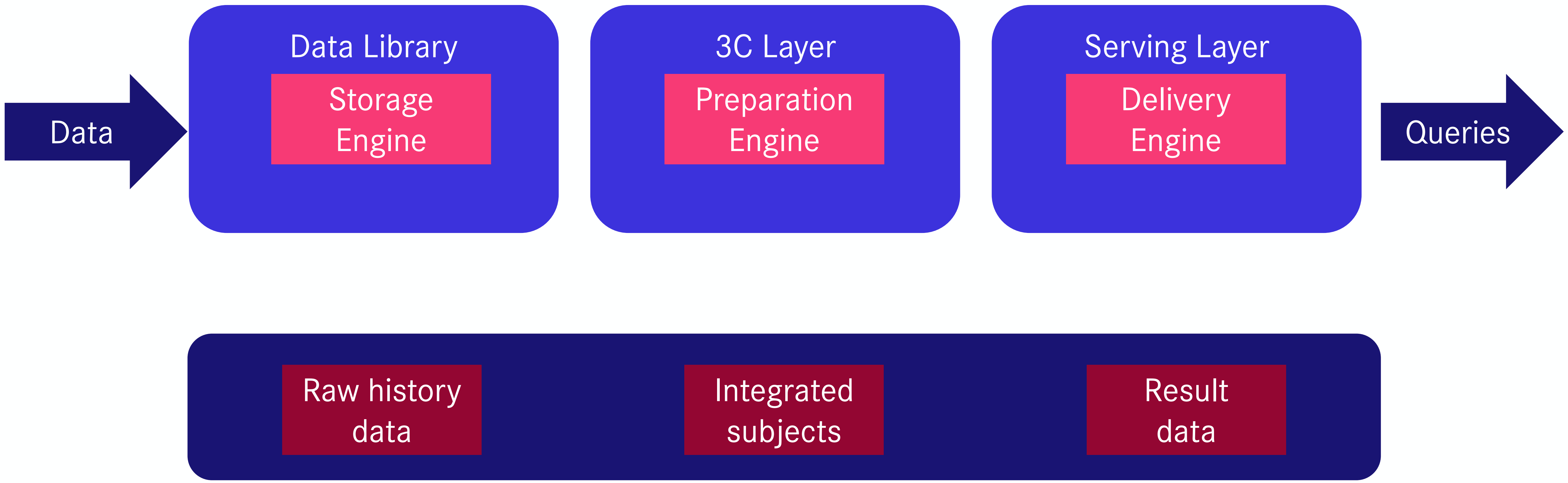
Source:



PS-3C ARCHITECTURE

- Architecture by Rogier Werschkull
- Store incoming data in Data Library Layer (Persistent staging = PS)
- Prepare data in a 3C layer for “Concept – Context – Connector”-model
- Concept + Connector can be virtualized on data in Data Library Layer

PS-3C ARCHITECTURE

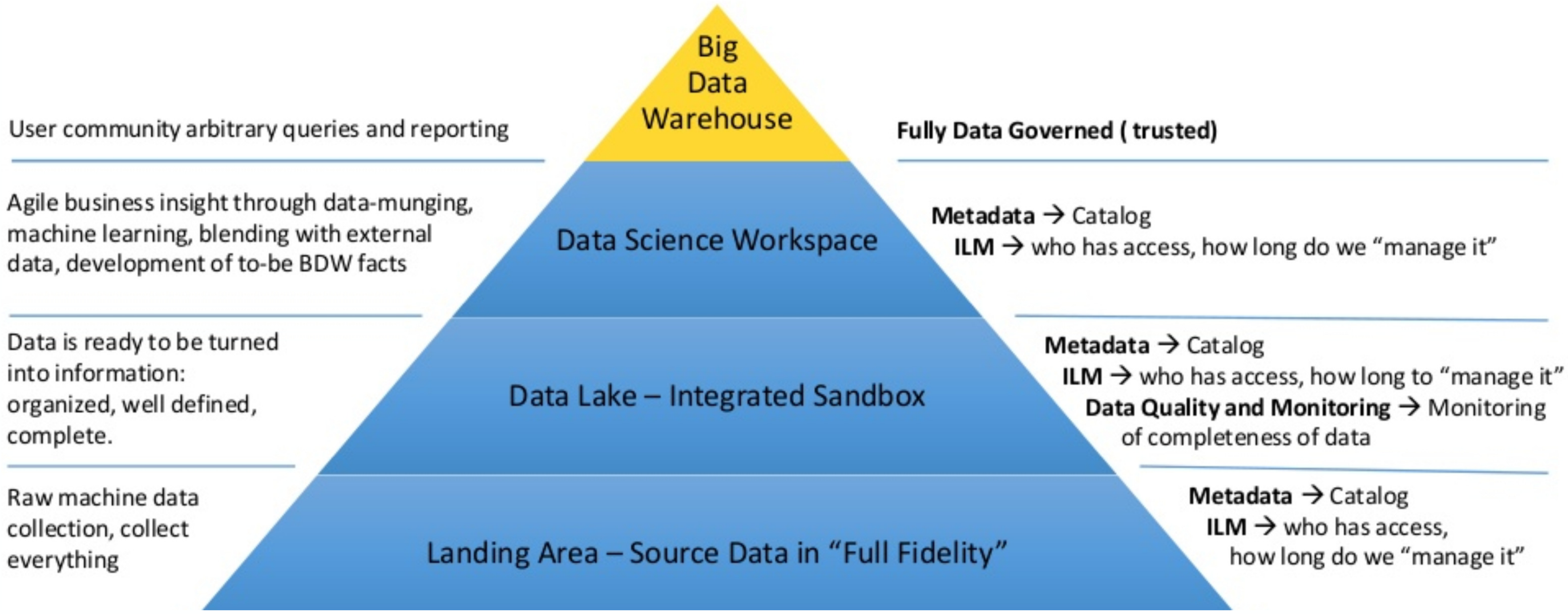


Source: TDWI 2016

POLYGLOT WAREHOUSE

- Architecture by Joe Caserta
- Big Data Warehouse may live in one or more platforms on premise or in the cloud
  - Hadoop only
  - Hadoop + MPP or RDBMS
  - Additionally NoSQL or Search

POLYGLOT WAREHOUSE

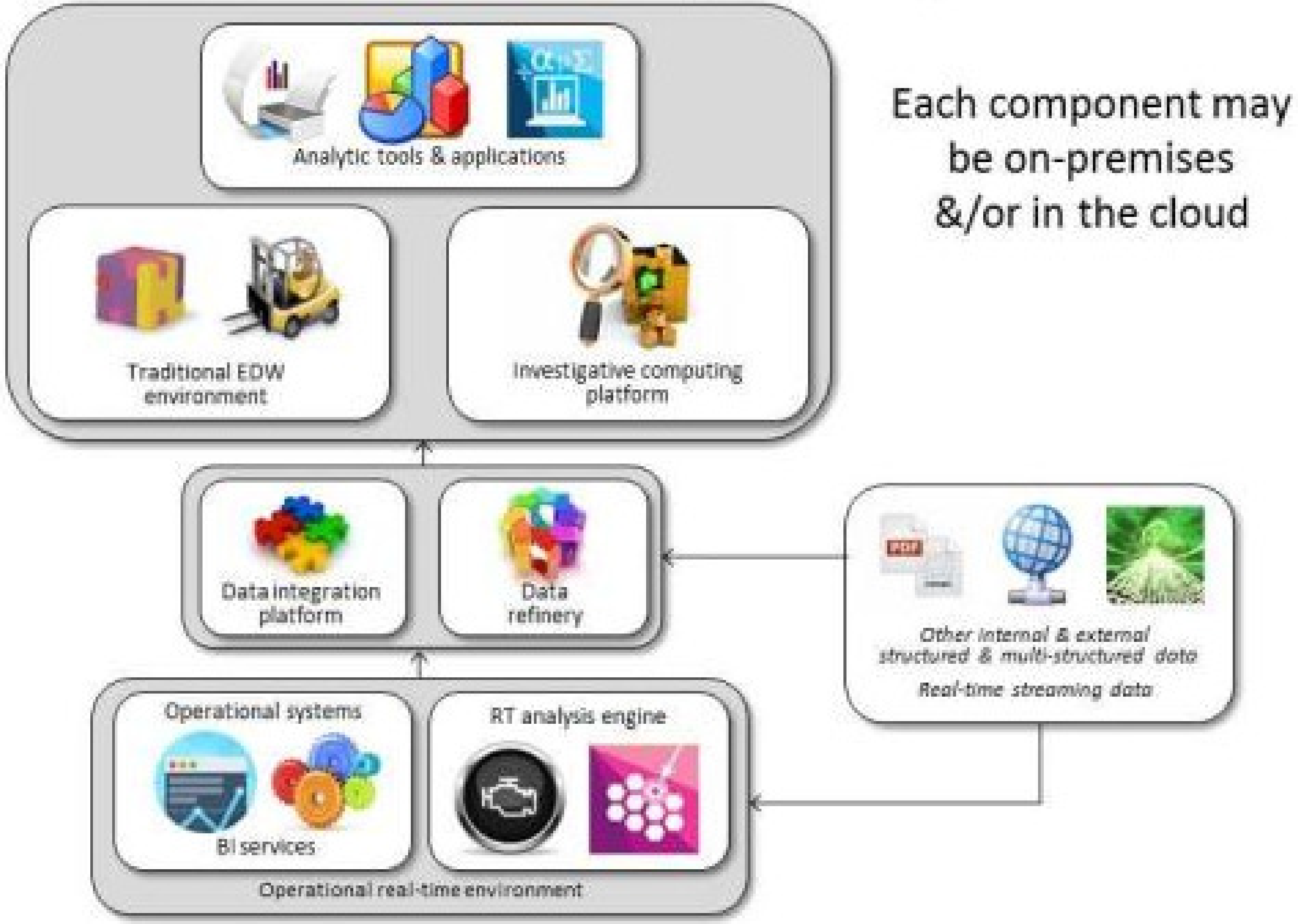


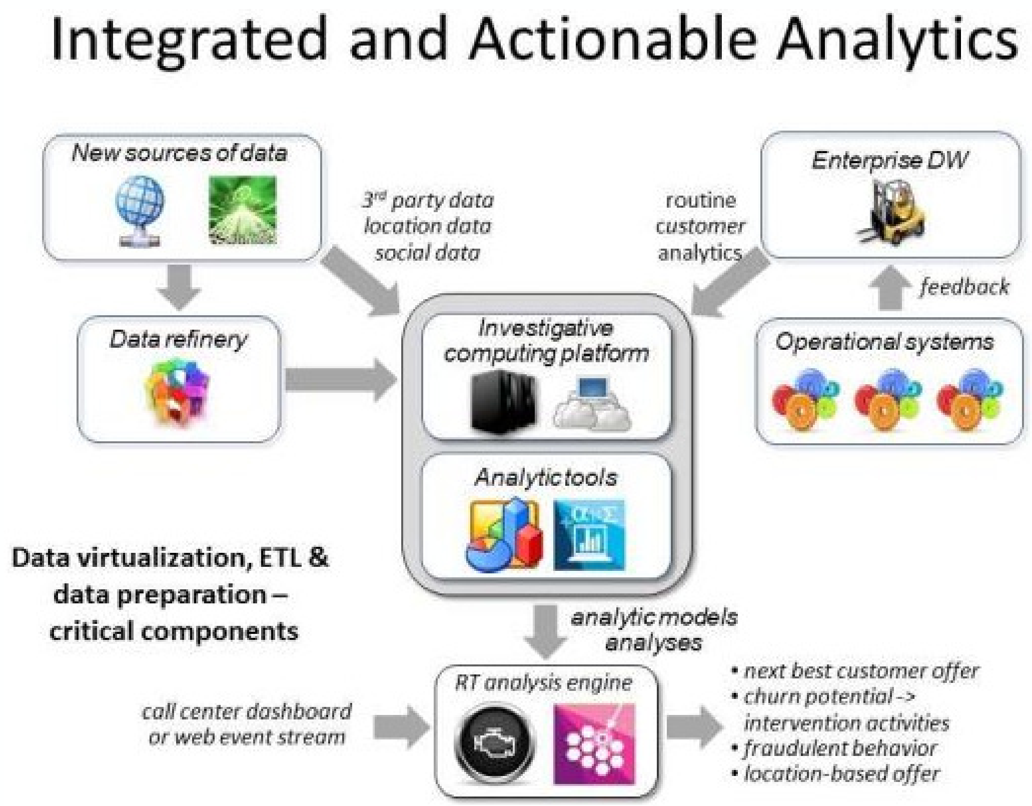
Source: https://www.slideshare.net/CasertaConcepts/hadoop-and-your-data-warehouse

THE EXTENDED DATA WAREHOUSE ARCHITECTURE (XDW) THE ENTERPRISE ANALYTICS ARCHITECTURE

- Architecture by Claudia Imhoff
- combine the stability and reliability of the BI architectures while embracing new and innovative technologies and techniques
- 3 components that extend the EDW environment
    - Investigative computing platform
    - Data refinery
    - Real-time (RT) analysis platform

THE EXTENDED DATA WAREHOUSE ARCHITECTURE (XDW) THE ENTERPRISE ANALYTICS ARCHITECTURE





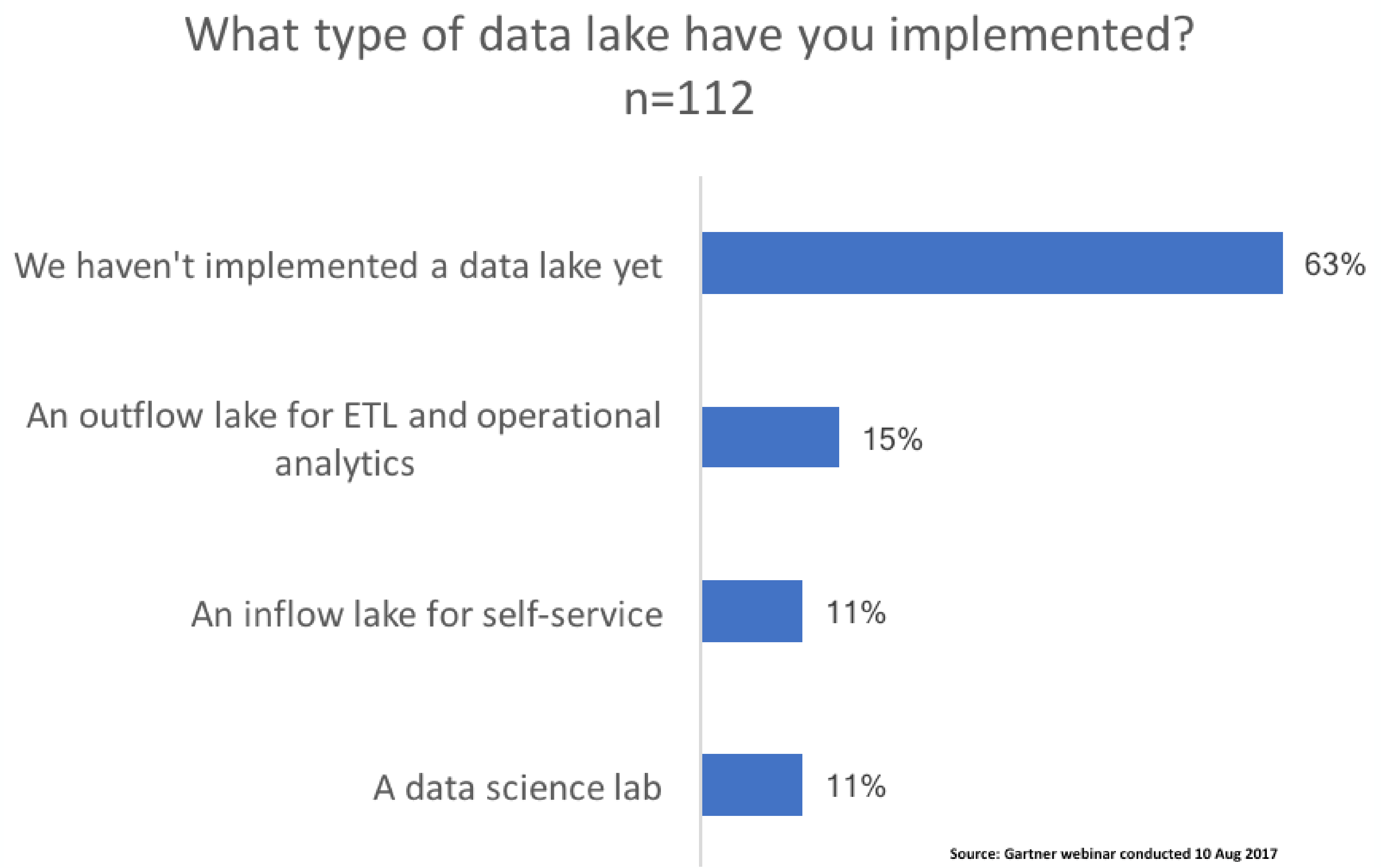
Source: https://upside.tdwi.org/articles/2016/03/15/extending-traditional-data-warehouse.aspx

GARTNER DATA LAKE ARCHITECTURE STYLES

- Inflow Lake: accommodates a collection of data ingested from many different sources that are disconnected outside the lake but can be used together by being colocated within a single place
- Outflow Lake: a landing area for freshly arrived data available for immediate access or via streaming. It employs schema-on-read for the downstream data interpretation and refinement.
- Data Science Lab: most suitable for data discovery and for developing new advanced analytics models

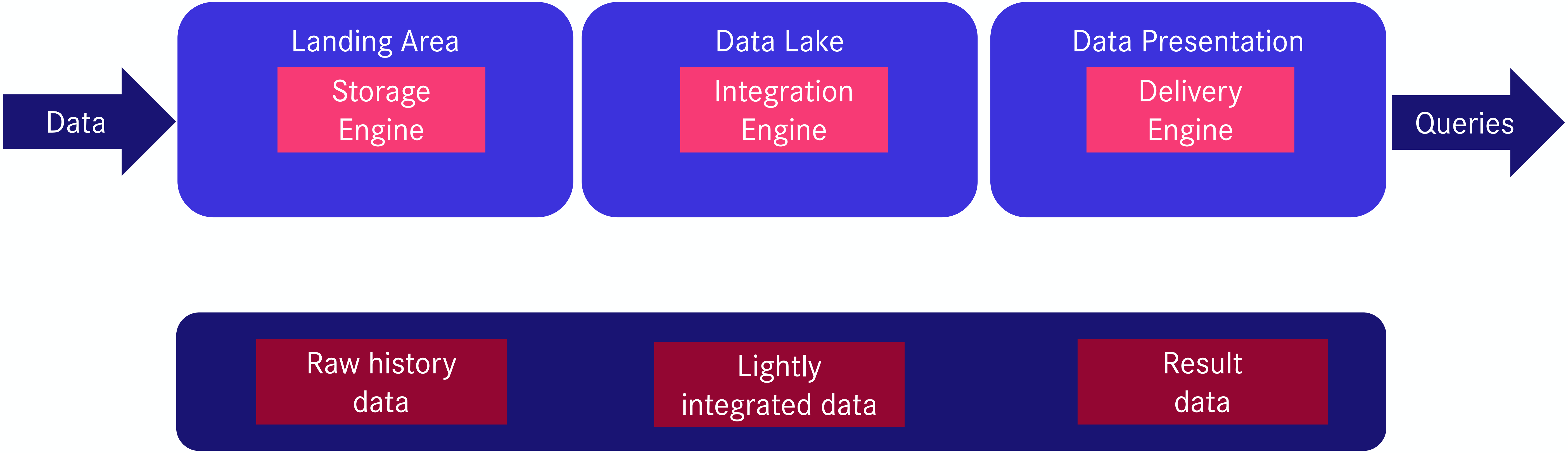
Source: http://blogs.gartner.com/nick-heudecker/data-lake-webinar-recap/

GARTNER DATA LAKE ARCHITECTURE STYLES



Source: http://blogs.gartner.com/nick-heudecker/data-lake-webinar-recap/

SUMMARY


Data
Landing Area
Storage Engine
Data Lake
Integration Engine
Data Presentation
Delivery Engine
Queries
Raw history data
Lightly integrated data
Result data

THANK YOU

Daimler TSS

Daimler TSS GmbH
Wilhelm-Runge-Straße 11, 89081 Ulm / Telefon +49 731 505-06 / Fax +49 731 505-65 99
tss@daimler.com / Internet: www.daimler-tss.com/ Intranet-Portal-Code: @TSS
Domicile and Court of Registry: Ulm / HRB-Nr.: 3844 / Management: Christoph Röger (CEO), Steffen Bäuerle