Curriculum for

Certified Professional for Software Architecture (CPSA)®
Advanced Level

Module
DATA

Designing, building, and maintaining data-centric software architectures

2024.1-RC3-EN-20240705

International Software Architecture
Qualification Board

Table of Contents

The abbreviation "e. V." is part of the iSAQB’s official name and stands for "eingetragener Verein" (registered association), which describes its status as a legal entity according to German law. For the purpose of simplicity, iSAQB e. V. shall hereafter be referred to as iSAQB without the use of said abbreviation.

List of Learning Goals

Introduction: General information about the iSAQB Advanced Level

What is taught in an Advanced Level module?

- The iSAQB Advanced Level offers modular training in three areas of competence with flexibly designable training paths. It takes individual inclinations and priorities into account.
- The certification is done as an assignment. The assessment and oral exam is conducted by experts appointed by the iSAQB.

What can Advanced Level (CPSA-A) graduates do?

CPSA-A graduates can:

- Independently and methodically design medium to large IT systems
- In IT systems of medium to high criticality, assume technical and content-related responsibility
- Conceptualize, design, and document actions to achieve quality requirements and support development teams in the implementation of these actions
- Control and execute architecture-relevant communication in medium to large development teams

Requirements for CPSA-A certification

- Successful training and certification as a Certified Professional for Software Architecture, Foundation Level® (CPSA-F)
- At least three years of full-time professional experience in the IT sector; collaboration on the design and development of at least two different IT systems
  - Exceptions are allowed on application (e.g., collaboration on open source projects)
- Training and further education within the scope of iSAQB Advanced Level training courses with a minimum of 70 credit points from at least three different areas of competence
- Successful completion of the CPSA-A certification exam

CPSA-A®

Certified Professional for Software Architecture
Advanced Level

Essentials

What does the module “DATA” convey?

The module presents DATA to the participants … At the end of the module, the participants know … and are able to …

Curriculum Structure and Recommended Durations

Content	Recommended minimum duration (minutes)
1. Motivation and Overview	90
2. Reference architectures for analytical systems	210
3. Data Sources	60
4. Ingestion	90
5. Storage	90
6. Queries	60
7. Transformation	90
8. Usage and Analysis	60
9. Data Pipelines	90
10. Data Mesh	120
11. Data Governance	120
Total	1080 (18h)

Duration, Teaching Method and Further Details

The times stated below are recommendations. The duration of a training course on the DATA module should be at least 3 days, but may be longer. Providers may differ in terms of duration, teaching method, type and structure of the exercises, and the detailed course structure. In particular, the curriculum provides no specifications on the nature of the examples and exercises.

Licensed training courses for the DATA module contribute the following credit points towards admission to the final Advanced Level certification exam:

Methodical Competence:	20 Points
Technical Competence:	10 Points
Communicative Competence:	0 Points

Prerequisites

Participants **should** have the following prerequisite knowledge:

- Prerequisite 1
- Prerequisite 2, etc.

Knowledge in the following areas may be **helpful** for understanding some concepts:

- Area 1:
  - Knowledge 1
  - Experience 2
  - Knowledge 3
  - Experience 4
  - Understanding 5

Structure of the Curriculum

The individual sections of the curriculum are described according to the following structure:

- **Terms/principles**: Essential core terms of this topic.
- **Teaching/practice time**: Defines the minimum amount of teaching and practice time that must be spent on this topic or its practice in an accredited training course.
- **Learning goals**: Describes the content to be conveyed including its core terms and principles.

This section therefore also outlines the skills to be acquired in corresponding training courses.

Supplementary Information, Terms, Translations

To the extent necessary for understanding the curriculum, we have added definitions of technical terms to the iSAQB glossary and complemented them by references to (translated) literature.

1. Motivation and overview

Duration: XXX min	Practice time: XXX min

1.1. Terms and Principles

Term 1, Term 2, Term 3

1.2. Learning Goals

LG 1-1: Data engineering

tbd.

LG 1-2: Differentiation beween operative and analytical data

tbd.

LG 1-3: Categories of data analysis

tbd.

LG 1-4: Challenges of analytical applications

tbd.

LG 1-5: Roles in data engineering

tbd.

LG 1-6: Monolitic and distributed data architectures

tbd.

LG 1-7: Lifecycle of analytical data

tbd.

1.3. References

[E. F. Codd 1990], [W. H. Inmon 2005], [R. Kimball 2011]

2. Reference architectures for analytical application systems

Duration: XXX min	Practice time: XXX min

2.1. Terms and Principles

Term 1, Term 2, Term 3

2.2. Learning Goals

LG 2-1: Overview architectural patterns

tbd.

LG 2-2: Architectural patterns for for unifying analytical data

tbd.

LG 2-3: Architecture decisions based on architectural patterns

tbd.

2.3. References

[J. Reis 2022], [M. Bornstein 2020], [W. H. Inmon 2005], [R. Kimball 2011], [D. Linstedt 2015], [C. Giebler et al. 2021], [P. Pääkkönen 2015]

3. Data Sources

Duration: XXX min	Practice time: XXX min

3.1. Terms and Principles

Term 1, Term 2, Term 3

3.2. Learning Goals

LG 3-1: Types of data sources and source systems

tbd.

LG 3-2: Properties of data sources and source systems

tbd.

LG 3-3: Provisioning of data through application-specific APIs

tbd.

LG 3-4: Provisioning of data through database systems

tbd.

LG 3-5: Provisioning of data through file systems

tbd.

LG 3-6: Provisioning of data through object stores

tbd.

LG 3-7: Provisioning of data through Message Queues and Event-Streaming

tbd.

3.3. References

[R. Castagna 2022]

4. Ingestion und Transport

Duration: XXX min	Practice time: XXX min

4.1. Terms and Principles

Term 1, Term 2, Term 3

4.2. Learning Goals

LG 4-1: Definition Data Ingestion

tbd.

LG 4-2: Identifying entities

tbd.

LG 4-3: Detecting changes

tbd.

LG 4-4: Connectors

tbd.

LG 4-5: Characteristics of Data Ingestion

tbd.

LG 4-6: Batch vs Stream Ingestion

tbd.

LG 4-7: Meta Data Ingestion

tbd.

4.3. References

[J. Reis 2022]

5. Storage

Duration: XXX min	Practice time: XXX min

5.1. Terms and Principles

Term 1, Term 2, Term 3

5.2. Learning Goals

LG 5-1: Storage systems

tbd.

LG 5-2: Database systems

tbd.

LG 5-3: Database systems for analytical applications

tbd.

LG 5-4: Concurrency Control

tbd.

LG 5-5: Versioning of data

tbd.

LG 5-6: Optimization and scaling

tbd.

LG 5-7: Data models for analytical data

tbd.

LG 5-8: Data Warehouse and Data Lake

tbd.

5.3. References

[starke]

6. Query und Processing

Duration: XXX min	Practice time: XXX min

6.1. Terms and Principles

Term 1, Term 2, Term 3

6.2. Learning Goals

LG 6-1: Analytical queries

tbd.

LG 6-2: Query programming models

tbd.

LG 6-3: Query processing & optimization

tbd.

6.3. References

[starke]

7. Transformation

Duration: XXX min	Practice time: XXX min

7.1. Terms and Principles

Term 1, Term 2, Term 3

7.2. Learning Goals

LG 7-1: Definition Data Transformation

tbd.

LG 7-2: Applications

tbd.

LG 7-3: Typical transformations

tbd.

LG 7-4: Staging Area

tbd.

LG 7-5: Robust transformations

tbd.

LG 7-6: Quality levels

tbd.

LG 7-7: Batch processing

tbd.

LG 7-8: Stream processing

tbd.

7.3. References

[starke]

8. Serving Data

Duration: XXX min	Practice time: XXX min

8.1. Terms and Principles

Term 1, Term 2, Term 3

8.2. Learning Goals

LG 8-1: Use Cases

tbd.

LG 8-2: Representation of mass data

tbd.

LG 8-3: Modularization

tbd.

LG 8-4: Data Analytics and Business Intelligence

tbd.

LG 8-5: Machine Learning

tbd.

LG 8-6: Reverse ETL

tbd.

9. Data Pipelines

Duration: XXX min	Practice time: XXX min

9.1. Terms and Principles

Term 1, Term 2, Term 3

9.2. Learning Goals

LG 9-1: Definition Data Pipelines

tbd.

LG 9-2: Applications of Data Pipelines

tbd.

LG 9-3: Types of Data Pipelines

tbd.

LG 9-4: Quality criteria for data pipelines

tbd.

LG 9-5: Building Blocks of Data Pipelines

tbd.

LG 9-6: Technologies and platforms for data pipelines

tbd.

LG 9-7: Operation of data pipelines

tbd.

9.3. References

[H. Varshney 2023], [E. Levy 2021], [B. Singhal 2022]

10. Data Mesh

Duration: XXX min	Practice time: XXX min

10.1. Terms and Principles

Term 1, Term 2, Term 3

10.2. Learning Goals

LG 10-1: Disadvantages of central data architectures

tbd.

LG 10-2: Definition Data Mesh

tbd.

LG 10-3: Domain Ownership

tbd.

LG 10-4: Data as a Product

tbd.

LG 10-5: Self-serve Data Platform

tbd.

LG 10-6: Federated Computational Governance

tbd.

LG 10-7: Top down vs. bottom up realization

tbd.

10.3. References

[J. Christ et al. 2018], [J. Majchrzak 2022], [Z. Dehghani 2023]

11. Cross Cutting Concerns

Duration: 120 min	Practice time: 20 min

11.1. Terms and Principles

Data Management, Data Governance, Data Contracts, Data Ownership, Data Quality, Data Security, Anonymization, Pseudonymization, Personalization, Metadata, Responsibility, DataOps

11.2. Learning Goals

LG 11-1: Definition

tbd.

LG 11-2 - Privacy, Compliance, Data Security

tbd.

LG 11-3 - Data Quality

tbd.

LG 11-4 - Data Access and Privileges

LG 11-5 - Data Stewardship und Ownership

tbd.

LG 11-6 - Data Contracts

tbd.

LG 11-7 - Policies

tbd.

LG 11-8 - Metadata

tbd.

LG 11-9 - Operational aspects

tbd.

11.3. References

[J. Reis 2022]

References

This section contains references that are cited in the curriculum.

A

- [R. Agarwal] R. Agarwal: Kafka Connectors — All you need to know to start using connectors in Kafka. https://medium.com/@the_infinity/kafka-connectors-all-you-need-to-know-to-start-using-connectors-in-kafka-d905cf8a371c, [Online; Stand: 2.06.2023].

B

- [F. Bachmann et al. 2000] F. Bachmann, L. Bass, J. Carriere, P. Clements, D. Garlan, J. Ivers, R. Nord, and R. Little: Software architecture documentation in practice: Documenting architectural layers. tech. rep., Carnegie-Mellon University Pittsburgh PA Software Engineering Inst, 2000.
- [M. Bornstein 2020] M. Bornstein, J. Li, and M. Casado: Emerging Architectures for Modern Data Infrastructure. https://a16z.com/emerging-architectures-for-modern-data-infrastructure, 2020.

C

- [E. F. Codd 1990] E. F. Codd: The relational model for database management: version 2. Addison-Wesley Longman Publishing Co., Inc., 1990.
- [R. Castagna 2022] R. Castagna: Strukturierte und unstrukturierte Daten: Die Unterschiede. https://www.computerweekly.com/de/feature/Strukturierte-und-unstrukturierte-Daten-Die-Unterschiede, November 2022, [Online; Stand: 2.06.2023].
- [B. Carnes 2020] B. Carnes: Basic SQL Commands - The List of Database Queries and Statements You Should Know. https://www.freecodecamp.org/news/basic-sql-commands/, 2020.
- [J. Christ et al. 2018] J. Christ, L. Visengeriyeva, S. Harrer: Data Mesh Architecture - Data Mesh From an Engineering Perspective. https://www.datamesh-architecture.com, 2022.

D

- [K. Dutta 2015] K. Dutta and M. Jayapal: Big Data Analytics for Real Time Systems. 02 2015.
- [Z. Dehghani 2023] Z. Dehghani, J. Christ, and S. Harrer: Data Mesh: Eine dezentrale Datenarchitektur entwerfen. O’Reilly Media, Inc., 2023.

E

- [W. Eckerson 2015] W. Eckerson: Which Data Warehouse Automation Tool is Right for You?. https://www.eckerson.com/register?content=which-data-warehouse-automation-tool-is-right-for-you, 2015.

G

- [C. Giebler et al. 2021] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, and B. Mitschang: The data lake architecture framework. BTW 2021, 2021.
- [M. Grellmann 2022] M. Grellmann: Die sechs Arten der Datenanalyse. https://martin-grellmann.de/die-sechs-arten-der-datenanalyse, 2022.

I

- [W. H. Inmon 2005] W. H. Inmon: Building the data warehouse. John wiley & sons, 2005.

K

- [R. Kimball 2011] R. Kimball and M. Ross: The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons, 2011.
- [J. Klump et al. 2021] J. Klump, L. Wyborn, M. Wu, J. Martin, R. R. Downs, A. Asmi, et al.: Versioning data is about more than revisions: A conceptual framework and proposed principles. 2021.
- [J. Kreps et al. 2011] J. Kreps, N. Narkhede, J. Rao, et al.: Kafka: A distributed messaging system for log processing. in Proceedings of the NetDB. vol. 11, pp. 1–7, Athens, Greece, 2011.
- [J. Kutay] J. Kutay: Change Data Capture (CDC): What it is and How it Works. https://www.striim.com/blog/change-data-capture-cdc-what-it-is-and-how-it-works/, [Online; Stand: 2.06.2023].

L

- [E. Levy 2021] E. Levy: Batch vs Stream vs Microbatch Processing: A Cheat Sheet. https://www.upsolver.com/blog/batch-stream-a-cheat-sheet, 2021.
- [D. Linstedt 2015] D. Linstedt and M. Olschimke: Building a scalable data warehouse with data vault 2.0. Morgan Kaufmann, 2015.
- [S. Luber 2018] S. Luber and N. Litzel: Was ist Data Profiling?. https://www.bigdata-insider.de/was-ist-data-profiling-a-691538/, 2018.

M

- [J. Majchrzak 2022] J. Majchrzak, S. Balnojan, M. Siwiak, M. Sieraczkiewicz: Data Mesh in Action. Manning Publication, 2022.
- [P. Mhatre 2021] P. Mhatre: Data Warehouse vs Data Vault vs Data Lake vs Delta Lake vs Data Fabric vs Data Mesh. https://medium.com/@mhatrep/data-warehouse-vs-data-vault-vs-data-lake-vs-delta-lake-vs-data-fabric-vs-data-mesh-1cf4c8991961, 2021.

P

- [P. Pääkkönen 2015] P. Pääkkönen and D. Pakkala: Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. Big Data Research, vol. 2, no. 4, pp. 166–186, 2015.
- [D. L. Parnas 2002] D. L. Parnas: The Secret History of Information Hiding. pp. 398–409. Berlin, Heidelberg, Springer Berlin Heidelberg, 2002.
- [T. B. Pedersen 2009] T. B. Pedersen: Multidimensional Modeling. pp. 1777–1784. Boston, MA: Springer US, 2009.

R

- [C. Richardson 2018] C. Richardson, Microservices patterns: with examples in Java. Simon and Schuster, 2018.
- [J. Reis 2022] J. Reis and M. Housley: Fundamentals of Data Engineering. O’Reilly Media, Inc., 2022.

S

- [Y. Sharvit 2022] Y. Sharvit: Data-oriented programming unlearning objects. Manning, 2022.
- [B. Singhal 2022] B. Singhal and A. Aggarwal: Etl, elt and reverse etl: A business case study. in Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering (ICATIECE), pp. 1–4, 2022.
- [A. Silberschatz 2011] A. Silberschatz, H. F. Korth, and S. Sudarshan: Database system concepts. 2011.

V

- [H. Varshney 2023] H. Varshney: What is a Data Staging Area? Staging Data Simplified 101. https://hevodata.com/learn/data-staging-area/, 2023.