Certificate of Advanced Studies

Data Engineering

Daten verschiedenster Form und Komplexität werden heute überall in grossen Mengen erzeugt. Ihre effiziente und zeitnahe Aufbereitung für operative, strategische und analytische Informationssysteme ist die Aufgabe des Data Engineerings. Dieses CAS vermittelt Ihnen eine umfassende Technologie- und Methoden-Kompetenz in der Nutzung moderner und leistungsfähiger Datenspeicher, Big Data-Lösungen mit Stream- und Event-Processing, Datenanalyse-Technologien und weiteren State-of-the-Art-Werkzeugen.

Berner
Fachhochschule

Inhaltsverzeichnis

Stand: 05.01.2024

1 Umfeld

Die Komplexität und der oft schnelle Lebenszyklus von Daten erfordern den Einsatz performanter und spezialisierter Werkzeuge und Methoden. Gegenüber der klassischen Business-Intelligence-Welt sind die Verschiedenartigkeit, Volatilität und notwendige Verarbeitungsgeschwindigkeit von Daten eine grosse Herausforderung. Im CAS Data Engineering erlernen Sie die methodischen Grundlagen, die Anforderungen an Software- und Hardware-Infrastruktur und den Einsatz ausgewählter Entwicklungswerkzeuge für ein erfolgreiches Data Engineering und die Umsetzung komplexer Big-Data-Projekte.

2 Zielpublikum

Das CAS Data Engineering richtet sich an Entwickler*innen, Fach- und technische Führungskräfte in Unternehmen und IT-Bereichen, die für den Aufbau, die Planung und/oder die Umsetzung von Data Engineering- und Big-Data-Projekten verantwortlich sind.

3 Ausbildungsziele

- Sie kennen die Methoden, Werkzeuge und Frameworks eines modernen Data-Engineering-Umfeldes.
- Sie kennen die aktuellen Speichertechnologien und Datenbanksysteme als Grundlage für den Umgang mit verschiedenen Datentypen und unterschiedlichen Anforderungen an deren Handhabung.
- Sie können Big-Data-Projekte in Ihrem Unternehmen planen, umsetzen und in die unternehmenseigene IT-Architektur einbetten.
- Sie kennen Architekturen und Werkzeuge zur Aufbereitung und Analyse von Echtzeit-Datenströmen.
- Sie können datenorientierte Anwendungen und Machine-Learning-Lösungen auf Apache Spark, einer modernen und hochgradig skalierbaren Plattform, entwerfen und implementieren.

4 Voraussetzungen

- Sie bringen Vorkenntnisse entsprechend einer Informatik- oder Wirtschaftsinformatik-Ausbildung mit, insbesondere mit Kenntnissen über Programmiersprachen, Datenbanksysteme und Abfragesprachen wie SQL.

- Programmier-Übungen finden mit Python statt, entsprechende Vorkenntnisse sind wünschenswert.

Für die Übungen wird ein Laptop mit Zugang zum Internet benötigt (für GitHub und Zugang zur AWS Cloud). Um die Umgebung mit den Werkzeugen nutzen zu können, ist eine virtuelle Umgebung in der AWS Cloud notwendig (AWS Lightsail Dienst). Die Kosten trägt hier der Teilnehmer, wobei sich diese auf ca. CHF 2.- / Kurstag belaufen (sofern die VM nach jedem Kurstag entsprechend abgebaut wird). Der richtige Umgang mit der Cloud-Umgebung werden wir am ersten Kurstag zeigen.

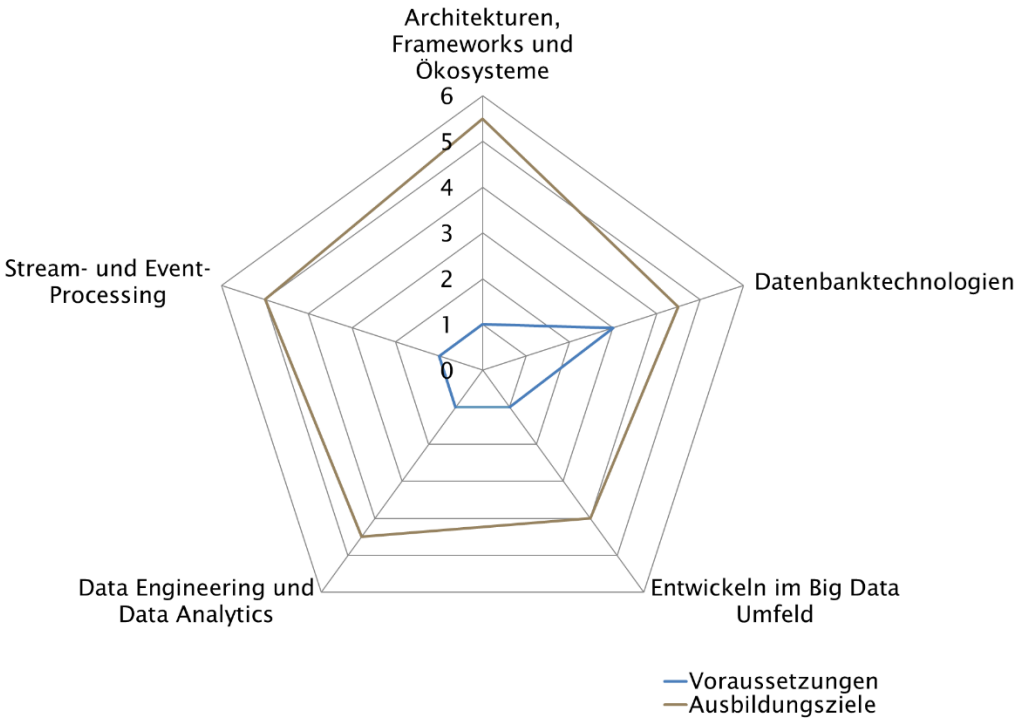
5 Unterrichtssprache

Die Unterrichtssprache ist Deutsch, die Unterlagen sind teilweise in Englisch.

6 Durchführungsort

Berner Fachhochschule, Weiterbildung, Aarbergstrasse 46 (Switzerland Innovation Park Biel/Bienne), 2503 Biel,
Telefon +41 31 848 31 11, E-Mail weiterbildung.ti@bfh.ch.

7 Kompetenzprofil



Kompetenzstufen

1. Kenntnisse/Wissen
2. Verstehen
3. Anwenden
4. Analyse
5. Synthese
6. Beurteilung

8 Kursübersicht

Kurs / Lehreinheit	Lektionen	Stunden	Dozierende
Datenbanktechnologien	24		Guido Schmutz
Grundlagen von Big Data und das Spark-Ökosystem	20		Guido Schmutz
Stream- and Event-Processing	24		Guido Schmutz
Hardware, Architektur, Cloud	12		Daniel Steiger
Data Engineering und Data Analytics für Spark	40		Jürgen Vogel
Projektarbeit	8	90	Verschiedene Betreuer*innen
Total	**128 / 16 Tage**	**90**	

Das CAS umfasst insgesamt 12 ECTS-Credits (~300 Std). Für die einzelnen Kurse ist Zeit für Selbststudium, Prüfungsvorbereitung etc. einzurechnen.

9 Kursbeschreibungen

Nachfolgend sind die einzelnen Kurse und Lehrveranstaltungen dieses Studienganges beschrieben.

Der Begriff Kurs schliesst alle Veranstaltungstypen ein, es ist ein zusammenfassender Begriff für verschiedene Veranstaltungstypen wie Vorlesung, Lehrveranstaltung, Fallstudie, Living Case, Fach, Studienreise, Semesterarbeiten usw.

9.1 Datenbanktechnologien

Lernziele	Die Teilnehmenden: – erwerben ein grundlegendes Verständnis von alternativen Datenbankkonzepten – verstehen die Konzepte hinter den neuen, modernen NoSQL- und NewSQL-Datenbanken – kennen die Unterschiede zu den relationalen Datenbanken – lernen die verschiedenen Arten von NoSQL kennen – können erfolgsversprechende Einsatzszenarien erkennen
Themen und Inhalte	– Was ist NoSQL? Was ist NewSQL? Warum gibt es diese neuen Datenbankarten? – Relevante Datenbankkonzepte wie BASE, ACID, CAP, Partitionierung, Sharding, Replikation usw. – Eigenschaften der NoSQL-Datenbanken – Klassifikation der NoSQL-Datenbanken – Anwendungsfälle für NoSQL-Datenbanken

	– Was geschieht mit den traditionellen, relationalen Datenbanken? – Schema-Less vs. Schema bzw. Schema-on-Write vs. Schema-on-Read – Ausgewählte, populäre NoSQL-Datenbanken: MongoDB, Redis, Cassandra, Elasticsearch, Neo4J und InfluxDB – NoSQL in einer modernen Datenplattform
Lehrmittel	– Folien/Skript – Literaturempfehlungen Nr. 1,2,3,4

9.2 Grundlagen von Big Data und das Spark-Ökosystem

Lernziele	Die Teilnehmenden: – lernen mögliche Architekturen von Big-Data-Lösungen kennen – kennen die Kernkomponenten von Hadoop und dem Hadoop-Ökosystem – lernen die Unterschiede zwischen Object Storage und HDFS kennen – lernen Apache Spark und das Spark-Ökosystem kennen – können mit Spark einfache Big-Data-Anwendungen entwickeln – sehen, warum SQL die «lingua franca» für die Datenverarbeitung ist – lernen, wie Datenflüsse in eine Big-Data-Plattform unterstützt werden können – sehen, wie man mit Query-Virtualisierung und SQL-Daten verarbeiten kann – lernen, wie die grossen Cloud-Anbieter Big Data unterstützen
Themen und Inhalte	– Anwendungsfälle für Big Data – Big-Data-Architektur – Historie von Big Data: Hadoop HDFS, MapReduce und das Hadoop Ecosystem – SQL on Big Data (Trino, Presto, …) – Datenserialisierung/Deserialisierung mit Avro, Parquet, usw. – Das Apache-Spark-Ökosystem mit Spark Core, Spark SQL, Dataframes und Datasets – Table Formats: Data Lake und Iceberg – Daten-Import und -Export in eine Big Data Platform – Automatisierung von Workflows – Big Data und Cloud – Data Lake vs. Data Lakehouse vs. Data Mesh

Lehrmittel	– Folien/Skript – Literaturempfehlungen Nr. 7, 8, 9, 10

9.3 Data Engineering und Data Analytics für Spark

Lernziele	Die Teilnehmenden: – lernen verschiedene Methoden und Verfahren zur Analyse von strukturierten (relationalen) Daten, unstrukturierten (textuellen) Daten (Textdaten) und soziale Netzwerke kennen und wenden diese exemplarisch an – erwerben ein grundlegendes Verständnis für die spezifischen Anforderungen in verschiedenen Anwendungsszenarien und den systematischen Lösungsentwurf – entwickeln mit Python anwendungsspezifische Lösungen für die Big-Data-Plattform Apache Spark – erhalten einen Einblick in die Entwicklung und Anwendung von Machine-Learning-basierten Lösungen
Themen und Inhalte	– Entwurf, Implementierung und Evaluation von Datenanalyse-Verfahren unter Verwendung von Apache Spark und der enthaltenen Kernbibliotheken Spark SQL, ML/MlLib und GraphX in Python – Einsatz von maschinellem Lernen auf Big Data am Beispiel der Analyse von relationalen Daten und Textdokumenten (Clustering, überwachtes Lernen/Klassifikation, Word Embeddings und Deep Learning) – Handhabung von Graphen-Daten am Beispiel der Analyse von sozialen Netzwerken (PageRank, Prestige)
Lehrmittel	– Folien/Skript – Eingebettete Übungsaufgaben mit Python, PySpark und Jupyter Notebooks – Literaturempfehlungen Nr. 5, 6, 9

9.4 Hardware, Architektur, Cloud

Lernziele	Die Teilnehmenden: – lernen Big-Data-spezifische Infrastrukturanforderungen zu formulieren – kennen die wesentlichen Architekturmerkmale eines Big-Data-Systems – kennen die aktuellen Big-Data-Plattformen und -Appliances der Marktleader – kennen die wichtigsten Integrationstechnologien – kennen die betrieblichen Aspekte einer Big-Data-Infrastruktur und sind in der Lage, ein Betriebskonzept zu erstellen – können die Unterschiede zwischen einer Cloud-basierten und einer On-Premise-Infrastruktur beschreiben und beurteilen – lernen die infrastrukturtechnischen Voraussetzungen für die Erfüllung von Sicherheitsanforderungen kennen – können den Reifegrad verschiedener Technologien beurteilen
Themen und Inhalte	– Anforderungen und Architekturtreiber im Big-Data-Umfeld

	– Reliability, Availability, Scaleability und Performance – Big-Data Infrastruktur-Blueprints – Big-Data-Plattformen – Integration von Big-Data-Systemen in die bestehende IT-Landschaft (Konnektoren) – Lifecycle einer Big-Data-Infrastruktur (Aufbau, Betrieb, Optimierung) – Überlegungen und Implikationen zu Big-Data in der Cloud – Secure Infrastructure – Leading-Edge-Technologien und Technologietrends
Lehrmittel	– Vorlesungsunterlagen (SlideDoc) – Aktuelle Whitepapers, Fachartikel und Hersteller-Unterlagen

9.5 Stream and Event Processing

Lernziele	Die Teilnehmenden: – lernen die Prinzipien des Stream und Event Processing kennen – können die Komponenten einer Event-Driven Architecture (EDA) beschreiben – lernen die unterschiedlichen Sprachen für die Erkennung und Verarbeitung von Events kennen – können Probleme mit Hilfe von Event-Processing lösen – können abschätzen, wann sich der Einsatz von Event Processing lohnt – kennen die Positionierung von Event Processing interhalb einer Big-Data-Architektur
Themen und Inhalte	– Was ist ein Event, was ist eine Message? – Was ist Complex Event Processing (CEP)? – Historie und Prinzipien von Stream und Complex Event Processing – Event Processing Design Patterns – Erkennen von Events – Aggregation von Events – wie können Business Events von den Raw Events abgeleitet werden – Internet of Things und Machine to Machine (M2M) – was hat dies mit Event-Processing zu tun? – Welche Sprachen für das Event-Processings gibt es? – Plattformen und Frameworks für Stream Processing: Apache Storm, Apache Flink, Kafka Streams, Spark Streaming usw.
Lehrmittel	– Folien/Skript – Literaturempfehlungen Nr. 4, 10, 13, 14

9.6 Projektarbeit

Zielsetzung und Thema	In der Semesterarbeit bearbeiten die Teilnehmenden ein Projekt oder eine Fragestellung aus ihrer Firma. Mit dem gewählten Thema vertiefen die Studierenden die im Studium erlernten Methoden. Das Thema der Semesterarbeit kann umfassen: – Machbarkeitsstudie – Lösungsentwicklung – Umsetzung oder Implementation von Analytics-Anforderungen – Evaluation und Projektierung – Algorithmen- oder Software-Entwicklung – IT-Architektur und Konzeption – Optimierung von Lösungen usw.
Ablauf	Die Semesterarbeit umfasst ca. 90 Stunden Arbeit und beinhaltet folgende Meilensteine (siehe auch Zeitplan): 1. In der Firma ein Thema und eine*n Betreuer*in suchen. 2. Kurzpräsentation des Themas vor einem Dozierendengremium. 3. Zuordnung von 1 Expert*in durch die Schule. 4. Durchführung der Arbeit in eigener Terminplanung. 5. 2–3 Meetings mit den Expert*innen. 6. Schlusspräsentation vor Klasse, Expert*innen und Dozierenden. 7. Abgabe des Berichtes an die Expert*innen und die Schule. 8. Bewertung durch die Expert*innen.

10 Kompetenznachweis

Für die Anrechnung der 12 ECTS-Credits ist das erfolgreiche Bestehen der Qualifikationsnachweise (Prüfungen, Projektarbeiten) erforderlich, gemäss folgender Aufstellung:

Kompetenznachweis	Gewicht	Art der Qualifikation	Erfolgsquote Studierende
Ein Prüfungsblock Grundlagen, Architektur, Hadoop Datenbanktechnologien, Stream- & Event-Processing (50 Pkte.) Hardware, Infrastruktur, Cloud (20 Pkte.) Data Engineering (30 Pkte.)	5	Schriftliche Prüfung 120 Minuten, Open Book, elektronisch (Moodle)	0 – 100 %
Semesterarbeit	5	Projektarbeit	0 – 100 %
Gesamtgewicht / Erfolgsquote	**10**		0 – 100 %

Die gewichtete Summe aus den Erfolgsquoten der Kompetenznachweise wird in eine Note zwischen 3 und 6 umgerechnet. Die Note 3 (gemittelte Erfolgsquote weniger als 50%) ist ungenügend. Die Noten 4 bis 6 (gemittelte Erfolgsquote zwischen 50% und 100%) sind genügend.

11 Lehrmittel

Die nachfolgend aufgeführten Lehrmittel sind wesentlich für das Lernen während des geführten Unterrichtes. Sie sind durch die Studierenden zu beschaffen.

Nr.	Titel	Autor*innen	Verlag	Jahr	ISBN-Nr.
1.	Designing Data-Intensive Applications	Martin Kleppmann	O'Reilly Media	2017	ISBN-10 1-4493-7332-1
2.	Cassandra: The Definitive Guide, 3rd Edition	Jeff Carpenter, Eben Hewitt	O'Reilly Media	2020	ISBN-13: 978-1-09-811516-6
3.	MongoDB: The Definitive Guide, 3rd Edition	Shannon Bradshaw, Eoin Brazil, Kristina Chodorow	O'Reilly Media	2019	ISBN-13: 978-1-4919-5446-1
4.	Graph Algorithms: Practical Examples in Apache Spark and Neo4j	Mark Needham, Amy Hodler	O'Reilly Media	2019	ISBN-13: 978-1492047681
5.	Natural Language Processing with Python	Steven Bird u.w.	O'Reilly Media	2009	ISBN-13: 978-0596516499
6.	Mining the Social Web	Matthew Russell	O'Reilly Media	2013	ISBN-10: 1449367615
7.	Learning Spark	Jules S. Damji, Brooke Wenig, Tathagata Das, Denny Lee	O'Reilly Media	2020	ISBN-10: 978-1-492-05004-9
8	Data Analysis with Python and PySpark	Jonathan Rioux	Manning Press	2022	ISBN-13: 978-1617297205
9.	Spark in Action, 2nd Edition	Jean-Georges Perrin	Manning Press	2020	ISBN-13: 978-1617295522
10. d	Big Data: Principles and best practices of scalable realtime data systems	Nathan Marz and James Warren	Manning Press	2015	ISBN-10: 9781617290343
11.	Kafka: The Definitive Guide, Second Edition	Gwen Shapira, Todd Palino, Rajini Sivaram and Krit Petty	O'Reilly Media	2021	ISBN-13: 9781492043089
12.	Kafka in Action	Viktor Gamov, Dylan Scott, Dave Klein	Manning Press	2022	ISBN-13: 9781617295232
13.	Mastering Kafka Streams and ksqlDB	Mitch Seymour	O'Reilly Press	2021	ISBN-13: 978-1-4920-6249-3
14.	Designing Cloud Data Platforms	Danil Zburivsky, Lynda Partner	Manning Press	2021	ISBN-13: 978-1617296444
15. S	Streaming Databases	Hubert Dulay, Ralph Matthias Debusmann	O'Reilly Media	2024	ISBN-13: 978-1-098-15477-6

12 Dozierende

Vorname Name	Firma	E-Mail
Guido Schmutz	Accenture	guido.schmutz@bfh.ch
Jürgen Vogel	Berner Fachhochschule	juergen.vogel@bfh.ch
Daniel Steiger	Accenture	daniel.steiger@trivadis.com

13 Organisation

CAS-Leitung:
Prof. Dr. Arno Schmidhauser
Tel: +41 31 848 32 75
E-Mail: arno.schmidhauser@bfh.ch

CAS-Administration:
Andrea Moser
Tel: +41 31 848 32 11
E-Mail: andrea.moser@bfh.ch

Während der Durchführung des CAS können sich Anpassungen bezüglich Inhalte, Lernzielen, Dozierenden und Kompetenznachweisen ergeben. Es liegt in der Kompetenz der Dozierenden und der Studienleitung, aufgrund der aktuellen Entwicklungen in einem Fachgebiet, aufgrund der konkreten Vorkenntnisse und Interessenslage der Teilnehmenden sowie aus didaktischen und organisatorischen Gründen Anpassungen im Ablauf eines CAS vorzunehmen.

Berner Fachhochschule
Technik und Informatik
Weiterbildung
Aarbergstrasse 46 (Switzerland Innovation Park Biel/Bienne)
2503 Biel

Telefon +41 31 848 31 11
E-Mail: weiterbildung.ti@bfh.ch

bfh.ch/ti/weiterbildung
bfh.ch/cas-data