Certificate of Advanced Studies

Big Data

Daten verschiedenster Form und Komplexität werden heute fast überall in grossen Mengen erzeugt. Ihre effiziente und zeitnahe Aufbereitung für operative, strategische und analytische Informationssysteme ist die Aufgabe von Big-Data-Systemen. Das CAS Big Data vermittelt Ihnen eine umfassende Technologie-Kompetenz, sowie gängige Methoden und Werkzeuge zur Realisierung von Big Data-Lösungen.

bfh.ch/cas-bgd

Inhaltsverzeichnis

Stand: 01.03.2022

1 Umfeld

Big Data ermöglicht die Nutzung grosser Datenmengen. Die Komplexität der Analysen und der oft schnelle Lebenszyklus der Daten erfordern den Einsatz performanter und spezialisierter Informatikmittel. Gegenüber der klassischen Business Intelligence-Welt stellen die Verschiedenartigkeit und die Volatilität der Datenquellen eine zusätzliche Herausforderung von Big Data-Systemen dar. Im CAS Big Data lernen Sie die methodischen Grundlagen, die Anforderungen an Software- und Hardware-Infrastruktur und ausgewählte Entwicklungswerkzeuge für die Realisierung erfolgreicher und komplexer Big-Data-Projekte kennen.

2 Zielpublikum

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

3 Ausbildungsziele

- Sie kennen die methodischen und technologischen Grundlagen von Big Data.
- Sie können Big Data-Projekte in Ihrem Unternehmen planen, umsetzen und betreiben.
- Sie können die Einbettung von Big Data-Lösungen in die unternehmenseigene IT-Architektur beurteilen und konzipieren.
- Sie kennen moderne Methoden und Werkzeuge zur Aufbereitung, Analyse und Darstellung von Echtzeit-Datenströmen.

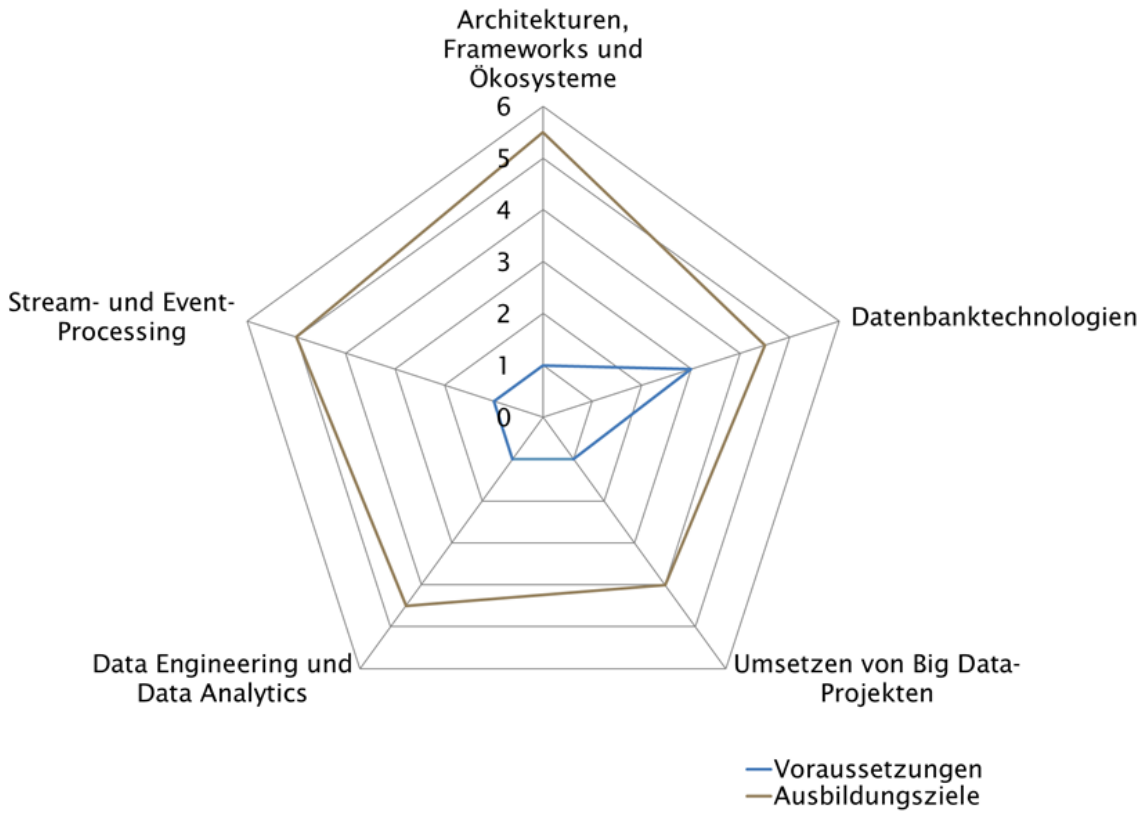
4 Voraussetzungen

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

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

- Für die Übungen wird ein kräftiger Laptop benötigt, typischerweise mehr als 16GB RAM und I7 CPU. Alternativ kann mit einer Azure oder AWS Cloud gearbeitet werden (gebührenpflichtig).

5 Kompetenzprofil



Kompetenzstufen

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

6 Kursübersicht

Kurs / Lehreinheit	Lektionen	Stunden	Dozierende
Datenbanktechnologien	16		Guido Schmutz
Grundlagen, Spark Ökosystem	24		Guido Schmutz
Hardware, Architektur, Cloud	12		Daniel Steiger
Stream- and Event-Processing	16		Guido Schmutz
Data Analytics (optional)	16		Werner Dähler
Data Engineering	32		Jürgen Vogel
Nutzenaspekte von Big Data Projekten	4		Heinz Steiner
Projektarbeit	12	90	Verschiedene Betreuer*innen
Total	**132**	**90**	

Das CAS umfasst insgesamt 12 ECTS-Credits. Für die einzelnen Kurse ist entsprechend Zeit für Selbststudium, Prüfungsvorbereitung etc. einzurechnen.

7 Kursbeschreibungen

Nachfolgend sind die einzelnen Kurse und Lehrveranstaltungen dieses Studienganges beschrieben.

7.1 Datenbanktechnologien

Lernziele	Die Teilnehmenden: – erwerben ein grundlegendes Verständnis von alternativen Datenbank-konzepten – 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 und NewSQL Datenbanken – Klassifikation der NoSQL und NewSQL Datenbanken – Anwendungsfälle für NoSQL und NewSQL 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, HBase, Redis, Cassandra, Neo4J, u.a. – NoSQL und NewSQL in einer Big Data Architektur
Lehrmittel	– Folien/Skript – Literaturempfehlungen Nr. 3, 11

7.2 Grundlagen, Spark Ökosystem

Lernziele	Die Teilnehmenden: – lernen mögliche Architekturen von Big Data-Lösungen kennen – kennen die Kernkomponenten von Hadoop und dem Hadoop Ökosystem – wissen wie HDFS und MapReduce funktioniert – können mit MapReduce einfache Big Data-Anwendungen entwickeln – lernen wie Datenflüsse in eine Big Data-Plattform unterstützt werden können – lernen für welche Art von Problemen sich MapReduce bzw. Hadoop eignet – lernen Apache Spark und das Spark Ökosystem kennen
Themen und Inhalte	– Warum und wozu Hadoop verwendet wird – Eigenschaften der Hadoop Architektur – Kernkonzepte von Hadoop – Funktionalität von MapReduce – Die wichtigsten Komponenten des Hadoop Ökosystem: HDFS, MapReduce, Pig, Hive, Zookeeper, Flume, usw. – Entwicklung von Hadoop Applikationen – SQL on Hadoop (Hive, Impala) – Daten-Import und -Export in eine Big Data Platform – Automatisierung von Workflows – Datenserialisierung/Deserialisierung mit Avro, Parquet, usw. – Kernkonzepte von Apache Spark – Das Apache Spark Ökosystem mit Spark Core, Spark SQL, Dataframes und Datasets – Architektur, Patterns und Best Practices
Lehrmittel	– Folien/Skript – Literaturempfehlungen Nr. 2, 7, 8, 9, 12, 14

7.3 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-Processing 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

7.4 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-premis 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

7.5 Data Analytics

Lernziele	Die Teilnehmenden: – kennen die analyserelevanten Phasen von CRISP-DM – erwerben ein grundlegendes Verständnis für die Anforderungen an statistisch zu analysierende strukturierte Daten – kennen die wichtigsten Methoden für Explorative Datenanalyse, Datenvorbereitung und Machine Learning – können diese Methoden mit Funktionen aus unterschiedlichen Python Bibliotheken selber anwenden
Themen und Inhalte	– CRISP-DM, die einzelnen Phasen – Ein erster Blick auf die Daten – Univariate und bivariate Methoden zum Erkennen von Datenanomalien – Multivariate Methoden (Unsupervised Learning) – Erstellen, validieren und Optimieren von Vorhersagemodellen (Supervised Learning) – Gegenüberstellung unterschiedlicher Python Bibliotheken, insb. pandas und sciki-learn vs. pyspark
Lehrmittel	– Folien/Skript – Literaturempfehlungen werden bei Kursbeginn bekannt gegeben

7.6 Data Engineering

Lernziele	Die Teilnehmenden: – entwerfen, implementieren und evaluieren verschiedene Daten-Analyseverfahren mit Hilfe der Big Data-Plattform Apache Spark – lernen verschiedene Verfahren zur Analyse von strukturierten Daten, unstrukturierter Daten (Textdaten), und Graphendaten (soziale Netzwerke) kennen
Themen und Inhalte	– Entwurf, Implementierung und Evaluation von Daten-Analyseverfahren 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 Textdokumenten (Clustering, Supervised Classification) und Recommender Systems (Collaborative Filtering) – Handhabung von Graphen-Daten am Beispiel der Analyse von sozialen Netzwerken (PageRank, Prestige) – Publizieren und konsumieren von frei zugänglichen Datensätzen im Web (Open Data)
Lehrmittel	– Folien/Skript – Eingebettete Übungsaufgaben mit Python, PySpark und Jupyter Notebooks – Literaturempfehlungen Nr. 5, 6, 15

7.7 Nutzenaspekte von Big Data Projekten

Lernziele	Die Teilnehmenden: – kennen den wirtschaftlichen Nutzen und die nicht technischen Herausforderungen von Big Data-Projekten – gewinnen eine Übersicht, wann der Einsatz von Big Data sinnvoll ist

Themen und Inhalte	– Was macht ein typisches Big Data-Projekt aus?
	– Nutzen und Abgrenzung von Big Data-Projekten beurteilen
	– Welche Fallen gilt es zu umschiffen?
	– Bei Big Data-Projekten stellt die Heterogenität der Teammitglieder eine besondere Herausforderung dar:
	  – Welche Vorteile bieten Pilot-Projekte beim Thema Big Data?
	  – Was ist bei der Kommunikation der Projekte zu beachten?
	– Big Data und Digitalisierung: Wie gehört das zusammen?

Lehrmittel	– Folien/Skript – Literaturempfehlung Nr. 1

7.8 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 suchen, und mit Vorteil eine*n Ansprechpartner*in / Betreuer*in in der Firma definieren. 2. Erstellen einer Projektskizze (Wordvorlage, 1 bis 2 Seiten) und einer Kurzpräsentation (Powerpoint, wenige Slides): a. Titel b. Umfeld c. Problemstellung d. Lösungsansatz (Vorgehen, Methoden) e. Name und Kontaktadressen der Gruppenmitglieder, und des Ansprechpartners / Betreuers in der Firma 3. Kurzpräsentation des Themas vor Dozierendengremium, 5-10' Präsentation, 5-10' Diskussion, max. 15'. 4. Eventuell Überarbeitung der Projektskizze gemäss Feedback. 5. Zuordnung von Expert*innen durch die Schule. 6. Durchführung der Arbeit in eigener Terminplanung. 7. 2-3 Meetings mit den Expert*innen (durch Studierende organisiert) 8. Schlusspräsentation vor Klasse, Expert*innen und Dozierenden. 15' Präsentation, 15' Diskussion. 9. Abgabe des Berichtes an die Expert*innen (per E-Mail, auf Wunsch in Papierform).

Ergebnis und Bewertung	Der Bericht ist in elektronischer Form als PDF-Dokument an die*den Betreuer*in zu schicken und auf der Moodle-Plattform des CAS zu hinterlegen. Bericht: ca. 20-30 Seiten, Source Code soweit notwendig für die Projektbeurteilung. Die Semesterarbeit wird nach folgenden Kriterien bewertet: – Themeneingabe Projektskizze rechtzeitig und vollständig eingereicht. Themenpräsentation sorgfältig vorbereitet. Idee oder Aufgabe durchdacht und abgegrenzt, Quellen recherchiert, Rahmenbedingungen definiert, Teilziele priorisiert. – Methodik und Ausführung Gewählte Methode(n) systematisch und korrekt angewendet. Kreativ und agil in der Ausführung. Entscheidungen präzis begründet. – Ergebnis Nachvollziehbares und dokumentiertes Ergebnis. Aufgabenstellung erfüllt. Ergebnisse validiert, getestet, verifiziert. Vergleich von Zielsetzung und Ergebnis vorgenommen. Learnings und Ausblick vorhanden. – Bericht und Dokumentation Vollständig und verständlich. Rechtschreibung korrekt. Kapiteleinteilung sinnvoll. Angemessene Darstellung. Grafiken auf das Wesentliche reduziert und beschriftet. – Schlusspräsentation Roter Faden, logisches Vorgehen, klare Aussagen. Identifikation mit dem Thema spür- und erkennbar. Professionelle Präsentationstechnik, Zeitvorgaben genutzt und eingehalten. Fragen präzis und sicher beantwortet.
Vertraulichkeit	Die Projektarbeiten werden als nicht-öffentlich behandelt. An den Präsentationen können jedoch auch interessierte Personen im Umfeld der Schule teilnehmen. Auf Wunsch steht ein kostenloses Standard-NDA der Schule zur Verfügung. Individuelle Vereinbarungen sind kostenpflichtig.

8 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 %

Jede*r Student*in kann in einem Kompetenznachweis eine Erfolgsquote von 0 bis 100% erreichen. Die gewichtete Summe aus den Erfolgsquoten pro Thema und dem Gewicht des Themas ergibt eine Gesamterfolgsquote zwischen 0 und 100%. Der gewichtete Mittelwert der Erfolgsquoten der einzelnen Kompetenznachweise wird in eine Note zwischen 3 und 6 umgerechnet. Die Note 3 (gemittelte Erfolgsquote weniger als 50%) ist ungenügend. Die Noten 4, 4.5, 5, 5.5 und 6 (gemittelte Erfolgsquote zwischen 50% und 100%) sind genügend.

9 Ergänzende Lehrmittel

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

Nr	Titel	Autoren	Verlag	Jahr	ISBN Nr.
1.	Big Data für Entscheider: Entwicklung und Umsetzung datengetriebener Geschäftsmodelle	Andreas Gadatsch, Holm Landrock	Springer-Verlag	2017	ISSN-2197-6716 (electronic)
2.	Hadoop: The Definitive Guide, 4th Edition	Tom White	O' Reilly Press	2015	ISBN-10: 1-4493-1152-0
3.	Next Generation Databases: NoSQL, NewSQL and BigData	Guy Harrison	Apress	2016	ISBN-10: 1-48-421330-0
4.	Event Processing in Action	Peter Niblett, Opher Etzion	Manning Press	2010	ISBN-10: 1-935182-21-8
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	Holden Karau u.w.	O'Reilly Media	2015	ISBN-10: 1-4493-5862-4
8.	Hadoop Application Architecture	Mark Grover u.w.	O'Reilly Media	2015	ISBN-10: 1-4919-0008-3
9.	Big Data: Principles and best practices of scalable realtime data systems	Nathan Marz and James Warren	Manning Press	2015	ISBN-10: 9781617290343
10.	Storm Applied	Sean T. Allen u.w.	Manning Press	2015	ISBN-10: 9781617291890
11.	NoSQL for Mere Mortals	Dan Sullivan	Addison Wesley	2015	ISBN-10: 0-13-402321-8
12.	Hadoop Operations, 2nd Edition	Eric Sammer	O'Reilly Press	2018	ISBN-10: 1-49-192383-0
13.	Streaming Data: Understanding the Real-Time Pipeline	Andrew Psaltis	Manning Press	2017	ISBN-10: 1-61-729228-1
14.	Kafka – The Definitive Guide	Neha Narkhede u.w.	O'Reilly Press	2016	ISBN-10: 1-4919-3616-9
15.	Spark in Action	P. Zecevic and M. Bonaci	Manning Press	2017	ISBN 9781617292606

10 Dozierende

Vorname Name	Firma	E-Mail
Guido Schmutz	Trivadis	guido.schmutz@bfh.ch
Daniel Steiger	Trivadis	daniel.steiger@trivadis.com
Werner Dähler	Berner Fachhochschule	werner.daehler@bfh.ch
Jürgen Vogel	Berner Fachhochschule	juergen.vogel@bfh.ch
Heinz Steiner	Trivadis	heinz.steiner@trivadis.com

11 Organisation

CAS-Leitung:
Prof. Dr. Arno Schmidhauser
Tel: +41 31 84 83 275
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 Inhalten, Lernzielen, Dozierenden und Kompetenznachweisen ergeben. Es liegt in der Kompetenz der Dozierenden und der Studienleitung, aufgrund der aktuellen Entwicklungen in einem Fachgebiet, 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

Telefon +41 31 848 31 11
Email: weiterbildung.ti@bfh.ch

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