

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение высшего образования
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»



РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
ПРИЕМ 2019 г.
ФОРМА ОБУЧЕНИЯ очная

Аналитика больших данных

Направление подготовки/ специальность	09.04.04 Программная инженерия
Образовательная программа (направленность (профиль))	Big Data Solutions / Технологии больших данных (язык обучения: английский)
Специализация	Big Data Solutions / Технологии больших данных
Уровень образования	высшее образование – магистратура

Курс	2	семестр	3
Трудоемкость в кредитах (зачетных единицах)	6		
Виды учебной деятельности	Временной ресурс		
Контактная (аудиторная) работа, ч	Лекции	32	
	Практические занятия	16	
	Лабораторные занятия	16	
	ВСЕГО	64	
Самостоятельная работа, ч	152		
ИТОГО, ч	216		

Вид промежуточной аттестации	Экзамен	Обеспечивающее подразделение	ОИТ
------------------------------	----------------	------------------------------	-----

Заведующий кафедрой- руководитель ОИТ на правах кафедры		Шерстнев В.С.
Руководитель ООП		Савельев А.О.
Преподаватель		Соколова В.В.

1. Цели освоения дисциплины

Целями освоения дисциплины является формирование у обучающихся определенного ООП (п. 5.4 Общей характеристики ООП) состава компетенций для подготовки к профессиональной деятельности.

Код компетенции	Наименование компетенции	Индикаторы достижения компетенций		Составляющие результатов освоения (дескрипторы компетенций)	
		Код индикатора	Наименование индикатора достижения	Код	Наименование
ПК(У)-2	Способен разрабатывать системы управления базами данных	И.ПК (У)-2.1	Разрабатывает компоненты системы управления базами данных	ПК(У)-2.1В1	Владеет навыками разработки структуры управления базами данных в целом и ее отдельных компонентов
				ПК(У)-2.1У1	Умеет идентифицировать класс разрабатываемой СУБД в зависимости от выполняемых ею задач и аппаратных средств, определенных в техническом задании на разработку систему СУБД
				ПК(У)-2.131	Знает теорию БД
				ПК(У)-2.132	Знает основные структуры и модели данных
				ПК(У)-2.133	Знает методы обработки данных
УК(У)-2	Способен управлять проектом на всех этапах его жизненного цикла	И.УК(У)-2.2	Осуществляет мониторинг хода реализации проекта, корректирует отклонения, вносит дополнительные изменения в план реализации проекта	УК(У)-2.2В1	Владеет способностью формирования плана разработки, управление рисками
				УК(У)-2.2У1	Умеет анализировать, оптимизировать и документировать проектные процессы, а также поддерживать их на всем протяжении ЖЦ разработки

2. Место дисциплины (модуля) в структуре ООП

Дисциплина относится к вариативной части междисциплинарного профессионального модуля учебного плана образовательной программы.

3. Планируемые результаты обучения по дисциплине

После успешного освоения дисциплины будут сформированы результаты обучения:

Планируемые результаты обучения по дисциплине		Индикатор достижения компетенции
Код	Наименование	
РД 1	Знание файловой системы HDFS, используемой для хранения больших наборов данных в хранилище Apache Hive.	И.ПК (У)-2.1
РД 2	Знание программной модели MapReduce, используемой для обработки больших наборов данных в хранилище Apache Hive.	И.ПК (У)-2.1
РД 3	Знание основ языка HiveQL (основные операторы и команды) для построения запросов на обработку данных, хранящихся в таблицах Apache Hive.	И.УК(У)- 2.2
РД 4	Умение применять встроенные функции языка HiveQL для осуществления анализа данных, хранимых в таблицах Apache Hive.	И.УК(У)- 2.2

Оценочные мероприятия текущего контроля и промежуточной аттестации представлены в календарном рейтинг-плане дисциплины.

4. Структура и содержание дисциплины

Основные виды учебной деятельности

Разделы дисциплины	Формируемый результат обучения по дисциплине	Виды учебной деятельности	Объем времени, ч.
--------------------	--	---------------------------	-------------------

Раздел 1. Хранение и обработка больших наборов данных в Apache Hive.	РД 1-4	Лекции	16
		Практические занятия	8
		Лабораторные занятия	8
		Самостоятельная работа	76
Раздел 2. Анализ больших наборов данных в Apache Hive.	РД 1-4	Лекции	16
		Практические занятия	8
		Лабораторные занятия	8
		Самостоятельная работа	76

Содержание разделов дисциплины:

Раздел 1. Хранение и обработка больших наборов данных в Apache Hive

Раздел формирует знания о методах анализа больших данных, навыки хранения наборов данных в файловой системе HDFS и обработки запросов, созданных с использованием языка HiveQL, на основе программной модели MapReduce.

Темы лекций:

1. Введение в анализ больших данных. Методы описательной статистики (4 ч.)
2. Свойства подхода к анализу данных – data mining. Этапы анализа больших данных согласно стандарту CRISP-DM (CRoss Industry Standard Process for Data Mining) (4 ч.)
3. Особенности архитектуры кластера Apache Hadoop, распределенная файловая система Hadoop (HDFS) (4 ч.).
4. Хранилище данных Apache Hive. Применение модели распределённых вычислений MapReduce при обработке данных в Apache Hive (4 ч.).

Темы практических занятий:

1. Установка и настройка программного обеспечения («виртуальный» кластер Hadoop, хранилище Apache Hive, Hue и др.) (2 ч.).
2. Проектирование и генерация датасета, загрузка в Apache Hive (2 ч.).
3. Проектирование и создание разделов (partitions) в таблице Apache Hive (2 ч.).
4. Проектирование и создание кластеров (buckets) в таблице Apache Hive (2 ч.).

Названия лабораторных работ:

1. Создание и заполнение «простой» таблицы в Apache Hive (2 ч.).
2. Использование строковых функций в Apache Hive (2 ч.).
3. Использование числовых функций в Apache Hive (2 ч.).
4. Использование функций для работы с датами в Apache Hive (2 ч.).

Раздел 2. Анализ больших наборов данных в Apache Hive

Раздел формирует навыки построения запросов к таблицам на языке HiveQL и анализа данных (корреляционный, дисперсионный и регрессионный и пр.) с использованием встроенных функций в Apache Hive.

Темы лекций:

1. Метод корреляционного анализа данных. Оценка параметров модели (4 ч.).
2. Метод дисперсионного анализа данных. Оценка параметров модели (4 ч.).
3. Метод линейного регрессионного анализа данных. Оценка параметров модели (4 ч.).
4. Метод ковариационного анализа данных. Оценка параметров модели (4 ч.).

Темы практических занятий:

1. Проектирование «сложной» таблицы и применение функций для работы с массивами, картами и структурами (2 ч.).
2. Использование функций преобразования типов в запросах на языке HiveQL (2 ч.).
3. Проектирование датасета, создание «управляемой» таблицы и применение аналитических функций для анализа данных (2 ч.).
4. Проектирование датасета, создание «управляемой» таблицы и применение функций для регрессионного анализа (2 ч.).

Названия лабораторных работ:

1. Использование функций преобразования типов в Apache Hive (2 ч.).
2. Использование функций задания условий в Apache Hive (2 ч.).
3. Использование групповых функций в Apache Hive (2 ч.).
4. Использование оконных функций в Apache Hive (2 ч.).

5. Организация самостоятельной работы студентов

Самостоятельная работа студентов при изучении дисциплины (модуля) предусмотрена в следующих видах и формах:

- Работа с лекционным материалом, поиск и обзор литературы и электронных источников информации по индивидуально заданной проблеме курса.
- Изучение тем, вынесенных на самостоятельную проработку.
- Поиск, анализ, структурирование и презентация информации.
- Перевод текстов с иностранных языков.
- Подготовка к лабораторным работам, к практическим и семинарским занятиям.
- Исследовательская работа и участие в научных студенческих конференциях, семинарах и олимпиадах.
- Подготовка к оценивающим мероприятиям.

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Учебно-методическое обеспечение

1. Макшанов, А. В. Технологии интеллектуального анализа данных: учебное пособие / А. В. Макшанов, А. Е. Журавлев. – 2-е изд., стер. – Санкт-Петербург: Лань, 2019. – 212 с. – ISBN 978-5-8114-4493-9. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/120063> (дата обращения: 11.01.2019). – Режим доступа: для авториз. пользователей.
2. Радченко, И. А. Технологии и инфраструктура Big Data: учебное пособие / И. А. Радченко, И. Н. Николаев. – Санкт-Петербург: НИУ ИТМО, 2018. – 52 с. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/136430> (дата обращения: 11.01.2019). – Режим доступа: для авториз. пользователей.
3. Чубукова, И. А. Data Mining: учебное пособие / И. А. Чубукова. – 2-е изд. – Москва: ИНТУИТ, 2016. – 470 с. – ISBN 978-5-94774-819-2. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/100582> (дата обращения: 11.01.2019). – Режим доступа: для авториз. пользователей.

Дополнительная литература

1. Григорьев, Ю. А. Реляционные базы данных и системы NoSQL: учебное пособие / Ю. А. Григорьев, А. Д. Плутенко, О. Ю. Плужникова. – Благовещенск: АмГУ, 2018. – 424 с. – ISBN 978-5-93493-308-2. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/156492> (дата обращения: 11.01.2019). – Режим доступа: для авториз. пользователей.
2. Стасышин, В. М. Практикум по языку SQL: учебное пособие / В. М. Стасышин, Л. Т. Стасышина. – Новосибирск: НГТУ, 2016. – 60 с. – ISBN 978-5-7782-2937-2. – Текст: электронный // Лань: электронно-библиотечная система. – URL: <https://e.lanbook.com/book/118207> (дата обращения: 11.01.2019). – Режим доступа: для авториз. пользователей.

6.2. Информационное и программное обеспечение

Internet-ресурсы (в т.ч. в среде LMS MOODLE и др. образовательные и библиотечные ресурсы):

1. Apache Hive [Электронный ресурс]: Official development web site / Дата обращения: 20.12.2018. – Режим доступа: <http://hive.apache.org/>.
2. Apache [Электронный ресурс]: Apache Hive Wiki / Дата обращения: 20.12.2018. – Режим доступа: <https://cwiki.apache.org/confluence/display/Hive/Home>.

Лицензионное программное обеспечение (в соответствии с **Перечнем лицензионного программного обеспечения ТПУ**): Zoom; 7-Zip; Adobe Acrobat Reader DC; Cisco Webex Meetings; Document Foundation LibreOffice; Google Chrome; Notepad++; Oracle VirtualBox.

7. Особые требования к материально-техническому обеспечению дисциплины

В учебном процессе используется следующее лабораторное оборудование для практических и лабораторных занятий:

№	Наименование специальных помещений	Наименование оборудования
1.	Аудитория для проведения учебных занятий всех типов, курсового проектирования, консультаций, текущего контроля и промежуточной аттестации (учебная лаборатория) 634034 Томская область, Томск, Советская улица, д. 84/3, учебный корпус «Кибернетический центр», аудитория 204.	Комплект оборудования для проведения лабораторных работ по основным разделам Информационных технологий: – ПК, 10 шт.

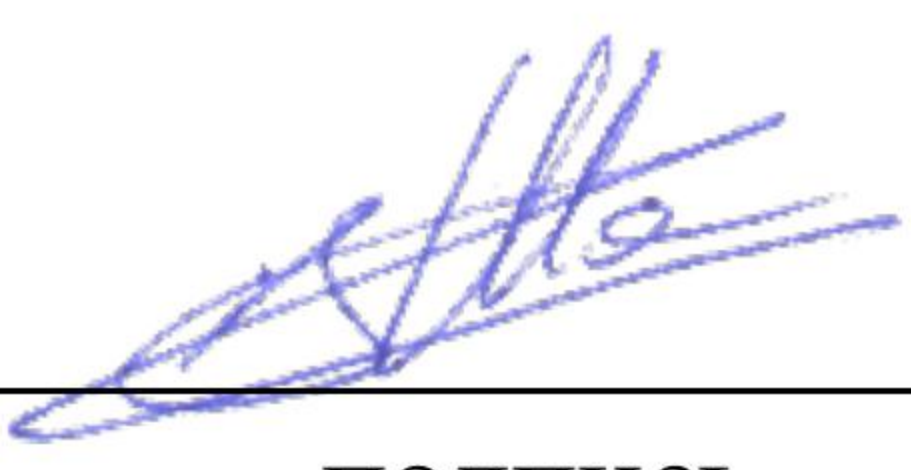
Рабочая программа составлена на основе Общей характеристики образовательной программы по направлению 09.04.04 Программная инженерия/ профиль Big Data Solutions / Технологии больших данных /специализации Big Data Solutions/Технологии больших данных (приема 2019 г., очная форма обучения).

Разработчик(и):

Должность	Подпись	ФИО
Доцент ОИТ ИШИТР		Соколова В.В.

Программа одобрена на заседании Отделения информационных технологий ИШИТР (протокол №13 от 19.06.2019 г.).

Заведующий кафедрой –
руководитель ОИТ

 / Шерстнев В.С.
подпись

Лист изменений рабочей программы дисциплины:

Учебный год	Содержание /изменение	Обсуждено на заседании ИШИТР (протокол)
2020/2021	1. Обновлено программное обеспечение 2. Обновлен список литературы, в том числе ссылок ЭБС	от 01.09.2020г. № 19