Минобрнауки России
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Факультет Электроники и вычислительной техники

Кафедра Электронно-вычислительные машины и системы

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к курсовой работе (проекту)

по дисциплине Системы обработки больших данных

на тему: Исследование датасета атмосферных осадков в Гонконге с использованием фреймворка Apache Spark

Студент Иванов Иван Иванович
(фамилия, имя, отчество)

Группа ЭВМ-1.1

Руководитель работы (проекта) _____ П.Д. Кравченя
(подпись и дата подписания) (инициалы и фамилия)

Члены комиссии:

_____ _____
(подпись и дата подписания) (инициалы и фамилия)

_____ _____
(подпись и дата подписания) (инициалы и фамилия)

_____ _____
(подпись и дата подписания) (инициалы и фамилия)

Нормоконтроллер _____ П.Д. Кравченя
(подпись и дата подписания) (инициалы и фамилия)

Волгоград 2025

Минобрнауки России
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Волгоградский государственный технический университет»

Факультет Электроники и вычислительной техники

Направление (специальность) Информатика и вычислительная техника

Кафедра Электронно-вычислительные машины и системы

Дисциплина Системы обработки больших данных

Утверждаю
Зав. кафедрой А.Е. Андреев

«____» _____ 20 ____ г.

ЗАДАНИЕ
на курсовую работу (проект)

Студент Иванов Иван Иванович
(фамилия, имя, отчество)

Группа ЭВМ-1.1

1. Тема: Исследование датасета атмосферных осадков в Гонконге с использованием фреймворка Apache Spark

Утверждена приказом от «___» _____ 20 ___ г., № _____.

2. Срок представления работы (проекта) к защите «___» _____ 20 ___ г.

3. Содержание расчётно-пояснительной записки: РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ PYSPARK; МАШИННОЕ ОБУЧЕНИЕ НА БОЛЬШИХ ДАННЫХ.

4. Перечень графического материала: _____

5. Дата выдачи задания «___» _____ 20 ___ г.

Руководитель работы (проекта) _____ П.Д. Кравченя
(подпись и дата подписания) (инициалы и фамилия)

Задание принял к исполнению _____ И.И. Иванов
(подпись и дата подписания) (инициалы и фамилия)

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Во введении сначала дается краткая характеристика области, в которой выполнена работа (1 – 3 предложения). Затем обосновывается актуальность работы.

Далее идут фразы, которые лучше повторить дословно:

В связи с этим целью данной работы являлось ... (цель должна быть одна).

Для достижения поставленной цели решались следующие задачи:

1. первая задача;
2. вторая задача;
3. третья задача;
4. ...

В конце введения следует добавить описание структуры курсовой работы. Например:

В первом разделе рассмотрена более подробно постановка задачи и проведен обзор литературы по ... Во втором разделе ... В третьем разделе ... В заключении работы сформулированы общие выводы ...

1 РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ PYSPARK

1.1 Постановка задачи разведочного анализа

Здесь нужно сформулировать постановку задачи разведочного анализа: что дано, и что нужно сделать.

1.2 Описание датасета

Здесь требуется описать выбранный датасет, привести ссылку на него, охарактеризовать его тематику, объем, количество признаков. Вкратце нужно описать признаки датасета (допускается описывать не все признаки, а только те, которые используются в исследовании). Также, можно сослаться на источники, например, в [1–4] рассматривается материал об ... Часть информации можно оформить в виде таблицы, но избегайте слишком длинных таблиц. На каждую таблицу должна быть ссылка в тексте, как, например, на таблицу 1, в которой приведен пример описания признаков.

1.3 Определение пропущенных значений

Обратите внимание, что приведенная здесь структура раздела не является жестким требованием, а служит примером оформления. При необходимости, её можно корректировать в разумных пределах.

При необходимости можно вставить рисунок и сослаться на него: на рисунке 1 приведена иллюстрация экосистемы Hadoop. Обратите внимание, что

Таблица 1 – Пример таблицы признаков

Признак	Расшифровка признака
Temperature	Среднемесячная температура в °C
Humidity	Влажность в процентах

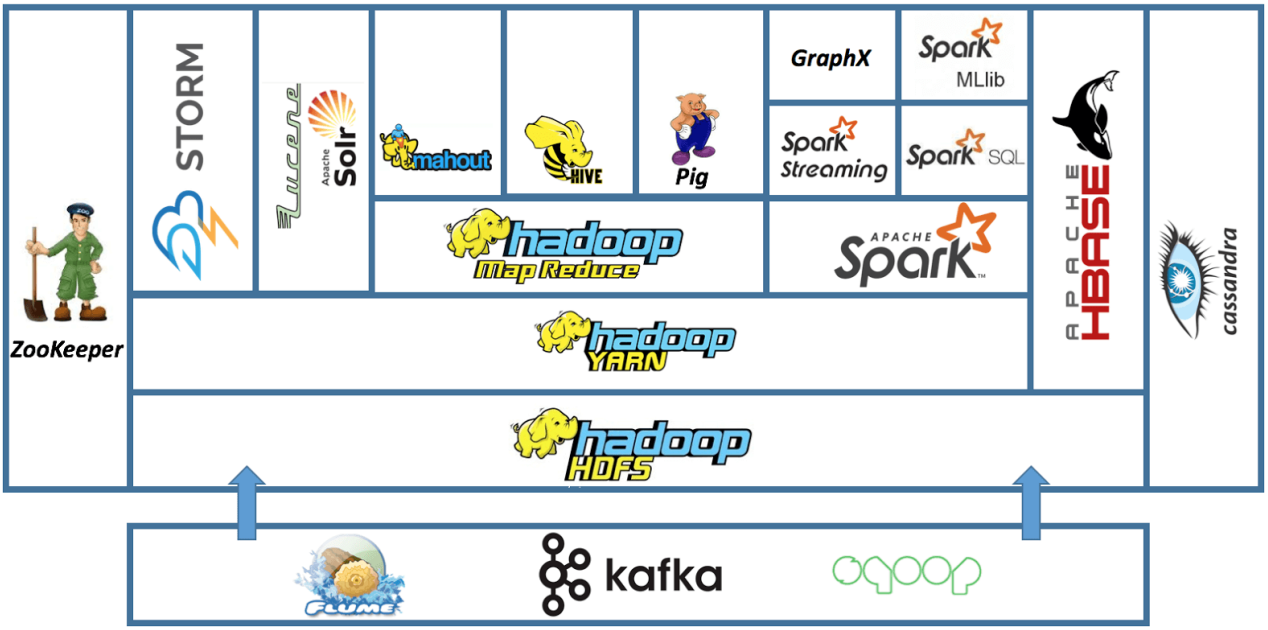

Рисунок 1 – Иллюстрация экосистемы Hadoop

таблицы и рисунки являются «плавающими» объектами: они могут располагаться не в месте их непосредственного объявления, а в некоторой близости от него.

Также, вот пример списка:

– первый элемент;
– второй элемент;
– третий элемент;
    – вложенный элемент.

А здесь – аналогичный нумерованный список:

1. первый элемент;
2. второй элемент;
3. третий элемент;
    a) вложенный элемент.

А вот формула, которая связывает между собой синус, косинус и тангенс.

$$\operatorname{tg}\alpha = \frac{\sin\alpha}{\cos\alpha}. \tag{1}$$

Ну, и ссылка на неё: формула (1) выражает связь тригонометрических функций.

1.4 Другая подглава

Здесь можно продемонстрировать пример включения фрагмента кода в текст работы. И заодно добавить еще пару ссылок [5; 6].

```
import matplotlib.pyplot as plt
import numpy as np
x = np.linspace(0, 10, 100)
y = np.sin(x)
plt.figure(figsize=(10, 6))
plt.plot(x, y, label='sin(x)', color='blue', linewidth=2)
plt.title('График функции sin(x)', fontsize=16)
plt.xlabel('x', fontsize=14)
plt.ylabel('sin(x)', fontsize=14)
plt.legend()
plt.grid(True)
plt.show()
```

Здесь продолжается текст. Не забывайте, что фрагмент кода не должен превышать половины страницы – затем должен следовать текст.

1.5 Выводы

В конце каждой главы должны быть выводы. Они представляют собой несколько предложений, которые кратко характеризуют проделанную в главе работу и полученные результаты.

2 МАШИННОЕ ОБУЧЕНИЕ НА БОЛЬШИХ ДАННЫХ

При необходимости, в главу можно добавить преамбулу с кратким введением в содержание этой главы.

2.1 Задача регресии

2.1.1 Постановка задачи регрессии

Здесь нужно сформулировать поставленную задачу регрессии, которая будет решаться дальше.

2.1.2 Решение задачи регрессии

Здесь подробно описывается решение задачи регрессии.

2.1.3 Анализ полученных результатов

После решения задачи и получения результатов их необходимо проинтерпретировать.

2.2 Задача бинарной классификации

2.2.1 Постановка задачи бинарной классификации

Аналогично задаче регрессии.

2.2.2 Решение задачи бинарной классификации

Аналогично задаче регрессии.

2.2.3 Анализ полученных результатов

Аналогично задаче регрессии.

2.3 Выводы

Сформулированные выводы по главе.

ЗАКЛЮЧЕНИЕ

В заключении коротко приводятся и анализируются полученные результаты, предлагаются дальнейшие направления развития темы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Изучаем Spark: молниеносный анализ данных / Х. Карау [и др.]. — ДМК Пресс, 2015. — 304 с. — ил.
2. Exploratory Data Analysis with pySpark. — 2020. — URL: https://github.com/roshankoirala/pySpark_tutorial/blob/master/Exploratory_data_analysis_with_pySpark.ipynb ; Дата обращения: 19.09.2022.
3. *Уайт Т.* Hadoop: Подробное руководство. — 3-е изд. — СПб. : Питер, 2013. — 672 с. — ил.
4. *Tekdogan T.*, *Cakmak A.* Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification // ICCBDC 2021. Association for Computing Machinery, New York, NY, USA, pages 15-20 (2021). — 2022. — 21 сент. — С. 15—20. — (ICCBDC 2021). — DOI: 10.1145/3481646.3481649. — arXiv: 2209.10637 [cs.DC].
5. Официальный сайт Apache Spark. — 2022. — URL: https://spark.apache.org/ ; Дата обращения: 19.09.2022.
6. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics / М. Захария [и др.] // Conference on Innovative Data Systems Research. — 2021. — URL: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf ; Дата обращения: 12.09.2024.

ПРИЛОЖЕНИЕ А

Пример листинга программного кода

Здесь можно привести полный листинг кода программы или модуля.

```
import matplotlib.pyplot as plt
import numpy as np

# Данные для графика
x = np.linspace(0, 10, 100)
y = np.sin(x)

# Создание графика
plt.figure(figsize=(10, 6))

plt.plot(x, y, label='sin(x)', color='blue', linewidth=2)

# Настройка графика
plt.title('График функции sin(x)', fontsize=16)
plt.xlabel('x', fontsize=14)
plt.ylabel('sin(x)', fontsize=14)
plt.legend()
plt.grid(True)

# Вывод графика
plt.show()
```

ПРИЛОЖЕНИЕ Б

Пример второго приложения

При необходимости, приложений может быть несколько.