Final Exam

You may use any written materials from class (textbook, slides, notes, etc.), as well as supplementary materials from the Internet. However, please *do not*:

- Work or consult with any other student on the exam (independent work only)
- Quote verbatim the Internet or ChatGPT in answering questions—if you use outside sources, please paraphrase in your own words.

Unless otherwise specified, please type your answers directly into this Word document. When you're done, submit the whole document (and any other supplemental files).

Section 1: Advanced SQL Functions and Views 15%

Write a Create view statement to transform the data from the table below into a new format according to the specifications listed below the table. You can use the SQL script below to create the table and insert the rows in your local database instance.

Patient_ ID	Patient_FName	Patient_ MName	Patient_LName	Patient_Phone	Patient_ Phone_ Type	Preferred_ Phone	Phone_Entry_ Date
999	April	R.	Wang	919-555-7777	Work	N	10 Feb 2017
999	April	R.	Wang	919-222-3333	Cell	Y	13 Mar 2022
999	April	R.	Wang	919-999-1111	Home	N	20 Feb 2020
777	Roger	*null*	Fallon	212-333-4444	Cell	Y	8 Apr 2021
888	Roger	*null*	Fallon	212-999-8888	Work	N	30 Apr 2021
222	Rena	*null*	Patel	440-555-2222	Work	Y	16 May 2022
444	Jenna	Shelley	Sonner	780-222-3333	Cell	Y	19 Aug 2023

- I would like to take the three name columns (Patient_FName, _MName, and LName) and concatenate their contents into a single column, Patient_Name. There should be spaces between first, middle, and last names, making sure the concatenation still works as expected when the middle name is null.
- In the view, I only want to see the patient's preferred phone number—not all phone numbers. Filter out non-preferred phone numbers, and store the preferred phone number in a new column, Preferred_Phone_Number. I would like to maintain the preferred phone type and entry date in their own columns (see schema below).
- The Phone_Entry_Date column in the original table is formatted as text. I would like the view to handle this as a real date instead.

The schema for your view is shown below.

Patient_ID	Patient_Name	Patient_Preferred_Phone	Preferred_Phone_Type	Preferred_Phone_Date

```
--Use this script to create the table in your local database. Reference this table when creating the view
CREATE TABLE ALL_PATIENT_PHONES (
    Patient_ID INT,
    Patient_FName VARCHAR(50),
    Patient_MName VARCHAR(50),
    Patient_LName VARCHAR(50),
    Patient_Phone VARCHAR(15),
    Patient_Phone_Type VARCHAR(10),
    Preferred_Phone CHAR(1),
    Phone_Entry_Date VARCHAR(20)
       );

INSERT INTO ALL_PATIENT_PHONES (Patient_ID, Patient_FName, Patient_MName, Patient_LName, Patient_Phone,
Patient_Phone_Type, Preferred_Phone, Phone_Entry_Date)
VALUES
```

```
(999, 'April', 'R.', 'Wang', '919-555-7777', 'Work', 'N', '10 Feb 2017'),
(999, 'April', 'R.', 'Wang', '919-222-3333', 'Cell', 'Y', '13 Mar 2022'),
(999, 'April', 'R.', 'Wang', '919-999-1111', 'Home', 'N', '20 Feb 2020'),
(777, 'Roger', NULL, 'Fallon', '212-333-4444', 'Cell', 'Y', '8 Apr 2021'),
(888, 'Roger', NULL, 'Fallon', '212-999-8888', 'Work', 'N', '30 Apr 2021'),
(222, 'Rena', NULL, 'Patel', '440-555-2222', 'Work', 'Y', '16 May 2022'),
(444, 'Jenna', 'Shelley', 'Sonner', '780-222-3333', 'Cell', 'Y', '19 Aug 2023');
```

ANSWER

```
CREATE VIEW PATIENTS_PREFERRED_PHONE AS
SELECT
        Patient_ID,
        CASE
        WHEN Patient_MName IS NOT NULL THEN
                CONCAT(Patient_FName, ' ', Patient_MName, ' ', Patient_LName)
        ELSE
                CONCAT(Patient_FName, ' ', Patient_LName)
        END AS Patient_Name,
        Patient_Phone AS Patient_Preferred_Phone,
        Patient_Phone_Type AS Preferred_Phone_Type,
        TO_DATE(Phone_Entry_Date, 'DD Mon YYYY') AS Preferred_Phone_Date
FROM ALL_PATIENT_PHONES
WHERE Preferred_Phone = 'Y';
```

Section 2: Regular Expressions 15%

1. UNC email addresses have many potential suffixes, including @live.unc.edu, @email.unc.edu, @med.unc.edu, and just plain @unc.edu. Write a regular expression to match any UNC email address *except* the plain @unc.edu variation. (E.g., your regex should match epfaff@email.unc.edu, but not epfaff@unc.edu.)

   [A-Za-z0-9._%+-]+@(live|email|med)\.unc\.edu

2. Write a regular expression that will match times written in any of the following formats:
   a. 14:56 (24-hour clock, no am/pm)
   b. 2:43 AM (12-hour clock, capitalized AM/PM)
   c. 9:21 pm (12-hour clock, lowercase am/pm)

   You do not need to match those exact times—just those formats, no matter what the time is. However, you must only match valid times on a 12- or 24-hour clock. (E.g., you should not match 65:67, which is not a valid time.)

(2[0-3]|[01]?[0-9]):[0-5][0-9](?:\s?(?:AM|PM|am|pm))?

Section 3: E-R Diagramming 20%

Draw an E-R diagram based on the following database schema. In your diagram, show all primary keys, attributes, relationships, and cardinalities using proper notation. If you make any assumptions, please state them. (Use whatever software you like to create your diagram, and either attach it as a separate file, or copy-paste the image into this document. Please make sure:

- It is legible
- There is no ambiguity in the relationship origin and destination lines and cardinality(avoid line crossings).
- Use a diagram format like the one in the next section.

PATIENT

Patient_ID	FName	LName	Gender

VISIT

Patient_ID	Visit_ID	Clinic_Name	Visit_Start_Date	Visit_End_Date

RACE

Race_ID	Race_Name

PATIENT_RACE

Patient_ID	Race_ID

DIAGNOSIS

Diagnosis_Cd	Diagnosis_Name

VISIT_DIAGNOSIS

Visit_ID	Diagnosis_Cd	Diagnosis_Priority

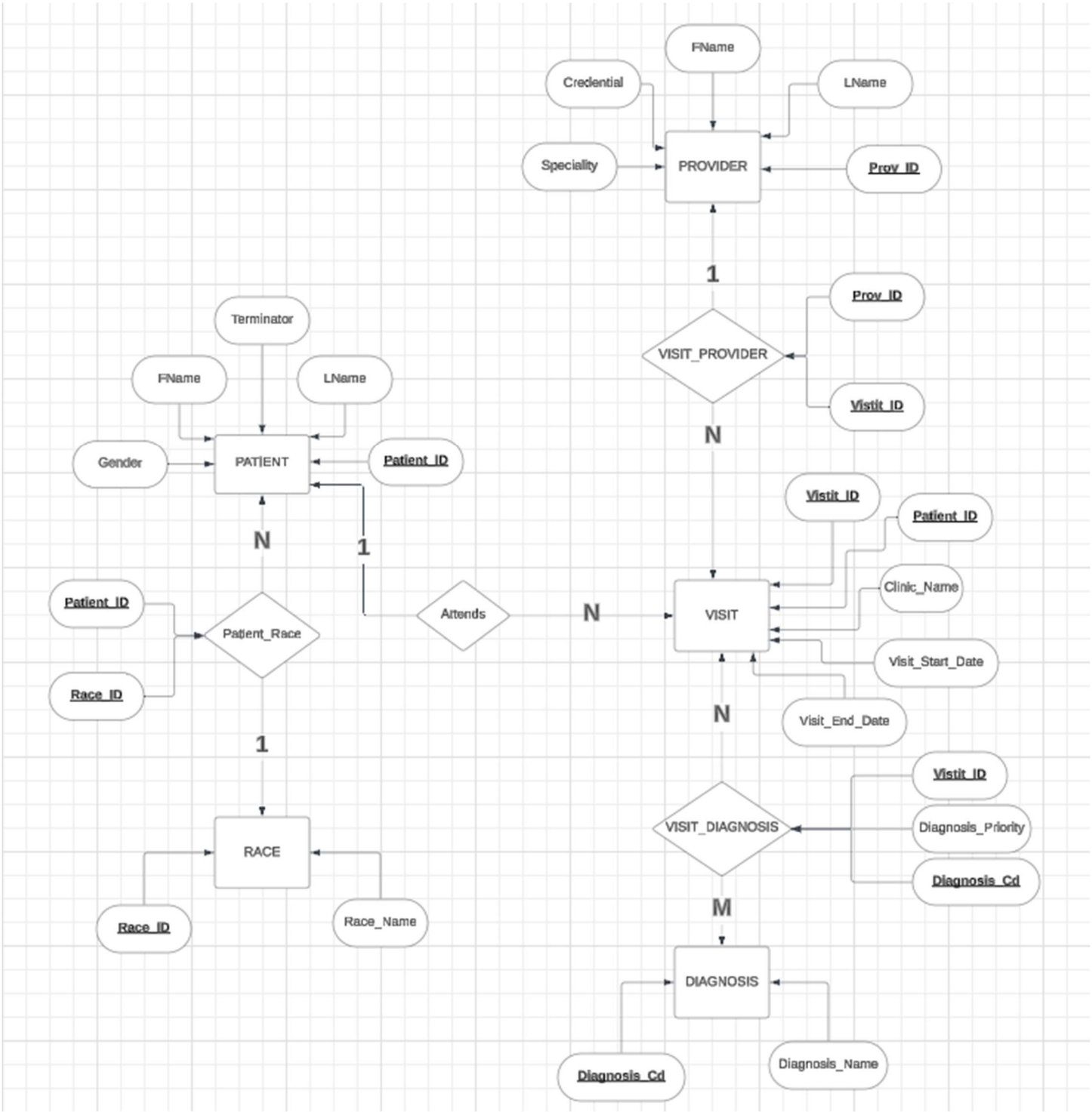
PROVIDER

Prov_ID	FName	LName	Credential	Specialty

VISIT_PROVIDER

Visit_ID	Prov_ID

ANSWER ON NEXT PAGE

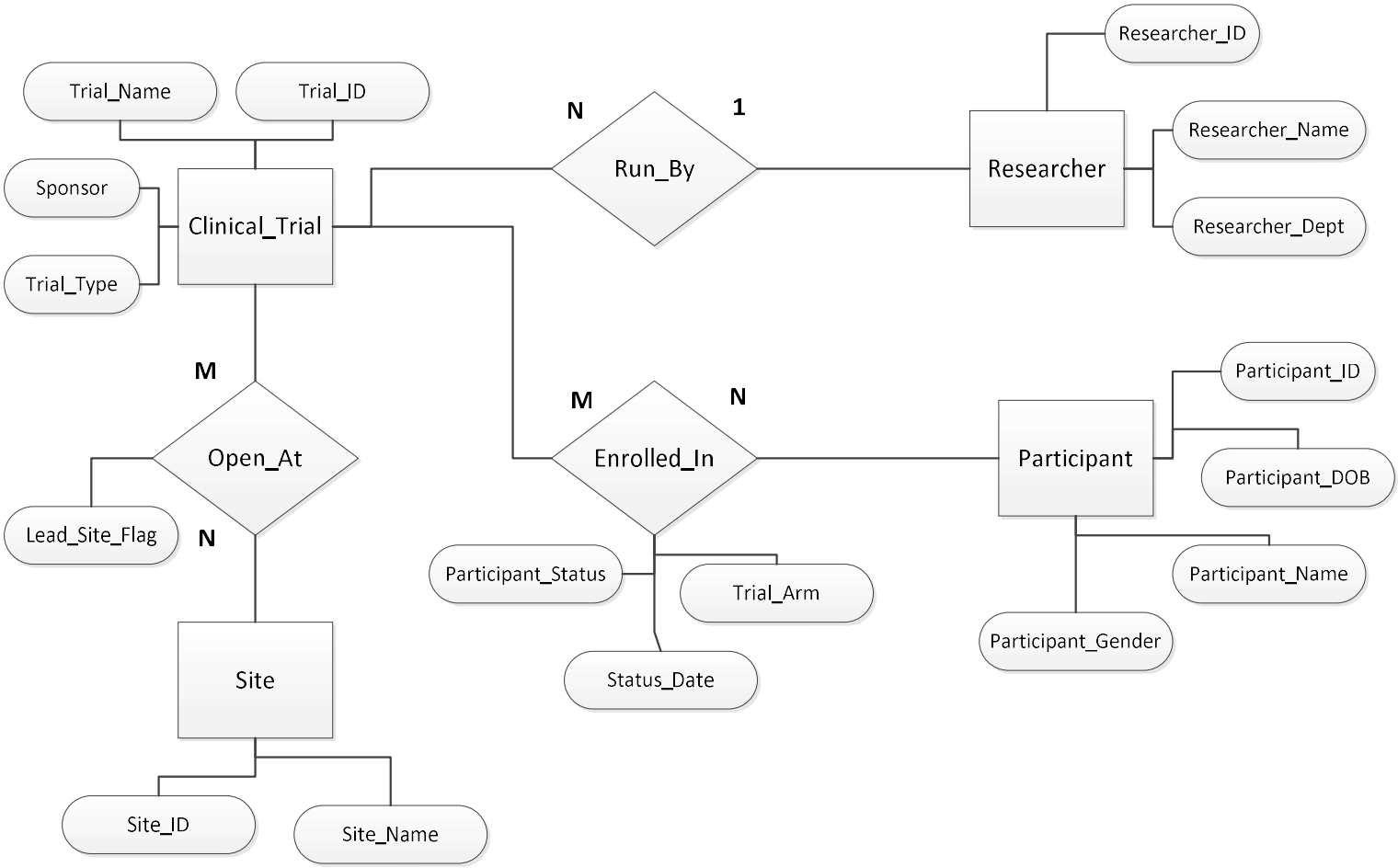


Notes:

I added the attends relationship to combine the tables since visit had the Patient_ID field. I assumed that mixed races had their own IDs, patients could be diagnosed with multiple health conditions in a single visit and a new visit would be created if a patient was transferred to a new provider.

Section 4: E-R Diagram Translation 25%

Write out a database schema based on the following ER diagram. Show primary keys and foreign keys using proper notation. If you make any assumptions, please state them as in line comments. (I have not specified keys in the interest of having you define the primary/foreign keys yourselves—I will respect any assumptions you make in doing so, so long as you state them.) Only represent the following 5 entities: CLINICAL_TRIAL, RUN_BY, RESEARCHER, PARTICIPANT and ENROLLED_IN



ANSWER

Clinical_Trial

Trial_ID	Trial_Name	Sponsor	Trial_Type

Researcher

Researcher_ID	Researcher_Name	Researcher_Dept

Run_By

Trial_ID	Researcher_ID

Participant

Participant_ID	Participant_Name	Participant_DOB	Participant_Gender

Enrolled_In

Participant_ID	Trial_ID	Participant_Status	Trial_Arm	Status_Date

Section 5: Normalization 10%

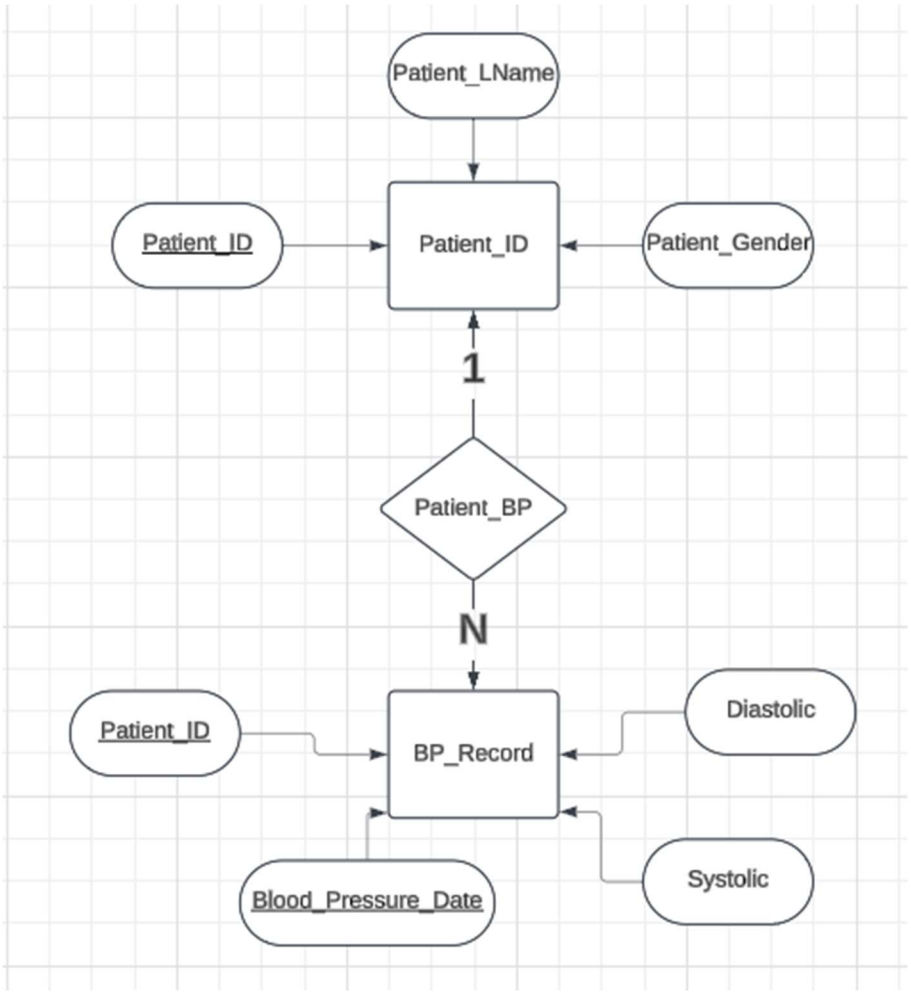
Considering the following table, please (a) state why the table is *not* currently in 3NF, and (b) using an E-R diagram, redesign the table(s) so that it is in 3NF. Identify primary key(s) in your resulting table(s).

Patient_ID	Patient_LName	Patient_Gender	Blood_Pressure	Blood_Pressure_Date
999	CHANG	F	110/70	1/1/2014
999	CHANG	F	114/72	2/5/2015
999	CHANG	F	117/70	6/2/2016
777	CHANG	M	135/80	7/1/2016
888	CHANG	M	120/80	8/1/2016

ANSWER

(a) The blood pressure column is non-atomic because it stores two values. There is a partial dependency because the Patient's Last name and gender solely depend on the Patient's ID and not on the ID, blood pressure and Date. A patient's name and gender can't change for different dates that the blood pressure is taken. There is redundancy because patient information isn't directly related to blood pressure.

(b)



Section 6: Short Answer Questions 15%

The Relational Model

1. Describe at least two things to watch out for when you're using aggregate functions on a column that contains null values.

When performing functions like AVG the null values are ignored so if you are expecting them to be zero than your average could be significantly off depending on the proportion of your data that is null.

If you count a column, it will only count the non-null values vs if you count * then all values will be counted. This difference could lead you to believe you have less records than you do.

2. Explain one advantage and one disadvantage of enforcing referential integrity via foreign keys.

Advantage: An advantage to using referential integrity is that it ensures that the data is consistent. There are no values that point to other tables in which the other table has no record for that value. If this were to happen in a hospital, then a patient could be given medicine without the patient being in the database and if forgotten about the patient could get a double dose.

Disadvantage: A disadvantage to using referential integrity is the strain any modification on the database can cause. When updating a single table, you can impact every other table through their foreign key connections. For the hospital example, giving a patient medicine will require you to add them to the patient table as well and update the insurance table and so on.

3. Define the following concepts – use examples as needed:
   - Insert anomaly

An insert anomaly is when you can't add a record to the database without causing redundant or incomplete information. This often happens when databases aren't in 3NF; an example would be someone tries to add a patient without having a blood pressure record for them yet.

   - Delete anomaly

A delete anomaly is when you unintentionally delete data by deleting records in a table that isn't 3NF. An example would be if someone tries to delete a blood pressure for a patient, but that patient only has 1 blood pressure in the database so as a result you delete all of the patients information.

Non-Relational Databases

1. Describe the difference between a labelled property graph and an RDF triplestore.

An LPG has nodes and edges to represent linked objects whereas RDF uses subject, predicate and object to store the same links but all in a single record. This makes LPGs much more flexible as relationships can be easily changed whereas RDF would require the new relationship to be completely recreated.

2. Briefly describe two different analytical/data storage use cases optimal for a graph database (either labelled property graph or triplestore). Be sure to explain why they are good use cases for the graph model. Use examples other than those we have discussed in class.

Fraud Detection: Fraud detection could use an LPG to represent bank accounts and transactions. This way it would be easier to discover circular transactions where money is sent between a couple of accounts but ends up back in the same account. The architecture of LPGs would make this very easy to spot as there would be circles.

Supply Chain Management: Supply Chain Managers could use RDF Triplestore to track the flow of products from suppliers to warehouses to customers. An example of one triple that would be part of the data would be SupplierA supplies product. This would help track where products came from to easily predict supply chain issues and reorder stock when warehouses are low.

3. Explain one advantage and one disadvantage of using Spark for data processing.

Advantage: Spark excels with large datasets since it can divide computations across multiple nodes allowing results to come much faster.

Disadvantage: Spark isn't as intuitive as simply running data on a laptop. The setup and optimization of Spark requires time and knowledge.