INDENG215 Final Project

Database of Sunshine Orchards in the United States

Team Number 1
Team Members: Jiayi Fang, Xinyu Hou, Xilin Tian, Yunqi Liang

Master of Analytics

IEOR
University of California, Berkeley
Fall 2023

Contents

1 Introduction

1.1 Client Description

Our fictional client is an orchard company called Sunshine Orchards, which grows, produces, and sells a variety of high-quality fruits, including peaches, oranges, and apples. The company's orchards are located in the U.S. and have multiple growing locations, carefully selected to cultivate various fruits based on weather conditions and land characteristics.

1.2 Database Description

The Sunshine Orchards orchard database includes entities such as fruit trees, fruit, farm workers, orders, customers, grown locations, weather conditions, harvest records, facilities, and suppliers, enabling comprehensive monitoring of orchard production, sales, and management through carefully designed relationships. The data requirements are described as follows:

- A unique ID, type/species category, age, location, and last pruning date identify each fruit tree. Fruit seedlings for the orchard are provided by the supplier. The vendors are identified by unique ID, name, contact details, resources provided, and start time. The environment/grown location in which each fruit tree grows, as well as climatic variations/weather conditions, also affects the final fruit that grows. Different growing environments, such as the farm's environmental materials, the farm's location and name, and the amount of ground and altitude occupied, as well as weather conditions, weather status, date, period, and moisture, are all factors taken into account.

- Each fruit from the fruit tree is identified by a unique ID, type, weight, harvest date, and check whether the fruit has a defect. Every fruit is harvested by farm workers. These farm workers need to record specific fruit information including unique record ID, workers' ID, equipment used, harvest date, and harvest amount, and record whether or not a quality audit has taken place. Workers' duties not only include harvesting fruit and recording but also managing the fruit tree. Workers have been identified by their unique ID, name, position, working location, and the date they started working. Equipment that has been in use is determined by specific ID, usage status, and type.

- When a customer wants to purchase fruit, a customer ID is generated that is unique to them, and in the account, there is a name, contact information, and shipping address. When an order is created, each order will have its unique ID, the time the order was created, the amount of fruit purchased, the order status, the shipping form, and the customer ID.

1.3 Database Summary

Below is a summary of entities included in the database and the attributes of each entity.

FruitTree Entity:

TreeID (Primary Key): Unique identifier for each fruit tree.

Type: Variety of the fruit tree (1 for peach tree, 2 for orange tree, 3 for apple tree).

Age: Age of the fruit tree.

Location: Specific planting location of the fruit tree. (Related to the LocationID in the GrownLocation entity)

PeriodWeather: Planting weather of the fruit tree during the period. (Related to the Period in the WeatherCondition entity)

LastPruningDate: Date of the last pruning for the fruit tree.

Fruit Entity:

FruitID (Primary Key): Unique identifier for each fruit.

Type: Variety of the fruit.Related to the Type in the Fruit entity.

Weight: Weight of the fruit.

HarvestDate: The date when the fruit was harvested.

HasDefect: Indicates whether the fruit has defects. (0: No damage, 1: Minor damage, 2: Severe damage)

FarmWorker Entity:

WorkerID (Primary Key): Unique identifier for each farm worker.

FirstName

MiddleName

LastName

Position: Job position of the farm worker. (1 for fruit picking, 2 for fruit transportation)

WorkingLocation: The location where the farm worker is assigned. (Related to the LocationID in the grown location entity)

StartDate: The date when the farm worker started working.

Order Entity:

OrderID (Primary Key): Unique identifier for each sales order.

Date: Creation date of the order.

TotalAmount: Total amount of the order.

CustomerID (Foreign Key): Links to the Customer entity.

OrderStatus: Status of the order. (1: Pending, 2: Processing, 3: Completed)

Shipment: Shipping method. (1: Land, 2: Air, 3: Sea).

Customer Entity:

CustomerID (Primary Key): Unique identifier for each customer.

FirstName

MiddleName

LastName

Address: Customer's address.

PhoneNumber: Customer's contact phone number.

GrownLocation Entity:

LocationID (Primary Key): Unique identifier for each planting location. (1:Dry and sunny place, 2:Dry and gloomy place, 3:Mosit and sunny place, 4:Moist and Glommy place)

Name: Name of the planting location.

Address: Specific address of the planting location.

Area: Acre of the planting location.

Altitude: Altitude of the planting location.

WeatherCondition Entity:

WeatherStatus (Primary Key): Description of the weather conditions. (1: Dry and sunny, 2: Dry and gloomy, 3: Moist and sunny, 4: Moist and gloomy)

Date: Date of the weather record.

Period: Three months a period for 2023. (Period 1 January, February, and March; Period 2 April, May, and June; Period 3 July, August, and September; Period 4 October, November, and December)

Temperature: Temperature on the recorded day.

Humidity: Humidity on the recorded day.

Condition: Description of the weather conditions.

HarvestRecord Entity:

RecordID (Primary Key): Unique identifier for each harvest record.

PickupID (Foreign Key): Link to the Equipment/Facility entity. (PickupID is the pickup equipment ID)

StorageID (Foreign Key): Link to the Equipment/Facility entity. (StorageID is the storage equipment ID

OrderID (Foreign Key): Link to the Order entity.

HarvestDate: Date of the harvest record.

HarvestAmount: Quantity harvested in the record.

HarvesterID (Foreign Key): Link to the FarmWorker entity. (HarvesterID is the FarmWorker's ID)

IsQualityApproved: Indicates whether the harvest is quality-approved. (1: Approved, 2: Not approved)

Equipment/Facility Entity:

FacilityID (Primary Key): Unique identifier for each storage facility.

Type: Type of the facility. (storage or pickup)

Capacity: Capacity of the storage facility.

UsageStatus: Current usage status of all facilities.

Efficiency: Efficiency of the pickup facility.

Supplier Entity:

SupplierID (Primary Key): Unique identifier for each supplier.

Name: Name of the supplier.

ContactInfo: Contact information for the supplier.

SuppliedTreeID: Fruit seedlings supplied by the supplier.

StartDate: The date when the collaboration with the supplier started.

2 Simplified EER

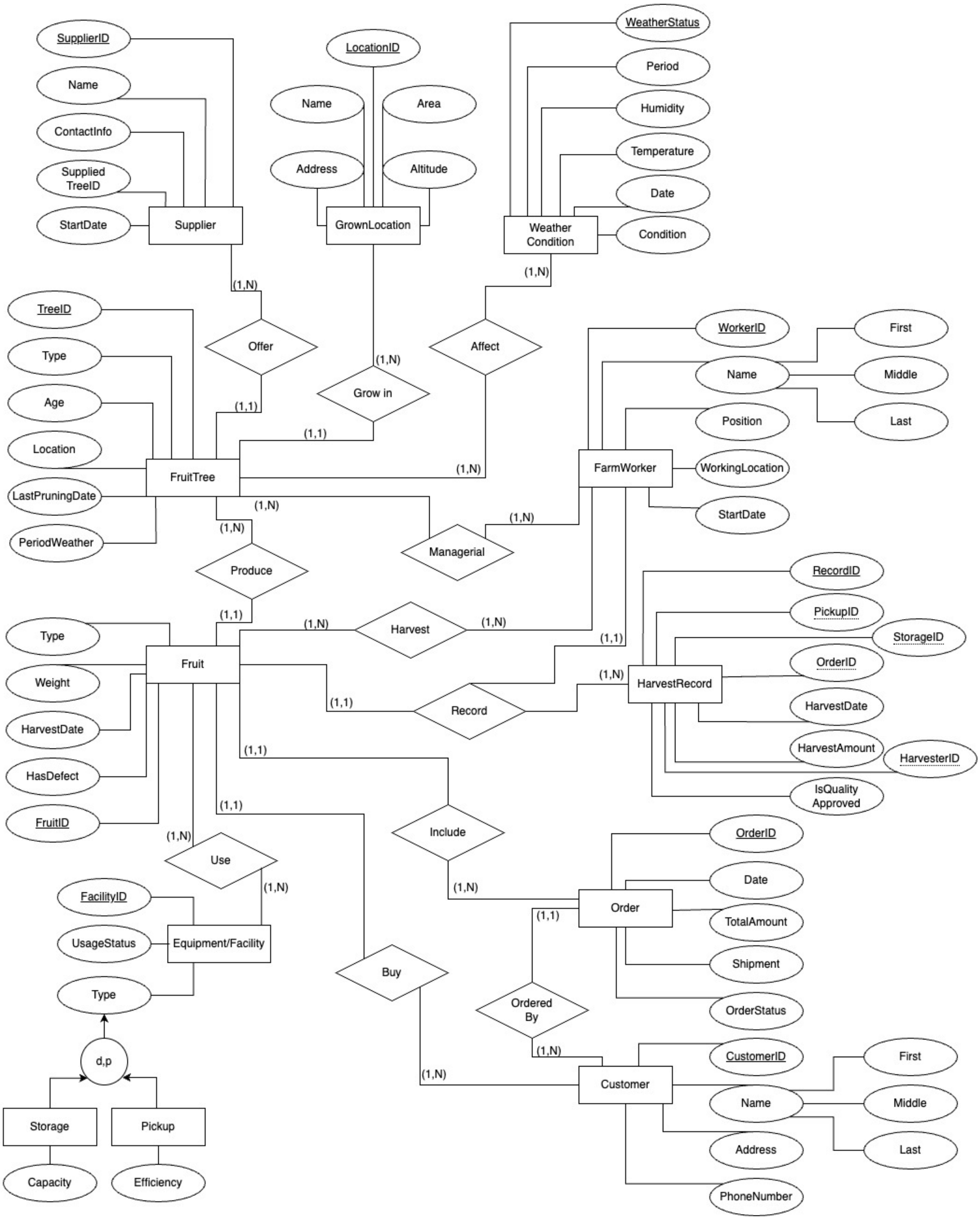
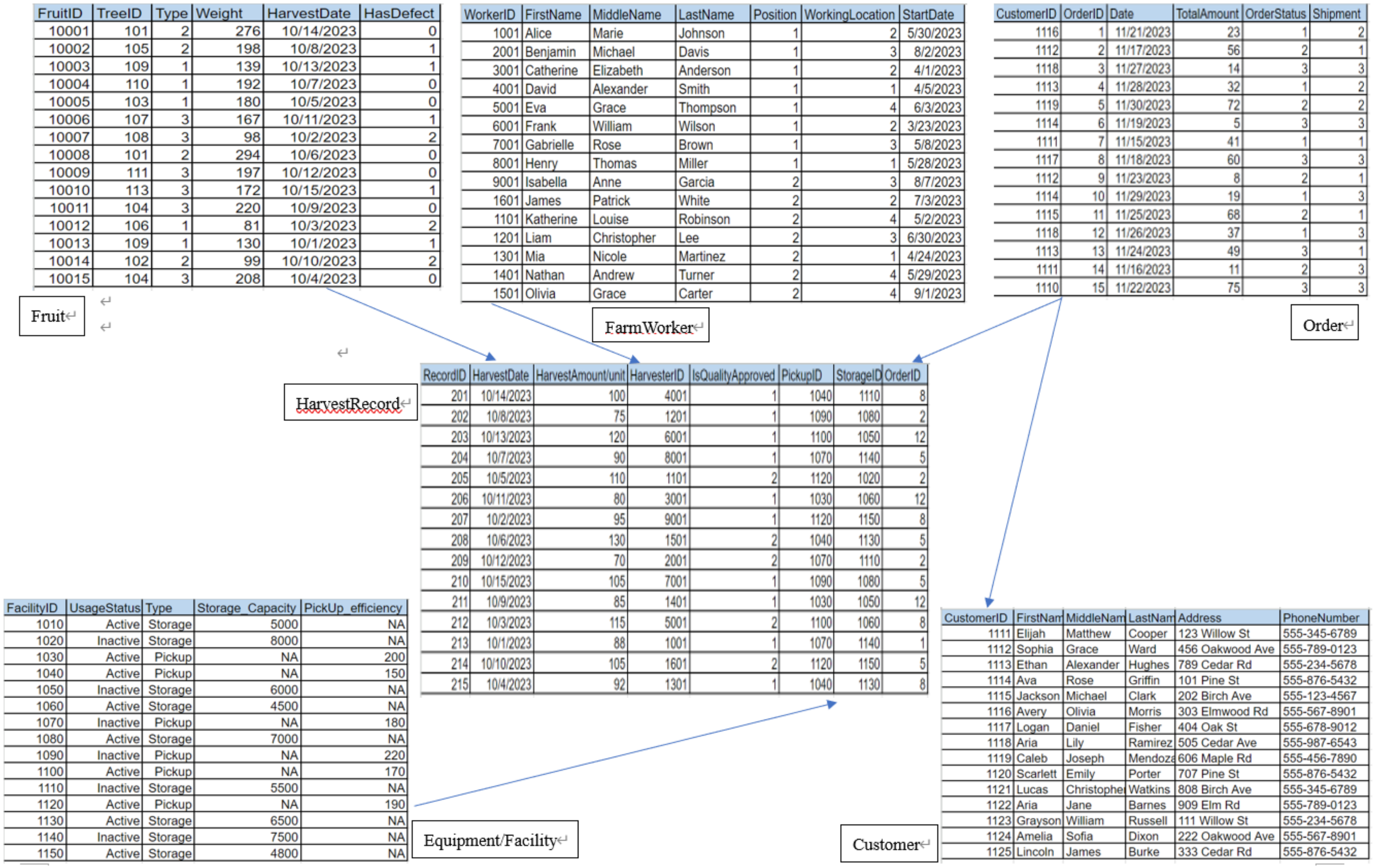


Figure 1: EER Diagram

3 Relational design (Schema)



FruitID	TreeID	Type	Weight	HarvestDate	HasDefect
10001	101	2	276	10/14/2023	0
10002	105	2	198	10/8/2023	1
10003	109	1	139	10/13/2023	1
10004	110	1	192	10/7/2023	0
10005	103	1	180	10/5/2023	0
10006	107	3	167	10/11/2023	1
10007	108	3	98	10/2/2023	2
10008	101	2	294	10/6/2023	0
10009	111	3	197	10/12/2023	0
10010	113	3	172	10/15/2023	1
10011	104	3	220	10/9/2023	0
10012	106	1	81	10/3/2023	2
10013	109	1	130	10/1/2023	1
10014	102	2	99	10/10/2023	2
10015	104	3	208	10/4/2023	0

Fruit

WorkerID	FirstName	MiddleName	LastName	Position	WorkingLocation	StartDate
1001	Alice	Marie	Johnson	1	2	5/30/2023
2001	Benjamin	Michael	Davis	1	3	8/2/2023
3001	Catherine	Elizabeth	Anderson	1	2	4/1/2023
4001	David	Alexander	Smith	1	1	4/5/2023
5001	Eva	Grace	Thompson	1	4	6/3/2023
6001	Frank	William	Wilson	1	2	3/23/2023
7001	Gabrielle	Rose	Brown	1	3	5/8/2023
8001	Henry	Thomas	Miller	1	1	5/28/2023
9001	Isabella	Anne	Garcia	2	3	8/7/2023
1601	James	Patrick	White	2	2	7/3/2023
1101	Katherine	Louise	Robinson	2	4	5/2/2023
1201	Liam	Christopher	Lee	2	3	6/30/2023
1301	Mia	Nicole	Martinez	2	1	4/24/2023
1401	Nathan	Andrew	Turner	2	4	5/29/2023
1501	Olivia	Grace	Carter	2	4	9/1/2023

FarmWorker

CustomerID	OrderID	Date	TotalAmount	OrderStatus	Shipment
1116	1	11/21/2023	23	1	2
1112	2	11/17/2023	56	2	1
1118	3	11/27/2023	14	3	3
1113	4	11/28/2023	32	1	2
1119	5	11/30/2023	72	2	2
1114	6	11/19/2023	5	3	3
1111	7	11/15/2023	41	1	1
1117	8	11/18/2023	60	3	3
1112	9	11/23/2023	8	2	1
1114	10	11/29/2023	19	1	3
1115	11	11/25/2023	68	2	1
1118	12	11/26/2023	37	1	3
1113	13	11/24/2023	49	3	1
1111	14	11/16/2023	11	2	3
1110	15	11/22/2023	75	3	3

Order

HarvestRecord

RecordID	HarvestDate	HarvestAmount/unit	HarvesterID	IsQualityApproved	PickupID	StorageID	OrderID
201	10/14/2023	100	4001	1	1040	1110	8
202	10/8/2023	75	1201	1	1090	1080	2
203	10/13/2023	120	6001	1	1100	1050	12
204	10/7/2023	90	8001	1	1070	1140	5
205	10/5/2023	110	1101	2	1120	1020	2
206	10/11/2023	80	3001	1	1030	1060	12
207	10/2/2023	95	9001	1	1120	1150	8
208	10/6/2023	130	1501	2	1040	1130	5
209	10/12/2023	70	2001	2	1070	1110	2
210	10/15/2023	105	7001	1	1090	1080	5
211	10/9/2023	85	1401	1	1030	1050	12
212	10/3/2023	115	5001	2	1100	1060	8
213	10/1/2023	88	1001	1	1070	1140	1
214	10/10/2023	105	1601	2	1120	1150	5
215	10/4/2023	92	1301	1	1040	1130	8

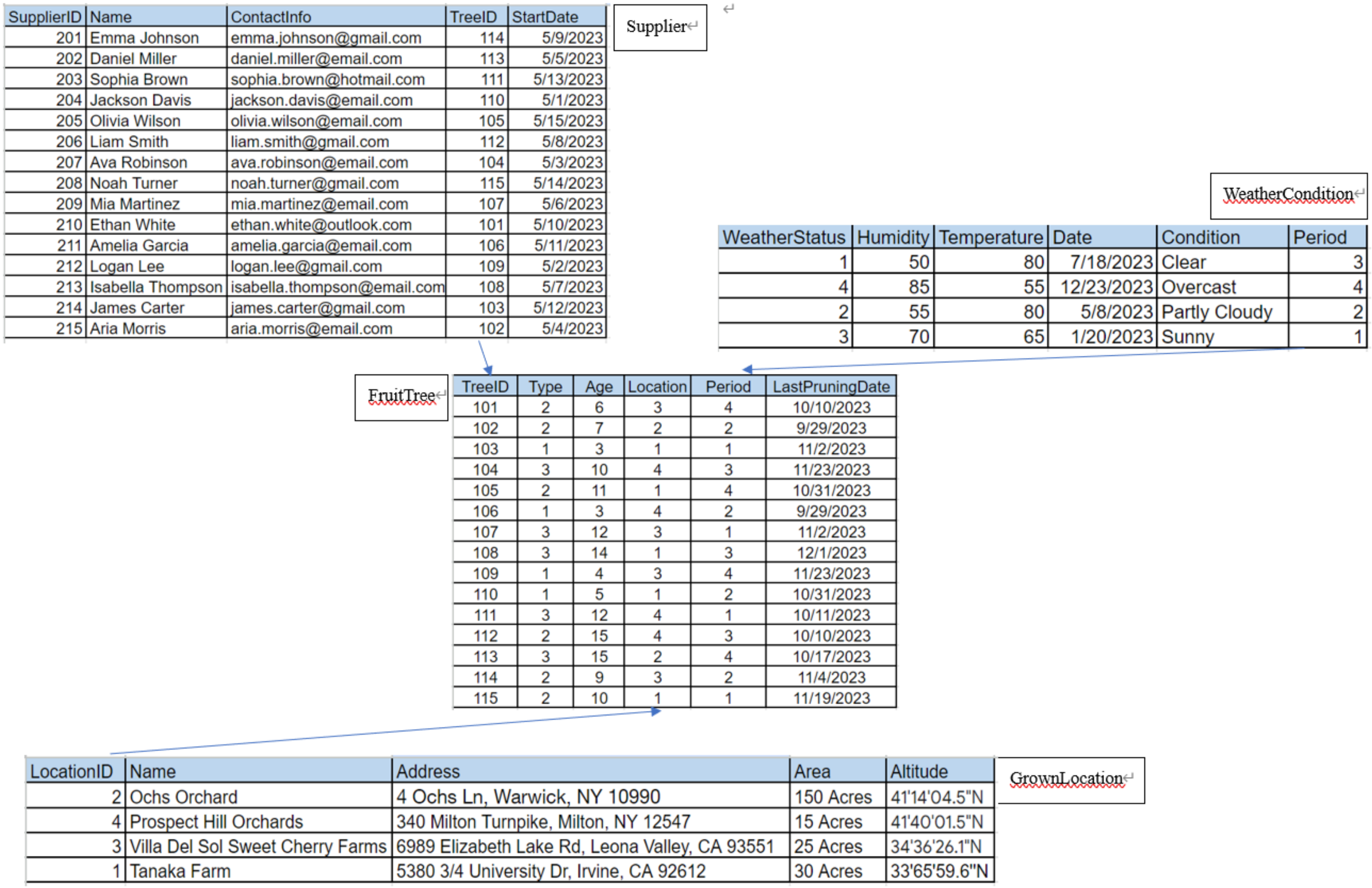
FacilityID	UsageStatus	Type	Storage_Capacity	PickUp_efficiency
1010	Active	Storage	5000	NA
1020	Inactive	Storage	8000	NA
1030	Active	Pickup	NA	200
1040	Active	Pickup	NA	150
1050	Inactive	Storage	6000	NA
1060	Active	Storage	4500	NA
1070	Inactive	Pickup	NA	180
1080	Active	Storage	7000	NA
1090	Inactive	Pickup	NA	220
1100	Active	Pickup	NA	170
1110	Inactive	Storage	5500	NA
1120	Active	Pickup	NA	190
1130	Active	Storage	6500	NA
1140	Inactive	Storage	7500	NA
1150	Active	Storage	4800	NA

Equipment/Facility

CustomerID	FirstNam	MiddleNam	LastNam	Address	PhoneNumber
1111	Elijah	Matthew	Cooper	123 Willow St	555-345-6789
1112	Sophia	Grace	Ward	456 Oakwood Ave	555-789-0123
1113	Ethan	Alexander	Hughes	789 Cedar Rd	555-234-5678
1114	Ava	Rose	Griffin	101 Pine St	555-876-5432
1115	Jackson	Michael	Clark	202 Birch Ave	555-123-4567
1116	Avery	Olivia	Morris	303 Elmwood Rd	555-567-8901
1117	Logan	Daniel	Fisher	404 Oak St	555-678-9012
1118	Aria	Lily	Ramirez	505 Cedar Ave	555-987-6543
1119	Caleb	Joseph	Mendoza	606 Maple Rd	555-456-7890
1120	Scarlett	Emily	Porter	707 Pine St	555-876-5432
1121	Lucas	Christopher	Watkins	808 Birch Ave	555-345-6789
1122	Aria	Jane	Barnes	909 Elm Rd	555-789-0123
1123	Grayson	William	Russell	111 Willow St	555-234-5678
1124	Amelia	Sofia	Dixon	222 Oakwood Ave	555-567-8901
1125	Lincoln	James	Burke	333 Cedar Rd	555-876-5432

Customer

Figure 2: Schema1



Supplier

SupplierID	Name	ContactInfo	TreeID	StartDate
201	Emma Johnson	emma.johnson@gmail.com	114	5/9/2023
202	Daniel Miller	daniel.miller@email.com	113	5/5/2023
203	Sophia Brown	sophia.brown@hotmail.com	111	5/13/2023
204	Jackson Davis	jackson.davis@email.com	110	5/1/2023
205	Olivia Wilson	olivia.wilson@email.com	105	5/15/2023
206	Liam Smith	liam.smith@gmail.com	112	5/8/2023
207	Ava Robinson	ava.robinson@email.com	104	5/3/2023
208	Noah Turner	noah.turner@gmail.com	115	5/14/2023
209	Mia Martinez	mia.martinez@email.com	107	5/6/2023
210	Ethan White	ethan.white@outlook.com	101	5/10/2023
211	Amelia Garcia	amelia.garcia@email.com	106	5/11/2023
212	Logan Lee	logan.lee@gmail.com	109	5/2/2023
213	Isabella Thompson	isabella.thompson@email.com	108	5/7/2023
214	James Carter	james.carter@gmail.com	103	5/12/2023
215	Aria Morris	aria.morris@email.com	102	5/4/2023

WeatherCondition

WeatherStatus	Humidity	Temperature	Date	Condition	Period
1	50	80	7/18/2023	Clear	3
4	85	55	12/23/2023	Overcast	4
2	55	80	5/8/2023	Partly Cloudy	2
3	70	65	1/20/2023	Sunny	1

FruitTree

TreeID	Type	Age	Location	Period	LastPruningDate
101	2	6	3	4	10/10/2023
102	2	7	2	2	9/29/2023
103	1	3	1	1	11/2/2023
104	3	10	4	3	11/23/2023
105	2	11	1	4	10/31/2023
106	1	3	4	2	9/29/2023
107	3	12	3	1	11/2/2023
108	3	14	1	3	12/1/2023
109	1	4	3	4	11/23/2023
110	1	5	1	2	10/31/2023
111	3	12	4	1	10/11/2023
112	2	15	4	3	10/10/2023
113	3	15	2	4	10/17/2023
114	2	9	3	2	11/4/2023
115	2	10	1	1	11/19/2023

GrownLocation

LocationID	Name	Address	Area	Altitude
2	Ochs Orchard	4 Ochs Ln, Warwick, NY 10990	150 Acres	41'14'04.5"N
4	Prospect Hill Orchards	340 Milton Turnpike, Milton, NY 12547	15 Acres	41'40'01.5"N
3	Villa Del Sol Sweet Cherry Farms	6989 Elizabeth Lake Rd, Leona Valley, CA 93551	25 Acres	34'36'26.1"N
1	Tanaka Farm	5380 3/4 University Dr, Irvine, CA 92612	30 Acres	33'65'59.6"N

Figure 3: Schema2

4 MySQL

4.1 MySQL Relationship View

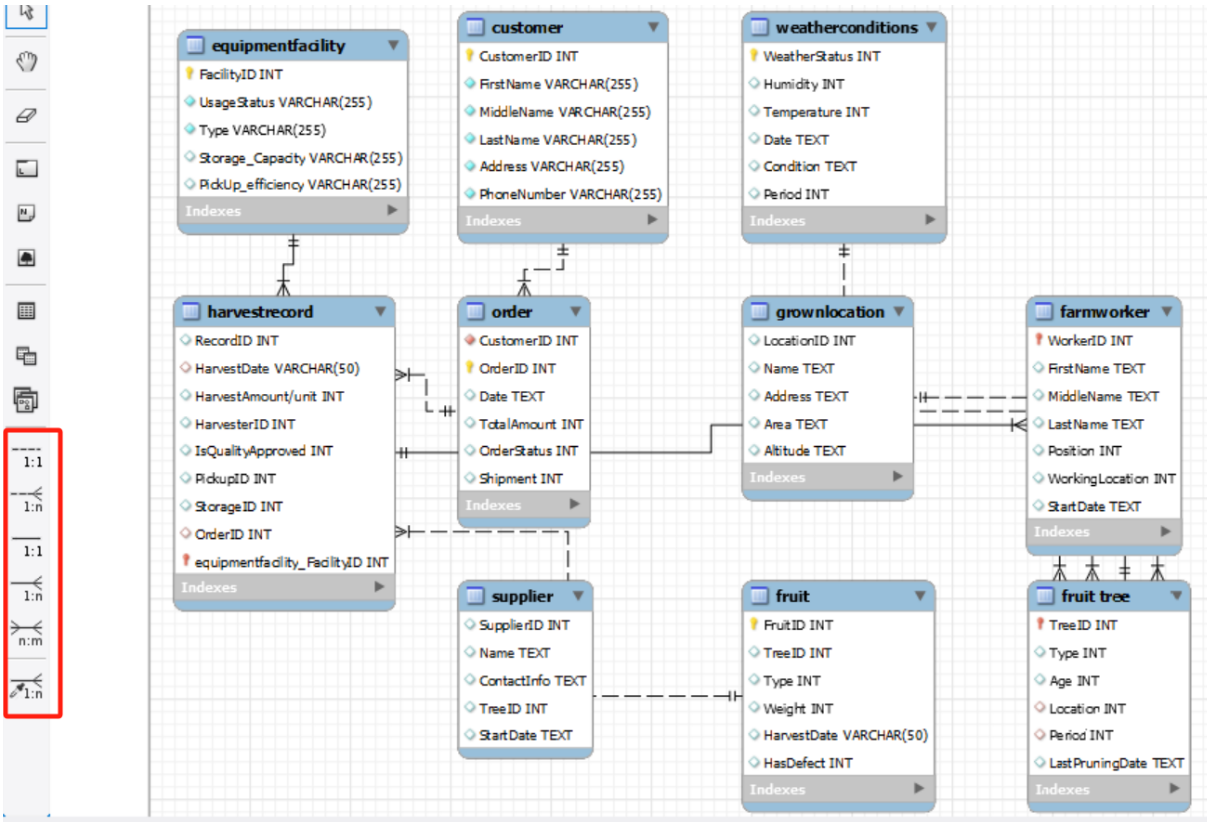


Figure 4: MySQL Relationship View

4.2 Create and Insert Synthetic Data

To compile some of the data, we searched the internet to find a large amount of real data, including actual weights, age of fruit trees, farm names, and farm addresses. For fields such as name, contact information, date, and address, we used Python to generate the data randomly to exclude the possible influence of subjective factors on the data. The IDs were then automatically created in ascending order. Following these steps, we created ten tables (CSV files), each containing 4 or 15 rows of data.

Next, we'll use MySQL Workbench to import the CSV file. First, create a database named 'Final Project'. Secondly, create 10 new tables under this database. Thirdly, In the third part, we open the table where we want to load the data. By clicking the Import button, select the CSV file, and click the Open button. Review the data and click the Apply button. When the MySQL Workbench displays the Apply SQL Script to Database dialog box, click the Apply button to insert the data into the table.

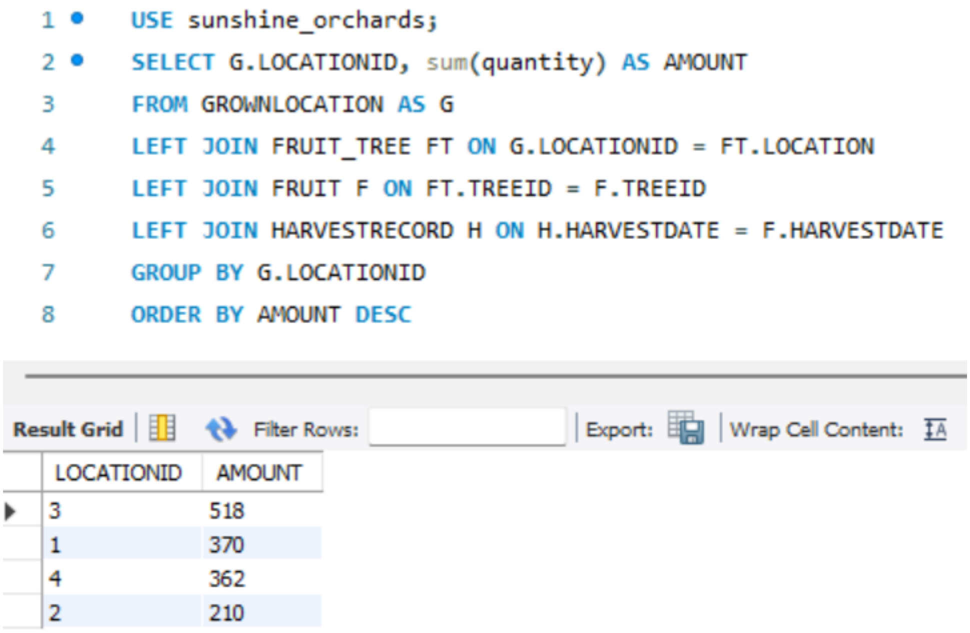
After importing the data, we also defined the primary key, foreign key, and data type. First double click on the table name to open the table design. In the "Columns" section, we define the data type for each column, such as INT, VARCHAR, etc. In the "Indexes" section, we define the Primary Key. In the "Foreign Keys" section, we define the Foreign Key.

5 “Interesting” Queries

5.1 Location Distribution

As an owner of this orchard, the choice of where to plant the fruit trees is also very important, as this will determine the final yield and quality of the fruit. So our client may be interested in the locations that produce the most fruits. Thus, the interesting query could be as follows:

Find the top 2 performing growing locations based on total harvested quantity



```sql
USE sunshine_orchards;
SELECT G.LOCATIONID, sum(quantity) AS AMOUNT
FROM GROWNLOCATION AS G
LEFT JOIN FRUIT_TREE FT ON G.LOCATIONID = FT.LOCATION
LEFT JOIN FRUIT F ON FT.TREEID = F.TREEID
LEFT JOIN HARVESTRECORD H ON H.HARVESTDATE = F.HARVESTDATE
GROUP BY G.LOCATIONID
ORDER BY AMOUNT DESC
```

```
LOCATIONID  AMOUNT
3           518
1           370
4           362
2           210
```

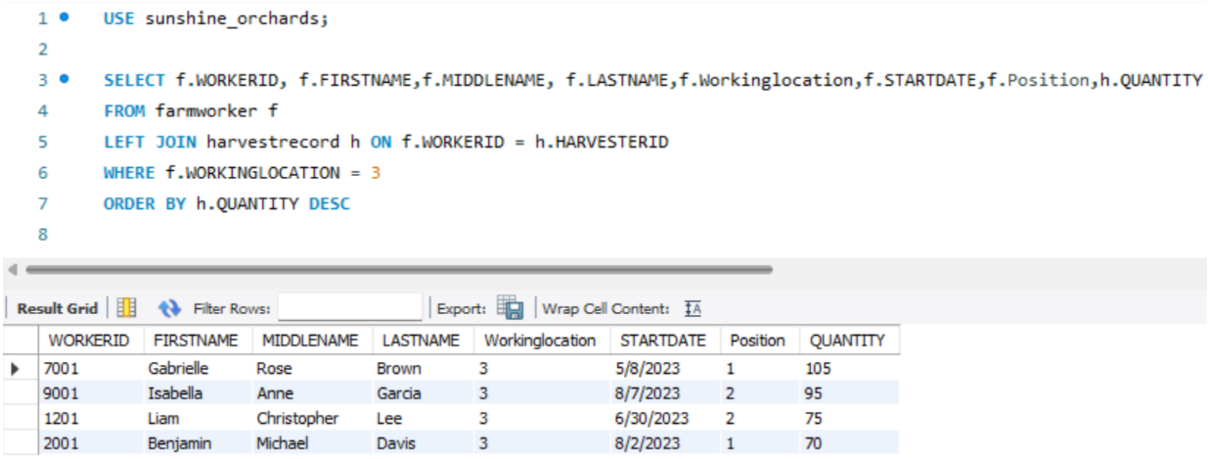
Figure 5: SQL and Result 1

The top 2 growing locations based on total harvested quantity are location NO.3, Villa Del Sol Sweet Cherry Farms, which harvested 518 units of fruit, and location NO.1, Tanaka Farm, which harvested 370 units of fruit.

5.2 Productive Farm Worker

Workers have many responsibilities such as managing fruit trees, harvesting fruits, and keeping records of these harvests. As an owner of this orchard, our client may need to have a reward system to manage our employees better. After the first query, we noticed that location No.3, Villa Del Sol Sweet Cherry Farms, has the highest harvested quantity. Hence, we want to find the most productive workers at location No.3. Thus, the interesting query could be as follows:

Identify the most productive farm worker who works at location No.3 by calculating the total quantity of fruits harvested by each worker



```sql
USE sunshine_orchards;

SELECT f.WORKERID, f.FIRSTNAME,f.MIDDLENAME, f.LASTNAME,f.Workinglocation,f.STARTDATE,f.Position,h.QUANTITY
FROM farmworker f
LEFT JOIN harvestrecord h ON f.WORKERID = h.HARVESTERID
WHERE f.WORKINGLOCATION = 3
ORDER BY h.QUANTITY DESC

```

WORKERID	FIRSTNAME	MIDDLENAME	LASTNAME	Workinglocation	STARTDATE	Position	QUANTITY
7001	Gabrielle	Rose	Brown	3	5/8/2023	1	105
9001	Isabella	Anne	Garcia	3	8/7/2023	2	95
1201	Liam	Christopher	Lee	3	6/30/2023	2	75
2001	Benjamin	Michael	Davis	3	8/2/2023	1	70

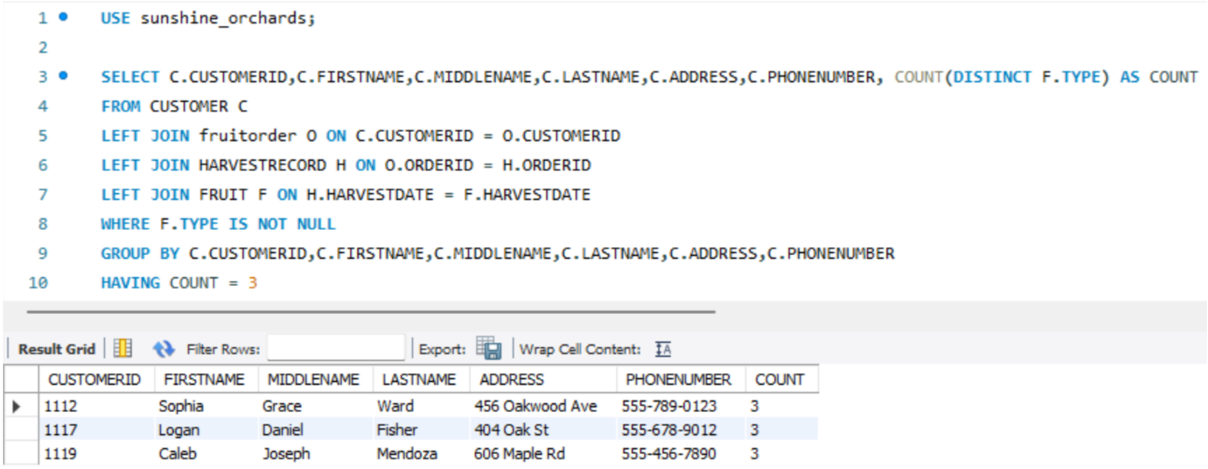
Figure 6: SQL and Result 2

At location NO.3, Villa Del Sol Sweet Cherry Farms, the most productive farm worker is Gabrielle Rose Brown who harvested 105 units of fruit as a pickup worker.

5.3 Customer Purchasing Needs

Every customer has different needs, some may only love one type of fruit, some fruits may not be an immediate need, and as someone who owns an orchard, our client may be more interested in which customers will want all of the fruit. Thus, the interesting query could be as follows:

Retrieve a list of customers who purchased all types of fruit in the last quarter.



```sql
USE sunshine_orchards;

SELECT C.CUSTOMERID,C.FIRSTNAME,C.MIDDLENAME,C.LASTNAME,C.ADDRESS,C.PHONENUMBER, COUNT(DISTINCT F.TYPE) AS COUNT
FROM CUSTOMER C
LEFT JOIN fruitorder O ON C.CUSTOMERID = O.CUSTOMERID
LEFT JOIN HARVESTRECORD H ON O.ORDERID = H.ORDERID
LEFT JOIN FRUIT F ON H.HARVESTDATE = F.HARVESTDATE
WHERE F.TYPE IS NOT NULL
GROUP BY C.CUSTOMERID,C.FIRSTNAME,C.MIDDLENAME,C.LASTNAME,C.ADDRESS,C.PHONENUMBER
HAVING COUNT = 3
```

CUSTOMERID	FIRSTNAME	MIDDLENAME	LASTNAME	ADDRESS	PHONENUMBER	COUNT
1112	Sophia	Grace	Ward	456 Oakwood Ave	555-789-0123	3
1117	Logan	Daniel	Fisher	404 Oak St	555-678-9012	3
1119	Caleb	Joseph	Mendoza	606 Maple Rd	555-456-7890	3

Figure 7: SQL and Result 3

During the last quarter, July, August, and September, Sophia Grace Ward, Logan Daniel Fisher, and Caleb Joseph Mendoza were the customers who purchased all types of fruit.

6 Potential Analysis

SQL is a standardized language used for managing and manipulating relational databases. It enables users to perform various tasks, including data retrieval, insertion, updating, and deletion, through queries and update statements.

Once we've queried or updated a database using SQL, we can use other tools and programming languages to further process and analyze the data. In tasks involving orchard data, Python is a powerful tool with a rich ecosystem of data-processing libraries.

For example, if we want to analyze the historical performance of different fruit varieties in orchards, use Python with Pandas (Use Pandas for data analysis and cleaning, facilitating operations on tabular data, such as filtering, transformation, and aggregation) and Matplotlib/Seaborn (Matplotlib and Seaborn libraries are used for data visualization, helping to gain a better understanding of data distributions and trends) to analyze historical data, visualize trends, and identify the performance of various fruit varieties over time.

Similarly, if we need to predict future harvest yields based on historical data and weather conditions, we can implement predictive modeling and leverage machine learning libraries like Scikit-learn (this library is used for machine learning tasks, including data preprocessing, model training, and evaluation). Train a model on historical harvest and weather data to make future yield predictions.

Other things we are curious about such as optimizing resource allocation among different orchard locations, visualizing the spatial distribution of fruit tree health in our orchards, and tools that can help in creating reports and visual dashboards for orchard performance, etc. could also be potentially wanted by the clients to get analyzed as well.