Updated Schema Design:

Since the IOT sensor is capturing the data for every store per second the no of rows in the table for storing all this data will be:

Number of rows: 3600*24*1000*365 = 31,536,000,000

Now for analyzing the sales performance with respect to weather data we will need the weather data only when the transaction is made. So, we will not store every reading of the IOT sensor in a separate table (which will be very large) in our DW instead we will accommodate the IOT data captured with respect to every transaction in the existing tables only.

Assuming we have all the data from IOT sensor is in numeric form (including the weather condition which is a categorical column) and is the outcome of a process of sensing the weather condition we can treat this data as a Facts.

Except for the Time of the Day, we will create a separate time dimension table for this metrics to avoid using calendar functions while building reports.
This table will be at minute level granularity (Assuming that we are not creating any reports at second level granularity)

2 possible ways of accommodating the IOT data:

1. Create another fact table at order level granularity with IOT sensor and transaction data.

Pros:
Less number of rows will be added in this fact table. Per transaction only 1 row will be added as compared to 2 rows in the exiting fact table since on an average each order has 2 order-line items.
Cons:
We cannot analyze the sales performance with respect to weather condition across product dimension since product information for each transaction is at order-line granularity.

2. Add the IOT data metrics as new attributes(columns) in the existing fact table which is at order-line level granularity.

Pros:
We can analyze the sales performance with respect to weather condition at product level along with other dimensions.
Cons:
Weather data will be duplicated for each product purchased in the same transaction on an average 2 rows per transaction. So, there will be more number of rows in this table.

3. Create a separate fact table only with IOT sensor data recorded for every Transaction and keep the existing fact table as is

Pros:

Less number of rows will be added in 2nd fact table. Per transaction only 1 row will be added as compared to 2 rows in the exiting fact table since on an average each order has 2 order-line items. And existing fact table will only have transaction data so overall storage cost for both the fact tables will be less as compared to 2nd approach.

Cons:
In this design for analyzing the sales performance with respect to weather data across different dimensions we must make multiple joins and have to join 2 fact tables.

After considering pros and cons for both the approaches it is better to choose the 2nd approach and add all the weather data in the existing fact table. Since even if this increases the overall size of the table (storage wise) number of rows in the fact table will be same as current design. And we don't have to perform to many joins for analyzing sales performance with respect to weather data across different dimensions.

New Fact Table Attributes:

Attribute	**Description**	**Data Type**	**Range**
DimDateID	FK from the date dimension table	Int	1 to 365
DimStoreID	FK from the Store dimension table	Int	1 to 1000
DimCustomerID	FK from the Customer dimension table	Int	1 to 1,000,000 (assuming these many new customers visits across all stores in 1 year)
DimProductID	FK from the Product dimension table	Int	1 to 255 (assuming there are 255 unique products across all the stores)
DimEmployeeID	FK from the Employee dimension table	Int	1 to 20000 (assuming 20 employees per store)
TransactionNo	Transaction number from operational database	Int	1 to 547500000 (1 number per transaction)
SalesAmount	Sales amount of the product	Float (3,2)	$1.00 to $100.00
Queue_Length	Average que length at register	Int	1 to 50

Indoor_Temperature	Indoor Temperature	Float (2,2)	-50.00 to 150.00
Outdoor_Temperature	Outdoor Temperature	Float (2,2)	-50.00 to 150.00
DimTimeID	FK from the Time dimension table	Int	1 to 86400
Humidity	Relative humidity	Float (2,2)	0.00 to 100.00
Cloud Cover	Cloud condition	Float (2,2)	0.00 to 100.00
Weather condition	Overall weather condition	Int	1 to 10
Pollen	Rating based on presence of pollen in the air.	Int	1 to 10
Traffic Conditions	Score based on condition of traffic	Int	-10000 to +10000
Seating	Available seating %	Int	0 to 100
Foot Traffic Score	Score based on footfall	Int	0 to 10000
Indoor Noise	Indoor noise level	Float (3,2)	0.00 to 140.00
Outdoor Noise	Outdoor noise level	Float (3,2)	0.00 to 140.00

Time dimension Table Attributes:

Attribute	Description	Data Type	Range
TimeID	Surrogate key for the	Int	1 to 86400
Hour	Hours of the day	Int	0 to 23
Minute	Minutes of the hour	Int	0 to 59

Estimate of new fact table size for 365 days transactions:

Attribute	Space Occupied
DimDateID	4 bytes
DimStoreID	4 bytes
DimCustomerID	4 bytes
DimProductID	4 bytes
DimEmployeeID	4 bytes
TransactionNo	4 bytes
SalesAmount	4 bytes
Queue_Length	4 bytes
Indoor_Temperature	4 bytes
Outdoor_Temperature	4 bytes
DimTimeID	4 bytes
Humidity	4 bytes

Cloud Cover	4 bytes
Weather condition	4 bytes
Pollen	4 bytes
Traffic Conditions	4 bytes
Seating	4 bytes
Foot Traffic Score	4 bytes
Indoor Noise	4 bytes
Outdoor Noise	4 bytes

Total space occupied by a single row:
4 (row header) + 20*4 = 84 bytes.

Total number of rows:
Number of stores * number of orders per day * average number of items per order * number of days in a year = 1000 * 1500 * 2 * 365 = 1,095,000,000

Total table size:
84 * 1095000000 rows = 91.98 GB ~= 92 GB

- If we choose 1st approach mentioned in the new schema design and add the IOT data along with transaction data in the separate fact table with granularity at order level instead of order-line level, then

size per row for IOT table:
4 bytes (row header) + 19 * 4 bytes = 80 bytes
total table size of IOT table:
80 bytes * 547,500,000 rows = 43.8 GB ~= 44 GB

- If we choose 3rd approach mentioned in the new schema design and add only IOT data in separate fact table with granularity at order level, then

size per row for current fact table:
4 bytes (row header) + 7 * 4 bytes = 32 bytes
total table size of current fact table:
32 bytes * 1,095,000,000 rows = 35.04 GB ~= 35 GB

size per row for IOT fact table:
4 bytes (row header) + 14 * 4 bytes = 60 bytes
total table size of IOT table:
60 bytes * 547,500,000 rows = 32.85 GB ~= 33 GB

So, there is a tradeoff between storage cost, performance, and the report availability.

If we go with 1st approach, we will save half of the storage per year as compared to 2nd approach but as explained earlier we won't be able to analyze sales performance with respect to weather data across product dimension.

If we go with 3rd approach, we will save 30% of the storage per year as compared to 2nd approach and we will be able to analyze the sales performance with respect to weather data across all dimensions, but as explained earlier we will have to make lot of joins on pretty big tables which is going to hamper the performance and also have to join 2 fact tables which is not a standard design and will cause more performance issues.

Since, in cloud Datawarehouse storage is cheaper as compared to heavy computation we should go ahead with the 2nd approach and add all the IOT sensor metrics to the existing fact table.

Estimate of time dimension table size

Attribute	**Space Occupied**
TimeID	4 bytes
Hour	4 bytes
Minute	4 bytes

Total space occupied by a single row:
4 (row header) + 3*4 bytes = 16 bytes.

Total number of rows:
Number of seconds in a day = 60 * 24 = 1440

Total table size:
16 * 1440 rows = 23.04 KB ~= 23 KB

Reports based on new schema:

a. How do sales vary across stores with varying cloud cover during the hours of 8:00am – 4:00pm and Months of May – Sep.

Column: StoreID from Store dimension table

Row: Cloud Cover value from Fact table

Filter: Month level filter on DateKey from Date dimension table & hour level filter on TimeKey from Time dimension table

Cell Values: sum of Sales_Amount from Fact table

b. How do sales vary across products with varying temperatures during the hours of 12:00pm – 5:00pm during the months of March – Sep.

Column: ProductID from Product dimension table

Row: Outdoor_Temperature value from Fact table

Filter: Month level filter on DateKey from Date dimension table & hour level filter on TimeKey from Time dimension table

Cell Values: sum of Sales_Amount from Fact table

Estimates for size of raw IoT data for 1 year

Attribute	**Space Occupied**
Queue_Length	4 bytes
Indoor_Temperature	4 bytes
Outdoor_Temperature	4 bytes
DimTimeID	4 bytes
Humidity	4 bytes
Cloud Cover	4 bytes
Weather condition	4 bytes
Pollen	4 bytes
Traffic Conditions	4 bytes
Seating	4 bytes
Foot Traffic Score	4 bytes
Indoor Noise	4 bytes
Outdoor Noise	4 bytes

Total space occupied by a single row:
4 (row header) + 13*4 = 56 bytes.

Total number of rows:
Number of stores * Number of seconds in a day * Number of days in a year = 1000 * 60 * 60 * 24 * 365 = 31,536,000,000

Total table size:
56 * 31536000000 rows = 1.76 TB ~= 2 TB

Relational database options:

AWS:

RDBMS Type:
MySQL db.t4g.medium instance deployed in multiple availability zones with only 1 standby instance

Price:
$0.129 / Hour for 2VCpu and 4 GB
$0.23 / month / GB for General Purpose SSD Storage

Total Price:

(0.129 * 24 * 365) + (0.23 * 2000 * 12) = 1130.04 + 5520 = $6650.04

GCP:

RDBMS Type:
MySQL

Price:
$45.2235 / CPU / Month
$7.665 / GB / Month for Memory
$0.34 / GB /month for SSD storage

(45.22 * 2 * 12) + (7.66 * 4 * 12) + (0.34 * 2000 * 12) = 1085.28 + 367.68 + 8160 = $9612.96

Total Price:

Object Storage Options:

AWS S3:

S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data

First 50 TB / Month	$0.023 per GB

Total cost for IOT raw data in AWS S3 for 1 year:

$0.023 * 2000 * 12 = $552

GCP Buckets:

Location	Standard storage (per GB per Month)	Nearline storage (per GB per Month)	Coldline storage (per GB per Month)	Archive storage (per GB per Month)
Iowa (`us-central1`)	$0.020	$0.010	$0.004	$0.0012
South Carolina (`us-east1`)	$0.020	$0.010	$0.004	$0.0012
Northern Virginia (`us-east4`)	$0.023	$0.013	$0.006	$0.0025
Columbus (`us-east5`)	$0.020	$0.010	$0.004	$0.0012
Oregon (`us-west1`)	$0.020	$0.010	$0.004	$0.0012

Total cost for IOT raw data in GCP Bucket for 1 year:

$0.022 * 2000 * 12 = $528

Recommendation:

- We are already putting required IOT sensor data in our DW.
- Now since we are dealing with all the raw sensor data, we can put it in a file and then store it in an object storage instead of relational database.

So, after comparing all the above storage options, we would recommend going with GCP Object Storage to store all the IOT sensor raw data.

Cost estimates for AWS Redshift cluster to support the DW

Fact table size: 92 GB
Time dimension table size: 23 KB
Size of all existing dimension tables: 1 GB

Total storage requirement for the Datawarehouse:
92 + 1 + 0.000023 = 93.000023 GB

Redshift pricing:

Current generation

Region: US West (Oregon)

	vCPU	Memory	Addressable storage capacity	I/O	Price
Dense Compute DC2					
dc2.large	2	15 GiB	0.16TB SSD	0.60 GB/s	$0.25 per Hour
dc2.8xlarge	32	244 GiB	2.56TB SSD	7.50 GB/s	$4.80 per Hour
RA3 with Redshift Managed Storage*					
ra3.xlplus	4	32 GiB	32TB RMS	0.65 GB/s	$1.086 per Hour
ra3.4xlarge	12	96 GiB	128TB RMS	2.00 GB/s	$3.26 per Hour
ra3.16xlarge	48	384 GiB	128TB RMS	8.00 GB/s	$13.04 per Hour

We will choose ra3.xlplus instance with Redshift Managed Storage since it has more than required storage and processing power for our DW.

Total Redshift cluster Cost for 1 year:

$1.086 * 24 * 365 = $ 9513.36