**1.-- Create a Hive table for the Sales Rep data:**

**create table SalesRep (**

**RepID INT,**

**RepName STRING,**

**Territory INT**

**)**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**LINES TERMINATED BY '\n';**

**-- and load the CSV into the Hive Sales Rep table:**

**LOAD DATA INPATH '/user/hue/hiveinput/reps.csv'**

**INTO TABLE SalesRep;**

**2. -- Create a Hive table for the PO's:**

**create table purchases (**

**SalesRepId INT,**

**PurchaseOrderId INT,**

**Amount INT**

**)**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**LINES TERMINATED BY '\n';**

**-- and load CSV into the Hive PO table:**

**LOAD DATA**

**LOCAL INPATH '/user/hue/hiveinput/purchases.csv'**

**INTO TABLE purchases;**

**3. Hive JOIN**  
  
So this is the underlining data that is being worked with:  
  
SELECT p.PurchaseOrderId, s.RepName, p.amount, s.Territory  
FROM purchases p JOIN SalesRep s  
WHERE p.SalesRepId = s.RepID;

|  |  |  |  |
| --- | --- | --- | --- |
| **PO ID** | **Rep** | **Amount** | **Territory** |
| 1 | Jana | 100 | 1 |
| 2 | Nadia | 200 | 1 |
| 3 | Nadia | 600 | 1 |
| 4 | Daniel | 80 | 2 |
| 5 | Jana | 120 | 1 |
| 6 | William | 170 | 2 |
| 7 | Daniel | 140 | 2 |

**4. Hive Rank by Volume only**  
  
SELECT   
  s.RepName, s.Territory, V.volume,   
**rank() over (ORDER BY V.volume DESC)** as rank  
FROM   
  SalesRep s  
  JOIN   
    ( SELECT  
      SalesRepId, SUM(amount) as Volume  
      FROM purchases  
      GROUP BY SalesRepId) V  
  WHERE V.SalesRepId = s.RepID  
  ORDER BY V.volume DESC;

|  |  |  |  |
| --- | --- | --- | --- |
| **Rep** | **Territory** | **Amount** | **Rank** |
| Nadia | 1 | 800 | 1 |
| Daniel | 2 | 220 | 2 |
| Jana | 1 | 220 | 2 |
| William | 2 | 170 | 4 |

The ranking over the entire data set - Daniel is tied for second among all Reps.

SELECT   
  s.RepName, s.Territory, V.volume,   
**dense\_rank() over (ORDER BY V.volume DESC)** as rank  
FROM   
  SalesRep s  
  JOIN   
    ( SELECT  
      SalesRepId, SUM(amount) as Volume  
      FROM purchases  
      GROUP BY SalesRepId) V  
  WHERE V.SalesRepId = s.RepID  
  ORDER BY V.volume DESC;

|  |  |  |  |
| --- | --- | --- | --- |
| **Rep** | **Territory** | **Amount** | **Rank** |
| Nadia | 1 | 800 | 1 |
| Daniel | 2 | 220 | 2 |
| Jana | 1 | 220 | 2 |
| William | 2 | 170 | 3 |

SELECT   
  s.RepName, s.Territory, V.volume,   
**row\_number() over (ORDER BY V.volume DESC)** as rank  
FROM   
  SalesRep s  
  JOIN   
    ( SELECT  
      SalesRepId, SUM(amount) as Volume  
      FROM purchases  
      GROUP BY SalesRepId) V  
  WHERE V.SalesRepId = s.RepID  
  ORDER BY V.volume DESC;

|  |  |  |  |
| --- | --- | --- | --- |
| **Rep** | **Territory** | **Amount** | **Rank** |
| Nadia | 1 | 800 | 1 |
| Daniel | 2 | 220 | 2 |
| Jana | 1 | 220 | 3 |
| William | 2 | 170 | 4 |

**5. Hive Rank within Territory, by Volume**  
  
SELECT   
  s.RepName, s.Territory, V.volume,   
  rank() over (**PARTITION BY s.Territory**ORDER BY V.volume DESC) as rank  
FROM   
  SalesRep s  
  JOIN   
    ( SELECT  
      SalesRepId, SUM(amount) as Volume  
      FROM purchases  
      GROUP BY SalesRepId) V  
  WHERE V.SalesRepId = s.RepID  
  ORDER BY V.volume DESC;

|  |  |  |  |
| --- | --- | --- | --- |
| **Rep** | **Territory** | **Amount** | **Rank** |
| Nadia | 1 | 800 | 1 |
| Jana | 1 | 220 | 2 |
| Daniel | 2 | 220 | 1 |
| William | 2 | 170 | 2 |

The ranking is within the territory - Daniel is the best is his territory.