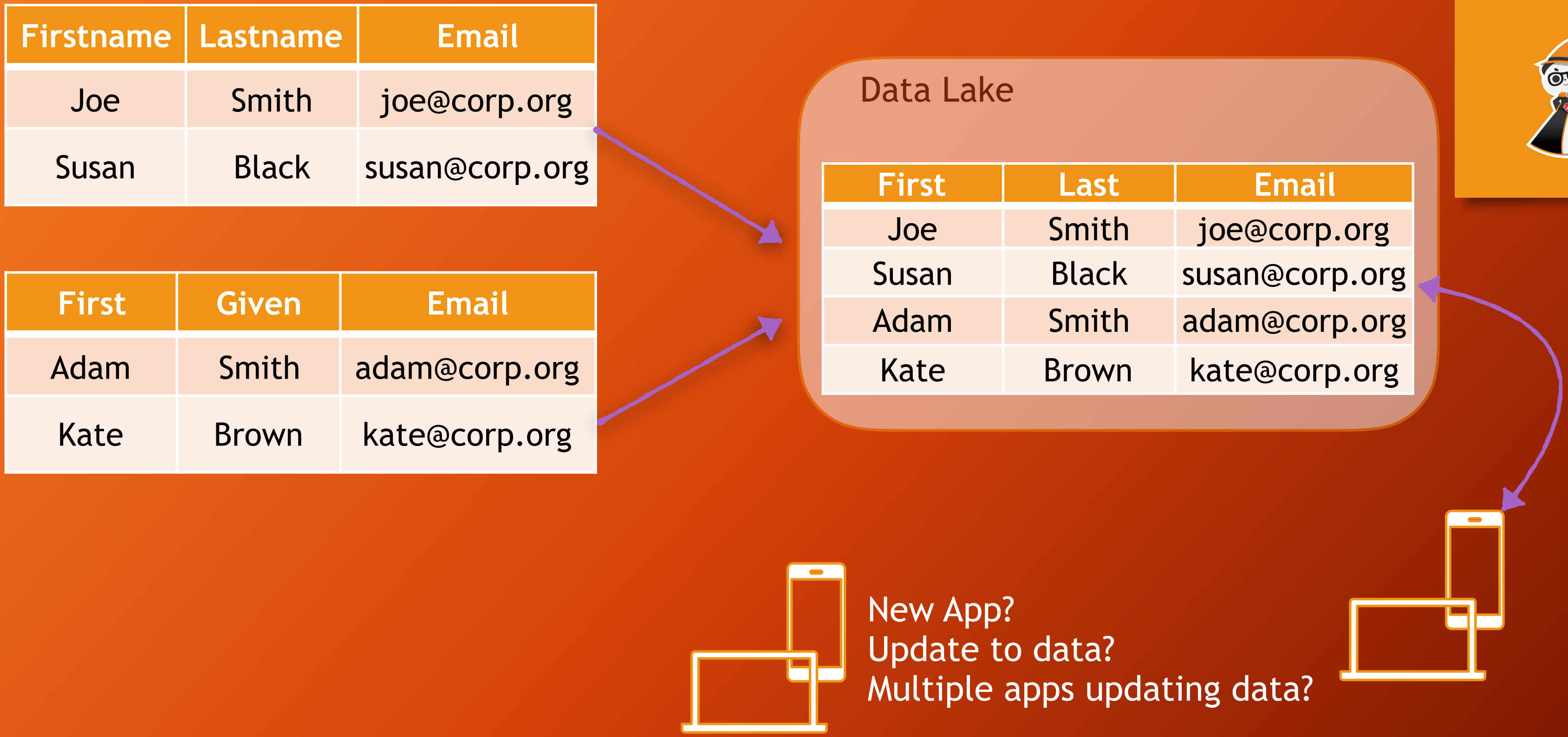
Data Warehouse

Data warehouse

- System used for reporting and data analysis
  - Data mining, analytical processing, market research, decision support
- Typically used as ETL
  - Extract
  - Transform
  - Load

Data marts

- Single focused
    - Collects specific data from certain systems
    - Usually used for a specific purpose (for a department)



Firstname	Lastname	Email
Joe	Smith	joe@corp.org
Susan	Black	susan@corp.org

First	Given	Email
Adam	Smith	adam@corp.org
Kate	Brown	kate@corp.org

Data Lake

First	Last	Email
Joe	Smith	joe@corp.org
Susan	Black	susan@corp.org
Adam	Smith	adam@corp.org
Kate	Brown	kate@corp.org

New App?
Update to data?
Multiple apps updating data?

Star schema

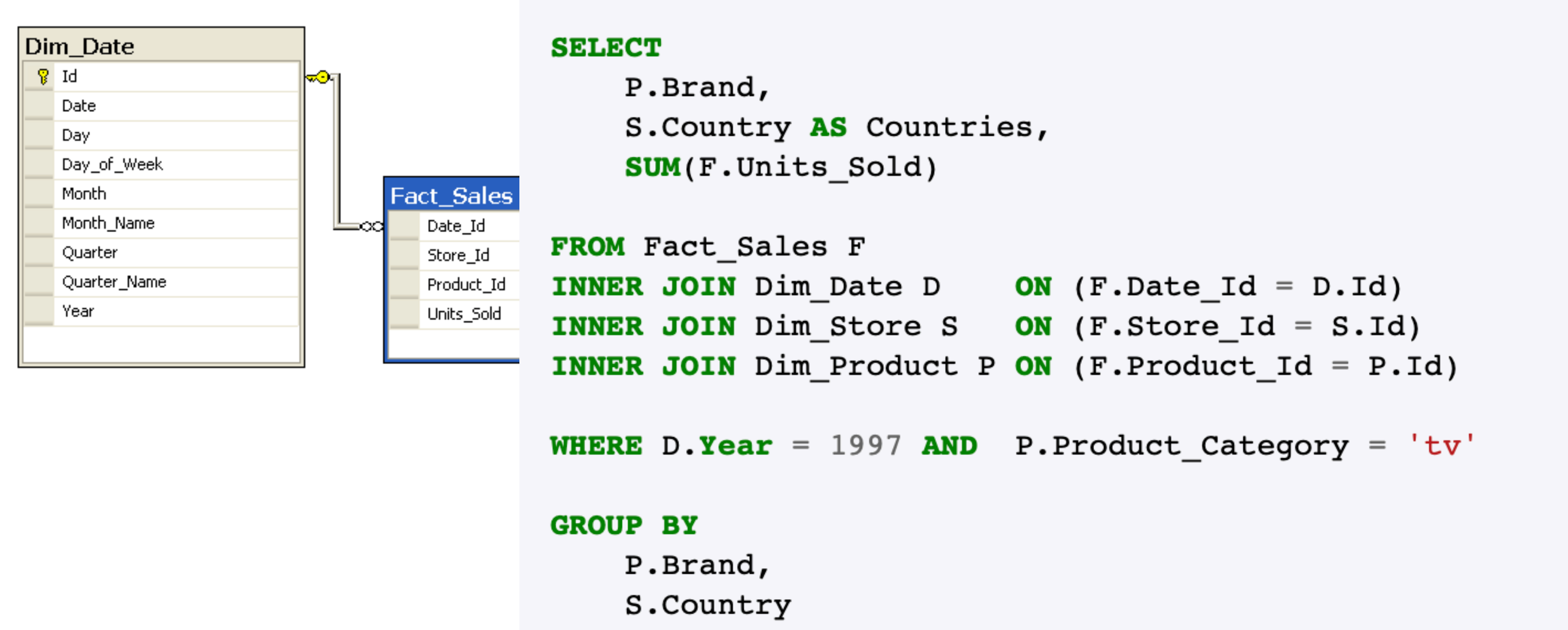
- Fact tables - dimension tables
  - Fact table: contain metrics, reference dimensional tables
    - Entries usually identified by a surrogate key (not derived from application data)
  - Dimension table: large set of attributes
    - Usually less data then fact tables

Star Schema: pros / cons

- Advantages
  - Denormalised data
  - Simpler queries
  - Simple business logic
  - Query performance & fast aggregations

- Disadvantages
  - Difficult to keep track of data integrity
  - Purpose built, less for complex analytics

Star Schema



Dim_Date

- Id
- Date
- Day
- Day_of_Week
- Month
- Month_Name
- Quarter
- Quarter_Name
- Year

Fact_Sales

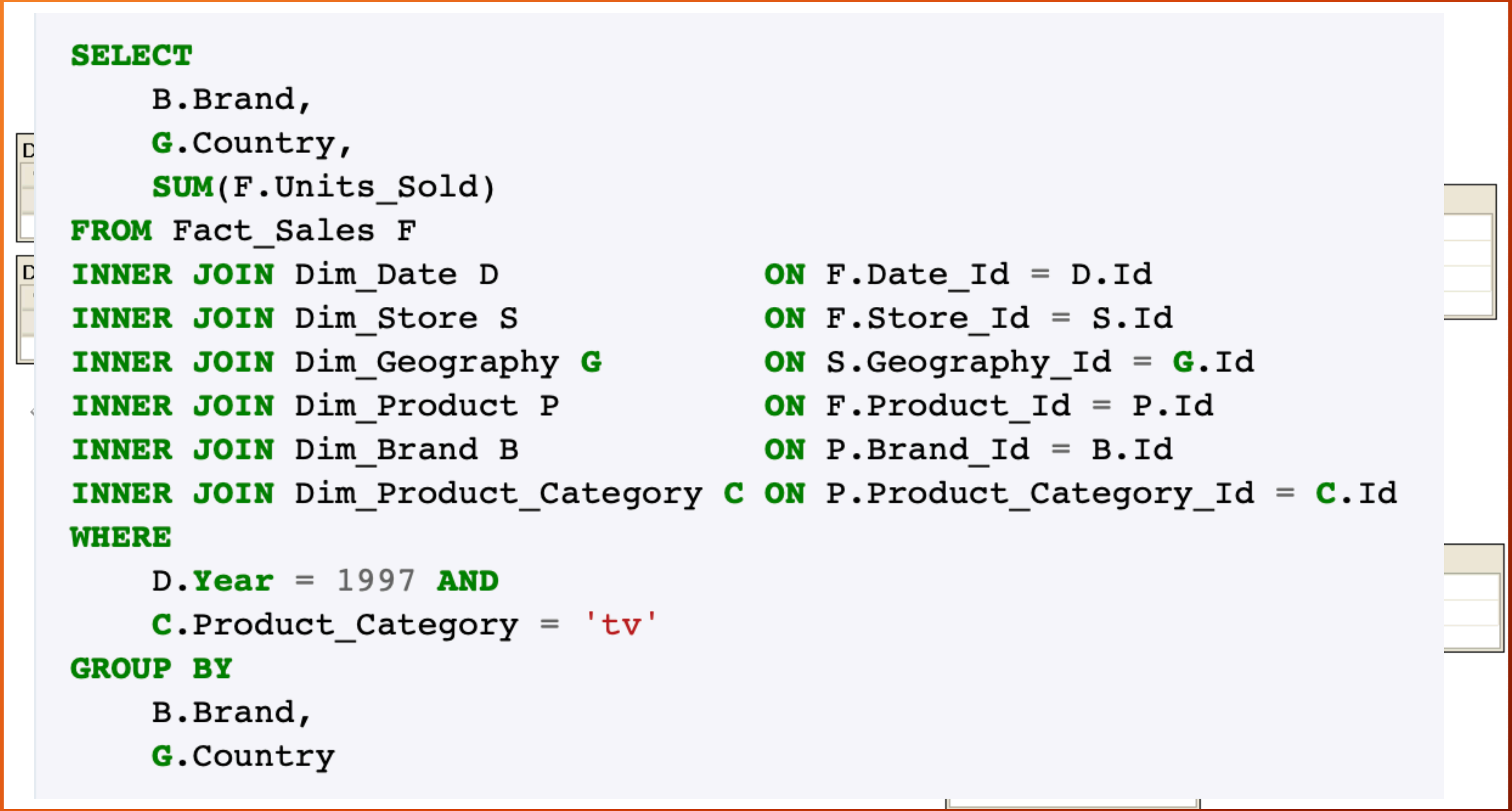
- Date_Id
- Store_Id
- Product_Id
- Units_Sold

```sql
SELECT
    P.Brand,
    S.Country AS Countries,
    SUM(F.Units_Sold)

FROM Fact_Sales F
INNER JOIN Dim_Date D    ON (F.Date_Id = D.Id)
INNER JOIN Dim_Store S   ON (F.Store_Id = S.Id)
INNER JOIN Dim_Product P ON (F.Product_Id = P.Id)

WHERE D.Year = 1997 AND  P.Product_Category = 'tv'

GROUP BY
    P.Brand,
    S.Country
```

*https://en.wikipedia.org/wiki/Star_schema#/media/File:Приклад_схеми_зірки.png

Snowflake

- “Snowflaking” is a method to normalise dimension tables
- “special” star schema
- However, complex joins



```sql
SELECT
    B.Brand,
    G.Country,
    SUM(F.Units_Sold)
FROM Fact_Sales F
INNER JOIN Dim_Date D                ON F.Date_Id = D.Id
INNER JOIN Dim_Store S               ON F.Store_Id = S.Id
INNER JOIN Dim_Geography G           ON S.Geography_Id = G.Id
INNER JOIN Dim_Product P             ON F.Product_Id = P.Id
INNER JOIN Dim_Brand B               ON P.Brand_Id = B.Id
INNER JOIN Dim_Product_Category C ON P.Product_Category_Id = C.Id
WHERE
    D.Year = 1997 AND
    C.Product_Category = 'tv'
GROUP BY
    B.Brand,
    G.Country
```

OLAP / OLTP

- Online Analytical Processing
  - Low volume transactions
  - Complex queries (usually with aggregations)
- Online Transaction Processing
  - Large number of short transactions