ARTIFICIAL INTELLIGENCE SOFTWARE DEVELOPMENT

Week 11 Lecture 1

Dr. Hari M Koduvely

Agenda for Today

- ❑ Theory:
  - ▪ Fundamentals of Data Engineering – Part 2

Database Normalization

❑ Normalization is a Database Design Technique

❑ Reduces Data Redundancy

❑ Eliminates Insertion, Update and Deletion anomalies

❑ Divides larger tables into smaller ones linked by relationships

❑ Ensure that data is stored logically

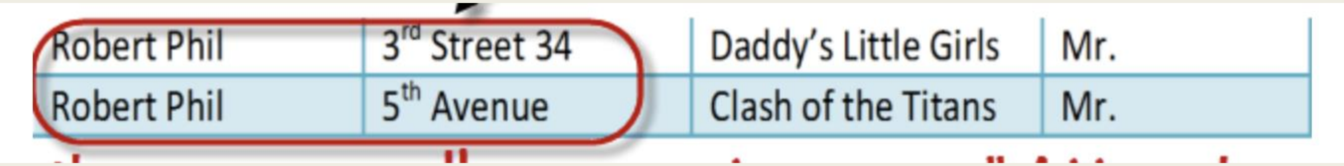
Database Normal Forms

- ❏ 1NF (First Normal Form)
- ❏ 2NF (Second Normal Form)
- ❏ 3NF (Third Normal Form)
- ❏ BCNF (Boyce-Codd Normal Form)
- ❏ 4NF (Fourth Normal Form)
- ❏ 5NF (Fifth Normal Form)
- ❏ 6NF (Sixth Normal Form)

In most practical applications, 3NF is sufficient

Database Normal Forms

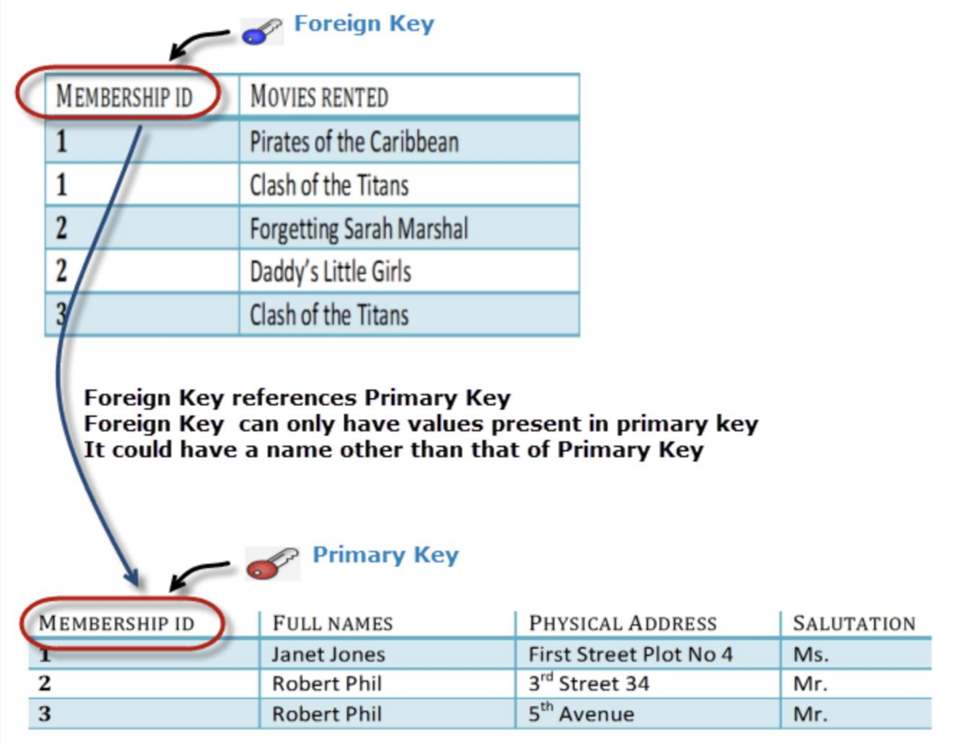
- ❑ A KEY is used to identify records in a database uniquely
- ❑ A Primary KEY is a single column value used to identify a database record uniquely
    - *A primary key cannot be NULL*
    - *A primary key value must be unique*
    - *The primary key values should rarely be changed*
    - *The primary key must be given a value when a new record is inserted*
- ❑ A Composite KEY is a primary key composed of multiple columns.



Robert Phil	3rd Street 34	Daddy's Little Girls	Mr.
Robert Phil	5th Avenue	Clash of the Titans	Mr.

Database Normal Forms

- ❑ Foreign Key references the primary key of another Table
- ❑ It helps connect the two Tables
- ❑ A foreign key can have a different name from its primary key
- ❑ It ensures rows in one table have corresponding rows in another
- ❑ Unlike the Primary key, most often they are not unique
- ❑ Foreign keys can be null even though primary keys can not



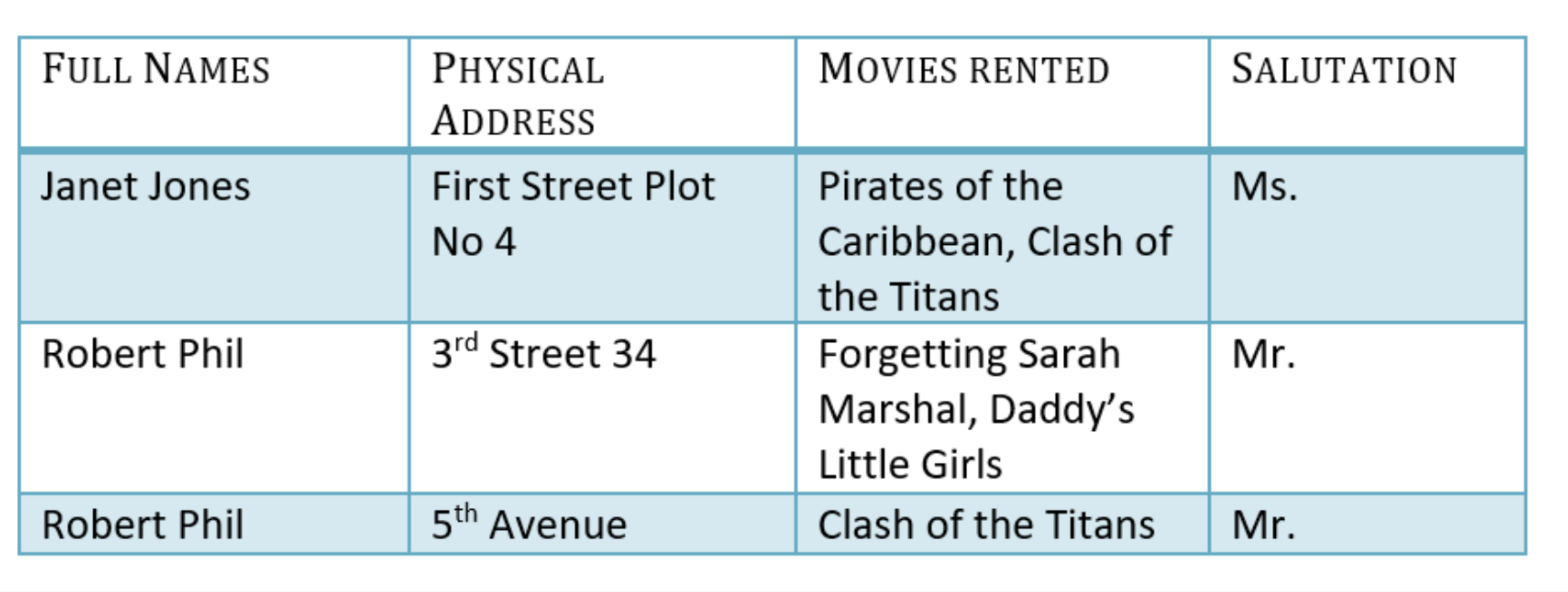
MEMBERSHIP ID	MOVIES RENTED
1	Pirates of the Caribbean
1	Clash of the Titans
2	Forgetting Sarah Marshal
2	Daddy's Little Girls
3	Clash of the Titans

Foreign Key references Primary Key
Foreign Key can only have values present in primary key
It could have a name other than that of Primary Key

MEMBERSHIP ID	FULL NAMES	PHYSICAL ADDRESS	SALUTATION
1	Janet Jones	First Street Plot No 4	Ms.
2	Robert Phil	3rd Street 34	Mr.
3	Robert Phil	5th Avenue	Mr.

Database Normal Forms Example

Movie Rental Database



Full Names	Physical Address	Movies rented	Salutation
Janet Jones	First Street Plot No 4	Pirates of the Caribbean, Clash of the Titans	Ms.
Robert Phil	3rd Street 34	Forgetting Sarah Marshal, Daddy’s Little Girls	Mr.
Robert Phil	5th Avenue	Clash of the Titans	Mr.

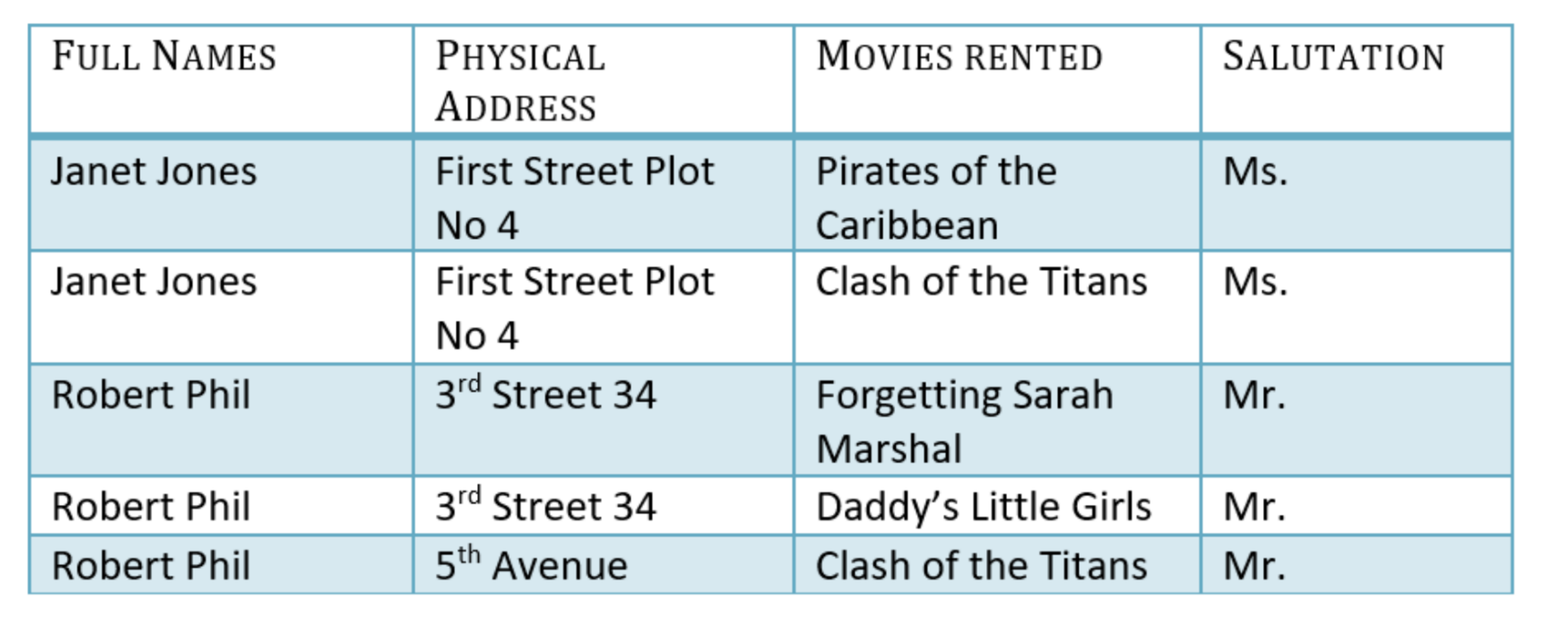
Database Normal Forms

1st Normal Form Rules

❑ Each table cell should contain a single value

❑ Each record need to be unique

❑ Each column name should be unique



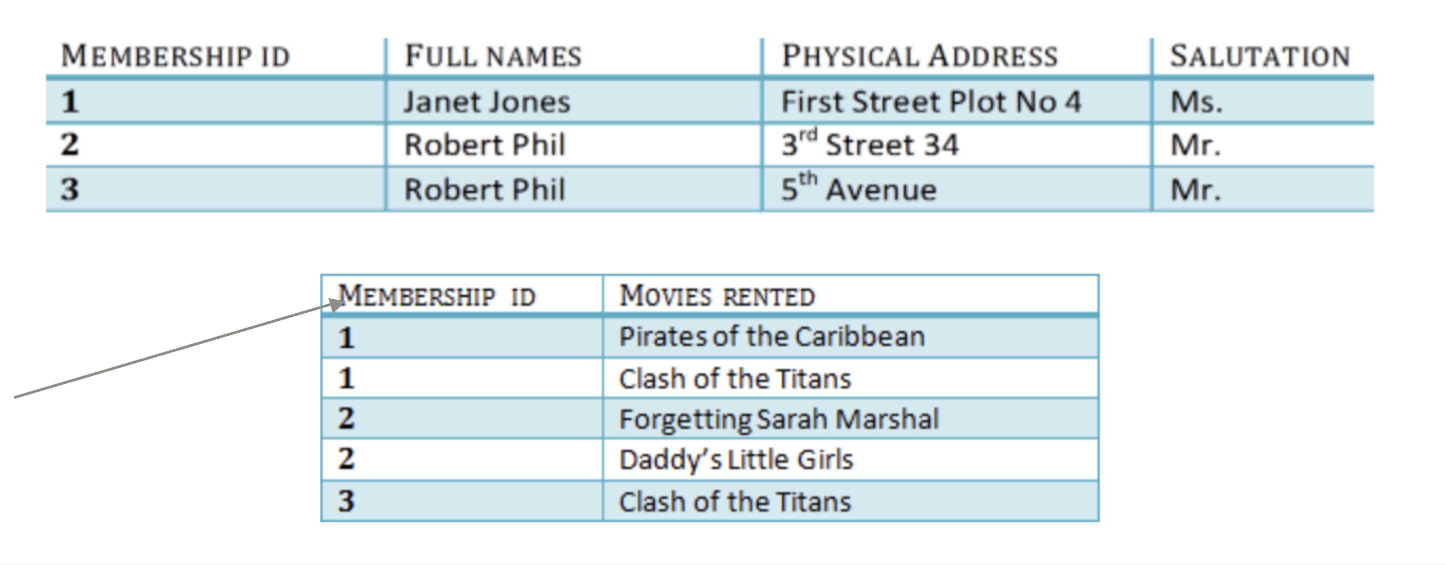
FULL NAMES	PHYSICAL ADDRESS	MOVIES RENTED	SALUTATION
Janet Jones	First Street Plot No 4	Pirates of the Caribbean	Ms.
Janet Jones	First Street Plot No 4	Clash of the Titans	Ms.
Robert Phil	3rd Street 34	Forgetting Sarah Marshal	Mr.
Robert Phil	3rd Street 34	Daddy's Little Girls	Mr.
Robert Phil	5th Avenue	Clash of the Titans	Mr.

Database Normal Forms

2nd Normal Form Rules

- □ Be 1NF
- □ Single Column Primary Key

Primary Key



MEMBERSHIP ID	FULL NAMES	PHYSICAL ADDRESS	SALUTATION
1	Janet Jones	First Street Plot No 4	Ms.
2	Robert Phil	3rd Street 34	Mr.
3	Robert Phil	5th Avenue	Mr.

MEMBERSHIP ID	MOVIES RENTED
1	Pirates of the Caribbean
1	Clash of the Titans
2	Forgetting Sarah Marshal
2	Daddy's Little Girls
3	Clash of the Titans

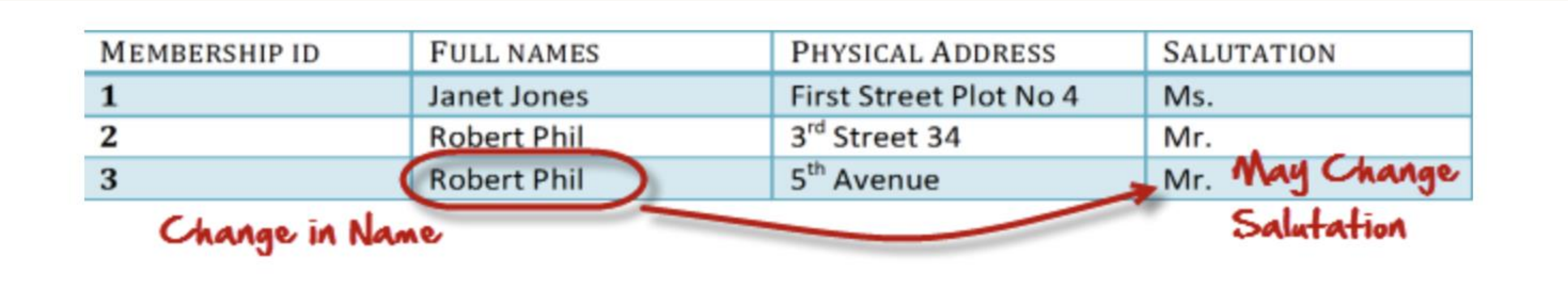
Foreign Key

Database Normal Forms

3rd Normal Form Rules

- Be 2NF
- No transactive functional dependence

- Transactive dependence is when changing a non-key column, might cause any of the other non-key columns to change



Membership ID	Full names	Physical Address	Salutation
1	Janet Jones	First Street Plot No 4	Ms.
2	Robert Phil	3rd Street 34	Mr.
3	Robert Phil	5th Avenue	Mr.

Change in Name
May Change Salutation

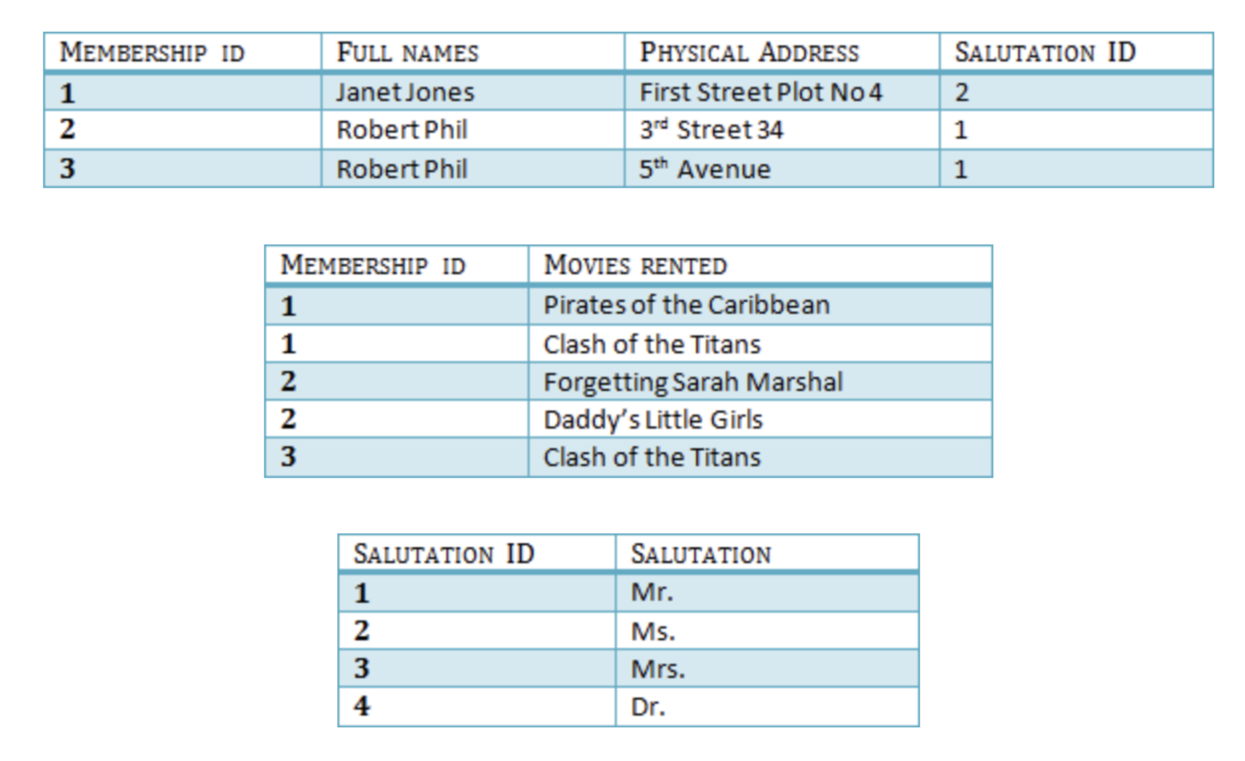
Image source https://www.guru99.com/database-normalization.html

Database Normal Forms

3rd Normal Form Rules

❑ Be 2NF

❑ No transactive functional dependence



Membership ID	Full names	Physical Address	Salutation ID
1	Janet Jones	First Street Plot No 4	2
2	Robert Phil	3rd Street 34	1
3	Robert Phil	5th Avenue	1

Membership ID	Movies rented
1	Pirates of the Caribbean
1	Clash of the Titans
2	Forgetting Sarah Marshal
2	Daddy's Little Girls
3	Clash of the Titans

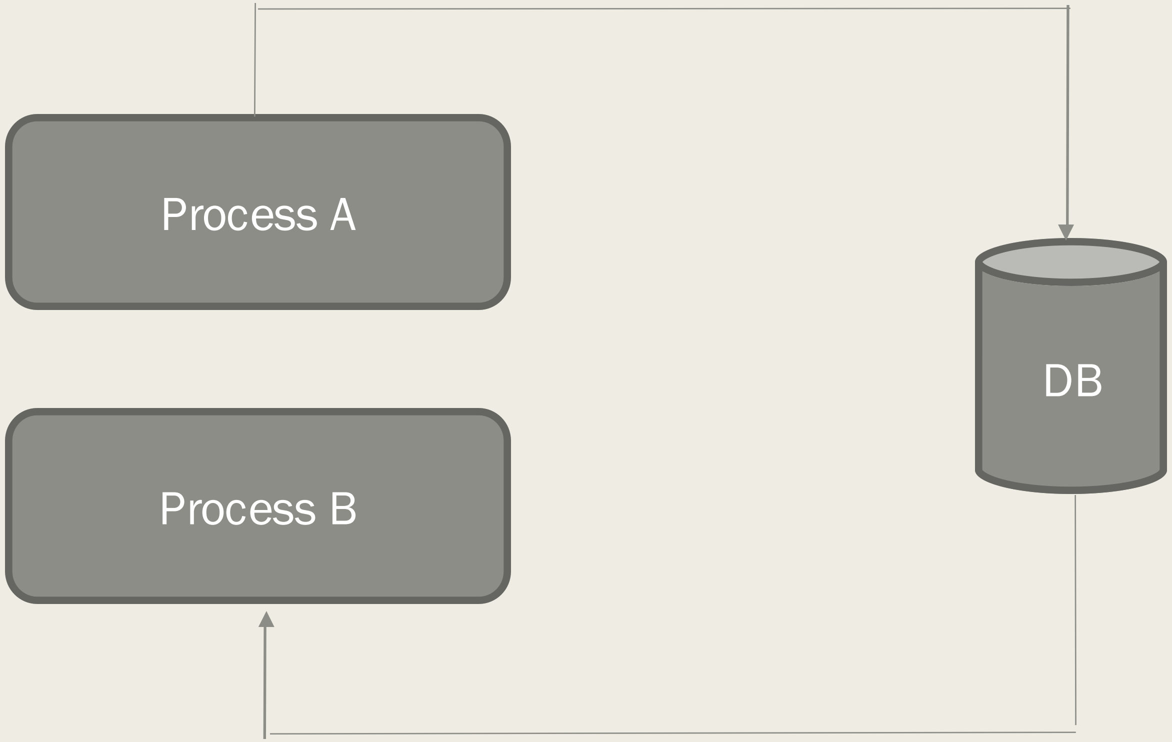
Salutation ID	Salutation
1	Mr.
2	Ms.
3	Mrs.
4	Dr.

Modes of Data Flow

- Typical production scenario:
  - *Multiple processes running simultaneously*
  - *Without sharing memory between them*
- How do we pass data between these processes?
- Data passing from one process to another is called **Data Flow**

Modes of Data Flow

Data Passing through Databases

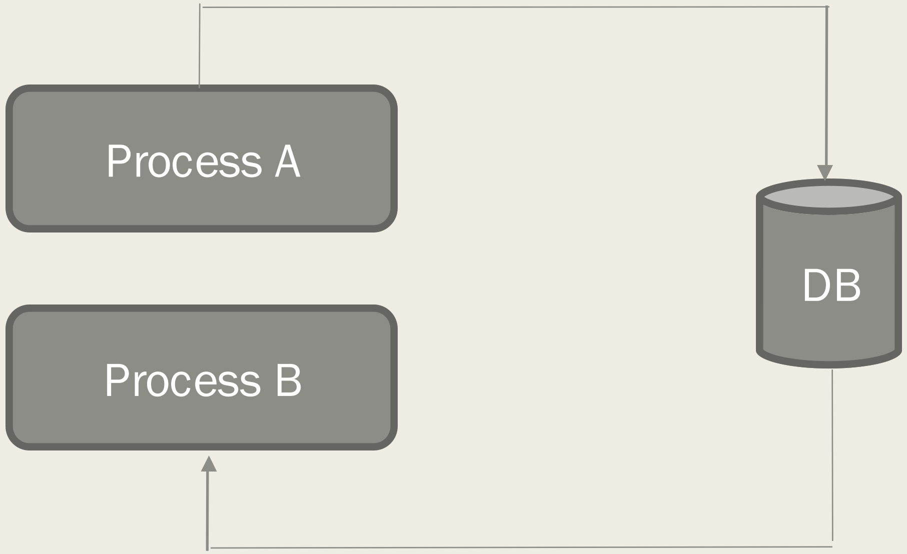


Modes of Data Flow

Data Passing through Databases

❑ Access issues
  - A and B can be part of different accounts

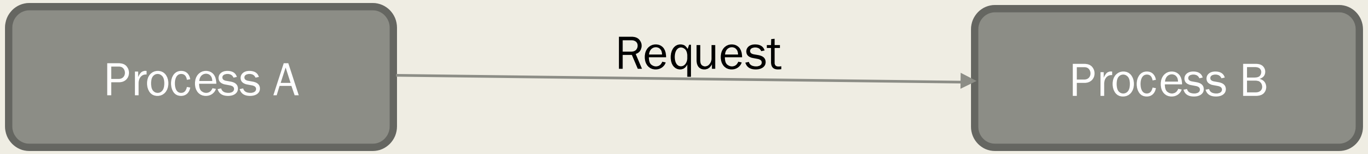
❑ Latency issues
  - Read and write on DB can be slow



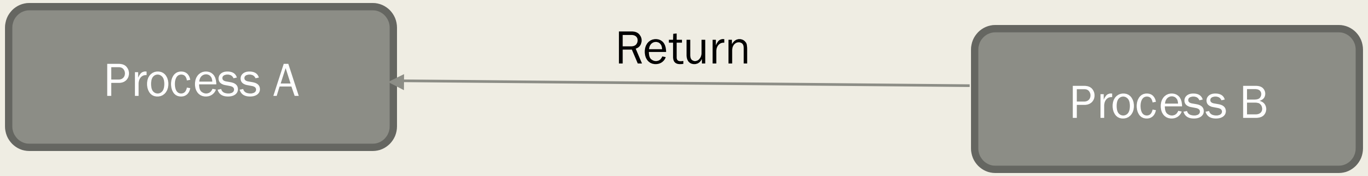
Modes of Data Flow

Data Passing through Services

❑ Process A send request to Process B for a particular data



❑ Process B returns the requested data through the same network



Modes of Data Flow

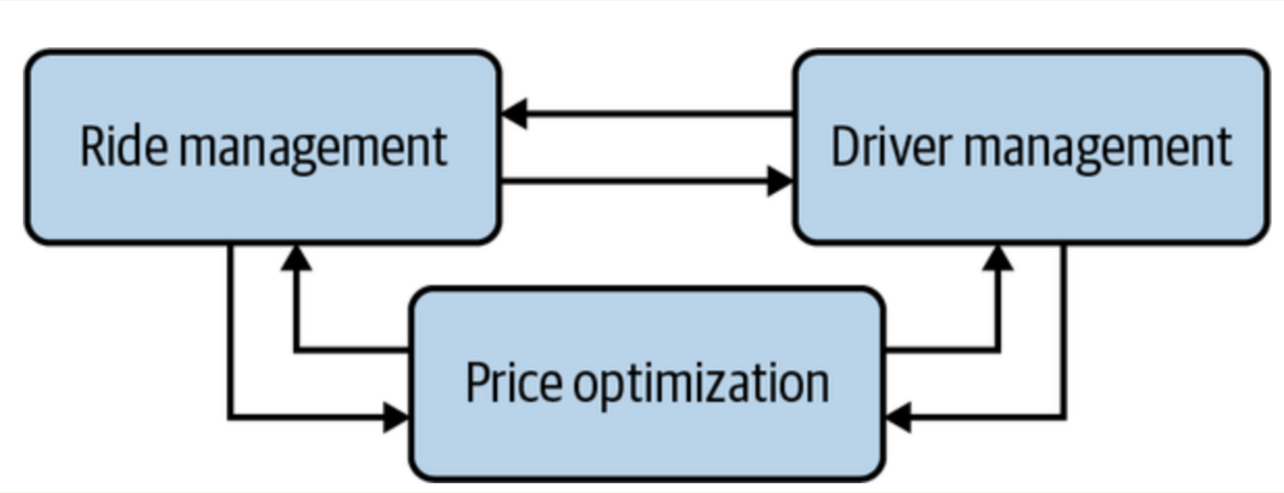
Data Passing through Services

- ❑ Two popular styles of passing data are
  - REST (Representational State Transfer)
    - Used for data request over a network
  - RPC (Remote Procedure Call)
    - Used for data request within a data center

Modes of Data Flow

Data Passing through Realtime Transport

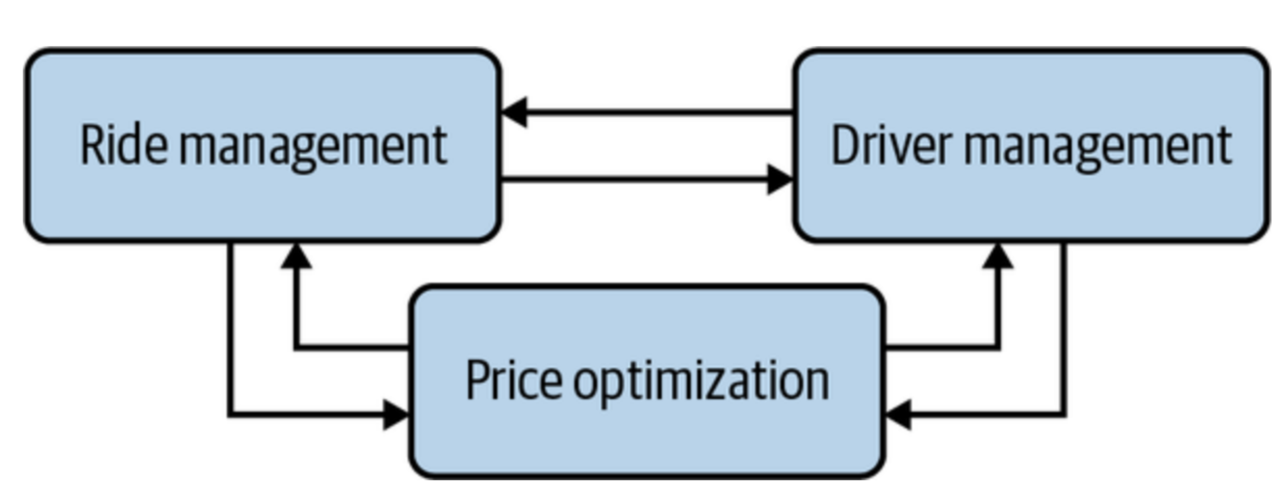
- ❑ Example scenario: Ride Sharing App
  - Ride management service
  - Driver management service
  - Price optimization service



Modes of Data Flow

Data Passing through Realtime Transport

❑ Example scenario: Ride Sharing App
  - Request driven data passing is synchronous.
  - A service that is down can cause all services that require data from it to be down.

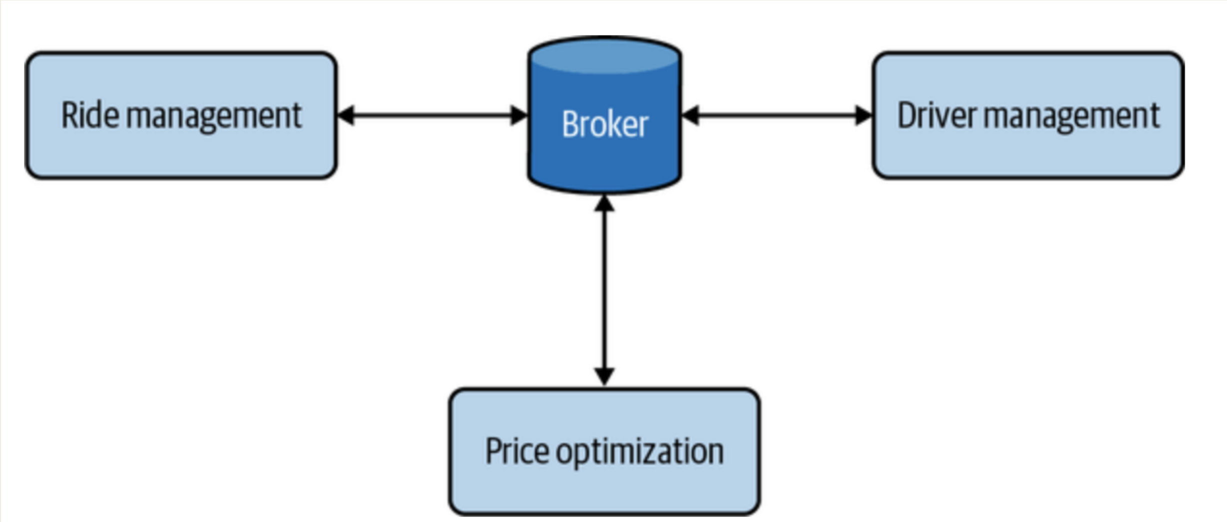


Modes of Data Flow

Data Passing through Realtime Transport

❑ Solution: A Broker that can co-ordinate data passing between services

- Each service only has to communicate with the broker
- Each service broadcast the data to broker as **events**



Modes of Data Flow

Data Passing through Realtime Transport

❑ Two models of Realtime Transport

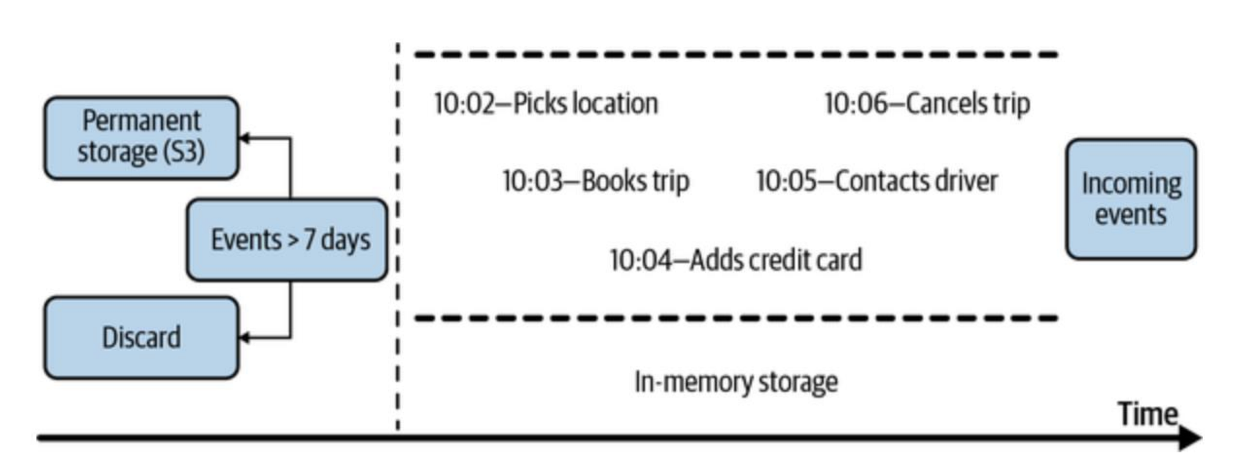
  - Publish-Subscribe (PubSub)
  - Message Queue

Modes of Data Flow

Data Passing through Realtime Transport

❑ PubSub Model

- Events are arranged into **Topics**
- A service can publish events to any number of topics
- A service that subscribe to a Topic can read all events in that topic
- The service publishing data is not concerned about who is subscribing
- Data is retained only for a finite interval of time



Modes of Data Flow

Data Passing through Realtime Transport

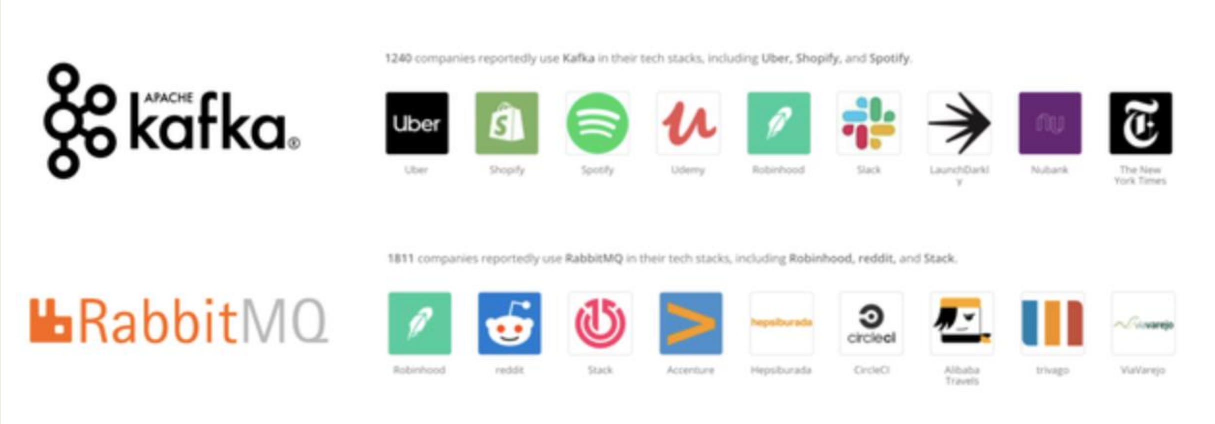
❑ Message Queue Model

  - Each event has an intended set of consumers (message).
  - message queue is responsible for getting the message to the right consumers.

Modes of Data Flow

Data Passing through Realtime Transport

- ❑ Examples of PubSub based services
  - Apache Kafka
  - Amazon Kinesis
- ❑ Examples of Message Que based services
  - Apache RocketMQ
  - RabbitMQ



1240 companies reportedly use **Kafka** in their tech stacks, including **Uber**, **Shopify**, and **Spotify**.

1811 companies reportedly use **RabbitMQ** in their tech stacks, including **Robinhood**, **reddit**, and **Stack**.

Batch Processing vs Stream Processing

❑ Historical Data are stored in:
- Databases
- Data lakes
- Data warehouses

❑ They are often processed in batches

❑ Using distributed computing frameworks like Hadoop or Spark

❑ Difference between Hadoop and Spark ?

Batch Processing vs Stream Processing

- ❑ Data are stored Realtime Transport are called **Streaming Data**

- ❑ Computations done on Streaming Data are called **Stream Processing**

- ❑ In ML Batch Processing is used to compute Static Features
  - ▪ E. g. Drivers ratings

- ❑ Stream Processing is used to compute Dynamic Features
  - ▪ E. g. How many drivers are available currently

Batch Processing vs Stream Processing

❑ In ML Batch Processing is used to compute Static Features

  ▪ E. g. Drivers ratings

❑ Stream Processing is used to compute Dynamic Features

  ▪ E. g. How many drivers are available currently

Example – Machine Learning with Kafka

Robust machine learning on streaming data using Kafka and Tensorflow-IO

https://www.tensorflow.org/io/tutorials/kafka

Google Colab Notebook