# Airbnb

## Big Data Implementation

**BS Data Science**
**PUCIT**
**Big Data Analytics**

BSDSF21A037: Duaa Mansur

# Airbnb
# Big Data Analytics Project

- Purpose:
- Analyze real-time Airbnb booking, cancellation, and viewing actions using big data tools.

# Tools and Goals:

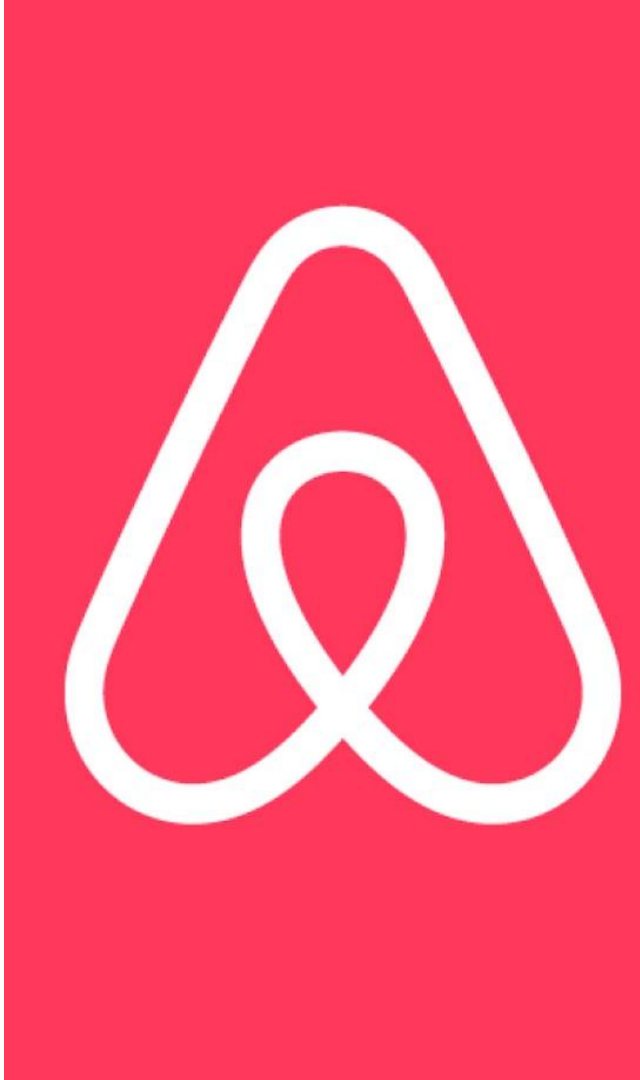
- Tools: Kafka, Spark, Cassandra, Streamlit.
- Provide live insights for better decision-making.

# Key Design Challenges

- • Scalability for high data volumes.
- • Real-time data streaming and processing.
- • User-friendly dashboard design.

# Architecture Overview

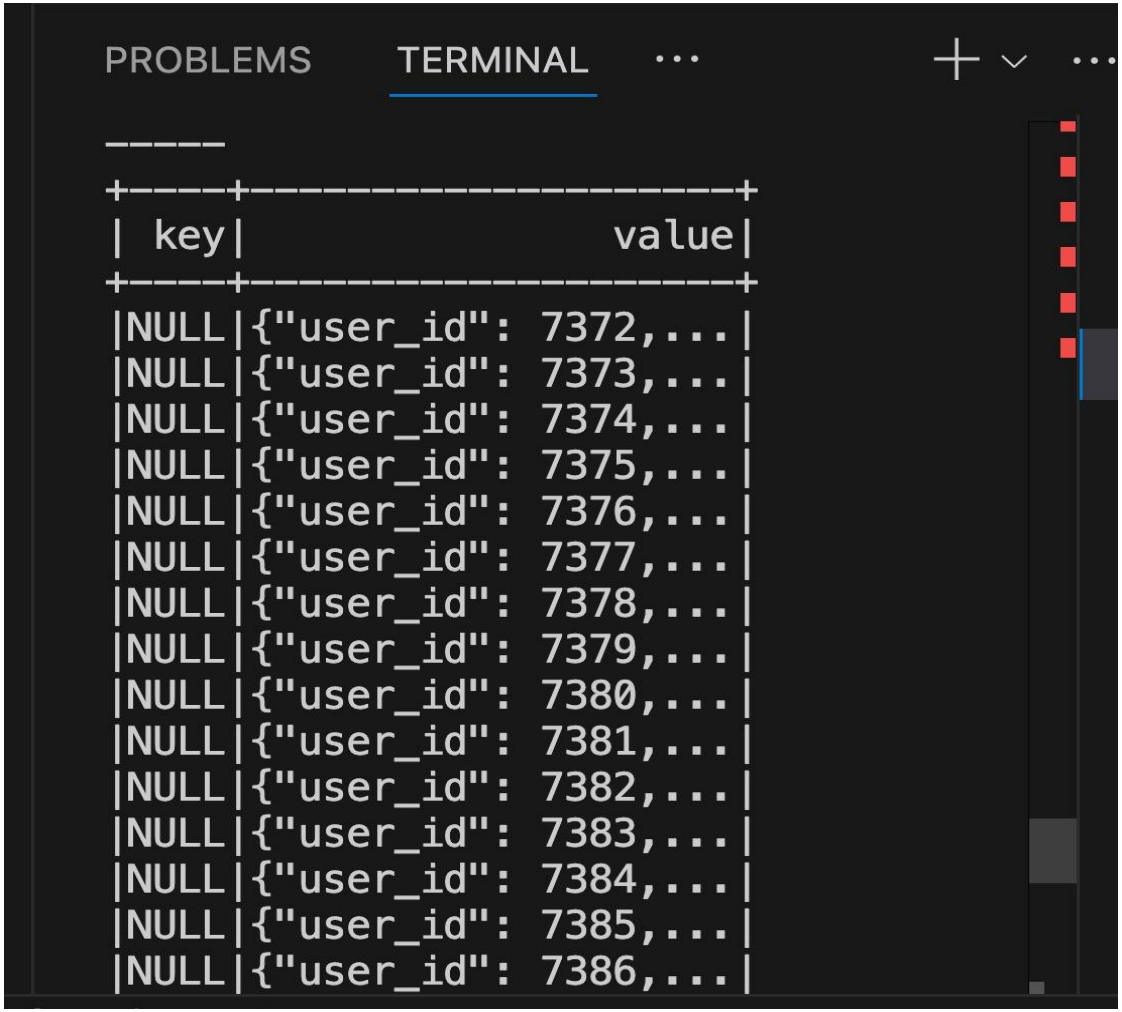
- Kafka: Streaming platform for real-time data ingestion.
- Spark Streaming: Processing and transforming data from Kafka.



```
 143, 'location': 'San Francisco', 'nights': 10, 'price': 301.73, 'room_type': '
Shared room'}
Sent event: {'user_id': 878, 'timestamp': '2025-01-26 23:14:03', 'action': 'canc
eled', 'cancellation_policy': 'moderate', 'guests': 1, 'host_id': 260, 'listing_
id': 357, 'location': 'San Francisco', 'nights': 1, 'price': 471.9, 'room_type':
 'Shared room'}
Sent event: {'user_id': 893, 'timestamp': '2025-01-26 23:26:23', 'action': 'canc
eled', 'cancellation_policy': 'strict', 'guests': 4, 'host_id': 62, 'listing_id'
: 368, 'location': 'Chicago', 'nights': 14, 'price': 121.43, 'room_type': 'Share
d room'}
Sent event: {'user_id': 769, 'timestamp': '2025-01-26 23:23:21', 'action': 'book
ed', 'cancellation_policy': 'strict', 'guests': 4, 'host_id': 6, 'listing_id': 2
29, 'location': 'Miami', 'nights': 1, 'price': 360.44, 'room_type': 'Private roo
m'}
Sent event: {'user_id': 893, 'timestamp': '2025-01-26 23:26:23', 'action': 'canc
eled', 'cancellation_policy': 'strict', 'guests': 4, 'host_id': 62, 'listing_id'
: 368, 'location': 'Chicago', 'nights': 14, 'price': 121.43, 'room_type': 'Share
d room'}
Sent event: {'user_id': 23, 'timestamp': '2025-01-26 23:12:08', 'action': 'cance
led', 'cancellation_policy': 'flexible', 'guests': 5, 'host_id': 74, 'listing_id
```

# Architecture Overview:

- Cassandra: NoSQL database for storing processed data.
- Streamlit: Frontend for live visualizations and insights.


PROBLEMS
TERMINAL
+----+--------------------+
| key|               value|
+----+--------------------+
|NULL|{"user_id": 7372,...|
|NULL|{"user_id": 7373,...|
|NULL|{"user_id": 7374,...|
|NULL|{"user_id": 7375,...|
|NULL|{"user_id": 7376,...|
|NULL|{"user_id": 7377,...|
|NULL|{"user_id": 7378,...|
|NULL|{"user_id": 7379,...|
|NULL|{"user_id": 7380,...|
|NULL|{"user_id": 7381,...|
|NULL|{"user_id": 7382,...|
|NULL|{"user_id": 7383,...|
|NULL|{"user_id": 7384,...|
|NULL|{"user_id": 7385,...|
|NULL|{"user_id": 7386,...|

# Data Flow

Kafka → Spark → Cassandra → Streamlit.

# Filtered Data

| | user_id | timestamp | action | cancellation_policy | guests | host_id | listing_id | location |
|---|---|---|---|---|---|---|---|---|
| 0 | 769 | 2025-01-26 23:23:21 | booked | strict | 4 | 6 | 229 | Miami |
| 1 | 23 | 2025-01-26 23:12:08 | canceled | flexible | 5 | 74 | 382 | New York |
| 2 | 114 | 2025-01-26 23:25:56 | canceled | moderate | 4 | 13 | 109 | Chicago |
| 3 | 660 | 2025-01-26 23:20:30 | booked | flexible | 1 | 214 | 236 | New York |
| 4 | 893 | 2025-01-26 23:26:23 | canceled | strict | 4 | 62 | 368 | Chicago |
| 5 | 53 | 2025-01-26 23:12:14 | canceled | flexible | 4 | 125 | 377 | Chicago |
| 6 | 987 | 2025-01-26 23:12:19 | viewed | flexible | 2 | 47 | 303 | Chicago |
| 7 | 878 | 2025-01-26 23:14:03 | canceled | moderate | 1 | 260 | 357 | San Francis |
| 8 | 110 | 2025-01-26 23:18:57 | viewed | strict | 3 | 110 | 408 | San Francis |
| 9 | 91 | 2025-01-26 23:16:47 | canceled | strict | 4 | 14 | 143 | San Francis |

# Examples and Key Insights

- Graphs:

Action counts (booked, canceled, viewed) across locations.

  - Average price by room type.
  - Booking trends over time.
  - Distribution of room types by action.

- Insights:
  - Popular room types.
  - Locations with the highest cancellations.
  - Top hosts by bookings.

# Conclusion and Future Work

- Summary:
- • Real-time analytics provide va
  platforms like Airbnb.
- Future Improvements:
- • Expanding dataset for higher
- • Integrating machine learning
  actions.
- • Scaling architecture for larger