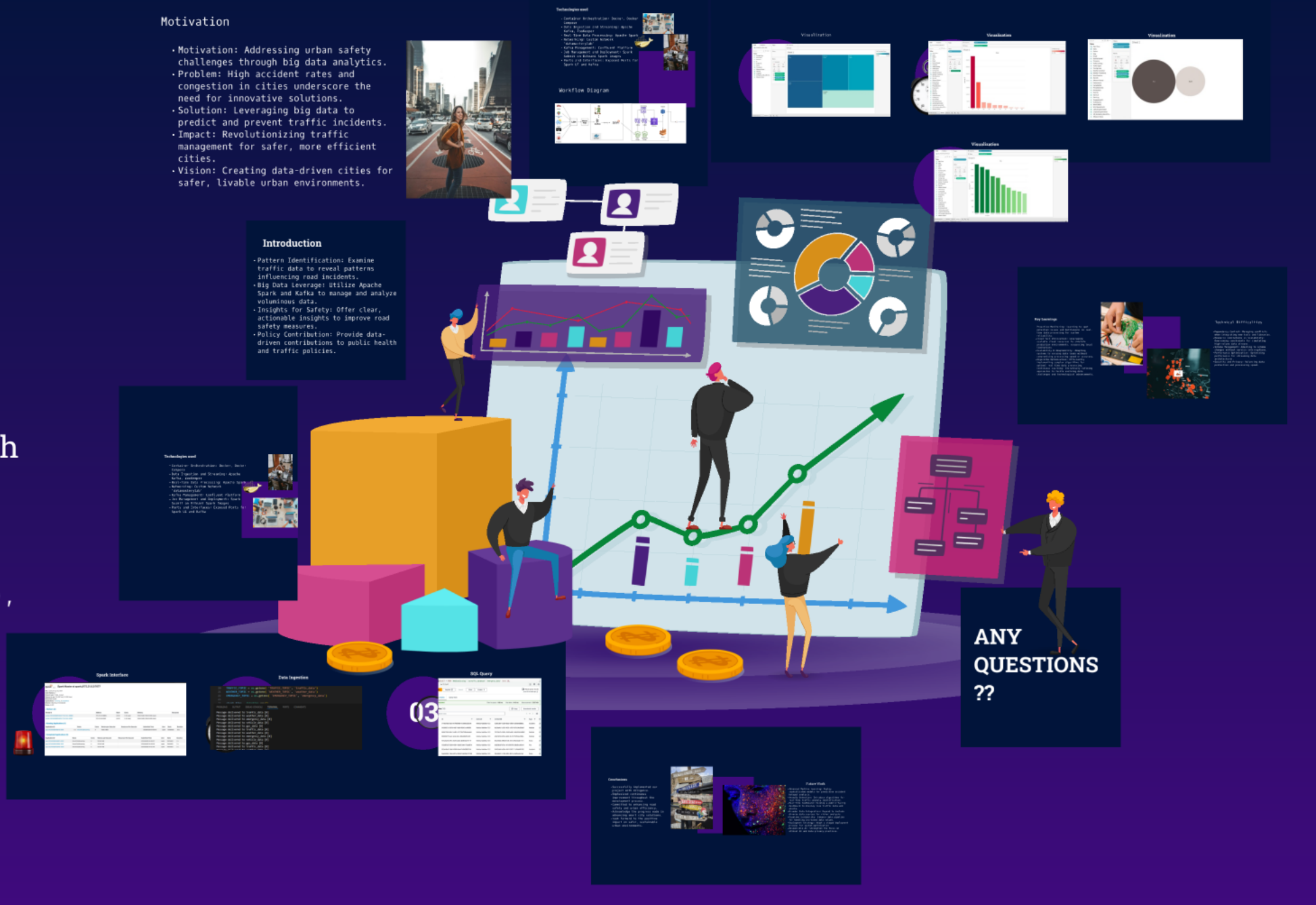
Beyond the Crash: Understanding US Accidents with Big Data

Data – 228 big data tech and app

Presented By:
Garima Singh (017428788),
Ishita Upadhyay (017431349),
Mukul Mahajan (017406701),
Vaibhav Shete (017443907)



Introduction

- Pattern Identification: Examine traffic data to reveal patterns influencing road incidents.
- Big Data Leverage: Utilize Apache Spark and Kafka to manage and analyze voluminous data.
- Insights for Safety: Offer clear, actionable insights to improve road safety measures.
- Policy Contribution: Provide data-driven contributions to public health and traffic policies.

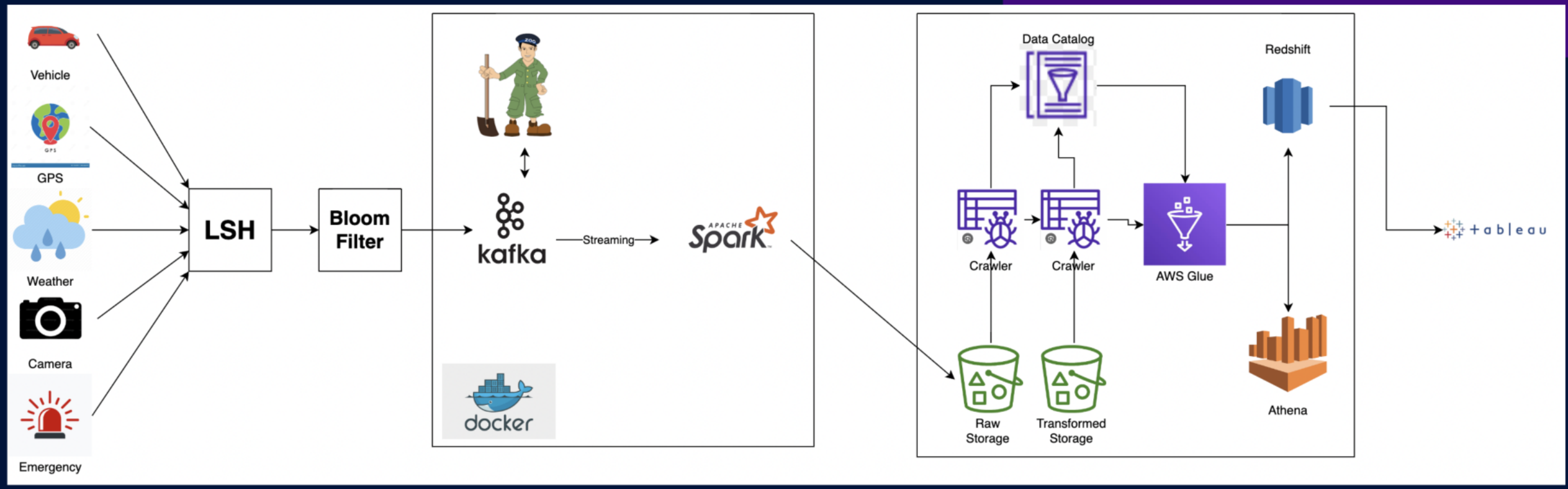
Motivation

- Motivation: Addressing urban safety challenges through big data analytics.
- Problem: High accident rates and congestion in cities underscore the need for innovative solutions.
- Solution: Leveraging big data to predict and prevent traffic incidents.
- Impact: Revolutionizing traffic management for safer, more efficient cities.
- Vision: Creating data-driven cities for safer, livable urban environments.

Technologies used

- Container Orchestration: Docker, Docker Compose
- Data Ingestion and Streaming: Apache Kafka, ZooKeeper
- Real-Time Data Processing: Apache Spark
- Networking: Custom Network 'datamasterylab'
- Kafka Management: Confluent Platform
- Job Management and Deployment: Spark Submit on Bitnami Spark Images
- Ports and Interfaces: Exposed Ports for Spark UI and Kafka

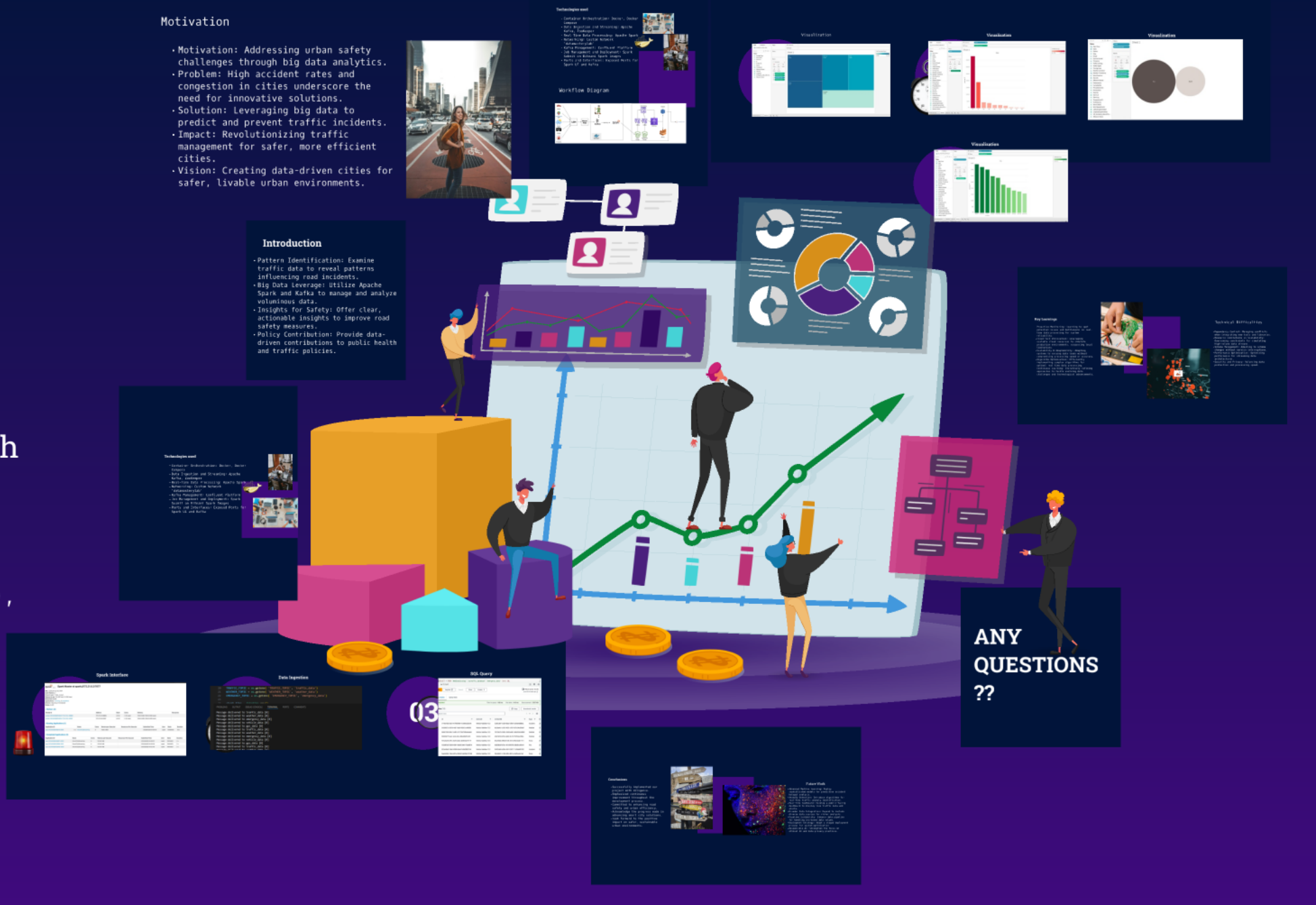
Workflow Diagram



Beyond the Crash: Understanding US Accidents with Big Data

Data – 228 big data tech and app

Presented By:
Garima Singh (017428788),
Ishita Upadhyay (017431349),
Mukul Mahajan (017406701),
Vaibhav Shete (017443907)



Spark Interface

Spark Master at spark://172.21.0.3:7077

URL: spark://172.21.0.3:7077
Alive Workers: 2
Cores in use: 4 Total, 4 Used
Memory in use: 2.0 GiB Total, 2.0 GiB Used
Resources in use:
Applications: 1 Running, 3 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240428014954-172.21.0.5-36865	172.21.0.5:36865	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	
worker-20240428014954-172.21.0.6-41907	172.21.0.6:41907	ALIVE	2 (2 Used)	1024.0 MiB (1024.0 MiB Used)	

Running Applications (1)

Application ID		Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240428045816-0003	(kill)	SmartCityStreaming	4	1024.0 MiB		2024/04/28 04:58:16	spark	RUNNING	33 s

Completed Applications (3)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240428044807-0002	SmartCityStreaming	4	1024.0 MiB		2024/04/28 04:48:07	spark	FINISHED	2 s
app-20240428043956-0001	SmartCityStreaming	4	1024.0 MiB		2024/04/28 04:39:56	spark	FINISHED	2 s
app-20240428043248-0000	SmartCityStreaming	4	1024.0 MiB		2024/04/28 04:32:48	spark	FINISHED	2 s

Data Ingestion

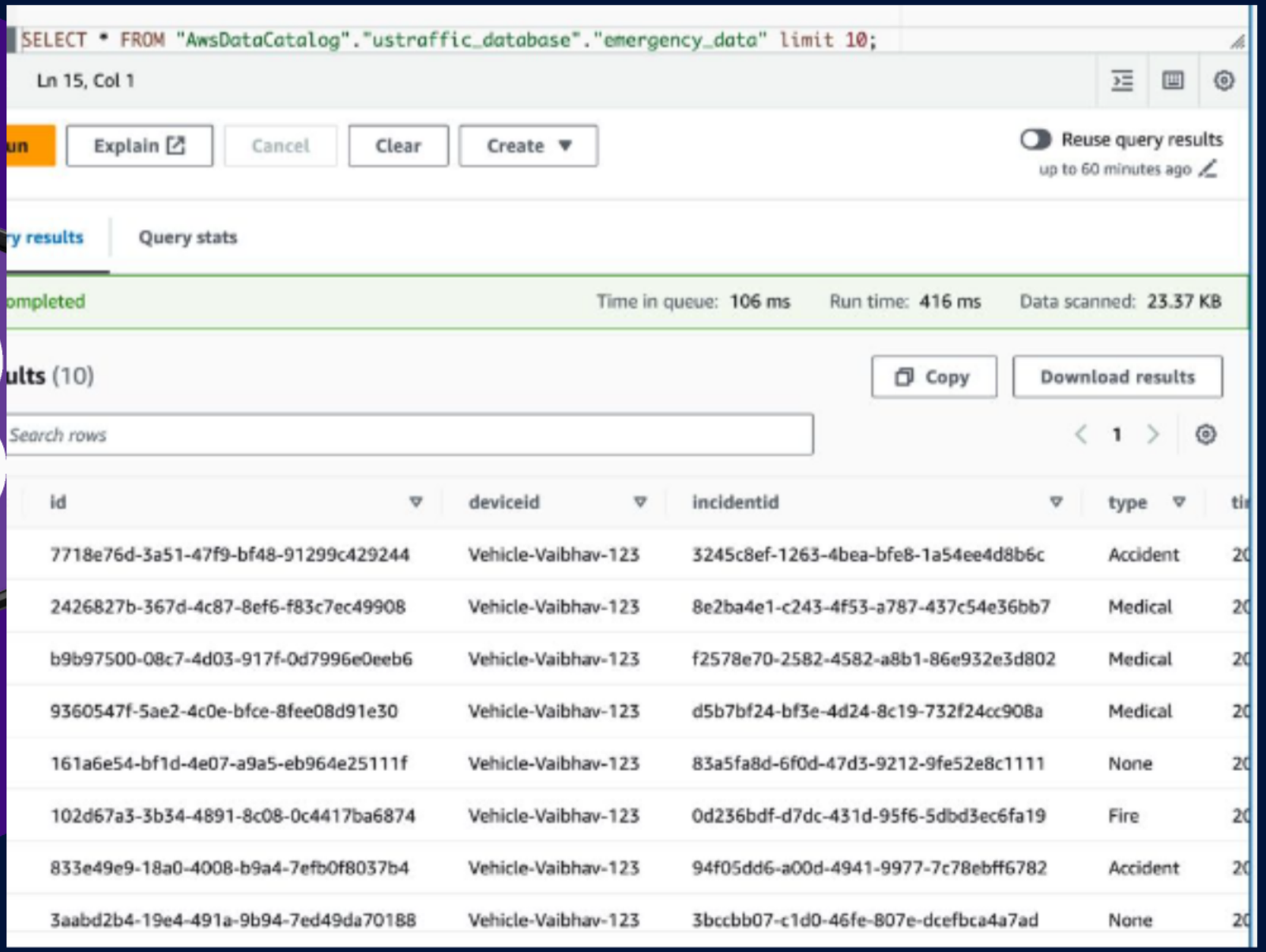
```
20  TRAFFIC_TOPIC = os.getenv( 'TRAFFIC_TOPIC', 'traffic_data')
21  WEATHER_TOPIC = os.getenv( 'WEATHER_TOPIC', 'weather_data')
22  EMERGENCY_TOPIC = os.getenv( 'EMERGENCY_TOPIC', 'emergency_data')
23
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS COMMENTS

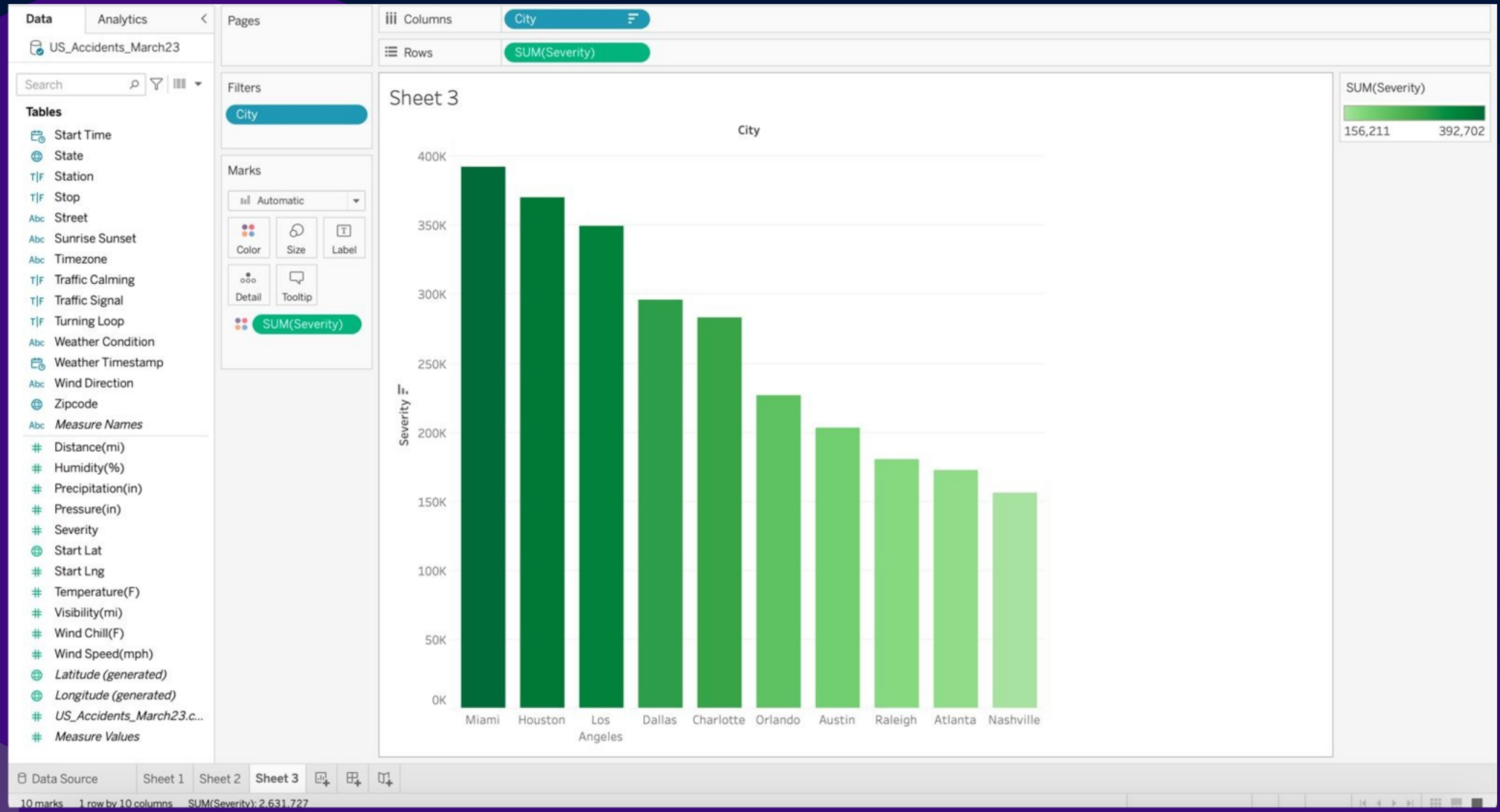
```
Message delivered to traffic_data [0]
Message delivered to weather_data [0]
Message delivered to emergency_data [0]
Message delivered to vehicle_data [0]
Message delivered to gps_data [0]
Message delivered to traffic_data [0]
Message delivered to weather_data [0]
Message delivered to emergency_data [0]
Message delivered to vehicle_data [0]
Message delivered to gps_data [0]
Message delivered to traffic_data [0]
```

SQL Query

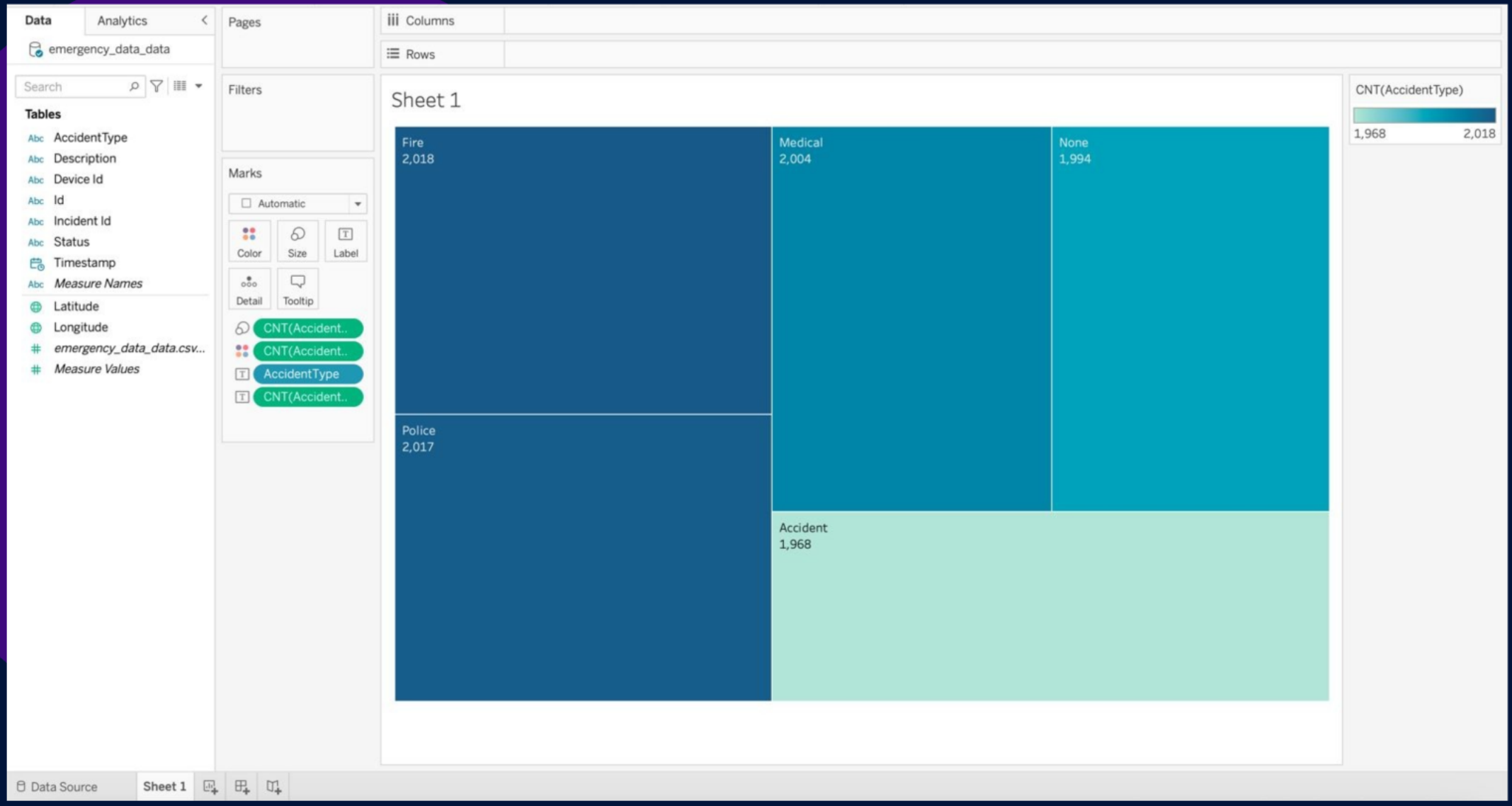
03


SELECT * FROM "AwsDataCatalog"."ustraffic_database"."emergency_data" limit 10;
Ln 15, Col 1
Explain
Cancel
Clear
Create
Reuse query results
up to 60 minutes ago
Query stats
Time in queue: 106 ms
Run time: 416 ms
Data scanned: 23.37 KB
Copy
Download results
Search rows
1
id
deviceid
incidentid
type
7718e76d-3a51-47f9-bf48-91299c429244 Vehicle-Vaibhav-123 3245c8ef-1263-4bea-bfe8-1a54ee4d8b6c Accident
2426827b-367d-4c87-8ef6-f83c7ec49908 Vehicle-Vaibhav-123 8e2ba4e1-c243-4f53-a787-437c54e36bb7 Medical
b9b97500-08c7-4d03-917f-0d7996e0eeb6 Vehicle-Vaibhav-123 f2578e70-2582-4582-a8b1-86e932e3d802 Medical
9360547f-5ae2-4c0e-bfce-8fee08d91e30 Vehicle-Vaibhav-123 d5b7bf24-bf3e-4d24-8c19-732f24cc908a Medical
161a6e54-bf1d-4e07-a9a5-eb964e25111f Vehicle-Vaibhav-123 83a5fa8d-6f0d-47d3-9212-9fe52e8c1111 None
102d67a3-3b34-4891-8c08-0c4417ba6874 Vehicle-Vaibhav-123 0d236bdf-d7dc-431d-95f6-5dbd3ec6fa19 Fire
833e49e9-18a0-4008-b9a4-7efb0f8037b4 Vehicle-Vaibhav-123 94f05dd6-a00d-4941-9977-7c78ebff6782 Accident
3aabd2b4-19e4-491a-9b94-7ed49da70188 Vehicle-Vaibhav-123 3bccbb07-c1d0-46fe-807e-dcefbca4a7ad None

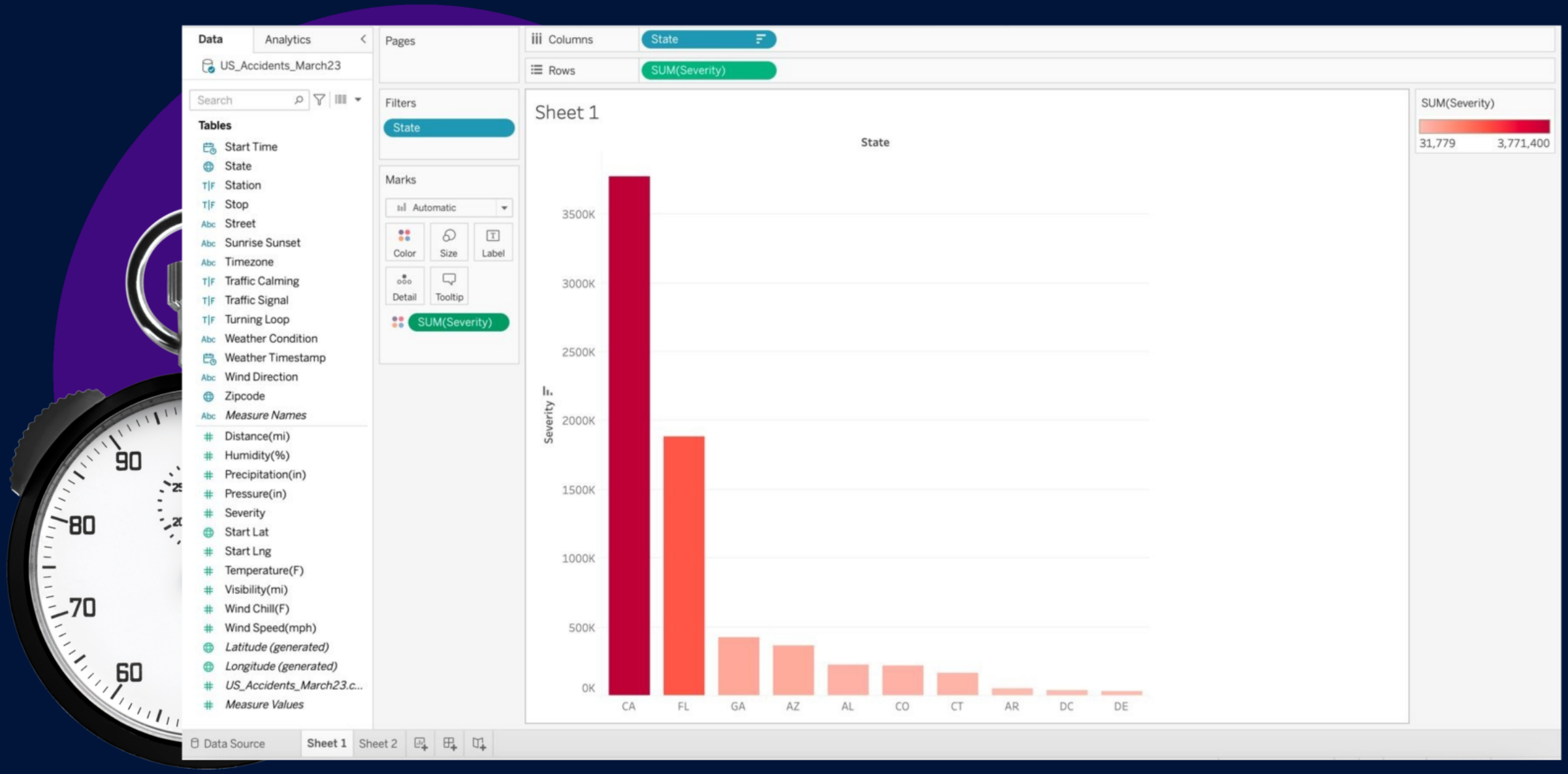
Visualization



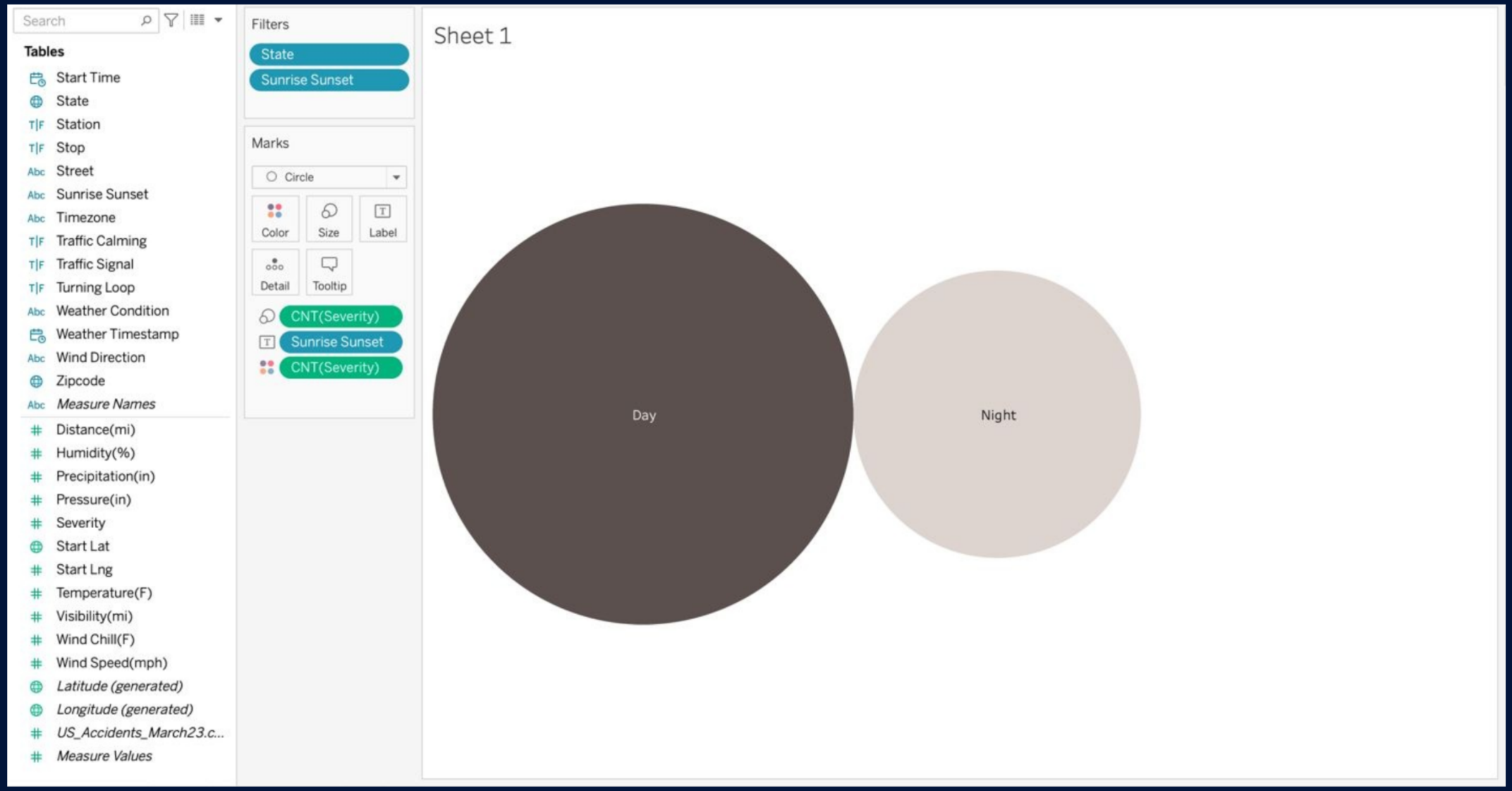
Visualization



Visualization

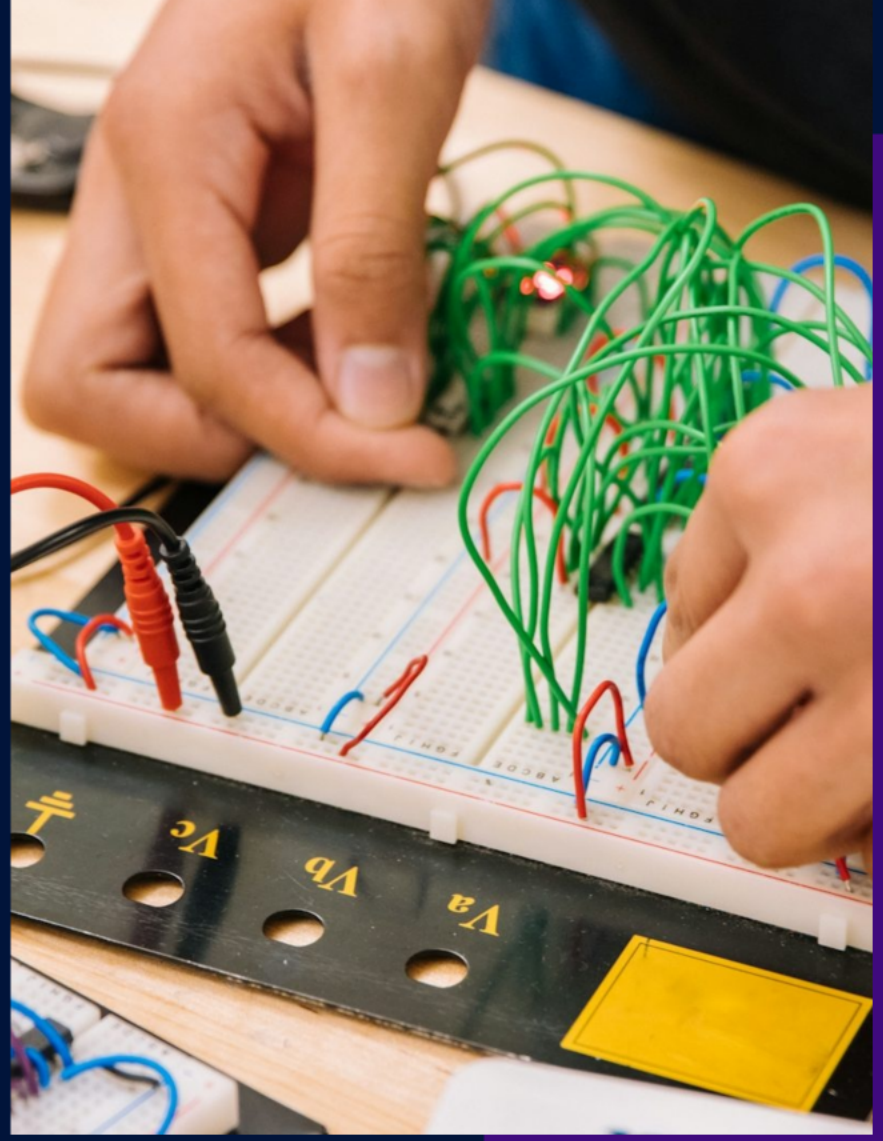


Visualization



Key Learnings

- Proactive Monitoring: Learning to spot potential issues and bottlenecks in real-time data processing for system reliability.
- Cloud Tech Utilization: Leveraging scalable cloud resources to simulate production environments, surpassing local limitations.
- Scalability & Adaptability: Adapting systems to varying data loads without compromising processing speed or accuracy.
- Algorithm Optimization: Efficiently implementing complex algorithms for optimal real-time data processing.
- Continuous Learning: Iteratively refining approaches to tackle evolving data challenges and technological advancements.



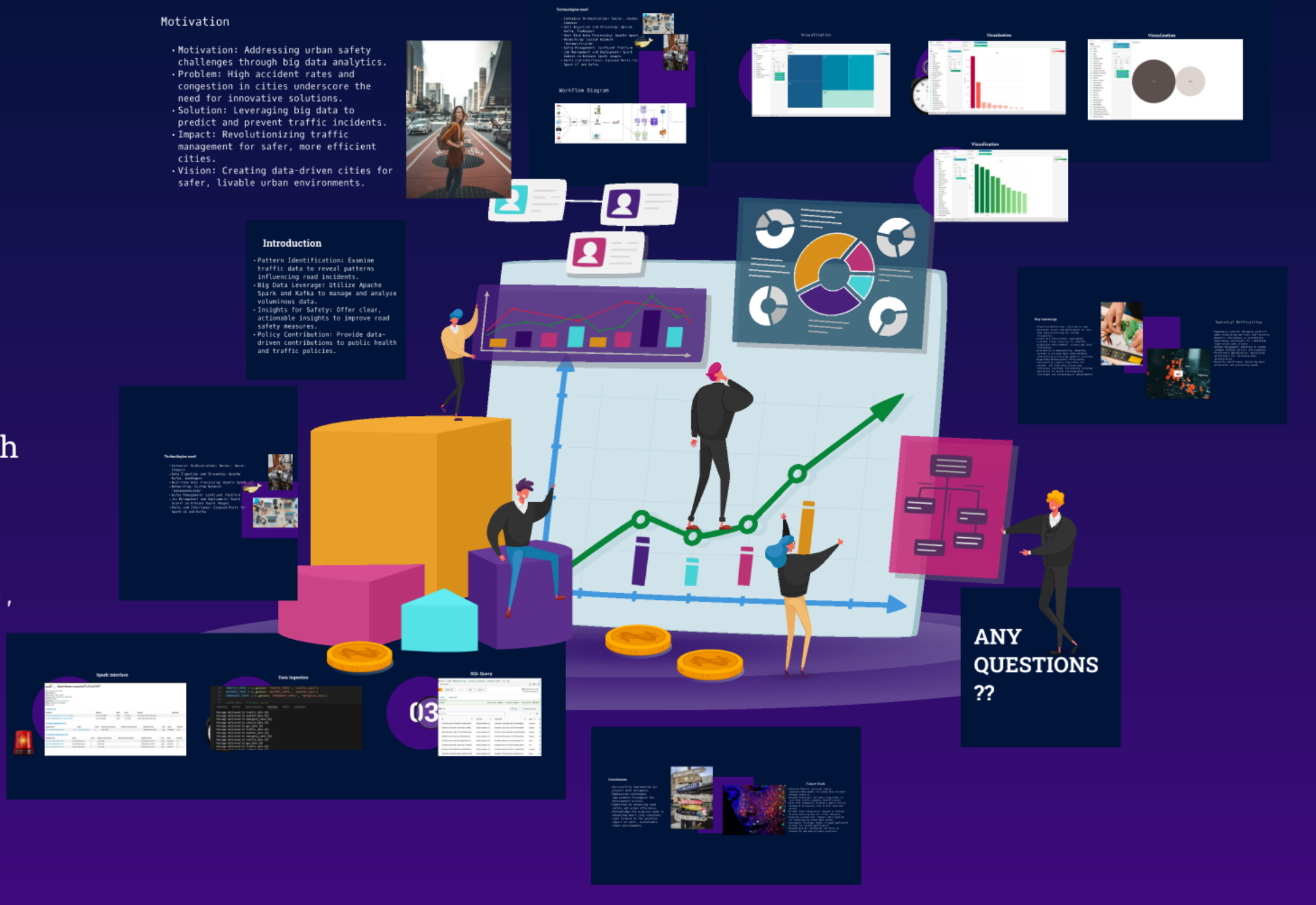
Technical Difficulties

- Dependency Control: Managing conflicts when integrating new tools and libraries.
- Resource Limitations vs Scalability: Overcoming constraints for simulating high-volume data streams.
- Schema Management: Adapting to schema changes without service interruptions.
- Performance Optimization: Optimizing performance for streaming data architectures.
- Security and Privacy: Balancing data protection and processing speed.

Beyond the Crash: Understanding US Accidents with Big Data

Data – 228 big data tech and app

Presented By:
Garima Singh (017428788),
Ishita Upadhyay (017431349),
Mukul Mahajan (017406701),
Vaibhav Shete (017443907)



Conclusions

- Successfully implemented our project with diligence.
- Emphasized continuous improvement throughout the development process.
- Committed to enhancing road safety and urban efficiency.
- Acknowledge the progress made in advancing smart city solutions.
- Look forward to the positive impact on safer, sustainable urban environments.

Future Work

- Advanced Machine Learning: Deploy sophisticated models for predictive accident hotspot analysis.
- Anomaly Detection: Introduce algorithms for real-time traffic anomaly identification.
- Real-Time Dashboard: Develop a public-facing dashboard to display live traffic data and alerts.
- Broader Data Integration: Expand to include diverse data sources for richer analysis.
- Pipeline Scalability: Enhance data pipeline for handling increased data volume.
- Deployment Strategy: Adopt a staged deployment process for system optimization.
- Responsible AI: Strengthen the focus on ethical AI and data privacy practices.


ANY
QUESTIONS
??