**Group Number:** 15

**Group Members:**

  Diveesh Sai Udayagiri, Sai Abhinav Alapati, Sai Venkat Choppara, Sriyuktha Sakhamuri

**Big Data Pipeline Planning Worksheet**

**Scenario: E-bike rental proof of concept**

Your startup e-bike company has partnered with a local government to pilot an e-bike rental program in suburban neighborhoods. The goal is to reduce carbon emissions and traffic congestion while collecting and analyzing usage data from IoT-equipped bikes. The proof-of-concept will demonstrate a pipeline to collect, process, and analyze this data.

**Section 1: Potential Data Sources**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Source** | **Format** | **Internal/Public** | **Description of Use** |
| IoT sensor data | JSON | Internal | Tracks bike speed, location, battery status |
| Rental transaction data | JSON | Internal | Logs bike rental start/end times, user ID |
| Weather API | JSON | Public | Provides weather conditions for trend analysis |
| Traffic congestion data | CSV/JSON | Public | Analyzes congestion patterns related to e-bike usage |
| User feedback data | JSON/Text | Internal | Collects user ratings, complaints, and comments |

**Section 2: Value, Veracity, and Variety Considerations**

**Value:**

* Identify peak usage times and areas with high demand.
* Predict maintenance needs based on battery and sensor data.
* Assess how weather impacts bike usage.
* Optimize bike distribution and rental pricing.

**Veracity:**

* IoT sensor data may have errors due to network latency.
* Rental transaction data is reliable but may contain user errors.
* Weather API data accuracy depends on third-party sources.
* Traffic congestion data may be outdated or incomplete.
* User feedback may be biased or inconsistent.

**Variety:**

* IoT sensor data in JSON format (real-time telemetry).
* Rental transaction data in structured JSON format.
* Weather data from APIs in JSON format.
* Traffic congestion data in CSV/JSON format.
* Unstructured text from user feedback.

**Section 3: Value, Veracity, and Variety - Detailed Considerations**

* **What types of analytics and processing might be needed to derive insights?**

 Machine learning for predictive maintenance, geospatial analysis for demand hotspots.

* **What queries could be made for analysis and visualization?**
  + Identify popular routes and rental frequency.
  + Detect patterns in battery performance and failures.
  + Analyze the impact of weather and traffic on rentals.
* **What privacy and security concerns need to be addressed?**
  + User data encryption and anonymization.
  + Secure transmission of IoT data.
* **What kind of veracity challenges need to be addressed?**
  + Sensor data noise and connectivity issues.
  + User-reported feedback inconsistencies.
* **What cleaning or transformations might be needed?**
  + Normalizing timestamps.
  + Filtering out duplicate or incorrect data.

**Section 4: Velocity and Volume Considerations**

* **What is the general volume of incoming data?**
  + Each of 50 bikes sends data every 2 minutes, generating significant real-time data.
  + Rental transaction logs grow with each ride.
* **How frequently is new data generated and ingested?**
  + IoT data updates in near real-time.
  + Rental data updates per transaction.
* **How might the data be transferred into your system?**
  + IoT data transmitted via cellular network.
  + Rental data stored in cloud-based databases.
* **Will the volume be steady or spiky?**
  + Higher data volume during peak commute times.
  + Low activity during night hours.
* **Does the data need to be processed in real time, or can it be analyzed periodically (e.g., nightly or weekly)?**
  + Real-time monitoring for bike health and availability.
  + Batch processing for historical trend analysis.

**Section 5: Velocity and Volume - Detailed Considerations**

* **Data Ingestion Cadence:**
  + IoT sensor data at a steady 2-minute interval.
  + Rental and user feedback data at irregular intervals.
* **Dataset Growth:**
  + Estimated growth: several GB per week.
  + Scalable cloud storage needed.
* **Impact of Inconsistent Data Patterns:**
  + Rental spikes during weekends and holidays require scalable infrastructure.
  + Unstable network connectivity could lead to missing sensor data.